UNIVERSITÉ DU
LUXEMBOURG

PhD-FSTC-2018-36
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defence held on 26/04/2018 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Dheeraj Reddy BOBBILI

Born on 28 July 1988 in Autonagar, (India)

UNRAVELING THE COMPLEX GENETICS OF NEUROLOGICAL DISORDERS

Dissertation defence committee

Dr. Rejko Krüger, dissertation supervisor
Professor, Université du Luxembourg

Dr. Enrico Glaab, Chairman
Senior Research Scientist, Université du Luxembourg

Dr. Reinhard Schneider, Vice-Chairman
Head of bioinformatics core facility, Université du Luxembourg

Dr. Federico Zara
Professor, Laboratory of Neurogenetics, Istituto Giannina Gaslini

Dr. Jose Antonio López Escámez
Hospital Universitario Virgen de las Nieves, Granada

Unraveling the complex genetics of neurological disorders

A dissertation

by

Dheeraj Reddy Bobbili

Completed in the

Bioinformatics Core Group, Luxembourg Centre for Systems Biomedicine

To obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG
EN *BIOLOGIE*

Dissertation Defence Committee:

Supervisor:	Prof. Dr. Rejko krüger
Committee members:	Dr. Enrico Glaab
	Dr. Reinhard Schneider
	Dr. Jose Antonio López-Escámez
	Dr. Federico Zara

2018

To my parents.

Declaration

I hereby declare that this dissertation has been written only by the undersigned and without any assistance from third parties. Furthermore, I confirm that no sources have been used in the preparation of this thesis other than those indicated herein.

Dheeraj Reddy Bobbili
Belval, Luxembourg
May 2, 2018

ACKNOWLEDGMENTS

I am thankful to Prof. Dr. Rejko krüger for being my supervisor. It is an honor to have you as my supervisor. You have been very supportive during my entire PhD. Your guidance and constructive ideas have helped me a lot during my PhD. I would like to thank Dr. Patrick May for his time and patience during the entire process of my PhD, right from hiring me to the end of my PhD. You have been a great inspiration and were always available whenever I needed you. You always believed in me and your enthusiasm towards research has motivated me to push myself harder. This work would not have been possible without you. Dr. Reinhard Schneider, I am very grateful to you for allowing me to be part of bioinformatics core and providing me with best of the facilities to conduct research. You are an inspiration, I have learnt a lot from you and it was a great opportunity to be part of your team.

Dr. Federico Zara, Dr. Jose Antonio López-Escámez and Dr. Enrico Glaab, thank you very much for taking time of your busy schedules to be part of my jury committee. I am deeply indebted to you for your efforts. I am extremely grateful to all my collaborators within and outside Luxembourg. Especially, Prof. Dr. Holger Lerche, Dr. Bernd A. Neubauer, Dr. Dennis Lal, Dr. Kamel Jabbari, Dr. Javier Simón-Sánchez and Anamika Giri for allowing me to access and work with the data. Thank you, Prof. Dr. Rudi Balling for being an inspiration not just for me, but to the entire LCSB. You have built a great place to work. I would like to thank all the colleagues from LCSB for your support during my entire period here.

A special thanks to the awesome biocore team, you are amazing and great people. During my PhD, each one of you have helped me one way or the other and I really appreciate your support. You guys have made biocore a wonderful place to work and It has been an amazing journey to work with you guys. Many of you are more than colleagues and have become good friends. Your suggestions in professional and personal life are invaluable. A special thanks to Venkata, Marek, Roland, Wei, Marie, Sarah, Kavita, Anshika, Anya, Joe, Susanne, Peter, Yohan, Noua, Adriano, Marx and Christophe. I would like to thank the High Performance Computing team of University of Luxembourg for their continuous support during the course of my PhD.

To all my good friends in Luxembourg, Aravind, Sunil, Aishwarya, Shaman, Maharshi and Serena. I cannot thank you guys enough for always being there for me, you made my life in Luxembourg a lifelong memory to cherish for. There are also many people in Luxembourg who helped me at various times, I owe them a lot. I thank all my friends back in India, each one of you have made a difference in my life and I have no words to express my feelings towards you. You guys never forget me and I am very lucky to have you guys as my friends.

To my extended family, especially Raghu anna thank you so much for always believing in me and at times the only person to support me to pursue my dreams. Geeta vadina, you were always welcoming and happy to see me. You guys were the closest relatives that I have here and I am really glad for that. Thanks to all my cousins and other family members who have been an enormous support and took care of amma whenever there is a need.

Finally, to my loving sister Chinni and her two beautiful kids ammu and duggu. You are the best sister, always positive and supportive even in difficult times. I am lucky to have you as my sister. Amma and Nanna, I have no words to thank you enough. It is only because of you, I am what I am today. Nanna, I miss you everyday of my life. You were a great inspiration and I always wish that you were with us today to see your kids achieving new heights in their lives. Amma, you are the strongest person that I ever met in my life and I feel blessed to be your son. Love you both.

ABSTRACT

Neurological disorders comprise a group of diseases that affect brain, spine and the nerves that connect them. Many people worldwide are affected by neurological disorders irrespective of their ethnicity, age or gender. Currently, >800 neurological disorders have been identified and the list is still growing. Parkinson's disease (PD) and epilepsy are two of the most common neurological disorders that cause a significant burden globally. It has been well established that in both PD and epilepsy, genetics play a significant role in the generation and progression of the disease, while both the diseases have a monogenic or polygenic origin. A review of literature shows that both PD and epilepsy are caused due to the symphony of common, rare and ultra-rare variants. However, there is a high degree of heterogeneity with regard to genetics, which is evident from the lack of confirmation from various studies and the minute overlap between linkage and genome wide association studies (GWAS). Hence, in order to understand the underlying mechanisms of disease generation and progress, there is a need for unification of results obtained from multiple studies and a multifaceted approach studying variants occurring with different allele frequencies. Advances in the field of next generation sequencing (NGS) provided us an opportunity to identify and characterize the genetic variants associated to a disease more efficiently. Two of such useful techniques are whole genome sequencing (WGS) and whole-exome sequencing (WES) where the DNA of an individual is sequenced to identify disease-causing variants. In this thesis, we aimed to uncover the role of rare/ultra-rare variants in PD and epilepsy, using WGS/WES data. To achieve the aforementioned goal, state-of-the-art bioinformatic tools and statistical methods were used on various datasets generated by different studies.

The work on epilepsy was divided into three parts. First, the burden analysis of rare variants in typical rolandic epilepsy (RE) and atypical rolandic epilepsy (ARE) was conducted, where we observed an increased burden of rare loss of function (LoF) variants across several disease genesets in RE/ARE cases. Whereas, in the second part, burden analyses of rare variants in several genesets were conducted in genetic generalized epilepsy (GGE). A significant burden was observed for rare nonsynonymous variants in GABA_{A} receptors in the discovery cohort. Furthermore, the observed burden was replicated in two independent datasets, of which one was a WES study while the other was a targeted panel sequencing of GABA_{A} receptor genes. From the identified variants in GABA_{A} receptors in the discovery cohort, selected variants were functionally validated. Third, in RE/ARE and GGE, analysis of rare copy number deletions showed a significant burden and several novel candidate genes were identified. In PD, firstly private variants in the Parkinson's Progression Markers Initiative (PPMI) dataset were studied, where we observed a genome-wide burden of private LoF variants and prediction models were constructed based on the burden score. Second, the genome-wide burden of U1 splice variants was observed in the PPMI dataset and the observation was confirmed in the Parkinson Disease Genetic Sequencing Consortium (PDGSC) dataset. Finally, we discovered several rare, novel variants (coding, non-coding and CNVs) belonging to multiple families from two familial PD studies (>50 families) that were segregating with PD. Altogether, this work demonstrates the utility of NGS in discovering novel genes and genesets found to be implicated in PD and epilepsy and show their heterogeneous contribution to the disease aetiology. These discoveries could improve the diagnostics of both PD and epilepsy by expanding the knowledge of molecular mechanisms underlying the disease and potentially help in establishing modern therapeutic applications.

SCIENTIFIC OUTPUT

Majority of this thesis has been adapted from the work that has either been published, is currently under peer-review and/or ready for submission with the candidate as the first/co-first author. All the projects in the thesis have been performed along with several collaborators and part of consortia and hence only the manuscripts with a major contributions have been incorporated as part of this thesis. In addition, the candidate has also co-authored several publications which are not included in the main part of the thesis. The full list of scientific outputs is listed in sections below. The original manuscripts are provided in the appendix.

Manuscripts used in this thesis

- **Dheeraj R. Bobbili***, Dennis Lal*, Patrick May* et al., Exome-wide analysis of mutational burden in patients with typical and atypical Rolandic epilepsy. *European Journal of Human Genetics*, page 1, January 2018 (*Published*)
- Patrick May*, Simon Girard*, Merle Harrer*, **Dheeraj R. Bobbili***, Julian Schubert*, Stefan Wolking* et al., Rare coding variants in $GABA_A$ receptor sub-unit encoding genes increase the risk for genetic generalized epilepsies. *Lancet Neurology*, January 2018 (*Under review*)
- Kamel Jabbari*, **Dheeraj R. Bobbili***, Dennis Lal, Eva M. Reinthaler, Julian Schubert, Stefan Wolking et al., Rare gene deletions in genetic generalized and Rolandic epilepsies. *PLOS ONE*, March 2018 (*Accepted*)
- **Dheeraj R. Bobbili***, Peter Banda*, Rejko Krüger, Patrick May, Excess of singleton loss-of-function variants in Parkinson's Disease. *Movement Disorders*, March 2018 (*Submitted*)
- Ibrahim Boussaad, Carolin Obermaier, Zoé Hanss, Enrico Glaab, **Dheeraj R. Bobbili**,

Katarzyna Wicher et al, A patient-based model of RNA mis-splicing uncovers novel targets in Parkinson's disease. *Cell*, March 2018. (*Submitted*)

- **Dheeraj R. Bobbili** et al., WGS of multiplex families reveals interesting candidate genes. (*In preparation*)

Manuscripts from other projects

- Patrick May, Sabrina Pichler, Daniela Hartl, **Dheeraj R. Bobbili**, Manuel Mayhaus, Christian Spaniol et al., Rare *ABCA7* variants in two German families with Alzheimer's disease. *Neurology:Genetics*, February 2018 (*Published*)
- Daniela Hartl, Patrick May, Wei Gu, Manuel Mayhaus, Sabrina Pichler, Christian Spaniol, Enrico Glaab, **Dheeraj R. Bobbili** et al., A rare loss-of-function variant of *ADAM17* is associated with familial Alzheimer disease. *Molecular Pyschiatry*, January 2018 (*Accepted*)
- Julia C. Fitzgerald, Alexander Zimprich, Daniel A. Carvajal Berrio, Kevin M. Schindler, Brigitte Maurer, Claudia Schulte, Christine Bus, Anne-Katrin Hauser, Manuela Kübler, Rahel Lewin, **Dheeraj R. Bobbili** et al., Metformin reverses *TRAP1* mutation-associated alterations in mitochondrial function in Parkinsons disease. *Brain*, 140(9):2444–2459, September 2017 (*Published*)
- Aishwarya Alex Namasivayam, Alejandro Ferreiro Morales, Ángela María Fajardo Lacave, Aravind Tallam, Borislav Simovic, David Garrido Alfaro, **Dheeraj R. Bobbili** et al., Community-Reviewed Biological Network Models for Toxicology and Drug Discovery Applications, *Gene regulation and systems biology* 2016; 10: 51–66 (*Published*)

Posters in international conferences

- Complex Genetics of Idiopathic Epilepsies (2015). *Big data in healthcare*, Luxembourg.
- Complex genetics of Parkinson's disease (2016). *GeoPD*, Luxembourg.
- Identification of genetic risk factors for Parkinson's disease (2016). *Life Science PhD days*, Luxembourg.
- Rare variant analysis of the PPMI dataset to uncover the complex genetic architecture of Parkinson's disease (2017). *Movements Disorders Society congress*, Vancouver.
- Methodical comparison of variant calling pipelines for next generation sequencing data. (2017). *European Society of Human Genetics*, Copenhagen.

** equal contributions*

Contents

Abstract iii

Scientific output v

List of Figures xiii

List of Tables xv

1 Introduction 1

- 1.1 Human Genome 1
- 1.2 Genetic code and variants 2
  - 1.2.1 Single Nucleotide Variants (SNVs) 3
  - 1.2.2 Insertion and Deletions (INDELs) 4
  - 1.2.3 Structural variants (SVs) 4
- 1.3 Technologies to detect genetic variants 5
  - 1.3.1 Sanger sequencing 5
  - 1.3.2 Microarrays 6
  - 1.3.3 NGS 6
- 1.4 Processing of NGS data 8
  - 1.4.1 Pre-processing and variant detection 9
  - 1.4.2 Variant Quality Control (QC) 12
  - 1.4.3 Sample QC 13
  - 1.4.4 Functional Annotation 14
  - 1.4.5 CNV detection 16
  - 1.4.6 QC and prioritization of CNVs 18
  - 1.4.7 Association testing 18
  - 1.4.8 Mode of Inheritance (MOI) filtering 19
- 1.5 Neurological disorders 19
- 1.6 Parkinson’s Disease 22
  - 1.6.1 Background 22
  - 1.6.2 Idiopathic PD 23

1.6.3 Mendelian/Familial PD . . . 24
1.6.4 Genetics . . . 26
1.7 Epilepsy . . . 27
1.7.1 Background . . . 27
1.7.2 Genetics . . . 29
1.7.3 Rolandic Epilepsy . . . 32
1.7.4 Genetic generalized epilepsy or Idiopathic generalized epilepsy (IGE) . . . 33
1.8 Aims of the thesis . . . 35
1.9 Contributions . . . 35

2 Rolandic Epilepsy **37**
2.1 Abstract . . . 37
2.2 Introduction . . . 37
2.3 Patients and Methods . . . 38
2.3.1 Study participants . . . 38
2.3.2 Data generation and processing . . . 38
2.3.3 Variant annotation and filtering . . . 38
2.3.4 Single variant and gene association analysis . . . 39
2.3.5 Selection of gene-sets . . . 39
2.3.6 Gene-set association analysis: . . . 40
2.4 Results . . . 40
2.4.1 Exome sequencing and variant filtering . . . 40
2.4.2 Association analysis . . . 40
2.4.3 Exome-wide and gene-set burden analysis . . . 40
2.5 Discussion . . . 42

3 Genetic Generalized epilepsy **44**
3.1 Abstract . . . 44
3.2 Introduction . . . 45
3.3 Patients and Methods . . . 45
3.3.1 Participants . . . 45
3.3.2 Procedures . . . 47
3.3.3 Population stratification . . . 48
3.3.4 Statistical analysis . . . 48
3.3.5 Functional analysis . . . 49
3.4 Results . . . 49
3.5 Discussion . . . 54
3.6 Additional methods and results . . . 56

3.6.1 Patient cohorts . . . 56
3.6.2 Exome sequencing analysis (Discovery cohort) . . . 58
3.6.3 Exome sequencing analysis (EpiPGX, replication cohort 1) . . . 67
3.6.4 $\mathrm{GABA_A}$ receptor gene panel (replication cohort 2) . . . 69
3.6.5 Functional analysis . . . 70
3.6.6 Annotation of functional tested $\mathrm{GABA_A}$ receptor variants . . . 71

4 CNVs in epilepsy **75**
4.1 Abstract . . . 75
4.2 Introduction . . . 75
4.3 Patients and Methods . . . 77
4.3.1 Data . . . 77
4.3.2 Data pre-processing . . . 78
4.3.3 Coverage . . . 79
4.3.4 Variant calling . . . 79
4.3.5 Sample QC . . . 79
4.3.6 Variant QC . . . 79
4.3.7 Variant annotation . . . 80
4.3.8 CNV detection . . . 80
4.3.9 CNV filtering . . . 80
4.3.10 Burden analysis of large and rare deletions . . . 80
4.3.11 Case-only CNVs . . . 81
4.3.12 Validation of CNVs . . . 81
4.3.13 Compound heterozygous mutations and protein-protein interactions . . . 81
4.3.14 Gene-set enrichment analysis . . . 81
4.3.15 Over-representation analysis . . . 81
4.3.16 CNV tolerance score analysis . . . 82
4.3.17 Overlap with different databases . . . 82
4.4 Results . . . 82
4.4.1 Epilepsy-associated microdeletions . . . 82
4.4.2 Comparative analysis of Rolandic and GGE candidate genes . . . 82
4.4.3 Deletion burden analysis . . . 83
4.4.4 Enrichment for known epilepsy and autism-associated genes . . . 83
4.4.5 Deletions of brain-critical exons . . . 84
4.4.6 Exome Aggregation Consortium deletions . . . 85
4.4.7 Compound heterozygous and first order protein-protein interaction mutations . . . 85
4.4.8 Over-representation of gene-disease associations . . . 88

4.4.9 Protein-protein interaction network analysis 88
4.4.10 Enrichment for likely disruptive *de-novo* mutations 91
4.5 Discussion 91

5 Excess of singleton loss-of-function variants in Parkinson's Disease 94
5.1 Abstract 94
5.2 Introduction 95
5.3 Patients and methods 96
5.3.1 Low quality samples filter 96
5.3.2 Relatedness filter 97
5.3.3 Ethnicity filter 97
5.3.4 Variant QC 97
5.3.5 Variant annotation 97
5.3.6 Excess of singleton variants 98
5.3.7 Geneset burden analysis 98
5.3.8 PRS generation 98
5.3.9 Construction of risk models 100
5.3.10 Input Features and Classification Models 101
5.4 Results 102
5.4.1 Population stratification and relatedness check 102
5.4.2 Excess of rare singleton LoF variants 104
5.4.3 Evidence of polygenic burden in PD 105
5.4.4 Prediction Performance 105
5.5 Discussion and Conclusion 107

6 Burden analysis of U1 splice variants 109
6.1 Abstract 109
6.2 Introduction 109
6.3 Patients and Methods 110
6.3.1 Discovery cohort (PPMI) 110
6.3.2 Sample QC 110
6.3.3 Variant QC 111
6.3.4 Variant annotation and filtering 111
6.3.5 Generation of MaxEntScan score 111
6.3.6 Benchmarking of MaxEntScan score 111
6.3.7 Splice site burden tests 112
6.3.8 Replication cohort (PDGSC) 112
6.3.9 Sample QC 113

6.3.10 Variant QC 113
6.3.11 Variant detection and annotation 113
6.3.12 Burden testing 113
6.3.13 Multiple testing adjustment 114
6.4 Results 114
6.4.1 Benchmarking 114
6.4.2 Burden analysis 116
6.4.3 Discovery cohort (PPMI) 116
6.4.4 Replication cohort (PDGSC) 116
6.5 Discussion 117

7 Familial-PD **119**
7.1 Abstract 119
7.2 Introduction 119
7.3 Identification of novel genes involved in nervous system development by whole genome sequencing in PD 120
7.3.1 Patients and methods 120
7.3.2 Results 124
7.3.3 Discussion 132
7.4 Integrated analysis of WGS data reveals potential candidate genes 134
7.4.1 Patients and methods 134
7.4.2 Results 138
7.4.3 Discussion 141

8 Conclusions and outlook **143**
8.1 Conclusions 143
8.2 Outlook 146
8.2.1 Larger cohort sizes for increased statistical power 146
8.2.2 Polygenic risk score (PRS) 146
8.2.3 Regulatory, splice variants and genome-wide deleterious score 147

Bibliography **149**

List of abbreviations **201**

Appendices **204**

Appendix A Supplementary material **205**
A.1 Rolandic Epilepsy 205
A.2 Excess of singleton loss-of-function variants in Parkinson's Disease 206

A.3 CNVs in epilepsy 207
A.4 Familial-PD 215

Appendix B Articles **225**

List of Figures

1.1 The reducing cost of sequencing per sample 2
1.2 Types of major genetic variants 3
1.3 A schematic representation of the different steps performed in a typical NGS analysis. 8
1.4 Various steps used in CNV detection 17
1.5 Different patterns of variants associated to neurological disorders 22
1.6 A figure describing the 2017 classification of epilepsy according to ILAE 29

2.1 Burden analysis of RE/ARE 41
2.2 Gene-set burden across different variant classes 42

3.1 Gene-set burden testing 51
3.2 Characterization of *GABRB2* missense variants associated with GGE 53
3.3 Characterization of *GABRA5* missense variants 54
3.4 qqplot of nonsynonymous variants 61
3.5 qqplot of synonymous variants 62
3.6 Percentage of genome shared across each pair of samples 63
3.7 Population stratification of discovery cohort 64
3.8 Population stratification of replication cohort 1 68

4.1 Flowchart of the analysis steps 78
4.2 Network analysis of brain-expressed genes 91

5.1 Distribution of PRS in cases vs controls 99
5.2 Population stratification of the PPMI samples 103
5.3 Excess of singleton variants in cases 104
5.4 UPDRS score versus the UPSIT score of the samples in PPMI dataset 105
5.5 The AUC values of various models 107

6.1 Determination of cut-offs for wild_score and maxentscan_change 115
6.2 A forest plot representing the burden analysis across different variant classes . . 117

7.0 Pedigrees analyzed in the study 127

7.1 Top scoring network from IPA network analysis 132
7.2 Flowchart of Courage-PD data analysis 137
7.3 Pedigrees carrying the variants. 139

List of Tables

1.1 Prevalance of PD and epilepsy 21
1.2 PARK loci discovered till date 25
1.3 Ion-channel genes mutated in epilepsy 30
1.4 Non-ion channel genes in epilepsy 31

3.1 Discovery cohort: GGE diagnosis, phenotype data. 56
3.2 Replication cohort 1: Patients with GGE diagnosis, phenotype data. 57
3.3 Replication cohort 2 for $\mathrm{GABA_A}$ receptor gene panel sequencing: GGE diagnosis, phenotype data. 57
3.4 Statistics of discovery cohort after quality filtering 60
3.5 Gene-sets for gene-set burden analysis 66
3.6 Statistics of replication cohort 1 after quality filtering. 69
3.7 Statistics of replication cohort 2 after QC 70
3.8 $\mathrm{GABA_A}$ receptor gene rare missense variants 74
3.9 Overview of tested $\mathrm{GABA_A}$ variants 74

4.1 Burden test of large deletions 83
4.2 Genes overlapping with specific gene sets. 84
4.3 Pathway analysis of deletions. 89
4.4 Network analysis of deletions. 90

5.1 Summary statistics and predictive ability 100
5.2 Metrics of PPMI dataset after QC. 104

7.1 MOI per family 128
7.2 Odds ratios of deleterious genes 129
7.3 Final list of prioritized genes 131

A.1 Deleterious variants identified in RE 205
A.2 Summary statistics of SNPs used to generate PRS 206
A.3 Deletions detected in our study 207

A.4 Deletions common to ExAC CNVs 213
A.5 Deletions detected via WES data and validated by array data. 214
A.6 List of prioritized genes 215
A.7 Name of the family and the mode of inheritance that was tested. 216
A.8 Variants annotated as disease causing with high confidence by HGMD 224

CHAPTER 1

INTRODUCTION

1.1 Human Genome

In humans, the genetic information is encoded in the form of deoxyribonucleic acid (DNA) which was discovered by a Swiss biochemist named Friedrich Miescher in 1869. The building blocks of DNA are four nucleotides namely Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). The human genome is comprised of ~3 billion nucleotides. Of the four nucleotides, A and G belong to a group of nitrogenous bases called purines, whereas C and T belong to pyrimidines. The DNA has a double helix structure, which was first discovered by Watson and Crick. The double helical structure of DNA is formed by a specific bonding between the nucleotides (A->T and G->C). DNA is present in each cell of the human body and it occurs in the form of a tightly coiled structure known as chromosome.

Human cells typically contain 23 pairs of chromosomes, out of which 22 pairs are called autosomes while the 23rd pair of chromosomes determine the sex of an individual and are referred to as either allosomes or sex chromosomes. The autosomes remain the same between male and females, whereas in allosomes, females have two X chromosomes and males have one X and Y chromosome each. Any change in the number of chromosomes may lead to chromosomal aberrations. For example, Turner syndrome in females, where one of the X chromosomes is lost or abnormally formed. Similarly, there are other chromosomal abnormalities such as Down syndrome, Klinefelter syndrome etc.

The human genome can be broadly divided into two parts namely the coding and non-coding part. The coding part comprises a set of nucleotides known as exons. The exons are interspersed by the non-coding regions known as introns. Both exons and introns together form a gene and human genome comprises of ~20,000 genes. The regions between genes are also non-coding and are known as intergenic regions.

The landmark achievement in the history of human genome was The Human Genome Project [1] (HGP), where the entire sequence of human genome was decoded in 2001. It changed the way scientists identify genetic causes of diseases in a minimal amount of time compared to the pre-HGP period. The HGP allowed us to compare the genomes of different species and helped us understand the conserved and key regions. The HGP is beginning to dramatically affect the way drugs are prescribed by enabling the prediction of side effects or benefits of a given drug on individuals. It took nearly 13 years and ~$2.7 billion for the project to complete, and since then the cost of sequencing has continued to reduce considerably. Current trends indicate that the price for sequencing a human genome is approaching ~$1000 as shown in Figure 1.1.

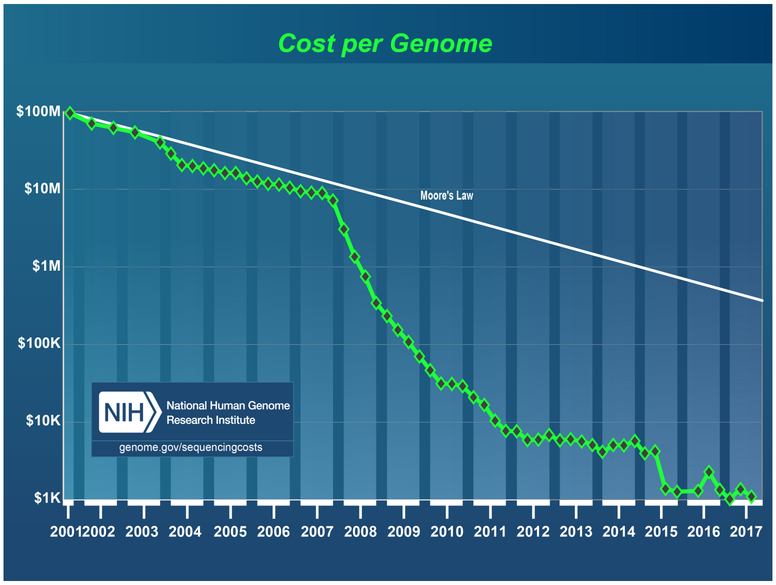


Figure 1.1: The reducing cost of sequencing per sample. The image was downloaded from `https://www.genome.gov/27541954/dna-sequencing-costs-data/`

1.2 Genetic code and variants

Genetic code is a collection of rules followed by the living cells to decipher the information from DNA to proteins. The conversion of DNA to protein occurs in two steps namely: 1) Transcription: DNA is transcribed to messenger RNA (mRNA) and 2) Translation: mRNA is converted to protein. During the process of translation, there are six possible ways to read a nucleotide sequence called reading frames and the resulting protein depends on the choice of reading frame. The reading frame divides information from mRNA into sets of three consecutive and non-overlapping nucleotides called codons. In total, there is a possibility of having 64 codons based on the permutation of four nucleotides (A, U, G and C). Out of 64 codons, 61 code for an amino acid whereas the remaining three act as stop signal for the translation and hence called “stop codons”. During translation, the codons from mRNA are read in a particular order beginning from “start codon” which is typically AUG and it codes for an amino acid named methionine. Human genetic code follows a principle called “degeneracy” which was first discovered by Lagerkvist [2]. According to this principle, human genetic code is redundant because multiple codons can code for the same amino acid, but there is no ambiguity as they always code for the same amino acid. Due to the degeneracy in human genetic code there are 61 codons, while they only code for 20 amino acids.

The term genetic variation refers to changes occurring in the DNA. Genetic variation occur as a result of errors during the process of DNA replication. It is an important driver of the evolution, as genetic variation inherits from one generation to the other. Genetic variation is what makes an organism unique, be it in humans or other organisms. Mutations are the irreversible changes occurring in the DNA and they are one of the major sources of genetic variation along with the recombination. An inverse relationship between the minor allele frequency (MAF) and the effect size of a mutation has been observed previously [3]. Hence, if a mutation is common it is assumed to be less harmful, on the other hand if it is rare it could be harmful and lead to a disease. As a result, for a mutation that is rare, a term “variant” is used instead of “mutation”. The major types of genetic variants are shown in the Figure 1.2 and are briefly described below.

Reference	**TTATTTCAACACACACAAAAAAAGTGTATATGCTCCACGATGCCTG**
Single nucleotide variants (SNVs)	**TTATTTCAACACACACAAAAAAAGTTTATATGCTCCACGATGCCTG**
Deletions	**TTATTTCAACACACACAAAAAA-------------CTCCACGATGCCTG**
Insertions	**TTATTTCAACACACACAAAAAAAGTTTGCCTGTATATGCTCCACGATGCCTG**
Copy number variants (CNVs)	**Large deletions or duplications >1kb**

Figure 1.2: Types of major genetic variants. They are shown with respect to the reference genome.

1.2.1 Single Nucleotide Variants (SNVs)

SNVs are single nucleotide changes in the DNA strand compared to a reference genome. If a SNV is common (occurs in >1% of the population), it is called a single nucleotide polymorphism

(SNP). Based on their functional effect, SNVs can be further classified into various functional categories as described below.

- **Coding variants:** As the name suggests, coding variants originate from the protein-coding regions of the genome. Based on the outcome of whether or not a given variant results in a change in amino acid, they can be broadly/further classified as follows:
    - **Synonymous variants:** According to the concept of degeneracy, multiple codons code for the same amino acid. Hence, if a SNV produces a new codon which codes for the same amino acid, then the translation process will go on normally and there will be no change in the protein production. Such variants are called synonymous variants and these are the variants that are assumed to be functionally neutral. But, certain synonymous variants can also be disease causing [4] based on their functions, for instance synonymous variants involved in splicing.
    - **Nonsynonymous variants:** However, if a SNV changes the amino acids of resulting proteins they are termed nonsynonymous variants. They are further subdivided into missense variants, where they just change the amino acid and nonsense variants where they cause a premature gain or loss of stop codon.
- **Non-coding variants:** These are the SNVs occurring at the non-coding regions of the genome. Their functional importance is vastly unknown and are currently a major focus of research in the field.

1.2.2 Insertion and Deletions (INDELs)

Insertions and deletions which are collectively called INDELs are small insertions or deletions occurring in the human genome with their size ranging from 1-10,000 bp [5]. They occur very frequently and are often detected along with the SNVs. Based on their effect on the reading frame, there are two major types of INDELs.

- **Frameshift:** An Indel is termed frameshift if the resulting change causes a shift in the reading frame during translation. The result of a frameshift Indel is that, the reading frame is not divisible by three anymore.
- **Non-frameshift:** These are INDELs which do not result in the shifting of reading frame and hence the reading frame length is divisible by three. They may cause amino acid insertions/deletions and might block the synthesis of proteins [6].

1.2.3 Structural variants (SVs)

Larger genomic alterations that are typically >1kb long are defined as SVs. There are several types of SVs, however compared to the SNVs, they are not studied extensively and currently represents and active area of research. SVs can be further classified into the categories described below:

Copy Number Variants (CNVs)

CNVs are the widely studied and common type of SVs. These are large insertions, deletions or duplications occurring in the human genome. CNVs are responsible for a considerable proportion of phenotypic variation [7] and they make up for ~12% of the human genome [8] and ~100 genes can be deleted completely without any supposed phenotypic effect [9]. Depending on their rate of occurrence they can be broadly divided into recurrent and non-recurrent CNVs. The probable reason for the generation of recurrent CNVs is homologous recombination among repeated sequences during meiosis. While, non-recurrent CNVs are generally caused by non-homologous mechanisms that arise in the entire genome and typically occur at sites with limited homology of 2 to 15 base pairs [10]. These CNVs can be either elementary, where a piece of DNA is eliminated from a position in the genome and the ends are merged, or they can be complex where a deletion is succeeded by a duplication or insertion of DNA.

CNVs vary in their size and based on a study comprising of a large collection of CNVs [9] , the mean lengths of copy number gains and copy number losses were 35,581 bp and 9,181bp respectively. CNVs can alter the expression of genes and induce phenotypic changes by varying the genome organization [11]. As a result, they can impact the susceptibility of a person to a particular disease or his/her response to a drug [12].

Not all the CNVs are disease causing in the human genome and based on their ability to cause a disease CNVs can be divided into various categories such as benign, likely benign, disease causing or CNVs of unknown significance [13]. CNVs are often linked to various complex and common nervous system disorders. Recently, there have been studies showing the role of CNVs in causing the diseases such as autism, schizophrenia and epilepsy [14, 15].

Inversions and other SVs

Inversions, as the name suggests are regions in the genome, where the DNA is reversed with respect to rest of the genome. Diseases caused by inversions include Hunter syndrome, Angelman syndrome, Sotos syndrome [16] etc. There are other SVs which include genomic translocations or segmental uniparentral disomy [16]. They are relatively rare and hence not well studied.

1.3 Technologies to detect genetic variants

1.3.1 Sanger sequencing

Sanger sequencing is a well known method to sequence the DNA. It was first developed by the British biochemist named Frederick Sanger and his colleagues in 1977 [17]. In HGP, Sanger sequencing was used to decode the human genome. Since then, it has been applied in many studies successfully to identify the nucleotide sequences. However, it has a very low throughput and is expensive to perform it on large-scale compared to next-generation sequencing (NGS). Due to the efficiency reasons, Sanger sequencing has been replaced by NGS platforms. But still, it is widely used to validate the variants identified via NGS and considered to be the gold standard due to its lower error rate.

1.3.2 Microarrays

To date, microarrays especially SNP arrays are being widely used to identify common genetic variants associated to the diseases/traits via genome-wide association studies (GWAS). Microarrays can be used to find SNP or large SVs. The main advantage of the microarrays is that, the genotyping quality is high and they are economical. However, novel variants cannot be detected through this technology and hence, cannot be used in the context of detecting novel disease causing variants.

However, it is possible to design a customized microarray, adding more variants such as the NeuroX chip [18] from Illumina which is customized for neurodegenerative diseases. Thus, can be used to replicate novel variants identified via NGS technology in a larger population for instance. It had already been employed in several studies related to Parkinson's disease (PD) [19].

1.3.3 NGS

Instead of defining the variants of interest a priori, with the aid of NGS a high throughput DNA sequencing can be achieved much efficiently. NGS generates millions of sequences per run, thus allowing researchers to sequence and if needed to resequence at a much faster pace compared to the pre-NGS era. Now a days, generating the data is often not the problem as it has become very fast and affordable to perform NGS (see Figure 1.1). Today, there are several platforms which offer NGS services such as 454, Illumina, Qiagen, Ion Torrent (Thermo Fisher) and Oxford Nanopore. Each of them has their own proprietary technologies, but Illumina holds the biggest chunk in the NGS market by holding upto 70% of the market share. Various NGS technologies used to identify genetic variants are described below.

Targeted panel sequencing (TPS)

TPS is a technique where, only a subset of genes or regions of the genome are isolated by employing different methods. The commonly used method is solution hybridization, where the probes are used to pull down the regions of interest. Other methods include, enrichment by applying polymerase chain reaction (PCR), during which every targeted region is amplified by a specific primer pair in a multiplexed reaction. There are also other methods which employ a different procedure for PCR multiplexing and hybridization. Targeted analysis can comprise of the exome, specific genes of interest (can be customizable), targets within genes and/or mitochondrial DNA. Hence, targeted sequencing enables researchers to focus on specific areas of interest, thereby saving time on data analysis and cost, as it is cheaper than whole genome sequencing (WGS) and whole exome sequencing (WES).

WES

WES is the process of sequencing only the coding regions of a genome instead of the whole genome. It is a cheaper alternative to the WGS and is a widely accepted technique. The application of WES has been shown in numerous studies to identify the causal coding variants in diseases such as Amylotropic Lateral Sclerosis (ALS), brain defects etc [20]. The major drawback and the feature of the WES is that, it can only sequence the coding regions. However, it is currently a widely applied technology as there is limited knowledge regarding the function of non-coding regions.

WGS

WGS is a way of determining the entire DNA sequence of an individual. With the advent of Illumina X10 machines the 1000$ genome has now become a reality and many research groups/clinics across the globe have started using WGS for a wide range of diseases and traits. Though it is expensive to perform WGS compared to WES or TPS, WGS has many advantages compared to the other two technologies. The first being the ability to analyze the entire genome. Recently, several studies focusing on the non-coding regions are being conducted and WES or TPS cannot be applied in such context. The other advantage being, WGS provides an uniform coverage across the exome [21] compared to WES. As a result, WGS is better at detecting more exonic variants and high-quality CNVs compared to WES [22].

1.4 Processing of NGS data

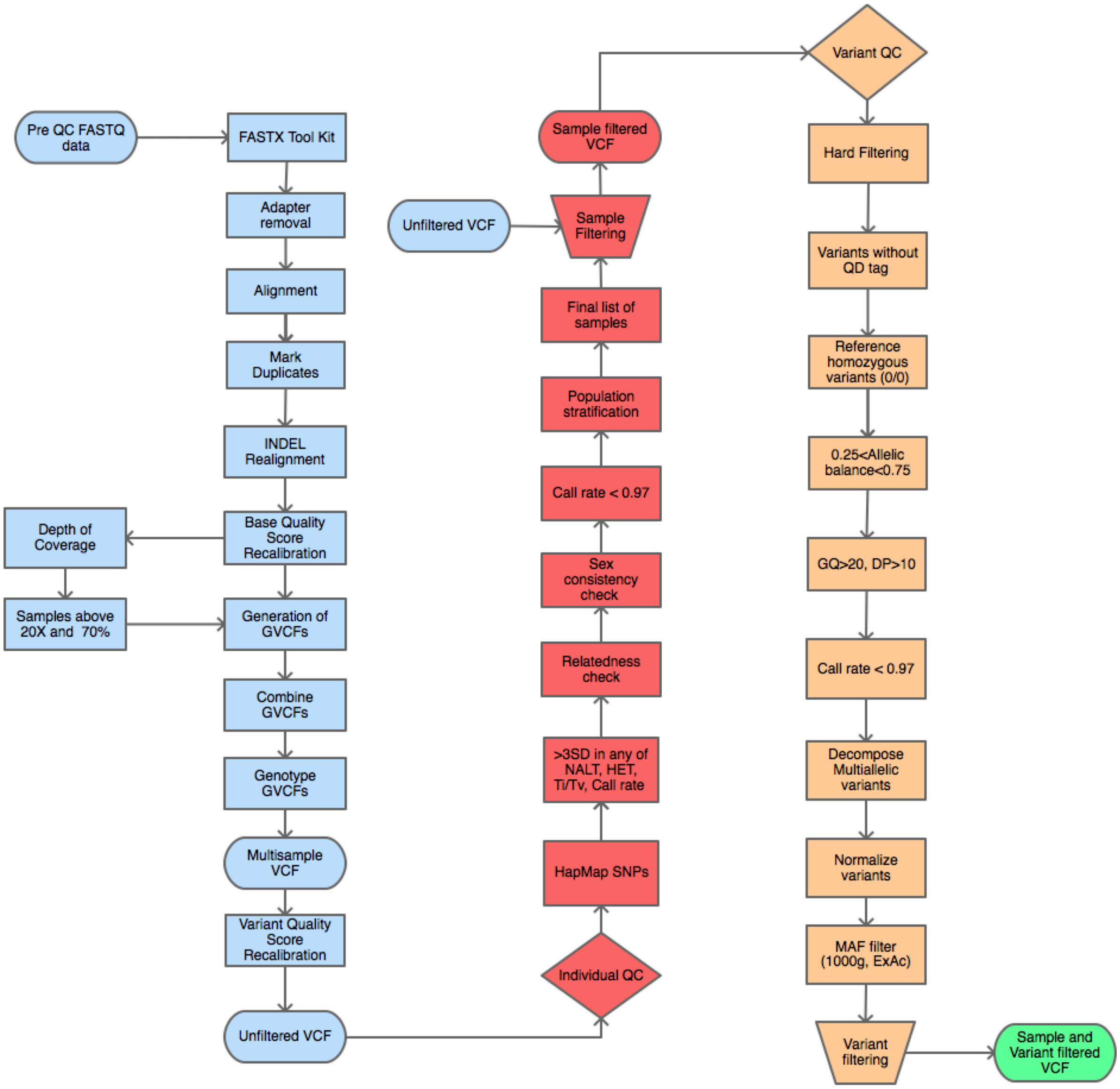


Figure 1.3: A schematic representation of the different steps performed in a typical NGS analysis.

The main goal of NGS data processing is to convert the raw sequencing reads to high-quality, annotated variant calls, in human readable format. The processing of human genome derived NGS data is both time consuming and computationally intensive, thus requiring high performance computing nodes (and facilities), especially in large scale (cohort) sequencing studies. There are multiple ways of processing NGS data, however in the recent years the best practices pipeline using Genome Analysis Toolkit (GATK) from the Broad institute has become a gold standard [23]. The main steps of NGS data processing are described below and shown in the Figure 1.3.

1.4.1 Pre-processing and variant detection

Quality Control (QC) of FastQ files

The high-throughput NGS platforms can generate millions of sequencing reads in a single run. The raw files obtained from the sequencers are called FastQ files which contain the stretches of short DNA sequences known as raw reads. NGS platforms generate two kinds of reads namely paired-end reads and single-end reads. Paired-end reads are generated by sequencing from both ends of the DNA and as a result, two FastQ files are generated per sample. Whereas, in the single-end sequencing DNA is sequenced only from one direction. In order to obtain meaningful and reliable biological results, one must ensure that the raw data is of high-quality, as biases exist within the data which may lead to unreliable results, affecting downstream analyses. Majority of the NGS vendors provide a summary report of their pipeline. However, it is specific to their proprietary pipeline and do not necessary reflect the quality of the data. Hence, a tool such as FastQC[24] provides a QC report which can detect and highlight the problems originating from the sequencing which are reflected in the quality of the data.

FastQC runs several tests on a FastQ file to generate a detailed QC report. FastQC assesses data quality by evaluating: read length, duplicated sequences, over-represented sequences, per sequence quality scores, nucleotide content, per base quality score and GC content. Based on the FastQC report one can setup the appropriate filtering steps for the downstream analyses. Contaminant oligonucleotide sequences such as, primers and adapters, can occur in both ends of NGS reads. These adapter sequences have to be removed as they may hinder correct mapping of the reads and influence the SNP calling and other downstream analyses. Two tools that are widely used for adapter removal are namely cutadapt [25] and Trimmomatic [26].

Mapping/Alignment

One of the crucial steps in performing the WES/WGS data analysis is the alignment of reads generated from the sequencer to the human genome. The outcome of read mapping may vary based on the read mapper that is used. Hence, it is crucial to choose a reliable mapper. A read mapper takes the FastQ files as input and produces either a sequence alignment map (SAM) or binary alignment format (BAM) file [27], based on the desired output. SAM/BAM format is a well accepted file format for storing the NGS data along with their mapping information. Two of the most widely used mappers are bowtie [28] and BWA-mem [29]. BWA-mem is part of the BWA suite of algorithms which uses Burrows-Wheeler algorithm to perform the read mapping. It is also the recommended aligner according to GATK best practices [23]. The algorithm is robust to sequencing errors and is shown to perform better compared to several other state of the art mappers. It is especially suited to the reads with a length of ~100bp, which is typically the read length generated by NGS platforms.

Removal of PCR duplicates

One major artifact of NGS procedures is the duplication of sequencing reads (defined as reads with the same start point and direction) generated as an artifact/effect of PCR. In an alignment scenario, the PCR duplicates tend to share the same DNA sequence and same alignment position. It is very important to identify and remove duplicate reads, as they may influence the downstream variant calling. The well known tool for this purpose is "mark duplicates" tool, which is part of the Picard suite (Picard `http://sourceforge.net/projects/picard/`). Other tools for this purpose include sambamba [30].

Picard tools are Java-based command-line tools to manipulate SAM or BAM files. It removes all the read pairs with identical coordinates, only retaining the pair with the highest mapping quality and examines aligned records in the supplied SAM/BAM file to locate duplicate molecules. In the end, it generates a SAM/BAM output file that includes all aligned reads without the duplicate records. Additionally, it also generates a file that contains information on the percentage of PCR duplicates found in the original aligned file.

INDEL Realignment

It is possible that even after read level and alignment level QC there might be some regions in the genome where the reads are misaligned due to various confounding factors such as the complexity in certain genomic regions. Mis-alignment in those regions may lead to the mismatch of many bases in those regions to the genome which might in turn lead to identification of those bases as SNPs. Also, there might me some regions where there is an Insertion or deletion in some reads, whereas the other reads might carry a SNP for the same position.

In order to mitigate this effect, one needs to correctly realign the reads in those regions. INDEL local realignment is recommended and can be performed by using the "IndelRealigner" tool from Genome Analysis Tool Kit [23] (GATK). Prior to using the "IndelRealigner" tools, a list of regions that require realignment has to be identified using the "RealignerTargetCreator" tool. These regions then undergo a local realignment which will alter the misaligned regions due to INDELs and are converted into higher quality reads containing a consensus INDEL, thereby increasing the reliability of downstream variant calling.

Base Quality Score recalibration (BQSR)

During the process of generation of FastQ reads, each base of the read is assigned a quality score generated on the phred-scale [31]. These scores estimate the errors generated by the sequencers. The scores generated by the sequencers are subject to various technical errors, leading to either over or under estimated base quality scores. GATK's BQSR [32] is one step where a machine learning model is applied to estimate the errors empirically and adjust the quality scores. Variant calling algorithms depend heavily on the per base quality score while identifying SNPs and INDELs. Hence, it is very important to adjust the quality scores in order to perform a reliable variant calling. This process of empirical adjustment allows users to obtain more accurate quality scores per base. The BQSR process is done in two major steps. First, a model is built based on a list of known variants and assumes that all the mismatches are errors and thus generates the estimates. In the second step the model is applied to all the variants and the variant quality scores are adjusted empirically.

Variant calling

The input for a variant caller is a SAM/BAM file and the output is a variant call format (VCF) file [33]. VCF is a generic file format that allows to store the information about genetic variants such as SNPs, INDELs and SVs along with their functional annotations. VCF is typically stored as a compressed file and can also be indexed to obtain the information quickly by providing a range of positions in the genome. The format was initially developed for the 1000 Genomes Project [34] and since then, it has been widely accepted for many studies. A typical VCF file contains a header line, the meta-information about the various steps employed in the variant calling followed by the information about genomic position and its respective genotypic information. There are two ways to detect a variant from the NGS data. The first is using one sample at a time and performing the variant calling. The most widely used tools for this step are samtools [27] and the unified genotyper from GATK. The main drawback of this approach is that, in the studies involving multiple samples, the variant quality and genotypic information is often lost, which makes it difficult to interpret the result.

Nowadays, this approach has been replaced by haplotype-based variant calling. Haplotype-based callers work by constructing a haplotype from the sequencing data instead of relying only on one position at a time [35]. This approach allows haplotype-based callers to identify variants in the regions which are difficult to analyze using a standard variant caller, especially to identify high-quality INDELs. Several tools are available following this approach such as HaplotypeCaller from GATK, freebayes [35] and platypus [36]. However, the HaplotypeCaller from GATK is the extensively used tool and it works by performing a local de-novo assembly of the regions of interest which makes HaplotypeCaller superior to other variant caller of GATK called UnifiedGenotyper. Another advantage of GATK haplotype caller is to generate an intermediate genomic VCF (gVCF) file which contains the genotype information for every position of a genome or an exome. Generation of gVCF allows the users to perform a more robust multi-sample variant calling compared to the traditional single sample variant calling. The advantage in multi-sample variant calling is that, it calls multiple samples together, due to which, the probability to call a variant increases even if some samples do not have enough coverage at a variant position or if it occurs in low allele frequencies. Further, multi-sample variant calling gives the leverage to identify additional variants of high-quality which cannot be identified via individual sample calling. The idea behind multiple sample calling is; if there is a strong evidence for a variant to be called in sample 1 and on the other hand, if there is a weak evidence in the sample 2, GATK's haplotype caller takes the evidence from sample 1 to call the variant in sample 2.

In exome sequencing, often many off-target (non-exonic) regions also have sufficient depth of coverage to call the variants. However, these off-target regions do not have similar coverage across all the samples. Henceforth, these variants are unreliable and including such regions during variant calling leads to low-quality variant calling. In order to exclude those low-quality variants, a common exonic interval file is used in case of WES based variant calling.

1.4.2 Variant Quality Control (QC)

Variant Quality Score Recalibration (VQSR)

After the variant calling by GATK's haplotype caller, the recommended way to filter low-quality variants is through VQSR tool from GATK. The advantage of VQSR is that, instead of defining hard thresholds in order to exclude low-quality variants, a continuous, covarying estimate of the relationship between SNP call annotations (such as QD, MQ, and ReadPosRankSum) and the probability of a SNP being a real variant versus an artifact. VQSR uses a list of "true sites" as one of the input. The commonly used sites for this purpose are HapMap3 sites [37] and those sites that are found to be polymorphic on the Omni 2.5M SNP chip array.

After building a model based on the "true sites" this adaptive error model can then be applied to both known and novel variations discovered in the call set of interest to evaluate the probability that each call is real. A variant quality logarithm of odds (VQSLOD) score is generated for each variant and added to INFO field of the VCF file. VQSLOD score represents the log odds of being a true variant versus being false under the trained Gaussian mixture model [38]. VQSR runs in two-steps: The first step works by generating a Gaussian mixture model based on the distribution of annotation values over a high-quality subset of the input call set and then scoring all the input variants according to the model. The second step consists of filtering the variants by applying the cut-offs based on the scores generated in the first step of VQSR.

Hard filtering of variants

Although VQSR is an efficient way to exclude low-quality variants, it requires a minimum of 30 samples to generate an efficient model. Also, there might be still some unreliable variants even after performing VQSR. In order to be more stringent a hard filtering approach is often performed as an extension to VQSR. In the hard filtering, variants are filtered based on various quality scores generated during the variant calling. This method of employing VQSR along with hard filtering has shown to be more efficient in reducing the false positives. There are no consensus hard filtering parameters, however the recommended criteria according to GATK best practices are [23]. a) For SNVs: Variants were filtered for QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, DP<10.0, GQ_MEAN<20.0, VQSLOD<0, ABHet >0.75 or <0.25 and Hardy Weinberg Phred scale P value of >20. b) For INDELs: QD < 2.0, FS > 200.0, ReadPosRankSum < -20.0, DP<10.0, GQ_MEAN<20.0, Hardy Weinberg Phred scale P value of >20, VQSLOD>0. However, these parameters tend to be adjusted according to the study, while it also depends on the methods used to generate the variant calls.

1.4.3 Sample QC

After the filtering of low-quality variants, the subsequent steps of sample QC include selecting the samples of high-quality. This is an important step because in an NGS study, there are various sources of errors, for example a gender or a relationship mislabeling could lead to completely erroneous results. Hence, in order to control for such errors, various steps are employed as part of a standard NGS data processing.

Sample filtering based on quality metrics

Number of alternate alleles, number of heterozygotes, transition/transversion ratio (Ti/Tv), number of singletons and call rate, serve as an evidence for the quality of the data. They can be calculated by tools such as PLINK/SEQ i-stats parameter at different stages of data filtering. One way to filter the low-quality samples, is to exclude any sample with >3 standard deviation (SD) from the mean in the above mentioned metrics. After excluding the low-quality variants and samples, only bi-allelic SNVs that are concordant with hapmap3 vcf (version 3.3) [37] are typically selected for checking the cryptic relatedness, deviations from reported sex and population stratification. The most widely used tool for this purpose is PLINK [39].

Relatedness and sex check

It is extremely important to check for unreported and cryptic relatedness in association studies. For relatedness and sex check, the well known tool is PLINK. For relationship detection, it works by identifying the fraction of genome shared between each pair of samples. Other tools available for this purpose are KING [40], genetic relationship by averaged blocks (GRAB) [41], etc. Typically, in an association study only one sample is selected from an identified pair of relatives based on, either it's quality, if there is any difference or one sample will be chosen randomly from the related pair. In a family based study, this approach can be used to check the reported relationships and take further QC steps if needed. To determine the gender of a sample, PLINK works by using the data from X chromosome. However, the cut-off to determine the sex has to be set based on the study, as each study has a different depth of coverage. In an association study, a sample is often excluded if there is a mismatch between the reported and calculated sex.

Population stratification

Population stratification is one of the main criteria that needs to be accounted in an association analysis. Otherwise, it might lead to spurious results. The widely used strategy is to merge the NGS data with 1000 genomes data [34] and then compare the ethnicity of the sample under current study with respect to the samples from 1000 genomes. Two tools are most commonly used for this purpose and they are multi dimensional scaling (MDS) available as part of PLINK and Eigenstrat [42] which is available as part of Eigensoft tool. Eigenstrat performs a principal component analysis (PCA) and produces a list of outliers with >6SD (default) iteratively based on the first ten principal components. Whereas typically in PLINK, the cut-off to determine outliers has to determined manually by visual inspection of first and second principal components.

Sample contamination

Along with the above mentioned criteria, contamination between different samples can be checked by using inbreeding coefficient as a means of measure. Similarly, missingness can also be used as one of the criteria to select high-quality samples. PLINK and vcftools [33] are the well known tools for calculating these metrics.

1.4.4 Functional Annotation

After generating a set of high-quality calls, the information that is obtained about the chromosomal position, the nucleotide level change and the genotype per sample. To utilize this information in a meaningful way and select the functionally relevant variants, it is requires to annotate the variants with information from multiple sources such as the the functional consequence (missense, synonymous etc.), location in the genome (intron or exon), name of the gene it is effecting (within the gene, upstream, downstream , regulator etc.,) and the frequency at which the variant is occurring in general population such as ExAC [43] or a disease related database such as Human Gene Mutation Database [44]. This entire process is called as variant annotation and the annotations are chosen depending on the application and the study. As, a first step in annotation the multi-allelic variants should be decomposed using tools such as variant-tests [45] and left normalized by bcftools [46]. In the next steps, tools such a ANNOVAR [47], variant effect predictor (VEP) [48] or Snpeff [49] can be used for annotation. Each tool has it's own naming convention for the predicted consequence of the variant and hence needs to be chosen based on the application. The main resources typically used for gene annotation are RefSeq of National Center for Biotechnology Information (NCBI), Ensembl and Consensus coding sequence (CCDS). Each of them has different number of coding transcripts and it is also possible that the variant predicted to be benign using one resource could be predicted as a damaging variant in the other resource. Hence, the choice of transcript has an important effect on the variant annotation [50]. Typically, if there are more than one transcripts for a variant, then the tools have an option to produce the annotation with the most severe consequence. Databases such as dbNSFP [51] provide various scores for nonsynonymous and splice site SNV consequences. To determine the rarity of a variant, databases like 1000 genomes, dbSNP [52], ExAC [43] (release 0.3, NFE and ALL), and the Exome variant server (EVS) `http://evs.gs.washington.edu/EVS/` are available. These databases include the genetic data from various ethnicities and hence, one can filter for ethnicity specific frequency when required. Further, the filtering criteria for rarity of a variant is often arbitrary and there is no single definition of the rarity. It should be adapted according to the study and it could range from (0.01% to 3%).

Variant prioritization and analysis

Depending on the aim of a study, variants can be prioritized in a different manners. However, a brief description on functional prioritization is given below:

- **Nonsynonymous variants:** These are amino acid changing variants. RefSeq, Ensemble and CCDS annotations can be used to define a variant as nonsynonymous variant.
- **Loss of function variants (LoF):** Any SNV annotated as “splicing”, “stop gain” or “stop loss” or any INDEL (especially frameshift) can be defined as a LoF variant. Although, they are named as LoF variants not all of them cause loss of function per se.
- **Synonymous variants:** These are the variants which do not change the amino acid. They were assumed to be functionally neutral and used in majority of studies a negative control. However, there might be some synonymous variants which are functionally important such as those involved in splicing [4].
- **Deleterious scores:** There are several deleteriousness prediction scores such as SIFT [53], PolyPhen2_HDIV [54], LRT [55], MutationTaster [56], PROVEAN [57], CADD [58], DANN [59], fathmm [60], GERP++_R [61] and SiPHy [62] are available for estimating the delteriousness of a variant. Some of them are available for the entire genome such as CADD, GERP++_R and SiPHy. It really depends how one can use these scores to prioritize a variant, sometimes a combination of these scores are used [63], while the other

times only one of them is used such as CADD [64, 65].

1.4.5 CNV detection

Various methods to detect CNVs include the traditional methodologies, such as karyotyping and fluorescence in situ hybridization (FISH) [66] or the array-based comparative genomic hybridization [67]. These approaches have several drawbacks including hybridization noise, constrained coverage, inferior resolution, and also similar to the detection SNPs using arrays the detection of novel and rare CNVs is not possible. With the advent of NGS, detection of CNVs has become much reliable and fast. The methods to detect CNVs from NGS data can be based on paired-end mapping, read depth, split reads, de-novo assembly of the genome and combination of the above approaches [68]. Each method has it strengths and weaknesses and out of the five methods the most well known method is by using read depth.

Typically, read depth based methods work by mapping the reads to a genome, normalization of the read count across the genome, calculation of exact copy number, and the final step is the segmentation. First, in the mapping step, the reads from FastQ files are aligned to the genome of interest and the number of reads covering each position of genome (Depth of coverage) is calculated. In the next steps the depth of coverage is normalized in order to avoid potential biases arising due to varying GC content or the repeat regions within the human genome. Once, the normalized read depths are obtained an estimation of copy number is performed in order to determine whether there is a gain or a loss. Finally, segmentation of genomic regions is performed to detect conflicting copy number regions. Further, these methods are divided into two based on whether the copy number detection is performed on one sample or multiple samples together. A detailed summary of all the different tools is provided in the study [68]. An example of a typical CNV detection workflow from WES data by employing XHMM [69, 70] is shown in Figure 1.4. Conceptually, XHMM works in a similar fashion as GATK's multiple sample calling where we give multiple files as input and based on the coverage in all the samples CNVs are discovered.

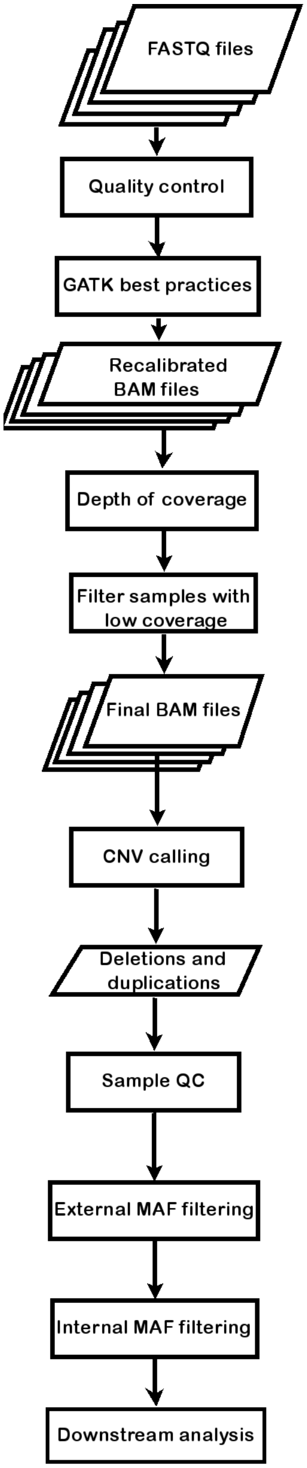


Figure 1.4: Various steps involved in the detection and processing of CNVs from WES data by using XHMM.

1.4.6 QC and prioritization of CNVs

In order to select high-quality CNVs each tool generates a quality score, for example CNVs detected by XHMM can be filtered using Z score and Q_SOME score. The main advantage of detecting CNVs via NGS is to discover rare and novel events. Hence, several publicly available databases such as CNVmap [9], the DGV gold standard data-set [71] and 1000 genomes SV [72] can be used for this purpose. Further, CNVs arising from the questionable regions of the human genome often have poor quality [73]. To mitigate this effect, the CNVs overlapping the regions such as centromeres, telomeres and Immunoglobulins can be excluded. Further, filtering criteria can be employed based on the study.

1.4.7 Association testing

Typically, association testing of SNVs/INDELs is performed in the case-control studies to determine the variants causing a significant burden. While studying the effect of rare variants, it is normal to not find any significant association of a variant to the disease mainly due to small sample sizes. Hence, it is a common practice to perform the association at gene, geneset and genome-wide levels. There are various methods to perform association testing. The common methods being burden analysis using Fisher method, kernel based methods such as SKAT or SKAT-O or a more recent method using linear models. Each of them have their own pros and cons. The standard burden tests work by collapsing the rare variants in a region such as gene or a pathway into a single burden variable and then regress the phenotype on the burden variable in order to test for the aggregated effects of rare variants in the defined region. However, the typical burden tests often tend loose their power if a region consists of both protective and deleterious variants or many non-causal variants acting in opposite direction.

In such conditions, advanced methods like sequence kernel association test (SKAT) [74] tend to be more powerful. Instead of aggregating variants, SKAT aggregates the associations between variants and the phenotype through a kernel matrix and can allow for SNP-SNP interactions. We used an optimized version of SKAT called SKAT-O [75] which is more powerful compared to SKAT, as SKAT-O behaves like the burden test by default when the burden test is more powerful than SKAT. Otherwise, if SKAT is more powerful than the burden test then SKAT is performed instead of burden test. These days several tools are available that take a VCF file as an input and perform a series of burden or kernel based tests such as rvtests [76], epacts (`https://genome.sph.umich.edu/wiki/EPACTS`) or PLINK/SEQ (`https://atgu.mgh.harvard.edu/plinkseq/`).

Association testing of CNVs is performed by burden testing and the most widely used tool for this purpose is PLINK [77]. It has a special module to perform the burden testing of CNVs with a combination of permutation.

Linear models were used to perform association analysis of genesets carrying SNVs/INDELs

[65, 78]. The main advantage of linear models is that, one can account for various confounding variables such as population differences, study wide coverage differences or other technical differences and it also allows to estimate the odds ratios. Based on the odds ratios, one can determine the direction of effects of variants. Similarly, linear models can also be used to perform association testing of CNVs to the trait of interest [79].

1.4.8 Mode of Inheritance (MOI) filtering

In contrast to association studies, family studies employ a MOI based variant filtering. Based on an individual study, different inheritance patterns can be tested for any kind of variants such as SNVs/INDELs or CNVs as described below.

- **Autosomal dominant inheritance:** Out of the two copies of a gene, if a mutation in one copy could induce the disease phenotype then such type of MOI is called Autosomal dominant inheritance. In this type of inheritance, as even one of the mutated gene can lead to disease phenotype, even if one of the parent is affected, there is 50% chance for the offspring to inherit the disease.
- **Autosomal recessive inheritance:** In the case of autosomal recessive MOI, both alleles of the gene have to be mutated in order to induce the disease phenotype. If only one allele of the gene is mutated the other other allele could compensate for the mutated allele thereby preventing the disease. However, the person carrying one mutated allele becomes a carrier and their off-springs can have three possible phenotypes: He/she could become a carrier themselves (50% chance) They could inherit the mutated gene from both the parents and become susceptible to the disease (25% chance) They could inherit the healthy alleles and stay normal with respect to the disease (25%)
- **De-novo inheritance:** These are the newly emerged mutations occurring either in germline cells of the parents or at some point in life after conception. These kind of mutations are commonly identified via trio based studies.
- **Compound heterozygous variants:** If an individual carries a variant in gene on one allele and another variant in the same gene on the second allele then the inheritance is called compound heterozygous inheritance.

1.5 Neurological disorders

Neurological disorders constitute a wide range of diseases affecting the nervous system. They affect several people worldwide irrespective of age, gender or race. Further, they not only damage the nervous system, but also affect the quality of life and cause a major financial burden [80]. More than 800 neurological disorders have been identified till date. They range in severity and the symptoms are often different from person to person ranging from cognitive dysfunction to manic behavior or depression [81]. An indicator of the disease burden is disability-adjusted life years (DALY). It is expressed as the the number of years lost due to disability, ill-health or early death. According to a study published in 2017 [82], the neurological disorders included in this analysis caused 250,692 million DALYs, comprising 10.2% of global DALYs, and 9,399 million deaths, comprising 16.8% of global deaths showing the burden caused by the neurological diseases alone. Large variations related to geographical and sex differences [83] for neurological disorders were also found.

A recent survey of world-wide literature [84] has shown that alzheimer's disease (AD), chronic low back pain (CLBP), stroke, traumatic brain injury (TBI), migraine headache, epilepsy, multiple sclerosis (MS), spinal cord injury, and Parkinson's disease have been found as the most common neurological disorders. There is no clear definition of age at onset (aao) for the neurological disorders as a whole. For example, autism spectrum disorder (ASD), Cerebral Palsy (CP) and Tourette syndrome are early-onset [85]. Whereas, disorders such as AD and PD affect mostly elderly people although with few exceptions. Similarly, there are several neurological disorders such as migraine, epilepsy, Multiple Sclerosis (MS), stroke and brain or spinal cord injuries can affect at any point of an individual's life time.

One of the main components that is believed to play a role in etiology of neurological disorders is genetics. By now, from various studies it has been well established that genetics and genomics play an important role in the etiology of neurological disorders [81]. Majority of the neurological disorders are found to be complex disorders, where often there is more than one factor that cause or aid in the progression of the disease [86]. Due to the complexity of these diseases, traditional methods studying limited genes and pathways cannot always give a full picture of the underlying mechanisms [81]. Despite the critical role of genetics in neurological disorders, the consequences of genetic variants are diverse. For instance, in Huntington's Disease (HD) the disease is caused by an extension of CAG repeat of huntingtin gene (*HTT*) which leads to the production of pathogenic huntingtin protein [87]. In the same way, the CAG repeat expansion of *ATXN1* produces abnormal ataxin-1 protein and leads to Spinocerebellar ataxia type 1 (SCA1). In contrast, in the case of diseases such as PD, AD, schizophrenia, epilepsy etc., the genetics is more complex and is often found that several genes contribute to the disease. Moreover, it is also multi-factorial and seen that there is a complex interplay of genes and environment [86, 88].

Measurement	All-age numbers (thousands)		Age-standardized rate (per 100000)	
	2015	Change from 1990 to 2015	2015	Change from 1990 to 2015
PD				
DALYs	2,059 (1,832 to 2,321)	111 · 2% (102.4 to 118.1)	33 (30 to 37)	10 · 8% (6.5 to 14.3)
Deaths	117 (114 to 121)	149 · 8% (135.0 to 161.4)	2 (2 to 2)	22 · 6% (15.7 to 28.4)
Prevalence	6,193 (5,726 to 6,777)	117 · 8% (113.2 to 122.8)	98 (90 to 107)	15 · 7% (13.3 to 18.3)
Epilepsy				
DALYs	12,418 (10,438 to 14,479)	2 · 5% (-5.7 to 11.2)	168 (141 to 195)	–22 · 5% (–28 · 2 to –16.8)
Deaths	125 (119 to 131)	18 · 9% (6.4 to 32.1)	2 (2 to 2)	–15 · 6% (–23 · 0 to –8.0)
Prevalence	23,415 (21,550 to 25,419	39 · 2% (33.4 to 45.2)	320 (295 to 347)	1 · 9% (–2 · 1 to 6 · 1)

Table 1.1: An increase in the number of DALYs from 1990 to 2015 due to epilepsy and PD. This table is modified from study [82]

According to a recent study [82], in the past few years there is a substantial increase in the DALYs due to PD and epilepsy Table 1.1. This emphasizes the fact that there is an increase in the global burden of neurological diseases and measures have to be taken in order to account for the increased burden. In the same study, it was shown that when stratified according to age, epilepsy caused the most burden in children and young adults. Whereas, the burden of PD increased along with age, similar to other neurological diseases [82].

In my current work, I have focused on genetics of two of the major neurological disorders namely, PD and epilepsy. Both, PD and epilepsy follow a common pattern where there is a symphony of common, less common, rare and ultra-rare variants with varying effect sizes as shown in Figure 1.5. The inception of Arrays, Next Generation Sequencing (NGS) and their combination with the latest systems biology approaches have discovered several novel risk genes, biomarkers and drug targets [89]. As, shown in Figure 1.5 common SNPs associated to the disease are usually identified by using arrays, whereas low frequency variants are identified using NGS technologies.

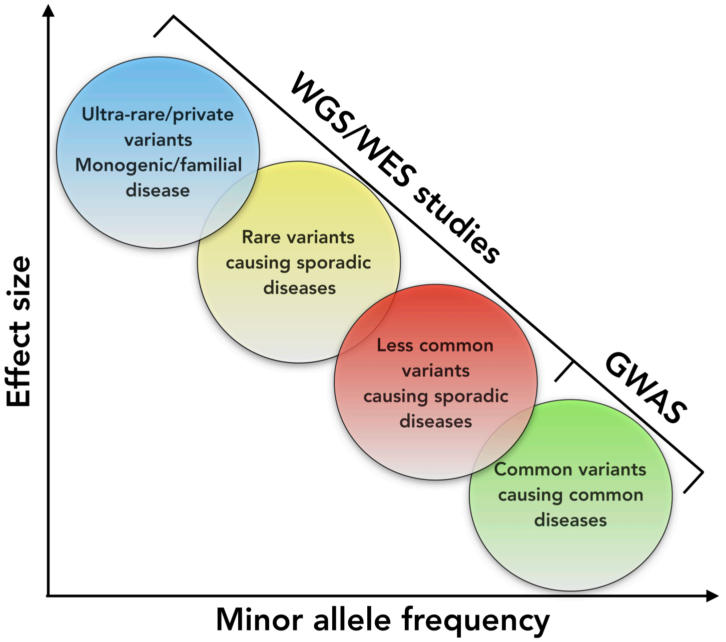

Effect size
Minor allele frequency
WGS/WES studies
GWAS
Ultra-rare/private variants Monogenic/familial disease
Rare variants causing sporadic diseases
Less common variants causing sporadic diseases
Common variants causing common diseases

Figure 1.5: A figure describing the pattern of various variants commonly found associated to neurological disorders and their varying effect sizes.

As PD and epilepsy are complex diseases with more than one gene effecting the disease, instead of studying variants belonging to one frequency spectrum, variants belonging to different allele frequencies and predicted biological effect were studied in this thesis. The work that has been done in this thesis provided us a glimpse of convoluted architecture of PD and epilepsy. Majority of this work has been conducted by using either Whole Exome Sequencing (WES) or Whole Genome Sequencing (WGS) data.

1.6 Parkinson's Disease

1.6.1 Background

PD is a severe neurodegenerative disease affecting several regions of brain, especially substantia nigra. Due to its substantial variability in phenotypic, neuropathological, and genotypic characteristics, it is being recognized as a heterogeneous disorder. It is a slow progressing disease, the average aao for PD is 60 years and it reaches a prevalence of 5% in individuals with an age >85 years [90], however some people (5%) were diagnosed with PD below 60 years. In brain, an important chemical messenger called dopamine is produced by the cells of substantia nigra and it aids to control the movement of human body. In PD, the loss of dopamine producing neurons occurs, which results in the uncontrolled movement of the patients. The cardinal symptoms of PD include resting tremor, rigidity and slowness of movement (bradykinesia). In PD, typically the symptoms start from one side of the body and then proceed to the other side. The motor symptoms are often accompanied by several non-motor symptoms. Some of the major non-motor symptoms for PD include loss of sense of smell, sleeplessness, speech problems, constipation, troubled swallowing, low blood pressure and drooling when standing [91]. Unfortunately, there is no confirmatory test like a blood test, EEG or brain scan to make a clear diagnosis of PD.

The diagnosis of PD usually depends on the expertise of neurologist who performs a thorough neurological examination. Especially, an investigation will be conducted for the presence of two or more of the cardinal symptoms. Also, the doctor could also check the patient's response to PD medications, which serves as a further evidence of PD. The first test approved by Food and Drug Administration (FDA) in order to diagnose PD is an imaging technique called DaTscan which serves as the measurement of dopamine activity in brain.

According to a recent report, more than 4 million individuals in Europe's five most and the world's ten most populous countries are currently afflicted with PD [92]. In United States alone PD is estimated to affect 630,000-1,000,000 people and by 2050 these numbers are projected to be doubled approximately. Ratio of men and women affected by PD is disproportional (2:1) as it more predominantly affects men [93] and is typically late-onset (>60 years). An estimated direct and indirect costs sum up to a total of \$15.5 billion per year for PD [84]. The estimated direct costs are at \$13,786 per patient, with an aggregate direct medical cost of \$8.1 billion. The aggregate direct cost includes outpatient and institutional care, retail prescriptions, supplies and equipment. Additionally, indirect costs are estimated to be \$10,816 per patient, or \$6.8 billion in aggregate, including number of working days lost due to illness, reduced employment and household income, adult day care, higher disability payment, any other formal care, and various household expenditures. PD is an incurable disease, apart from the financial burden, it mainly affects the quality of life. However, few medications and techniques such as deep brain simulation can help in the management of PD. Based on the cause, PD can be broadly classified into two types namely Idiopathic PD and PD with mendelian inheritance. They are described below.

1.6.2 Idiopathic PD

The most common type of PD is Idiopathic PD (IPD), affecting >2% of those over 75 years. IPD occurs in individuals having no family history of PD. The etiology of IPD is incompletely understood. Hence, in IPD various factors could contribute to the disease etiology. Ageing is one of the major risk factor, similar to other neurodegenerative diseases. Smoking is one of the factors that is believed to play an important role in IPD, a negative association has been observed between PD and smoking. It is believed that smoking has a neuroprotective affect, as people who smoke cigarette are less likely to develop PD [94]. However, the findings linking smoking and IPD are very controversial as the studies did not account for smokers dying younger, and therefore being less likely to develop a condition that is more common in old age.

Several weak associations between PD and head injury [95], use of psychoactive medication, middle-age obesity, lack of exercise, rural living, well-water ingestion, and pesticide exposure (paraquat, organophosphates, and rotenone) have also been reported. Environmental toxins such as 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine MPTP [96] can produce a similar but not identical clinical scenario [97]. But, majority of studies lack a large sample size and thus remain inconclusive. In order to estimate the genetic contribution to the pathogenesis of PD, several twin studies have been conducted [98–100]. However, majority of them showed low concordance rates in monozygotic and dizygotic twins. A major criticism of these studies was that the cross-sectional study designs used did not exclude the possibility of a later disease onset in unaffected siblings. This obstacle has been overcome by using positron emission tomography studies (PET), which is sufficiently sensitive to identify pre-symptomatic subjects by detecting decreased striatal (18F)6- fluorodopa as a metric for decreased striatonigral dopaminergic function. Based on PET scan data, the concordance rate was significantly higher for monozygotic twins, than for dizygotic twins [101–103] (55% versus 18%), suggesting a substantial genetic contribution to the PD pathogenesis.

1.6.3 Mendelian/Familial PD

PD occurs as a sporadic disorder in vast majority of patients, but in 5%-10% of cases, the disease occurs as a Mendelian disorder. It has been previously thought that there is no genetic basis for PD. However, the discovery of mutations in *SNCA* (α-synuclein) [104] has changed the perception of PD etiology. Since, the discovery of *SNCA* as a PD associated gene, a substantial number of genes related to PD were identified as shown in Table 1.2, including mutations in genes responsible for rare monogenic forms of PD. It could be seen in Figure 1.5 that, PD is caused by variants in multiple genes and they vary in the conferred risk depending on their allele frequency. Previous methods that considered PD as a monogenic disease were successful to develop therapies that compensate for the dopaminergic deficit responsible for the motor symptoms of PD. However, they fall short in terms of developing neuroprotective treatment strategies. Focusing on patho-mechanisms and understanding the underlying molecular pathology of neurodegeneration is essential, and genetic stratification of patients into subgroups provide an important entry point for precision medicine.

The monogenic forms of PD have become a valuable resource for PD research, as patient-based cell models display disease-specific cellular phenotypes that recapitulate the phenotypes found in post-mortem brain tissue. According to this concept, the validation of clinicogenetic subtypes of PD may be achieved based on rare but strong molecular signatures and subsequently applied to the different pathophysiological tiers within each disease subtype. As genes were identified to cause monogenic forms of PD, they were assigned PARK loci and numbered in chronological order of their identification. However, the PARK loci do not contain genes which only cause monogenic forms of PD, but also loci identified from genome-wide linkage screens, some of which have not been replicated in subsequent studies. To date, 18 PD-associated loci (PARK1-18) have been described Table 1.2.

Locus	**Position**	**Gene**	**Inheritance**	**Function**	**Implications**	**age at onset**
PARK1 & 4	4q21-23	α-synuclein (*SNCA*)	Dominant	Unclear (presynaptic protein)	Protein aggregation	LOPD/EOPD
PARK2	6q25.2-27	Parkin	Recessive	Ubiquitin ligase	Aberrant protein & mitochondrial homeostasis	EOPD
PARK3	2p13	Unknown	Dominant	-	-	LOPD
PARK5	4p14	*UCHL1*	Dominant	Ubiquitin hydrolase	Aberrant protein homeostasis	LOPD
PARK6	1p35-36	*PINK1*	Recessive	Putative serine/threonine kinase	Aberrant mitochondrial homeostasis	EOPD
PARK7	1p36	*DJ-1*	Recessive	Redox sensor	Oxidative stress	EOPD
PARK8	12p11.2-q13.1	*LRRK2*	Dominant	Putative serine/threonine kinase	Aberrant phosphorylation	LOPD
PARK9	1p36	*ATP13A2*	Recessive	Lysosomal P-type ATPase	Aberrant protein homeostasis	EOPD
PARK10	1p32	Unknown	Unknown	-	-	LOPD
PARK11	2q37.1	*GIGYF2*?	Dominant	-	-	LOPD
PARK12	Xq21-q25	Unknown	X-linked	-	-	LOPD
PARK13	2p12	Omi/*HtrA2*	Dominant	Mitochondrial serine protease	Aberrant mitochondrial homeostasis?	LOPD
PARK14	22q13.1	*PLA2G6*	Recessive	Phosopholipase	Aberrant lipid homeostasis?	EOPD
PARK15	22q12-q13	*FBXO7*	Recessive	Component of SCF E3 complex	Aberrant protein homeostasis?	EOPD
PARK16	1q32	Unknown	Unknown	-	-	LOPD
PARK17	16p12.1-q12.1	*VPS35*	Risk	-	Aberrant endosomal recycling?	LOPD
PARK18	3q27, 1q21, 4p15, 17q21, 12q12 and 14q32	*EIF4G1*, *GBA*, *BST1*, *MAPT*, *ATXN2* and *ATXN3*	Risk	-	-	LOPD

Table 1.2: A table describing all the PARK loci that were discovered till date and their implications. This table is a combination of two tables from the studies [92, 105]. LOPD = Late onset Parkinson's disease and EOPD = Late onset Parkinson's disease

1.6.4 Genetics

Familial PD genes

Linkage analysis is a powerful approach for the discovery of disease associated genes in families and it has led to the discovery of two genes *SNCA* and *LRRK2* which were further strongly supported by the evidence from GWAS. It has been functionally shown that expression levels of *SNCA* are inversely correlated to the aao [106]. For *LRRK2* the mutation p.G2019S was seen to be major risk factor for PD. The penetrance seems to be ethnicity specific varying from 22% Ashkenazi Jews (residents in US), 45% for Norwegians and 80% for Arab-Berbers [106]. This ethnicity related penetrance is very important, especially while conducting genetic counselling. However, only ~1% of PD seems to be explained by the p.G2019S mutation. Whereas, in the Asian population, another mutation p.G2385 has more penetrance (11.37%) compared to the previous one [107]. Several genes were identified recently that are shown to be causing Familial-PD. The autosomal dominant genes causing PD include *SNCA, LRRK2, EIF4G1, VPS35, DNAJC13, CHCHD2, TMEM230* and *RIC3*. Whereas, the monogenic causes for autosomal recessive or X-linked PD include *PARK2, DJ-1, PINK1, FBXO7, 22q11.2del, SYNJ1, RAB39B, DNAJC6, PODXL, VPS13C* and *PTHRHD1* [108]. However, functional studies are further required in order to confirm the role of these genes in the familial forms of PD.

GWAS genes

Several GWAS have been conducted to discover the PD associated genes including the recent large-scale meta-analysis where they used 26,035 cases and 403,190 controls and discovered 17 novel associated loci. In total, there are >50 loci that are found to be associated to PD [109]. The large-scale GWAS have supported the association of *LRRK2* and *SNCA* to PD. Similarly, several genes such as *UCHL1, PARK16, GAK, MAPT, GBA, NAT2, INOS2A, GAK, HLA-DRA* and *APOE* have also been identified as risk factors for PD [110].

Modifiers

As described previously in section 1.6.4 the mutation p.G2019S accounts for 1% of PD in Caucasians, whereas it is higher in other populations. However, it has been noted that the carriers of p.G2019S mutations have varying phenotypes. Some of them have early-onset parkinsonism, while some remain asymptomatic of PD despite their old age. Hence, it has been claimed that there must be some modifiers which alter the aao and the development of PD. A recent study [111] has identified that *DNM3* might be a genetic modifier of aao for parkinsonism by *LRRK2* p.G2019S.

CNVs in PD

There have several studies which detected the genomic triplications and duplications in *SNCA*. The first study to report the genomic triplicationss in chromosome 4q21 22 of *SNCA* was performed in a large family with PD following AD inheritance. The triplication was confirmed by PCR and FISH technology [112]. Since then, several studies have reported the duplications and triplications in *SNCA* [113]. Interestingly, all of the CNVs in *SNCA* suggest a gain of function [114]. However several forms of PD are early onset, suggesting a LoF mechanism in PD and one of the major genes that was shown to be associated to PD via LoF is *PARK2*. There have been some reports where heterozygous CNVs in *PARK2* have been found to be associated to the increased risk of PD [115, 116]. While, some other studies did not confirm such an association [117, 118]. Hence, it remains to be seen whether any rare CNVs in *PARK2* will be found in the near future. The other PD associated gene that is found to harbor a heterozygous deletion is *PINK1*. In a study [119], it was found that the entire *PINK1* is deleted and the deletions span a length of 56kb. Similarly, other studies have also found heterozygous deletions in *PINK1* [120, 121]. The deletions in *DJ1* have been found in genetically segregated cohorts of Netherlands and Italy [122, 123]. Further, deletion involving *DJ1* were also found in a family of Iranian origin[124]. All the families carrying these deletions are consanguineous. In one family of Iranian origin, a deletion was reported in *ATP13A2*, but no other CNVs were reported in the same gene till date [124]. Other CNVs affecting single gene include *TH, VPS35, PGRN* and *HMOX1* [113].

1.7 Epilepsy

1.7.1 Background

Epilepsy is a chronic neurological condition affecting over 65 million people worldwide. 2.8 million Americans are affected by epilepsy which is approximately 1% of the general population [84]. Epilepsy disproportionately affects black men (`https://www.cdc.gov/mmwr/pdf/wk/mm6145.pdf`) as well as the elderly [125]. Epilepsy is one of the less studied neurological disease despite of it high prevalence and economic burden. According to a study, direct medical costs in both children and adults in the year 2009 epilepsy is estimated to a cause a burden of $9.6 billion. Another study from 2004 also estimated the prevalence and burden of epilepsy, however they did not assume indirect costs and considering the static disease prevalence from 2004 to 2014, adjusting to 2014 [84], the direct medical costs of epilepsy for 2014 are estimated at $13.4 billion. However, the same study showed that by including the indirect costs for epilepsy patients, the total cost burden (direct and indirect) is estimated at $36.8 billion.

The primary feature of epilepsy is the spontaneous seizure activity, which occurs due to the sudden bursts of electrical activity in the brain. Epilepsy is a disease with a combination of multiple syndromes [126]. More than 50 syndromes have been found to be part of epilepsy

[127]. Based on their location of origin, epilepsy can be broadly divided into focal epilepsy and generalized epilepsy. Focal epilepsies originate from a specific area of a brain, whereas the origin of epileptogenesis is unclear in generalized epilepsies. According to the International League Against Epilepsy (ILAE) [128], the practical definition of epilepsy is described as below.
Epilepsy is a disease of the brain defined by any of the following conditions

- *At least two unprovoked (or reflex) seizures occurring >24h apart*
- *One unprovoked (or reflex) seizure and a probability of further seizures similar to the general recurrence risk (at least 60%) after two unprovoked seizures, occurring over the next 10 years*
- *Diagnosis of an epilepsy syndrome: Epilepsy is considered to be resolved for individuals who had an age-dependent epilepsy syndrome but are now past the applicable age or those who have remained seizure-free for the last 10 years, with no seizure medicines for the last 5 years.*

Although, epilepsy can occur at any age it is more common amongst children and people above 65 years of age. Not all epilepsy sub-types are life long, some forms are confined to childhood. The prevalence of epilepsy in general population is 3.3-7.8/1000 and 3.4-5.8/1000 in pediatric studies with an age limit ranging from 0 to 18 years [129]. There are several sub-types of Epilepsy. A recent classification of epilepsy according to the ILAE 2017 (Instruction manual for the ILAE 2017) [130] can be seen in the Figure 1.6.

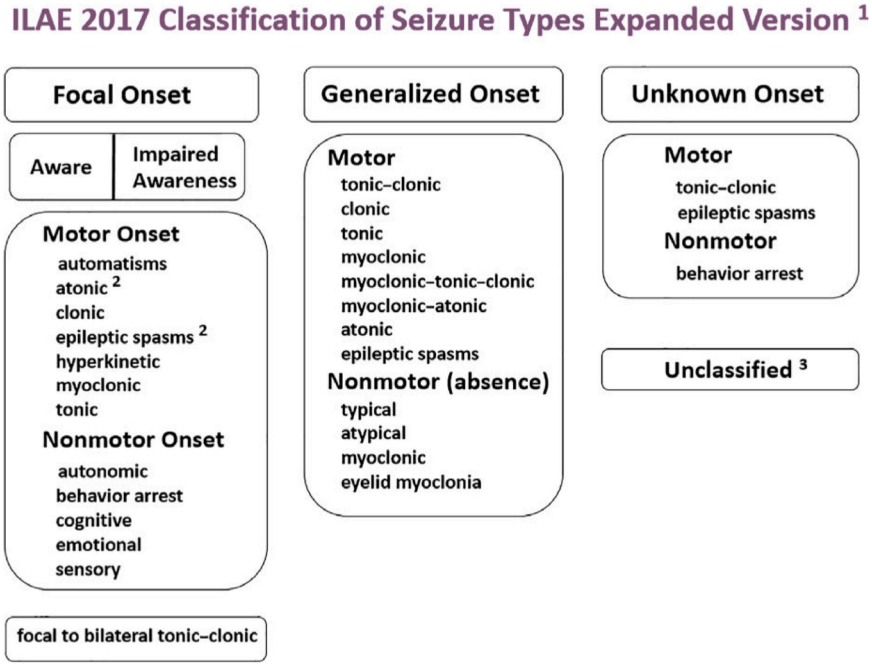


Figure 1.6: A figure describing the 2017 classification of epilepsy according to ILAE

1.7.2 Genetics

One of the factors that is believed to play an important role in many epilepsy syndromes is genetics. Several genes were identified by Genome-wide association studies (GWAS) [131, 132], Trio-based studies [133] and segregation-based studies in the families. They are briefly described below.

Ion-channel and non ion-channel variants in epilepsy

In the central nervous system and excitable tissues such as skeletal and heart muscle, ion channels form the basis of excitability regulation. There are various types of ion-channels such as sodium, potassium, calcium or chloride channels depending on the ions they allow to pass through them. The ion-channels play a major part in controlling the excitability and any defect in the ion-channels could lead to hyper or hypo-excitability of the concerned tissue [134]. The change in excitability might ultimately lead to epileptogenesis. About 25% of the genes that are known to be mutated in epilepsy are in the ion-channels [127]. The first concept of "channelopathy" in epilepsy has been led by the discovery of variants in *KCNQ2* and *SCN1A* [135]. Till date, several variants occurring in the voltage or ligand gated ion-channels were found to be major cause of epilepsy. Table 1.3 shows the ion-channel genes that are mutated in diverse forms of epilepsies. The mutations in ion-channels are known to cause rare monogenic idiopathic epilepsies, however they were also found in some common epilepsies such as juvenile myoclonic epilepsy or childhood and juvenile absence epilepsies [136].

Gene	Phenotype	Protein
	Voltage-Gated	
SCN1A	Dravet syndrome; GEFS+	NaV 1.1
SCN1B	GEFS+, temporal lobe epilepsy, an early infantile epileptic encephalopathy	NaVb1
SCN2A	BFNIE, early-onset epileptic encephalopathies, neurodevelopmental disorders	NaV1.2
SCN8A	BFIE, epileptic encephalopathy	Nav1.6
KCNA1	Partial epilepsy and episodic ataxia	KV1.1
KCNA2	Epileptic encephalopathy	KV1.2
KCNB1	Epileptic encephalopathy	KV2.1
KCNC1	Progressive myoclonus epilepsy	KV3.1
KCNMA1	epilepsy and paroxysomal dyskinesia	KCal.1
KCNQ2	BFNE, epileptic encephalopathy	KV7.2
KCNQ3	BFNE	KV7.3
KCNT1	ADNFLE, EIMFS	KNal.1
KCTD7	Progressive myoclonus epilepsy	KCTD7
HCN1	GGE	HCN1
CACNA1A	Epilepsy, episodic ataxia, epileptic encephalopathy	CaV2.1
CACNA1H	GGE	CaV3.2
	Ligand-Gated	
GRIN1	Epileptic encephalopathy	GluNl
GRIN2A	Epileptic encephalopathy	GluN2A
GRIN2B	Epileptic encephalopathy	GluN2B
GRIN2D	Epileptic encephalopathy	GluN2D
GABRA1	GGE, Epileptic encephalopathy	GABRA1
GABRB3	CAE, Epileptic encephalopathy	GABRB3
GABRG2	FS/GEFS+, epileptic encephalopathy	GABRG2
CHRNA2	ADNFLE	CHRNA2
CHRNA4	ADNFLE	CHRNA4
CHRNB2	ADNFLE	CHRNB2

Table 1.3: Various ion-channel genes known to be involved in idiopathic epilepsies and epileptic encephalopathies. The table has been adapted from the study [127]. BFIE, benign familial infantile epilepsy; BFNIE, benign familial neonatal-infantile epilepsy; EIMFS, epilepsy of infancy with migrating focal seizures; FS, febrile seizures.

Apart from ion-channels, several mutations in the non-ion channel genes have been discovered in epilepsy and they are shown in the Table 1.4.

Gene	**Gene Name**
NHLRC1	NHL repeat containing E3 ubiquitin protein ligase 1
SLC6A1	solute carrier family 6 member 1
KCTD7	potassium channel tetramerization domain containing 7
CPA6	carboxypeptidase A6
SLC25A22	solute carrier family 25 member 22
CLN8	CLN8, transmembrane ER and ERGIC protein
CDKL5	cyclin dependent kinase like 5
SNIP1	Smad nuclear interacting protein 1
RELN	reelin
PNKP	polynucleotide kinase 3'-phosphatase
EPM2A	EPM2A, laforin glucan phosphatase
SRPX2	sushi repeat containing protein, X-linked 2
STXBP1	syntaxin binding protein 1
WWOX	WW domain containing oxidoreductase
ST3GAL5	ST3 beta-galactoside alpha-2,3-sialyltransferase 5
SZT2	SZT2, KICSTOR complex subunit
PRICKLE1	prickle planar cell polarity protein 1
ASAH1	N-acylsphingosine amidohydrolase 1
STRADA	STE20-related kinase adaptor alpha
IER3IP1	immediate early response 3 interacting protein 1
SPTAN1	spectrin alpha, non-erythrocytic 1
PCDH19	protocadherin 19
SCARB2	scavenger receptor class B member 2
SLC35A2	solute carrier family 35 member A2
ARX	aristaless related homeobox
ARHGEF9	Cdc42 guanine nucleotide exchange factor 9
CNTNAP2	contactin associated protein like 2
PRICKLE2	prickle planar cell polarity protein 2
CSTB	cystatin B
PLCB1	phospholipase C beta 1
SYN1	synapsin I
SLC13A5	solute carrier family 13 member 5
SIK1	salt inducible kinase 1
DEPDC5	DEP domain containing 5
CHD2	chromodomain helicase DNA binding protein 2
GNAO1	G protein subunit alpha o1
ST3GAL3	ST3 beta-galactoside alpha-2,3-sialyltransferase 3
PRRT2	proline rich transmembrane protein 2
DNM1	dynamin 1
GOSR2	golgi SNAP receptor complex member 2
MEF2C	myocyte enhancer factor 2C
SLC2A1	solute carrier family 2 member 1
STX1B	syntaxin 1B
ALDH7A1	aldehyde dehydrogenase 7 family member A1
HNRNPU	heterogeneous nuclear ribonucleoprotein U
TBC1D24	TBC1 domain family member 24
LGI1	leucine rich glioma inactivated 1
PNPO	pyridoxamine 5'-phosphate oxidase
ALG13	ALG13, UDP-N-acetylglucosaminyltransferase subunit
EEF1A2	eukaryotic translation elongation factor 1 alpha 2 height

Table 1.4: A list of non-ion channel genes that were found to carry mutations in epilepsy. This table was modified from the study [137].

CNVs in epilepsy

Various studies carried out in GGE, EE or focal epilepsies have shown a clear role of CNVs in epilepsy. The CNVs involved in epilepsies are rare and consist of genes causing the epilepsies. One such example is deletion Xp22 which disrupts the gene *CDKL5* [138]. In the same study they found a deletion 5q33-q34 which spans the *GABRA1* and *GABRG2* genes both of which were shown to be associated to different forms of epilepsy [139, 140]. Similarly, various CNVs spanning genes such as *NRXN1, SCN1A, SCN2A, BMP5, AUTS2, PODXL, CNTNAP2, NIPA2, CYFIPI, CHNRNA7, NDE1, GRIN2A* and *PRRT2* have also found to be identified in various forms of epilepsies [141].

1.7.3 Rolandic Epilepsy

Background

Rolandic epilepsy (RE), is also known as benign epilepsy with centrotemporal spikes (BECT). It is one of the most common epilepsies occurring during childhood and it accounts for about 10–20% of the childhood epilepsies. The typical aao for RE is 3–13 years, with a peak incidence between 7–9 years old, and invariably shows remission by 14 years. The core clinical characteristic includes a focal seizure with sensorimotor symptoms, involving the face and laryngeal muscle, or secondary generalized tonic–clonic seizures, mainly during sleep. Characteristic centrotemporal spikes (CTS) and typical seizures are sufficient for diagnosis. The prognosis of RE is relatively benign, as the name indicates; however, moderate behavior and learning problems may exist in some patients. Compared to typical RE, atypical RE (ARE) includes more severe symptoms such as atypical benign partial epilepsy (ABPE), Landau–Kleffner syndrome (LKS), and epileptic encephalopathy with continuous spike and waves during sleep (CSWS). The symptoms observed in ARE occur together with speech and language dysfunction. The genetic origin of RE has been the subject of much speculation but remains largely unknown as most RE patients do not show a simple Mendelian inheritance pattern. Given their overlapping clinical characteristics, RE and ARE are presumed to have a shared genetic etiology. Hence, in this thesis they were studied together.

Genetics

There have been several family studies in the 1990s showing the genetic basis of RE. However, the twin studies did not provide a strong support to this hypothesis [142, 143]. All the studies that have been conducted till date on RE have remained inconclusive and the different genes emerging from different studies are pointing RE to be a complex disease [144].

Familial genes: A previous genome-wide linkage analysis of 38 families has shown that a chromosomal region 11p13 showed a significant linkage [145]. A further candidate SNP analysis across the linkage region identified three SNPs rs964112, rs11031434 and rs986527 all present in the intron region of *ELP4* gene with significant association. The result was replicated in a separate set of 120 controls and 40 cases from Canada, additionally a novel SNP rs2104246 also showed significant association. However, no such association for *ELP4* gene was found in a separate study [146] indicating that further larger sample sizes and functional studies are required in order to corroborate the findings.

Candidate gene studies: A study conducted on six CNVs has shown that a duplication in the chromosomal locus 16p11.2 increases the risk of RE and ARE which was supported by the association analysis. However, no such association was seen for temporal lobe epilepsy and Genetic generalized epilepsy (GGE)indicating an enrichment selectively towards RE and ARE. Further, candidate gene studies on RE cases with severe symptoms identified that mutations in *GRIN2A* and GABAergic receptors could play a potential role [147–149].

1.7.4 Genetic generalized epilepsy or Idiopathic generalized epilepsy (IGE)

Background

GGE is one of the most common types of epilepsy and constitute about one-third of epilepsies [150]. GGEs are believed to have a strong genetic component underlying the disease development and progression. According to the ILAE classification, the following are sub-types of GGE.

1. Benign myoclonic epilepsy in infancy (BMEI)
2. Generalized epilepsies with febrile seizures plus (GEFS+)
3. Epilepsy with myoclonic–astatic seizures (EMAS)
4. Childhood absence epilepsy (CAE)
5. IGEs with variable phenotypes (IGEVP)
    - Juvenile absence epilepsy(JAE)
    - Juvenile myoclonic epilepsy(JME)
    - Epilepsy with generalized tonic–clonic seizures only (EGTCSO)

Genetics

GGEs are considered to be of genetic origin and till date several twin studies have been conducted with high concordance [151]. About 2-8% of GGEs are considered to be monogenic and majority of the genes that are implicated in GGEs include ion-channel sub-units [152].

Although few genes apart from ion-channel sub-units are shown to be responsible for developing GGEs [153] their role remains inconclusive as it is often difficult to define the mechanism of epileptogenesis and assumed to have a functional interference by the ion-channel proteins [153].

Familial genes: Several genes implicated in GGE have been successfully identified by studying the families. Such genes include *SCN1A, KCNQ2, KCNQ3, EFHC1, GABRA2* and *CHRNA4*. Early studies based on twins have been conducted and further strengthened the argument that GGEs are genetic disorders. Till date about 2-8% of GGEs are considered to be monogenic [151]. Vast majority of the monogenic GGEs are associated to variants in voltage gated ion-channel receptors (Na and K channel subunits). The largest linkage study of 379 multiplex families has identified only two loci namely 5q34 and 13q31.3 as linked to GGE, whereas suggestive evidence was found for additional six loci 1p36.22, 3p14.2, 5q34, 13q12.12, 13q31.3, and 19q13.42 [154]. However, the linkage peak at 5q34 is interesting as it encodes for several GABA receptor sub-units(*GABRB2, GABRA6, GABRA1, GABRG2*).

Association studies: A GWAS study based on 3020 GGE cases and 3975 controls identified two loci 2p16.1 and 17q21.32 that are significantly associated to the GGE [155]. Other than that, the GWAS studies did not identify many significant genes. A large-scale meta-analysis comprising of GGEs identified a locus 4p15.1 harboring the gene *PCDH7* [156]. Although there are few common variants that were found to be associated to GGEs, a large proportion of genetic components is still missing and this could be potentially explained by the rare variants. But, there has not been much success in identifying the rare variants associated to GGEs via association studies. Especially, the studies using exome sequencing have failed to identify any significant rare variants [157]. Recently, a large-scale WES based association study comprising of 640 cases with familial GGE and 3877 controls did not detect any gene with genome wide statistical significance. However, an increased burden was observed when rare, deleterious variants in a group of genes with analogous function were collapsed.

CNVs in GGE: A large-scale association study including 1223 cases and 3,669 controls showed that micro deletions in the 15q13.3 region were present in 12/1223 cases but not present in any of the controls [158]. In the same line several micro deletions (1q21.1, 15q11.2, 15q13.3, 16p11.2, 16p13.11 and 22q11.2) were found to be associated to GGE [159, 160]. Another association study involving 1,408 GGE cases and 2,256 controls have identified large rare deletions in the gene RBFOX1 [161] that are associated to GGE.

1.8 Aims of the thesis

Although several rare variants and genes associated to PD and epilepsy have been identified till date, there is still a large portion of missing heritability. Hence, we hypothesized that both PD and epilepsy are genetically heterogeneous and in my thesis I aimed to fill in that missing gap by applying state-of-the art statistical and analytic methods to the WES and WGS data to both the diseases. My main focus was on rare/ultra-rare variants as they have the highest effect size compared to the common ones Figure 1.5. As part of it, I developed several pipelines in order to detect and analyze the variants from WES/WGS data and in the next steps they were applied to various data-sets belonging to epilepsy and PD.

The specific aims of my thesis were:

1. To discover disease-causing variants in different kinds of epilepsies namely RE/ARE and GGE.
2. To identify the differential burden of genetic variants in sporadic PD.
3. To discover potential novel disease causing variants in familial-PD.
4. To build disease prediction models based on WES/WGS data.

1.9 Contributions

- Chapter 1
    - Description: This chapter provides the background information regarding human genome and various neurological disorders.
    - Contributions: I wrote the text in full.
- Chapter 2
    - Description: Role of rare variants in Typical and Atypical rolandic epilepsy using the WES data were studied in this chapter. This chapter is a full reprint of the article published in European Journal of Human Genetics.
    - Contributions: Data analysis, interpretation of results, writing and revision of manuscript.
- Chapter 3
    - Description: Rare variants in a group of GABA_A receptors in GGE were studied using the WES data in this chapter. This chapter is currently under revision in Lancet Neurology.

  - Contributions: Data analysis, interpretation of results, writing and revision of manuscript.

- Chapter 4
  - Rare copy number variants in RE/ARE and GGE using the WES data were studied in this chapter. This chapter is currently submitted in PLOS ONE.
  - Contributions: Data analysis, interpretation of results, writing and revision of manuscript.
- Chapter 5
  - Description: Role of ultra-rare variants in PD using the WES of 367 cases and 159 controls were studied in this chapter. This chapter is currently submitted in Movement disorders.
  - Contributions: Data analysis, interpretation of results, writing and revision of manuscript.
- Chapter 6
  - Description: Burden analysis of 5' splice variants in PD using the WES data is described in this chapter. This chapter is a major part of a manuscript currently submitted in Cell.
  - Data analysis, interpretation of results, writing and revision of manuscript.
- Chapter 7
  - Description: Rare variants identified via WGS in >50 families are described in this chapter. This chapter is a major part of manuscript in preparation.
  - Contributions: Data analysis, interpretation of results, writing and revision of manuscript.
- Chapter 8
  - Description: Conclusions and outlook.
  - Contributions: I wrote the text in full.

CHAPTER 2

ROLANDIC EPILEPSY

2.1 Abstract

Rolandic Epilepsy (RE) is the most common focal epilepsy in childhood. To date no hypothesis-free exome-wide mutational screen has been conducted for RE and Atypical RE (ARE). Here we report on whole-exome sequencing of 194 unrelated patients with RE/ARE and 567 ethnically matched population controls. We identified an exome-wide significantly enriched burden for deleterious and loss-of-function variants only for the established RE/ARE gene *GRIN2A*. The statistical significance of the enrichment disappeared after removing ARE patients. A nominally significant enrichment for loss-of-function variants was detected for several disease-related gene-sets.

2.2 Introduction

Rolandic Epilepsy (RE), or epilepsy with Centro-Temporal Spikes (CTS), is one of the most common epilepsy syndromes of childhood. RE is related to rarer, and less benign epilepsy syndromes, including atypical benign partial epilepsy, Landau-Kleffner syndrome and epileptic encephalopathy with continuous spike-and-waves during sleep, referred to as RE related syndromes, or Atypical Rolandic Epilepsy (ARE) [162]. In up to 20% sib pairs or families, mutations affecting *GRIN2A*, a subunit of the excitatory glutamate receptor NMDA, were found implicated as major risk factor for RE and ARE by us and others [147, 148]. Recently, the association of the genes *RBFOX1*, *RBFOX3*, *DEPDC5*, *GABRG2* and genomic duplications at 16p11.2 in 1,5-2,0% was identified in patients with RE and ARE [149, 161, 163] through candidate gene and loci screens. In the current study, an unbiased exome-wide survey was conducted in the

RE/ARE cohort.

2.3 Patients and Methods

2.3.1 Study participants

204 unrelated European Rolandic cases (182 RE, 22 ARE) and 728 population control subjects were included [149]. Written informed consent was obtained from participating subjects and, if appropriate, from both patients and adolescents.

2.3.2 Data generation and processing

Exome sequencing of all individuals was performed with the Illumina HiSeq 2000 using the EZ Human Exome Library kit (NimbleGen, Madison, WI). Sequencing adapters were trimmed and samples with <30X mean depth or <70% total exome coverage at 20X mean depth of coverage were excluded from further analysis. Variant calling was performed in targeted exonic intervals with 100bp padding using the GATK best practices pipeline [23] against the GRCh37 human reference genome followed by multi-allelic variant decomposition and left normalization. Samples were excluded from further analysis if they (i) were not ethnically matched, (ii) were related, (iii) showed discrepancy with reported sex, (iv) had an excess heterozygosity >3SD in any of the quality metrics (NALT, NMIN, NHET, NVAR, RATE and SINGLETON statistics as calculated by PLINKseq i-stats parameter [164]. The genotypes of variants with read depth <10 or genotype quality <20 were set to missing. Variants were excluded if they (i) failed variant quality score recalibration (VQSR) or GATK recommended hard filter, (ii) showed missingness >3%, (iii) were present in repeat regions or (iv) had an average read depth <10 in either cases or controls. The ExAC variants were restricted to the exonic intervals used for variant calling in this study, not present in the repeat regions and passed the VQSR threshold.

2.3.3 Variant annotation and filtering

Variants were annotated using ANNOVAR [165] version 2015 Mar 22 with RefSeq and Ensembl, Combined Annotation Dependent Depletion (CADD) scores [58], allele frequencies and dbNSFP (v3.0) annotations. The sample used in this study are of NFE ancestry, hence to investigate rare variants, we selected variants having a minor allele frequency (MAF) <0.005 in the European populations of the 1000 genomes, Exome Variant Server (EVS) and the Non-Finnish European (NFE) data from ExAC. We generated three classes of variants for further analyses: (1) deleterious variants (CADD15) which were defined as missense variants with a CADD Phred score >15, (2) loss-of-function (LOF) variants comprising all rare indels, stop gain, stop loss and splice site variants (2nt plus/minus the exon boundary), (3) CADD15+LOF variants as the union of the above two datasets, and (4) rare synonymous variants.

2.3.4 Single variant and gene association analysis

For the statistical analysis, we employed two independent control cohorts (available in-house and ExAC) to increase reliability and power of the statistical tests. For single variant burden analysis, we applied the single score method in RVTESTS [76] to cases and in-house controls, for which individual genotypes were available. For gene burden analysis, a 2x2 contingency table was constructed by counting the number of alternate allele counts per gene in patients vs. controls (in-house controls and NFE ExAC controls). We then obtained a one-sided p-value, odds ratios and the 95% confidence intervals [166] by using Fisher's exact test. Resulting p-values were corrected for 18,668 RefSeq protein-coding genes [133] by Bonferroni approach. Finally, to ensure the exclusion of false positive association results and following the "rare variant of large effect hypothesis", we selected those genes that are present in the first quartile of the Residual Variant Intolerance Score (RVIS) distribution [167].

2.3.5 Selection of gene-sets

We investigated the following four neuron-related gene-sets: (1) genes encoding proteins at synapses downloaded from the SynaptomeDB [168] database ("SYNAPTIC_GENES", N=1887), (2) genes of postsynaptic signalling complexes including N-methyl-D-aspartate receptors (NMDARs) and the neuronal activity-regulated cytoskeleton-associated protein (ARC) [169] ("NMDAR_ARC_COMPLEX", N=80), (3) genes encoding proteins at the inhibitory synapses ("INHIBITORY", N=5,941) and excitatory synapses ("EXCITATORY", N=5,261) [170], (4) glutamate receptor subunit encoding genes ("GLUTAMATE_RECEPTORS", N=18). In addition, we included five gene-sets associated with disease and/or mutational intolerance: (1) genes encoding targets of Fragile-X-Mental-Retardation-1-Protein [171] ("FMRP_TARGETS_DARNELL", N=1,772), (2) genes intolerant for mutations from ExAC ("EXAC_CONSTRAINED_GENES", N=3,230), (3) genes intolerant for loss-of-function mutations [172] ("constrained") ("CONSTRAINED_GENES_SAMOCHA", N=1,004), (4) a curated list of dominant genes associated with developmental delay obtained from the DECIPHER database [173] ("DDG2P_MONOALLELIC", N=299), and (5) genes found related before to epileptic encephalopathies [174] (EPILEPTIC_ENCEPHALOPATHY, N=73). As control data sets we used (1) for each dataset the corresponding set of synonymous variants, and (2) the 'non-constraint' gene-set including RefSeq genes that have been found tolerant to loss-of-function mutations ("GENES_WITHOUT_CONSTRAINT", N=14,417). *GRIN2A*, as the most significant single gene from the burden analysis, it was excluded from all gene-sets in order to test if other genes also contribute to the disease association.

2.3.6 Gene-set association analysis:

The gene-set association analysis for the different types of variants was performed by using a logistic regression approach using R (version 3.2) and adjusting for the following confounding variables: the total number of called genotypes per sample, the total number of rare coding variants per sample, the total number of rare coding singletons (variants observed only once in the entire dataset) per sample, calculated sex, the first four principal components and the total number of variants per sample for each variant class.

2.4 Results

2.4.1 Exome sequencing and variant filtering

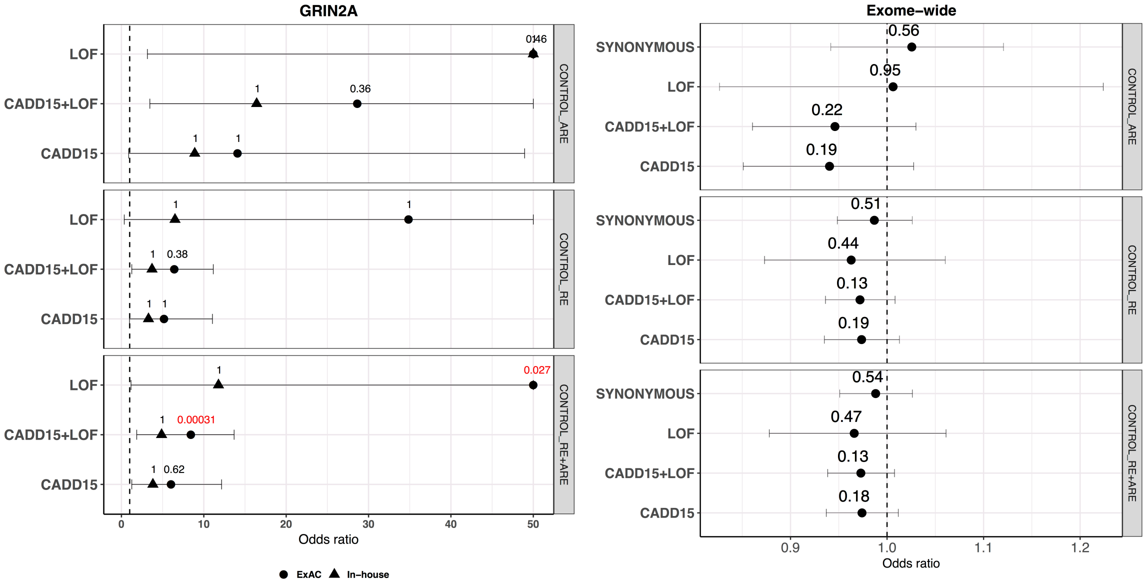
We performed whole-exome sequencing on 204 patients with RE/ARE and 728 population controls. After QC, the final dataset consisted of 19 ARE, 175 RE and 567 control samples. From the total of 761 samples, 226,521 exonic and splice site variants were called. The mean transition/transversion ratio equalled 3.39 per sample. After the final filtering 45,881 CADD15, 10,326 LOF and 38,802 synonymous variants were analyzed.

2.4.2 Association analysis

To investigate the mutational burden within the RE spectrum, all associations were assessed for both RE and ARE separately and by combining cases from both phenotypes while assuming them to be a single disease. In comparison to 567 in-house controls we did not observe statistically significant burden in any of the variants or genes in cases after multiple-testing correction. In order to increase the statistical power, we used the Non-Finnish European (NFE) ExAC cohort as an additional control dataset. Association testing against the much larger NFE-ExAC cohort (N=33,370) identified an exome-wide significant burden for CADD15, CADD15+LOF and LOF variants for *GRIN2A* within the combined typical and atypical (RE+ARE) cohort. No other variant-intolerant gene (i.e. being present in the first quartile of RVIS) was significantly enriched for variants in any of the tested patient groups. Although, variant enrichment for *GRIN2A* was not found to be significant after correction for RE and/or ARE separately, the odds ratio for *GRIN2A* consistently exceeded unity in all considered datasets (Figure 2.1A).

2.4.3 Exome-wide and gene-set burden analysis

Assuming a shared mutational burden in patients across gene-sets of convergent function and/or pathways, we performed gene-set burden analyses by using the in-house controls. A logistic regression approach was used to account for various confounding variables (see Methods). No significant exome-wide burden was observed across the different variant classes (Figure 2.1B). Despite the fact that none of the gene-sets showed a significant result after multiple-testing correction, we found several gene-sets with an odds ratio >1 for the CADD15, CADD15+LOF and LOF variant classes, especially for the LOF variants, but not for synonymous variants (Figure 2.2). A similar result was seen when we performed the analysis with ARE and RE independently.



(A)

(B)

Figure 2.1: Burden analysis of RE/ARE. Typical Rolandic Epilepsy is represented as RE, Atypical Rolandic Epilepsy as ARE and RE plus ARE as ROLANDIC. On the x-axis, the odds ratios in cases vs controls are given. The names of the variant classes are given on the y-axis. Each panel represents a different dataset. The dashed vertical line represents the expected odds ratio of 1. The horizontal lines indicate 95% confidence intervals. (A) Assessment of risk for deleterious mutations in GRIN2A against two control groups (ExAC and In-house). The values on top of each point represent multiple-testing corrected p-values, the ones in red are significant p-values and the ones in black are the p-values that are not significant after multiple-testing correction. The odds ratios are restricted to a maximum value of 50. (B) Exome-wide burden analysis by different variant classes. The values on top of each point represent the p-value. Synonymous variants serve as a control functional group.

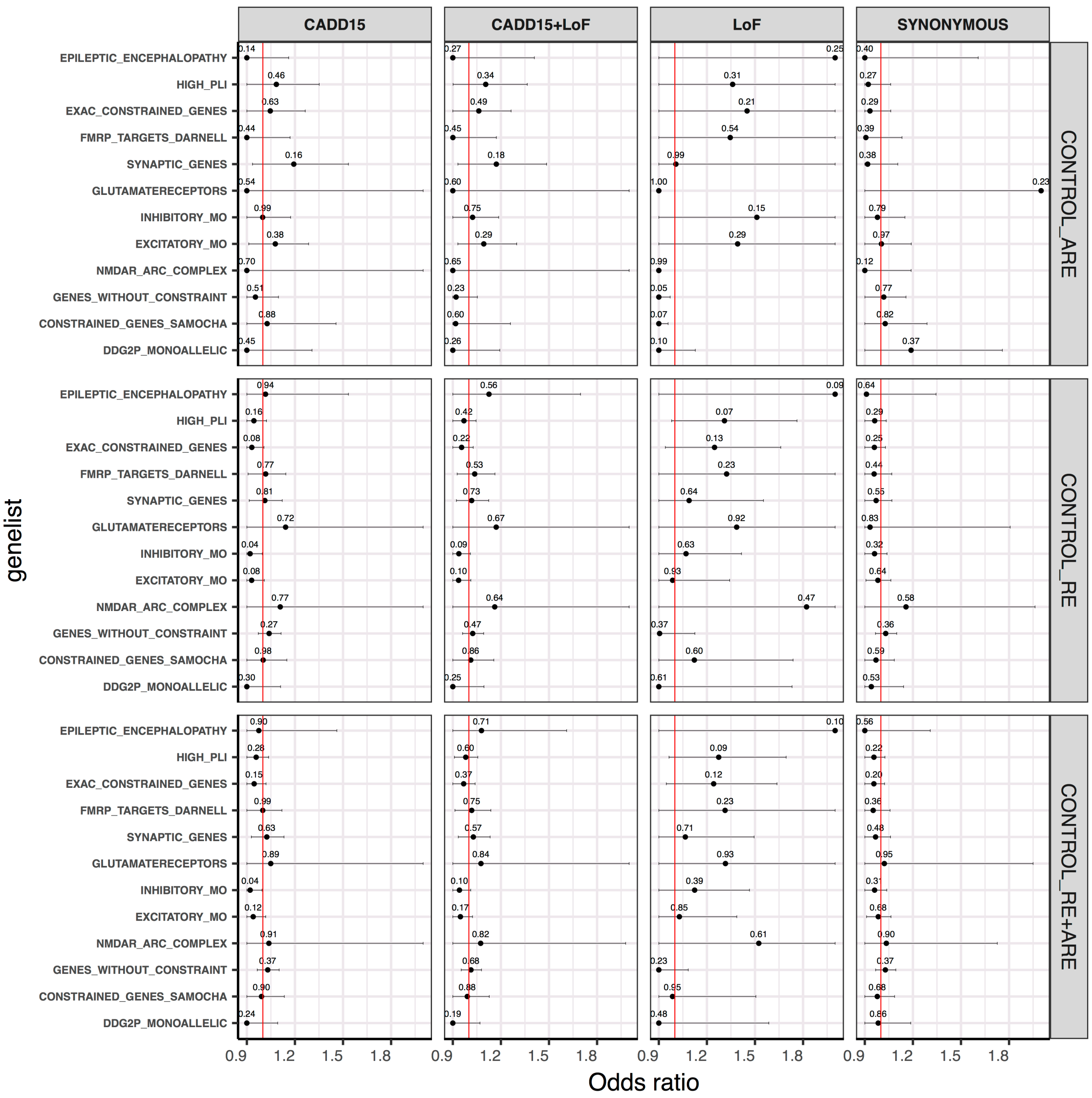


Figure 2.2: Gene-set burden across different variant classes. Each panel represent a different variant class. The synonymous variants serve as a control variant class. *GRIN2A* was removed from all gene-sets to identify other contributing genes. On the x-axis, the odds ratios in cases vs controls are given. On the y-axis the names of different gene-sets are given. The red vertical line represents the expected odds ratio of 1. The horizontal lines indicate 95% confidence intervals and are restricted to the maximum of odds ratios over all gene-sets. In that case, points are represented as the points without error bars to their right. The uncorrected p-values are shown on top of each point. CADD15=Deleterious predicted missense variants. LOF=Loss-of-function variants.

2.5 Discussion

We performed the first exome-wide association study investigating rare genetic variants of large effect in 194 patients with childhood focal epilepsies with centro-temporal spikes in comparison with 567 in-house and online available 33,370 population controls from the ExAC database.

By performing an unbiased gene-burden analysis of patients against the in-house and ExAC controls (Figure 2.1A), we show that only for *GRIN2A* rare CADD15, CADD15+LOF and LOF variants are significantly more frequent in typical and Atypical Rolandic Epilepsy (RE and ARE, respectively, odds ratio >1). Owing to the small sample size and genetic heterogeneity, no other gene or gene-set was significantly enriched for variants after correction for multiple-testing (Figure 2.2). However, we could observe a consistent trend in the odds ratios for the enrichment of LOF variants in several diseases associated gene-sets comprising genes under negative selection, glutamate receptors and genes associated with epileptic encephalopathies (Figure 2.2). These patterns indicate that besides the major disease gene *GRIN2A*, we identified several novel variants (that have not been seen in ExAC) in Table 1 from other less frequently mutated genes such as *DEPDC5, GABRG2* etc., indicating that they also contribute to RE and ARE. Availability of larger cohorts in the future should allow us to identify these other genes associated to RE/ARE.

CHAPTER 3

GENETIC GENERALIZED EPILEPSY

3.1 Abstract

Generalized epilepsy with genetic etiology (GGE) is the most common type of inherited epilepsy characterized by absence, myoclonic and generalized tonic-clonic seizures typically occurring with generalized spike-and-wave discharges on electroencephalography. Despite a high concordance rate of 80% in monozygotic twins, the genetic background is still largely unknown. Individuals included in the study were clinically evaluated for GGE. Whole-exome sequencing (WES) was performed for the discovery case cohort, the first replication case cohort and for two independent control cohorts. A second replication case cohort underwent targeted next-generation sequencing of the 19 known genes encoding subunits of GABA_A receptors and was compared to the respective GABA_A receptor variants of a third independent control cohort. Functional investigations were performed using automated two-microelectrode voltage clamping in Xenopus oocytes.

Statistical comparison of 152 familial index cases with GGE in the discovery case cohort to 549 ethnically matched controls revealed significant enrichment of rare missense variants in the ensemble of GABA_A receptor encoding genes in cases. The enrichment for these genes could be replicated in a second WES cohort of 357 sporadic and familial GGE cases and 1485 independent controls. Comparison of these genes in a second independent replication cohort of 635 familial and sporadic GGE index cases, based on candidate-gene panel sequencing, to a third independent control cohort confirmed the overall enrichment of rare missense variants in cases. Functional studies for two selected genes (*GABRB2, GABRA5*) showed significant loss-of-function effects with reduced current amplitudes in five of seven tested variants compared to

wild-type receptors. Our results suggest that functionally relevant variants in $\mathrm{GABA_A}$ receptor subunit encoding genes constitute a significant risk factor for GGE. This conclusion is based on an enrichment of rare variants in those genes in three independent case-control datasets and physiological studies revealing a loss of function for tested variants which are supposed to favor a neuronal dis-inhibition which is a well-known mechanism in epilepsy.

3.2 Introduction

In the recent past, gene discovery in monogenetic diseases, including familial and severe epilepsy syndromes, has been boosted by next generation sequencing yielding a steadily increasing number of disease-causing genetic defects. Unraveling the genetic origin of complex inherited disorders has, however, been more difficult. GGE comprises common epilepsies with generalized absence, myoclonic and tonic-clonic seizure [175]. It has a high heritability, as has been shown in twin studies [176] and represents a kind of ‘prototype’ of genetic epilepsy with complex inheritance.

A few single nucleotide polymorphisms in genome-wide association studies and altered copy number variations have been the major common risk factors identified so far in GGE. These, however, only explain a small part of the high heritability. Single gene defects in larger families with autosomal dominantly inherited GGE have been identified as disease-causing, e.g. in *GABRA1* or *GABRG2* encoding subunits of $\mathrm{GABA_A}$ receptors [139, 177], or in *SLC2A1* encoding the glucose transporter type 1 [178, 179]. Larger candidate gene or whole exome sequencing (WES) studies have not revealed a significant burden of mutations in single genes or groups of genes thus far [157, 180]. Only recently, a study running in parallel to the one reported here demonstrated mutational burdens of ultra-rare variants in gene-sets related to epilepsy [133].

We set out to investigate the burden of genetic mutations in mainly familial GGE by first testing hypothesis-free sets of genes related to the disease and disease-relevant pathways, validate the findings and follow-up with hypothesis-driven functional studies. We demonstrate the presence of such a genetic burden in one gene-set encoding the main inhibitory receptors in the mammalian brain, replicate the finding in two independent GGE cohorts and prove their functional significance by physiological investigations.

3.3 Patients and Methods

3.3.1 Participants

The discovery GGE exome sequencing case cohort included 152 subjects (after quality control (QC) of the exome sequencing data) with GGE from multiplex families which were collected by the Epicure and the EuroEPINOMICS-CoGIE consortia. All subjects were of European descent

(Italian n=69, German n=51, Dutch n=11, Danish n=8, British n=6, Finnish n=4, Swedish n=2, Greek n=1). The cohort included 88 females (58%). The primary GGE-diagnoses were childhood absence epilepsy (CAE, n=68), juvenile absence epilepsy (JAE, n=16), juvenile myoclonic epilepsy (JME, n=37), GGE with generalized tonic-clonic seizures alone (EGTC, n=24), early-onset absence epilepsy (EOAE, defined as beginning below 3 years of age, n=4), epilepsy with myoclonic absences (EMA, n=1) and unclassified GGE (n=2) (see Section 3.6 Table 3.1).

The age of epilepsy onset ranged from 1.5 to 38 years with a median of 10 years and all subjects had a normal development without obvious developmental delay or intellectual disability, although most were not formally tested. We included the few cases with EOAE, EMA and unclassified GGE since these entities in our view are close to classical GGE. For EOAE it has been recently suggested by a large study that it is likely genetically similar to classical CAE [181]. EMA may also have genetic overlaps with GGE [182] and we often find in family studies both well classified and unclassified GGE cases in the same pedigrees. The majority of cases (n=143, 94%) derived from multiplex families with at least two affected family members, thereof 76 families with three or more affected members. All cases had EEG changes consistent with GGE (see Table 3.1 Section 3.6).

The replication case cohort 1 consisted of 357 GGE cases (after QC) that were collected in six European countries (Belgium n=5, Germany n=174, Ireland n=22, Italy n=23, Netherlands n=61, and UK n=72) by the EpiPGX consortium. The cohort included 225 females (63%) and 132 males (37%). GGE diagnosis included CAE (n=55), JAE (n=28), JME (n=157), EGTC (n=19), and unclassified GGE (n=98). 92 cases (26%) derived from multiplex families with at least two affected members. 131 cases were sporadic, for the remaining 134 cases, familial history was not known. Age of epilepsy onset ranged from 0 to 49 years with a median of 13 years. All cases had EEG changes consistent with GGE (see Table 3.2 Section 3.6).

Two independent control cohorts for the case discovery and the replication cohort 1 were obtained from two independently sequenced cohorts from the Rotterdam study [183, 184] which were matched for ethnicity and sex (Section 3.6). All the control samples were at least 55 years old or older and were checked for several neurological conditions at baseline. As GGE is a disease with typical onset from childhood to adolescence, it is unlikely that people from this older control cohort could still develop GGE.

For the GABA_A receptor panel cohort (replication cohort 2), individuals were collected by referral from neurologists or pediatricians in Quebec, Canada, and in Europe by Epicure or CoGIE partners. The replication cohort 2 consisted of 631 subjects (after QC) with GGE that were collected from Canada (n=290) and five European countries (Germany n=153, Denmark n=58, Belgium n=71, Netherlands n=58, and Finland n=1). They included 390 females (62%) and 241 males (38%). Subjects were diagnosed with CAE (n=109), JAE (n=92), JME (n=189), EGTC (n=104), or unclassified absence epilepsy (n=137) not otherwise specified according to ILAE definitions [175] (Section 3.6 Table 3.3). 154 cases were familial with at least 2 affected family members, for 51 there was a positive family history of epilepsy but only one affected member was available, and the remaining 426 cases were sporadic. All cases had EEG changes consistent with GGE. A third independent set of controls was obtained from the UK10K project consortium [185]. A full list of the investigators who contributed to the generation of the UK10K data is available from `www.UK10K.org`. Funding for UK10K was provided by the Wellcome Trust under award WT091310 (EGAS00001000101,129,130,131,242,306). Data transfer agreements were made between the CRCHUM and the appropriate instances. A total of 639 ethnically matched individuals were selected from the exome control cohort (324 females and 315 males). The diagnosis of GGE in all case cohorts was based on detailed clinical interview, a full neurological examination and respective EEGs. Written informed consent was obtained from all subjects or their respective relatives and the study was approved by the local Ethical Committees. One affected individual of each family was selected for sequencing.

3.3.2 Procedures

For the discovery stage, paired-end whole-exome sequencing (WES) of cases and controls was performed with the Illumina HiSeq 2000 using the EZ Human Exome Library v2.0 kit (NimbleGen, Madison, WI). Cases and controls were sequenced at different locations, cases at the Cologne Center for Genomics, the controls in Rotterdam [183]. Sequencing adapters were trimmed and samples with <30X mean depth or <70% total exome coverage at 20X mean depth of coverage were excluded from further analysis. Variant calling was performed by using the GATK [23] best practices pipeline with the GRCh37 human reference genome (see Section 3.6).

The replication case cohort 1 was paired-end whole-exome sequenced at deCODE genetics (Iceland) on the Illumina HiSeq 2500 using the Nextera Rapid Capture Expanded Exome kit (Illumina). A second set of Rotterdam control samples was sequenced again in Rotterdam [184] using the EZ Human Exome Library kit (NimbleGen, Madison, WI). For all WES samples, we applied standard procedures for assessing potential population stratification for the European population as well as a relatedness check (Figure 3.73.83.6). To exclude low quality variants, we performed an additional filtering based on quality metrics of individual genotypes, using read depth and genotype quality as the filtering criteria. We excluded any variant position with mean depth of <10 in either cases or controls. For all WES samples the same exome regions file from the EZ Human Exome Library v2.0 kit was used. For the WES analysis, only samples with more than 30X mean coverage or more than 70% of the exome intervals covered by at least 20x mean coverage were included for the analysis (Section 3.6). For replication case cohort 2, a total of 19 genes encoding for all known subunits of $\mathrm{GABA_A}$ receptors were selected for deep sequencing (*GABRA1, GABRA2, GABRA3, GABRA4, GABRA5, GABRA6, GABRB1,* *GABRB2, GABRB3, GABRD, GABRE, GABRG1, GABRG2, GABRG3, GABRP, GABRQ, GABRR1, GABRR2, GABRR3*, altogether referred to as *GABRX* herein).

Exon targets were generated based on RefSeq, representing 184 exons from 19 genes. Primer design was made using the Primer3 oligonucleotide design tool and in silico PCR tool for validating the specificity of each amplicon. Target regions were enriched by PCR using the 48.48 Access Array Integrated Fluidic Circuit (IFC) (Fluidigm, San Francisco, CA). In the final assay, 185 amplicons targeted the protein-coding sequence of 19 *GABRX* genes with an overhang at exon boundaries in order to capture splice site variants. *GABRX* exon-specific primers with Fluidigm tags were tested along with materials and reagents as recommended in the Access Array System User Guide (Fluidigm, South San Francisco, CA). Finally, GABRR3 had to be dropped because of QC reasons having not enough good quality reads covering this gene. After quality trimming the reads were mapped against the GRCh37 human reference genome using the GATK [23] suite and the MUGQIC pipelines (`https://bitbucket.org/mugqic/mugqic_pipelines`). Data from the control cohort were processed using the same pipelines. Coverage comparisons were made to keep bases covered in at least 95% of the subjects as well as the control cohort. RefSeq gene annotation information was used for the classification into missense and synonymous variants and to filter for rare (allele frequency smaller than 0.5%) variants using the ExAC database [43] (for details see Section 3.6).

3.3.3 Population stratification

We applied principal-components analysis (PCA) to assess potential population substructure separately for each case-control cohort, using the implementation in Eigenstrat [42]. Population outliers were defined as SD of >3 based on the first 10 PC and excluded from further analysis (Section 3.6).

3.3.4 Statistical analysis

Due to the limited sample size, single-gene collapsing analysis for the discovery stage was performed using Combined and Multivariate Collapsing [186] (CMC) method for collapsing and combining rare variants together with a two-sided Fisher's exact test, as implemented in the Exact CMC method in rvtests [76] (Section 3.6). P-values for single-gene collapsing tests were corrected for multiple testing by use of the Bonferroni method (as implemented in the R function p.adjust) for 18,668 protein-coding genes. For all three stages, gene-set collapsing tests were performed using the regression-based two-sided SKAT-O test method [75], as implemented in rvtests [76]. For the two WES cohorts, SKAT-O was used and we included sex and the first 10 PC from the Eigenstrat analysis as covariates to account for possible gender and population substructure effects. Gene-set collapsing tests were applied separately to missense and to synonymous variants of specific sets of candidate genes. Seven different disease- and process-specific gene sets were constructed based on their relation to GGE together with a control gene set not related to GGE.

A description of the gene-set construction is given in the Section 3.6 and the gene sets are given in the Section 3.6 in Table 3.5. In order to control the family-wise error rate, we applied Holm's correction for multiple testing 14 hypotheses, namely seven gene sets combined with two sets of variant type (missense and synonymous), in the discovery cohort, while correction was done for only two hypotheses in each of the two replication cohorts, since only the $GABA_A$ receptor gene set was carried forward the replication (Section 3.6). The odds ratio (OR) for a given gene-set was determined by comparing the presence of qualifying rare (nonsynonymous or synonymous) variants in all genes within each gene-set between cases and controls.

3.3.5 Functional analysis

Functional experiments were performed using automated two-microelectrode voltage clamping in *Xenopus* oocytes. All methods for functional studies have been described in Section 3.6.

3.4 Results

We first performed WES in a cohort of 238 independent, mainly familial cases of classical forms of GGE, i.e. childhood or juvenile absence epilepsy (CAE/JAE), juvenile myoclonic epilepsy (JME) and epilepsy with generalized tonic-clonic seizures on awakening (EGTCA), collected by the Epicure and EuroEPINOMICS-CoGIE consortia. As controls, we used ethnically and sex matched (Section 3.6 Figure 3.6, 3.7) population control individuals from the Rotterdam Study [183], that underwent WES using the same enrichment and sequencing procedures, albeit with a somewhat lower coverage than in the GGE cohort. After quality control (QC) and population outlier removal, the final dataset consisted of 152 unrelated GGE and 549 unrelated control samples. To adjust for the different coverage between case and control samples, we considered only variants with an average read depth of >10 both in case and in control samples (see Section 3.6, Figure 3.7).

From the total of 701 samples, 204,023 exonic and splice site variants were called. The mean exonic transition/transversion ratio equaled 3.46, indicating good data quality. Rare variants ($MAF<0.005$) were classified as missense (Nonsyn) and silent (Syn) variants. 82,579 Nonsyn and 48,450 Syn variants constituted the analysis data set (see Section 3.6, Table 3.4). First, we tested hypothesis-free all individual RefSeq genes separately for association but could not identify any single gene enriched for any variant type (Section 3.6). Therefore, we next applied an independent hypothesis-driven analysis by testing the enrichment of rare variants in seven gene sets related to epilepsy and its underlying molecular processes.

These gene sets represented (i) all voltage-gated cation channels, (ii) all excitatory postsynaptic receptors, (iii) all $GABA_A$ receptors as the main inhibitory postsynaptic receptors, (iv) more broadly the GABAergic pathway (since such genes have been associated specifically with generalized epilepsies), and genes associated (v) with generalized epilepsies, (vi) with epileptic encephalopathies, (vii) with focal epilepsies (Section 3.6 Table 3.5). We tested separately for each variant type; silent variants were expected to show no difference between cases and controls. We found a significant enrichment for missense variants in a set of $\mathrm{GABA_A}$ receptor genes (19 genes, p_{Nonsyn}=0.019, OR=2.40, 95% CI=[1.41,4.10]) by use of the SKAT-O test after multiple-testing correction (Figure 3.1). None of the other gene sets showed a significantly increased burden of rare variants. Synonymous variants, used as a negative control, did not show a significant enrichment in any of the gene sets (Section 3.6, Tables 3.6.6, 3.8).

To replicate the finding for the $\mathrm{GABA_A}$ receptor encoding genes, we first used the replication case cohort 1 collected by the EpiPGX consortium, consisting of 724 individuals with GGE from six European countries. They were mainly sporadic (n=268) or of unknown familial history (n=265) and diagnosed with classical forms of GGE (Section 3.6, Table 3.2). For the analysis of this cohort, an independent matched subset of control samples from the Rotterdam Study [184] was used. After applying the same QC steps as applied to the discovery cohort, the final dataset consisted of 357 unrelated GGE and 1,485 unrelated control samples [184]. We confirmed the significant enrichment of rare missense variants in $\mathrm{GABA_A}$ receptor genes in cases compared to controls after multiple-testing correction for two sets of variants (nonsyn and syn; p_{Nonsyn}=0.016, OR=1.46, 95% CI=[1.05,2.03]) by use of the SKAT-O test (Section 3.6 Table 3.6). Synonymous variants showed again no significant enrichment.

For a second replication cohort, we designed a targeted enrichment panel comprising all 19 $\mathrm{GABA_A}$ receptor genes. All genes were sequenced in an independent cohort of 631 cases with familial or sporadic GGE (Section 3.6, Table 3.3). *GABRR3* was excluded for QC reasons. We obtained control samples from the UK10K project (`https://www.uk10k.org/`) and selected 639 gender matched individuals after sample QC. Additional variant QC reduced the number of individuals to 583 cases and 635 controls in the final sample set. We found a significant enrichment of rare missense variants for the $\mathrm{GABA_A}$ receptor genes in cases compared to controls (p_{Nonsyn}=0.027, OR=1.46, 95% CI=[1.02,2.08]) by use of a SKAT-O test and after correction for two sets of variants (Nonysn, Syn). No significant enrichment was observed for synonymous variants. Thus, we can conclude that enrichment of rare missense variants in $\mathrm{GABA_A}$ receptor genes are reproducibly present in individuals with GGE when compared to controls. All detected case-only variants are given in the Section 3.6, Tables 3.6.6,3.8. Case-only rare missense variants were found in all $\mathrm{GABA_A}$ receptor genes except in GABRR3 (Table 3.8).

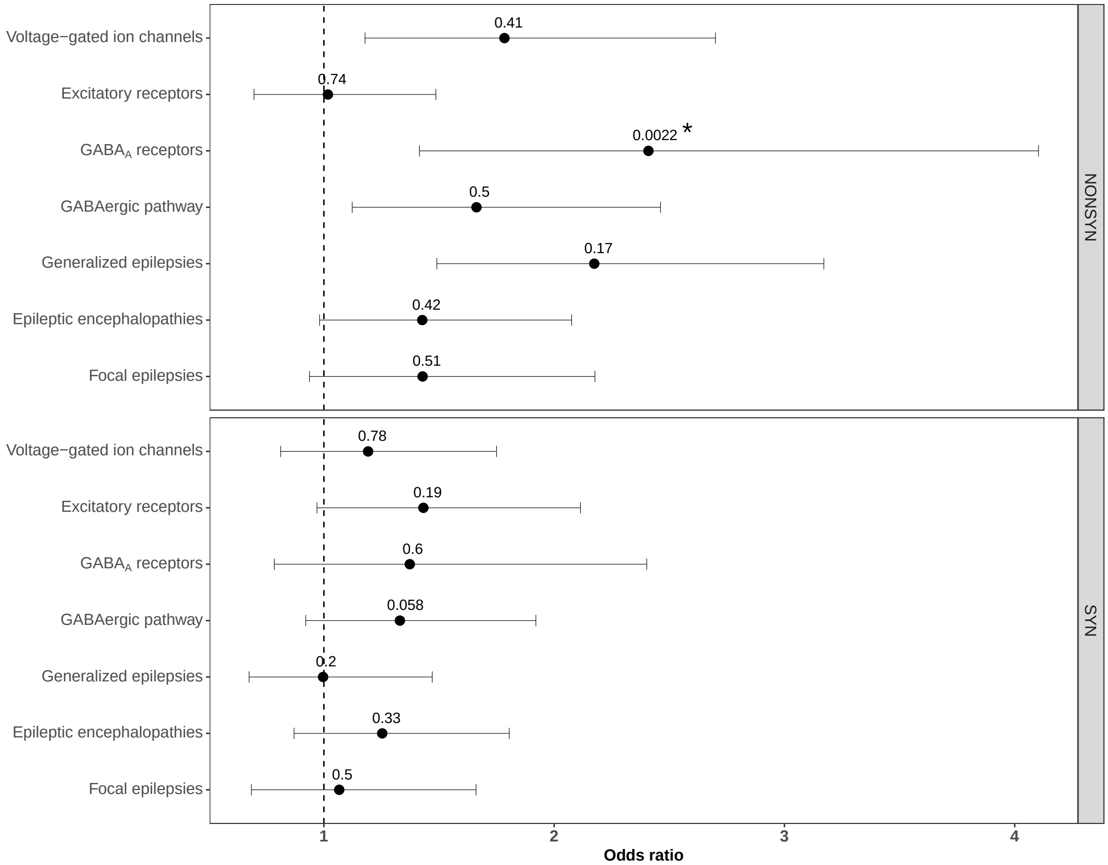


Figure 3.1: Rare variant gene-set odds ratios and burden enrichment for rare variants in the whole-exome sequencing GGE discovery cohort. Cases from the CoGIE discovery cohort, matched with controls from the Rotterdam study. Gene-set collapsing analysis by use of a SKAT-O test was performed on seven epilepsy-related gene sets for missense (NONSYN) and synonymous (SYN) variants. The gene sets are described in the Section 3.6. The star denotes the enriched gene-set collapsing p-value after Holm correction.

The combination of two α_1-, two β_2- and one γ_2- subunit (genes *GABRA1*, *GABRB2*, *GABRG2*) represents the most common form of a functional GABA$_\text{A}$ receptor in the brain [187], and variants in *GABRA1* and *GABRG2* have been shown to play an important role in familial GGE, febrile seizures and EE [139, 177, 188–191].

It is important to note that the enrichment of missense variants in in GABA$_\text{A}$ receptor genes was not driven by variants in those known epilepsy genes. The signal was no longer significant when reducing the analysis only to those two genes (Table 3.8). Instead, the qualifying variants were evenly distributed over all GABA$_\text{A}$ receptor encoding genes. The α_5 subunit (gene *GABRA5*) is supposed to mediate extrasynaptic tonic inhibition [192], and tonic inhibition has been described to be altered in genetic mouse models of epilepsy [193, 194]. *GABRB2* and *GABRA5* have not previously been associated with GGE, although *GABRB2* mutations were described recently in patients with intellectual disability and epilepsy [195–197].

For functional studies, we selected seven missense variants in *GABRB2* and *GABRA5* (Section 3.6, Table 3.9) for electrophysiological studies in *Xenopus* oocytes (Section 3.6). Five of these variants were selected since they co-segregated with the phenotype of available members in nuclear families. Another variant (p.R3S) was found in three different French-Canadian pedigrees, so we hypothesized whether this could be a more common causal variant in a specific population (Figures 3.2a and 3.3a). The last variant, p.P453L, did not co-segregate, but was selected as another variant in *GABRA5* which is localized in a different protein region (the C-terminus) than the other variants. All missense variants were predicted to be deleterious by at least three out of seven missense prediction tools and were highly conserved (Table 3.9). Three of these variants were consistently of ultra-low frequency in the European population in different public databases (dbGAP, 1000G, ExAC; Section 3.6,Table 3.9).

The localization of the variants in the $\mathrm{GABA_A}$ subunits is shown in Figures 3.2b and 3.3b. After application of 1 mM GABA, we observed a significant reduction in current amplitudes of $\mathrm{GABA_A}$ receptors containing either p.K221R or p.V316I variants in the β_2-subunit, and p.M1I, p.S238N, or p.E243K in the α_5-subunit, in comparison to respective compositions of WT receptors. No significant reductions were observed for p.R3S in the β_2- and for p.P453L in the α_5-subunit (Figures 3.3d, 3.3e, 3.2d, 3.2e,). The GABA sensitivity was studied by applying different GABA concentrations with no significant changes observed for receptors containing any of the studied variants (Figures 3.2e and 3.3e). Thus, five out of seven variants suggest a loss of receptor function predicting postsynaptic or extrasynaptic neuronal disinhibition.

All variants showing significantly reduced current amplitudes co-segregated with the disease phenotype in family members that were available for testing (Figures 3.2a and 3.3a), corroborating their contribution to the disease phenotypes. In two families, we observed co-segregating variants in two different $\mathrm{GABA_A}$ receptor subunits: p.V316I in the β_2- and p.M1I in the α_5-subunit co-occurred in the same nuclear family, and p.E243K in the α_5-subunit was accompanied by a deleterious frameshift mutation in *GABRG2* in another family (Figures 3.2a and 3.3a). Variants with altered receptor function were all located in the N-terminus containing GABA binding sites or in the pore region. p.M1I suppresses the start codon such that translation starts six amino acids later, which shortens the signalling peptide consisting of the first 20 amino acids.

While the peptide is removed and not part of the mature $\mathrm{GABA_A}$ receptor in the plasma membrane, this alteration could still affect the protein biogenesis and lead to reduced expression of functional receptors. p.R3S, which also affects the signalling peptide, and p.P453L, located in the functionally less relevant C-terminus, did not lead to a significant change in receptor function. p.R3S recurred in three French-Canadian families and p.P453L was detected in only one of several affected members of a larger family indicating that they might represent benign polymorphisms.

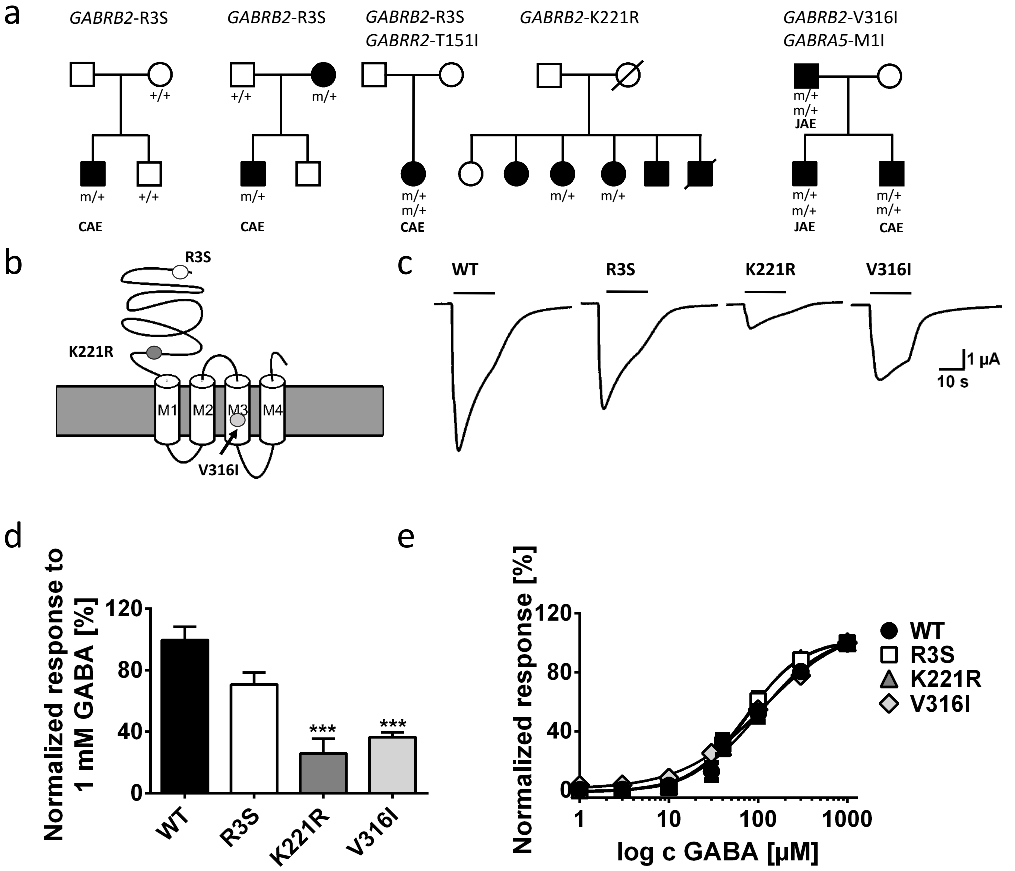


Figure 3.2: *GABRB2* mutations associated with GGE. (a) Pedigree of the families. (b) Schematic representation of the β_2 subunit of the $\mathrm{GAB_AR}$ and predicted positions of the R3S and K221R mutations located in the N-terminal domain and V316I located in the transmembrane domain 3. (c) Examples of GABA-induced currents after 1 mM GABA application for WT, R3S, K221R and V316I mutations. (d) Current responses normalized to 1 mM GABA application for WT (n=30), R3S (n=24), K221R (n=21) and V316I (n = 16); ***$p<0.001$, Kruskal Wallis test, with Dunn´s comparison test. (e) Dose-response curve for α_1 β_2 γ_2s WT (n=30), R3S (n=14), K221R (n=10), V316I (n=7) obtained using application of different GABA concentrations and normalization to the maximal GABA response for each cell.

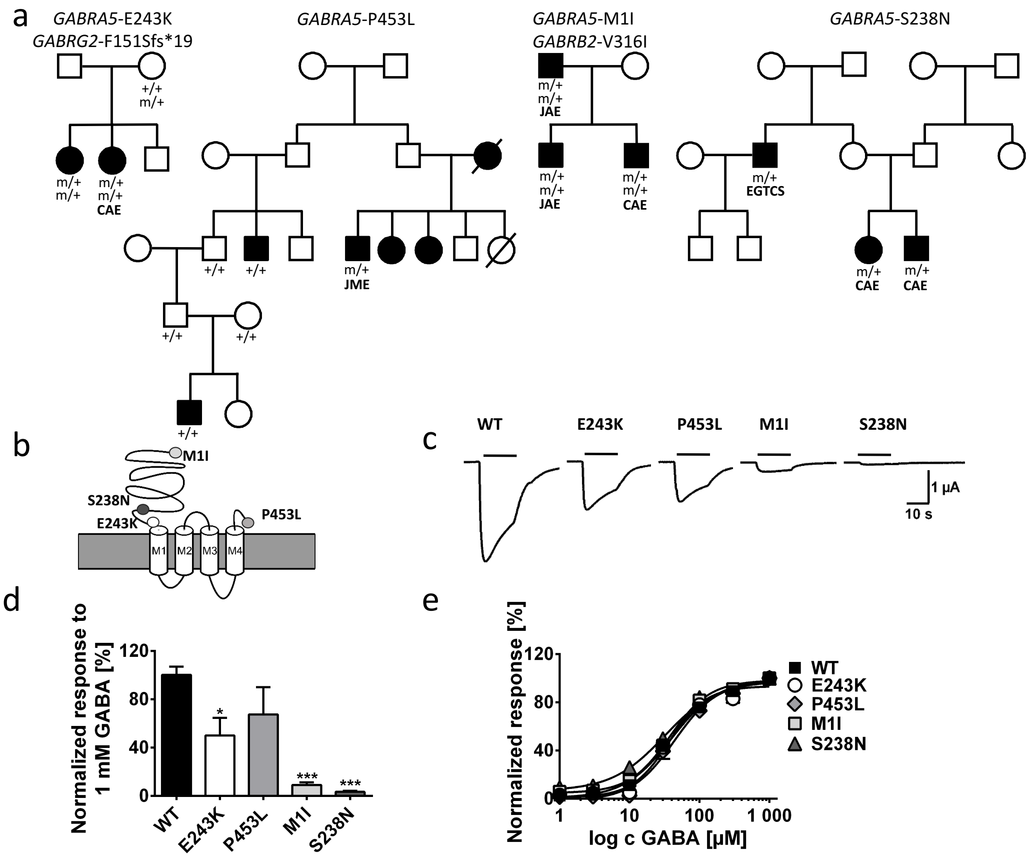


Figure 3.3: *GABRB2* mutations associated with GGE. (a) Pedigree of the families. (b) Schematic representation of the α_5 subunit of the $\mathrm{GABA_AR}$ and predicted positions of the M1I, S238N and E243K mutations located in the N-terminal domain and P453L located in the C-terminal domain. (c) Examples of GABA-induced currents after application of 1 mM GABA for WT, M1I, S238N, E243K and P453L mutations. (d) Normalized current responses to 1 mM GABA application for WT (n=43), M1I (n=10), S238N (n=13), E243K (n=14) and P453L (n=11); $*<p0.05$, $***p<0.0001$, Kruskal Wallis test, with Dunn´s comparison test. (e) Dose-response curve for α_1 β_2 γ_2s WT (n=37), M1I (n=15), S238N (n=11), E243K (n=8) and P453L (n=8) obtained after application of different GABA concentrations and normalization to the maximal GABA response for each cell.

3.5 Discussion

Our results show an excess of rare missense variants in $\mathrm{GABA_A}$ receptor subunit encoding genes in three independent cohorts of altogether >1000 familial and sporadic GGE index cases. Five selected variants in two genes, GABRB2 and GABRA5, previously not associated with GGE (i) clearly changed receptor function and (ii) co-segregated in nuclear families, suggesting an important contribution to the disease phenotype and inheritance in those pedigrees. Previous studies in smaller cohorts failed to show a significant excess of variants in cases versus controls either in a test for the set of all ion channel encoding genes [180] or in whole exomes [157], Our findings indicate that the enrichment of rare genetic variants in a set of inhibitory GABA receptors does play a significant role in the pathogenesis of GGE. The difference between these previous studies and ours could be explained by (i) a larger sample size in our study across all cohorts and (ii) by testing different gene-sets that had not been considered before.

In a parallel study [133], a similar effect could be shown for ultra-rare deleterious variants in gene-sets comprising known epilepsy genes or genes associated with epileptic encephalopathies. Due to our smaller sample size and the associated low number of ultra-rare variants, we here chose a different approach considering all variants with a MAF<0.005, which proved to yield significant results in other studies [65, 198, 199]. Both studies failed to identify single genes with a genome-wide significant burden of rare variants in individuals with GGE. It will be interesting in future studies to combine different cohorts to increase sample size and power for such analyses to shed further light on the complex genetic architecture of GGE.

One limitation of the current study is that the cohorts, due to funding restrictions of the individual projects, were sequenced at different locations using different technologies. Combining and analyzing such data in an unbiased way is still a major challenge in large genetic sequencing projects. An a priori selection bias for the targeted genes yielding a false significance can also not be completely ruled out. The careful choice of gene sets was based on purely biological and published evidence and did not change the selection afterwards. This approach should minimize any potential selection bias and associated false-positive findings. We addressed these issues by using a stringent QC and consistent processing of all datasets (Section 3.6), and by using two different GGE cohorts to replicate our data in independent case and control datasets.

One of the variants detected and functionally examined in our study (p.V316I in *GABRB2*) has been described in the meantime as a de novo mutation in a different dataset of cases with severe developmental and epileptic encephalopathies, in which whole genome sequencing of parent-patient trios was used [195]. This finding clearly corroborates the pathogenicity of this variant. The association of genetic variants with different phenotypes is well-known as the phenomenon of pleiotropy, and has also been described in other $\mathrm{GABA_A}$ receptor encoding genes [140, 188] including large phenotypic variability within one extended pedigree [140].

We have also recently characterized the variant p.T336M in GABRA3 – which was detected in our discovery cohort (Table 3.8) – as part of another study in which we identified GABRA3 as a new epilepsy gene associated with highly heterogeneous epileptic phenotypes including asymptomatic variant carriers [200]. This variant also causes a severe loss-of-function effect but does not co-segregate in the respective pedigree, so that other factors must contribute to the GGE at least in two family members. While co-segregation is a strong indicator for the pathogenicity of genetic variants, we have to be aware that GGE is a common disease with complex inheritance. Variants in $\mathrm{GABA_A}$ receptor encoding genes could therefore still contribute to the disease in the carriers, whereas other family members not carrying the respective variants must have other genetic causes of their epilepsy. Similarly, copy number variations often do not co-segregate within nuclear families but have been replicated several times as a significant risk factor for GGE [158, 160, 201, 202].

Given the reproducibility of our results in three independent datasets together with co-segregation and functional evidence for GABA_A receptor dysfunction, many but not all of the detected variants probably contribute to the etiology of GGE in our three cohorts. This disease-relevant contribution may range from a major gene effect – as observed in 'monogenic' Mendelian epilepsies – to relatively small effect sizes in the variant carriers, depending on the amount of the electrophysiological dysfunction and probably other unknown factors, such as the genetic background. Overall, we therefore consider the detected increase in GABA_A receptor variants in cases vs. controls as a significant risk factor to develop GGE.

Lastly, our results indicate a genetic overlap among rare and common forms of epilepsy, since there is increasing evidence that de novo variants in GABA_A receptor encoding genes cause severe forms of epileptic encephalopathies [187, 188, 190, 191, 200, 203, 204] and they re-iterate a central role of GABAergic mechanisms in generalized epilepsies [139, 140, 177, 188, 191–194, 197, 203–206]

3.6 Additional methods and results

3.6.1 Patient cohorts

CoGIE (discovery cohort (European))

Gender	64 males, 88 females
Age of Onset	Mean 9.98 years, Median 10 years
Affected family members (n)	1 (9), 2 (67), 3 (43), 4 (23), 5 (5), 6 (3), 7 (1), 8 (1)
Origin (n)	Italian (69), German (51), Finnish (4), Dutch (11), British (6), Danish (8), Swedish (2), Greek (1)
Epilepsy diagnosis (n)	CAE (68), JME (37), EGTC (24), JAE (16), EOAE (4), EMA (1), unclassified GGE (2)

Table 3.1: Discovery cohort: GGE diagnosis, phenotype data.

EpiPGX (Replication cohort 1, European)

Gender	132 males, 225 females
Age of Onset	Mean 9.98 years, Median 13 years
Affected family members (n) 0	92
Origin (n)	Belgium (5), Germany (174), Ireland (22), Italy (23), Netherlands (61), UK (72)
Epilepsy diagnosis (n)	CAE (55), EGTC (19), JAE (28), JME (157), unclassified GGE (98)

Table 3.2: Replication cohort 1: Patients with GGE diagnosis, phenotype data.

GABA_A panel cohort (Replication cohort 2, European/French-Canadian)

Gender	132 males, 225 females
Age of Onset	Mean 9.98 years, Median 13 years
Affected family members (n) 0	92
Origin (n)	Belgium (5), Germany (174), Ireland (22), Italy (23), Netherlands (61), UK (72)
Epilepsy diagnosis (n)	CAE (55), EGTC (19), JAE (28), JME (157), unclassified GGE (98)

Table 3.3: Replication cohort 2 for GABA_A receptor gene panel sequencing: GGE diagnosis, phenotype data.

Rotterdam Study control samples

The WES control samples used in this study were both part of the Rotterdam study [183, 184] but were sequenced independently. The Rotterdam samples were obtained from Ommoord district in the city of Rotterdam in The Netherlands. All the control samples were at least 55 years old or older. They were checked for several neurological conditions at baseline. Only population matched control samples of European origin were selected for the discovery and replication cohort 1, respectively, using the Eigenstrat [42] selection procedure described above. The control samples for the discovery cohort included 177 males and 372 females (68%), the control samples for the EpiPGX replication cohort 1 included 596 males and 889 females (60%).

3.6.2 Exome sequencing analysis (Discovery cohort)

Data pre-processing

Sequencing adapters were removed from the FastQ files by using cutadapt [207]. GATK [23] best practices (`https://software.broadinstitute.org/gatk/best-practices`, version 3.2). were followed for the next steps of data pre-processing and variant calling. Alignment was performed using bwa-mem [29] with default parameters to the GRCh37 human reference genome. Conversion of sam to bam files was done by samtools [27]. Sorting of bam files, marking of duplicate reads that remain after PCR amplification and addition of read group information was done by using picard with default parameters. Using GATK version 3.2, base quality scores recalibration, local realignment for small insertions and deletions (InDels) was performed. All samples with less than 30X mean coverage or less than 70% of the exome intervals covered by at least 20x mean coverage were excluded from the analysis. A multiple sample calling approach was employed using GATK.

Sample filtering based on quality metrics

Number of alternate alleles, number of heterozygotes, number of variants called, number of minor alleles, number of singletons and call rate served as data quality parameters. They were calculated by using PLINK/SEQ (`https://atgu.mgh.harvard.edu/plinkseq`) i-stats command. Any sample with >3 standard deviations (SD) from the mean in any of the used metrics was excluded from the analysis. Next, we selected the variants that are common between hapmap [37] (version 3.3) and the current dataset. The selected variants were further filtered to be: 1) Only bi-allelic SNVs, 2) with a call rate >98% and 3) not in linkage disequilibrium. The variants selected above were used to check cryptic relatedness, deviations from reported sex and to perform correction of population stratification by using Eigenstrat [42].

Relatedness and sex check

In order to check the relatedness between each pair of samples within a cohort, the PLINK [77] "–genome" command was used to identify the fraction of genome shared identical-by-descent (IBD). For pairs with PI_HAT score of >0.25 (see Figure 3.6), the sample with lower mean depth of coverage was removed from further analysis.

Sample contamination

We checked for sample contamination between different samples by using the inbreeding coefficient as measure. PLINK [77] "–het" command was used with default parameters to calculate the inbreeding coefficient. Any sample exceeding >3SD in the output "F" value was excluded from the analysis. Individuals with high missingness can lead to bias in the results. High heterozygosity is an indicator of possible sample contamination. Hence, to mitigate this effect we excluded samples with missingness > 10%.

Population stratification

We merged our data with the 1000genomes [34] (1000g) data and assessed individual ancestry by use of a principal-components analysis, using Eigenstrat [42] with default parameters. Except for few outliers, cases and controls clustered with the samples of European origin in 1000g data (Figure 3.7). We then merged our data with only the Central European (CEU) and Toscanian Italian samples (TSI) from the 1000genomes cohort. Then, by using Eigenstrat with a sigma value of 3, which excludes all the samples with a SD of >3 based on the first 10 principal components, we excluded population outliers (Figure 2B, bottom).

Filtering of low quality variants

Initial filtering of variants was performed based on quality metrics over all the samples with the parameters below, for VQSR: Tranches chosen, VQSRTrancheSNV99.90to100.00. The QC parameters for hard filtering over all samples were: a) for SNVs: QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, DP<10.0, GQ_MEAN<20.0, VQSLOD<0, <5% missingness, ABHet > 0.75 or < 0.25 and Hardy Weinberg Phred scale P value > 20. b) for InDels: QD < 2.0, FS > 200.0, ReadPosRankSum < -20.0, DP < 10.0, GQ_MEAN < 20.0, Hardy Weinberg Phred scale P value>20, VQSLOD>0.

In-order to exclude low quality variants, we performed an additional filtering based on quality metrics of individual genotypes, using read depth and genotype quality as the filtering criteria. Therefore, variant genotypes with a read depth of <10 and GQ of <20 were converted to missing by using bcftools [46]. In addition, we calculated the mean read depth of each position by using vcftools [33] 12 in cases and controls separately, and any variant position with mean depth of <10 in either cases or controls was excluded. Different quality metrics were calculated by the GATK 'varianteval' tool and are shown in Table 3.4.

Variant annotation and filtering

Multi-allelic variants were decomposed using variant-tests and left normalized by bcftools. Variants were annotated using ANNOVAR [208](version 2015Mar22) using RefSeq and Ensembl versions 20150322 and the dbNSFP [51] (version 2.6) annotations and pathogenicity scores. As we were interested in rare variants, we filtered the variants for a minor allele frequency (MAF) < 0.005 in the European populations of public databases like 1000 genomes, dbSNP, ExAC [43] (release 0.3, NFE and ALL), and the Exome variant server (EVS). In addition, we excluded all variants with an AF > 0.005 in the dataset from this study. We created two variant subsets on which we performed the statistical analysis: 1) nonsynonymous variants (NONSYN) and 2) synonymous variants (SYN).

Variant statistics

Quality metric	After QC
Number of samples	152 cases 549 controls
Number of variants	472,970
Number of exonic/splicing variants	204,023
Ti/Tv ratio per sample	2.84
Ti/Tv ratio of exonic/splicing variants per sample	3.46
Number of rare (MAF<=0.005) exonic/splicing variants	147,941
Number of rare nonsynonymous variants	82,579
Number of rare synonymous variants	48,450

Table 3.4: Statistics of discovery cohort after quality filtering

Gene collapsing analysis

In the current study, we were interested in rare-variant associations and, due to our limited sample size, for some genes the variant count per variant class was very low. Kernel-based methods such as SKAT and SKAT-O [75] tend to be anti-conservative in such cases [209, 210]. Hence, for the hypothesis-free single-gene collapsing analysis we used the two-sided Exact CMC [186] test, as implemented in the rvtests [76] package. The CMC19 method collapses and combines rare variants, followed by a Fisher's exact test. The method has the drawback that we could not adjust for covariates as we can do using the SKAT-O test. We used Bonferroni's correction for multiple testing, which tends to be conservative for correlated hypotheses, using the R function p.adjust(method="bonferroni"), R version 3.30, namely for 18,668 protein-coding genes. Quantile-quantile plots (Q-Q plots) were generated for the p-values obtained from the single-gene collapsing analysis. Top 5 genes with the lowest p-values were labelled. R version 3.30 was used to generate the qqplots (Figures 3.4, 3.5). The observed p-value distribution was overwhelmingly very close to the expected distribution under the null hypothesis of no association, indicating good control for potentially confounding factors.

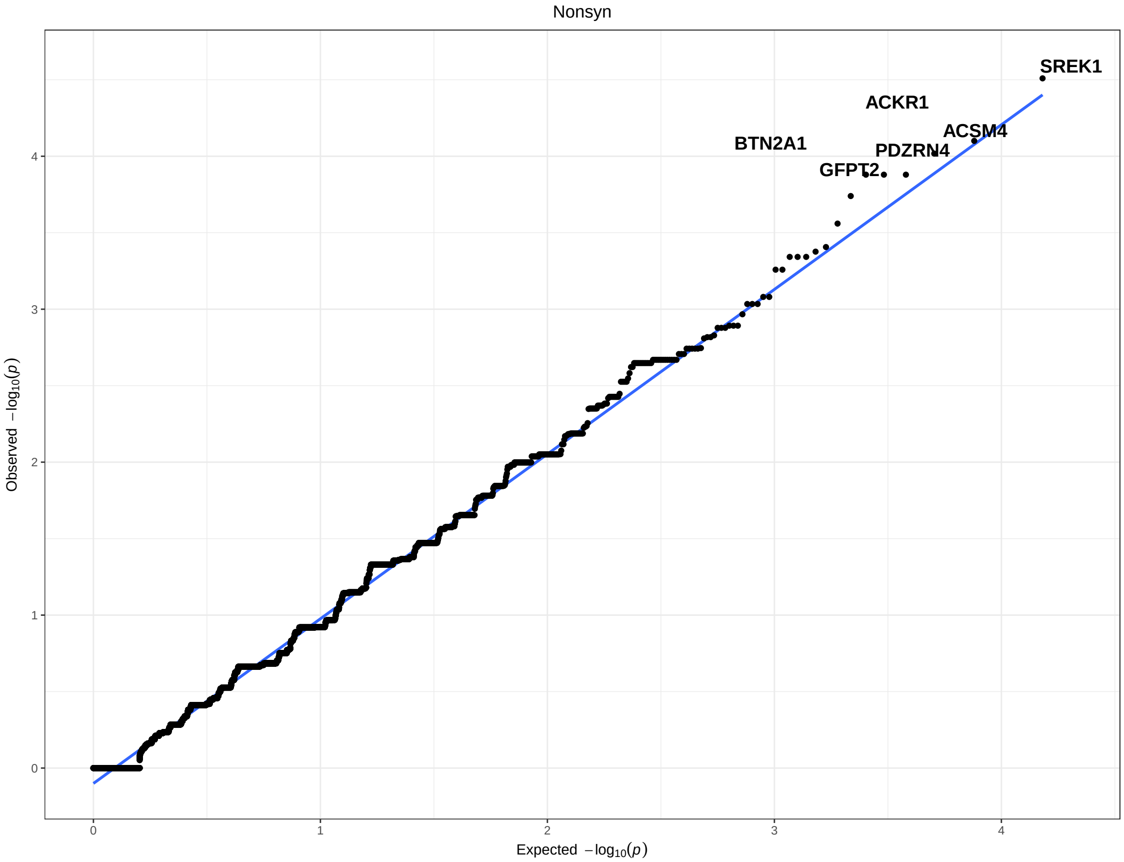


Figure 3.4: Q-Q plot of single-gene collapsing analysis for nonsynonymous variants of the discovery cohort

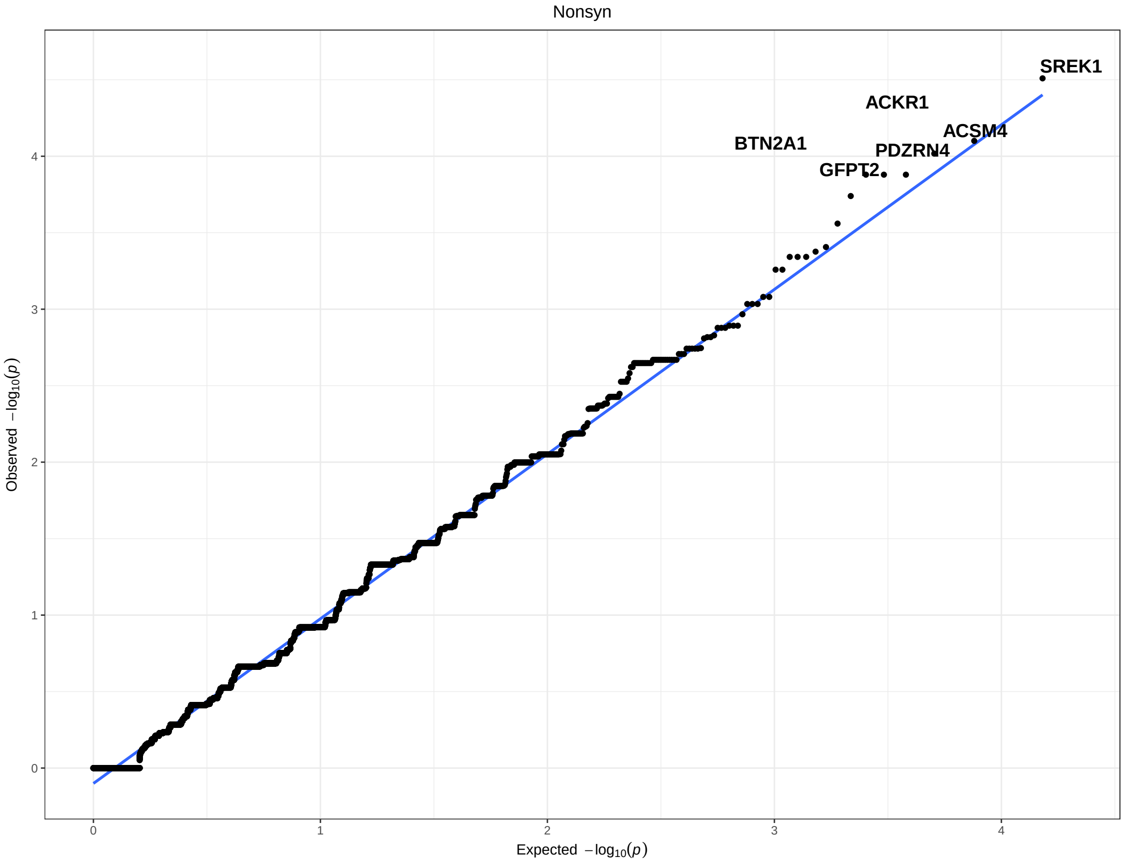


Figure 3.5: Q-Q plot of single-gene collapsing analysis for synonymous variants of the discovery cohort.

Gene-set burden analysis

Gene-set collapsing analysis was performed using the SKAT-O [75] test, as implemented in rvtests [76] while also including the first 10 principal components from the Eigenstrat [42] 3 analysis and sex as covariates. A detailed description of the seven gene lists, including the gene names and literature sources, are given in Table 3.5. We compiled panels of candidate genes for generalized (GGE) and focal epilepsies (FE), as well as epileptic encephalopathies (EE) on the basis of the published literature. For the gene lists of FE and GGE we selected genes that cause epilepsy as the main symptom without severe intellectual disability or any other predominant syndromic symptoms. If genes preferentially predispose to either FE or GGE, they were only assigned to one of the groups. For the list of EE genes, we only included genes that cause epilepsy as the predominant disease phenotype and excluded those genes that display severe developmental disorders with facultative associated seizures. The GABAergic pathway gene list was compiled on the human "GABAergic synapse" pathway defined in Kyoto Encyclopedia of Genes and Genomes [211] (KEGG) (`http://www.genome.jp/dbget-bin/www_bget?pathway+hsa04727`), and also included ion channels specifically expressed in inhibitory neurons.

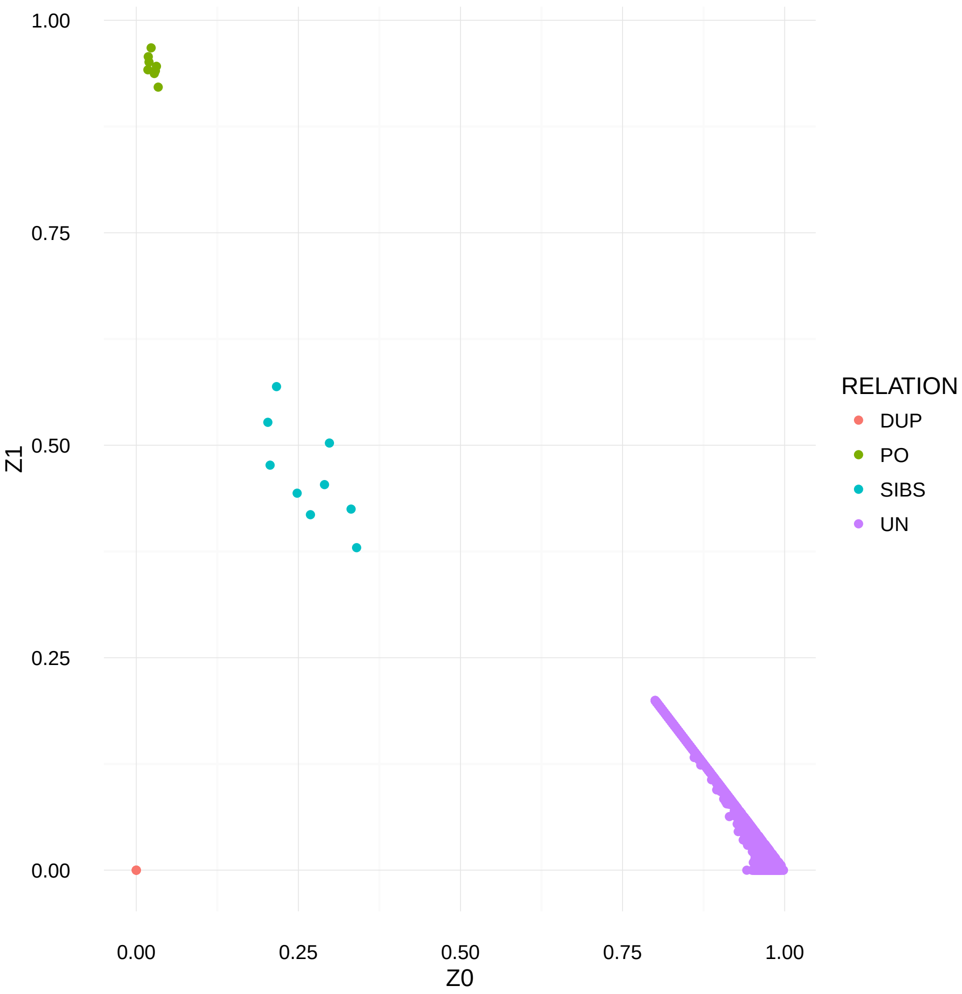


Figure 3.6: Percentage of genome shared across each pair of samples. Each dot in the plot represents a pair of samples. DUP = duplicate or monozygotic twins samples, PO = parent offspring pair, SIBS = siblings pair, UN = unrelated pair of samples.

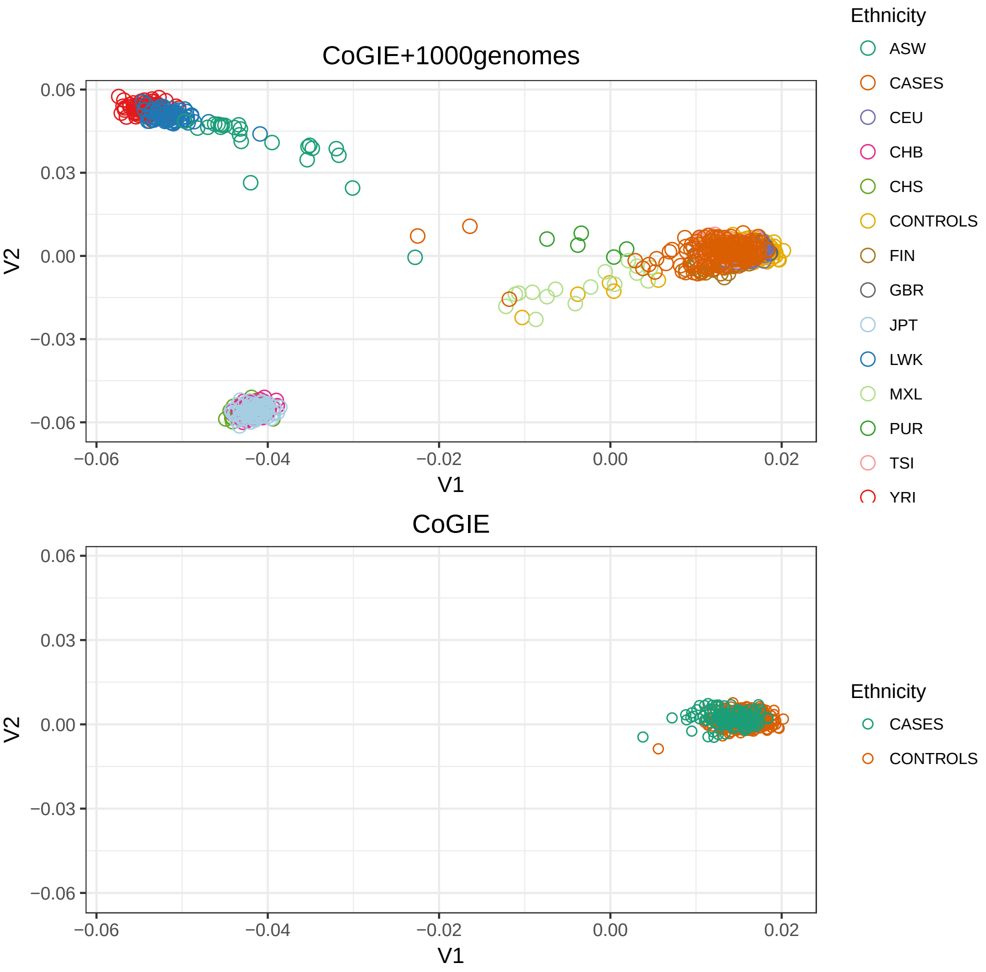


Figure 3.7: Top: Population stratification of the discovery cohort together with samples from the 1000 genomes study. Each color represents different ethnicities and each shape represents the super population to which the samples belong to. The abbreviations of the legend are given below. ASW: Americans of African Ancestry in SW USA, CEU, CHB: Han Chinese in Beijing, China, CHS: Southern Han Chinese, FIN: Finnish in Finland, GBR: British in England and Scotland, JPT: Japanese in Tokyo, Japan, LWK: Luhya in Webuye, Kenya, MXL: Mexican Ancestry from Los Angeles, PUR: Puerto Ricans from Puerto Rico, TSI: Toscani in Italia, YRI: Yoruba in Ibadan, Nigeria. AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European. Bottom) Samples included in the analyses after final QC.

Gene-set (size)	**Genes**	**Source**
Focal epilepsies (20 genes)	*CHRNA2, CHRNA4, CHRNB2, CPA6, DEPDC5, GRIN2A, KCNA1, KCNQ2, KCNQ3, KCNT1, LGI1, PRRT2, RBFOX1, RBFOX3, SCN2A, SCN8A, TBC1D24, NPRL2, NPRL3, GRIN2B*	Literature (see Suppl. Table S5 E)
Generalized epilepsies (28 genes)	*ALDH7A1,CACNA1A, CACNA1H, CACNB4, CASR, CNTN2, EFHC1, EPM2A, GABRA1, GABRB3, GABRD, GABRG2, GPHN, KCNA2, KCNC1, KCNMA1, NIPA2, NRXN1, PCDH7, PLCB1, RBFOX1, RORB, SCN1A, SCN1B, SCN9A, SLC2A1, STX1B, TBC1D24*	Literature (see Suppl. Table S5 GGE)
Epileptic encephalopathies (53 genes)	*AARS, ALDH7A1, ALG13, ARHGEF9, ARX, BOLA3, CACNA1A, CDKL5, CHD2, COQ4, DNM1, DOCK7, EEF1A2 ,GABRA1, GABRB3, GABRG2, GNAO1, GRIN1, GRIN2A, GRIN2B, HCN1, KCNA2, KCNB1, KCNC1, KCNQ2, KCNT1, MEF2C, NECAP1, NRXN1, PCDH19, PIGA, PLCB1, PNKP, ROGDI, SCN1A, SCN1B, SCN2A, SCN8A, SIK1, SIK2, SLC13A5, SLC25A22, SLC2A1, SLC35A2, SLC6A1,SPTAN1,ST3GAL3, STX1B, STXBP1, SYNGAP1, SZT2, TWNK, WWOX*	Literature (see Suppl. Table S5 EE)
GABAA receptors (19 genes)	*GABRA1, GABRA2, GABRA3, GABRA4, GABRA5, GABRA6, GABRB1, GABRB2, GABRB3, GABRD, GABRE, GABRG1, GABRG2, GABRG3, GABRP, GABRQ, GABRR1, GABRR2, GABRR3*	`http://www.genenames.org/cgi-bin/genefamilies/set/563`
GABAergic pathway (113 genes)	*ABAT, ADCY1, ADCY2, ADCY3, ADCY4, ADCY5, ADCY6, ADCY7, ADCY8, ADCY9, ANK2, ANK3, ARHGEF9, DISC1, DLC1, DLC2, DNAI1, FGF13, GABARAP, GABARAPL1, GABARAPL2, GABBR1, GABBR2, GABRA1, GABRA2, GABRA3, GABRA4, GABRA5, GABRA6, GABRB1, GABRB2, GABRB3, GABRD, GABRE, GABRG1, GABRG2, GABRG3, GABRP, GABRQ, GABRR1, GABRR2, GABRR3, GAD1, GAD2, GLS, GLS2, GLUL, GNAI1, GNAI2, GNAI3, GNAO1, GNB1, GNB2, GNB3, GNB4, GNB5, GNG10, GNG11, GNG12, GNG13, GNG2, GNG3, GNG4, GNG5, GNG7, GNG8, GNGT1, GNGT2, GPHN, HAP1, KCNB2, KCNC1, KCNC2, KCNC3, KCNJ6, KIF5A, KIF5B, KIF5C, MAGI, MKLN1, MYO5A, NLGN2, NRXN1, NSF, PFN1, PLCL1, PRKACA, PRKACB, PRKACG, PRKCA, PRKCB, PRKCG, RAFT1, RDX, SCN1A, SCN1B, SCN2B, SCN3A, SCN8A, SEMA4D, SLC12A2, SLC12A5, SLC32A1, SLC38A1, SLC38A2, SLC38A3, SLC38A5, SLC6A1,SLC6A11, SLC6A13, SRC, TRAK1, TRAK2*	Literature `http://www.genome.jp/dbget-bin/www_bget?pathway+hsa04727`

Voltage-gated ion channels (86 genes)	*SCN10A, SCN11A, SCN1A, SCN1B, SCN2A2, SCN2B, SCN3A, SCN3B, SCN4A, SCN4B, SCN5A, SCN7A, SCN8A, SCN9A, CACNA1A, CACNA1B, CACNA1C, CACNA1D, CACNA1E, CACNA1F, CACNA1G, CACNA1H, CACNA1I, CACNA1S, CACNA2D1, CACNA2D2, CACNA2D3, CACNA2D4, CACNB1, CACNB2, CACNB3, CACNB4, KCNA1, KCNA10, KCNA2, KCNA3, KCNA4, KCNA5, KCNA6,KCNA7, KCNAB1, KCNAB2, KCNAB3, KCNB1, KCNB2, KCNC1, KCNC2, KCNC3, KCNC4, KCND1, KCND2, KCND3, KCNE1, KCNE1L, KCNE2, KCNE3, KCNE4, KCNF1, KCNG1, KCNG2, KCNG3, KCNG4, KCNH1, KCNH2, KCNH3, KCNH4, KCNH5, KCNH6, KCNH7, KCNH8, KCNQ1, KCNQ2, KCNQ3, KCNQ5, KCNQ4, KCNRG, KCNS1, KCNS2, KCNS3, KCNT1, KCNV1, KCNV2, HCN1, HCN2, HCN3, HCN4*	Voltage-gated sodium channels (`http://www.genenames.org/cgi-bin/genefamilies/set/184`) Voltage-gated calcium channels (`http://www.genenames.org/cgi-bin/genefamilies/set/253`) Voltage-gated potassium channels (`http://www.genenames.org/cgi-bin/genefamilies/set/274`) Hyperpolarization-activated cyclic nucleotide–gated channels 22
Excitatory receptors (34 genes)	*CHRNA1, CHRNA10, CHRNA2, CHRNA3, CHRNA4, CHRNA5, CHRNA6, CHRNA7, CHRNA9, CHRNB1, CHRNB2, CHRNB3, CHRNB4, CHRND, CHRNE, CHRNG, GRIA1, GRIA2, GRIA3, GRIA4, GRIK1, GRIK2, GRIK3, GRIK4, GRIK5, GRIN1, GRIN2A, GRIN2B, GRIN2C, GRIN2D, GRIN3A, GRIN3B, GRID1, GRID2*	Ionotropic glutamate receptors (`http://www.guidetopharmacology.org/GRAC/FamilyDisplayForward?familyId=75`) Cholinergic receptors (`http://www.genenames.org/cgi-bin/genefamilies/set/173`)

Table 3.5: Gene-sets for gene-set burden analysis. All gene sets were tested for the neutral synonymous variants in order to exclude technical bias. P-value correction was performed by Holm's procedure, as implemented in the R function p.adjust(method="holm"), R version 3.30. Gene-set analysis p-values were adjusted for 14 tests (7 gene-sets and 2 types of variants, non-synonymous and synonymous).

3.6.3 Exome sequencing analysis (EpiPGX, replication cohort 1)

Alignment and variant calling

Reads were mapped to the human genome version GRCh37 for EpiPGX samples and version hg19 for Rotterdam controls, which were further converted to GRCh37 version. From the BAM files we generated gVCFs using the bcbio-nextgen (`https://github.com/chapmanb/bcbio-nextgen`) pipeline framework (version 0.8.9). The variant calling pipeline used picard (`http://broadinstitute.github.io/picard`)(version 1.96) to clean BAM files, marked duplicates with biobambam2 (`https://github.com/gt1/biobambam2`) (version 2.0.8), recalibrated, realigned with GATK and gVCFs were generated by using GATK HaplotypeCaller. Finally, joint calling was performed using GATK GenotypeGVCFs (version 3.5) with the bcbio-nextgen reference data (dbSNP build 138 [5]) and the options --read_filter BadCigar --read_filter NotPrimaryAlignment --standard_min_confidence_threshold_for_calling 30.0 --downsample_to_coverage 2000 --downsampling_type BY_SAMPLE.

QC (sample and variant)

The QC at sample and variant level were performed similar to the discovery cohort as described above. Majority of the samples were of European descent as it could be seen in Figure 3.8 (top) and outliers were excluded the same way as described for the discovery cohort. Only ethnically matched samples were used in the subsequent burden analysis Figure 3.8 (bottom). Various QC metrics are shown in Table 3.6. Variant annotation was performed by using ANNOVAR [208] and the variants were further divided into various classes (Nonsynonymous and Synonymous) as described above.

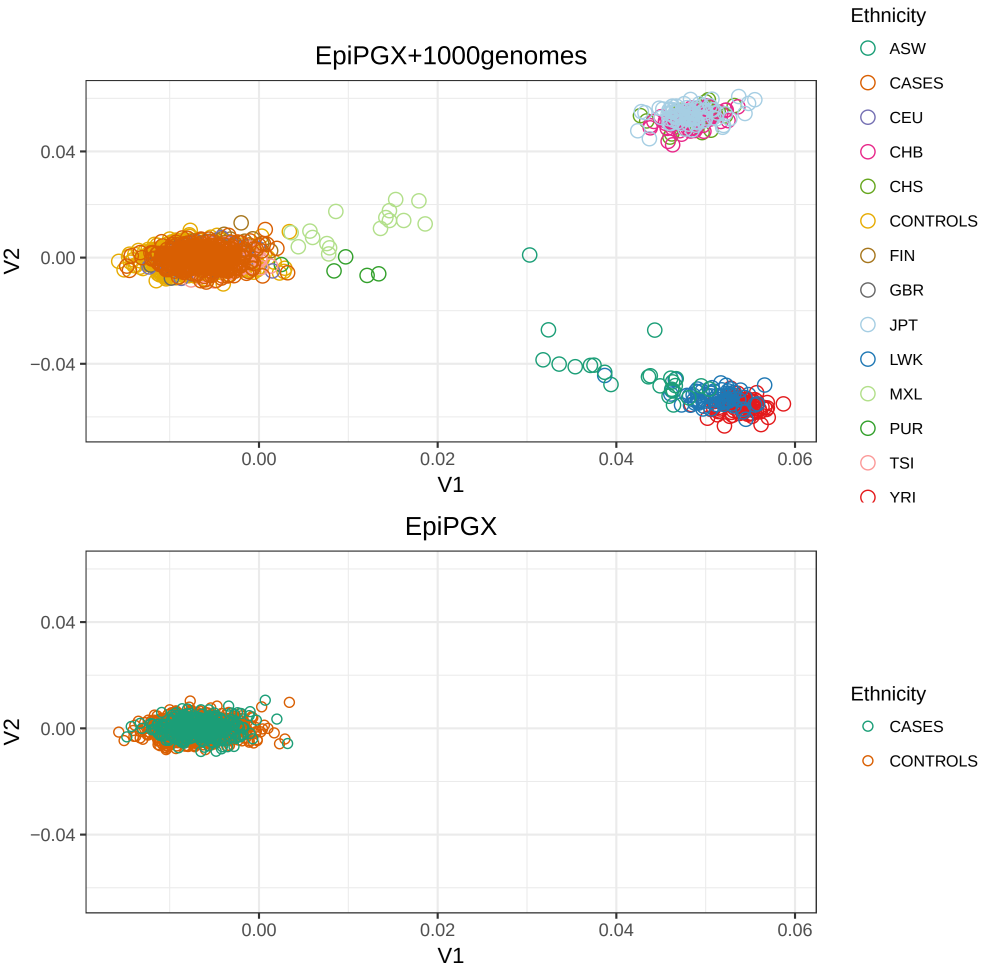


Figure 3.8: Ethnicity of EpiPGX samples with samples within 1000 genomes study. Each color represents different ethnicities and each shape represents the super population to which the samples belong to. The abbreviations of the legend are given below. ASW: Americans of African Ancestry in SW USA, CEU, CHB: Han Chinese in Beijing, China, CHS: Southern Han Chinese, FIN: Finnish in Finland, GBR: British in England and Scotland, JPT: Japanese in Tokyo, Japan, LWK: Luhya in Webuye, Kenya, MXL: Mexican Ancestry from Los Angeles, PUR: Puerto Ricans from Puerto Rico, TSI: Toscani in Italia, YRI: Yoruba in Ibadan, Nigeria. AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European. Bottom) Samples included in the analyses after final QC.

Quality metric	**After QC**
Number of samples	357 cases
	1,485 controls
Number of variants	353,173
Number of exonic/splicing variants	335,630
Ti/Tv ratio per sample	3.28
Ti/Tv ratio of exonic/splicing variants per sample	3.38
Number of rare (MAF<=0.005) exonic/splicing variants	278,184
Number of rare nonsynonymous variants	164,233
Number of rare synonymous variants	92,129

Table 3.6: Statistics of replication cohort 1 after quality filtering.

Gene-set collapsing analysis

Similar to the discovery cohort, association analysis was performed for the GABA$_A$ receptor gene set using SKAT-O, as implemented in rvtests [76], using the first ten principal components from the Eigenstrat [42] analysis as covariates. P-values for missense and synonymous were corrected using the Holm procedure.

3.6.4 GABA$_A$ receptor gene panel (replication cohort 2)

Alignment, enrichment assessment and variant calling

Using Trimmomatic [26] all reads were trimmed and clipped to remove Illumina MiSeq adapters and bad quality bases. BWA [29] mem was used to align reads to the GRCh37 human reference genome. All alignments files were stored in the BAM format and the Picard suite was used to merge all alignment in a single file (http://picard.sourceforge.net). GATK [23] suite was used to produce metrics file and to perform variant calling using UnifiedGenotyper.

QC (sample and variant) and annotation

Samples were filtered based on the PLINK/SEQ QC metrics as described above and outlier samples (>3SD) were excluded in a similar way as described in the discovery cohort. Variants were filtered by using the same parameters as discovery cohort. Finally, ANNOVAR was used to annotate the variants and they were divided into various classes as defined above.

Quality metric	After QC
Number of samples	583 cases
	635 controls
Number of variants	260
Number of exonic/splicing variants	260
Number of rare (MAF<=0.005) exonic/splicing variants	212
Number of rare nonsynonymous variants	102
Number of rare synonymous variants	95

Table 3.7: Statistics of replication cohort 2 after QC

Gene-set collapsing analysis

Similar to the discovery cohort, association analysis was performed for $GABA_A$ receptor genes using Skat-O test. Multiple-testing correction was performed using the Holm procedure.

3.6.5 Functional analysis

Mutagenesis and RNA preparation

We used the Quick Change kit (Stratagene) to engineer the missense mutations in the *GABRB2* and *GABRA5* cDNAs (NM_021911 and NM_000810, respectively) inserted in the pcDNA3 vector (kind gift from Dr. Patrick Cossette and Dr. Steven Petrou, Melbourne). Primers are available upon request. Mutations were confirmed and additional changes were excluded by Sanger sequencing. cRNA was prepared using the T7 RNA polymerase kit from Ambion.

Oocyte preparation and injection

Oocytes were obtained from the Institute of Physiology I, Tübingen, or purchased from EcoCyte Bioscience (Castrop-Rauxel). Experiments were approved by local authorities (Regierungspräsidium Tübingen, Germany). The preparation of oocytes for two-microelectrode voltage-clamp recordings included treatment with collagenase (1mg/ml of type CLS II collagenase, Biochrom KG) in OR-2 solution (mM: 82.5 NaCl, 2.5 KCl, 1 $MgCl_2$ and 5 Hepes, pH 7.6), followed by thorough washing and storing at 17°C in Barth solution (mM: 88 NaCl, 2.4 $NaHCO_3$, 1 KCl, 0.41 $CaCl_2$, 0.82 $MgSO_4$ and 5 Tris/HCl, pH 7.4 with NaOH) supplemented with 50μg/ml gentamicin (Biochrom KG) as described previously [24,25]. We injected a total amount of 70 nl at a concentration of 2 μg/μl of cRNA encoding respective mixtures of WT or mutant subunits into oocytes using the Robooinject® (Multi Channel Systems). Oocytes were stored for 1-3 days at 17°C before the experiment. The combination of the different subunits used was α_1 β_2 γ_{2s} or α_5 β_2 γ_{2s} in 1:1:2 ratios. Current amplitudes of WT and mutant receptors were recorded and compared on the same day using the same batch of oocytes so that data of different days could be pooled when normalized to the WT.

Automated oocyte two-microelectrode voltage clamp

GABA-evoked ionic currents in oocytes were recorded at room temperature (20-22°C) using a Roboocyte2® system (Multi Channel Systems). Prepulled and prepositioned intracellular glass microelectrodes had a resistance of 0.3–1 MΩ when filled with 1 M KCl and 1.5 M KAc. ND96 was used as extracellular bath solution (in mM: 93.5 NaCl, 2 KCl, 1.8 $CaCl_2$, 2 $MgCl_2$, 5 Hepes, pH 7.5). Currents were sampled at 1 kHz. To activate the receptors, increasing GABA concentrations (in μM: 1, 3, 10, 40, 100, 300, 1000) diluted in ND96 solution were applied for 15 s each.

Electrophysiological data analysis

Oocytes were held at -70 mV. The amplitude of the GABA-induced currents was analyzed using Roboocyte2+ (Multi Channel Systems), Clampfit (pClamp 8.2, Axon Instruments), Microsoft Excel (Microsoft) and GraphPad Prism (GraphPad Software). The current response of each GABA concentration was normalized to the maximum response evoked by the highest GABA concentration (1 mM). The following four parameter logistic equation:

$$Y\ (X) = \min + \frac{(max - min)}{1 + 10^{((LogEC50 - X)*nh}}$$

was fit to the normalized GABA responses of each oocyte, with *max* and *min* being the maximum and minimum evoked responses, *X* the corresponding GABA concentration, EC_{50} the concentration of the agonist at which half of the maximum response is achieved, and *nh* the Hill coefficient reflecting the steepness of the dose-response curve. For each oocyte, EC_{50} values were determined and then averaged for each combination of receptor subunits used. Current amplitudes in response to 1 mM GABA application for mutant channels were normalized to the mean value of the WT channel response recorded on the same day.

Statistical analysis

For statistical evaluation, GraphPad Prism 6 was used. Normalized current amplitudes were compared between different groups (WT and different variants) using one-way ANOVA on ranks (Kruskal Wallis rank sum test) with Dunn´s *post-hoc* test. All data are as mean ± standard error of the mean (SEM). Statistical differences are indicated in the figure legends using the following symbols: *$p<0.05$, **$p<0.001$, ***$p<0.0001$.

3.6.6 Annotation of functional tested $GABA_A$ receptor variants

As described above, the functionally tested $GABA_A$ variants were annotated with ANNOVAR [47] using the RefSeq gene annotations, allele frequencies from ExAC [43], 1000g

[34] and ESP, dbSNP (`https://www.ncbi.nlm.nih.gov/projects/SNP`), and pathogenicity and conservation scores from dbNSFP [51] (Supplemental Table S7). The following missense pathogenicity prediction scores and thresholds for pathogenicity were used: SIFT [53] (D deleterious), PolyPhen2_HDIV [54](D damaging), LRT [55] (D deleterious), MutationTaster [56] (A disease causing automatic, D disease causing), PROVEAN [57] (D deleterious), CAD [58] phred score $>$ 10, fathmm [60](D deleterious). For conservation, we used GERP++_R [61] (>3) and SiPHy [62] (>10) for conservation evaluation. Details on different scores and their prediction classes, ranges and thresholds can be found on the ANNOVAR webpage: `http://annovar.openbioinformatics.org/en/latest/user-guide/filter`.

Geneset	Variant class	Number of variants	Number of cases	Number of controls	Skat-O P-value uncorrected	Skat-O P-value (corrected by Holm)	OR	lowerCI	upperCI
Epileptic encephalopathies	SYN	284	152	549	0.292299	1	1.25	0.87	1.80
Excitatory receptors	SYN	151	152	549	0.212472	1	1.43	0.97	2.11
Focal epilepsies	SYN	110	152	549	0.416506	1	1.066	0.68	1.66
GABAergic pathway	SYN	444	152	549	0.0628653	0.8172489	1.33	0.92	1.92
GABA-A receptors	SYN	64	152	549	0.574707	1	1.37	0.78	2.40
Generalized epilepsies	SYN	177	152	549	0.237372	1	0.99	0.67	1.47
Voltage-gated ion channels	SYN	471	152	549	0.771146	1	1.19	0.81	1.74
Epileptic encephalopathies	NONSYN	259	152	549	0.528544	1	1.42	0.98	2.07
Excitatory receptors	NONSYN	241	152	549	0.708729	1	1.01	0.69	1.48
Focal epilepsies	NONSYN	142	152	549	0.523553	1	1.42	0.93	2.17
GABAergic pathway	NONSYN	564	152	549	0.442513	1	1.66	1.12	2.46
GABA-A receptors	NONSYN	63	152	549	0.0013633	0.0190862	2.40	1.41	4.10
Generalized epilepsies	NONSYN	194	152	549	0.166314	1	2.17	1.49	3.17
Voltage-gated ion channels	NONSYN	664	152	549	0.601852	1	1.78	1.17	2.70

EpiPGX

Geneset	Variant class	Number of variants	Number of cases	Number of controls	Skat-O P-value uncorrected	Skat-O P-value (corrected by Holm)	OR	lowerCI	upperCI
GABA-A receptors	SYN	99	357	1485	0.587127	0.58712700	0.82	0.54	1.25
GABA-A receptors	NONSYN	107	357	1485	0.00805992	0.01611984	1.46	1.05	2.03

GABA_panel

Geneset	Variant class	Number of variants	Number of cases	Number of controls	Skat-O P-value uncorrected	Skat-O P-value (corrected by Holm)	OR	lowerCI	upperCI
GABA-A receptors	SYN	95	583	635	0.0613778	0.0613778	0.86	0.60	1.24
GABA-A receptors	NONSYN	103	583	635	0.0133277	0.0266554	1.458	1.019	2.08 height

`https://dropit.uni.lu/invitations?share=0c2fe26106c4920e2630&dl=0`

Table 3.8: For all cohorts, the rare, case-only missense variants are listed together with their RefSeq annotations and ExAC (NFE=Non-Finnish European) allele frequencies. An overview table gives information for which GABA_A gene variants were found in which cohort and if variants from this gene were functional tested.

`https://dropit.uni.lu/invitations?share=bc14f52de9820cfd80bf&dl=0`

Table 3.9: Annotations for the tested GABA_A variants for which sample material was available. Genome position in hg19, allele counts (AC) and frequencies (AF) information in ExAC (ALL=all populations, NFE=Non-Finnish European population), allele frequencies in 1000g and ESPdbSNP identifier, type of variant (snp/insertion) and exonic type (frameshift insertion/missense), and for missense mutations the prediction scores from SIFT, Polyphen2, LRT, MutationTaster, Provean, CADD and fathmm as well as the conservation scores from GERP and SiPhy are given. Deleterious predictions are given as D, conserved sites are shown in bold text. Additionally, the number of prediction tools with a deleterious prediction per variant is given as well as the number of conservation scores showing conservation per variant.

CHAPTER 4

CNVS IN EPILEPSY

4.1 Abstract

Genetic Generalized Epilepsy (GGE) and benign epilepsy with centro-temporal spikes or Rolandic Epilepsy (RE) are common forms of genetic epilepsies. Rare copy number variants have been recognized as important risk factors in brain disorders. We performed a systematic survey of rare deletions affecting protein-coding genes derived from exome data of patients with common forms of genetic epilepsies. We analysed exomes from 390 European patients (196 GGE and 194 RE) and 572 population controls to identify low-frequency genic deletions. We found that 75 (32 GGE and 43 RE) patients out of 390, i.e. ~19%, carried rare genic deletions. In particular, large deletions (>400 kb) represent a higher burden in both GGE and RE syndromes as compared to controls. The detected low-frequency deletions (1) share genes with brain-expressed exons that are under negative selection, (2) overlap with known autism and epilepsy-associated candidate genes, (3) are enriched for CNV intolerant genes recorded by the Exome Aggregation Consortium (ExAC) and (4) coincide with likely disruptive *de novo* mutations from the NPdenovo database. Employing several knowledge databases, we discuss the most prominent epilepsy candidate genes and their protein-protein networks for GGE and RE.

4.2 Introduction

Epilepsies are among the most widespread neurological disorders with a lifetime incidence of ~3% [212]. They represent a heterogeneous group of different disease entities that, with regard to aetiology, can be roughly divided in epilepsies with an exogenous/symptomatic cause and those with a genetic cause. Genetic generalized epilepsies (GGE; formerly idiopathic generalized epilepsies) are the most common genetic epilepsies accounting for 30% of all epilepsies. They comprise syndromes such as juvenile myoclonic epilepsy, childhood absence epilepsy and juvenile absence epilepsy. In general, they tend to take a benign course and show a good response to pharmacotherapy. Among focal genetic epilepsies, benign epilepsy with centro-temporal spikes or Rolandic epilepsy (RE) is the most common form. RE has its onset in childhood or early adolescence and usually tapers off around the age of 15.

High-throughput genomic studies raised the number of epilepsy-associated candidate genes to hundreds; nowadays, frequently mutated ones are included in diagnostic gene panels (for recent reviews see [14, 213]. Large consortia initiatives such as Epi4k [190] enrolled 1,500 families, in which two or more affected members displayed epilepsy, as well as 750 individuals, including 264 trios, with epileptic encephalopathies and infantile spasms, Lennox-Gastaut syndrome, polymicrogyria or periventricular heterotopias. In addition to the detection of known and unknown risk factors, the consortium found a significant overlap between the gene network of their epilepsy candidate genes and the gene networks for autism spectrum disorder (ASD) and intellectual disability. Intriguingly, epilepsy is the medical condition most highly associated with genetic autism syndromes [214].

Genomic disorders associated with copy number variations (CNVs) appear to be highly penetrant, occur on different haplotype backgrounds in multiple unrelated individuals and seem to be under strong negative selection [215–217]. A number of chromosomal locations suspected to contribute to epilepsy have been identified [218–222].

A genome-wide screen for CNVs using array comparative genomic hybridization (aCGH) in patients with neurological abnormalities and epilepsy led to the identification of recurrent microdeletions on 6q22 and 1q22.31 [223]. A deletion on 15q13.3 belongs to the most frequent recurrent microdeletions in epilepsy patients; it is associated with intellectual disability, autism, schizophrenia, and epilepsy [224, 225]. The recurrence of some CNVs seems to be triggered by the genome structure, namely by the chromosomal distribution of interspersed repetitive sequences (like Alu transposons) or recently duplicated genome segments (large blocks of sequences >10with >95% sequence identity that constitute five to six percent of the genome) that give rise to nonallelic homologous recombination [215, 226].

CNV screening in large samples showed that 34% of heterozygous deletions affect genes associated with recessive diseases [227]. CNVs are thought to account for a major proportion of human genetic variation and have an important role in genetic susceptibility to common disease, in particular neuropsychiatric disorders [228]. Genome-wide surveys have demonstrated that rare CNVs altering genes in neuro-developmental pathways are implicated in epilepsy, autism spectrum disorder and schizophrenia [14, 15].

Considering all types of CNVs across two analysed cohorts, the total burden was not significantly different between subjects with epilepsy and subjects without neurological disease [229]; however, when considering only genomic deletions affecting at least one gene, the burden was significantly higher in patients. Likewise, using Affymetrix SNP 6.0 array data, it has recently been shown that there is an increased burden of rare large deletions in GGE [221]. The drawback of the latter approach is that smaller CNVs cannot be detected. Systematic searches of CNVs in epilepsy cohorts using whole-exome sequencing (WES) data, which provides the advantage to identify smaller deletions along with the larger ones, are still missing.

In the present study, we provide the CNV results of the largest WES epilepsy cohort reported so far. We aimed at (1) identifying the genome-wide burden of large deletions (>400kb), (2) studying the enrichment for deletions of brain-expressed exons, in particular those under negative selection, (3) detecting deletions that overlap with previously defined autism and epilepsy candidate genes, and (4) browsing knowledge databases to help understand the disease aetiology.

4.3 Patients and Methods

The study protocol was approved by the local institutional review boards of the contributing clinical centres. Written informed consent was obtained from participating subjects and, if appropriate, from both patients and adolescents.

4.3.1 Data

GGE cohort: This cohort included 196 subjects with genetic generalized epilepsy. All subjects were of European descent (Italian 81, German 54, Finnish 22, Dutch 11, British 9, Danish 8, Turkish 6, Swedish 3, French 1, Greek 1). The cohort included 117 female subjects (60%). The GGE-diagnoses were childhood absence epilepsy (CAE; n=94), juvenile absence epilepsy (JAE; 21), juvenile myoclonic epilepsy (JME; 47), genetic generalized epilepsy with generalized tonic-clonic seizures (EGTCS, 27), early-onset absence epilepsy (EOAE, 4), epilepsy with myoclonic absences (EMA, 1), and unclassified GGE (2). Age of epilepsy onset ranged from 1 year to 38 years with a median of 8 years. The majority of subjects derived from multiplex families with at least 2 affected family members (n=183), thereof 90 families with 3 or more affected members.

RE cohort: This cohort included 204 unrelated Rolandic patients of European ancestry which were recruited from centers in Austria (n = 107), Germany (n = 84), and Canada (n = 13).

Control cohort: We used 445 females and 283 males (728 in total) from the Rotterdam Study as population control subjects [230]. The same cohort was recently used for the screening of 18 GABA_A-receptor genes in RE and related syndromes [149].

Our primary analysis workflow included three major steps as shown in Figure 4.1. These are

1) data pre-processing, 2) SNV/INDEL analysis and 3) copy number variant analysis.

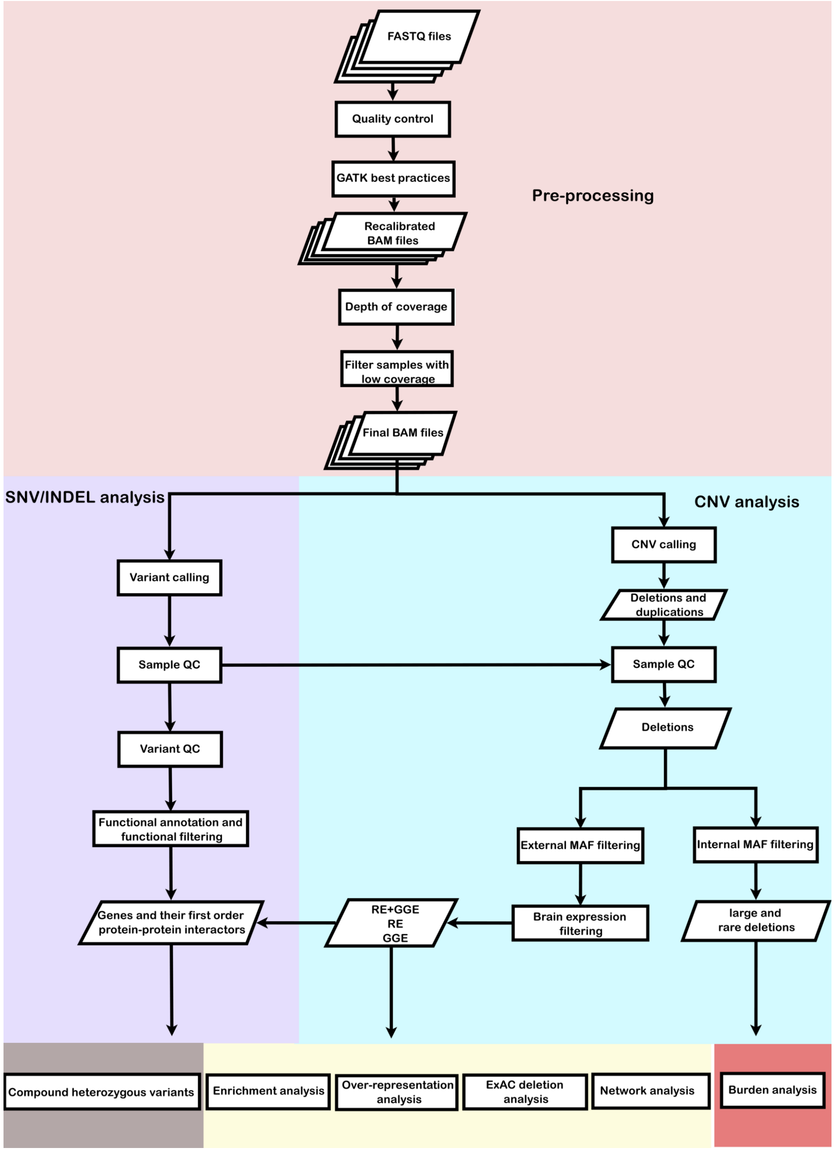


Figure 4.1: Flowchart of the analysis steps. Parameters used in each step are described in detail in the methods section.

4.3.2 Data pre-processing

Sequencing adapters were removed from the FASTQ files with cutadapt [25] and sickle [231]. GATK best practices were followed for the next steps of data pre-processing and variant calling [23]. Alignment to the GRCh37 human reference genome was performed using BWA-MEM [29] with default parameters. Conversion of SAM to BAM files was done with SAMtools [27]. Sorting of BAM files, marking of duplicate reads due to PCR amplification and addition of read group information were done using Picard (https://github.com/broadinstitute/picard) tools with default parameters. Base quality score recalibration and local realignment for INDELs was performed using GATK version 3.2.

4.3.3 Coverage

Mean depth of coverage and target coverage of exons were calculated from the BAM files using the depth of coverage tool from GATK. The same files were also used as input for calling of CNVs.

4.3.4 Variant calling

The GATK haplotype caller (version 3.2) was chosen to perform multiple sample variant calling and genotyping with default parameters. To include splice site variants in the flanking regions of the exons, exonic intervals were extended by 100each upstream and downstream. Multiple sample calling is advantageous in deciding whether a variant can be identified confidently as it provides the genotype for every sample. It allows filtering variants based on the rate of missing genotypes across all samples and also according to the individual genotype.

4.3.5 Sample QC

Samples were excluded from the analysis based on the following criteria: 1) Samples with a mean depth <30x or <70% of exon targets covered at <20x were excluded from further analysis; 2) samples with >3 standard deviations from mean in number of alternate alleles, number of heterozygotes, transition/transversion ratio, number of singletons and call rate as calculated with the PLINK/SEQ i-stats tool (`https://atgu.mgh.harvard.edu/plinkseq/`); 3) call rate <97%; 4) ethnically unmatched samples as identified by multi-dimensional scaling analysis with PLINK version 1.9 [39]; 5) PI_HAT score>0.25 as calculated by PLINK version 1.9 to exclude related individuals.

4.3.6 Variant QC

Initial filtering of variants was performed based on quality metrics over all the samples with the following parameters for VQSR: Tranches chosen, VQSRTrancheSNV99.90to100.00. QC over all samples (INFO column) was done as follows: a) for SNVs, variants were filtered for QD < 2.0, FS > 60.0, MQ< 40.0, MQRankSum <=12.5, ReadPosRankSum <=8.0, DP <10.0, GQ_MEAN < 20.0, VQSLOD < 0, more than 5% missingness, ABHet > 0.75 or < 0.25 and deviation from Hardy-Weinberg equilibrium (Phred scale p-value of > 20); b) for INDELs, the same was done as for SNVs except for the following parameters for variant filtration: QD <2.0, FS >200.0, ReadPosRankSum <=20.0, DP <10.0, GQ_MEAN <20.0, missingness <5%, Hardy-Weinberg Phred scale value of >20, VQSLOD >0.

To further exclude low quality variants, we also applied filtering based on quality metrics for each genotype using read depth and quality of individual genotypes. Genotypes with a read depth of <10 and GQ of <20 were converted to missing by using BCFtools [27]. Multi-allelic variants were decomposed using variant-tests [232] and left-normalized using BCFtools.

4.3.7 Variant annotation

Variants were annotated with ANNOVAR [47] version 2015, Mar22 using RefSeq and Ensembl versions 20150322 and the dbNSFP [51] version 2.6 annotations including nine scores for missense mutations (SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, MetaSVM, MetaLR), the CADD score, and three conservation-based scores from GERP++, PhyloP and SiPhy. Splicing variants were defined to include 2 bp before and after the exon boundary position. To obtain rare variants, we filtered the variants for a minor allele frequency (MAF) of <0.005 in public databases such as 1000 genomes [34], dbSNP [52], ExAC (release 0.3) and the exome variant server (EVS). We defined deleterious variants as those variants that fulfil any of the following three criteria: 1) all the variants except the synonymous variants predicted to be deleterious by at least 5 out of 8 missense prediction scores, CADD score >4.5, or 2 out of 3 conservation scores (GERP>3, PhyloP>0.95, SiPHy>10) show high conservation; 2) variants annotated as “splicing”, “stop gain” or “stop loss”; 3) any insertion or deletion.

4.3.8 CNV detection

In the remaining high quality samples, CNVs were detected by using XHMM as described in [69]. In the current study, we focused only on deletions, as the false positive rate for duplications is too high to allow for meaningful interpretation. CNV calls were annotated using bedtools version 2.5 [233]. NCBI RefSeq (hg19, 20150322) was used to identify the genes that lie within the deletion boundaries.

4.3.9 CNV filtering

The detected deletions were filtered based on the following criteria: 1) Z score <∼–3, given by XHMM; 2) Q_SOME score   60, given by XHMM.

4.3.10 Burden analysis of large and rare deletions

Excess deletion rate of the large deletions (length >400 kb) in subjects with epilepsy compared to the controls was measured as described in [221] using PLINK version 1.9 [39]. We set the overlap fraction to 0.7 (70%) and the internal allele frequency cut-off <0.5% and evaluated the significance empirically by 10,000 case-control label permutations.

4.3.11 Case-only CNVs

The CNVs that are unique for cases (not present in any of the in-house controls) and occur at a low frequency, i.e., present in 2 independent cases, while having a frequency of 1% in the CNVmap, the DGV gold standard dataset [71] and 1000 genomes SV [72] were selected and subjected to further analysis as described below.

4.3.12 Validation of CNVs

We proceeded by visual inspection of depth variation across exons of the filtered deletions; we also performed qPCR validations of three small deletions, two of which, *NCAPD2* and *CAPN1*, stood the filtering procedure (see Table A.3). For RE patients, genomic DNA samples were analysed using the Illumina OmniExpress Beadchip (Illumina, San Diego, CA, USA) [221]. Twenty-three of 60 CNVs present in the RE patients were validated by available array data (Table A.5). Generally, small CNVs cannot be reliably identified with SNP arrays [234]. Indeed, of the 37 CNVs that were not identified in the beadchip data, 23 have a size of <10, whereas only 2 of the 23 validated CNVs have a size of less than 10according to the array data.

4.3.13 Compound heterozygous mutations and protein-protein interactions

We checked for concurrence of a deletion in one allele and a deleterious variant in the second allele. We included the first order interacting partners from the protein-protein interaction network (PPIN) in this analysis [235] and assessed if any gene or its first order interacting partner carries a deletion in one allele and a deleterious variant in the other. We excluded all genes that had no HGNC (HUGO Gene Nomenclature Committee) entry resulting in a network of 13,364 genes and 140,902 interactions. This network was then further filtered for interactions likely to occur in brain tissues using a curated data set of brain-expressed genes [236]. The final brain-specific PPIN consisted of 10,469 genes and 114,533 interactions.

4.3.14 Gene-set enrichment analysis

Genes that were expressed in brain [236]. and located within deletion boundaries were used as input for an enrichment analysis using the Ingenuity Pathway Analyser (IPA®) [237]. We performed the enrichment analysis with all deleted genes from the RE and GGE samples together as well as for each phenotype separately.

4.3.15 Over-representation analysis

To assess whether the deleted set of genes were enriched in known epilepsy-associated genes, we retrieved genes that were associated with the disease term "epilepsy" from the DisGeNET database [238]. Then we compared the overlap between the brain-expressed genes that are deleted in RE (n=85), GGE (n=49) and RE+GGE (n=134) against the brain-expressed epilepsy-related genes in DisGeNet (n=674). We used the total number of brain-expressed genes (n=14,177) as the background. The R GeneOverlap package (`https://bioconductor.org/packages/release/bioc/html/GeneOverlap.html`) was used to compute the p-value.

4.3.16 CNV tolerance score analysis

The CNV tolerance score was used as defined in [239]. The CNV tolerance and deletion scores for the genes that are deleted in our study were obtained from the ExAC database [43] and their enrichment in GGE and RE cases was assessed by the Wilcoxon rank sum test.

4.3.17 Overlap with different databases

The overlap between the different data sets was obtained by gene symbol matches between the detected gene deletions and the gene lists from different databases; more details are given in the discussion section. A workflow depicting the steps above is shown in Figure 4.1.

4.4 Results

After quality control, exomes of 390 epilepsy cases (196 GGE, 194 RE) and 572 controls were used for downstream analyses (Figure 4.1). The final RE and GGE datasets comprised 26,476 and 30,207 variants, respectively.

4.4.1 Epilepsy-associated microdeletions

75 out of 390 epilepsy patients (~19%) carried a total of 104 case-only deletions spanning 260 genes (see Table A.3), which covered a wide size range between 915 bp and 3.11 Mbp. 43 out of 194 RE patients carried deletions compared to 32 out of 196 patients with GGE, thus, we did not observe any significant difference in the total number of deletions between the two disease entities (p-value = 0.68). In the combined dataset, 35 out of 73 were large multigene deletions. Among them were several recurrent deletions (see Table A.3), including those located on 15q13.3 and 16p11.2 that were previously reported to be associated with epilepsy and other brain disorders.

4.4.2 Comparative analysis of Rolandic and GGE candidate genes

Because our cohort is composed of GGE and RE patients, we sought to compare the functional differences between the two subtypes of epilepsies by studying the pathways and functions that are enriched in the respective deleted genes (see Table 4.4). Initially we performed GO term enrichment without applying any additional filter to the deletion calls. As shown in Table 4.3, synaptic and receptor functions are more prominent in RE cases. If the deletion calls were filtered for brain-specific gene expression, we observed that, separately and together, GGE and RE-deleted genes are enriched for the functional terms “nervous system development and function”, "behavior" and "tissue morphology"; this functional convergence might have been expected when selecting for brain-expressed genes.

When analyzing GGE and RE datasets separately, the top PPIN enriched in GGE is associated with "carbohydrate metabolism", "small molecule biochemistry" and "cell signaling", whereas the top networks associated with RE are "neurological disease", "organismal injury and abnormalities" and "psychological disorders" (see Table 4.4). The enriched network including GGE and RE-deleted genes (Figure 4.2) is described below.

4.4.3 Deletion burden analysis

We performed 10,000 case-control label permutations to test whether there is an increased burden of large and rare deletions in cases as compared to the controls (Table 4.1). We noticed that (1) the deletion rate per individual with at least one deletion in cases compared to the controls showed statistical significance in both GGE and RE (p-value = 1e-04, p-value = 0.011) and (2), considering cumulative length of all large and small deletions, no significant difference between cases and controls was observed in both GGE and RE (p-value = 0.16, p-value = 0.41), indicating that there is no difference in the length of CNVs in cases and controls.

Dataset	Deletion rate per person	Proportion of samples with at least one deletion	Total length of deletions	Average length of deletions
IGE+RE	1,0E-04	1,0E-04	2,7E-01	2,8E-01
IGE	1,0E-04	1,0E-04	1,7E-01	1,8E-01
RE	1,1E-02	3,0E-03	4,1E-01	2,3E-01

Table 4.1: Burden test showing empirical p values of cases/controls permutation statistics. RE = Rolandic epilepsy (Typical/Atypical), IGE = Idiopathic generalized epilepsy.

4.4.4 Enrichment for known epilepsy and autism-associated genes

To check the overlap between the deletions detected in our study and genes known to be associated with epilepsy, we searched for overlap with the genes listed (n=499) in the Epilepsy-Genes database [240]. This led to the following set of 8 genes: *CHRFAM7A, CHRNA7, SCN1A, CNTNAP2, GABRB3, GRIN2A, IGSF8, ITPR1. The GRIN2A* deletion is from the same patient published earlier [148] and which we used as one of the positive controls in our primary CNV detection pipeline [241]. One should notice that genes such as *CHRNA7* and *GABRB3* are located within larger deletions containing other genes; so they might be questionable as *bonafide* epilepsy-associated genes.

Using the core autism candidate genes (n=455 genes) present in *brainspan*, [242], we identified 13 deleted genes: *APBA2, ATP10A, CDH22, CDH8, GABRA5, GABRG3, NDN, NDNL2, CNTNAP2, GABRB3, GRIN2A, SCN1A and SHANK1* (Table 4.2). This set is particularly enriched in GO terms “neuron parts” and “transporter complexes”. Note that *GABRB3* and *GABRG3* belong to multigenic large deletions (Table A.3).

PSD genes	BCG genes	Autism brainSpan	EpilepsyDB	clinVar
NDUFS3	APBA2	APBA2	CHRFAM7A	SACS
RIMBP2	ATRNL1	ATP10A	CHRNA7	CNTNAP2
TJP1	CDH22	CDH22	SCN1A	GABRB3
CNTN1	CSMD1	CDH8	CNTNAP2	GRIN2A
CNTNAP2	ETV1	GABRA5	GABRB3	ITPR1
GABRB3	FAN1	GABRG3	GRIN2A	SCN1A
GRIN2A	GMFB	NDN	IGSF8	
HSPA1L	IGSF8	NDNL2	ITPR1	
IGSF8	NPR2	CNTNAP2		
PTPRZ1	OTUD7A	GABRB3		
SHANK1	PLXDC2	GRIN2A		
	SCN1A	SCN1A		
	ZFAND1	SHANK1		
	ZNF343			
	ZNF568			
	CNTN1			
	CNTNAP2			
	GABRB3			
	GRIN2A			
	ITPR1			
	PTPRZ1			
	SHANK1			

Table 4.2: Overlap with specific sets. In grey are genes common to at least 2 of the compared sets. PSD (post-synatric density); BCG (Brain Critical Genes).

4.4.5 Deletions of brain-critical exons

Reduced fecundity associated with disorders such as autism, schizophrenia, mental retardation and epilepsy puts negative selection pressure on risk alleles. A recent report [216] combined exome and transcriptome data from large human population samples and defined a class of brain-expressed exons that are under purifying selection, namely those that are highly expressed in brain tissues and at the same time exhibiting suppressed accumulation of missense mutations in population controls (low mutation burden). These exons were called “brain-critical exons” (n=3,955), the associated genes were accordingly called “brain-critical genes” (BCG, n=1,863) [14].

Twenty-two deleted genes are in common with the BCG set (see Table 4.2). The *SHANK1* deletion is found in a single RE case. It spans 7,339 bp (8 exons out of 9). There is only one report on the possible implication of the deletion of this gene in childhood epilepsy [243]. A deletion of *ITPR1* is observed in another RE case; this deletion affects also *SUMF1*, but this gene was filtered out by the BCG overlap selection. The deletion of *CNTN1* in a GGE patient encompasses in addition *MUC19* and *LRRK2*, the latter is a known Parkinson candidate gene [244].

4.4.6 Exome Aggregation Consortium deletions

The ExAC data comprise 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. Deletions annotated in ExAC (release 0.3.1 of 23/08/16) were identified, similar to the present study, by read depth analysis using XHMM [239]. We sought to compare those CNV calls with the ones detected in the present work. Out of the 260 deleted genes detected in our study, 164 genes (67%) showed deletions in ExAC too (see Table A.4). Several genes highlighted in the previous paragraphs were ranked high using the CNV tolerance score defined by [239]. However, we did not identify a significant difference, neither in CNV tolerance scores (p-value = 0.53) nor in CNV deletion scores (p-value = 0.22), between GGE and RE-deleted genes. This may indicate that GGE and RE deletions are equally likely to fall into the same category of ExAC deletion calls.

4.4.7 Compound heterozygous and first order protein-protein interaction mutations

Compound heterozygous mutations play a role in many disease aetiologies such as autism and Parkinson's disease [245–247]. We searched for possibly deleterious non-synonymous changes in the parental undeleted gene copy, but we did not detect any hemizygous variant that had a critical intolerance score (see Methods). Subsequently, we hypothesised that simultaneous mutations in proteins which interact directly (first-order protein interactors) may increase the associated deleterious effect. Within a curated brain-specific PPIN (see Methods, [235]), we inspected first order interacting proteins with potentially deleterious mutations or exon losses (see Table 4.4.7) and found a few interesting hits, including *SPTAN1* that interacts directly with *SHANK1*; *SPTAN1* encodes alpha-II spectrin and is known to be associated with epilepsy [248, 249]. A remarkable and unique case of multiple hits was observed in a patient who accumulated four hits: the originally detected *ITPR1* deletion and three potentially deleterious non-synonymous SNVs in *RYR2*, *HOMER2* and *STARD13*. *RYR2* (ryanodine receptor 2) and *ITPR1* (inositol-1,4,5-trisphosphate receptor 1) have been independently reported to be implicated in brain disorders. *RYR2 de novo* mutations have been identified in patients with intellectual disability [250] and activation of *ITPR1* and *RYR2* can lead to the release of Ca^{2+} from intracellular stores affecting propagating Ca^{2+} waves [251]. *HOMER2*, a brain-expressed gene, has been reported to be involved in signalling defects in neuropsychiatric disorders [252]. The *STARD13* locus has been reported to be associated with aneurysm and sporadic brain arteriovenous malformations [253, 254].

Sample	Gene (deleterious SNV/Indel)	Gene in deletion boundaries	Case	Chr	Position	Ref	Alt	Annotation
SN10600087_4671_E145b_1	LACTB	MRPS27	RE	15	63421767	T	C	exonic
SN7640113_5312_E677d_1	SPEN	SF3B3;SNORD111;SNORD111B	RE	1	16254645	G	A	exonic
SN7640113_5312_E677d_1	NRG1	SF3B3;SNORD111;SNORD111B	RE	8	32406278	A	G	exonic
SN7640113_5314_S97_1	SPTAN1	SHANK1	RE	9	131367308	T	G	splicing
SN7640113_5548_ROL_0451_1	STARD13	ITPR1;ITPR1-AS1;SUMF1	RE	13	33700223	C	T	exonic
SN7640113_5548_ROL_0451_1	RYR2	ITPR1;ITPR1-AS1;SUMF1	RE	1	237730032	A	G	exonic
SN7640113_5548_ROL_0451_1	HOMER2	ITPR1;ITPR1-AS1;SUMF1	RE	15	83561556	G	C	exonic
SN7640113_5558_ROL_0481_1	EPS15L1	AGFG2	RE	19	16528403	C	T	exonic
SN0000000_8623_PND5133937_1	DDX41	U2SURP	IGE	5	176939650	G	C	splicing

4.4.8 Over-representation of gene-disease associations

DisGeNET is a discovery platform integrating information on gene-disease associations from public data sources and literature [255]. The current version (DisGeNET v4.0) contains 429,036 associations between 17,381 genes and 15,093 diseases ranked according to supporting evidence. Over-representation analysis of genes that are deleted in both GGE and RE together (134 genes) showed significant over-representation (empirical p-value = 0.012) of epilepsy-associated genes (*APBA2, CHRNA7, CNTNAP2, F5, GABRA5, GABRB3, GRIN2A, KCNQ1, MT1E, PTPRZ1, SCN1A, SGCG, SSTR4*). We observed a similar result for GGE (49 genes; empirical p-value = 0.009; overlapping genes: *CNTNAP2*, *F5, MT1E, PTPRZ1, SCN1A, SGCG,* and *SSTR4*), but we did not see an over-representation in RE (85 genes; empirical p-value = 0.217; overlapping epilepsy genes are *APBA2, CHRNA7, GABRA5, GABRB3, GRIN2A,* and *KCNQ1*). This may reflect the heterogeneous risk factors in adulthood epilepsies compared to RE.

4.4.9 Protein-protein interaction network analysis

We searched for network modules carrying a higher deletion burden with Ingenuity Pathway Analyser (IPA®). Considering GGE and RE together and using brain-expressed genes as an input for IPA we identified a total of 12 networks. The identified network scores ranged from two to 49 and the number of focus molecules in each network ranged from one to 24. Of all the 12 identified networks (see Supplementary Material), the network shown in Figure 4.2 is the top-ranked network with a score of 49 and 24 focus molecules. It is associated to the terms "Nervous system development and function", "Neurological disease" and "Behavior". The network reveals an interesting module where the genes *CAPN1, GRIN2A, ITPR1, SCNA1* and *CHRNA7* are central. Interestingly, *CAPN1* is well ranked (no deletion or duplication) in the ExAC CNV records (Table A.4) and it is not covered by BCG, epilepsy and autism data sets used in this study.

Physiological System Development and Function	
Name	p-value
IGE+RE	
Nervous System Development and Function	2.74E-02 - 3.36E-06
Tissue Morphology	2.62E-02 - 4.20E-06
Behavior Auditory and Vestibular System Development and Function	2.37E-02 - 3.63E-05
Organ Morphology	2.43E-02 - 5.29E-04
RE	
Nervous System Development and Function	4.90E-02 - 3.89E-05
Tissue Morphology	4.90E-02 - 1.34E-04
Behavior	4.90E-02 - 2.56E-04
Auditory and Vestibular System Development and Function	4.53E-02 - 2.59E-04
Organ Morphology and Vestibular System Development and Function	4.90E-02 - 2.59E-04
IGE	
Nervous System Development and Function	4.91E-02 - 2.28E-04
Tissue Morphology	4.07E-02 - 2.28E-04
Behavior	4.47E-02 - 4.62E-04
Hematological System Development and Function	3.81E-02 - 6.79E-04
Immune Cell Trafficking	3.81E-02 - 6.79E-04

Table 4.3: Physiological System Development and Function

Top Networks	
IGE+RE	Associated Network Functions
1	Nervous System Development and Function, Neurological Disease, Behavior
2	Connective Tissue Disorders, Developmental Disorder, Skeletal and Muscular Disorders
3	Cell-To-Cell Signaling and Interaction, Molecular Transport, Small Molecule Biochemistry
4	Cancer, Organismal Injury and Abnormalities, Reproductive System Disease
5	Carbohydrate Metabolism, Lipid Metabolism, Small Molecule Biochemistry
RE	
1	Neurological Disease, Organismal Injury and Abnormalities, Psychological Disorders
2	Cell Morphology, Nervous System Development and Function, Tissue Morphology
3	Cellular Development, Cellular Growth and Proliferation, Hematological System Development and Function
4	Embryonic Development, Organismal Development, Tissue Morphology
5	Cellular Compromise, Cell Cycle, Amino Acid Metabolism
IGE	
1	Carbohydrate Metabolism, Small Molecule Biochemistry, Cell Signaling
2	Cancer, Organismal Injury and Abnormalities, Endocrine System Disorders
3	Cancer, Dermatological Diseases and Conditions, Organismal Injury and Abnormalities
4	Lymphoid Tissue Structure and Development, Tissue Morphology, Behavior

Table 4.4: Top enriched networks from the IPA analysis.

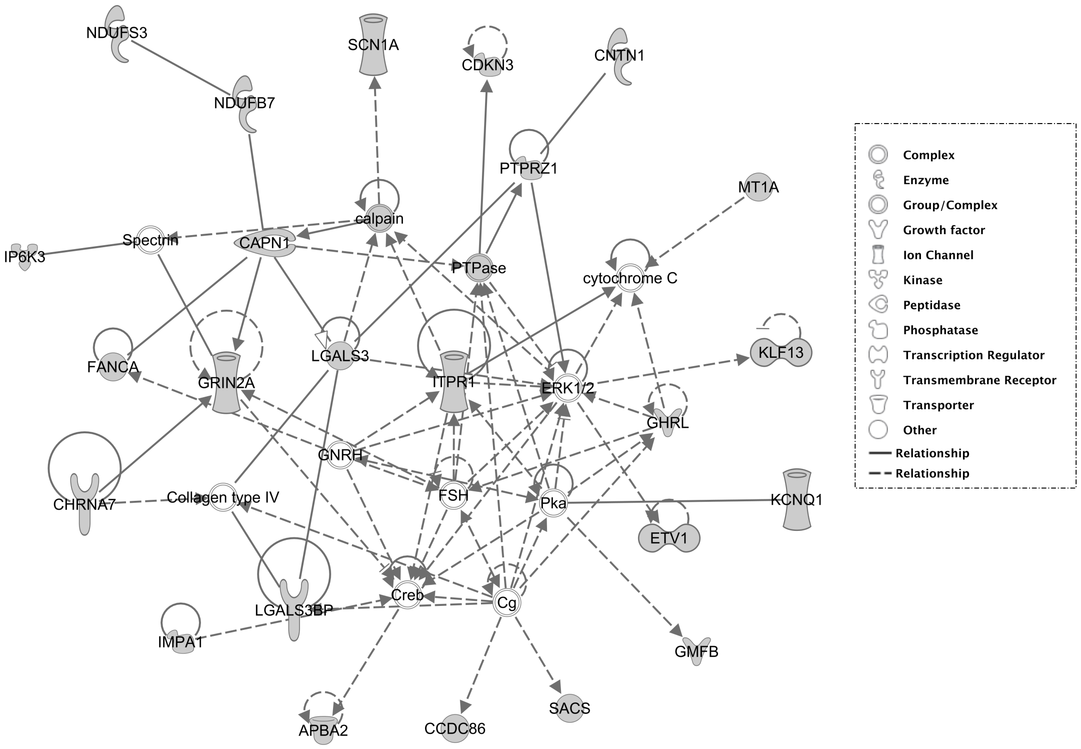


Figure 4.2: Network analysis of brain-expressed genes. The genes were filtered by the CNVs identified in both GGE and RE together. The top network from the pathway analysis generated by Ingenuity Pathway Analyser (IPA®) is shown.

4.4.10 Enrichment for likely disruptive *de-novo* mutations

Many studies on neuropsychiatric disorders such as autism spectrum disorder, epileptic encephalopathy, intellectual disability and schizophrenia have utilized massive trio-based whole-exome sequencing (WES) and whole-genome sequencing (WGS). Epilepsy candidate genes with *de novo* mutations (DNMs) were searched in the NeuroPsychiatric De Novo Database, NPdenovo [256]. DNMs were found in *GABRB3, SHANK1, ITPR1, GRIN2A, SCN1A, PCDHB4* and *IQGAP2*.

4.5 Discussion

We analysed a WES dataset of 390 epilepsy patients (196 GEE, 194RE) for microdeletions. The deletion rate per individual with at least one deletion in cases compared to 572 controls showed statistical significance in both GGE and RE. Enrichment for known epilepsy and autism genes led to gene sets with synaptic and receptor functions which were mainly represented in Rolandic cases (Table 4.3). The top PPIN enriched in GGE was associated with "carbohydrate metabolism", "small molecule biochemistry" and "cell signaling", whereas the top networks associated with RE are “neurological disease”, “organismal injury and abnormalities” and “psychological disorders”, this is reminiscent of our previous attempt to classify metabolic and developmental epilepsies [14].

Among single-gene deletions, *CDH22, CDH12* and *CDH8* are of particular interest; *CDH12* is a cadherin expressed specifically in the brain and its temporal pattern of expression seems to be consistent with a role during a critical period of neuronal development [257]. Moreover, a group of cadherins, *CDH7, CDH12, CDH18* and *PCDH12*, are reported to be associated with bipolar disease and schizophrenia [258]. The smallest deletion (1,166 bp) that we could detect in this study concerns *NCAPD2*; this gene is annotated in the autismkb database [259]. It is an important component of the chromatin-condensing complex, which is highly conserved across metazoan. This gene was previously found to be associated with Parkinson’s disease [234] and its paralog *NCAPD3* is associated with developmental delay [260].

Deletions of brain-critical exons pointed to the *ITPR1* deletion, which has been reported to be associated with spinocerebellar ataxia type 16 [261, 262]. *CNTN1* is another deletion of interest, the gene is highly expressed in fetal brain, it encodes a neural membrane protein which functions as a cell adhesion molecule and may be involved in forming axonal connections/growth and in neuronal migration in the developing nervous system [263, 264]. Moreover, its paralogs *CNTN2* and *CNTN4* are associated with epilepsy [265] and autism [266], respectively. Interestingly, in the ExAC data, the brain-expressed genes *ITPR1* and *CNTN1* show the third and fourth highest intolerance score ranks, respectively (Table A.4).

Protein-Protein interaction network analysis revealed the *CAPN1* deletion as an interesting candidate gene; this is a double gene loss (4,270 bp) spanning *CAPN1* (exon 17 to 22 out of 22 exons) and *SLC22A1* (exon 1 out of 10 exons). *SLC22A1*, a transporter of organic ions across cell membranes, is lowly expressed in the brain, whereas *CAPN1* is highly expressed in the brain. Calpain1 (*CAPN1*) belongs to the calcium-dependent proteases, which play critical roles in both physiological and pathological conditions in the central nervous system. They are also recognized for their synaptic and extra-synaptic neurotoxicity and neuro-protection [267]. Several ion channels, including *GRIN2A*, [268] are calpain substrates. Further, a missense mutation in *CAPN1* is associated with spino-cerebellar ataxia in the Parson Russell terrier dog breed [269] and has recently been reported in humans with cerebellar ataxia and limb spasticity [270].

Additional candidate genes can be identified on the periphery of the IPA network (see Fig 2): 1) *CNTN1* (commented on above), 2) *SACS*, for which a large deletion (> 1Mb) was found, and 3) the single gene deletion of *KCNQ1* (~ 57 kb). For *SACS*, a SNV is reported to be associated with spastic ataxia [271] and epilepsy [272]. *KCNQ1* and its paralog *KCNQ3* are subunits forming an expressed neuronal voltage-gated potassium channel. Further, hypomorphic mutations in either *KCNQ2*, an established epilepsy-associated gene [273], or *KCNQ3* are reported to be highly penetrant [274]. *KCNQ1* is co-expressed in heart and brain; it is found in forebrain neuronal networks and brainstem nuclei, regions in which a defect in the ability of neurons to repolarize after an action potential can produce seizures and dysregulate autonomic control of the mouse heart [275], yet one should be cautious as no validation is available for human.

Enrichment for likely disruptive *de novo* mutations in several genes suggests that deletions of these genes could cause a similar phenotype as in the NPdenovo and consequently will be penetrant in the heterozygotic state. This is indeed the case for *ITPR1*, for which recessive and dominant *de novo* mutations causing Gillespie syndrome [276], a rare variant form of aniridia characterized by non-progressive cerebellar ataxia, intellectual disability and iris hypoplasia, have been described. Two of the genes, which we have identified as *ITPR1* interactors, *RYR2* and *SPTAN1*, are also DNM genes in DPdenovo.

In summary, by filtering and comparison to genes that are (1) evolutionary constrained in the brain, (2) implicated in autism and epilepsy, (3) spanned by ExAC deletions, or (4) affected by neuropsychiatric associated *de novo* mutations, we observed a significant enrichment of deletions in genes potentially involved in neuropsychiatric diseases, namely *GRIN2A*, *GABRB3, SHANK1, ITPR1, CNTN1, SCN1A, PCDHB4*, *IQGAP2, SACS, KCNQ1* and *CAPN1.* Interaction network analysis identified a hub connecting many of the epilepsy candidate genes identified in this and previous studies. The extended search for likely deleterious mutations in the first order protein-protein interactions and NPdenovo database pointed to the potential importance of *ITPR1* deletion alone or in combination with *RYR2* and *SPTAN1* deleterious mutations.

We are aware that the set of epilepsy exomes that we screened for CNVs in the present study, although the largest analyzed so far, is still small given the genetic complexity of the disease and its population frequency. However, this study appears to provide a contrasting view to the genetic bases of childhood and juvenile epilepsies, as the top protein–protein interactions showing that GGE deleted proteins are preferentially associated with metabolic pathways, whereas in RE cases the association is biased towards neurological processes. Scrutinizing of additional patients' exomes/genomes and transcriptomes should provide an efficient way to understand the disease aetiology and the biological processes underlying it. The results presented here may contribute to the understanding of epilepsy genetics and provide a resource for future validations to improve diagnostics.

CHAPTER 5

EXCESS OF SINGLETON LOSS-OF-FUNCTION VARIANTS IN PARKINSON'S DISEASE

5.1 Abstract

Parkinson's disease (PD) is a complex disease. Besides variants in high-risk genes, multiple other genes associated to sporadic PD were discovered via genome-wide association studies. Yet, there are a large number of genetic factors that remain unexplored. In order to unravel the genetic factors that play a role in PD, we studied the whole-exome sequencing data available as a part of Parkinson Progression Markers Initiative (PPMI). After quality filtering, the final dataset comprised of 352 PD cases and 149 ethnically matched controls. We performed burden tests at exome-wide level for different variant classes. We observed a significant exome-wide burden of singleton loss-of-function variants in cases compared to the controls (corrected P_{glm}=0.01, OR=1.09, CI=1.03-1.16 and corrected P_{emp}=0.002) but not in the singleton synonymous variants (corrected P_{glm}= 1, OR=0.99, CI=0.97-1.02 and corrected P_{emp} = 0.55). Furthermore, no burden of singleton loss-of-function (LoF) was identified in a group of genes identified via genome-wide associated genes, pointing into the direction of polygenic burden. Additionally, no significant exome-wide burden of rare variants was detected either. Our study supports the complex disease notion of PD by highlighting its convoluted architecture. Finally, we built a prediction model with an AUC=0·709 ± ·0047 (95% CI) based on logistic regression with a combination of singleton LoF variants, common poly risk variants, and family history of PD as the features. Our results outperform the state-of-the-art classification model for the PPMI data set [277], which reached an AUC=0.639 based on common variants. By just adding two more features we reached an AUC=0·709 and we show that the addition of a novel singleton LoF score

per individual substantially improves the AUC. The main finding of this study is to discover the complex genetics of PD at an exome-wide level and to show that prediction models based on rare/ultra-rare variants plus common variants perform better. Such prediction models could aid the clinicians in decision making during the diagnosis of PD.

5.2 Introduction

Parkinson's disease is a neurodegenerative disorder and is linked to several genetic and environmental factors. Several genes were identified by sequencing studies that were conducted under familial design [278]. Large scale meta-analyses have identified several genes that are associated to PD [109] in the case-control setting [279, 280]. As the common variants alone lack to explain the entire heritability of PD, there might be other causes such as DNA methylation levels [281], rare [19], ultra-rare or singleton variants (seen in only one sample in the cohort), variants which could fill in the missing gap [282].

In order to identify the disease associated variants/genes, an array of burden tests [199, 283] have been developed to aggregate the signal from rare or common variants acting in a similar direction or with different directions. Even after aggregating the variants at the level of genes, there is still a limited power to attain genome-wide statistical significance and we often require larger sample sizes to uncover novel disease associations. To increase the statistical power, variants can be aggregated at a higher level instead of gene sets and pathways, or across the whole genome. For instance, it has been previously shown that, in schizophrenia there is an excess of genome-wide ultra-rare variants [78] in cases versus controls and also in a group of genes [284]. Whereas, in sudden unexpected death in epilepsy [285] there is a genome-wide excess of rare disruptive variants. In this study, we investigated the whole exome sequencing (WES) data from PPMI consortium [286] and performed an exome-wide burden analysis by aggregating the rare and singleton variants in the entire exome.

Previous studies have also conducted the analyses based on the genetic data from PPMI to build predictive models in order to differentiate PD cases from the healthy controls [287–289] and to sub-type the PD cases [290]. Similarly, the PPMI exome sequencing data has been employed as a replication dataset to show a significant burden in a group of 54 lysosomal genes in PD [291] and to test the burden of rare loss-of-function (LoF) variants in 27 candidate genes [19]. Further, it was utilized to describe LoF variants in *TRAP1* [292]. But, an unbiased exome-wide study to test the burden in PD cases versus healthy controls is still missing. A previous study also aimed to identify the rare variants in PD by conducting the burden analyses [293] and showed the partial role of rare variants in PD. In our study, we performed burden analyses at exome-wide level and showed an increased burden of singleton LoF variants in cases versus controls. Our findings implicate non-synonymous as well as stop-altering and splice site variants at a genome-wide level and highlight the polygenic nature of PD.

On the basis of polygenic risk score (PRS), singleton count, and the family history of PD, we trained a logistic regression and a random forest to classify PD and healthy controls with a relatively high AUC of 0·709. This approach highlights that, rare/ultra-rare variants along with the common variants confer a risk for PD and they should also be included in generating the PRS for PD. The significance of singleton count alongside standard polygenic risk could translate to improved prediction models for PD.

5.3 Patients and methods

The Parkinson's Progression Markers Initiative (PPMI) study is an effort to identify the biomarkers of PD progression [286]. Detailed information about this initiative and the data can be found on their website (`http://www.ppmi-info.org`). Exome sequencing was performed on whole-blood extracted DNA samples collected according to the PPMI Research Biomarkers Laboratory Manual using Illumina Nextera Rapid Capture Expanded Exome Kit. Nextera Expanded Exome targets 201,121 Exons, UTRs and miRNA and covers 95.3% of Refseq exome. >340,000 probes are constructed against the human NCBI37/hg19 reference genome. Targeted genomic footprint is 62Mb. Library preparation for next-generation sequencing using Nextera Rapid Capture Expanded Exome Kit was performed per manufacturer's protocol (Illumina, Inc. San Diego). Exome- enriched libraries (multiplexed sets of 12 samples) were sequenced on the Illumina HiSeq 2500 sequencing platform using 2 x 100 bp paired-end read cycles. Briefly, the variants were called following GATK [23] best practices. The initial PPMI exome dataset comprised of 404 PD and 183 healthy controls, which were filtered by several criteria as described below.

5.3.1 Low quality samples filter

Number of alternate alleles, number of heterozygotes, Ti/Tv ratio, number of singletons and call rate served as data quality parameters. They were calculated by PLINK/SEQ (`https://atgu.mgh.harvard.edu/plinkseq`) i-stats command. Any sample with >3 standard deviation (SD) from the mean in any of the above mentioned metrics was excluded from the analysis. Next, we selected the variants that were common between HapMap(version 3.3) [37] and the current dataset. The selected variants were further filtered to be: 1) Only bi-allelic SNVs, 2) with a call rate >98% and 3) not in linkage disequilibrium. The variants filtered above were included to check cryptic relatedness, deviations from reported sex and to perform population stratification analysis via eigenstrat [42].

5.3.2 Relatedness filter

Cryptic relatedness check was performed via both PLINK [39] and KING [40] algorithms based on the same set of SNVs as described above. We checked up to second degree relatedness (Pi_Hat score >0.25) and randomly chose one sample of the identified relative pairs to be included in the final analyses.

5.3.3 Ethnicity filter

We merged our data with the 1000 genomes (1000g) data and performed population stratification employing eigenstrat with default parameters to confirm the ethnicity of our samples. Except for few outliers, both cases and controls were clustered with the samples of European origin in the 1000g data Figure 5.2. In order to determine the ethnicity outliers from the eigenstrat analysis, a sigma value of 3 was applied as a cut-off (which excludes all the samples with a SD of >3 based on the first 10 principal components). Additionally, we excluded the samples >3SD based on the first and second principal components from the eigenstrat analysis.

5.3.4 Variant QC

The downloaded PPMI vcf file had already been filtered for high quality variants according to the variant quality score recalibration approach as part of GATK best practices by the authors of original study. In order to be more stringent, we applied additional filters as described below: 1) For SNVs: Variants were filtered for QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, DP<10.0, GQ_MEAN<20.0, VQSLOD<0, ABHet >0.75 or <0.25 and Hardy Weinberg Phred scale P-value of >20. 2) For indels: Parameters for variant filtration were QD < 2.0, FS > 200.0, ReadPosRankSum < -20.0, DP<10.0, GQ_MEAN<20.0. Additionally, filtering based on individual genotype quality and read depth is performed by converting the variant genotypes with a read depth of <10 and GQ of <20 to missing by the bcftools [27]. Finally, only variants with a call rate of >0.9 were kept for further analyses.

5.3.5 Variant annotation

Multi-allelic variants were decomposed based on variant-tests [45] and left normalized by bcftools [27]. Variants were annotated by ANNOVAR [47] version 2016 June17 using RefSeq and Ensembl gene annotations, the dbNSFP v3.0 [51] prediction and conservation scores as well as genome-wide CADD [58] scores. Exonic and splice site variants (EXONSPLICING) were selected according to RefSeq and Ensembl annotations. Rare variants were defined as variants with minor allele frequency(MAF) < 0.005 in European population of public databases such as 1000 genomes [34], ExAC (release 0.3) [43], and the Exome variant server (`http://evs.gs.washington.edu/EVS`). Singleton variants were defined as the variants present in only one sample in the entire dataset (AC=1). We divided the rare and singleton exonic and splicing variants into different variant classes such as: a) nonsynonymous+LoF variants (NONSYN+LoF), b) loss-of-function (LoF) variants defined as stop gain, stop loss, splice site variants and all insertions/deletions (LoF), c) NONSYN variants with a CADD phred score >10 (CADD10), d) NONSYN variants with a CADD phred score >20 (CADD20) and e) synonymous variants (SYN) as a control variant set, as they are assumed to be functionally neutral. All the analyses described below were performed separately for each variant class for both rare and singleton variants.

5.3.6 Excess of singleton variants

We checked whether rare and singleton variants (variants present in only one sample, AC=1) were overrepresented in cases compared to the controls. In order to do that, first we generated an individual burden score for each sample by counting number of variants in each variant class. Then we compared the individual burden score of cases and controls by two different approaches: First, for each variant class, we constructed a generalized linear model by correcting for total number of singleton variants called in that sample, gender and first 10 principal components from the eigenstrat analysis as covariates and a P-value(P_{glm}) was generated. Second, coverage or sample size bias could lead to an increased number of singletons, in order to account for this bias, we performed 10,000 sample label permutations and for each permutation we computed the one-sided Wilcoxon rank sum test [294, 295] to calculate a P-value, by comparing the individual burden score per sample between cases and controls. Then, the permutation P-values were compared with the original P-value to generate an empirical P-value (P_{emp}). We chose the Wilcoxon rank sum test because it accounts for differences in sample sizes and the presence of any outlier samples [296]. R version 3.4.2 was employed to calculate all the P-values. We corrected for 10 comparisons for multiple variant classes (5 variant classes in rare and singleton groups) according to the “bonferroni” method implemented in function “p.adjust” in R version 3.4.1.

5.3.7 Geneset burden analysis

In order to identify whether there was a polygenic burden or only a few genes contribute to the observed burden, we restricted our burden analysis as described above to a group of 74 PD associated genes that were identified previously in a large-scale meta-analysis [109].

5.3.8 PRS generation

After the QC, the final dataset comprised of 352 cases and 149 controls, the summary of the samples along with the clinical scores are given in Table 5.1. In order to generate PRS per sample, summary statistics of 43 SNPs that were found to be genome wide significant from the meta analysis [109] were selected are given in Table A.2. PRSice [297] with default parameters was used to calculate PRS per each sample. In addition, we also included the *LRRK2* p.G2019S into the PRS calculation (PRS_LRRK2). However, we did not include it in the final analysis as it showed inferior predictive ability (Table 5.1) compared to PRS without accounting for *LRRK2* p.G2019S variant. A clear difference in the distribution of PRS in cases and controls can be seen in Figure 5.1.

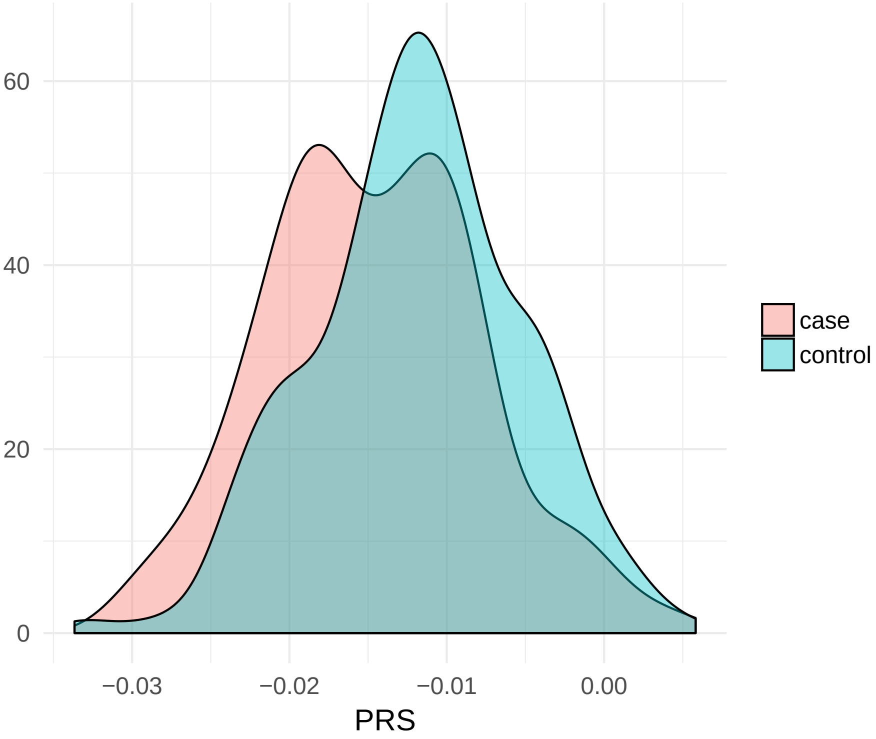


Figure 5.1: Distribution of PRS. There is a clear shift in PRS in the cases compared to the controls.

Feature	**Cases (n=352)**	**Controls (n=149)**	**ANOVA/Chi-sq P-value**	**F-statistics**
Clinical features				
QUIP Score	0.285 (0.621)	0.262 (0.728)	7.157 E-1	0.13
ESS score	5.949 (3.500)	5.581 (3.470)	2.837 E-1	1.15
Benton Summary Score	12.835 (2.127)	13.148 (1.919)	1.234 E-1	2.38
Total Semantic Fluency Score	48.855 (11.544)	51.987 (11.009)	5.182 E-3	7.88
T Anxiety	32.137 (9.284)	28.456 (6.773)	1.571 E-5	19.02
Family history of PD (%)	87 (24.71)	7 (4.69)	1.553 E-7	27.52
MoCA Total Score	27.182 (2.260)	28.242 (1.127)	8.673 E-8	29.52
SCOPA-AUT Total Autonomic	12.273 (8.757)	7.824 (6.840)	6.014 E-8	30.27
REM Sleep Behavior Score	4.233 (2.690)	2.812 (2.236)	2.551 E-8	32.04
S Anxiety	32.906 (10.087)	27.349 (7.516)	3.069 E-9	36.45
Symbol Digit Modalities Total Correct	41.392 (9.655)	47.456 (10.814)	1.302 E-9	38.24
UPDRS Score Part I	5.500 (4.021)	2.716 (2.518)	3.986 E-14	60.65
UPDRS Score Part II	5.923 (4.171)	0.392 (0.984)	0.000 E+0	253.26
UPSIT Raw Score	22.429 (8.251)	34.443 (4.394)	0.000 E-0	280.91
UPDRS Total Score	32.125 (12.808)	4.304 (4.114)	0.000 E+0	666.62
UPDRS Score Part III	20.719 (8.751)	1.196 (2.195)	0.000 E+0	714.66
Non-clinical features				
Male (%)	235 (66.76)	97 (65.10)	7.193 E-1	0.13
age at onset/age of last examination	61.841 (9.584)	60.934 (10.463)	3.472 E-1	0.89
PRS_LRRK2	0.093 (0.008)	0.091 (0.007)	1.168 E-2	6.41
Singleton Count	12.165 (4.326)	10.631 (3.739)	1.861 E-4	14.18
PRS	-0.012 (0.008)	-0.015 (0.007)	2.268 E-5	18.3
Family history of PD (%)	87 (24.71)	7 (4.69)	1.553 E-7	27.52

Table 5.1: Summary statistics and predictive ability of various clinical scores available from the PPMI consortium and the features generated in this study. For independence/significance testing we applied ANOVA for continuous data and Chi-square for binary data. The values in brackets indicate SD values unless stated otherwise.

5.3.9 Construction of risk models

Several PD risk models were built previously [277] by utlilizing a PRS which is generated based on common variants. However, in our study we observed a significant difference in the count of singleton LoF variants between cases and controls (Figure 5.3). Hence, as an additional variable the count of singleton LoF variant per individual was applied as an additional variable and built an improved prediction model. Two state of the art approaches namely logistic regression and random forest were chosen to construct and test the prediction models. All the analyses were performed using Ada, a novel data exploration and analytic platform developed at Luxembourg Centre for Systems Biomedicine (publication in progress).

Ada is a performant and highly configurable system for secured integration, visualization, and analysis of heterogeneous clinical and omics data sets. Ada allows users to conveniently explore and filter data and produce dynamic and personalized “views” containing charts and widgets for various statistics. Ada currently harbors around 1300 data sets from diverse studies including LuxPark, DeNoPa, PPMI, TREND, GBA, ADNI, and mPower.

For more advanced statistical analysis and machine learning, Ada employs Spark ML library (`https://spark.apache.org`), which is a performant and scalable distributed computing library. This covers a wide variety of classification, regression, clusterization, feature selection, normalization, and time-series processing routines. Ada is available for registered users at `https://ada.parkinson.lu`.

5.3.10 Input Features and Classification Models

A final list of input features was generated by evaluating various clinical and non-clinical features for their predictive ability by employing one way ANOVA for continuous features and Chi-square test for categorical variables. A one way ANOVA compares the means from two independent (unrelated) groups by using the F-distribution. The principle behind ANOVA is that, according to null hypothesis, the means of different groups being compared are equal. Hence, a significant P-value (0.05 in our case) shows that the means of two groups are unequal. The F-statistics and P-values obtained from ANNOVA/Chi-sq test are shown in the Table 5.1. After the selection of input features, we built four models as described below:

- A model based on PRS only ($\text{model}_{\text{PRS}}$)
- A model based on singleton LoF score only ($\text{model}_{\text{singleton}}$)
- A model based on singleton LoF score and PRS ($\text{model}_{\text{singleton\_PRS}}$)
- Finally, an integrated model comprising of singleton LoF score, PRS and PD family history ($\text{model}_{\text{integrated}}$).

The parameters of our classification models were set to defaults provided by Spark ML library:

- Binomial logistic regression - L2 regularization, fitting the intercept, max. 100 iterations, and tolerance of 10E-6.
- Random forest with depth 2 - max. 32 bins, 20 trees, without subsampling of training data.

As we discuss in Section 5.4.4 the reason for a rather shallow architecture of the random forest is a small amount of input features, which leads to overfitting. Before training we normalized the features to z-scores and obtained two sets: with and without 50% subsampling of cases. Each iteration we split the sets randomly with 0.9 training-test ratio and fed the training part to our classifiers. We repeated this process 1000 times and reported the mean test AUC as a target evaluation metric.

5.4 Results

5.4.1 Population stratification and relatedness check

As it can be seen from Figure 5.2, except for a few outliers both the cases and controls were clustered together with the European samples of the 1000 genomes data. This observation is in line with the previous observations from another study based on PPMI data which was performed on genotype array data [287]. After the filtering based on ethnicity, cryptic relatedness and quality parameters the final dataset comprised of 367 PD and 159 control samples. The quality metrics are shown in Table 5.2, the Ti/Tv ratios of exonic/splicing variants is >3 indicating the good quality.

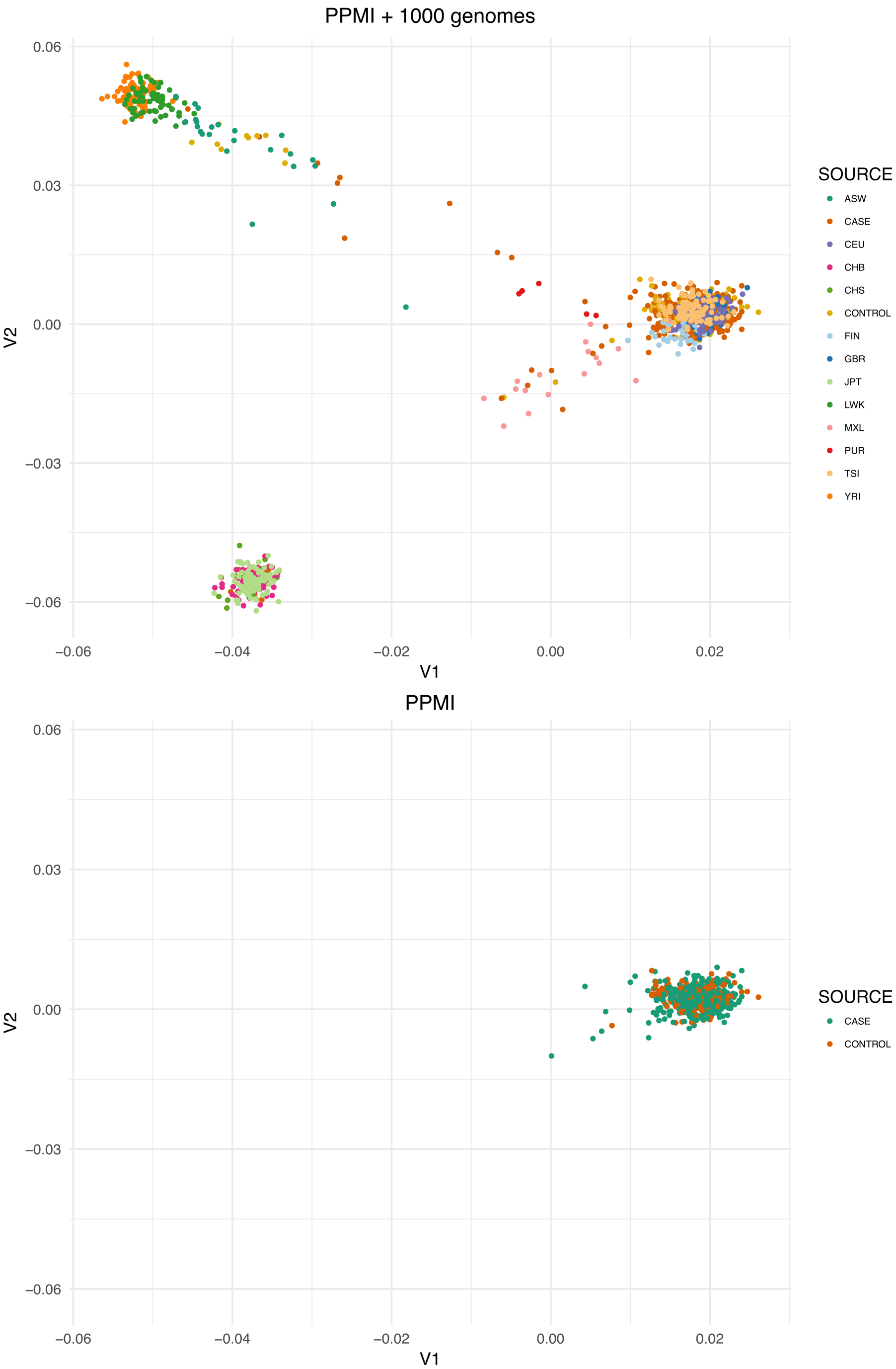


Figure 5.2: A) A figure representing the ethnicity of samples in the current study. The sample were represented along with samples within 1000 genomes study. Each colour represents different ethnicities and each shape represents the super population to which the samples belong to. The abbreviations of the legend are given below. ASW: Americans of African Ancestry in SW USA, CEU, CHB: Han Chinese in Beijing, China, CHS: Southern Han Chinese, FIN:Finnish in Finland, GBR: British in England and Scotland, JPT: Japanese in Tokyo, Japan, LWK: Luhya in Webuye, Kenya, MXL: Mexican Ancestry from Los Angeles, PUR: Puerto Ricans from Puerto Rico, TSI: Toscani in Italia, YRI: Yoruba in Ibadan, Nigeria. AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European. B) Samples included in the analyses after final QC.

Number of cases	367
Number of controls	159
Number of variants	4,87,024
Number of exonic/splicing variants	2,32,762
Ti/Tv ratio of exonic/splicing variants	3.07

Table 5.2: Metrics of PPMI dataset after QC.

5.4.2 Excess of rare singleton LoF variants

Exome-wide burden was not seen when we performed the burden analysis of rare variants, however as shown in Figure 5.3, when we restricted our analysis only to singleton variants it could be seen that there is an excess of singleton LoF (corrected P_{emp}=0.002, corrected P_{glm}=0.01) variants in cases compared to controls [298]. Whereas, no significant difference was seen between cases and controls in neither the SYN variants (corrected P_{emp} =0.55, corrected P_{glm}=1) nor in the other variant classes.

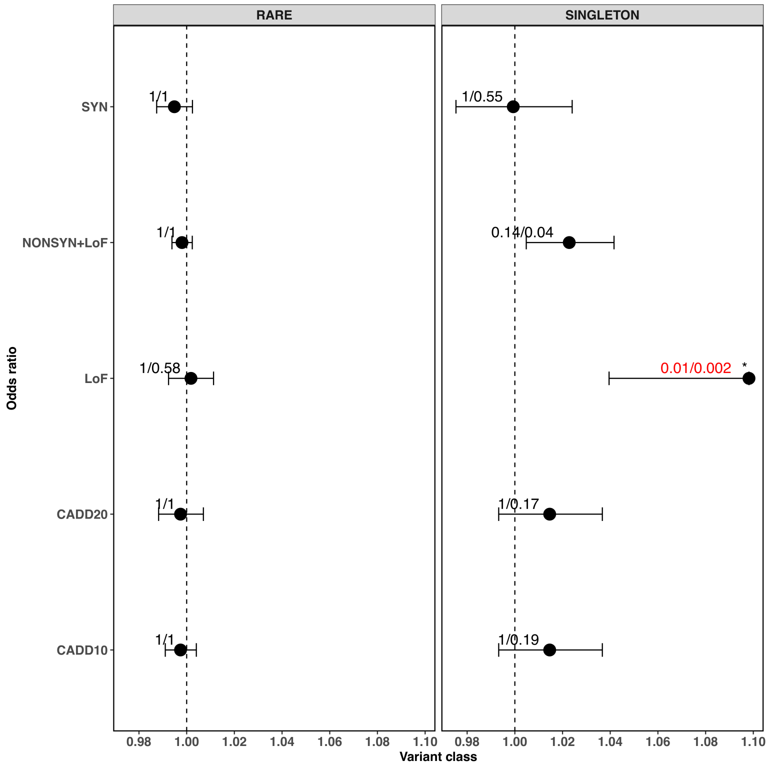


Figure 5.3: Plot representing the excess of singleton variants in cases versus the controls. Each dot represents the odds ratio generated by the glm. The values on top of each point represents corrected P-value from glm, empirical P-value from wilcoxon rank sum test respectively, they are separated by "/". If both the corrected P-values were below 0.05 they are highlighted in red with an "*" on top.

5.4.3 Evidence of polygenic burden in PD

When we restricted our analysis to a group of genes that were significantly associated to PD, we did not detect an increased burden of singleton LoF variants (P_{emp}= 0.5, P_{glm}=0.49). This shows the polygenic nature of PD where there is a distribution of burden across the genome rather than being confined to a group of already known PD associated genes.

5.4.4 Prediction Performance

We trained our prediction models on non-clinical features only. It is due to the fact that, undoubtedly, the clinical scores are designed to distinguish the PD cases from healthy controls, the very classification problem we aim to predict. Therefore, they make the prediction rather trivial.

For instance, the clinical scores of University of Pennsylvania Smell Identification Test (UPSIT) and Unified Parkinson's disease rating scale (UPDRS), which describe certain aspects of PD phenotypes, separate the cases and controls into two distinct groups as can be seen in Figure 5.4. In our experiments the prediction models based on these two features could easily reach an AUC>0.95 (results not reported here). Additionally, by performing ANOVA/chi-square test we demonstrated that a majority of the clinical features have a very low P-value and thus posses a high predictive power (Table 5.1).

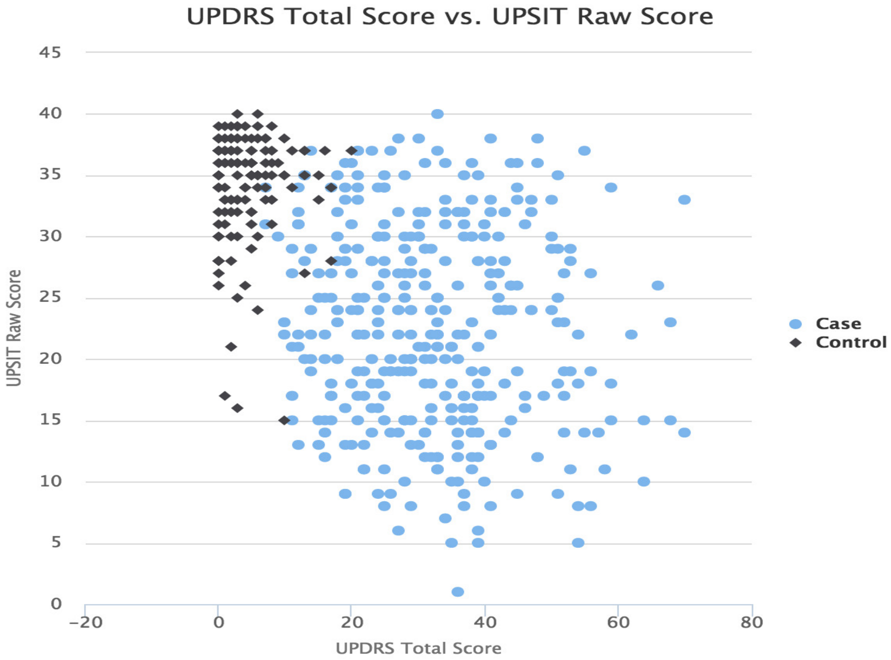


Figure 5.4: UPDRS score versus the UPSIT score of the samples in PPMI dataset. The cases and controls are separated into two distinct groups.

In our final construction we shortlisted our features to three: PRS, singleton count and family history of PD as they are shown to have the most significant predictive power out of all non-clinical features we considered (Table 5.1). Moreover, since we aimed for a minimal set of predictive features we did not feed gender and age of onset to our models, which were commonly employed in previous studies [277, 287].

Our main result is that by combining PRS, singleton LoF count and PD family history we reached an AUC of 0·709 ± ·0047 (95% CI). Performance of the partial models with the PRS only, the singleton LoF count only, as well as the PRS and singleton LoF count combined were substantially lower: 0·621 ± ·0054 (95% CI), 0·604 ± ·0051 (95% CI), and 0·654 ± ·0053 (95% CI) respectively. The reported AUC is a mean over 1000 repetitions on test sets randomly drawn with 0.9 training-test split for the binomial logistic regression.

Our predictor that is built on the combination of common and singleton count with an AUC of 0·653 ± ·0·005 (95% CI) outperforms the state-of-the-art classification model for PPMI dataset built on the basis of PRS [277] with an AUC=0.639, which also employed a logistic regression. By feeding solely the PRS to our logistic regression we are reaching an AUC=0·621 ± ·0054, which is comparable to the previous study [277]. The difference in performance could be due to different utilization of SNPs, samples, and methods to generate the PRS. Finally, by adding the family history of PD an AUC climbs to 0·709 ± ·0047. That is more than 10% performance improvement compared to the state-of-the-art by applying only 3 non-clinical variables.

The study [277] also presented an UPSIT-score-only model with a very high performance (AUC = 0·901 ± ·027 (95% CI)). By adding the demographic features and PRS they attained an AUC = 0·923 ± ·23 (95% CI). Even though, it is a significant increase as shown in the study based on DeLong's test for correlated ROC curves ($|z| = 3.027$, p-value = 0.002), in relative terms the PRS could increase the AUC only marginally and thus, the prediction is almost fully dominated by the UPSIT score. We wanted to avoid that and perform a more challenging prediction without including any clinical scores as discussed on the top of this section.

Besides the logistic regression we trained also another machine learning classifier, a random forest. As presented in Figure 5.5, the logistic regression performs better than the random forest. This is due to the fact that our classifiers were fed with a very few variables (1-3), which makes the task too simple for the random forest. As opposed to the logistic regression, which has almost identical performance on the training and test sets, the random forest overfits the training data (AUC=0.739). This would even worsen for random forests with larger depths (hence the shallow setting). In future work, instead of being minimalistic we will utilize dozens or hundreds of partial or intermediate genetic variables, which are expected to favor the random forest.

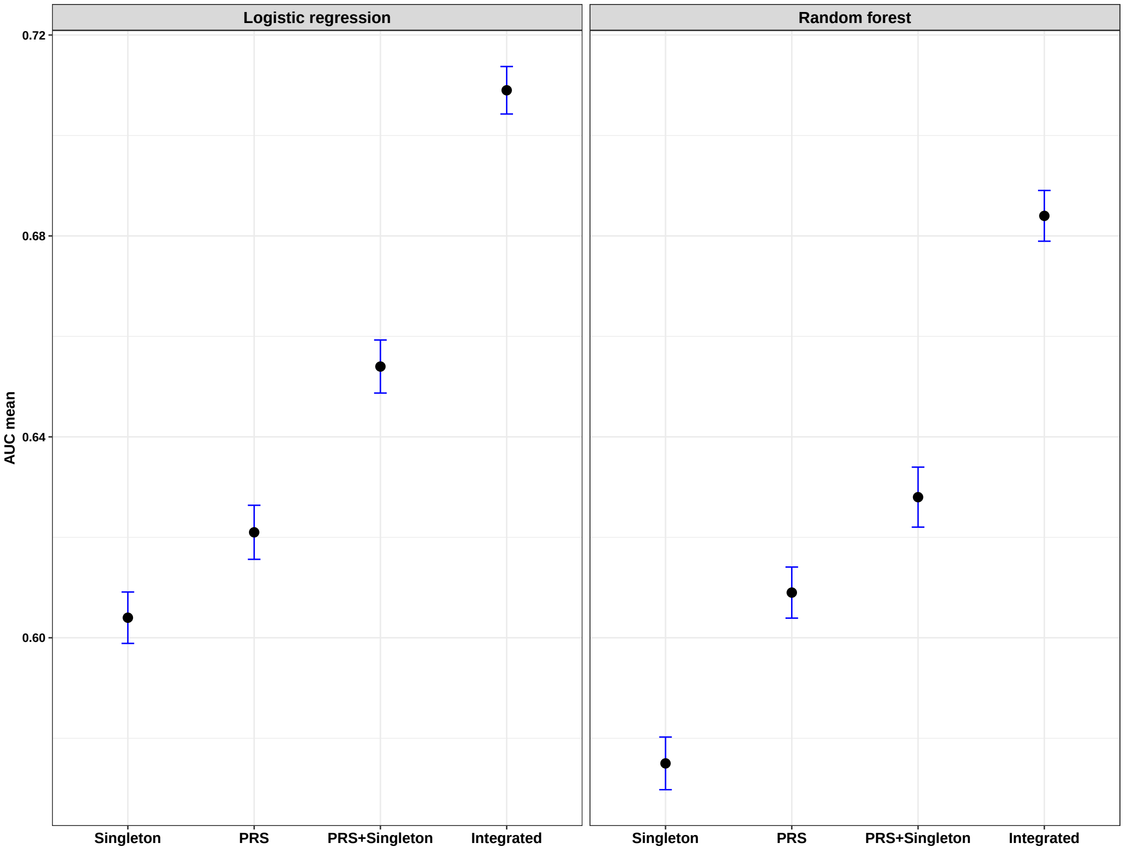


Figure 5.5: The AUC values of various models at 95%CI. Two predictors namely logistic regression and random forest were applied. PRS = PRS only model, Singleton = A model based on singleton LoF score only, PRS+Singleton = A model based on singleton LoF score and PRS and Integrated = Model comprising of singleton LoF score, PRS and PD family history.

5.5 Discussion and Conclusion

Even 200 years after the first description of PD by James Parkinson, its diagnosis is still a challenge and no curative treatment available. By studying the WES data of 367 PD cases and 159 controls we have shown a polygenic burden increases risk for PD. This burden mainly consists of multiple singleton LoF variants distributed across the exome.

Identification of individual genes that show a genome-wide significance is often difficult primarily due to the small sample sizes and multiple testing correction. However, our results indicate the additive contribution of singleton LoF variants of an individual to the aetiology of PD. This finding cannot be attributed to a bias as we have corrected for various confounding variants by applying the generalized linear models and additionally by performing sample label permutations. Moreover, to further strengthen our findings, we see a significant burden of singleton LoF variants but not in functionally neutral singleton synonymous variants in PD. Based on the evidence from the current study, we speculate that the genetic risk of PD is not confined to a group of genes but instead, is distributed across the exome. Hence, in summary our results support the polygenic inheritance and complex genetic architecture of PD.

In the second logical part of our paper, we trained two classifiers, binomial logistic regression and random forest, on three features: singleton LoF variants, common variants, and family history of PD. Our logistic regression model with an AUC of 0·709 ± ·0047 (95% CI) outperforms the state-of-the-art classification model for PPMI data set for non-clinical features. Also, we have shown that the predictive models built on the features based on rare and common variants perform better compared to the models built on common variants alone. In PD research, a general consensus is that, in very broad terms, PD is triggered by a combination of genetic and environmental factors. Nevertheless, because acquiring clinical scores is expensive and laborious, by limiting ourselves to genetics we make potential diagnostic applications of our models more practical and scalable, acknowledging the evident deficiency of the information provided.

Despite the fact that there is a exome-wide significance of singleton LoF variants, our study should be considered preliminary and needs replication in larger PD cohorts. Identification of variants associated to PD along with the integration of PD specific pathway information that is represented in resources such as PD map [283, 299] could lead to a genetic diagnosis of PD and there is an imperative need to decipher such variants to understand the PD aetiology.

The major limitation of the current study is the small sample size. When studying rare and singleton variants, larger samples sizes are needed to adequately pinpoint certain genes or variants that are associated with the disorder. Another limitation of our study and of WES studies in general is that we could only perform burden analyses of coding variants. However, there might be additional factors such as variants in the non-coding regions which could also contribute to the progression of PD. Clearly, this could be only tested when WGS data is made available. We expect that with an increasing number of samples more accurate predictive models can be constructed and contribution of rare variants in generating these models will improve significantly. In the future more refined strategies to include rare variants in the construction of PRS is warranted. It is our hope that we can extend this work and refine our strategy in order to build an accurate diagnostic model that can be employed in the clinical setting. The PRS could be also applied to stratify the patients for a personalized medical treatment.

CHAPTER 6

BURDEN ANALYSIS OF U1 SPLICE VARIANTS

6.1 Abstract

Parkinson's disease (PD) is a heterogeneous neurodegenerative disorder with monogenic forms representing prototypes of the underlying molecular pathology and reproducing to variable degrees the sporadic forms of the disease. There have been several reports of variants causing an abnormal splicing in PD especially the U1 splice variants. However, a large-scale study measuring the effect of U1 splice variants in PD is still missing. In our study, we performed an exome-wide burden analysis of less common U1 splice variants predicted to be deleterious in the PPMI cohort comprising of 372 cases and 161 controls. Our analysis of exomes revealed that U1 splice-site mutations were enriched in sporadic PD patients compared to the healthy controls and majority of the signal is coming from the genes that are expressed in brain. The observed finding was replicated in a larger independent cohort.

6.2 Introduction

Parkinson's disease (PD) is increasingly recognized as a heterogeneous disorder, as reflected by its substantial phenotypic, neuropathological, and genotypic variability [300]. Therefore, previous models that considered PD as a single disease entity, although successful for developing symptomatic therapies that compensate for the dopaminergic deficit responsible for the motor symptoms of PD, fall short in terms of developing neuroprotective treatment strategies [301]. Focusing on pathomechanisms and understanding the underlying molecular pathology of neurodegeneration is essential, and genetic stratification of patients into subgroups provides an important entry point for precision medicine [302]. During the last 20 years, a substantial number of genes related to PD were identified, including mutations in genes responsible for rare monogenic forms of PD. These monogenic forms of PD have become a valuable resource for PD research, as patient-based cell models display disease-specific cellular phenotypes that recapitulate the phenotypes found in *post-mortem* brain tissue [303]. According to this concept, the validation of clinicogenetic subtypes of PD may be achieved based on rare but strong molecular signatures and subsequently applied to the different pathophysiological tiers within each disease subtype [304].

Mutations disrupting splicing in monogenic PD have recently come into focus, and variants predicted *in silico* to cause aberrant splicing have been described for *PINK1*, *GBA*, *PARK7,* and *PARK2* [121, 305–307]. We show for the first time for the common sporadic form of PD that yet unrecognized mutations in U1 splicing sites are overrepresented in exomes from patients compared to controls. Our findings are in line with large-scale characterization of disease-associated mutations that found splicing mutations largely underestimated and open the door for the mechanisms involving splicing aberrations in PD [308].

6.3 Patients and Methods

6.3.1 Discovery cohort (PPMI)

The Parkinson's Progression Markers Initiative (PPMI) study is an effort to identify the biomarkers of PD progression [286]. We used the whole exome sequencing (WES) data available as part of this project. Detailed information about this initiative and the data can be found on the project website (`http://www.ppmi-info.org/`). Briefly, the variants were called following GATK [23] best practices by the authors of the original study. The data was obtained in the form of a Variant Call Format file (VCF).

6.3.2 Sample QC

Samples with >3 standard deviation (SD) from the QC metrics (number of alternate alleles, number of heterozygotes, Ti/Tv ratio, number of singletons and call rate) that were calculated by using PLINK/SEQ i-stats (`https://atgu.mgh.harvard.edu/plinkseq/`) were excluded from the analysis. For population stratification we selected the variants that were common between our dataset and hapmap version 3.30 [37], present in autosomal chromosomes, not in linkage equilibrium, call rate $> 80\%$, allele frequency $>5\%$ and Hardy-Weinberg equilibrium P-value $<$ 0.001 and used PLINK [39] multi-dimensional scaling (MDS) as described in the study [284] to identify outliers. Each sample that was >3 SD of the first and the second principal components was considered as ethnicity outlier and excluded from further analyses. By using the same set of variants as described above, relatedness check was performed up to second degree applying PLINK [39] and KING [40] algorithms. From the identified related sample pairs one sample was chosen randomly to be included in the final analyses.

6.3.3 Variant QC

Multi-allellic variants were decomposed by using variant-tests [45] and left normalized by bcftools [27]. The authors of the PPMI study used the variant quality score recalibration (VQSR) method as recommended by GATK best practices [23] to filter out low quality variants. Additionally, we used GATK hard filtering to select only high quality SNVs. Variant genotypes with a read depth (DP) <10 and genotype quality (GQ) < 20 [133] were converted to missing by using bcftools [27] and only variants with a call rate of >0.9 were kept for further analyses.

6.3.4 Variant annotation and filtering

As the current study is focused on U1 splice site variants we restricted our further analyses to the 5' consensus splice site positions, i.e., +3 to -6 from the exon/intron boundary. The exon-intron intervals were obtained from the UCSC table browser based on hg19 reference genome. Variants were annotated by using ANNOVAR [47] version 2016December05 using RefSeq gene annotations and the dbNSFP v3.0 [51] prediction scores. Only rare variant [309], as defined by variants with a minor allele frequency of $< 5\%$ in the European population of 1000 genomes [309], ExAC (NFE (non-finnish Europeans), release 0.3) [43], and the Exome variant server (http://evs.gs.washington.edu/EVS) were selected. In order to prioritize the 5' splice variants based on their deleteriousness, we used three different scores. The first score is generated by using the MaxEntScan method [310] which is based on the maximum entropy principle. The other two scores were ensemble scores (*dbscSNV_ADA* and *dbscSNV_RF*) generated from multiple splice site prediction tools [311] which are available as part of dbNSFP database [51].

6.3.5 Generation of MaxEntScan score

To prioritize variants using MaxEntScan method, for each SNV that lies in the consensus splice site region a wild type 9 mer (WT) was extracted from the reference genome (hg19). Then, the variant was introduced within the WT sequence by using the python module pyfaidx [312], hence creating a mutated consensus splice site (MUT) sequence for each variant. In the next steps, the scores were calculated for both WT and MUT sequences by using the scripts provided in the MaxEntScan website (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html). The relative percentage change (*maxentscan_change*) was calculated by using:

$$maxentscan\_change = \left(\frac{wild\_score - mut\_score}{wild\_score}\right) * 100$$

6.3.6 Benchmarking of MaxEntScan score

We were interested in the highly deleterious splice variants and, in line with our hypothesis, one recent study has shown that, 21 variants out 30 variants tested within BRCA1 genes were predicted by MaxEntScan method and were later confirmed by the functional validation. Out of the 21 variants that were predicted to be deleterious 18 of them had a *wild_score*>5 and a *maxentscan_change*>70. In order to benchmark our methods and determine reliable cut-offs, we used two datasets: 1) The professional version of Human Gene Mutation Database (HGMD) [44] version February 2017, and 2) gnomAD [43], which comprises of variant data from 123,136 exome sequences and 15,496 whole-genome sequences from individuals which were sequenced as part of various disease-specific and population genetic studies.

We only selected the variants annotated as high confidence and pathogenic ("DM" flag) in HGMD (HGMDpatho) variants. VCF files were generated for HGMD and gnomAD datasets for only those variants that were present within the U1 consensus splice regions annotated in a similar way as we did for the discovery cohort. Density plots based on various scores were generated for HGMDpatho variants and gnomAD splice variants (see Figure 6.1).

6.3.7 Splice site burden tests

The *wild_score* generated from the wild type 9mer by MaxEntScan is used to identify a true splice site. The higher the *wild_score* the higher the probability of being a true splice site [311]. We separated the variants into different classes: 1) All the deleterious splicing variants (DEL.splicing), 2) DEL.splicing variants in coding regions (DEL.exonic.splicing), 3) DEL.splicing variants within intronic regions (DEL.intronic.splicing), 4) DEL.exonic.splicing variants present in the genes that are expressed in brain (DEL.exonic.brain.splicing) [236], 5) DEL.exonic.splicing variants present in the genes that are not expressed in brain (DEL.exonic.nonbrain.splicing), and 6) rare synonymous variants as a negative class. We used a previously published list of brain expressed genes [236] to test if there is an increased burden in brain expressed genes (n= 14,177) compared to the non-brain expressed genes (n=6,428). Our hypothesis was that cases carry a higher number of DEL.splicing variants compared to the controls. For each variant class a VCF file was generated and the variant counts per sample was calculated by using bcftools [27] stats command.

We performed burden testing by constructing the generalized linear regression models using R version 3.4.1 while correcting for various confounding factors for each sample such as: 1) Sex 2) total number of variants remaining after final QC, 3) TiTv ratio of novel variants relative to the dbSNP version 138 [52], 4) TiTv ratio of variants present in dbSNP version 138, 5) heterozygous variants to homozygous variants ratio, 6) first ten principal components derived from the multi-dimensional scaling.

6.3.8 Replication cohort (PDGSC)

We used the WES data available as part of the ongoing Parkinson's Disease Genome Sequencing Consortium (PDGSC) project. The PDGSC dataset is an effort to integrate PD WES data generated from multiple studies across different sequencing centres. The variant calling was performed by the consortium using GATK best practices version 3.4. Similar to the discovery cohort, we obtained the data in the form of VCF file. Since the PPMI samples are also part of the PDGSC cohort, all samples overlapping between PPMI and PDGSC were excluded in beforehand from the PDGSC dataset. PPMI samples within PDGSC were identified based on their sample ids as well as using relatedness test (see above).

6.3.9 Sample QC

Sample QC was performed by the PDGSC consortium. Briefly, samples were excluded based on the following parameters: 1) <15x mean coverage 2) discordance between genetic and reported sex, 3) <85% call rate, 4) outliers for various parameters such as variant counts (all, non-reference genotypes, hets, singletons, mean minor allele rates), TiTv ratio, mean quality scores for non-reference variants and mean depth for non-reference variants, 5) heterozygosity outliers (-0.1< F<0.1) , 6) ancestry outliers >6 SD from means of CEU and TSI for PC1 and PC2 , 7) extract probands randomly from pairs related at >12.5% and 8) exclude samples<18 years of age or with missing age data.

6.3.10 Variant QC

Similar to the discovery cohort a VQSR filtering method was employed by the authors of original study. In addition, we used the same filtering procedure as described above for the discovery cohort with one difference in the threshold for call rate. As the data was generated at multiple centres by employing different sequencing protocols we might lose true positive variants if we would filter too stringently leading to loss of statistical power ultimately. Hence, we used a less stringent, although a standard threshold [64] of call rate >0.8 for a variant to be included in the analysis.

6.3.11 Variant detection and annotation

Variants were annotated and splice variants were scored using the same procedure as for the discovery cohort.

6.3.12 Burden testing

We employed the same procedure for burden testing by adjusting for all the covariates that were described above for discovery cohort. In order to further adjust for study wide differences, we used the total number of sites that were fully called within each sample as an additional covariate along with the other covariates. This approach allowed us to account for any exome-wide biases arising due to different sequencing protocols that were employed at different sequencing centres and other confounding factors arising from technical differences. The same can also be noted from the fact that there is no statistically significant difference between the number of synonymous variants (neutral variants) between cases and controls (Fig 6).

6.3.13 Multiple testing adjustment

The P-values from burden analysis of both discovery and replication cohorts were corrected for multiple testing by the function "p.adjust" (R version 3.4.1) using the false discovery rate (FDR) method for discovery and replication cohort separately.

6.4 Results

6.4.1 Benchmarking

In the current study, we were interested in variants having a high likelihood causing splice changes. Hence, in Figure 6.1 (A) and (B) based on HGMDpatho variants, it could be seen that there is a clear separation in the distribution of majority of variants at a *wild_score* of 5 (red-dashed line) and at a *maxentscan_change* of 70% (blue-dashed line). Whereas, a reversed distribution could be seen for the gnomAD variants Figure 6.1 (C) and (D). HGMDpatho variants (Figure 6.1 (A) and (B) showed *dbscnv_RF* and *dbscnv_ADA* scores of >0.9 GnomAD variants in Figure 6.1 (C) and (D) showed scores on the opposite part of the distribution. Based on the above inferences and the results based on previous study [313], we choose the following cut-offs for further processing: Deleterious splice site variants (DEL.splicing) were defined as SNVs with the following criteria: *wild_score>5* and *maxentscan_change>70* and *dbscSNV_ADA score>0.9* and *dbscSNV_RF score>0.9*. If the ensemble scores were not available for any particular variant only MaxEntScan method (*wild_score>5* and *maxentscan_change>70*) was used.

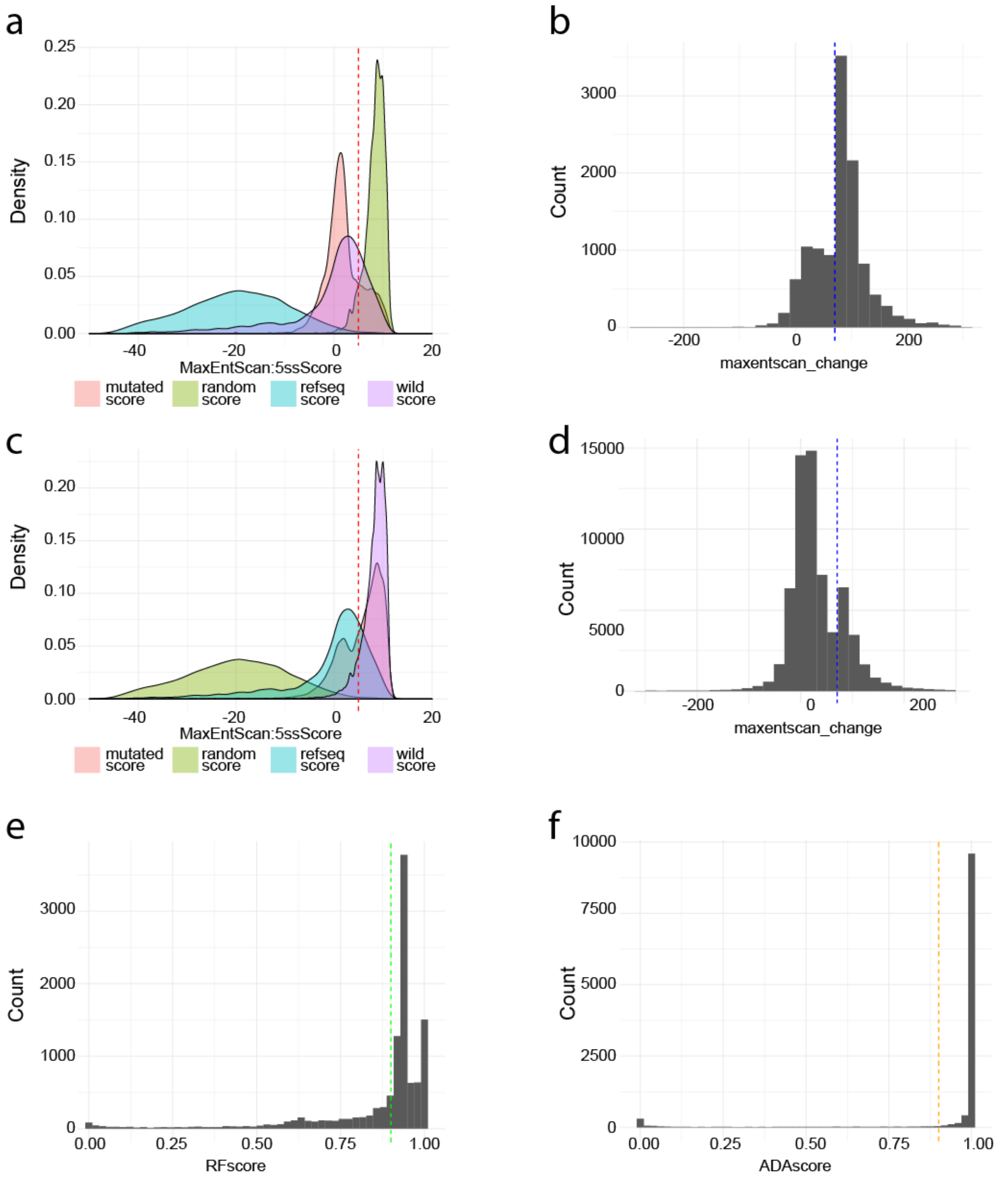


Figure 6.1: Determination of cut-offs for wild_score and maxentscan_change. Dashed lines in each plot indicate the cut-offs that were used to define a variant as deleterious (DEL.splicing). (a) Distribution of wild_score and mutated_score of HGMDpatho variants, (b) distribution of maxentscan_change of HGMDpatho variants, (c) distribution of wild_score, and mutated_score of gnomAD variants, (d) distribution of maxentscan_change of gnomAD variants, (e) distribution of dbscnv_RF score of HGMDpatho variants, and (f) distribution of dbscnv_ADA score of HGMDpatho variants. mutated_score = maxentscan score of mutated 9mers and wild_score = maxentscan score of all wild type 9mers

6.4.2 Burden analysis

6.4.3 Discovery cohort (PPMI)

After filtering based on ethnicity, cryptic relatedness, quality parameters and without missing information for sex, the final dataset comprised of 372 PD and 161 control samples. A total of 128 DEL.splicing variants (Supplemental Table 2) were included in the final analysis. We observed a genome-wide burden in cases compared to the controls Figure 6.2 (P-value=0.012, OR=1.39, CI=1.08 - 1.82, P-value FDR corrected=0.074). The signal is coming mainly from the DEL.exonic.splicing variants (P-value=0.028, OR =1.37, CI=1.04-1.84, P-value FDR corrected=0.08) rather than from the DEL.intronic.splicing variants (P-value=0.25, OR=1.50, CI=0.76- 3.21, P-value FDR corrected=0.25). For the DEL.exonic.splicing variants, the majority of the burden is caused by the DEL.exonic.brain.splicing variants (P-value=0.06, OR=1.47, CI=0.99-2.24, P-value FDR corrected=0.12), compared to the DEL.exonic.nonbrain.splicing variants (P-value=0.24, OR=1.24, CI=0.86-1.85, P-value FDR corrected=0.25).

6.4.4 Replication cohort (PDGSC)

The final dataset comprised of 2,710 cases and 5,713 controls. A total of 2,328 DEL.splicing variants were included in the final analysis. Similar to the discovery cohort we observed an overall burden of DEL.splicing variants in cases compared to the controls (P-value=0.007, OR=1.04, CI=1.01-1.08, P-value FDR corrected=0.014) here even after multiple testing correction. The majority of burden after FDR correction is due to the DEL.exonic.splicing variants (P-value=0.003, OR=1.11, CI=1.03-1.19, P-value FDR corrected=0.011) rather than the DEL.intronic.splicing variants (P-value=0.09, OR=1.03, CI=0.99-1.08, P-value FDR corrected=0.138). In the DEL.exonic.splicing variants, the burden after FDR correction is coming from the DEL.exonic.brain.splicing variants (P-value=5.774e-05, OR=1.20, CI=1.10-1.32, P-value FDR corrected=0.0003) compared to the DEL.exonic.nonbrain.splicing variants (P-value=0.69, OR=0.97, CI=0.86-1.09, P-value FDR corrected=0.69).

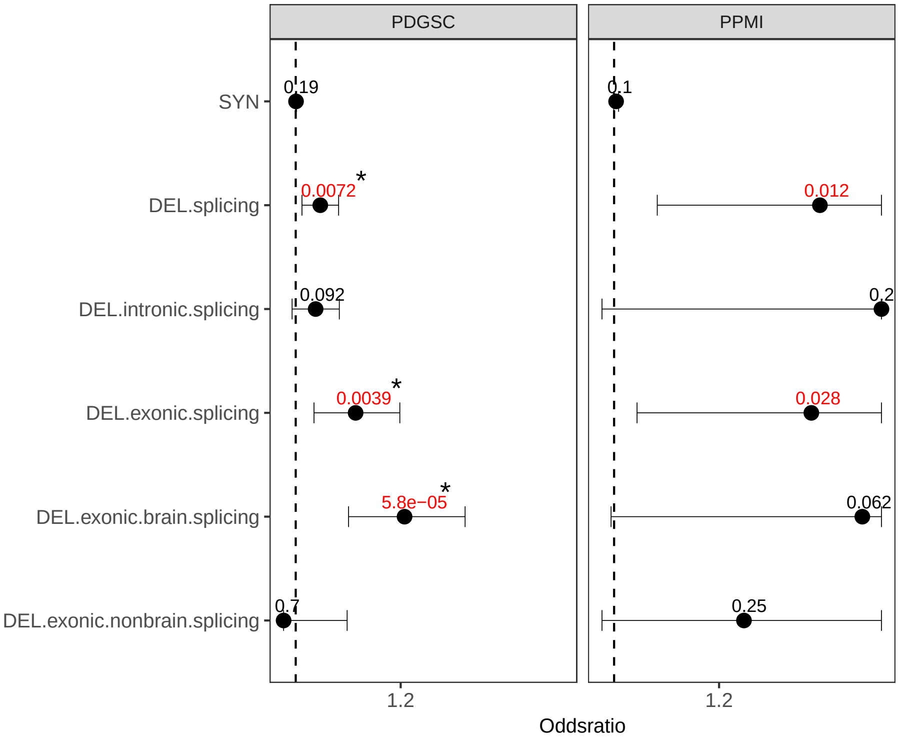


Figure 6.2: A forest plot representing the burden analysis across different variant classes. Each dot represents the odds ratio, the value on top of each dot represents the corresponding uncorrected P-value. Values in red indicate FDR corrected P-values<0.05. (left) Results of the PPMI discovery cohort analysis, the upper limit of confidence interval in the plot is restricted to the maximum of odds ratios, (right) results for the PDGSC replication cohort. SYN = rare and low frequency synonymous variants, DEL.splicing = deleterious splicing variants, DEL.intronic.splicing = deleterious variants in intronic regions, DEL.exonic.splicing = deleterious variants in coding regions, DEL.exonic.brain.splicing = DEL.exonic.splicing variants present in genes expressed in the brain, DEL.exonic.nonbrain.splicing = DEL.exonic.splicing variants present in genes that are not expressed in the brain.

6.5 Discussion

Herein, we describe a novel mechanistic concept for the pathogenesis of PD related to U1 splice-site mutations. Our findings indicate that the pathogenic relevance of exonic splicing mutations was underestimated in PD. These results are in line with a recent study showing that approximately 10% of pathogenic missense variants predicted to alter protein coding essentially disrupt splicing [308]. Although defective pre-mRNA processing is known to represent a common cause of human diseases, with approximately 15% of all mutations causing aberrant splicing [314], for PD pathogenesis, the dysregulation of splicing as an alternative mechanism contributing to the neurodegenerative process was not systematically addressed [315]. Our analysis of exonic mutations affecting U1-mediated splicing using a large dataset for sporadic PD, including WES results from the PPMI study and from the PDGSC cohort consistently revealed a higher burden of rare and low frequency exonic variants affecting U1 snRNA binding sites among sporadic PD patients [286]. Together, these data indicate an enrichment of disease-associated variants in the exon-intron boundary of brain expressed genes in PD and underscore the therapeutic potential of compounds acting on pathological splicing also in sporadic PD cases.

Our study illustrates the promise for treatment approaches in precision medicine in PD that focus on genetic and molecular stratification. To account for the increasingly recognized heterogeneity in PD and other neurodegenerative disorders, new strategies need to be developed for the stratification of patients along shared pathogenic mechanisms. By employing a text mining approach one can identify the candidate drugs based on the abnormal splicing which may translate into basket studies referring to patients sharing the same underlying mechanism, as already shown for precision medicine approaches in cancer, and will allow for clinical trials in patients across groups that share certain molecular signatures [316].

CHAPTER 7

FAMILIAL-PD

7.1 Abstract

Until today, several variants associated to PD were discovered via large-scale GWAS. Familial studies conducted via NGS approaches provide an advantage to identify the true cause of the disease and hence are more powerful. However, a comprehensive study of rare variants to understand the etiology of PD by large-scale genome sequencing was still missing. Hence, we conducted a Whole genome sequencing (WGS) based study of familial-PD by analyzing two independent familial-PD cohorts and a replication case control cohort. By employing WGS and prioritizing variants based on various functional annotations, we identified several likely candidate variants that are rare, predicted to be deleterious and co-segregating with PD. Some, of them were found in the genes already associated to PD, but majority of them are novel with regards to their association to PD. Hence, the list of variants generated in this study could serve as reference to perform functional validations in the future.

7.2 Introduction

Parkinson's disease (PD) is one of the common neurological disorders in the elderly patients. In majority of the individuals it is late onset (>58), however there are also some early onset forms of PD. Till date >70 loci have been shown to be associated to PD [109] via large-scale meta-analysis and family based studies but the genetic architecture of PD still remains complicated. Genome-wide association studies (GWAS) have been successful in deciphering novel regions which increase the genetic risk of PD. The main drawback of GWAS is that, identification of causal variants is highly unlikely and often, it is necessary to have large cohorts of individuals to increase the statistical power to discover rare variants associated to the disease. In contrast, whole genome sequencing (WGS) or whole exome sequencing (WES) provides an opportunity to locate rare variants in genes with medium to large effects, especially by sequencing the individuals in a family and thereby reducing the number of samples that need to be sequenced. Upto 20% of PD cases are believed to have a familial origin and the genes identified via classic linkage analysis in families include *PINK1, PARK2, PARK7, LRRK2,* and *SNCA* [104, 122, 123, 244, 317, 318]. WGS/WES studies have been highly successful in identifying several novel variants in PD. One such example is *VPS35* [319] in which two novel variants p.Asp620Asn and p.Pro316Ser were found via WES. A recent study [320] has identified likely deleterious variants in two genes *TNK2* and *TNR* that are segregating with PD and present in multiple families. Similarly, a WES based study has identified rare variants in *PLXNA4* [321] in PD although their role in PD remain inconclusive. Hence, family studies are of paramount importance in identifying the causal variants of PD. Further, studying the families could provide a chance to unravel the complexity of the disease by showing how multiple variants may act together to influence the disease risk. For example, in a recent study [111], the rs2421947 variant in *DNM3* was found to reduce the age at onset (aao) of PD by ~12.5 years in the carriers of *LRRK2* p.G2019S variant.

In the current study, we analyzed the WGS data available from two familial-PD studies without any known genetic cause in the families. The first is a two-stage study in which the discovery cohort comprised of 16 families consisting a minimum of two affected siblings with PD and the replication study was conducted with the WES data from 369 PD cases and 159 ethnically matched controls provided by the PPMI consortium [286]. In our second familial-PD study, WGS of 90 samples from 36 families with both autosomal dominant and recessive inheritance patterns was performed. We prioritized the variants by a combination of mode of inheritance analysis, burden analysis, machine learning and pathway enrichment based approaches and identified potential candidate genes.

7.3 Identification of novel genes involved in nervous system development by whole genome sequencing in PD

7.3.1 Patients and methods

Data generation

Discovery cohort: The discovery cohort comprised of 44 samples of which 16 samples were female and 28 were male. The mean aao was 57.67 years and all the families were of German origin. WGS of affected samples was performed by the Complete Genomics, Inc (CGI). Whereas, the control samples were sequenced by the illumina Hiseq. The complete genomics data was processed as described in the study [322]. In brief, WGS was performed by Complete Genomics using a proprietary paired-end, nanoarray-based sequencing-by-ligation technology [323, 324]. The pedigrees of families selected for the WGS are given in the Figure 7.0

Replication cohort: The replication cohort comprised of WES data of 369 cases and 159 controls generated as part of PPMI [286] consortium. WES was performed on whole-blood extracted DNA samples that were collected according to the PPMI Research Biomarkers Laboratory Manual. Illumina Nextera Rapid Capture Expanded Exome Kit which targets 201,121 Exons, UTRs and miRNA and covers 95.3% of Refseq exome was used to perform the WES. >340,000 probes were constructed against the human NCBI37/hg19 reference genome with a targeted genomic footprint of 62Mb. Exome- enriched libraries (multiplexed sets of 12 samples) were sequenced on the Illumina HiSeq 2500 sequencing platform using 2 x 100 bp paired-end read cycles.

Variant detection and quality control (QC)

Discovery cohort: After QC, DNA samples were sent to Complete Genomics for sequencing. Next steps of QC, mapping and variant calling for the sequencing data were performed by Complete Genomics as part of their sequencing service using the Standard Sequencing Service pipeline version 2.0 (`http://www.completegenomics.com/documents/Standard_Sequencing_Service_Getting_Started_Guide_2.4-2.5.pdf`). Sequencing reads were mapped against NCBI Build 37. For the samples sequenced by illumina, genomic variant call format (gVCF, `https://support.illumina.com/help/BaseSpace_App_WGS_BWA_help/Content/Vault/Informatics/Sequencing_Analysis/BS/swSEQ_mBS_gVCF.html`) files for each sample were provided by the vendor. In addition to the family controls, we selected an additional 17 controls from the study [325]. These controls are super centenarians and these are the individuals who survived beyond 110 years without any known neurological disease. As these samples were healthy with an age beyond 100 years, we assumed that if a variant is present in these samples it is not likely to be disease causing. Hence, we excluded all the variants that were concordant between our study and super centenarians. Furthermore, we excluded the variants present in the low confidence regions of the human genome such as repeat regions etc., according to the study [65].

Replication cohort: Variant calling was performed by the PPMI consortium and provided a variant call format file (VCF) [33]. In brief, the multi-sample VCF was generated by following the GATK best practices [23] which applies the standard bwa-picard-GATK haplotype caller pipeline. In order to select only the high quality, unrelated and the samples whose calculated gender was matched to the reported sex we employed the QC procedure as employed in [65]. Population stratification analysis to select only the European samples was performed via eigenstrat [42]. In order to determine the ethnicity outliers from the eigenstrat analysis, a sigma value of 3 was applied as a cut-off (which excludes all the samples with a SD of >3 based on the first 10 principal components). Additionally, we excluded the samples >3SD based on the first and second principal components from the eigenstrat analysis. The downloaded PPMI vcf file had already been filtered for high quality variants by the authors of original study according to the variant quality score recalibration (VQSR) approach as part of GATK best practices. In order to be more stringent, we applied additional filters as described in the study [65]. Finally, only variants with a call rate of > 0.9 were kept for further analyses.

Mode of inheritance (MOI) analysis and detection of shared genomic regions

As an intitial step in the discovery cohort, we first combined all the variants from all affected samples into the union of variants for each set using the CGAtools listvariant command and CG var-files as input. CGAtools (CG Analysis Tools) version 1.5 was used as provided by CG and available under (`http://cgatools.sourceforge.net`). We used the CGAtools testvariant command to test each genome for the presence of each variant. Only variants that were called in all genomes within one family were selected for further analysis. We first removed variants that were not called in at least one genome as high-quality calls (VQHIGH) by CG.

The WGS data of parents was not available in any pedigree, hence we defined the segregating variants as those that were present in all the PD samples per family and not present in any of the control samples in the cohort. Due to the fact that, the control samples in our study were sequenced using a different technology (Illumina), for every variant found to be present in all the cases per family, we checked for its presence in the control samples and excluded it if present in any of the sample. Additionally, any variant present in the super centenarians was also excluded from further analysis. In our study, variants that were shared between two individuals within the same pedigree should be located within a region that shares one or two identical haplotypes between the two genomes and is inherited from the same ancestor, a concept, which is also called identity by descent (IBD). ISCA version 0.1.9 [323, 326, 327] was employed to search for identical haplotype blocks between all pairs of genomes within each of two sets. Afterwards we built the intersection of all regions between all of pairs of genomes to determine the regions that were shared by all genomes from one set. For each set, we filtered out all variants following autosomal dominant inheritance outside the shared regions and excluded the variants outside the IBD intervals thereby reducing our variants of interest.

Annotation

The remaining high quality segregating variants in the discovery and all the high quality variants of the replication cohort were functionally annotated by ANNOVAR [47] version 2016 June17 with RefSeq and Ensembl gene annotations. The dbNSFP v3.0 [51] prediction and conservation scores as well as genome-wide CADD [58] scores were also applied. Exonic and splice site variants were selected according to RefSeq and Ensembl annotations. Rare variants were defined as variants with minor allele frequency (MAF) < 0.01 and < 0.05 for autosomal dominant and autosomal recessive hypothesis respectively. In order to determine the MAF, European population of public databases such as 1000 genomes [34], ExAC (release 0.3) [43], and the Exome variant server (`http://evs.gs.washington.edu/EVS`) were used. From the rare exonic and splicing variants variants, we selected the variants for downstream analysis based on the following criteria: 1) Loss of function (LoF) variants were defined as stop gain, stop loss, splice site variants and all insertions/deletions or 2) Nonsynonymous variants with a CADD phred score >15 according to the study [65].

Phenolyzer

The variants prioritized using above approaches were collapsed into genes and for each family the list of candidate genes was generated. The generated list of genes per family was given as an input to Phenolyzer [328] for ranking them based on their relevance to PD. Phenolyzer works in three steps, first it converts the phenotype of interest which is “parkinson disease” in our case into a group of professional disease names based on the Human Phenotype Ontology (HPO), a resource developed to define a standard ontology for human phenotypes [329]. Second, the entire list of genes having an association to all the diseases will be generated; Third, it finds more genes by generating a database of gene-gene relation and as a last step it provides a score for each gene by integrating all the information together. The score provided by Phenolyzer can be used to rank the given list of genes. From the phenolyzer output, we selected the top 5 genes per family for downstream analysis.

Statistical analysis of replication data-set

In the PPMI data-set, in order to identify the genes carrying higher number of deleterious variants in cases versus the controls we collapsed all the selected variants as described in the Section 7.3.1 into genes. Then, for each gene we calculated the odds ratios (OR) based on the samples carrying atleast one variant in that gene. The analysis was performed using R version 3.4. An OR >1 for a gene means more number of cases carry a variant in that gene compared to the controls. Hence, we only selected those genes with an OR > 1 for further analysis.

Ingenuity Pathway Analyser

An intersection of genes prioritized in the discovery cohort using Phenolyzer and the genes with an OR > 1 in the replication data-set were selected for the Ingenuity Pathway and network analysis (IPA®) [237]. IPA is a commercial tool comprising of manually curated interactions and hence is more reliable. The IPA relies on their proprietary database called Ingenuity Knowledge Base which is an exhaustively curated resource composing of high quality knowledge on functional annotations and biological interactions.

7.3.2 Results

Data generation

Discovery cohort: The pedigrees of the families sequenced in this study are shown in Figure 7.0. We identified a total of 19,165,307 variants that were present in all the affected samples per family. When we excluded all the variants present in the super centenarians, we reduced the list to 6,003,751 variants. Finally, by excluding the variants present in any of the control samples and not present in the regions of low confidence in the human genome we were left with a total of 6,21,221 variants.

Replication cohort: The final data-set after sample QC comprised of 369 PD cases and 159 controls. They are both ethnically and sex matched. After the variant QC, 73,904 variants have remained in the analysis, of which 68,036 were SNVs and 5,868 were indels.

Segregating variants

After the functional annotation, we conducted the analysis per family. In total, 9 families followed autosomal dominant (AD), one family followed autosomal recessive and 6 families followed either AR/AD hypothesis. After the QC and functional annotation an average of 57 SNVs per family and 9 indels per family were collapsed to the genes and included in the phenolyzer analysis. The number of SNVs and indels that were functionally prioritized and co-segregating with PD were different between each family as shown in the Table 7.1. Based on the output from Phenolyzer, top 5 genes per family were selected for the subsequent analysis in the replication data-set. Some of the families did not have five segregating genes. In total, 76 genes were assessed for their ORs in the PPMI data-set. Out of the 76 genes, 71 genes carried at least one variant fulfilling our criteria (rare, nonsynoymous variants with a CADD phred > 15 or LoF) in the replication PPMI data-set. In the PPMI data-set 33 unique genes had an OR > 1 as shown in the Table 7.2 and were selected for Ingenuity pathway analysis.

Enrichment analysis

33 genes were selected after restricting the analysis to rare deleterious variants and prioritization using Phenolyzer. Enrichment analysis of the 33 genes using IPA revealed interesting pathways and functions such as "Carbohydrate metabolism", "Nervous System Development and Function" and "Tissue morphology". Further, we identified that the top enriched network has a function related to the nervous system development as shown in Figure 7.1 and in total 14 of the 33 genes as seen in the Table 7.3 were present in the network from the input genes.

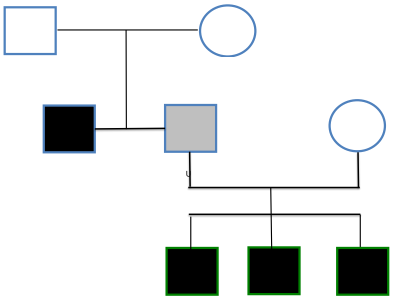
Variants in novel genes

By employing a two-stage approach, we identified several novel candidate gene/variants which are co-segregating with PD, highly conserved, predicted to be damaging and occurring at very low allele frequencies. Due to the different effect sizes of the variants, we might not be able to replicate the association of all segregating genes from our discovery cohort. The main reason is that our discovery cohort comprised of families with PD, whereas in the replication cohort, majority of the patients do not have a family history of PD. Hence, in addition to including genes with an OR $>$ 1 and prioritized based on IPA, we also included the variants present in the genes that were selected to be involved in various neurological disorders based on literature search. The main rationale behind this approach is to not miss any potential PD associated variant. The entire list of variants selected by both the approaches is given in the Table 7.3.

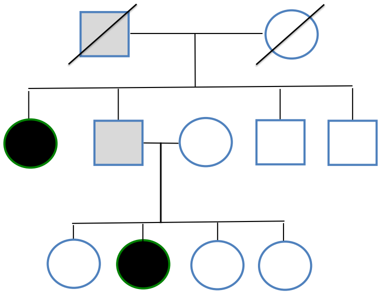
It could be seen that *NOS1* and *IRS1* were present in the both the categories of prioritized genes Tbale 7.3. Further, some of the families have more than one proposed candidate genes showing the complexity of PD and highlighting the difficulty in identifying the causal variants in small families. Few genes were prioritized in more than one family. However, no variant was identified in more than one family indicating the heterogeneity of PD.

Majority of the prioritized genes were found to be mutated in a single family, suggesting a low incidence of novel mutations in PD Table 7.3. *USP25* is one of the genes carrying variants in two families, it is present in our PD candidate gene list but not prioritized by IPA. It carries a p.P784L mutation in Family FN9984 and p.V846I mutation in Family pd_M009_M023. *USP25* has been previously identified to be in a suggestive locus in a GWAS comprising of 3,426 cases and 29,624 controls [330]. However, it failed to reach a genome wide significance in the replication study. Similarly, another gene *ITSN2* carries two variants p.I995T and p.A1515V in two different families FN17908 and M013 respectively. *ITSN2* has been shown to be associated to Schizophrenia according to Disgenet database [238]. However, in the family FN17908 there is another variant p.T335I prioritized by IPA and the gene harboring that variant is *SLC2A1*. *GRIN2A* has been associated to various forms of epilepsy[65, 147, 148] and in the current study we found a variant p.N1076K in *GRIN2A* of family FN10364. The same variant has been found in another familial-PD study [320], however the variant could not make it to their final list of candidate variants. Even in our study *GRIN2A* did not have an OR $>$ 1 in the PPMI data-set, but it could also be due to the fact that the penetrance of this variant is low and sample size of our replication dataset is small. Hence, larger sample sizes are needed to replicate such associations.

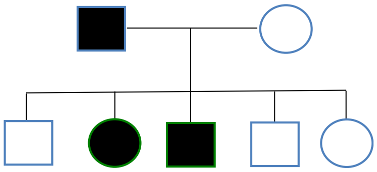
FN17908



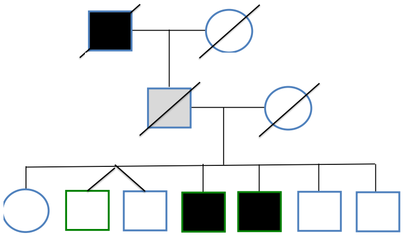
M009/M023



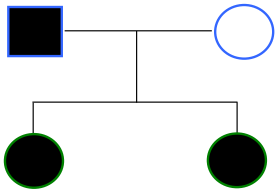
FN10364



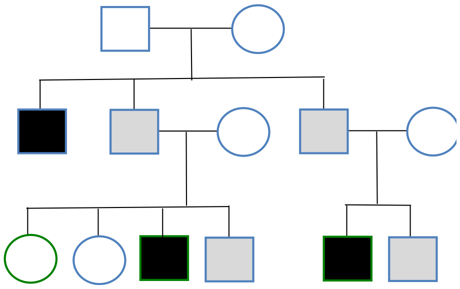
M013



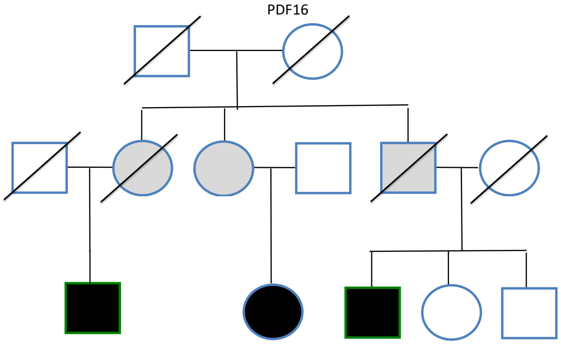
M040



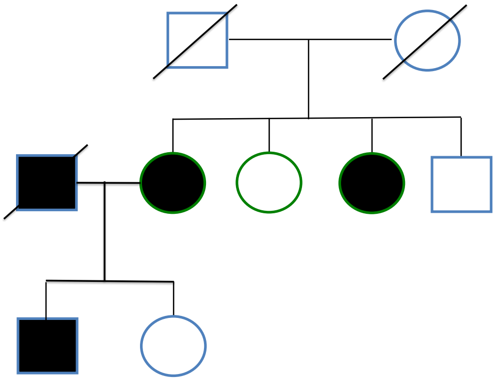
M014/M015



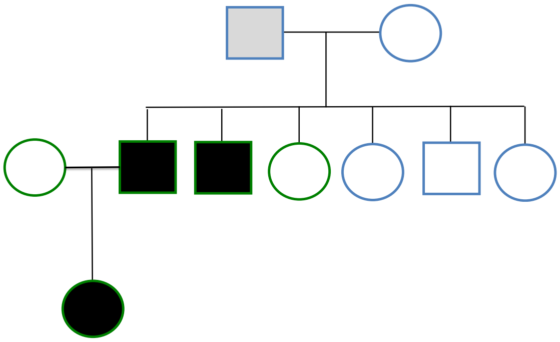
PDF16



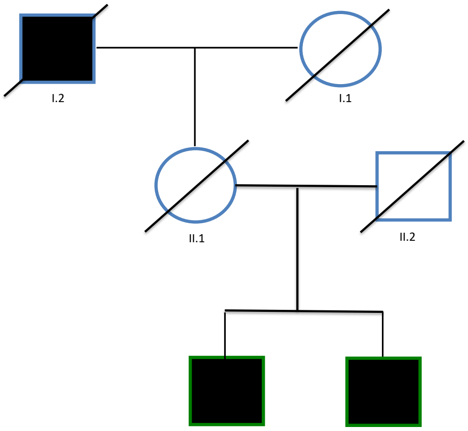
DE43



T10381



FN9984



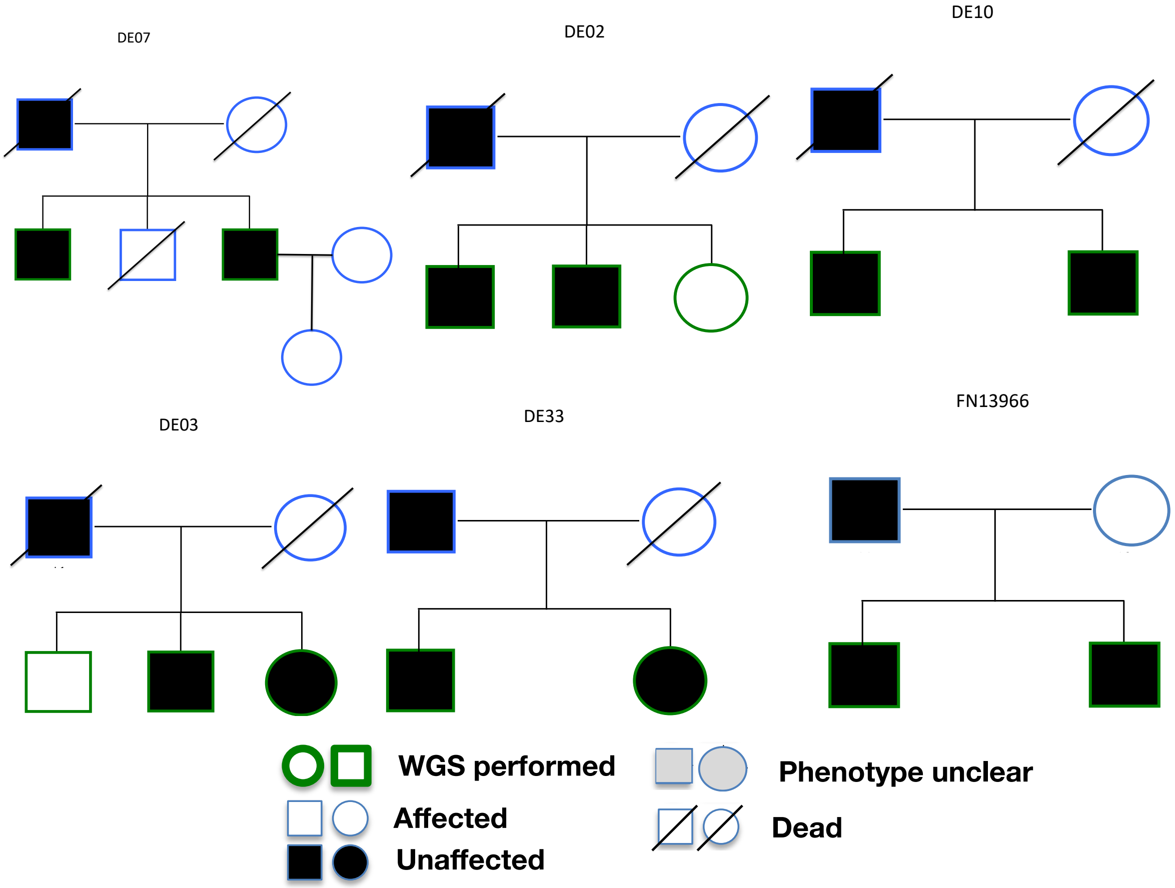


Figure 7.0: Pedigrees analyzed in the study.

Variant prioritization

Family_name	MOI	AD_SNVs	AR_SNVs	AD_indels	AR_indels
DE02	AD	40	2	8	0
DE03	AD	38	1	6	0
DE07	AD	123	3	23	0
DE10	AD	137	4	14	0
DE33	AD	4	0	0	0
DE43	AR	40	1	11	1
FN10364	AD	59	2	14	0
FN13966	AD	96	1	14	0
FN17908	AR/AD	50	1	8	6
FN9984	AD	105	5	8	0
M009_M023	AR/AD	43	1	9	0
M013	AR/AD	60	1	10	0
M014_M015	AR/AD	17	0	7	0
M040	AD	53	1	7	0
F16	AR/AD	29	3	5	0
T10381	AR/AD	1	0	0	0

Table 7.1: Possible MOI per family and the number of different kinds of prioritized variants segregating per family. AD = Autosomal dominant, AR = Autosomal recessive, AD/AR = Both types of inheritance is possible, AD_SNVs = SNVs following AD inheritance, AR_SNVs = SNVs following AR inheritance, AD_indels = Indels following AD inheritance and AR_indels = Indels following AR inheritance

Gene	LowerCI	UpperCI	OR
AKT2	0.103	45.45	2.17
AP1G1	0.052	32.04	1.29
APC	0.521	2.74	1.19
DDIT3	0.210	73.38	3.92
EEF1D	0.380	2.67	1.00
FGF6	0.542	159.97	9.31
GORASP1	0.832	7.24	2.45
GPI	0.264	87.57	4.81
HSPA9	0.103	45.45	2.17
IRS1	0.853	4.60	1.98
ITPR3	0.578	2.76	1.26
LTA4H	0.419	9.18	1.96

MAPKAP1	0.434	28.23	3.50
MYC	0.251	18.72	2.17
N6AMT1	0.052	32.04	1.29
NFRKB	0.451	4.38	1.40
NOS1	0.493	2.62	1.139
NOTCH1	0.573	3.27	1.37
NUP205	0.823	9.67	2.82
PSMB11	0.963	8.22	2.81
PTRH2	0.052	32.04	1.29
SEC16A	0.985	3.17	1.77
SLC2A1	0.133	12.54	1.29
SNX1	0.133	12.54	1.29
SP1	0.156	59.32	3.04
SPTAN1	0.411	2.83	1.08
SPTBN1	0.619	5.65	1.87
SPTBN2	0.545	3.14	1.31
SPTBN5	1.004	2.81	1.68
SRMS	0.744	14.57	3.29
TSC2	0.526	2.55	1.15
UBA7	0.613	4.55	1.67
USP42	0.496	31.44	3.95

Table 7.2: Genes harboring variants that were predicted to be deleterious and co-segregating with the PD. Only genes carrying higher number of variants in cases versus controls (OR > 1) are shown here. OR = odds ratio, LowerCI = Lower confidence interval and UpperCI = Upper confidence interval.

Family__name	**Chrom**	**Position**	**aa.change**	**Function**	**Gene**	**Category**
DE02	2	54858253	p.E1010D	exonic	*SPTBN1*	IPA
DE03	1	111861782	p.D158fs	exonic	*CHIA*	PD genes
DE03	17	76823418	p.H200N	exonic	*USP36*	PD genes
DE07	4	188924601	p.P214A	exonic	*ZFP42*	PD genes
DE07	11	66488550	.	intronic,splicing	*SPTBN2*	IPA
DE07	17	17700037	p.1259__1260del	exonic	*RAI1*	PD genes
DE10	5	112179359	p.A2672T	exonic	*APC*	IPA
DE10	9	18776904	p.T893P	exonic	*ADAMTSL1*	PD genes
DE10	15	49031277	p.I1375fs	exonic	*CEP152*	PD genes
DE10	15	62238002	p.A1644V	exonic	*VPS13C*	PD genes
DE10	19	40742161	.	intronic,splicing	*AKT2*	IPA
DE43	2	228884530	p.S347F	exonic	*SPHKAP*	PD genes
DE43	8	139838989	p.A294D	exonic	*COL22A1*	PD genes
DE43	15	48726830	p.E2193K	exonic	*FBN1*	PD genes
DE43	15	89872343	.	splicing,intronic	*POLG*	PD genes
FN10364	9	131380339	p.A1826V	exonic	*SPTAN1*	IPA
FN10364	12	4543445	p.R188Q	exonic	*FGF6*	IPA
FN10364	16	9858173	p.N1076K	exonic	*GRIN2A*	PD genes
FN13966	8	11710985	.	splicing,UTR5	*CTSB*	PD genes
FN13966	9	121971061	p.R361G	exonic	*BRINP1*	PD genes
FN13966	12	57910666	p.E169K	exonic	*DDIT3*	IPA
FN13966	12	53777215	p.447__448del	exonic	*SP1*	IPA
FN13966	21	43274890	p.T145I	exonic	*PRDM15*	PD genes
FN17908	1	43394673	p.T335I	exonic	*SLC2A1*	IPA
FN17908	2	24475269	p.I995T	exonic	*ITSN2*	PD genes
FN9984	6	75855921	p.G322V	exonic	*COL12A1*	PD genes

FN9984	7	146818170	p.G285A	exonic	*CNTNAP2*	PD genes
FN9984	9	128321981	p.S68N	exonic	*MAPKAP1*	IPA
FN9984	9	35236570	p.R86H	exonic	*UNC13B*	PD genes
FN9984	16	56692993	p.S35P	exonic	*MT1F*	PD genes
FN9984	20	57019129	.	splicing,intronic	*VAPB*	PD genes
FN9984	21	17222109	p.P784L	exonic	*USP25*	PD genes
M009_M023	2	227663330	p.A42G	exonic	*IRS1*	IPA, PD genes
M009_M023	8	119391814	p.Q150X	exonic	*SAMD12*	PD genes
M009_M023	21	17238604	p.V846I	exonic	*USP25*	PD genes
M013	2	24431159	p.A1515V	exonic	*ITSN2*	PD genes
M013	9	139405111	p.R912W	exonic	*NOTCH1*	IPA
M013	14	23511778	p.R115Q	exonic	*PSMB11*	IPA
M014_M015	17	17718592	p.R812Q	exonic	*SREBF1*	PD genes
M040	12	117710246	p.G259S	exonic	*NOS1*	IPA, PD genes
F16	17	67282371	p.Y708C	exonic	*ABCA5*	PD genes

Table 7.3: A list of genes and variants prioritized either by using the IPA or present in the candidate PD genes list. aa.change = predicted amino change based on RefSeq by ANNOVAR, Function = function of the variant predicted by ANNOVAR based on RefSeq, Gene = HGNC symbol of the gene harboring the variant, Category = method by which the variant was prioritized.

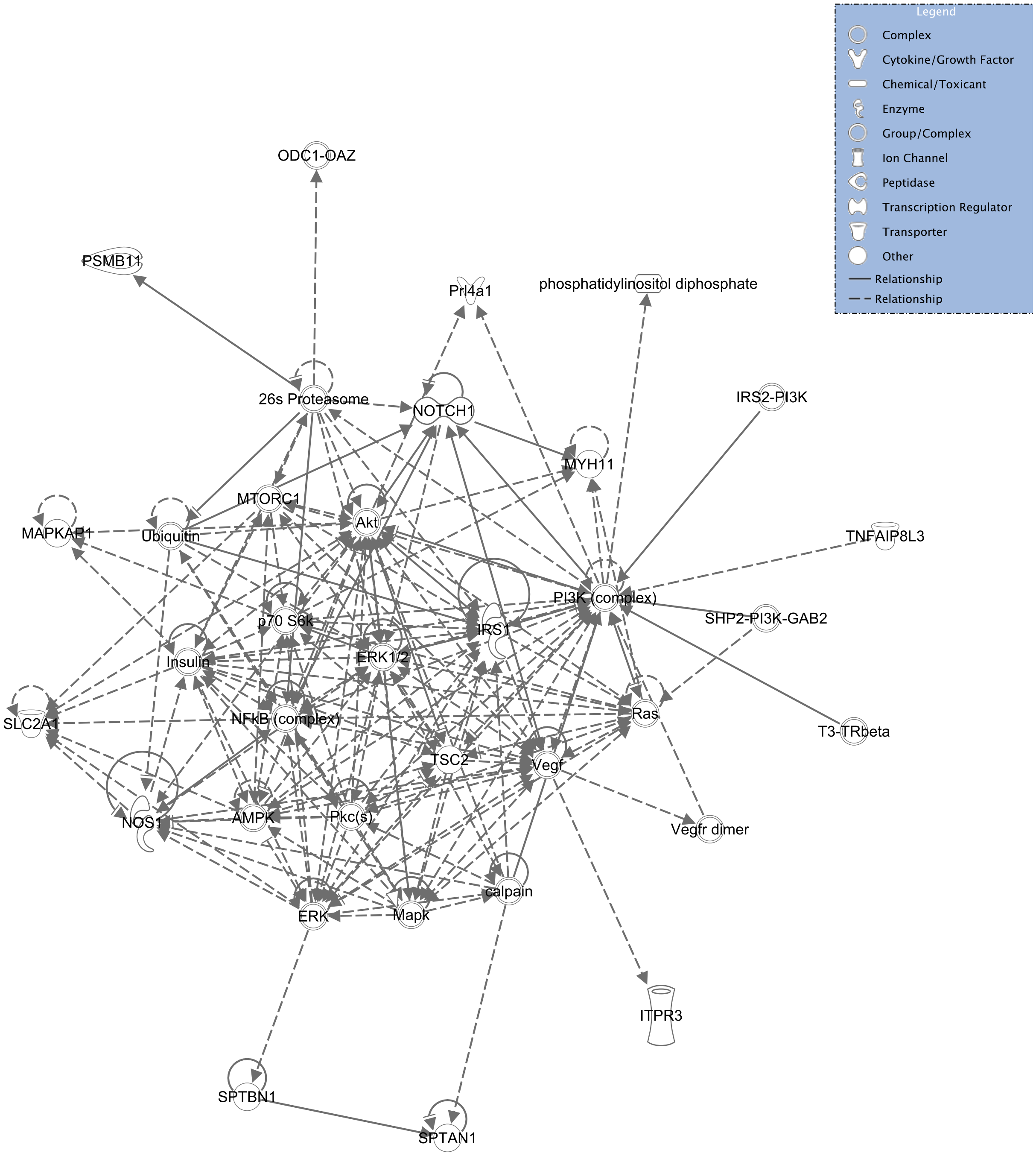


Figure 7.1: Top scoring network from IPA network analysis. It is enriched with functions related to nervous system development and function.

7.3.3 Discussion

In this study, we adopted a two-step strategy in order to identify and replicate the genes that might harbor rare deleterious variants contributing to PD. All the 14 variants identified in this study via prioritization by IPA were from a different gene and present in only one family. This distribution of variants in different genes further supports the complexity of PD where there is a presence of variants with very small effect size and emphasizes the fact that we need much larger sample sizes in order to detect meaningful association of variants to diseases.

However, we mitigated this lack of larger sample size by using a two-stage approach and could show that the genes prioritized in this study were both segregating with the disease and have an increased burden in cases compared to controls in an independent case-control study. Together, all the 14 genes *SPTBN1, SPTBN2, APC, AKT2, SPTAN1, FGF6, DDIT3, SP1 ,SLC2A1 ,MAPKAP1, IRS1, NOTCH1, PSMB11* and *NOS1* are involved in the "nervous system development and function" indicating that the results are not obtained by chance. Additionally, it could also be seen that one of the top enriched pathway is "carbohydrate metabolism". Dysregulation of glucose metabolism was found to be an early sign of sporadic PD [331]. A previous study [332] has shown that there is an association between alpha-synuclein and *IRS1* expression suggesting a novel mechanism for alpha-synuclein associated pathogenesis. Variants in *NOS1* has been linked to PD [333, 334] and suggested that mutations in *NOS1* could be a potential risk factor for PD and other pyschiatric disorders [335].

Mutations in *SLC2A1* have been previously associated to Glucose Transporter Type 1 Deficiency Syndrome and epilepsy [336] and the symptoms of GLUT1 deficiency syndrome are similar to PD including parkinsonism [337]. Hence it is possible that *SLC2A1* is the candidate gene in the family FN17908 rather than *ITSN2*. But, we need to perform functional validation in order to solidify these results. *SPTAN1* has been known to be involved in the disease pathology of early infantile epileptic encephalopathy [249, 338, 339] and could be a potential link between PD and epilepsy. In a recent study, it has been shown that PD gene *ATP13A2* [340] regulates *SYT11* via ubiquitination of *TSC2* thereby causing an impairment of autophagy-lysosomal pathway, hence the variants in *TSC2* could be of functional importance. In total, we identified likely disease causing variants in ~87% (14/16) of the families. Our intention in this study was to provide the list of potential candidate genes which could help accelerate the future research aiming for genetic diagnosis of PD. Overall, we show that a two stage approach can be employed in order to identify the plausible candidates especially in case of variants occurring at very low allele frequencies. However, this study should be considered preliminary and further replication in families affected with PD is warranted.

7.4 Integrated analysis of WGS data reveals potential candidate genes

7.4.1 Patients and methods

Data generation and processing

A total of 180 samples, were sequenced via WGS at macrogen (`www.macrogen.com/eng`). Briefly, samples were prepared according to the Illumina TruSeq Nano DNA library preparation guide and libraries were sequenced using Illumina HiSeqX sequencer (`www.illumina.com`). However, only the results from 36 families comprising of 90 samples were presented in this work due to the collaborative agreement. The results from the families excluded here were already published elsewhere [341, 342]. 50 females and 40 male samples were included in the final analysis. The mean aao was 62.22 years. Our study comprised of ethnically different families (Italy (4), Dutch (13), Portugal (1), Spain (6), Tunisia (3) and Turkey (9)). Out of the 36 families, 23 were tested for AD inheritance and 13 families were tested for AR inheritance mainly.

A multi-sample variant calling approach was employed on all the samples together. The standard BWA-mem-picard-GATK pipeline was applied according to GATK best practices [23] for variant calling. The procedure we employed in this study is similar to the study [65]. Further, we performed variant level QC by using GATK's VQSR and hard filtering to select the high quality variants as described previously in the study [65]. We annotated the variants following the same procedures as described in the Section 7.3.1. Briefly, ANNOVAR [47] was used to annotate the variants with their respective allele frequencies, predicted functions and deleterious score. Furthermore, to identify functionally important non-coding variants, additional databases were also used for annotation such as 1. CAGE clusters identified in the frontal lobe of 119 control individuals as described in the study [343]. This data is generated by performing an eQTL analysis based on cap analysis gene expression sequencing (CAGEseq) data which was created from human postmortem frontal lobe tissue and it is combined with genotypes obtained through genotyping arrays, exome sequencing, and CAGEseq. 2. CAGE clusters identified in 7 brain regions (Frontal, Temporal, Caudate, putamen, Cerebellum, Occipital) in FTD cases and controls (60 individuals in total) and 3. Bidirectionally expressed enhancers (eRNA) in the previous data-set. Additionally, the GnomAD data-set for allele frequency filtering (`http://gnomad.broadinstitute.org/`) was also taken into account. GnomAD is an extension of ExAC database [43] which includes the data from 123,136 exomes and 15,496 genomes.

Functional prioritization

We defined a variant as rare if it has a MAF of $< 1\%$ in the GnomAD [43]. Once the rare variants were selected, different strategies for coding and non-coding variants were applied in order to prioritize them. They are described below.

- **Coding variants:** Only nonsynonymous variants with a CADD phred score $>$ 20 were selected.
- **Non-coding variants:** In order to select the potentially deleterious variants, the criteria mentioned below were chosen and the qualifying variant has to satisfy at least one of them. Subsequently, only the qualifying variants with a CADD phred score $>$ 20 were selected similar to coding variants. We chose a CADD threshold of 20 because the variants above that threshold were supposed to be the top 1% most deleterious variants [58].
    - If it was already identified in the human gene mutation database (HGMD) [44] or ClinVar [344]
    - Annotated as an UTR5 or UTR3 variant based on RefSeq
    - If it was present in the upstream or downstream of a gene based on RefSeq
    - If it was present in the transcription factor binding sites according to ENCODE and ANNOVAR annotations
    - If it was present in the DNAse hyper sensitivity regions according to ENCODE annotations
    - If the variant was identified in any known GWAS previously
    - If the variant was a known miRNA target
    - If the variant is present in the regions sequenced by CAGE technology using the brain tissues from the study [343]

In addition to the variants prioritized by above mentioned criteria, we also generated a list of coding and non-coding variants annotated as disease causing with high confidence in HGMD and given in Table A.8.

CNV calling

CNV calling was performed by using RCP [345]. It works by using the depth of coverage information that is available in the BAM file [27] that is generated after the mapping of raw reads to the genome. For each sample of interest, RCP detects the CNVs by comparing the depth of coverage in the sample of interest to joint profiles that were pre-computed from a broad set of $>$ 6000 high quality genomes sequenced at a depth of $>$ 40X. In order to account for various confounding factors such as %GC, multi-genome profiles were constructed which represent the observed or inferred diploid depth of coverage for every position in the genome and they are called Reference Coverage Profiles (RCPs). By normalizing the scaled coverage of sample of interest to the RCP and employing a Hidden Markov model (HMM) segmentation approach, the CNVs were detected for each sample.

QC and filtering of CNVs

From the UCSC browser, we obtained a list of CNVs that were found previously and present in the questionable regions of the genome. They are described below and any CNV overlapping with the regions below was excluded using bedtools [233]. CNVs that are small or found on sex chromosomes (X and Y) are often unreliable and the false positive rate is high. Hence, we only selected CNVs with a minimum length of $>$ 10kb and present only on autosomes for subsequent analyses. Although, we might loose potential candidates by this stringent approach, we wanted to focus only on the highly reliable candidates for prioritization. Further, we only selected the complete deletions (copy number=0) and duplications with a minimum copy number of 4 for the downstream analysis.

- Centromeres (extracted from Chromosome band file): start co-ord - 500 kb ; end co-ord +500 kb
- Telomeres
- Immunoglobulins
- Mappability: DAC blacklist , Duke excluded and wgEncode CrgMappabilityAlign100mer (datavalue <= 0.25)
- GC percent (>=90 and <=10) : positions are merged by 10 bp with bedtools
- Common CNVs from DGV
- Repeat masker
- Gap locations
- Hiseq depth of 0.1%

Prioritization of genes

We employed various steps in order to identify the potential candidate genes as shown in the Figure 7.2 and obtained the list of genes harboring rare and variants co-segregating with disease. In the next step, we used Phenolyzer [328] as described in the Section 7.3.1 with the search term “parkinson” in order to narrow down the candidate genes.

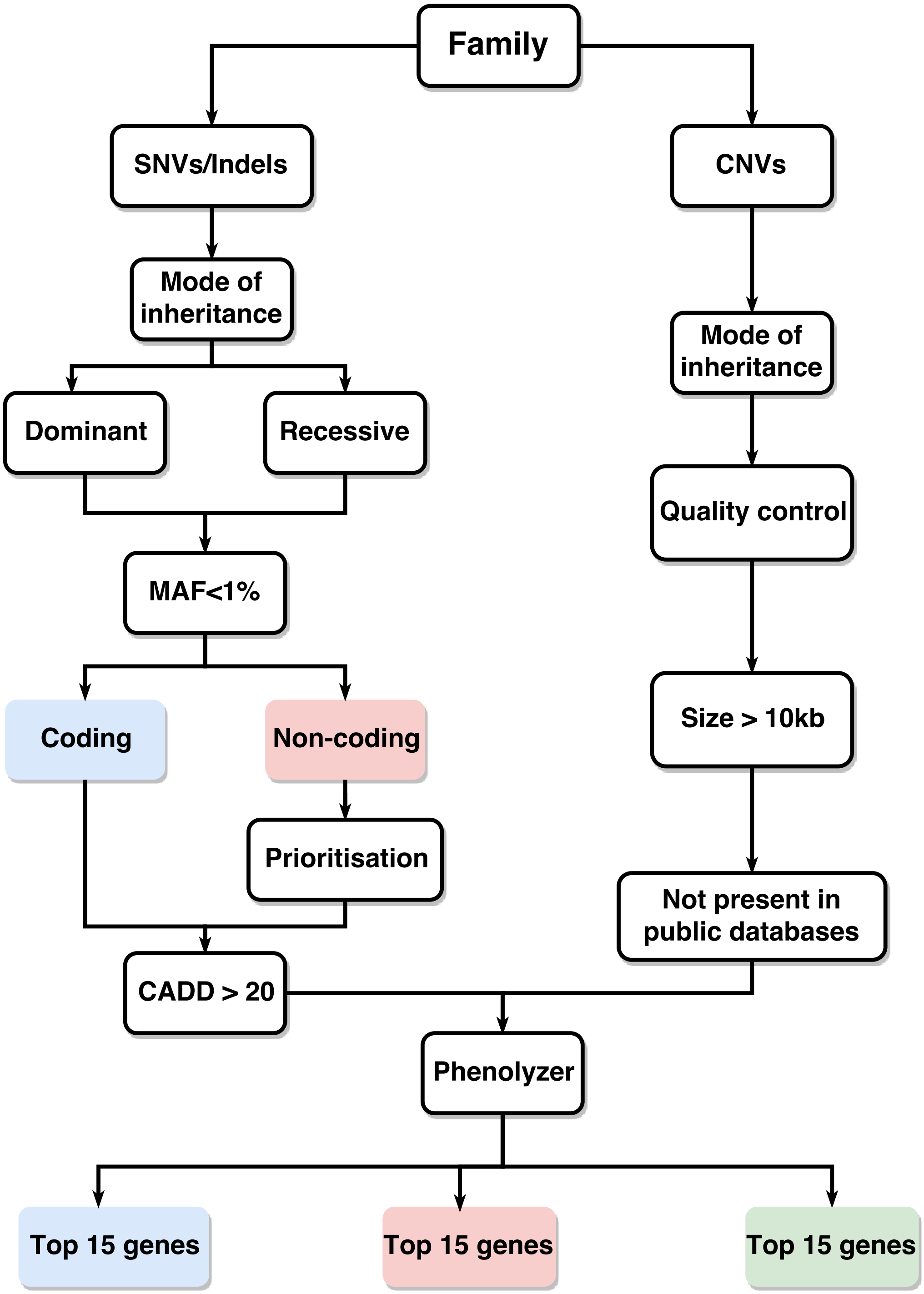


Figure 7.2: Different steps employed for the variant prioritization in the Courage-PD data-set.

7.4.2 Results

Variant calling and QC

A total of 23,646,530 variants were called across all the samples and the Ti/Tv ratio equaled to 1.95 before performing QC. After the QC, 21,609,473 remained in the final data-set and the Ti/Tv ratio has increased to 2.08. On an average, each sample had 3,662,556 SNPs and 1,243,632 indels after QC.

Selection of candidate genes

As shown in the Figure 7.2, we prioritized the genes based on multiple evidences and selected the top 15 ranked genes carrying coding, non-coding and CNVs. Some families did not have segregating genes in all the coding, non-coding and CNVs categories. Of the 36 families that we analyzed, there were 164 genes following autosomal dominant inheritance and 26 genes following autosomal recessive inheritance and carrying coding deleterious variants. In the similar lines, for genes carrying non-coding deleterious variants there are 149 genes following autosomal dominant and 15 genes following autosomal recessive inheritance respectively. From the CNV analysis, only 12 genes were found to carry a deletion and only one gene was found to carry a duplication. An entire list of all the prioritized genes is given in the Table A.6. A few interesting findings were mentioned below.

Heterozygous variants in ***LRRK2***, ***TRAP1*** and ***GRIN2A***

The most studied variant in *LRRK2* gene is p.G2019S with regard to PD. However, we identified two additional variants in *LRRK2* in two families. The first variant p.R1514Q was found in a pedigree from Turkey which comprises of an affected and an unaffected sibling as shown in the PD296 of Figure 7.3. This variant was already studied in two previous studies in PD [320, 346]. In the study [346] it was found in 6/98 PD patients and in the study [320] it was segregating with the disease, however it did not withstand their prioritization step. In our study, the variant was present in only one individual effected with PD whereas it was absent in the unaffected sibling. The affected sibling had an early age-of-onset of 36 years. The second variant in *LRRK2* which fulfilled our criteria is a nonsynonymous variant p.M96T found in a pedigree of Spanish ethnicity. It has not been linked to PD in any previous study. It is present in heterozygous state in both the affected siblings and they had an aao of 65 and 53 years. The pedigree is shown in the HCB4 of Figure 7.3. The same *GRIN2A* variant as identified in the Section 7.3.2 p.N1076K was also identified in another family in this part of the study making it a total of three families in which the variant was detected. Recently, in a separate study we found a LoF variant in *TRAP1* [292] in a PD case. In this study, we identified a heterozygous nonsynonymous variant p.R469H which is present in two affected individuals and not present in the unaffected individual as shown in the HCB5 of Figure 7.3.

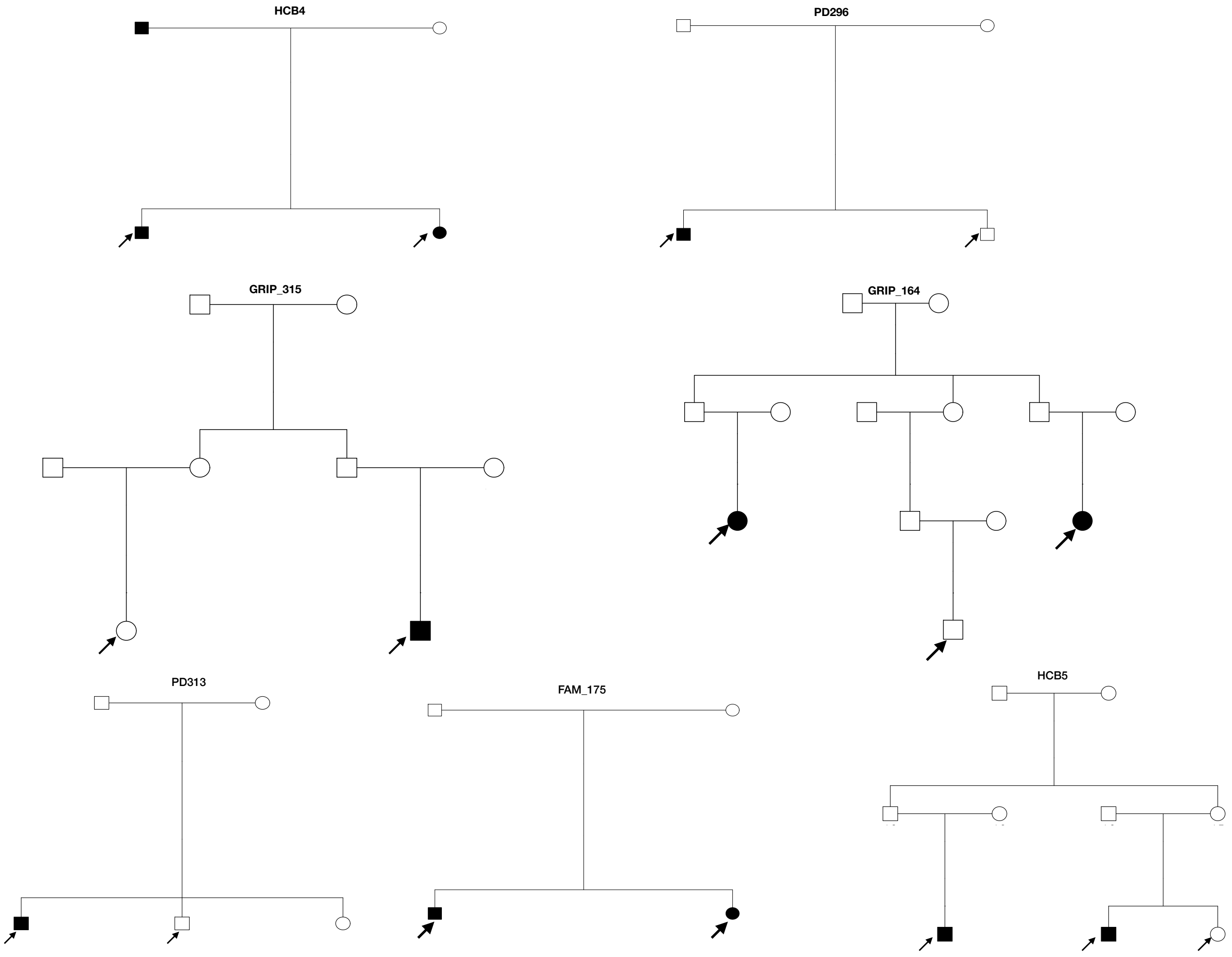


Figure 7.3: Pedigrees carrying variants in LRRK2: HCB4 carries *LRRK2* p.M96T variant, PD296 carries *LRRK2* p.R1514Q variant, HCB5 carries TRAP1 p.R469H variant, GRIP315 carries *MECP2* p.T311M variant, PD313 carries *SBF1* p.R1053W, GRIP_164 carries 13:47812108 A>G variants and FAM_175 carries p.G199S and p.S438G variants in *ATP13A2* and *FBXO7* respectively. The arrows indicate the samples that underwent WGS.

Homozygous variants in *MECP2* and *SBF1*

We identified a recessive p.T311M nonsynonymous variant in *MECP2*. It is present in one affected male individual and not in the unaffected individual. The same mutation was identified in Rett syndrome in previous studies [347, 348] and the pedigree is shown in GRIP315 of Figure 7.3. We identified a nonsynonymous homozygous variant p.R1053W in *SBF1*, it is present as a homozygous variant in the affected sample and as a heterozygous variant in the unaffected. The pedigree is shown in the PD313 of Figure 7.3.

Compound heterozygous variants in *PARK2*

By using our integrated approach, we identified a compound heterozygous variant in *PARK2* in a family comprising of an affected and unaffected sibling. One of the variant is an intronic variant (6:162855008-T>C) predicted to be affecting the Transcription factor binding site and the other one is p.R256C which is a nonsynonymous variant and has been previously studied [349].

Heterozygous variants in *FBXO7* and *ATP13A2* in the same family

We discovered two heterozygous variants p.G199S and p.S438G in two recessive genes names *FBXO7* and *ATP13A2* respectively in an Italian family FAM_175 of Figure 7.3. Previously, it was also seen in an Italian family, that the affected samples carry homozygous variant in *ATP13A2* and also a heterozygous variant in *FBXO7* [350].

Variants in *ERBB4* and *KIF2A*

In addition to the coding variants identified in this study, we also identified several non-coding variants according to our criteria as defined in the Section 7.4.1. The genes harboring deleterious non-coding variants are associated to various neurological disorders. One such example is *ERBB4*, which was found to be involved in Schizophrenia [351]. In the same family another variant is found in the *KIF2A*. In a previous study, it has been shown that *KIF2A* which is an anterograde motor protein [352] could serve as a biomarker for PD. Further, *KIF2A* was also found to carry a mutation in another family in our study.

Promoter variant in ATG7

In a previous study in PD, a promotor variant of *ATG7* was found (3:11313449 G>A) and in our study also we identified the same variant based on the HGMD annotation. However, it has a CADD phred score 1.74 and moreover it is present in the same family in which the LRRK2 p.M96T. All the variants annotated as disease causing with high confidence are given in the Table A.8.

Variants in *HTR2A*

A variant (13:47812108 A>G) affecting the transcription factor binding site of *HTR2A* is discovered in our study. Previously, a genetic variant in *HTR2A* is associated to the repetitive behavior and impulse control [353] in PD. The pedigree carrying this variant is given in GRIP_164 of Figure 7.3.

7.4.3 Discussion

Familial studies have been successful in the past to identify several genes related to PD [108], still there is a paucity in the discovery of novel genes underlying familial-PD. In this study, our objective was to identify meaningful genes by employing WGS in several small families and a multiple evidence based approach. In our quest to identify potential candidate genes, we discovered various interesting variants/genes (Table A.6). The most promising candidates are discussed in this section. A variant p.R1514Q in *LRRK2* is found in our study and it has been previously found in more than one study [320, 346] but has not been followed up, it is especially interesting because the variant is found in a pedigree of Turkish origin but a previous study mainly composing of North European samples [346] found that this variant has higher frequency in cases. This indicates that p.R1514Q might not be just restricted to North European population and additional studies including samples from other countries are warranted. Another variant in *LRRK2* p.M96T with no previous association to PD was found in our study, but in the same family another variant in the promotor region of *ATG7* was also discovered. The variant 11313449G>A in *ATG7* has been previously shown to be implicated in PD [354] by significantly decreasing transcriptional activities of the gene promotor of *ATG7*. The proposed mechanisms of actions were probably by either completely abolishing, modifying and creating binding sites for putative transcription factors. Hence, it is probable that 11313449G>A variant is the causal gene in the family HCB4 rather than p.M96T, but further functional studies are needed to be done to validate this finding.

PARK2 is known to have an autosomal recessive inheritance requiring both the alleles to carry a variant to induce an early-onset PD. In our study, we detected an exonic variant and a deleterious intronic variant in the same individual constituting a compound heterozygous variant. The exonic heterozygous variant p.R256C was also identified in a previous study comprising of early onset sporadic PD cases of French origin [355], but the pedigree in which the variant was identified is of Dutch origin. One, shortcoming of the previous study was that they excluded the intronic variants from the analysis and maybe it is worth to also look into the intronic *PARK2* variants. Interestingly the variant p.R256C is present within RING finger 1 of Parkin protein and could cause a gain-of-function [356]. Another study showed that this variant is a risk factor for nigrostriatal dysfunction [357]. In view of this findings it would be interesting to see whether p.R256C acts alone or if it acts in conjunction with the intronic variant discovered in this study.

TRAP1 is shown to be involved in mitochondrial dysfunction and in our previous study we found an individual affected with PD carrying a LoF variant in *TRAP1* [292]. We identified a recessive variant in *MECP2* and the variant has been previously identified in the context of Rett syndrome [347, 348].

A combination of *ATP13A2* and *FBXO7* heterozygous variants was found in an Italian family. Interestingly, in an earlier study also the same combination of genes were found to be mutated in an Italian pedigree [350]. Although, in that study the variant in *ATP13A2* was homozygous, it has been shown recently that heterozygous variants in *ATP13A2* could also implicate PD by causing the cellular dysfunction [358]. In this context, it could be possible that this combination is specific to Italian population and maybe the variants in *ATP13A2* and *FBXO7* together increase the risk of PD, but it is speculative and remains to be checked. *MECP2* acts on neuronal development and function and it has been proposed to be a drug target for PD [359] based on the functional studies. Previous studies conducted on mice have shown that *MECPH2* knock-out has resulted in PD like symptoms such as loss of motor deficits indicating the disruption of nigrostriatal pathway [360]. Hence, it is a very prominent candidate and needs to be investigated in the future.

SBF1 mutations have been previously associated to Charcot-Marie-Tooth disease [361, 362] and axonal neuropathy [363] and the patient carrying the homozygous *SBF1* variant has a very early aao of 27 years and hence it is possible that it is associated to the early onset PD. In summary, this study shows the power of integrating multiple evidences to discover putative novel candidate genes. All the variants identified in this study need to be further followed up either in a familial setting or in a large case-control cohort.

CHAPTER 8

CONCLUSIONS AND OUTLOOK

8.1 Conclusions

NGS has provided us with a unique opportunity to identify genetic variants underlying a disease. Currently, generating the massive amounts of data is not the issue, however the processing and interpretation of data is the bottleneck. Especially, the interpretation of variants found via familial studies is often quite challenging because we often end up with several variants that co-segregate with the disease. This problem is similar to finding a needle in a haystack and it gets even severe if the families that are being analyzed are small, which was also the case in our study. Hence, to overcome this problem we integrated several sources of knowledge in this thesis to find meaningful associations. Our main goal was to focus on variants occurring at low frequency with larger effect size and identify those causing a significant burden in the case-control studies and in familial-PD we aimed to find the putative candidate variants/genes causing the disease. In order to achieve these goals, we developed the variant analysis pipelines and processed several data-sets generated via WES and WGS belonging to two of the most common diseases PD and epilepsy. Although PD and epilepsy are two distinct brain diseases, their underlying genetic architecture is complex and polygenic. By employing state of the art statistical and analytic methods, we shed further light on the genetics of both PD and epilepsy and found several imperative and interesting findings in this thesis.

In RE/ARE, several twin studies have been performed and majority of them remained discordant [142, 143]. Similarly, various candidate gene studies have been performed [148, 149, 163], where the emphasis was only on one or few genes. However, an unbiased genome wide study focusing on the burden of rare variants in RE/ARE was absent. In this work (Chapter 2), we carried out a first ever exome-wide association study in RE/ARE. In this study, we investigated the genetic variants occurring at very low frequency (0.5%) and causing large effect in RE/ARE cases compared with the in-house and the ExAC database. We performed the gene-level burden analysis and showed that *GRIN2A* is the only gene reaching a genome-wide significance. Although there might be other genes also contributing to the disease, we could not detect such genes mainly due to the sample size. Hence, in order to compensate the smaller sample size of our cohort, we extended the burden analysis to geneset level and showed that there is a consistent increase in the odds ratios for the LoF variants in several disease associated gene-sets. The genesets are comprised of genes under negative selection, glutamate receptors and genes associated with epileptic encephalopathies. *GRIN2A* was excluded from the selected genesets to show that besides *GRIN2A*, there are other genes also which are contributing to the disease etiology. Additionally, we identified several interesting novel LoF variants (Not present in ExAC) and provided in the Table A.1, which could be used as a future references for other research groups to perform functional validations.

GGE has always been believed to have a strong underlying genetic cause and many variants in ion-channel genes have been identified via various familial studies. Nonetheless, no study has shown an excess of variants in ion-channel genes via whole-exome sequencing, keeping the genetic cause of GGE under the wraps. In this study (Chapter 3), we performed the burden analysis in a cohort comprising of familial and sporadic GGE cases (>1000 cases combined) and showed, for the first time there is a significant burden of rare nonsynonymous variants in a group of 19 $\mathrm{GABA_A}$ receptor genes. In total, similar findings were observed in 3 independent cohorts (1 discovery and 2 replication cohorts). Furthermore, functional studies of selected segregating variants in $\mathrm{GABA_A}$ receptors showed that they induced a change in the receptor function, providing a compelling evidence that $\mathrm{GABA_A}$ receptors and GABAergic mechanism play a role in the GGE supporting the previous studies [139, 177, 188, 191, 192, 204–206].

To our knowledge, we performed the largest WES based CNV study in RE/ARE and GGE in Chapter 4. A significant burden of rare and large deletions were observed in cases compared to the controls was observed. This finding was similar to a study performed using array data [221] comprising a larger number of samples. We performed the functional enrichment analysis of the deletions belonging to RE/ARE and GGE separately and showed that there is a difference between both the diseases at the biological pathway level. This finding was in line with a previous effort to classify metabolic and developmental epilepsies [14]. In an effort to identify interesting deletions which could be either be followed up by functional validation or as a list for other researchers, we found several candidate single-gene deletions which are described in detail in the section 4.5. We also integrated the SNV/Indels identified in chapters 2 and 3 along with the deletions identified in Chapter 4 and found several interesting candidates which are also described in detail.

By using the WES data in Chapter 5 we have shown a genome-wide burden of multiple singleton LoF variants. Similar to the studies in epilepsy as described above, it is difficult to identify single genes with a genome-wide significance mainly due to the multiple testing correction. Hence, alternate approaches are a need of the hour to compensate for the loss of power in studies with small sample sizes. In the current study, we employed an alternative approach and our results show that there is an increase of singleton LoF variants within the PD samples compared to the controls. Based on the evidence provided by our study, it could be contemplated that in PD, the burden is distributed across the entire exome rather than being confined to a group of PD associated genes (Section 5.4.3). In summary, our findings support the complex genetic architecture of PD and show that there is still a lot of missing heritability in PD which needs to be unearthed. Along with the major limitation of our study which is the small sample size, there might be additional factors such as variants in the non-coding regions which could also contribute to the progression of PD. The second limitation would be the non-availability of a replication cohort, therefore despite the fact that there is a exome-wide significance of singleton LoF variants, our study should be considered preliminary and needs replication in larger PD cohorts.

The work from Chapter 6 indicated an enrichment of disease-associated variants in the exon-intron boundary of brain expressed genes in PD and underscore the therapeutic potential of compounds acting on pathological splicing also in sporadic PD cases. Our study illustrates the promise for treatment approaches in precision medicine in PD that focus on genetic and molecular stratification. To account for the increasingly recognized heterogeneity in PD and other neurodegenerative disorders, new strategies need to be developed for the stratification of patients along shared pathogenic mechanisms. Our study highlights the importance of variants regulating splicing mechanism in PD, especially the U1 splice variants.

In the first part of Chapter 7, we used a two-stage approach of a discovery and replication cohort where we identified 14 genes involved in the nervous system pathway. The variants harbored by those genes are predicted to be rare and well conserved. However, neither the variants nor the genes were present in more than 1 family in the discovery cohort warranting further evidence of their involvement in PD, although it is partially supported by their odds ratio of >1 in an independent case-control study. In the second part, we identified various interesting variants harboring heterozygous, compound heterozygous and homozygous variants in genes that were either associated to PD or related diseases. We identified several interesting coding, non-coding and CNVs from the WGS of familial-PD. The list of variants could be further prioritized and functional validation can be performed on the most promising candidates.

Taken together, the studies presented in this thesis involve a broad range of methods and topics that are expected to become increasingly important in the genetic study of PD and epilepsy, as well as other common diseases, in the years to come. In my opinion the key message of this work is that the genetics of PD and epilepsy is more bewildering than we expected.

8.2 Outlook

8.2.1 Larger cohort sizes for increased statistical power

It could be seen clearly from the results in Chapter 2, that there are increased odds ratios of LoF variants in several disease related genesets in RE/ARE. In the future, more patients should be recruited and more comprehensive burden analyses at single variant, gene and geneset levels needs to be performed to get significant results. For the GGE, a meta-analysis of various cohorts described in the study or a combined analysis where a multiple-sample calling on all the samples together could be performed in order to have an increased power. This approach will increase the sample size significantly, thereby providing more evidence and not just at the geneset level, but also at a single gene level. Additionally, a new statistical approach could be developed in-order to combine the SNV/Indel and CNV data to untangle the genetic complexity of RE/ARE and GGE. In the near future the findings from Chapter 7 should be replicated in larger datasets such as the latest dataset from Courage-PD comprising of >5000 PD cases that are currently being genotyped on the Neurochip platform [364].

8.2.2 Polygenic risk score (PRS)

Although, the clinicians are very well trained and follow stringent classification criteria, it could be still possible that patient might be wrongly classified into a different disease. One possible reason of the complexity of genetics in PD and epilepsy is the phenotypisation of patients. Hence, instead of looking at one type of disease maybe we are actually looking at a combination of different diseases. More precise phenotypes together with higher samples sizes would allow to untangle the genetic architecture of complex diseases. To help classify the patients into more precise phenotypes, one could take advantage of the avalanche of clinical data that is available along with the genetic information in the form of PRS. Robust machine learning and clustering algorithms could be built in order to tackle this problem. At the current stage, PRS is not utilizing the complete genetic information because it is generated based on the common variants only. A more robust method to generate PRS based on rare and common variants needs to be developed. The benefits of generating an accurate PRS are plenty, as it can be generated for any trait of interest. One possible application of PRS would be to build better prediction models in order to detect the predisposition for a trait/disease earlier, as one could be better prepared for the engagement in early stage or treatment prevention strategies. The models built on the basis of PRS could aid the clinicians in decision support and counseling the patients. Another example would be to predict the age at onset of a disease or other continuous variables such as any clinical scores. To achieve this one could use genetic data, already existing clinical data or a combination of both, thereby reducing the medical treatment costs as well as the burden for patients due to the diagnostics.

8.2.3 Regulatory, splice variants and genome-wide deleterious score

In Chapter 6 we only looked at variants disrupting the U1 splice sites. However, there is a possibility that a variant could generate a new U1 splice site. Hence, in the future, variants that generate putative splice sites should also be predicted. Similarly, the pipelines developed in this work (Chapter 6) can also be extended to predict deleterious 3' splice site variants. Splicing is not a PD specific mechanism and it has been shown to be disrupted in various diseases [365]. Hence, the same methodology developed in Chapter 6 can also be applied to other disorders. The current state of variant prioritization methods are focused mostly on identifying disease causing variants in coding regions. However, from the recent studies it is clear that, we have begun only scratching the surface of the problem and there is an enormous dearth of knowledge with regard to the non-coding variants and their potential role in the diseases. One such example of functionally important non-coding variants are regulatory variants. They drive the gene expression by acting on enhancers, gene promoters, or binding sites for RNA or proteins. Due to their importance in the functioning of cells in the body, variants disrupting or creating new regulatory regions should be investigated in more depth. One resource that could be of great help to achieve this goal is the ENCODE project [366]. It is a valuable resource comprising of information about several important regulatory regions such as transcription factor binding sites, DNAse hyper sensitivity sites, histone modification regions etc. Currently, we have several WGS/WES datasets belonging to various disorders and the pipelines that will be developed in the future to predict disease causing non-coding variants could be applied to all the in-house datasets.

The predicted disease causing ability of a variant can be converted into a single metric called as deleterious score. Few examples of such scores that are widely used include SIFT [53], PolyPhen2_HDIV [54], LRT [55], MutationTaster [56], PROVEAN [57], CADD [58], fathmm [60], GERP++_R [61], DANN [59] and SiPHy [62]. Some of them such as CADD, DANN, GERP++_R and SiPHy are available for all the variants in the human genome including the non-coding variants. Nonetheless, they are not yet reliable for non-coding variants mainly due to the fact that there is no gold standard dataset available to evaluate their performance, unlike the coding variants. Albeit discovering meaningful associations and variants in this thesis, it is quite possible that this work missed several variants due to the lack of appropriate standards to evaluate them, especially the variants occurring in the non-coding regions of the human genome which is often referred to as the “dark genome”. This is an area where there is an enormous potential to be improved and one can take the advantage of large collection of data that is publicly available and use sophisticated machine learning algorithms to predict the potential deleterious non-coding variants and generate a genome-wide score. Experimental biologists should also come up with new methods to functionally validate the non-coding variants, such that they are able to generate and provide more experimental data which can then be used in formulating a robust genome-wide deleterious score.

A statistical framework needs to be established in order to integrate the information from coding, non-coding variants and CNVs. A network guided approached could be used to identify important biological pathways. One way to perform network analysis is to use the p-values generated from the association analysis as weights for the genes and project the genes onto a protein-protein interaction network (PPIN) and find the enriched modules. Such modules could show us which pathways are over represented in a particular disease or across different diseases. In order to be more specific to neurodegenerative disease, we can use a PPIN generated from brain as a whole or specific sub-tissues like substantia nigra or cell types like dopaminergic neurons or astrocytes. Such integration of transcriptomic and genomic data from the human brain tissue that is available publicly from repositories such as The Genotype-Tissue Expression project (GTEx) [367] or from single-cell transcriptomic data available for various cell types of brain [368] would help us to delineate the complexity of PD and epilepsy.

Personalized medicine in practice

We are living in exciting times, technologies such as NGS and artificial intelligence (AI) driven by machine learning have absolutely changed the way we look at the data. They enabled researchers to make sense of big data by revealing interesting patterns in the human genome. In my opinion, by combining NGS, AI and CRISPR/cas technologies, one can achieve wonders in the field of personalized medicine. For instance, we perform the WGS of a patient, utilize either or a combination of: risk scores, regression models, machine learning algorithms to identify the genetic cause and correct the genetic mutation by applying CRISPR/Cas technology, thereby restoring the original function of a gene. Although, we are still at the beginning of this phase, with the dropping NGS costs and the amount of research that is being performed in these areas, it is quite possible that we will witness the era of personalized medicine in the near future.

To achieve the goal of precision medicine in epilepsy and PD, we need to address the following points: Much larger cohorts are need to be built such as Epi25 for epilepsy (`http://epi-25.org/` which aims to conduct WES of 25,000 patients) and they have to be cautiously defined both phenotypically and genomically; functional characterization of the mutations identified via various studies have to be conducted and results should be carefully interpreted by the domain experts; standard guidelines should be established to define the pathogenicity of variants. These goals can only be achieved through the collaboration and integration of research groups and by bringing the researchers with genetic, clinical and biological expertise under one umbrella.

BIBLIOGRAPHY

[1] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K. Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui Wei, Ron Wides, Chunlin Xiao, Chunhua Yan, Alison Yao, Jane Ye, Ming Zhan, Weiqing Zhang, Hongyu Zhang, Qi Zhao, Liansheng Zheng, Fei Zhong, Wenyan Zhong, Shiaoping C. Zhu, Shaying Zhao, Dennis Gilbert, Suzanna Baumhueter, Gene Spier, Christine Carter, Anibal Cravchik, Trevor Woodage, Feroze Ali, Huijin An, Aderonke Awe, Danita Baldwin, Holly Baden, Mary Barnstead, Ian Barrow, Karen Beeson, Dana Busam, Amy Carver, Angela

Center, Ming Lai Cheng, Liz Curry, Steve Danaher, Lionel Davenport, Raymond Desilets, Susanne Dietz, Kristina Dodson, Lisa Doup, Steven Ferriera, Neha Garg, Andres Gluecksmann, Brit Hart, Jason Haynes, Charles Haynes, Cheryl Heiner, Suzanne Hladun, Damon Hostin, Jarrett Houck, Timothy Howland, Chinyere Ibegwam, Jeffery Johnson, Francis Kalush, Lesley Kline, Shashi Koduru, Amy Love, Felecia Mann, David May, Steven McCawley, Tina McIntosh, Ivy McMullen, Mee Moy, Linda Moy, Brian Murphy, Keith Nelson, Cynthia Pfannkoch, Eric Pratts, Vinita Puri, Hina Qureshi, Matthew Reardon, Robert Rodriguez, Yu-Hui Rogers, Deanna Romblad, Bob Ruhfel, Richard Scott, Cynthia Sitter, Michelle Smallwood, Erin Stewart, Renee Strong, Ellen Suh, Reginald Thomas, Ni Ni Tint, Sukyee Tse, Claire Vech, Gary Wang, Jeremy Wetter, Sherita Williams, Monica Williams, Sandra Windsor, Emily Winn-Deen, Keriellen Wolfe, Jayshree Zaveri, Karena Zaveri, Josep F. Abril, Roderic Guigó, Michael J. Campbell, Kimmen V. Sjolander, Brian Karlak, Anish Kejariwal, Huaiyu Mi, Betty Lazareva, Thomas Hatton, Apurva Narechania, Karen Diemer, Anushya Muruganujan, Nan Guo, Shinji Sato, Vineet Bafna, Sorin Istrail, Ross Lippert, Russell Schwartz, Brian Walenz, Shibu Yooseph, David Allen, Anand Basu, James Baxendale, Louis Blick, Marcelo Caminha, John Carnes-Stine, Parris Caulk, Yen-Hui Chiang, My Coyne, Carl Dahlke, Anne Deslattes Mays, Maria Dombroski, Michael Donnelly, Dale Ely, Shiva Esparham, Carl Fosler, Harold Gire, Stephen Glanowski, Kenneth Glasser, Anna Glodek, Mark Gorokhov, Ken Graham, Barry Gropman, Michael Harris, Jeremy Heil, Scott Henderson, Jeffrey Hoover, Donald Jennings, Catherine Jordan, James Jordan, John Kasha, Leonid Kagan, Cheryl Kraft, Alexander Levitsky, Mark Lewis, Xiangjun Liu, John Lopez, Daniel Ma, William Majoros, Joe McDaniel, Sean Murphy, Matthew Newman, Trung Nguyen, Ngoc Nguyen, Marc Nodell, Sue Pan, Jim Peck, Marshall Peterson, William Rowe, Robert Sanders, John Scott, Michael Simpson, Thomas Smith, Arlan Sprague, Timothy Stockwell, Russell Turner, Eli Venter, Mei Wang, Meiyuan Wen, David Wu, Mitchell Wu, Ashley Xia, Ali Zandieh, and Xiaohong Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351.

[2] U. Lagerkvist. "two out of three": an alternative method for codon reading. *Proceedings of the National Academy of Sciences*, 75(4):1759–1762.

[3] Ju-Hyun Park, Mitchell H. Gail, Clarice R. Weinberg, Raymond J. Carroll, Charles C. Chung, Zhaoming Wang, Stephen J. Chanock, Joseph F. Fraumeni, and Nilanjan Chatterjee. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences of the United States of America*, 108(44):18026–18031, .

[4] Orion J. Buske, AshokKumar Manickaraj, Seema Mital, Peter N. Ray, and Michael Brudno.

Identification of deleterious synonymous variants in human genomes. *Bioinformatics*, 29 (15):1843–1850.

[5] Bo-Young Kim, Jung Hoon Park, Hye-Yeong Jo, Soo Kyung Koo, and Mi-Hyun Park. Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data. *PLoS ONE*, 12(8), .

[6] Ning Zhang, Tao Huang, and Yu-Dong Cai. Discriminating between deleterious and neutral non-frameshifting indels based on protein interaction networks and hybrid properties. *Molecular Genetics and Genomics*, 290(1):343–352, .

[7] Peter H. Sudmant, Swapan Mallick, Bradley J. Nelson, Fereydoun Hormozdiari, Niklas Krumm, John Huddleston, Bradley P. Coe, Carl Baker, Susanne Nordenfelt, Michael Bamshad, Lynn B. Jorde, Olga L. Posukh, Hovhannes Sahakyan, W. Scott Watkins, Levon Yepiskoposyan, M. Syafiq Abdullah, Claudio M. Bravi, Cristian Capelli, Tor Hervig, Joseph T. S. Wee, Chris Tyler-Smith, George van Driem, Irene Gallego Romero, Aashish R. Jha, Sena Karachanak-Yankova, Draga Toncheva, David Comas, Brenna Henn, Toomas Kivisild, Andres Ruiz-Linares, Antti Sajantila, Ene Metspalu, Jüri Parik, Richard Villems, Elena B. Starikovskaya, George Ayodo, Cynthia M. Beall, Anna Di Rienzo, Michael Hammer, Rita Khusainova, Elza Khusnutdinova, William Klitz, Cheryl Winkler, Damian Labuda, Mait Metspalu, Sarah A. Tishkoff, Stanislav Dryomov, Rem Sukernik, Nick Patterson, David Reich, and Evan E. Eichler. Global diversity, population stratification, and selection of human copy number variation. *Science (New York, N.Y.)*, 349(6253):aab3761, .

[8] Global variation in copy number in the human genome. *Nature*, 444(7118):444.

[9] Mehdi Zarrei, Jeffrey R. MacDonald, Daniele Merico, and Stephen W. Scherer. A copy number variation map of the human genome. *Nature Reviews Genetics*, 16(3):172–183.

[10] PJ Hastings, James R Lupski, Susan M Rosenberg, and Grzegorz Ira. Mechanisms of change in gene copy number. *Nature reviews. Genetics*, 10(8):551–564.

[11] Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853.

[12] Barkur S. Shastry. Copy number variation and susceptibility to human disorders (review). *Molecular Medicine Reports*, 2(2):143–147.

[13] de Leeuw Nicole, Dijkhuizen Trijnie, Hehir-Kwa Jayne Y., Carter Nigel P., Feuk Lars, Firth Helen V., Kuhn Robert M., Ledbetter David H., Martin Christa Lese, van Ravenswaaij-Arts Conny M. A., Scherer Steven W., Shams Soheil, Van Vooren Steven, Sijmons Rolf, Swertz Morris, and Hastings Ros. Diagnostic interpretation of array data using public databases and internet sources. *Human Mutation*, 33(6):930–940.

[14] Kamel Jabbari and Peter Nürnberg. A genomic view on epilepsy and autism candidate genes. *Genomics*, 108(1):31–36.

[15] Matthew F. Pescosolido, Ece D. Gamsiz, Shailender Nagpal, and Eric M. Morrow. Distribution of disease-associated copy number variants across distinct disorders of cognitive development. *Journal of the American Academy of Child and Adolescent Psychiatry*, 52 (4):414–430.e14.

[16] Lars Feuk, Andrew R. Carson, and Stephen W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97.

[17] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467.

[18] Mike A. Nalls, Jose Bras, Dena G. Hernandez, Margaux F. Keller, Elisa Majounie, Alan E. Renton, Mohamad Saad, Iris Jansen, Rita Guerreiro, Steven Lubbe, Vincent Plagnol, J. Raphael Gibbs, Claudia Schulte, Nathan Pankratz, Margaret Sutherland, Lars Bertram, Christina M. Lill, Anita L. DeStefano, Tatiana Faroud, Nicholas Eriksson, Joyce Y. Tung, Connor Edsall, Noah Nichols, Janet Brooks, Sampath Arepalli, Hannah Pliner, Chris Letson, Peter Heutink, Maria Martinez, Thomas Gasser, Bryan J. Traynor, Nick Wood, John Hardy, and Andrew B. Singleton. NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiology of Aging*, 36(3):1605.e7–1605.e12, .

[19] Iris E. Jansen, Hui Ye, Sasja Heetveld, Marie C. Lechler, Helen Michels, Renée I. Seinstra, Steven J. Lubbe, Valérie Drouet, Suzanne Lesage, Elisa Majounie, J. Raphael Gibbs, Mike A. Nalls, Mina Ryten, Juan A. Botia, Jana Vandrovcova, Javier Simon-Sanchez, Melissa Castillo-Lizardo, Patrizia Rizzu, Cornelis Blauwendraat, Amit K. Chouhan, Yarong Li, Puja Yogi, Najaf Amin, Cornelia M. van Duijn, Huw R. Morris, Alexis Brice, Andrew B. Singleton, Della C. David, Ellen A. Nollen, Shushant Jain, Joshua M. Shulman, and Peter Heutink. Discovery and functional prioritization of parkinson's disease candidate genes from large-scale whole exome sequencing. *Genome Biology*, 18:22.

[20] Gerald Goh and Murim Choi. Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. *Genomics & Informatics*, 10(4):214–219.

[21] Janine Meienberg, Rémy Bruggmann, Konrad Oexle, and Gabor Matyas. Clinical sequencing: is WGS the better WES? *Human Genetics*, 135:359–362.

[22] Aziz Belkadi, Vincent Pedergnana, Aurélie Cobat, Yuval Itan, Quentin B. Vincent, Avinash Abhyankar, Lei Shang, Jamila El Baghdadi, Aziz Bousfiha, the Exome/Array Consor-

tium, Alexandre Alcais, Bertrand Boisson, Jean-Laurent Casanova, Laurent Abel, Waleed Al-Herz, Cigdem Arikan, Peter Arkwright, Cigdem Aydogmus, Olivier Bernard, Lizbeth Blancas-Galicia, Stéphanie Boisson-Dupuis, Damien Bonnet, Omar Boudghene Stambouli, Lobna Boussofara, Jeannette Boutros, Jacinta Bustamante, Michael Ciancanelli, Theresa Cole, Antonio Condino-Neto, Mukesh Desai, Claire Fieschi, José Luis Franco, Philippe Ichai, Emmanuelle Jouanguy, Melike Keser-Emiroglu, Sara S. Kilic, Seyed Alireza Mahdaviani, Nizar Mahlhoui, Davood Mansouri, Nima Parvaneh, Capucine Picard, Anne Puel, Didier Raoult, Nima Rezaei, Ozden Sanal, Silvia Sanchez Ramon, François Vandenesch, Guillaume Vogt, and Shen-Ying Zhang. Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *Proceedings of the National Academy of Sciences*, 113(24):6713–6718.

[23] Mark A. DePristo, Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, Guillermo del Angel, Manuel A. Rivas, Matt Hanna, Aaron McKenna, Tim J. Fennell, Andrew M. Kernytsky, Andrey Y. Sivachenko, Kristian Cibulskis, Stacey B. Gabriel, David Altshuler, and Mark J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498.

[24] Andrew. FastQC: a quality control tool for high throughput sequence data.

[25] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–10.

[26] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

[27] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, .

[28] Ben Langmead. Aligning short sequencing reads with bowtie. *Current protocols in bioinformatics*, CHAPTER:Unit–11.7.

[29] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, 00(0):3–3.

[30] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31(12):2032–2034.

[31] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Research*, 8(3):175–185.

[32] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Chris Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 11(1110):11.10.1–11.10.33.

[33] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, 27(15):2156–8.

[34] 1000 Genomes Project Consortium, Goncalo R. Abecasis, Adam Auton, Lisa D. Brooks, Mark A. DePristo, Richard M. Durbin, Robert E. Handsaker, Hyun Min Kang, Gabor T. Marth, and Gil A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

[35] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio]*.

[36] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, Andrew O M Wilkie, Gil McVean, and Gerton Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics*, 46(8):912–8.

[37] International HapMap Consortium. The international HapMap project. *Nature*, 426(6968): 789–796.

[38] Ryan F. McCormick, Sandra K. Truong, and John E. Mullet. RIG: Recalibration and interrelation of genomic sequence data with the GATK. *G3: Genes, Genomes, Genetics*, 5(4):655–665.

[39] Christopher C Chang, Carson C Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):1–16, .

[40] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873.

[41] Hong Li, Gustavo Glusman, Chad Huff, Juan Caballero, and Jared C. Roach. Accurate and robust prediction of genetic relationship from whole-genome sequences. *PLOS ONE*, 9(2):e85437, .

[42] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):ng1847.

[43] Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen, Daniel P. Birnbaum, Jack A. Kosmicki, Laramie E. Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N. Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I. Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M. Peloso, Ryan Poplin, Manuel A. Rivas, Valentin Ruano-Rubio, Samuel A. Rose, Douglas M. Ruderfer, Khalid Shakir, Peter D. Stenson, Christine Stevens, Brett P. Thomas, Grace Tiao, Maria T. Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M. Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C. Florez, Stacey B. Gabriel, Gad Getz, Stephen J. Glatt, Christina M. Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M. Neale, Aarno Palotie, Shaun M. Purcell, Danish Saleheen, Jeremiah M. Scharf, Pamela Sklar, Patrick F. Sullivan, Jaakko Tuomilehto, Ming T. Tsuang, Hugh C. Watkins, James G. Wilson, Mark J. Daly, and Daniel G. MacArthur. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291.

[44] Peter D. Stenson, Edward V. Ball, Matthew Mort, Andrew D. Phillips, Jacqueline A. Shiel, Nick S. T. Thomas, Shaun Abeysinghe, Michael Krawczak, and David N. Cooper. Human gene mutation database (HGMD): 2003 update. *Human Mutation*, 21(6):577–581.

[45] Adrian Tan, Gonçalo R. Abecasis, and Hyun Min Kang. Unified representation of genetic variants. *Bioinformatics*, 31(13):2202–2204, .

[46] Vagheesh Narasimhan, Petr Danecek, Aylwyn Scally, Yali Xue, Chris Tyler-Smith, and Richard Durbin. BCFtools/RoH: a hidden markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, 32(11):1749–1751.

[47] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16): e164–e164, .

[48] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biology*, 17.

[49] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2):80–92.

[50] Davis J McCarthy, Peter Humburg, Alexander Kanapin, Manuel A Rivas, Kyle Gaulton, Jean-Baptiste Cazier, and Peter Donnelly. Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, 6(3):26.

[51] Xiaoming Liu, Chunlei Wu, Chang Li, and Eric Boerwinkle. dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Human Mutation*, 37(3):235–241.

[52] S. T. Sherry. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29 (1):308–311.

[53] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, 4(7): 1073–1081.

[54] Ivan Adzhubei, Daniel M. Jordan, and Shamil R. Sunyaev. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics*, 0 7: Unit7.20.

[55] Sung Chun and Justin C. Fay. Identification of deleterious mutations within three human genomes. *Genome Research*, 19(9):1553–1561.

[56] Jana Marie Schwarz, Christian Rödelsperger, Markus Schuelke, and Dominik Seelow. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods*, 7(8):575–6.

[57] Yongwook Choi, Gregory E. Sims, Sean Murphy, Jason R. Miller, and Agnes P. Chan. Predicting the functional effect of amino acid substitutions and indels. *PLOS ONE*, 7(10): e46688.

[58] Martin Kircher, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315.

[59] Daniel Quang, Yifei Chen, and Xiaohui Xie. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics (Oxford, England)*, 31(5): 761–763.

[60] Hashem A. Shihab, Julian Gough, Matthew Mort, David N. Cooper, Ian NM Day, and Tom R. Gaunt. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Human Genomics*, 8(1):11.

[61] Eugene V. Davydov, David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, 6(12):e1001025.

[62] Manuel Garber, Mitchell Guttman, Michele Clamp, Michael C. Zody, Nir Friedman, and Xiaohui Xie. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, 25(12):i54–i62.

[63] Najim Ameziane, Patrick May, Anneke Haitjema, Henri J. van de Vrugt, Sari E. van Rossum-Fikkert, Dejan Ristic, Gareth J. Williams, Jesper Balk, Davy Rockx, Hong Li, Martin A. Rooimans, Anneke B. Oostra, Eunike Velleuer, Ralf Dietrich, Onno B. Bleijerveld, A. F. Maarten Altelaar, Hanne Meijers-Heijboer, Hans Joenje, Gustavo Glusman, Jared Roach, Leroy Hood, David Galas, Claire Wyman, Rudi Balling, Johan den Dunnen, Johan P. de Winter, Roland Kanaar, Richard Gelinas, and Josephine C. Dorsman. A novel fanconi anaemia subtype associated with a dominant-negative mutation in *RAD51. Nature Communications*, 6:8829.

[64] Tarjinder Singh, Mitja I. Kurki, David Curtis, Shaun M. Purcell, Lucy Crooks, Jeremy McRae, Jaana Suvisaari, Himanshu Chheda, Douglas Blackwood, and Gerome Breen. Rare loss-of-function variants in SETD1a are associated with schizophrenia and developmental disorders. *Nature neuroscience*, 19(4):571.

[65] Dheeraj R. Bobbili, Dennis Lal, Patrick May, Eva M. Reinthaler, Kamel Jabbari, Holger Thiele, Michael Nothnagel, Wiktor Jurkowski, Martha Feucht, Peter Nürnberg, Holger Lerche, Fritz Zimprich, Roland Krause, Bernd A. Neubauer, Eva M. Reinthaler, Fritz Zimprich, Martha Feucht, Hannelore Steinböck, Birgit Neophytou, Julia Geldner, Ursula Gruber-Sedlmayr, Edda Haberlandt, Gabriel M. Ronen, Janine Altmüller, Dennis Lal, Peter Nürnberg, Thomas Sander, Holger Thiele, Roland Krause, Patrick May, Rudi Balling, Holger Lerche, and Bernd A. Neubauer. Exome-wide analysis of mutational burden in patients with typical and atypical rolandic epilepsy. *European Journal of Human Genetics*, page 1.

[66] Karen Buysse, Barbara Delle Chiaie, Rudy Van Coster, Bart Loeys, Anne De Paepe, Geert Mortier, Frank Speleman, and Björn Menten. Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *European Journal of Medical Genetics*, 52(6):398–403.

[67] Nigel P. Carter. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*, 39(7):S16–S21.

[68] Min Zhao, Qingguo Wang, Quan Wang, Peilin Jia, and Zhongming Zhao. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, 14:S1.

[69] Satoko Miyatake, Eriko Koshimizu, Atsushi Fujita, Ryoko Fukai, Eri Imagawa, Chihiro Ohba, Ichiro Kuki, Megumi Nukui, Atsushi Araki, Yoshio Makita, Tsutomu Ogata, Mitsuko Nakashima, Yoshinori Tsurusaki, Noriko Miyake, Hirotomo Saitsu, and Naomichi Matsumoto. Detecting copy-number variations in whole-exome sequencing data using the eXome hidden markov model: an 'exome-first' approach. *Journal of human genetics*, 60 (4):175–82.

[70] Menachem Fromer, Jennifer L Moran, Kimberly Chambert, Eric Banks, Sarah E Bergen, Douglas M Ruderfer, Robert E Handsaker, Steven A McCarroll, Michael C O'Donovan, Michael J Owen, George Kirov, Patrick F Sullivan, Christina M Hultman, Pamela Sklar, and Shaun M Purcell. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *American journal of human genetics*, 91(4):597–607.

[71] Jeffrey R MacDonald, Robert Ziman, Ryan K C Yuen, Lars Feuk, and Stephen W Scherer. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic acids research*, 42:D986–92.

[72] Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, Miriam K. Konkel, Ankit Malhotra, Adrian M. Stütz, Xinghua Shi, Francesco Paolo Casale, Jieming Chen, Fereydoun Hormozdiari, Gargi Dayama, Ken Chen, Maika Malig, Mark J. P. Chaisson, Klaudia Walter, Sascha Meiers, Seva Kashin, Erik Garrison, Adam Auton, Hugo Y. K. Lam, Xinmeng Jasmine Mu, Can Alkan, Danny Antaki, Taejeong Bae, Eliza Cerveira, Peter Chines, Zechen Chong, Laura Clarke, Elif Dal, Li Ding, Sarah Emery, Xian Fan, Madhusudan Gujral, Fatma Kahveci, Jeffrey M. Kidd, Yu Kong, Eric-Wubbo Lameijer, Shane McCarthy, Paul Flicek, Richard A. Gibbs, Gabor Marth, Christopher E. Mason, Androniki Menelaou, Donna M. Muzny, Bradley J. Nelson, Amina Noor, Nicholas F. Parrish, Matthew Pendleton, Andrew Quitadamo, Benjamin Raeder, Eric E. Schadt, Mallory Romanovitch, Andreas Schlattl, Robert Sebra, Andrey A. Shabalin, Andreas Untergasser, Jerilyn A. Walker, Min Wang, Fuli Yu, Chengsheng Zhang, Jing Zhang, Xiangqun Zheng-Bradley, Wanding Zhou, Thomas Zichner, Jonathan Sebat, Mark A. Batzer, Steven A. McCarroll, Ryan E. Mills, Mark B. Gerstein, Ali Bashir, Oliver Stegle, Scott E. Devine, Charles Lee, Evan E. Eichler, and Jan O. Korbel. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, .

[73] Detection and correction of artefacts in estimation of rare copy number variants and analysis of rare deletions in type 1 diabetes. *Human Molecular Genetics*, 24(6):1774–1790.

[74] Iuliana Ionita-Laza, Seunggeun Lee, Vlad Makarov, Joseph D. Buxbaum, and Xihong Lin. Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*, 92(6):841–853.

[75] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah a Nickerson, David C Christiani, Mark M Wurfel, and Xihong Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.*, 91(2):224–237, .

[76] Xiaowei Zhan, Youna Hu, Bingshan Li, Goncalo R. Abecasis, and Dajiang J. Liu. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics*, 32(9):1423–1426.

[77] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575, .

[78] Giulio Genovese, Menachem Fromer, Eli A. Stahl, Douglas M. Ruderfer, Kimberly Chambert, Mikael Landén, Jennifer L. Moran, Shaun M. Purcell, Pamela Sklar, Patrick F. Sullivan, Christina M. Hultman, and Steven A. McCarroll. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nature neuroscience*, 19(11):1433–1441.

[79] O. Tšuiko, M. Nõukas, O. Žilina, K. Hensen, J.S. Tapanainen, R. Mägi, M. Kals, P.A. Kivistik, K. Haller-Kikkatalo, A. Salumets, and A. Kurg. Copy number variation analysis detects novel candidate genes involved in follicular growth and oocyte maturation in a cohort of premature ovarian failure cases. *Human Reproduction (Oxford, England)*, 31(8): 1913–1925.

[80] Monica DiLuca and Jes Olesen. The cost of brain diseases: A burden or a challenge? *Neuron*, 82(6):1205–1208.

[81] Guangchun Han, Jiya Sun, Jiajia Wang, Zhouxian Bai, Fuhai Song, and Hongxing Lei. Genomics in neurological disorders. *Genomics, Proteomics & Bioinformatics*, 12(4):156–163.

[82] GBD 2015 Neurological Disorders Collaborator Group. Global, regional, and national burden of neurological disorders during 1990-2015: a systematic analysis for the global burden of disease study 2015. *The Lancet. Neurology*, 16(11):877–897.

[83] Emanuela Zagni, Lucia Simoni, and Delia Colombo. Sex and gender differences in central nervous system-related disorders. *Neuroscience Journal*, 2016.

[84] Clifton L. Gooch, Etienne Pracht, and Amy R. Borenstein. The burden of neurological disease in the united states: A summary report and call to action. *Annals of Neurology*, 81(4):479–484.

[85] Michael H. Bloch and James F. Leckman. Clinical course of tourette syndrome. *Journal of psychosomatic research*, 67(6):497–501.

[86] Jia-Nee Foo, Jian-Jun Liu, and Eng-King Tan. Whole-genome and whole-exome sequencing in neurological diseases. *Nature Reviews Neurology*, 8(9):508–517.

[87] Audrey S. Dickey and Albert R. La Spada. Therapy development in huntington disease: From current strategies to emerging opportunities. *American Journal of Medical Genetics. Part A*.

[88] Mohammad Faghihi, Salim Mottagui-Tabar, and Claes Wahlestedt. *Genetics of neurological disorders*, volume 4.

[89] Thomas Werner. Next generation sequencing in functional genomics. *Briefings in Bioinformatics*, 11(5):499–511.

[90] Amy Reeve, Eve Simcox, and Doug Turnbull. Ageing and parkinson's disease: Why is advancing age the biggest risk factor? *Ageing Research Reviews*, 14:19–30.

[91] Shey-Lin Wu, Rajka M. Liscic, SangYun Kim, Sandro Sorbi, and Yuan-Han Yang. Nonmotor symptoms of parkinson's disease. *Parkinson's Disease*, 2017.

[92] Chou Chai and Kah-Leong Lim. Genetic insights into sporadic parkinson's disease pathogenesis. *Current Genomics*, 14(8):486–501.

[93] Glenda E. Gillies, Ilse S. Pienaar, Shiv Vohra, and Zahi Qamhawi. Sex differences in parkinson's disease. *Frontiers in Neuroendocrinology*, 35(3):370–384.

[94] Leah R. Miller, Shyamali Mukherjee, Twum A. Ansah, and Salil K. Das. Cigarette smoke and dopaminergic system. *Journal of Biochemical and Molecular Toxicology*, 21(6):325–335.

[95] F D Dick, G De Palma, A Ahmadi, N W Scott, G J Prescott, J Bennett, S Semple, S Dick, C Counsell, P Mozzoni, N Haites, S Bezzina Wettinger, A Mutti, M Otelea, A Seaton, P Söderkvist, and A Felice. Environmental risk factors for parkinson's disease and parkinsonism: the geoparkinson study. *Occupational and Environmental Medicine*, 64(10):666–672.

[96] J. W. Langston, P. Ballard, J. W. Tetrud, and I. Irwin. Chronic parkinsonism in humans due to a product of meperidine-analog synthesis. *Science (New York, N.Y.)*, 219(4587): 979–980.

[97] Samuel M. Goldman. Environmental toxins and parkinson's disease. *Annual review of pharmacology and toxicology*, 54:141–164.

[98] Georgia Xiromerisiou, Henry Houlden, Anna Sailer, Laura Silveira-Moriyama, John Hardy, and Andrew J. Lees. Identical twins with leucine rich repeat kinase type 2 mutations discordant for parkinson's disease. *Movement Disorders: Official Journal of the Movement Disorder Society*, 27(10):1323.

[99] D. Dickson, M. Farrer, S. Lincoln, R. P. Mason, T. R. Zimmerman, L. I. Golbe, and J. Hardy. Pathology of PD in monozygotic twins with a 20-year discordance interval. *Neurology*, 56(7):981–982.

[100] C. M. Tanner, R. Ottman, S. M. Goldman, J. Ellenberg, P. Chan, R. Mayeux, and J. W. Langston. Parkinson disease in twins: an etiologic study. *JAMA*, 281(4):341–346.

[101] D. J. Burn, M. H. Mark, E. D. Playford, D. M. Maraganore, T. R. Zimmerman, R. C. Duvoisin, A. E. Harding, C. D. Marsden, and D. J. Brooks. Parkinson's disease in twins studied with 18f-dopa and positron emission tomography. *Neurology*, 42(10):1894–1900.

[102] A. Laihinen, H. Ruottinen, J. O. Rinne, M. Haaparanta, J. Bergman, O. Solin, M. Koskenvuo, R. Marttila, and U. K. Rinne. Risk for parkinson's disease: twin studies for the detection of asymptomatic subjects using [18f]6-fluorodopa PET. *Journal of Neurology*, 247(2):110–113.

[103] David J. Brooks. Neuroimaging in parkinson's disease. *NeuroRx*, 1(2):243–254.

[104] M. H. Polymeropoulos, C. Lavedan, E. Leroy, S. E. Ide, A. Dehejia, A. Dutra, B. Pike, H. Root, J. Rubenstein, R. Boyer, E. S. Stenroos, S. Chandrasekharappa, A. Athanassiadou, T. Papapetropoulos, W. G. Johnson, A. M. Lazzarini, R. C. Duvoisin, G. Di Iorio, L. I. Golbe, and R. L. Nussbaum. Mutation in the alpha-synuclein gene identified in families with parkinson's disease. *Science (New York, N.Y.)*, 276(5321):2045–2047.

[105] Chao-Dong Wang and Piu Chan. Clinicogenetics of parkinson s disease: A drawing but not completed picture. *Neuroimmunology and Neuroinflammation*, 1(3):115.

[106] Michelle K. Lin and Matthew J. Farrer. Genetics and genomics of parkinson's disease. *Genome Medicine*, 6:48.

[107] Taku Hatano, Manabu Funayama, Shin-ichiro Kubo, Ignacio F. Mata, Yutaka Oji, Akio Mori, Cyrus P. Zabetian, Sarah M. Waldherr, Hiroyo Yoshino, Genko Oyama, Yasushi Shimo, Ken-ichi Fujimoto, Hirokazu Oshima, Yasuto Kunii, Hirooki Yabe, Yoshikuni Mizuno, and Nobutaka Hattori. Identification of a japanese family with LRRK2 p.r1441g-related parkinson's disease. *Neurobiology of Aging*, 35(11):2656.e17–2656.e23.

[108] Andreas Puschmann. New genes causing hereditary parkinson's disease or parkinsonism. *Current Neurology and Neuroscience Reports*, 17(9):66.

[109] Diana Chang, Mike A. Nalls, Ingileif B. Hallgrímsdóttir, Julie Hunkapiller, Marcel van der Brug, Fang Cai, International Parkinson's Disease Genomics Consortium, 23andMe Research Team, Geoffrey A. Kerchner, Gai Ayalon, Baris Bingol, Morgan Sheng, David Hinds, Timothy W. Behrens, Andrew B. Singleton, Tushar R. Bhangale, and Robert R. Graham. A meta-analysis of genome-wide association studies identifies 17 new parkinson's disease risk loci. *Nature Genetics*, 49(10):1511–1516, .

[110] Christine Klein and Ana Westenberger. Genetics of parkinson's disease. *Cold Spring Harbor Perspectives in Medicine*, 2(1).

[111] Joanne Trinh, Emil K. Gustavsson, Carles Vilariño-Güell, Stephanie Bortnick, Jeanne Latourelle, Marna B. McKenzie, Chelsea Szu Tu, Ekaterina Nosova, Jaskaran Khinda, Austen Milnerwood, Suzanne Lesage, Alexis Brice, Meriem Tazir, Jan O. Aasly, Laura Parkkinen, Hazal Haytural, Tatiana Foroud, Richard H. Myers, Samia Ben Sassi, Emna Hentati, Fatma Nabli, Emna Farhat, Rim Amouri, Fayçal Hentati, and Matthew J. Farrer. DNM3 and genetic modifiers of age of onset in LRRK2 gly2019ser parkinsonism: a genome-wide linkage and association study. *The Lancet Neurology*, 15(12):1248–1256.

[112] A. B. Singleton, M. Farrer, J. Johnson, A. Singleton, S. Hague, J. Kachergus, M. Hulihan, T. Peuralinna, A. Dutra, and R. Nussbaum. alpha-synuclein locus triplication causes parkinson's disease. *Science*, 302(5646):841–841.

[113] Valentina La Cognata, Velia D'Agata, and Francesca Cavalcanti {and} Sebastiano Cavallaro. Genetics of parkinson's disease: The role of copy number variations.

[114] Valentina La Cognata, Giovanna Morello, Velia D'Agata, and Sebastiano Cavallaro. Copy number variability in parkinson's disease: assembling the puzzle through a systems biology approach. *Human Genetics*, 136(1):13–37, .

[115] Nathan Pankratz, Alexandra Dumitriu, Kurt N. Hetrick, Mei Sun, Jeanne C. Latourelle, Jemma B. Wilk, Cheryl Halter, Kimberly F. Doheny, James F. Gusella, William C. Nichols, Richard H. Myers, Tatiana Foroud, Anita L. DeStefano, the PSG-PROGENI

{and} GenePD Investigators, and Coordinators {and} Molecular Genetic Laboratories. Copy number variation in familial parkinson disease. *PLOS ONE*, 6(8):e20988, .

[116] N. Pankratz, D. K. Kissell, M. W. Pauciulo, C. A. Halter, A. Rudolph, R. F. Pfeiffer, K. S. Marder, T. Foroud, W. C. Nichols, and Parkinson Study Group-PROGENI Investigators. Parkin dosage mutations have greater pathogenicity in familial PD than simple sequence mutations. *Neurology*, 73(4):279–286, .

[117] Rowena J. Keyser, Debbie Lombard, Rene Veikondis, Jonathan Carr, and Soraya Bardien. Analysis of exon dosage using MLPA in south african parkinson's disease patients. *Neurogenetics*, 11(3):305–312.

[118] D. M. Kay, C. F. Stevens, T. H. Hamza, J. S. Montimurro, C. P. Zabetian, S. A. Factor, A. Samii, A. Griffith, J. W. Roberts, and E. S. Molho. A comprehensive analysis of deletions, multiplications, and copy number variations in PARK2. *Neurology*, 75(13):1189–1194.

[119] Roberta Marongiu, Francesco Brancati, Angelo Antonini, Tamara Ialongo, Caterina Ceccarini, Oronzo Scarciolla, Anna Capalbo, Riccardo Benti, Gianni Pezzoli, and Bruno Dallapiccola. Whole gene deletion and splicing mutations expand the PINK1 genotypic spectrum. *Human mutation*, 28(1):98–98.

[120] Ji-feng Guo, Xue-wei Zhang, Li-luo Nie, Hai-nan Zhang, Bin Liao, Jing Li, Lei Wang, Xin-xiang Yan, and Bei-sha Tang. Mutation analysis of parkin, PINK1 and DJ-1 genes in chinese patients with sporadic early onset parkinsonism. *Journal of neurology*, 257(7):1170–1175.

[121] Lluís Samaranch, Oswaldo Lorenzo-Betancor, José M. Arbelo, Isidre Ferrer, Elena Lorenzo, Jaione Irigoyen, Maria A. Pastor, Carmen Marrero, Concepción Isla, Joanna Herrera-Henriquez, and Pau Pastor. PINK1-linked parkinsonism is associated with lewy body pathology. *Brain: A Journal of Neurology*, 133:1128–1142.

[122] V. Bonifati, P. Rizzu, F. Squitieri, E. Krieger, N. Vanacore, J. C. van Swieten, A. Brice, C. M. van Duijn, B. Oostra, G. Meco, and P. Heutink. DJ-1(PARK7), a novel gene for autosomal recessive, early onset parkinsonism. *Neurological Sciences: Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, 24 (3):159–160.

[123] C. M. Van Duijn, M. C. J. Dekker, V. Bonifati, R. J. Galjaard, J. J. Houwing-Duistermaat, PJLM Snijders, L. Testers, G. J. Breedveld, MWIM Horstink, and L. A. Sandkuijl. Park7, a novel locus for autosomal recessive early-onset parkinsonism, on chromosome 1p36. *The American Journal of Human Genetics*, 69(3):629–634.

[124] Hossein Darvish, Abolfazl Movafagh, Mir Davood Omrani, Saghar Ghasemi Firouzabadi, Eznollah Azargashb, Javad Jamshidi, Ali Khaligh, Leyla Haghnejad, Nilofar Safavi Naeini, and Atefeh Talebi. Detection of copy number changes in genes associated with parkinson's disease in iranian patients. *Neuroscience letters*, 551:75–78.

[125] Edward Faught, Joshua Richman, Roy Martin, Ellen Funkhouser, R Foushee, Polly Kratt, Yongin Kim, Kay Clements, N Cohen, D Adoboe, Robert Knowlton, and Maria Pisu. Incidence and prevalence of epilepsy among older US medicare beneficiaries. *Neurology*, 78: 448–53.

[126] Hirak Kumar Mukhopadhyay, Chandi Charan Kandar, Sanjoy Das, Lakshmi Kanta Ghosh, and Bijan Gupta. Epilepsy and its management: A review. *Journal of PharmSciTech*, 1: 20–26.

[127] Julia Oyrer, Snezana Maljevic, Ingrid E. Scheffer, Samuel F. Berkovic, Steven Petrou, and Christopher A. Reid. Ion channels in genetic epilepsy: From genes and mechanisms to disease-targeted therapies. *Pharmacological Reviews*, 70(1):142–173.

[128] Robert S. Fisher, Carlos Acevedo, Alexis Arzimanoglou, Alicia Bogacz, J. Helen Cross, Christian E. Elger, Jerome Engel, Lars Forsgren, Jacqueline A. French, Mike Glynn, Dale C. Hesdorffer, B.i. Lee, Gary W. Mathern, Solomon L. Moshé, Emilio Perucca, Ingrid E. Scheffer, Torbjörn Tomson, Masako Watanabe, and Samuel Wiebe. ILAE official report: A practical clinical definition of epilepsy. *Epilepsia*, 55(4):475–482, .

[129] C. Behr, M. A. Goltzene, G. Kosmalski, E. Hirsch, and P. Ryvlin. Epidemiology of epilepsy. *Revue Neurologique*, 172(1):27–36.

[130] Robert S. Fisher, J. Helen Cross, Carol D'Souza, Jacqueline A. French, Sheryl R. Haut, Norimichi Higurashi, Edouard Hirsch, Floor E. Jansen, Lieven Lagae, Solomon L. Moshé, Jukka Peltola, Eliane Roulet Perez, Ingrid E. Scheffer, Andreas Schulze-Bonhage, Ernest Somerville, Michael Sperling, Elza Márcia Yacubian, and Sameer M. Zuberi. Instruction manual for the ILAE 2017 operational classification of seizure types. *Epilepsia*, 58(4): 531–542, .

[131] Michael Steffens, Costin Leu, Ann-Kathrin Ruppert, Federico Zara, Pasquale Striano, Angela Robbiano, Giuseppe Capovilla, Paolo Tinuper, Antonio Gambardella, Amedeo Bianchi, Angela La Neve, Giovanni Crichiutti, De Kovel, Carolien G.f, Dorothée Kasteleijn-Nolst Trenité, Gerrit-Jan de Haan, Dick Lindhout, Verena Gaus, Bettina Schmitz, Dieter Janz, Yvonne G. Weber, Felicitas Becker, Holger Lerche, Bernhard J. Steinhoff, Ailing A. Kleefuß-Lie, Wolfram S. Kunz, Rainer Surges, Christian E. Elger, Hiltrud Muhle, Sarah von Spiczak, Philipp Ostertag, Ingo Helbig, Ulrich Stephani, Rikke S. Møller, Helle Hjalgrim, Leanne M. Dibbens, Susannah Bellows, Karen Oliver, Saul Mullen, Ingrid E. Scheffer,

Samuel F. Berkovic, Kate V. Everett, Mark R. Gardiner, Carla Marini, Renzo Guerrini, Anna-Elina Lehesjoki, Auli Siren, Michel Guipponi, Alain Malafosse, Pierre Thomas, Rima Nabbout, Stephanie Baulac, Eric Leguern, Rosa Guerrero, Jose M. Serratosa, Philipp S. Reif, Felix Rosenow, Martina Mörzinger, Martha Feucht, Fritz Zimprich, Claudia Kapser, Christoph J. Schankin, Arvid Suls, Katrin Smets, Peter De Jonghe, Albena Jordanova, Hande Caglayan, Zuhal Yapici, Destina A. Yalcin, Betul Baykan, Nerses Bebek, Ugur Ozbek, Christian Gieger, Heinz-Erich Wichmann, Tobias Balschun, David Ellinghaus, Andre Franke, Christian Meesters, Tim Becker, Thomas F. Wienker, Anne Hempelmann, Herbert Schulz, Franz Rüschendorf, Markus Leber, Steffen M. Pauck, Holger Trucks, Mohammad R. Toliat, Peter Nürnberg, Giuliano Avanzini, Bobby P. C. Koeleman, and Thomas Sander. Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. *Human Molecular Genetics*, 21(24): 5359–5372.

[132] Dalia Kasperaviciūte, Claudia B Catarino, Erin L Heinzen, Chantal Depondt, Gianpiero L Cavalleri, Luis O Caboclo, Sarah K Tate, Jenny Jamnadas-Khoda, Krishna Chinthapalli, Lisa M S Clayton, Kevin V Shianna, Rodney a Radtke, Mohamad a Mikati, William B Gallentine, Aatif M Husain, Saud Alhusaini, David Leppert, Lefkos T Middleton, Rachel a Gibson, Michael R Johnson, Paul M Matthews, David Hosford, Kjell Heuser, Leslie Amos, Marcos Ortega, Dominik Zumsteg, Heinz-Gregor Wieser, Bernhard J Steinhoff, Günter Krämer, Jörg Hansen, Thomas Dorn, Anne-Mari Kantanen, Leif Gjerstad, Terhi Peuralinna, Dena G Hernandez, Kai J Eriksson, Reetta K Kälviäinen, Colin P Doherty, Nicholas W Wood, Massimo Pandolfo, John S Duncan, Josemir W Sander, Norman Delanty, David B Goldstein, and Sanjay M Sisodiya. Common genetic variation and susceptibility to partial epilepsies: a genome-wide association study. *Brain*, 133:2136–2147.

[133] Epi4K consortium and Epilepsy Phenome/Genome Project. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. *The Lancet. Neurology*, 16(2):135–143.

[134] JA Armijo, M Shushtarian, Elsa Valdizán, Antonio Cuadrado, Isabel De las Cuevas, and J Adín. *Ion Channels and Epilepsy*, volume 11.

[135] Candace T. Myers and Heather C. Mefford. Advancing epilepsy genetics in the genomic era. *Genome Medicine*, 7(1).

[136] Ortrud K. Steinlein. Genetics and epilepsy. *Dialogues in Clinical Neuroscience*, 10(1): 29–38.

[137] S. F. Berkovic, I. E. Scheffer, S. Petrou, N. Delanty, T. J. Dixon-Salazar, D. J. Dlugos, I. Helbig, W. N. Frankel, D. B. Goldstein, E. L. Heinzen, D. H. Lowenstein, H. C. Mefford,

J. M. Parent, A. Poduri, and S. F. Traynelis. A roadmap for precision medicine in the epilepsies. *The Lancet Neurology*, 14(12):1219–1228, .

[138] Heather C. Mefford, Simone C. Yendle, Cynthia Hsu, Joseph Cook, Eileen Geraghty, Jacinta M. McMahon, Orvar Eeg-Olofsson, Lynette G. Sadleir, Deepak Gill, Bruria Ben-Zeev, Tally Lerman-Sagie, Mark Mackay, Jeremy L. Freeman, Eva Andermann, James T. Pelakanos, Ian Andrews, Geoffrey Wallace, Evan E. Eichler, Samuel F. Berkovic, and Ingrid E. Scheffer. Rare copy number variants are an important cause of epileptic encephalopathies. *Annals of Neurology*, 70(6):974–985.

[139] Patrick Cossette, Lidong Liu, Katéri Brisebois, Haiheng Dong, Anne Lortie, Michel Vanasse, Jean-Marc Saint-Hilaire, Lionel Carmant, Andrei Verner, Wei-Yang Lu, Yu Tian Wang, and Guy Rouleau. Mutation of GABRA1 in an autosomal dominant form of juvenile myoclonic epilepsy. *Nature genetics*, 31:184–9.

[140] R. H. Wallace, C. Marini, S. Petrou, L. A. Harkin, D. N. Bowser, R. G. Panchal, D. A. Williams, G. R. Sutherland, J. C. Mulley, I. E. Scheffer, and S. F. Berkovic. Mutant GABA(a) receptor gamma2-subunit in childhood absence epilepsy and febrile seizures. *Nature Genetics*, 28(1):49–52.

[141] Tian Chen, Mohan Giri, Zhenyi Xia, Yadu Nanda Subedi, and Yan Li. Genetic and epigenetic mechanisms of epilepsy: a review. *Neuropsychiatric disease and treatment*, 13: 1841, .

[142] Lata Vadlamudi, Marianne J. Kjeldsen, Linda A. Corey, Marit H. Solaas, Mogen L. Friis, John M. Pellock, Karl O. Nakken, Roger L. Milne, Ingrid E. Scheffer, A. Simon Harvey, John L. Hopper, and Samuel F. Berkovic. Analyzing the etiology of benign rolandic epilepsy: a multicenter twin collaboration. *Epilepsia*, 47(3):550–555, .

[143] Lata Vadlamudi, A. Simon Harvey, Mary M. Connellan, Roger L. Milne, John L. Hopper, Ingrid E. Scheffer, and Samuel F. Berkovic. Is benign rolandic epilepsy genetically determined? *Annals of Neurology*, 56(1):129–132, .

[144] Weixi Xiong and Dong Zhou. Progress in unraveling the genetic etiology of rolandic epilepsy. *Seizure*, 47:99–104.

[145] Lisa J Strug, Tara Clarke, Theodore Chiang, Minchen Chien, Zeynep Baskurt, Weili Li, Ruslan Dorfman, Bhavna Bali, Elaine Wirrell, Steven L Kugler, David E Mandelbaum, Steven M Wolf, Patricia McGoldrick, Huntley Hardison, Edward J Novotny, Jingyue Ju, David a Greenberg, James J Russo, and Deb K Pal. Centrotemporal sharp wave EEG trait in rolandic epilepsy maps to elongator protein complex 4 (ELP4). *Eur. J. Hum. Genet.*, 17(9):1171–1181.

[146] Eva M. Reinthaler, Dennis Lal, Wiktor Jurkowski, Martha Feucht, Hannelore Steinböck, Ursula Gruber-Sedlmayr, Gabriel M. Ronen, Julia Geldner, Edda Haberlandt, Birgit Neophytou, Andreas Hahn, Janine Altmüller, Holger Thiele, Mohammad R. Toliat, EuroEPINOMICS Consortium, Holger Lerche, Peter Nürnberg, Thomas Sander, Bernd A. Neubauer, and Fritz Zimprich. Analysis of ELP4, SRPX2, and interacting genes in typical and atypical rolandic epilepsy. *Epilepsia*, 55(8):e89–93, .

[147] Gemma L. Carvill, Brigid M. Regan, Simone C. Yendle, Brian J. O'Roak, Natalia Lozovaya, Nadine Bruneau, Nail Burnashev, Adiba Khan, Joseph Cook, Eileen Geraghty, Lynette G. Sadleir, Samantha J. Turner, Meng-Han Tsai, Richard Webster, Robert Ouvrier, John A. Damiano, Samuel F. Berkovic, Jay Shendure, Michael S. Hildebrand, Pierre Szepetowski, Ingrid E. Scheffer, and Heather C. Mefford. GRIN2a mutations cause epilepsy-aphasia spectrum disorders. *Nature Genetics*, 45(9):1073–1076.

[148] Mutations in GRIN2a cause idiopathic focal epilepsy with rolandic spikes. *Nature Genetics*, 45(9):1067–1072.

[149] Eva M. Reinthaler, Borislav Dejanovic, Dennis Lal, Marcus Semtner, Yvonne Merkler, Annika Reinhold, Dorothea A. Pittrich, Christoph Hotzy, Martha Feucht, Hannelore Steinböck, Ursula Gruber-Sedlmayr, Gabriel M. Ronen, Birgit Neophytou, Julia Geldner, Edda Haberlandt, Hiltrud Muhle, M. Arfan Ikram, Cornelia M. van Duijn, Andre G. Uitterlinden, Albert Hofman, Janine Altmüller, Amit Kawalia, Mohammad R. Toliat, EuroEPINOMICS Consortium, Peter Nürnberg, Holger Lerche, Michael Nothnagel, Holger Thiele, Thomas Sander, Jochen C. Meier, Günter Schwarz, Bernd A. Neubauer, and Fritz Zimprich. Rare variants in gamma-aminobutyric acid type a receptor genes in rolandic epilepsy and related syndromes. *Annals of Neurology*, 77(6):972–986, .

[150] C. P. Panayiotopoulos. *Idiopathic Generalised Epilepsies*. Bladon Medical Publishing.

[151] D. K. V. Prasad, U. Satyanarayana, and Anjana Munshi. Genetics of idiopathic generalized epilepsy: An overview. *Neurology India*, 61(6):572.

[152] Yvonne G. Weber and Holger Lerche. Genetic mechanisms in idiopathic epilepsies. *Developmental Medicine & Child Neurology*, 50(9):648–654.

[153] Francesco Nicita, Paola De Liso, Federica Rachele Danti, Laura Papetti, Fabiana Ursitti, Antonella Castronovo, Federico Allemand, Elena Gennaro, Federico Zara, Pasquale Striano, and Alberto Spalice. The genetics of monogenic idiopathic epilepsies and epileptic encephalopathies. *Seizure*, 21(1):3–11, .

[154] Costin Leu, Carolien de Kovel, Federico Zara, Pasquale Striano, Marianna Pezzella, Angela Robbiano, Amedeo Bianchi, Francesca Bisulli, Antonietta Coppola, Anna Giallonardo,

Francesca Beccaria, Dorothee Kasteleijn, Dick Lindhout, Verena Gaus, Bettina Schmitz, Dieter Janz, Yvonne G Weber, Felicitas Becker, Holger Lerche, and Thomas Sander. Genome-wide linkage meta-analysis identifies susceptibility loci at 2q34 and 13q31.3 for genetic generalized epilepsies. *Epilepsia*, 53:308–18, .

[155] EPICURE Consortium, EMINet Consortium, Michael Steffens, Costin Leu, Ann-Kathrin Ruppert, Federico Zara, Pasquale Striano, Angela Robbiano, Giuseppe Capovilla, Paolo Tinuper, Antonio Gambardella, Amedeo Bianchi, Angela La Neve, Giovanni Crichiutti, Carolien G. F. de Kovel, Dorothée Kasteleijn-Nolst Trenité, Gerrit-Jan de Haan, Dick Lindhout, Verena Gaus, Bettina Schmitz, Dieter Janz, Yvonne G. Weber, Felicitas Becker, Holger Lerche, Bernhard J. Steinhoff, Ailing A. Kleefuß-Lie, Wolfram S. Kunz, Rainer Surges, Christian E. Elger, Hiltrud Muhle, Sarah von Spiczak, Philipp Ostertag, Ingo Helbig, Ulrich Stephani, Rikke S. Møller, Helle Hjalgrim, Leanne M. Dibbens, Susannah Bellows, Karen Oliver, Saul Mullen, Ingrid E. Scheffer, Samuel F. Berkovic, Kate V. Everett, Mark R. Gardiner, Carla Marini, Renzo Guerrini, Anna-Elina Lehesjoki, Auli Siren, Michel Guipponi, Alain Malafosse, Pierre Thomas, Rima Nabbout, Stephanie Baulac, Eric Leguern, Rosa Guerrero, Jose M. Serratosa, Philipp S. Reif, Felix Rosenow, Martina Mörzinger, Martha Feucht, Fritz Zimprich, Claudia Kapser, Christoph J. Schankin, Arvid Suls, Katrin Smets, Peter De Jonghe, Albena Jordanova, Hande Caglayan, Zuhal Yapici, Destina A. Yalcin, Betul Baykan, Nerses Bebek, Ugur Ozbek, Christian Gieger, Heinz-Erich Wichmann, Tobias Balschun, David Ellinghaus, Andre Franke, Christian Meesters, Tim Becker, Thomas F. Wienker, Anne Hempelmann, Herbert Schulz, Franz Rüschendorf, Markus Leber, Steffen M. Pauck, Holger Trucks, Mohammad R. Toliat, Peter Nürnberg, Giuliano Avanzini, Bobby P. C. Koeleman, and Thomas Sander. Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. *Human Molecular Genetics*, 21(24):5359–5372.

[156] International League Against Epilepsy Consortium on Complex Epilepsies. Electronic address: epilepsy-austin@unimelb.edu.au. Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *The Lancet. Neurology*, 13(9):893–903.

[157] Erin L Heinzen, Chantal Depondt, Gianpiero L Cavalleri, Elizabeth K Ruzzo, Nicole M Walley, Anna C Need, Dongliang Ge, Min He, Elizabeth T Cirulli, Qian Zhao, Kenneth D Cronin, Curtis E Gumbs, C Ryan Campbell, Linda K Hong, Jessica M Maia, Kevin V Shianna, Mark McCormack, Rodney A Radtke, Gerard D O'Conner, Mohamad A Mikati, William B Gallentine, Aatif M Husain, Saurabh R Sinha, Krishna Chinthapalli, Ram S Puranam, James O McNamara, Ruth Ottman, Sanjay M Sisodiya, Norman Delanty, and David B Goldstein. Exome sequencing followed by large-scale genotyping fails to identify

single rare variants of large effect in idiopathic generalized epilepsy. *American journal of human genetics*, 91(2):293–302.

[158] Ingo Helbig, Heather C Mefford, Andrew J Sharp, Michel Guipponi, Marco Fichera, Andre Franke, Hiltrud Muhle, Carolien de Kovel, Carl Baker, Sarah von Spiczak, Katherine L Kron, Ines Steinich, Ailing a Kleefuss-Lie, Costin Leu, Verena Gaus, Bettina Schmitz, Karl M Klein, Philipp S Reif, Felix Rosenow, Yvonne Weber, Holger Lerche, Fritz Zimprich, Lydia Urak, Karoline Fuchs, Martha Feucht, Pierre Genton, Pierre Thomas, Frank Visscher, Gerrit-Jan de Haan, Rikke S Møller, Helle Hjalgrim, Daniela Luciano, Michael Wittig, Michael Nothnagel, Christian E Elger, Peter Nürnberg, Corrado Romano, Alain Malafosse, Bobby P C Koeleman, Dick Lindhout, Ulrich Stephani, Stefan Schreiber, Evan E Eichler, and Thomas Sander. 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat. Genet.*, 41(2):160–162, .

[159] Bobby P. C. Koeleman. What do genetic studies tell us about the heritable basis of common epilepsy? polygenic or complex epilepsy? *Neuroscience Letters*.

[160] Carolien G F de Kovel, Holger Trucks, Ingo Helbig, Heather C Mefford, Carl Baker, Costin Leu, Christian Kluck, Hiltrud Muhle, Sarah von Spiczak, Philipp Ostertag, Tanja Obermeier, Ailing a Kleefuss-Lie, Kerstin Hallmann, Michael Steffens, Verena Gaus, Karl M Klein, Hajo M Hamer, Felix Rosenow, Eva H Brilstra, Dorothée Kasteleijn-Nolst Trenité, Marielle E M Swinkels, Yvonne G Weber, Iris Unterberger, Fritz Zimprich, Lydia Urak, Martha Feucht, Karoline Fuchs, Rikke S Møller, Helle Hjalgrim, Peter De Jonghe, Arvid Suls, Ina-Maria Rückert, Heinz-Erich Wichmann, Andre Franke, Stefan Schreiber, Peter Nürnberg, Christian E Elger, Holger Lerche, Ulrich Stephani, Bobby P C Koeleman, Dick Lindhout, Evan E Eichler, and Thomas Sander. Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain*, 133:23–32.

[161] Dennis Lal, Eva M. Reinthaler, Janine Altmüller, Mohammad R. Toliat, Holger Thiele, Peter Nürnberg, Holger Lerche, Andreas Hahn, Rikke S. Møller, Hiltrud Muhle, Thomas Sander, Fritz Zimprich, and Bernd A. Neubauer. RBFOX1 and RBFOX3 mutations in rolandic epilepsy. *PLOS ONE*, 8(9):e73323, .

[162] Natalio Fejerman. Atypical rolandic epilepsy. *Epilepsia*, 50 Suppl 7:9–12.

[163] Dennis Lal, Eva M. Reinthaler, Julian Schubert, Hiltrud Muhle, Erik Riesch, Gerhard Kluger, Kamel Jabbari, Amit Kawalia, Christine Bäumel, Hans Holthausen, Andreas Hahn, Martha Feucht, Birgit Neophytou, Edda Haberlandt, Felicitas Becker, Janine Altmüller, Holger Thiele, Johannes R. Lemke, Holger Lerche, Peter Nürnberg, Thomas Sander, Yvonne Weber, Fritz Zimprich, and Bernd A. Neubauer. DEPDC5 mutations in genetic focal epilepsies of childhood. *Annals of Neurology*, 75(5):788–792, .

[164] Silvia De Rubeis, Xin He, Arthur P. Goldberg, Christopher S. Poultney, Kaitlin Samocha, A. Ercument Cicek, Yan Kou, Li Liu, Menachem Fromer, Susan Walker, Tarjinder Singh, Lambertus Klei, Jack Kosmicki, Shih-Chen Fu, Branko Aleksic, Monica Biscaldi, Patrick F. Bolton, Jessica M. Brownfeld, Jinlu Cai, Nicholas G. Campbell, Angel Carracedo, Maria H. Chahrour, Andreas G. Chiocchetti, Hilary Coon, Emily L. Crawford, Lucy Crooks, Sarah R. Curran, Geraldine Dawson, Eftichia Duketis, Bridget A. Fernandez, Louise Gallagher, Evan Geller, Stephen J. Guter, R. Sean Hill, Iuliana Ionita-Laza, Patricia Jimenez Gonzalez, Helena Kilpinen, Sabine M. Klauck, Alexander Kolevzon, Irene Lee, Jing Lei, Terho Lehtimäki, Chiao-Feng Lin, Avi Ma'ayan, Christian R. Marshall, Alison L. McInnes, Benjamin Neale, Michael J. Owen, Norio Ozaki, Mara Parellada, Jeremy R. Parr, Shaun Purcell, Kaija Puura, Deepthi Rajagopalan, Karola Rehnström, Abraham Reichenberg, Aniko Sabo, Michael Sachse, Stephan J. Sanders, Chad Schafer, Martin Schulte-Rüther, David Skuse, Christine Stevens, Peter Szatmari, Kristiina Tammimies, Otto Valladares, Annette Voran, Li-San Wang, Lauren A. Weiss, A. Jeremy Willsey, Timothy W. Yu, Ryan K. C. Yuen, Edwin H. Cook, Christine M. Freitag, Michael Gill, Christina M. Hultman, Thomas Lehner, Aarno Palotie, Gerard D. Schellenberg, Pamela Sklar, Matthew W. State, James S. Sutcliffe, Christopher A. Walsh, Stephen W. Scherer, Michael E. Zwick, Jeffrey C. Barrett, David J. Cutler, Kathryn Roeder, Bernie Devlin, Mark J. Daly, and Joseph D. Buxbaum. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515 (7526):209–15.

[165] Hui Yang and Kai Wang. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature Protocols*, 10(10):1556–1566.

[166] Steffen Syrbe, Ulrike B. S. Hedrich, Erik Riesch, Tania Djémié, Stephan Müller, Rikke S. Møller, Bridget Maher, Laura Hernandez-Hernandez, Matthis Synofzik, Hande S. Caglayan, Mutluay Arslan, José M. Serratosa, Michael Nothnagel, Patrick May, Roland Krause, Heidrun Löffler, Katja Detert, Thomas Dorn, Heinrich Vogt, Günter Krämer, Ludger Schöls, Primus E. Mullis, Tarja Linnankivi, Anna-Elina Lehesjoki, Katalin Sterbova, Dana C. Craiu, Dorota Hoffman-Zacharska, Christian M. Korff, Yvonne G. Weber, Maja Steinlin, Sabina Gallati, Astrid Bertsche, Matthias K. Bernhard, Andreas Merkenschlager, Wieland Kiess, EuroEPINOMICS Res, Michael Gonzalez, Stephan Züchner, Aarno Palotie, Arvid Suls, Peter De Jonghe, Ingo Helbig, Saskia Biskup, Markus Wolff, Snezana Maljevic, Rebecca Schüle, Sanjay M. Sisodiya, Sarah Weckhuysen, Holger Lerche, and Johannes R. Lemke. De novo loss- or gain-of-function mutations in KCNA2 cause epileptic encephalopathy. *Nature Genetics*, 47(4):393–399.

[167] Slavé Petrovski, Quanli Wang, Erin L. Heinzen, Andrew S. Allen, and David B. Goldstein.

Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS genetics*, 9(8):e1003709.

[168] Mehdi Pirooznia, Tao Wang, Dimitrios Avramopoulos, David Valle, Gareth Thomas, Richard L. Huganir, Fernando S. Goes, James B. Potash, and Peter P. Zandi. SynaptomeDB: an ontology-based knowledgebase for synaptic genes. *Bioinformatics*, 28(6):897–899.

[169] G. Kirov, A. J. Pocklington, P. Holmans, D. Ivanov, M. Ikeda, D. Ruderfer, J. Moran, K. Chambert, D. Toncheva, L. Georgieva, D. Grozeva, M. Fjodorova, R. Wollerton, E. Rees, I. Nikolov, L. N. van de Lagemaat, A. Bayés, E. Fernandez, P. I. Olason, Y. Böttcher, N. H. Komiyama, M. O. Collins, J. Choudhary, K. Stefansson, H. Stefansson, S. G. N. Grant, S. Purcell, P. Sklar, M. C. O'Donovan, and M. J. Owen. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular Psychiatry*, 17(2):142–153.

[170] Alisa Mo, Eran A. Mukamel, Fred P. Davis, Chongyuan Luo, Gilbert L. Henry, Serge Picard, Mark A. Urich, Joseph R. Nery, Terrence J. Sejnowski, Ryan Lister, Sean R. Eddy, Joseph R. Ecker, and Jeremy Nathans. Epigenomic signatures of neuronal diversity in the mammalian brain. *Neuron*, 86(6):1369–1384.

[171] Jennifer C. Darnell, Sarah J. Van Driesche, Chaolin Zhang, Ka Ying Sharon Hung, Aldo Mele, Claire E. Fraser, Elizabeth F. Stone, Cynthia Chen, John J. Fak, Sung Wook Chi, Donny D. Licatalosi, Joel D. Richter, and Robert B. Darnell. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, 146(2):247–261.

[172] Kaitlin E. Samocha, Elise B. Robinson, Stephan J. Sanders, Christine Stevens, Aniko Sabo, Lauren M. McGrath, Jack A. Kosmicki, Karola Rehnström, Swapan Mallick, Andrew Kirby, Dennis P. Wall, Daniel G. MacArthur, Stacey B. Gabriel, Mark DePristo, Shaun M. Purcell, Aarno Palotie, Eric Boerwinkle, Joseph D. Buxbaum, Edwin H. Cook, Richard A. Gibbs, Gerard D. Schellenberg, James S. Sutcliffe, Bernie Devlin, Kathryn Roeder, Benjamin M. Neale, and Mark J. Daly. A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, 46(9):944–950.

[173] Helen V. Firth, Shola M. Richards, A. Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M. Pettett, and Nigel P. Carter. DECIPHER: Database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics*, 84(4):524–533.

[174] EpiPM Consortium. A roadmap for precision medicine in the epilepsies. *The Lancet. Neurology*, 14(12):1219–1228.

[175] Ingrid E. Scheffer, Samuel Berkovic, Giuseppe Capovilla, Mary B. Connolly, Jacqueline French, Laura Guilhoto, Edouard Hirsch, Satish Jain, Gary W. Mathern, Solomon L. Moshé, Douglas R. Nordli, Emilio Perucca, Torbjörn Tomson, Samuel Wiebe, Yue-Hua Zhang, and Sameer M. Zuberi. ILAE classification of the epilepsies: Position paper of the ILAE commission for classification and terminology. *Epilepsia*, 58(4):512–521.

[176] Samuel F. Berkovic, R. Anne Howell, David A. Hay, and John L. Hopper. Epilepsies in twins: Genetics of the major epilepsy syndromes. *Annals of Neurology*, 43(4):435–445, .

[177] Stéphanie Baulac, Gilles Huberfeld, Isabelle Gourfinkel-An, Georgia Mitropoulou, Alexandre Beranger, Jean-François Prud'homme, Michel Baulac, Alexis Brice, Roberto Bruzzone, and Eric Leguern. First genetic evidence of GABAA receptor dysfunction in epilepsy: A mutation in the gamma2-subunit gene. *Nature genetics*, 28:46–8.

[178] Todor Arsov, Saul A. Mullen, Sue Rogers, A. Marie Phillips, Kate M. Lawrence, John A. Damiano, Hadassa Goldberg-Stern, Zaid Afawi, Sara Kivity, Chantal Trager, Steven Petrou, Samuel F. Berkovic, and Ingrid E. Scheffer. Glucose transporter 1 deficiency in the idiopathic generalized epilepsies. *Annals of Neurology*, 72(5):807–815.

[179] P Striano, Y G Weber, M R Toliat, J Schubert, C Leu, R Chaimana, S Baulac, R Guerrero, E LeGuern, A-E Lehesjoki, A Polvi, A Robbiano, J M Serratosa, R Guerrini, P Nürnberg, T Sander, F Zara, H Lerche, and C Marini. GLUT1 mutations are a rare cause of familial idiopathic generalized epilepsy. *Neurology*, 78(8):557–62.

[180] Tara Klassen, Caleb Davis, Alica Goldman, Dan Burgess, Tim Chen, David Wheeler, John McPherson, Traci Bourquin, Lora Lewis, Donna Villasana, Margaret Morgan, Donna Muzny, Richard Gibbs, and Jeffrey Noebels. Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. *Cell*, 145(7):1036–48.

[181] Agostinelli Sergio, Accorsi Patrizia, Beccaria Francesca, Belcastro Vincenzo, Canevini Maria Paola, Capovilla Giuseppe, Cappanera Silvia, Bernardina Bernardo Dalla, Darra Francesca, Gaudio Luigi Del, Elia Maurizio, Falsaperla Raffaele, Giordano Lucio, Gobbi Giuseppe, Minetti Carlo, Nicita Francesco, Parisi Pasquale, Pavone Piero, Pezzella Marianna, Sesta Michela, Spalice Alberto, Striano Salvatore, Tozzi Elisabetta, Traverso Monica, Vari Stella, Vignoli Aglaia, Zamponi Nelia, Zara Federico, Striano Pasquale, Verrotti Alberto, and null null. Clinical dissection of early onset absence epilepsy in children and prognostic implications. *Epilepsia*, 54(10):1761–1770.

[182] Elizabeth C. Galizia, Candace T. Myers, Costin Leu, Carolien G. F. de Kovel, Tatiana Afrikanova, Maria Lorena Cordero-Maldonado, Teresa G. Martins, Maxime Jacmin, Suzanne Drury, V. Krishna Chinthapalli, Hiltrud Muhle, Manuela Pendziwiat, Thomas

Sander, Ann-Kathrin Ruppert, Rikke S. Møller, Holger Thiele, Roland Krause, Julian Schubert, Anna-Elina Lehesjoki, Peter Nürnberg, Holger Lerche, EuroEPINOMICS Co-GIE Consortium, Aarno Palotie, Antonietta Coppola, Salvatore Striano, Luigi Del Gaudio, Christopher Boustred, Amy L. Schneider, Nicholas Lench, Bosanka Jocic-Jakubi, Athanasios Covanis, Giuseppe Capovilla, Pierangelo Veggiotti, Marta Piccioli, Pasquale Parisi, Laura Cantonetti, Lynette G. Sadleir, Saul A. Mullen, Samuel F. Berkovic, Ulrich Stephani, Ingo Helbig, Alexander D. Crawford, Camila V. Esguerra, Dorothee G. A. Kasteleijn-Nolst Trenité, Bobby P. C. Koeleman, Heather C. Mefford, Ingrid E. Scheffer, and Sanjay M. Sisodiya. CHD2 variants are a risk factor for photosensitivity in epilepsy. *Brain: A Journal of Neurology*, 138:1198–1207.

[183] Albert Hofman, Guy G. O. Brusselle, Sarwa Darwish Murad, Cornelia M. van Duijn, Oscar H. Franco, André Goedegebure, M. Arfan Ikram, Caroline C. W. Klaver, Tamar E. C. Nijsten, Robin P. Peeters, Bruno H. Ch Stricker, Henning W. Tiemeier, André G. Uitterlinden, and Meike W. Vernooij. The rotterdam study: 2016 objectives and design update. *European Journal of Epidemiology*, 30(8):661–708, .

[184] M. Arfan Ikram, Guy G. O. Brusselle, Sarwa Darwish Murad, Cornelia M. van Duijn, Oscar H. Franco, André Goedegebure, Caroline C. W. Klaver, Tamar E. C. Nijsten, Robin P. Peeters, Bruno H. Stricker, Henning Tiemeier, André G. Uitterlinden, Meike W. Vernooij, and Albert Hofman. The rotterdam study: 2018 update on objectives, design and main results. *European Journal of Epidemiology*, 32(9):807–850.

[185] The UK10K Consortium. The UK10k project identifies rare variants in health and disease. *Nature*, 526(7571):82–90.

[186] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American journal of human genetics*, 83(3):311–21.

[187] H. Möhler. GABA(a) receptor diversity and pharmacology. *Cell and Tissue Research*, 326 (2):505–516.

[188] Rikke S. Møller, Thomas V. Wuttke, Ingo Helbig, Carla Marini, Katrine M. Johannesen, Eva H. Brilstra, Ulvi Vaher, Ingo Borggraefe, Inga Talvik, Tiina Talvik, Gerhard Kluger, Laurence L. Francois, Gaetan Lesca, Julitta de Bellescize, Susanne Blichfeldt, Nicolas Chatron, Nils Holert, Julia Jacobs, Marielle Swinkels, Cornelia Betzler, Steffen Syrbe, Marina Nikanorova, Candace T. Myers, Line H. G. Larsen, Sabina Vejzovic, Manuela Pendziwiat, Sarah von Spiczak, Sarah Hopkins, Holly Dubbs, Yuan Mang, Konstantin Mukhin, Hans Holthausen, Koen L. van Gassen, Hans A. Dahl, Niels Tommerup, Heather C. Mefford, Guido Rubboli, Renzo Guerrini, Johannes R. Lemke, Holger Lerche, Hiltrud

Muhle, and Snezana Maljevic. Mutations in GABRB3: From febrile seizures to epileptic encephalopathies. *Neurology*, 88(5):483–492.

[189] Christopher T. Johansen, Jian Wang, Matthew B. Lanktree, Henian Cao, Adam D. McIntyre, Matthew R. Ban, Rebecca A. Martins, Brooke A. Kennedy, Reina G. Hassell, Maartje E. Visser, Stephen M. Schwartz, Benjamin F. Voight, Roberto Elosua, Veikko Salomaa, Christopher J. O'Donnell, Geesje M. Dallinga-Thie, Sonia S. Anand, Salim Yusuf, Murray W. Huff, Sekar Kathiresan, and Robert A. Hegele. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature Genetics*, 42 (8):684–687.

[190] Epi4K Consortium, Epilepsy Phenome/Genome Project, Andrew S. Allen, Samuel F. Berkovic, Patrick Cossette, Norman Delanty, Dennis Dlugos, Evan E. Eichler, Michael P. Epstein, Tracy Glauser, David B. Goldstein, Yujun Han, Erin L. Heinzen, Yuki Hitomi, Katherine B. Howell, Michael R. Johnson, Ruben Kuzniecky, Daniel H. Lowenstein, Yi-Fan Lu, Maura R. Z. Madou, Anthony G. Marson, Heather C. Mefford, Sahar Esmaeeli Nieh, Terence J. O'Brien, Ruth Ottman, Slavé Petrovski, Annapurna Poduri, Elizabeth K. Ruzzo, Ingrid E. Scheffer, Elliott H. Sherr, Christopher J. Yuskaitis, Bassel Abou-Khalil, Brian K. Alldredge, Jocelyn F. Bautista, Samuel F. Berkovic, Alex Boro, Gregory D. Cascino, Damian Consalvo, Patricia Crumrine, Orrin Devinsky, Dennis Dlugos, Michael P. Epstein, Miguel Fiol, Nathan B. Fountain, Jacqueline French, Daniel Friedman, Eric B. Geller, Tracy Glauser, Simon Glynn, Sheryl R. Haut, Jean Hayward, Sandra L. Helmers, Sucheta Joshi, Andres Kanner, Heidi E. Kirsch, Robert C. Knowlton, Eric H. Kossoff, Rachel Kuperman, Ruben Kuzniecky, Daniel H. Lowenstein, Shannon M. McGuire, Paul V. Motika, Edward J. Novotny, Ruth Ottman, Juliann M. Paolicchi, Jack M. Parent, Kristen Park, Annapurna Poduri, Ingrid E. Scheffer, Renée A. Shellhaas, Elliott H. Sherr, Jerry J. Shih, Rani Singh, Joseph Sirven, Michael C. Smith, Joseph Sullivan, Liu Lin Thio, Anu Venkat, Eileen P. G. Vining, Gretchen K. Von Allmen, Judith L. Weisenberg, Peter Widdess-Walsh, and Melodie R. Winawer. De novo mutations in epileptic encephalopathies. *Nature*, 501 (7466):217–221.

[191] Dingding Shen, Ciria C. Hernandez, Wangzhen Shen, Ningning Hu, Annapurna Poduri, Beth Shiedley, Alex Rotenberg, Alexandre N. Datta, Steffen Leiz, Steffi Patzer, Rainer Boor, Kerri Ramsey, Ethan Goldberg, Ingo Helbig, Xilma R. Ortiz-Gonzalez, Johannes R. Lemke, Eric D. Marsh, and Robert L. Macdonald. De novo GABRG2 mutations associated with epileptic encephalopathies. *Brain : a journal of neurology*, 140:49–67.

[192] Valerie B. Caraiscos, Erin M. Elliott, Kong E. You-Ten, Victor Y. Cheng, Delia Belelli, J. Glen Newell, Michael F. Jackson, Jeremy J. Lambert, Thomas W. Rosahl, Keith A. Wafford, John F. MacDonald, and Beverley A. Orser. Tonic inhibition in mouse hippocam-

pal CA1 pyramidal neurons is mediated by alpha5 subunit-containing gamma-aminobutyric acid type a receptors. *Proceedings of the National Academy of Sciences*, 101(10):3662–3667.

[193] Ulrike B. S. Hedrich, Camille Liautard, Daniel Kirschenbaum, Martin Pofahl, Jennifer Lavigne, Yuanyuan Liu, Stephan Theiss, Johannes Slotta, Andrew Escayg, Marcel Dihné, Heinz Beck, Massimo Mantegazza, and Holger Lerche. Impaired action potential initiation in GABAergic interneurons causes hyperexcitable networks in an epileptic mouse model carrying a human na(v)1.1 mutation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 34(45):14874–14889.

[194] Nianhui Zhang, Weizheng Wei, Istvan Mody, and Carolyn R. Houser. Altered localization of GABA(a) receptor subunits on dentate granule cell dendrites influences tonic and phasic inhibition in a mouse model of epilepsy. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 27(28):7520–7531, .

[195] Fadi F. Hamdan, Candace T. Myers, Patrick Cossette, Philippe Lemay, Dan Spiegelman, Alexandre Dionne Laporte, Christina Nassif, Ousmane Diallo, Jean Monlong, Maxime Cadieux-Dion, Sylvia Dobrzeniecka, Caroline Meloche, Kyle Retterer, Megan T. Cho, Jill A. Rosenfeld, Weimin Bi, Christine Massicotte, Marguerite Miguet, Ledia Brunga, Brigid M. Regan, Kelly Mo, Cory Tam, Amy Schneider, Georgie Hollingsworth, Deciphering Developmental Disorders Study, David R. FitzPatrick, Alan Donaldson, Natalie Canham, Edward Blair, Bronwyn Kerr, Andrew E. Fry, Rhys H. Thomas, Joss Shelagh, Jane A. Hurst, Helen Brittain, Moira Blyth, Robert Roger Lebel, Erica H. Gerkes, Laura Davis-Keppen, Quinn Stein, Wendy K. Chung, Sara J. Dorison, Paul J. Benke, Emily Fassi, Nicole Corsten-Janssen, Erik-Jan Kamsteeg, Frederic T. Mau-Them, Ange-Line Bruel, Alain Verloes, Katrin Õunap, Monica H. Wojcik, Dara V. F. Albert, Sunita Venkateswaran, Tyson Ware, Dean Jones, Yu-Chi Liu, Shekeeb S. Mohammad, Peyman Bizargity, Carlos A. Bacino, Vincenzo Leuzzi, Simone Martinelli, Bruno Dallapiccola, Marco Tartaglia, Lubov Blumkin, Klaas J. Wierenga, Gabriela Purcarin, James J. O'Byrne, Sylvia Stockler, Anna Lehman, Boris Keren, Marie-Christine Nougues, Cyril Mignot, Stéphane Auvin, Caroline Nava, Susan M. Hiatt, Martina Bebin, Yunru Shao, Fernando Scaglia, Seema R. Lalani, Richard E. Frye, Imad T. Jarjour, Stéphanie Jacques, Renee-Myriam Boucher, Emilie Riou, Myriam Srour, Lionel Carmant, Anne Lortie, Philippe Major, Paola Diadori, François Dubeau, Guy D'Anjou, Guillaume Bourque, Samuel F. Berkovic, Lynette G. Sadleir, Philippe M. Campeau, Zoha Kibar, Ronald G. Lafrenière, Simon L. Girard, Saadet Mercimek-Mahmutoglu, Cyrus Boelman, Guy A. Rouleau, Ingrid E. Scheffer, Heather C. Mefford, Danielle M. Andrade, Elsa Rossignol, Berge A. Minassian, and Jacques L. Michaud. High rate of recurrent de novo mutations in developmental and epileptic encephalopathies. *American Journal of Human Genetics*, 101(5):664–685, .

[196] Atsushi Ishii, Jing-Qiong Kang, Cara C. Schornak, Ciria C. Hernandez, Wangzhen Shen, Joseph C. Watkins, Robert L. Macdonald, and Shinichi Hirose. Ade novomissense mutation ofGABRB2causes early myoclonic encephalopathy. *Journal of Medical Genetics*, 54(3):202–211.

[197] Siddharth Srivastava, Julie Cohen, Jonathan Pevsner, Swaroop Aradhya, Dianalee McKnight, Elizabeth Butler, Michael Johnston, and Ali Fatemi. A novel variant in GABRB2 associated with intellectual disability and epilepsy. *American Journal of Medical Genetics. Part A*, 164A(11):2914–2921.

[198] Paul L. Auer and Guillaume Lettre. Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, 7(1):16.

[199] Seunggeung Lee, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin. Rare-variant association analysis: Study designs and statistical tests. *American journal of human genetics*, 95(1):5–23, .

[200] Cristina Elena Niturad, Dorit Lev, Vera M. Kalscheuer, Agnieszka Charzewska, Julian Schubert, Tally Lerman-Sagie, Hester Y. Kroes, Renske Oegema, Monica Traverso, Nicola Specchio, Maria Lassota, Jamel Chelly, Odeya Bennett-Back, Nirit Carmi, Tal Koffler-Brill, Michele Iacomino, Marina Trivisano, Giuseppe Capovilla, Pasquale Striano, Magdalena Nawara, Sylwia Rzonca, Ute Fischer, Melanie Bienek, Corinna Jensen, Hao Hu, Holger Thiele, Janine Altmüller, Roland Krause, Patrick May, Felicitas Becker, EuroEPINOMICS Consortium, Rudi Balling, Saskia Biskup, Stefan A. Haas, Peter Nürnberg, Koen L. I. van Gassen, Holger Lerche, Federico Zara, Snezana Maljevic, and Esther Leshinsky-Silver. Rare GABRA3 variants are associated with epileptic seizures, encephalopathy and dysmorphic features. *Brain: A Journal of Neurology*, 140(11):2879–2894.

[201] Leanne M Dibbens, Saul Mullen, Ingo Helbig, Heather C Mefford, Marta a Bayly, Susannah Bellows, Costin Leu, Holger Trucks, Tanja Obermeier, Michael Wittig, Andre Franke, Hande Caglayan, Zuhal Yapici, Thomas Sander, Evan E Eichler, Ingrid E Scheffer, John C Mulley, and Samuel F Berkovic. Familial and sporadic 15q13.3 microdeletions in idiopathic generalized epilepsy: precedent for disorders with complex inheritance. *Hum. Mol. Genet.*, 18(19):3626–3631.

[202] Heather C. Mefford. CNVs in epilepsy. *Current Genetic Medicine Reports*, 2:162–167.

[203] Katrine Johannesen, Carla Marini, Siona Pfeffer, Rikke S. Møller, Thomas Dorn, Cristina Elena Niturad, Elena Gardella, Yvonne Weber, Marianne Søndergård, Helle Hjalgrim, Mariana Nikanorova, Felicitas Becker, Line H. G. Larsen, Hans A. Dahl, Oliver Maier, Davide Mei, Saskia Biskup, Karl M. Klein, Philipp S. Reif, Felix Rosenow, Abdallah F.

Elias, Cindy Hudson, Katherine L. Helbig, Susanne Schubert-Bast, Maria R. Scordo, Dana Craiu, Tania Djémié, Dorota Hoffman-Zacharska, Hande Caglayan, Ingo Helbig, Jose Serratosa, Pasquale Striano, Peter De Jonghe, Sarah Weckhuysen, Arvid Suls, Kai Muru, Inga Talvik, Tiina Talvik, Hiltrud Muhle, Ingo Borggraefe, Imma Rost, Renzo Guerrini, Holger Lerche, Johannes R. Lemke, Guido Rubboli, and Snezana Maljevic. Phenotypic spectrum of GABRA1: From generalized epilepsies to severe epileptic encephalopathies. *Neurology*, 87(11):1140–1151.

[204] Jing-Qiong Kang and Robert L. Macdonald. Molecular pathogenic basis for GABRG2 mutations associated with a spectrum of epilepsy syndromes, from generalized absence epilepsy to dravet syndrome. *JAMA neurology*, 73(8):1009–1016.

[205] Holger Lerche, Mala Shah, Heinz Beck, Jeff Noebels, Dan Johnston, and Angela Vincent. Ion channels in genetic and acquired forms of epilepsy. *The Journal of Physiology*, 591: 753–764.

[206] Robert L. Macdonald, Jing-Qiong Kang, and Martin J. Gallagher. Mutations in GABAA receptor subunits associated with genetic epilepsies. *The Journal of Physiology*, 588(11): 1861–1869.

[207] Chuming Chen, Sari S. Khaleel, Hongzhan Huang, and Cathy H. Wu. Software for preprocessing illumina next-generation sequencing short read sequences. *Source Code for Biology and Medicine*, 9:8, .

[208] Kai Wang, Mingyao Li, and Hakon Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature reviews. Genetics*, 11(12):843–54, .

[209] Seunggeun Lee, Christian Fuchsberger, Sehee Kim, and Laura Scott. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies. *Biostatistics (Oxford, England)*, 17(1):1–15, .

[210] Clement Ma, Tom Blackwell, Michael Boehnke, and Laura J. Scott. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic epidemiology*, 37(6):539–550.

[211] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

[212] D. C. Hesdorffer, G. Logroscino, E. K. T. Benn, N. Katri, G. Cascino, and W. A. Hauser. Estimating risk for developing epilepsy: a population-based study in rochester, minnesota. *Neurology*, 76(1):23–27.

[213] Katherine L. Helbig, Kelly D. Farwell Hagman, Deepali N. Shinde, Cameron Mroske, Zöe Powis, Shuwei Li, Sha Tang, and Ingo Helbig. Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 18(9):898–905, .

[214] Ashura W. Buckley and Gregory L. Holmes. Epilepsy and autism. *Cold Spring Harbor Perspectives in Medicine*, 6(4):a022749–a022749.

[215] Santhosh Girirajan, Catarina D. Campbell, and Evan E. Eichler. Human copy number variation and complex genetic disease. *Annual Review of Genetics*, 45:203–226.

[216] Mohammed Uddin, Kristiina Tammimies, Giovanna Pellecchia, Babak Alipanahi, Pingzhao Hu, Zhuozhi Wang, Dalila Pinto, Lynette Lau, Thomas Nalpathamkalam, Christian R. Marshall, Benjamin J. Blencowe, Brendan J. Frey, Daniele Merico, Ryan K. C. Yuen, and Stephen W. Scherer. Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. *Nature Genetics*, 46(7):742–747, .

[217] Mohammed Uddin, Giovanna Pellecchia, Bhooma Thiruvahindrapuram, Lia D'Abate, Daniele Merico, Ada Chan, Mehdi Zarrei, Kristiina Tammimies, Susan Walker, Matthew J. Gazzellone, Thomas Nalpathamkalam, Ryan K. C. Yuen, Koenraad Devriendt, Géraldine Mathonnet, Emmanuelle Lemyre, Sonia Nizard, Mary Shago, Ann M. Joseph-George, Abdul Noor, Melissa T. Carter, Grace Yoon, Peter Kannu, Frédérique Tihy, Erik C. Thorland, Christian R. Marshall, Janet A. Buchanan, Marsha Speevak, Dimitri J. Stavropoulos, and Stephen W. Scherer. Indexing effects of copy number variation on genes involved in developmental delay. *Scientific Reports*, 6:28663, .

[218] Costin Leu, Antonietta Coppola, and Sanjay M. Sisodiya. Progress from genome-wide association studies and copy number variant studies in epilepsy. *Current Opinion in Neurology*, 29(2):158–167, .

[219] John C. Mulley and Heather C. Mefford. Epilepsy and the new cytogenetics. *Epilepsia*, 52 (3):423–432.

[220] Laura Addis, Richard E. Rosch, Antonio Valentin, Andrew Makoff, Robert Robinson, Kate V. Everett, Lina Nashef, and Deb K. Pal. Analysis of rare copy number variation in absence epilepsies. *Neurology. Genetics*, 2(2):e56.

[221] Dennis Lal, Ann-Kathrin Ruppert, Holger Trucks, Herbert Schulz, Carolien G de Kovel, Dorothée Kasteleijn-Nolst Trenité, Anja C M Sonsma, Bobby P Koeleman, Dick Lindhout, Yvonne G Weber, Holger Lerche, Claudia Kapser, Christoph J Schankin, Wolfram S Kunz, Rainer Surges, Christian E Elger, Verena Gaus, Bettina Schmitz, Ingo Helbig, Hiltrud Muhle, Ulrich Stephani, Karl M Klein, Felix Rosenow, Bernd A Neubauer,

Eva M Reinthaler, Fritz Zimprich, Martha Feucht, Rikke S Møller, Helle Hjalgrim, Peter De Jonghe, Arvid Suls, Wolfgang Lieb, Andre Franke, Konstantin Strauch, Christian Gieger, Claudia Schurmann, Ulf Schminke, Peter Nürnberg, and Thomas Sander. Burden analysis of rare microdeletions suggests a strong impact of neurodevelopmental genes in genetic generalised epilepsies. *PLoS genetics*, 11(5):e1005226–e1005226, .

[222] Ingrid E. Scheffer and Heather C. Mefford. Epilepsy: Beyond the single nucleotide variant in epilepsy genetics. *Nature Reviews. Neurology*, 10(9):490–491.

[223] Przemyslaw Szafranski, Gretchen K. Von Allmen, Brett H. Graham, Angus A. Wilfong, Sung-Hae L. Kang, Jose A. Ferreira, Sheila J. Upton, John B. Moeschler, Weimin Bi, Jill A. Rosenfeld, Lisa G. Shaffer, Sau Wai Cheung, Paweł Stankiewicz, and Seema R. Lalani. 6q22.1 microdeletion and susceptibility to pediatric epilepsy. *European journal of human genetics: EJHG*, 23(2):173–179.

[224] John A. Damiano, Saul A. Mullen, Michael S. Hildebrand, Susannah T. Bellows, Kate M. Lawrence, Todor Arsov, Leanne Dibbens, Heather Major, Hans-Henrik M. Dahl, Heather C. Mefford, Benjamin W. Darbro, Ingrid E. Scheffer, and Samuel F. Berkovic. Evaluation of multiple putative risk alleles within the 15q13.3 region for genetic generalized epilepsy. *Epilepsy Research*, 117:70–73.

[225] Johanna A. Jähn, Sarah von Spiczak, Hiltrud Muhle, Tanja Obermeier, Andre Franke, Heather C. Mefford, Ulrich Stephani, and Ingo Helbig. Iterative phenotyping of 15q11.2, 15q13.3 and 16p13.11 microdeletion carriers in pediatric epilepsies. *Epilepsy Research*, 108 (1):109–116.

[226] James R. Lupski. Clinical genomics: from a truly personal genome viewpoint. *Human Genetics*, 135(6):591–601.

[227] Philip M. Boone, Bo Yuan, Ian M. Campbell, Jennifer C. Scull, Marjorie A. Withers, Brett C. Baggett, Christine R. Beck, Christine J. Shaw, Pawel Stankiewicz, Paolo Moretti, Wendy E. Goodwin, Nichole Hein, John K. Fink, Moon-Woo Seong, Soo Hyun Seo, Sung Sup Park, Izabela D. Karbassi, Sat Dev Batish, Andrés Ordóñez-Ugalde, Beatriz Quintáns, María-Jesús Sobrido, Susanne Stemmler, and James R. Lupski. The alu-rich genomic architecture of SPAST predisposes to diverse and functionally distinct disease-associated CNV alleles. *American Journal of Human Genetics*, 95(2):143–161.

[228] Dheeraj Malhotra and Jonathan Sebat. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell*, 148(6):1223–1241.

[229] Ian M Campbell, Mitchell Rao, Sean D Arredondo, Seema R Lalani, Zhilian Xia, Sung-Hae L Kang, Weimin Bi, Amy M Breman, Janice L Smith, Carlos a Bacino, Arthur L

Beaudet, Ankita Patel, Sau Wai Cheung, James R Lupski, Paweł Stankiewicz, Melissa B Ramocki, and Chad a Shaw. Fusion of large-scale genomic knowledge and frequency data computationally prioritizes variants in epilepsy. *PLoS Genet.*, 9(9):e1003797–e1003797.

[230] Albert Hofman, Cornelia M. van Duijn, Oscar H. Franco, M. Arfan Ikram, Harry L. A. Janssen, Caroline C. W. Klaver, Ernst J. Kuipers, Tamar E. C. Nijsten, Bruno H. Ch. Stricker, Henning Tiemeier, André G. Uitterlinden, Meike W. Vernooij, and Jacqueline C. M. Witteman. The rotterdam study: 2012 objectives and design update. *European Journal of Epidemiology*, 26(8):657–686, .

[231] najoshi. sickle: Windowed adaptive trimming for fastq files using quality. original-date: 2011-02-09T01:18:45Z.

[232] Nigel C K Tan, John C Mulley, and Samuel F Berkovic. Genetic association studies in epilepsy: "the truth is out there". *Epilepsia*, 45(11):1429–1442, .

[233] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2.

[234] Ping Zhang, Ling Liu, Jinsha Huang, Liang Shao, Hongcai Wang, Nian Xiong, and Tao Wang. Non-SMC condensin i complex, subunit d2 gene polymorphisms are associated with parkinson's disease: a han chinese study. *Genome*, 57(5):253–257, .

[235] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Disease networks. uncovering disease-disease relationships through the incomplete interactome. *Science (New York, N.Y.)*, 347(6224): 1257601.

[236] Dalila Pinto, Elsa Delaby, Daniele Merico, Mafalda Barbosa, Alison Merikangas, Lambertus Klei, Bhooma Thiruvahindrapuram, Xiao Xu, Robert Ziman, Zhuozhi Wang, Jacob A S Vorstman, Ann Thompson, Regina Regan, Marion Pilorge, Giovanna Pellecchia, Alistair T Pagnamenta, Bárbara Oliveira, Christian R Marshall, Tiago R Magalhaes, Jennifer K Lowe, Jennifer L Howe, Anthony J Griswold, John Gilbert, Eftichia Duketis, Beth A Dombroski, Maretha V De Jonge, Michael Cuccaro, Emily L Crawford, Catarina T Correia, Judith Conroy, Inês C Conceição, Andreas G Chiocchetti, Jillian P Casey, Guiqing Cai, Christelle Cabrol, Nadia Bolshakova, Elena Bacchelli, Richard Anney, Steven Gallinger, Michelle Cotterchio, Graham Casey, Lonnie Zwaigenbaum, Kerstin Wittemeyer, Kirsty Wing, Simon Wallace, Herman van Engeland, Ana Tryfon, Susanne Thomson, Latha Soorya, Bernadette Rogé, Wendy Roberts, Fritz Poustka, Susana Mouga, Nancy Minshew, L Alison McInnes, Susan G McGrew, Catherine Lord, Marion Leboyer, Ann S Le Couteur, Alexander Kolevzon, Patricia Jiménez González, Suma Jacob, Richard Holt, Stephen

Guter, Jonathan Green, Andrew Green, Christopher Gillberg, Bridget A Fernandez, Frederico Duque, Richard Delorme, Geraldine Dawson, Pauline Chaste, Cátia Café, Sean Brennan, Thomas Bourgeron, Patrick F Bolton, Sven Bölte, Raphael Bernier, Gillian Baird, Anthony J Bailey, Evdokia Anagnostou, Joana Almeida, Ellen M Wijsman, Veronica J Vieland, Astrid M Vicente, Gerard D Schellenberg, Margaret Pericak-Vance, Andrew D Paterson, Jeremy R Parr, Guiomar Oliveira, John I Nurnberger, Anthony P Monaco, Elena Maestrini, Sabine M Klauck, Hakon Hakonarson, Jonathan L Haines, Daniel H Geschwind, Christine M Freitag, Susan E Folstein, Sean Ennis, Hilary Coon, Agatino Battaglia, Peter Szatmari, James S Sutcliffe, Joachim Hallmayer, Michael Gill, Edwin H Cook, Joseph D Buxbaum, Bernie Devlin, Louise Gallagher, Catalina Betancur, and Stephen W Scherer. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *American journal of human genetics*, 94(5):677–94.

[237] Andreas Krämer, Jeff Green, Jack Pollard, and Stuart Tugendreich. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics (Oxford, England)*, 30(4):523–30.

[238] DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45:D833–D839, .

[239] Douglas M Ruderfer, Tymor Hamamsy, Monkol Lek, Konrad J Karczewski, David Kavanagh, Kaitlin E Samocha, Mark J Daly, Daniel G MacArthur, Menachem Fromer, and Shaun M Purcell. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nature Genetics*.

[240] Xia Ran, Jinchen Li, Qianzhi Shao, Huiqian Chen, Zhongdong Lin, Zhong Sheng Sun, and Jinyu Wu. EpilepsyGene: a genetic resource for genes and mutations related to epilepsy. *Nucleic Acids Research*, 43:D893–899.

[241] Amit Kawalia, Susanne Motameny, Stephan Wonczak, Holger Thiele, Lech Nieroda, Kamel Jabbari, Stefan Borowski, Vishal Sinha, Wilfried Gunia, Ulrich Lang, Viktor Achter, and Peter Nürnberg. Leveraging the power of high performance computing for next generation sequencing data analysis: tricks and twists from a high throughput exome workflow. *PloS One*, 10(5):e0126321.

[242] Ahmed Mahfouz, Mark N. Ziats, Owen M. Rennert, Boudewijn P. F. Lelieveldt, and Marcel J. T. Reinders. Shared pathways among autism candidate genes determined by co-expression network analysis of the developing human brain transcriptome. *Journal of molecular neuroscience: MN*, 57(4):580–594.

[243] Sarra Dimassi, Audrey Labalme, Gaetan Lesca, Gabrielle Rudolf, Nadine Bruneau, Edouard Hirsch, Alexis Arzimanoglou, Jacques Motte, Anne de Saint Martin, Nadia

Boutry-Kryza, Robin Cloarec, Afaf Benitto, Agnès Ameil, Patrick Edery, Philippe Ryvlin, Julitta De Bellescize, Pierre Szepetowski, and Damien Sanlaville. A subset of genomic alterations detected in rolandic epilepsies contains candidate or known epilepsy genes including GRIN2a and PRRT2. *Epilepsia*, 55(2):370–378.

[244] Alexander Zimprich, Saskia Biskup, Petra Leitner, Peter Lichtner, Matthew Farrer, Sarah Lincoln, Jennifer Kachergus, Mary Hulihan, Ryan J. Uitti, Donald B. Calne, A. Jon Stoessl, Ronald F. Pfeiffer, Nadja Patenge, Iria Carballo Carbajal, Peter Vieregge, Friedrich Asmus, Bertram Müller-Myhsok, Dennis W. Dickson, Thomas Meitinger, Tim M. Strom, Zbigniew K. Wszolek, and Thomas Gasser. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron*, 44(4):601–607, .

[245] Susan Shur-Fen Gau, Hsiao-Mei Liao, Chao-Chun Hong, Wei-Hsien Chien, and Chia-Hsiang Chen. Identification of two inherited copy number variants in a male with autism supports two-hit and compound heterozygosity models of autism. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, 159B(6):710–7.

[246] Christian Gilissen, Jayne Y Hehir-Kwa, Djie Tjwan Thung, Maartje van de Vorst, Bregje W M van Bon, Marjolein H Willemsen, Michael Kwint, Irene M Janssen, Alexander Hoischen, Annette Schenck, Richard Leach, Robert Klein, Rick Tearle, Tan Bo, Rolph Pfundt, Helger G Yntema, Bert B A de Vries, Tjitske Kleefstra, Han G Brunner, Lisenka E L M Vissers, and Joris A Veltman. Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509):344–7.

[247] Heterozygote carriers for CNVs in PARK2 are at increased risk of parkinson's disease. *Human molecular genetics*, 24(19):5637–43.

[248] Francesco Nicita, Fiorenza Ulgiati, Laura Bernardini, Giacomo Garone, Laura Papetti, Antonio Novelli, and Alberto Spalice. Early myoclonic encephalopathy in 9q33-q34 deletion encompassing STXBP1 and SPTAN1. *Annals of Human Genetics*, 79(3):209–217, .

[249] Hirotomo Saitsu, Jun Tohyama, Tatsuro Kumada, Kiyoshi Egawa, Keisuke Hamada, Ippei Okada, Takeshi Mizuguchi, Hitoshi Osaka, Rie Miyata, Tomonori Furukawa, Kazuhiro Haginoya, Hideki Hoshino, Tomohide Goto, Yasuo Hachiya, Takanori Yamagata, Shinji Saitoh, Toshiro Nagai, Kiyomi Nishiyama, Akira Nishimura, Noriko Miyake, Masayuki Komada, Kenji Hayashi, Syu-Ichi Hirai, Kazuhiro Ogata, Mitsuhiro Kato, Atsuo Fukuda, and Naomichi Matsumoto. Dominant-negative mutations in alpha-II spectrin cause west syndrome with severe cerebral hypomyelination, spastic quadriplegia, and developmental delay. *American Journal of Human Genetics*, 86(6):881–891.

[250] Fadi F. Hamdan, Myriam Srour, Jose-Mario Capo-Chichi, Hussein Daoud, Christina Nassif, Lysanne Patry, Christine Massicotte, Amirthagowri Ambalavanan, Dan Spiegelman, Ousmane Diallo, Edouard Henrion, Alexandre Dionne-Laporte, Anne Fougerat, Alexey V. Pshezhetsky, Sunita Venkateswaran, Guy A. Rouleau, and Jacques L. Michaud. De novo mutations in moderate or severe intellectual disability. *PLoS Genetics*, 10(10):e1004772–e1004772, .

[251] D. N. Hertle and M. F. Yeckel. Distribution of inositol-1,4,5-trisphosphate receptor isotypes and ryanodine receptor isotypes during maturation of the rat hippocampus. *Neuroscience*, 150(3):625–638.

[252] Lisa Foa and Robert Gasperini. Developmental roles for homer: more than just a pretty scaffold. *Journal of neurochemistry*, 108(1):1–10.

[253] P. H. C. Kremer, B. P. C. Koeleman, L. Pawlikowska, S. Weinsheimer, N. Bendjilali, S. Sidney, J. G. Zaroff, G. J. E. Rinkel, L. H. van den Berg, Y. M. Ruigrok, G. a. P. de Kort, J. H. Veldink, H. Kim, and C. J. M. Klijn. Evaluation of genetic risk loci for intracranial aneurysms in sporadic arteriovenous malformations of the brain. *Journal of Neurology, Neurosurgery, and Psychiatry*, 86(5):524–529.

[254] Katsuhito Yasuno, Kaya Bilguvar, Philippe Bijlenga, Siew-Kee Low, Boris Krischek, Georg Auburger, Matthias Simon, Dietmar Krex, Zulfikar Arlier, Nikhil Nayak, Ynte M. Ruigrok, Mika Niemelä, Atsushi Tajima, Mikael von und zu Fraunberg, Tamás Dóczi, Florentina Wirjatijasa, Akira Hata, Jordi Blasco, Agi Oszvald, Hidetoshi Kasuya, Gulam Zilani, Beate Schoch, Pankaj Singh, Carsten Stüer, Roelof Risselada, Jürgen Beck, Teresa Sola, Filomena Ricciardi, Arpo Aromaa, Thomas Illig, Stefan Schreiber, Cornelia M. van Duijn, Leonard H. van den Berg, Claire Perret, Carole Proust, Constantin Roder, Ali K. Ozturk, Emília Gaál, Daniela Berg, Christof Geisen, Christoph M. Friedrich, Paul Summers, Alejandro F. Frangi, Matthew W. State, H. Erich Wichmann, Monique M. B. Breteler, Cisca Wijmenga, Shrikant Mane, Leena Peltonen, Vivas Elio, Miriam C. J. M. Sturkenboom, Patricia Lawford, James Byrne, Juan Macho, Erol I. Sandalcioglu, Bernhard Meyer, Andreas Raabe, Helmuth Steinmetz, Daniel Rüfenacht, Juha E. Jääskeläinen, Juha Hernesniemi, Gabriel J. E. Rinkel, Hitoshi Zembutsu, Ituro Inoue, Aarno Palotie, François Cambien, Yusuke Nakamura, Richard P. Lifton, and Murat Günel. Genome-wide association study of intracranial aneurysm identifies three new risk loci. *Nature Genetics*, 42(5):420–425.

[255] DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, page gkw943, .

[256] Jinchen Li, Tao Cai, Yi Jiang, Huiqian Chen, Xin He, Chao Chen, Xianfeng Li, Qianzhi Shao, Xia Ran, Zhongshan Li, Kun Xia, Chunyu Liu, Zhong Sheng Sun, and Jinyu Wu.

Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Molecular Psychiatry*, 21(2):290–297, .

[257] Justin R Meyer, Devin T Dobias, Joshua S Weitz, Jeffrey E Barrick, Ryan T Quick, and Richard E Lenski. Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science (New York, N.Y.)*, 335(6067):428–32.

[258] Christoph Redies, Nicole Hertel, and Christian A. Hübner. Cadherins and neuropsychiatric disorders. *Brain Research*, 1470:130–144.

[259] Li-Ming Xu, Jia-Rui Li, Yue Huang, Min Zhao, Xing Tang, and Liping Wei. AutismKB: an evidence-based knowledgebase of autism genetics. *Nucleic Acids Research*, 40:D1016–1022.

[260] Taoyun Ji, Ye Wu, Huifang Wang, Jingmin Wang, and Yuwu Jiang. Diagnosis and fine mapping of a deletion in distal 11q in two chinese patients with developmental delay. *Journal of Human Genetics*, 55(8):486–489, .

[261] A. Iwaki, Y. Kawano, S. Miura, H. Shibata, D. Matsuse, W. Li, H. Furuya, Y. Ohyagi, T. Taniwaki, J. Kira, and Y. Fukumaki. Heterozygous deletion of ITPR1, but not SUMF1, in spinocerebellar ataxia type 16. *Journal of Medical Genetics*, 45(1):32–35.

[262] Joyce van de Leemput, Jayanth Chandran, Melanie A. Knight, Lynne A. Holtzclaw, Sonja Scholz, Mark R. Cookson, Henry Houlden, Katrina Gwinn-Hardy, Hon-Chung Fung, Xian Lin, Dena Hernandez, Javier Simon-Sanchez, Nick W. Wood, Paola Giunti, Ian Rafferty, John Hardy, Elsdon Storey, R. J. McKinlay Gardner, Susan M. Forrest, Elizabeth M. C. Fisher, James T. Russell, Huaibin Cai, and Andrew B. Singleton. Deletion at ITPR1 underlies ataxia in mice and spinocerebellar ataxia 15 in humans. *PLoS genetics*, 3(6): e108.

[263] E. O. Berglund and B. Ranscht. Molecular cloning and in situ localization of the human contactin gene (CNTN1) on chromosome 12q11-q12. *Genomics*, 21(3):571–582.

[264] Transgenic mice expressing f3/contactin from the TAG-1 promoter exhibit developmentally regulated changes in the differentiation of cerebellar neurons. *Development (Cambridge, England)*, 130(1):29–43.

[265] Elisabeth Stogmann, Eva Reinthaler, Salwa Eltawil, Mohammed A. El Etribi, Mahmoud Hemeda, Nevine El Nahhas, Ahmed M. Gaber, Amal Fouad, Sherif Edris, Anna Benet-Pages, Sebastian H. Eck, Ekaterina Pataraia, Davide Mei, Alexis Brice, Suzanne Lesage, Renzo Guerrini, Friedrich Zimprich, Tim M. Strom, and Alexander Zimprich. Autosomal recessive cortical myoclonic tremor and epilepsy: association with a mutation in the potassium channel associated gene CNTN2. *Brain: A Journal of Neurology*, 136:1155–1160.

[266] Malene B. Rasmussen, Jakob V. Nielsen, Charles M. Lourenço, Joana B. Melo, Christina Halgren, Camila V. L. Geraldi, Wilson Marques, Guilherme R. Rodrigues, Mads Thomassen, Mads Bak, Claus Hansen, Susana I. Ferreira, Margarida Venâncio, Karen F. Henriksen, Allan Lind-Thomsen, Isabel M. Carreira, Niels A. Jensen, and Niels Tommerup. Neurodevelopmental disorders associated with dosage imbalance of ZBTB20 correlate with the morbidity spectrum of ZBTB20 candidate target genes. *Journal of Medical Genetics*, 51(9):605–613.

[267] Ping Wang, Ping Zhang, Ji Huang, Min Li, and Xia Chen. Trichostatin a protects against cisplatin-induced ototoxicity by regulating expression of genes related to apoptosis and synaptic function. *Neurotoxicology*, 37:51–62, .

[268] Konstantin I. Piatkov, Jang-Hyun Oh, Yuan Liu, and Alexander Varshavsky. Calpain-generated natural protein fragments as short-lived substrates of the n-end rule pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 111(9): E817–826.

[269] Oliver P. Forman, Luisa De Risio, and Cathryn S. Mellersh. Missense mutation in CAPN1 is associated with spinocerebellar ataxia in the parson russell terrier dog breed. *PloS One*, 8(5):e64627.

[270] Yubin Wang, Joshua Hersheson, Dulce Lopez, Monia Hammer, Yan Liu, Ka-Hung Lee, Vanessa Pinto, Jeff Seinfeld, Sarah Wiethoff, Jiandong Sun, Rim Amouri, Faycal Hentati, Neema Baudry, Jennifer Tran, Andrew B. Singleton, Marie Coutelier, Alexis Brice, Giovanni Stevanin, Alexandra Durr, Xiaoning Bi, Henry Houlden, and Michel Baudry. Defects in the CAPN1 gene result in alterations in cerebellar development and cerebellar ataxia in mice and humans. *Cell Reports*, 16(1):79–91, .

[271] J. C. Engert, P. Bérubé, J. Mercier, C. Doré, P. Lepage, B. Ge, J. P. Bouchard, J. Mathieu, S. B. Melançon, M. Schalling, E. S. Lander, K. Morgan, T. J. Hudson, and A. Richter. ARSACS, a spastic ataxia common in northeastern québec, is caused by mutations in a new gene encoding an 11.5-kb ORF. *Nature Genetics*, 24(2):120–125.

[272] Mikko Muona, Samuel F. Berkovic, Leanne M. Dibbens, Karen L. Oliver, Snezana Maljevic, Marta A. Bayly, Tarja Joensuu, Laura Canafoglia, Silvana Franceschetti, Roberto Michelucci, Salla Markkinen, Sarah E. Heron, Michael S. Hildebrand, Eva Andermann, Frederick Andermann, Antonio Gambardella, Paolo Tinuper, Laura Licchetta, Ingrid E. Scheffer, Chiara Criscuolo, Alessandro Filla, Edoardo Ferlazzo, Jamil Ahmad, Adeel Ahmad, Betul Baykan, Edith Said, Meral Topcu, Patrizia Riguzzi, Mary D. King, Cigdem Ozkara, Danielle M. Andrade, Bernt A. Engelsen, Arielle Crespel, Matthias Lindenau, Ebba Lohmann, Veronica Saletti, João Massano, Michael Privitera, Alberto J. Espay,

Birgit Kauffmann, Michael Duchowny, Rikke S. Møller, Rachel Straussberg, Zaid Afawi, Bruria Ben-Zeev, Kaitlin E. Samocha, Mark J. Daly, Steven Petrou, Holger Lerche, Aarno Palotie, and Anna-Elina Lehesjoki. A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. *Nature Genetics*, 47(1):39–46.

[273] Zaid Afawi, Karen L. Oliver, Sara Kivity, Aziz Mazarib, Ilan Blatt, Miriam Y. Neufeld, Katherine L. Helbig, Hadassa Goldberg-Stern, Adel J. Misk, Rachel Straussberg, Simri Walid, Muhammad Mahajnah, Tally Lerman-Sagie, Bruria Ben-Zeev, Esther Kahana, Rafik Masalha, Uri Kramer, Dana Ekstein, Zamir Shorer, Robyn H. Wallace, Marie Mangelsdorf, James N. MacPherson, Gemma L. Carvill, Heather C. Mefford, Graeme D. Jackson, Ingrid E. Scheffer, Melanie Bahlo, Jozef Gecz, Sarah E. Heron, Mark Corbett, John C. Mulley, Leanne M. Dibbens, Amos D. Korczyn, and Samuel F. Berkovic. Multiplex families with epilepsy: Success of clinical and molecular genetic characterization. *Neurology*, 86(8): 713–722.

[274] Edward C. Cooper. Made for "anchorin": Kv7.2/7.3 (KCNQ2/KCNQ3) channels and the modulation of neuronal excitability in vertebrate axons. *Seminars in Cell & Developmental Biology*, 22(2):185–192.

[275] A. M. Goldman, E. Glasscock, J. Yoo, T. T. Chen, T. L. Klassen, and J. L. Noebels. Arrhythmia in heart and brain: KCNQ1 mutations link epilepsy and sudden unexplained death. *Science Translational Medicine*, 1(2):2ra6.

[276] Sylvie Gerber, Kamil J. Alzayady, Lydie Burglen, Dominique Brémond-Gignac, Valentina Marchesin, Olivier Roche, Marlène Rio, Benoit Funalot, Raphaël Calmon, Alexandra Durr, Vera Lucia Gil-da Silva-Lopes, Maria Fernanda Ribeiro Bittar, Christophe Orssaud, Bénédicte Héron, Edward Ayoub, Patrick Berquin, Nadia Bahi-Buisson, Christine Bole, Cécile Masson, Arnold Munnich, Matias Simons, Marion Delous, Helene Dollfus, Nathalie Boddaert, Stanislas Lyonnet, Josseline Kaplan, Patrick Calvas, David I. Yule, Jean-Michel Rozet, and Lucas Fares Taie. Recessive and dominant de novo ITPR1 mutations cause gillespie syndrome. *American Journal of Human Genetics*, 98(5):971–980.

[277] Mike A. Nalls, Cory Y. McLean, Jacqueline Rick, Shirley Eberly, Samantha J. Hutten, Katrina Gwinn, Margaret Sutherland, Maria Martinez, Peter Heutink, Nigel M. Williams, John Hardy, Thomas Gasser, Alexis Brice, T. Ryan Price, Aude Nicolas, Margaux F. Keller, Cliona Molony, J. Raphael Gibbs, Alice Chen-Plotkin, Eunran Suh, Christopher Letson, Massimo S. Fiandaca, Mark Mapstone, Howard J. Federoff, Alastair J. Noyce, Huw Morris, Vivianna M. Van Deerlin, Daniel Weintraub, Cyrus Zabetian, Dena G. Hernandez, Suzanne Lesage, Meghan Mullins, Emily Drabant Conley, Carrie A. M. Northover, Mark Frasier, Ken Marek, Aaron G. Day-Williams, David J. Stone, John P. A. Ioannidis, Andrew B. Singleton, and Parkinson's Disease Biomarkers Program and Parkinson's Progression Marker

Initiative investigators. Diagnosis of parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study. *The Lancet. Neurology*, 14(10): 1002–1009, .

[278] Celia Kun-Rodrigues, Christos Ganos, Rita Guerreiro, Susanne A. Schneider, Claudia Schulte, Suzanne Lesage, Lee Darwent, Peter Holmans, Andrew Singleton, International Parkinson's Disease Genomics Consortium (IPDGC), Kailash Bhatia, and Jose Bras. A systematic screening to identify de novo mutations causing sporadic early-onset parkinson's disease. *Human Molecular Genetics*, 24(23):6711–6720.

[279] Jose M. Bras and A. B. Singleton. Exome sequencing in parkinson's disease. *Clinical Genetics*, 80(2):104–109.

[280] Andrew Singleton and John Hardy. The evolution of genetics: Alzheimer's and parkinson's diseases. *Neuron*, 90(6):1154–1163.

[281] Yu-Hsuan Chuang, Kimberly C. Paul, Jeff M. Bronstein, Yvette Bordelon, Steve Horvath, and Beate Ritz. Parkinson's disease is associated with DNA methylation levels in human blood and saliva. *Genome medicine*, 9(1):76.

[282] Ying-Chao Lin, Ai-Ru Hsieh, Ching-Lin Hsiao, Shang-Jung Wu, Hui-Min Wang, Ie-Bin Lian, and Cathy S J Fann. Identifying rare and common disease associated variants in genomic data using parkinson's disease as a model. *J Biomed Sci*, 21:88–88.

[283] Loukas Moutsianas, Vineeta Agarwala, Christian Fuchsberger, Jason Flannick, Manuel A. Rivas, Kyle J. Gaulton, Patrick K. Albers, GoT2D Consortium, Gil McVean, Michael Boehnke, David Altshuler, and Mark I. McCarthy. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLOS Genetics*, 11(4):e1005165.

[284] Shaun M Purcell, Jennifer L Moran, Menachem Fromer, Douglas Ruderfer, Nadia Solovieff, Panos Roussos, Colm O'Dushlaine, Kimberly Chambert, Sarah E Bergen, Anna Kähler, Laramie Duncan, Eli Stahl, Giulio Genovese, Esperanza Fernández, Mark O Collins, Noboru H Komiyama, Jyoti S Choudhary, Patrik K E Magnusson, Eric Banks, Khalid Shakir, Kiran Garimella, Tim Fennell, Mark DePristo, Seth G N Grant, Stephen J Haggarty, Stacey Gabriel, Edward M Scolnick, Eric S Lander, Christina M Hultman, Patrick F Sullivan, Steven A McCarroll, and Pamela Sklar. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487):185–90, .

[285] Costin Leu, Simona Balestrini, Bridget Maher, Laura Hernández-Hernández, Padhraig Gormley, Eija Hämäläinen, Kristin Heggeli, Natasha Schoeler, Jan Novy, Joseph Willis, Vincent Plagnol, Rachael Ellis, Eleanor Reavey, Mary O'Regan, William O. Pickrell,

Rhys H. Thomas, Seo-Kyung Chung, Norman Delanty, Jacinta M. McMahon, Stephen Malone, Lynette G. Sadleir, Samuel F. Berkovic, Lina Nashef, Sameer M. Zuberi, Mark I. Rees, Gianpiero L. Cavalleri, Josemir W. Sander, Elaine Hughes, J. Helen Cross, Ingrid E. Scheffer, Aarno Palotie, and Sanjay M. Sisodiya. Genome-wide polygenic burden of rare deleterious variants in sudden unexpected death in epilepsy. *EBioMedicine*, 2(9):1063–1070, .

[286] Parkinson Progression Marker Initiative. The parkinson progression marker initiative (PPMI). *Progress in Neurobiology*, 95(4):629–635.

[287] Mike A Nalls, Margaux F Keller, Dena G Hernandez, Lan Chen, David J Stone, and Andrew B Singleton. Baseline genetic associations in the parkinson's progression markers initiative (PPMI). *Movement disorders : official journal of the Movement Disorder Society*, 31(1):79–85, .

[288] Mike A Nalls, Nathan Pankratz, Christina M Lill, Chuong B Do, Dena G Hernandez, Mohamad Saad, Anita L DeStefano, Eleanna Kara, Jose Bras, Manu Sharma, Claudia Schulte, Margaux F Keller, Sampath Arepalli, Christopher Letson, Connor Edsall, Hreinn Stefansson, Xinmin Liu, Hannah Pliner, Joseph H Lee, Rong Cheng, M Arfan Ikram, John P A Ioannidis, Georgios M Hadjigeorgiou, Joshua C Bis, Maria Martinez, Joel S Perlmutter, Alison Goate, Karen Marder, Brian Fiske, Margaret Sutherland, Georgia Xiromerisiou, Richard H Myers, Lorraine N Clark, Kari Stefansson, John A Hardy, Peter Heutink, Honglei Chen, Nicholas W Wood, Henry Houlden, Haydeh Payami, Alexis Brice, William K Scott, Thomas Gasser, Lars Bertram, Nicholas Eriksson, Tatiana Foroud, and Andrew B Singleton. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for parkinson's disease. *Nature genetics*, 46(9):989–93, .

[289] Ivo D. Dinov, Ben Heavner, Ming Tang, Gustavo Glusman, Kyle Chard, Mike Darcy, Ravi Madduri, Judy Pa, Cathie Spino, Carl Kesselman, Ian Foster, Eric W. Deutsch, Nathan D. Price, John D. Van Horn, Joseph Ames, Kristi Clark, Leroy Hood, Benjamin M. Hampstead, William Dauer, and Arthur W. Toga. Predictive big data analytics: A study of parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLoS ONE*, 11(8).

[290] Seyed-Mohammad Fereshtehnejad, Yashar Zeighami, Alain Dagher, and Ronald B. Postuma. Clinical criteria for subtyping parkinson's disease: biomarkers and longitudinal progression. *Brain*, 140(7):1959–1976.

[291] Excessive burden of lysosomal storage disorder gene variants in parkinson's disease. *Brain*, 140(12):3191–3203.

[292] Julia C. Fitzgerald, Alexander Zimprich, Carvajal Berrio, Daniel A, Kevin M. Schindler, Brigitte Maurer, Claudia Schulte, Christine Bus, Anne-Kathrin Hauser, Manuela Kübler, Rahel Lewin, Dheeraj Reddy Bobbili, Lisa M. Schwarz, Evangelia Vartholomaiou, Kathrin Brockmann, Richard Wüst, Johannes Madlung, Alfred Nordheim, Olaf Riess, L. Miguel Martins, Enrico Glaab, Patrick May, Katja Schenke-Layland, Didier Picard, Manu Sharma, Thomas Gasser, and Rejko Krüger. Metformin reverses TRAP1 mutation-associated alterations in mitochondrial function in parkinson's disease. *Brain*, 140(9):2444–2459.

[293] Cynthia Sandor, Frantisek Honti, Wilfried Haerty, Konrad Szewczyk-Krolikowski, Paul Tomlinson, Sam Evetts, Stephanie Millin, Thomas Keane, Shane A. McCarthy, Richard Durbin, Kevin Talbot, Michele Hu, Caleb Webber, Chris P. Ponting, and Richard Wade-Martins. Whole-exome sequencing of 228 patients with sporadic parkinson's disease. *Scientific Reports*, 7:41188.

[294] Xiao Ji, Rachel L. Kember, Christopher D. Brown, and Maja Bućan. Increased burden of deleterious variants in essential genes in autism spectrum disorder. *Proceedings of the National Academy of Sciences*, 113(52):15054–15059, .

[295] Ryan KC Yuen, Bhooma Thiruvahindrapuram, Daniele Merico, Susan Walker, Kristiina Tammimies, Ny Hoang, Christina Chrysler, Thomas Nalpathamkalam, Giovanna Pellecchia, and Yi Liu. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nature medicine*, 21(2):185–191.

[296] Loes M Olde Loohuis, Jacob A S Vorstman, Anil P Ori, Kim A Staats, Tina Wang, Alexander L Richards, Ganna Leonenko, James T Walters, Joseph DeYoung, Rita M Cantor, and Roel A Ophoff. Genome-wide burden of deleterious coding variants increased in schizophrenia. *Nature communications*, 6:7501–7501.

[297] Jack Euesden, Cathryn M. Lewis, and Paul F. O'Reilly. PRSice: Polygenic risk score software. *Bioinformatics (Oxford, England)*, 31(9):1466–1468.

[298] Excess of rare novel loss-of-function variants in synaptic genes in schizophrenia and autism spectrum disorders. *Molecular psychiatry*, 19(8):872–9.

[299] Kazuhiro A. Fujita, Marek Ostaszewski, Yukiko Matsuoka, Samik Ghosh, Enrico Glaab, Christophe Trefois, Isaac Crespo, Thanneer M. Perumal, Wiktor Jurkowski, Paul M. A. Antony, Nico Diederich, Manuel Buttini, Akihiko Kodama, Venkata P. Satagopam, Serge Eifes, Antonio Del Sol, Reinhard Schneider, Hiroaki Kitano, and Rudi Balling. Integrating pathways of parkinson's disease in a molecular interaction map. *Molecular Neurobiology*, 49(1):88–102.

[300] Serge Przedborski. The two-century journey of parkinson disease research. *Nature Reviews Neuroscience*, 18(4):251–259.

[301] Alberto J. Espay, Michael A. Schwarzschild, Caroline M. Tanner, Hubert H. Fernandez, David K. Simon, James B. Leverenz, Aristide Merola, Alice Chen-Plotkin, Patrik Brundin, Marcelo A. Kauffman, Roberto Erro, Karl Kieburtz, Daniel Woo, Eric A. Macklin, David G. Standaert, and Anthony E. Lang. Biomarker-driven phenotyping in parkinson's disease: A translational missing link in disease-modifying clinical trials. *Movement Disorders*, 32(3): 319–324.

[302] Rejko Krüger, Jochen Klucken, Daniel Weiss, Lars Tönges, Pierre Kolber, Stefan Unterecker, Michael Lorrain, Horst Baas, Thomas Müller, and Peter Riederer. Classification of advanced stages of parkinson's disease: translation into stratified treatments. *Journal of Neural Transmission*, 124(8):1015–1027.

[303] Paul M. A. Antony, Nico J. Diederich, Rejko Krüger, and Rudi Balling. The hallmarks of parkinson's disease. *FEBS Journal*, 280(23):5981–5993.

[304] Brenna Cholerton, Eric B. Larson, Joseph F. Quinn, Cyrus P. Zabetian, Ignacio F. Mata, C. Dirk Keene, Margaret Flanagan, Paul K. Crane, Thomas J. Grabowski, Kathleen S. Montine, and Thomas J. Montine. Precision medicine: Clarity for the complexity of dementia. *The American Journal of Pathology*, 186(3):500–506.

[305] Rosanna Asselta, Valeria Rimoldi, Chiara Siri, Roberto Cilia, Ilaria Guella, Silvana Tesei, Giulia Soldà, Gianni Pezzoli, Stefano Duga, and Stefano Goldwurm. Glucocerebrosidase mutations in primary parkinsonism. *Parkinsonism & Related Disorders*, 20(11):1215–1220.

[306] Jose M. Bras, Rita J. Guerreiro, James T.H. Teo, Lee Darwent, Jenny Vaughan, Sophie Molloy, John Hardy, and Susanne A. Schneider. Atypical parkinsonism-dystonia syndrome caused by a novel DJ1 mutation. *Movement Disorders Clinical Practice*, 1(1):45–49.

[307] Farzaneh Ghazavi, Zeinab Fazlali, Setareh Sadat Banihosseini, Sayed-Rzgar Hosseini, Mohammad Hossein Kazemi, Seyedmehdi Shojaee, Khosro Parsa, Homa Sadeghi, Farzad Sina, Mohammad Rohani, Gholam-Ali Shahidi, Nasser Ghaemi, Mostafa Ronaghi, and Elahe Elahi. PRKN, DJ-1, and PINK1 screening identifies novel splice site mutation in PRKN and two novel DJ-1 mutations. *Movement Disorders*, 26(1):80–89.

[308] Rachel Soemedi, Kamil J. Cygan, Christy L. Rhine, Jing Wang, Charlston Bulacan, John Yang, Pinar Bayrak-Toydemir, Jamie McDonald, and William G. Fairbrother. Pathogenic variants that alter protein code often disrupt splicing. *Nature Genetics*, 49(6):848–855.

[309] Eirini Marouli, Mariaelisa Graff, Carolina Medina-Gomez, Ken Sin Lo, Andrew R. Wood, Troels R. Kjaer, Rebecca S. Fine, Yingchang Lu, Claudia Schurmann, Heather M. Highland, Sina Rüeger, Gudmar Thorleifsson, Anne E. Justice, David Lamparter, Kathleen E. Stirrups, Valérie Turcot, Kristin L. Young, Thomas W. Winkler, Tõnu Esko, Tugce Karaderi, Adam E. Locke, Nicholas G. D. Masca, Maggie C. Y. Ng, Poorva Mudgal, Manuel A. Rivas, Sailaja Vedantam, Anubha Mahajan, Xiuqing Guo, Goncalo Abecasis, Katja K. Aben, Linda S. Adair, Dewan S. Alam, Eva Albrecht, Kristine H. Allin, Matthew Allison, Philippe Amouyel, Emil V. Appel, Dominique Arveiler, Folkert W. Asselbergs, Paul L. Auer, Beverley Balkau, Bernhard Banas, Lia E. Bang, Marianne Benn, Sven Bergmann, Lawrence F. Bielak, Matthias Blüher, Heiner Boeing, Eric Boerwinkle, Carsten A. Böger, Lori L. Bonnycastle, Jette Bork-Jensen, Michiel L. Bots, Erwin P. Bottinger, Donald W. Bowden, Ivan Brandslund, Gerome Breen, Murray H. Brilliant, Linda Broer, Amber A. Burt, Adam S. Butterworth, David J. Carey, Mark J. Caulfield, John C. Chambers, Daniel I. Chasman, Yii-Der Ida Chen, Rajiv Chowdhury, Cramer Christensen, Audrey Y. Chu, Massimiliano Cocca, Francis S. Collins, James P. Cook, Janie Corley, Jordi Corominas Galbany, Amanda J. Cox, Gabriel Cuellar-Partida, John Danesh, Gail Davies, Paul I. W. de Bakker, Gert J. de Borst, Simon de Denus, Mark C. H. de Groot, Renée de Mutsert, Ian J. Deary, George Dedoussis, Ellen W. Demerath, Anneke I. den Hollander, Joe G. Dennis, Emanuele Di Angelantonio, Fotios Drenos, Mengmeng Du, Alison M. Dunning, Douglas F. Easton, Tapani Ebeling, Todd L. Edwards, Patrick T. Ellinor, Paul Elliott, Evangelos Evangelou, Aliki-Eleni Farmaki, Jessica D. Faul, Mary F. Feitosa, Shuang Feng, Ele Ferrannini, Marco M. Ferrario, Jean Ferrieres, Jose C. Florez, Ian Ford, Myriam Fornage, Paul W. Franks, Ruth Frikke-Schmidt, Tessel E. Galesloot, Wei Gan, Ilaria Gandin, Paolo Gasparini, Vilmantas Giedraitis, Ayush Giri, Giorgia Girotto, Scott D. Gordon, Penny Gordon-Larsen, Mathias Gorski, Niels Grarup, Megan L. Grove, Vilmundur Gudnason, Stefan Gustafsson, Torben Hansen, Kathleen Mullan Harris, Tamara B. Harris, Andrew T. Hattersley, Caroline Hayward, Liang He, Iris M. Heid, Kauko Heikkilä, Øyvind Helgeland, Jussi Hernesniemi, Alex W. Hewitt, Lynne J. Hocking, Mette Hollensted, Oddgeir L. Holmen, G. Kees Hovingh, Joanna M. M. Howson, Carel B. Hoyng, Paul L. Huang, Kristian Hveem, M. Arfan Ikram, Erik Ingelsson, Anne U. Jackson, Jan-Håkan Jansson, Gail P. Jarvik, Gorm B. Jensen, Min A. Jhun, Yucheng Jia, Xuejuan Jiang, Stefan Johansson, Marit E. Jørgensen, Torben Jørgensen, Pekka Jousilahti, J. Wouter Jukema, Bratati Kahali, René S. Kahn, Mika Kähönen, Pia R. Kamstrup, Stavroula Kanoni, Jaakko Kaprio, Maria Karaleftheri, Sharon L. R. Kardia, Fredrik Karpe, Frank Kee, Renske Keeman, Lambertus A. Kiemeney, Hidetoshi Kitajima, Kirsten B. Kluivers, Thomas Kocher, Pirjo Komulainen, Jukka Kontto, Jaspal S. Kooner, Charles Kooperberg, Peter Kovacs, Jennifer Kriebel, Helena Kuivaniemi, Sébastien Küry, Johanna Kuusisto, Martina La Bianca, Markku Laakso, Timo A. Lakka, Ethan M. Lange, Leslie A. Lange, Carl D. Langefeld,

Claudia Langenberg, Eric B. Larson, I.-Te Lee, Terho Lehtimäki, Cora E. Lewis, Huaixing Li, Jin Li, Ruifang Li-Gao, Honghuang Lin, Li-An Lin, Xu Lin, Lars Lind, Jaana Lindström, Allan Linneberg, Yeheng Liu, Yongmei Liu, Artitaya Lophatananon, Jian'an Luan, Steven A. Lubitz, Leo-Pekka Lyytikäinen, David A. Mackey, Pamela A. F. Madden, Alisa K. Manning, Satu Männistö, Gaëlle Marenne, Jonathan Marten, Nicholas G. Martin, Angela L. Mazul, Karina Meidtner, Andres Metspalu, Paul Mitchell, Karen L. Mohlke, Dennis O. Mook-Kanamori, Anna Morgan, Andrew D. Morris, Andrew P. Morris, Martina Müller-Nurasyid, Patricia B. Munroe, Mike A. Nalls, Matthias Nauck, Christopher P. Nelson, Matt Neville, Sune F. Nielsen, Kjell Nikus, Pål R. Njølstad, Børge G. Nordestgaard, Ioanna Ntalla, Jeffrey R. O'Connel, Heikki Oksa, Loes M. Olde Loohuis, Roel A. Ophoff, Katharine R. Owen, Chris J. Packard, Sandosh Padmanabhan, Colin N. A. Palmer, Gerard Pasterkamp, Aniruddh P. Patel, Alison Pattie, Oluf Pedersen, Peggy L. Peissig, Gina M. Peloso, Craig E. Pennell, Markus Perola, James A. Perry, John R. B. Perry, Thomas N. Person, Ailith Pirie, Ozren Polasek, Danielle Posthuma, Olli T. Raitakari, Asif Rasheed, Rainer Rauramaa, Dermot F. Reilly, Alex P. Reiner, Frida Renström, Paul M. Ridker, John D. Rioux, Neil Robertson, Antonietta Robino, Olov Rolandsson, Igor Rudan, Katherine S. Ruth, Danish Saleheen, Veikko Salomaa, Nilesh J. Samani, Kevin Sandow, Yadav Sapkota, Naveed Sattar, Marjanka K. Schmidt, Pamela J. Schreiner, Matthias B. Schulze, Robert A. Scott, Marcelo P. Segura-Lepe, Svati Shah, Xueling Sim, Suthesh Sivapalaratnam, Kerrin S. Small, Albert Vernon Smith, Jennifer A. Smith, Lorraine Southam, Timothy D. Spector, Elizabeth K. Speliotes, John M. Starr, Valgerdur Steinthorsdottir, Heather M. Stringham, Michael Stumvoll, Praveen Surendran, Leen M. 't Hart, Katherine E. Tansey, Jean-Claude Tardif, Kent D. Taylor, Alexander Teumer, Deborah J. Thompson, Unnur Thorsteinsdottir, Betina H. Thuesen, Anke Tönjes, Gerard Tromp, Stella Trompet, Emmanouil Tsafantakis, Jaakko Tuomilehto, Anne Tybjaerg-Hansen, Jonathan P. Tyrer, Rudolf Uher, André G. Uitterlinden, Sheila Ulivi, Sander W. van der Laan, Andries R. Van Der Leij, Cornelia M. van Duijn, Natasja M. van Schoor, Jessica van Setten, Anette Varbo, Tibor V. Varga, Rohit Varma, Digna R. Velez Edwards, Sita H. Vermeulen, Henrik Vestergaard, Veronique Vitart, Thomas F. Vogt, Diego Vozzi, Mark Walker, Feijie Wang, Carol A. Wang, Shuai Wang, Yiqin Wang, Nicholas J. Wareham, Helen R. Warren, Jennifer Wessel, Sara M. Willems, James G. Wilson, Daniel R. Witte, Michael O. Woods, Ying Wu, Hanieh Yaghootkar, Jie Yao, Pang Yao, Laura M. Yerges-Armstrong, Robin Young, Eleftheria Zeggini, Xiaowei Zhan, Weihua Zhang, Jing Hua Zhao, Wei Zhao, Wei Zhao, He Zheng, Wei Zhou, The EPIC-InterAct Consortium, CHD Exome+ Consortium, ExomeBP Consortium, T2D-Genes Consortium, GoT2D Genes Consortium, Global Lipids Genetics Consortium, ReproGen Consortium, Magic Investigators, Jerome I. Rotter, Michael Boehnke, Sekar Kathiresan, Mark I. McCarthy, Cristen J. Willer, Kari Stefansson, Ingrid B. Borecki, Dajiang J. Liu, Kari E. North, Nancy L. Heard-Costa, Tune H.

Pers, Cecilia M. Lindgren, Claus Oxvig, Zoltán Kutalik, Fernando Rivadeneira, Ruth J. F. Loos, Timothy M. Frayling, Joel N. Hirschhorn, Panos Deloukas, and Guillaume Lettre. Rare and low-frequency coding variants alter human adult height. *Nature*, 542(7640):186–190.

[310] Gene Yeo and Christopher B. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 11(2):377–394.

[311] Xueqiu Jian, Eric Boerwinkle, and Xiaoming Liu. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Research*, 42(22):13534–13544.

[312] Matthew D. Shirley, Zhaorong Ma, Brent S. Pedersen, and Sarah J. Wheelan. Efficient ”pythonic” access to FASTA files using pyfaidx.

[313] Barbara Wappenschmidt, Alexandra A. Becker, Jan Hauke, Ute Weber, Stefanie Engert, Juliane Köhler, Karin Kast, Norbert Arnold, Kerstin Rhiem, Eric Hahnen, Alfons Meindl, and Rita K. Schmutzler. Analysis of 30 putative BRCA1 splicing mutations in hereditary breast and ovarian cancer families identifies exonic splice site mutations that escape in silico prediction. *PLOS ONE*, 7(12):e50800.

[314] Malka Nissim-Rafinia and Batsheva Kerem. Splicing regulation as a potential genetic modifier. *Trends in genetics: TIG*, 18(3):123–127.

[315] Valentina La Cognata, Velia D’Agata, Francesca Cavalcanti, and Sebastiano Cavallaro. Splicing: is there an alternative contribution to parkinson’s disease? *Neurogenetics*, 16:245–263, .

[316] Nicholas J. Schork. Personalized medicine: Time for one-person trials. *Nature News*, 520 (7549):609.

[317] T. Kitada, S. Asakawa, N. Hattori, H. Matsumine, Y. Yamamura, S. Minoshima, M. Yokochi, Y. Mizuno, and N. Shimizu. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature*, 392(6676):605–608.

[318] Coro Paisán-Ruíz, Shushant Jain, E. Whitney Evans, William P. Gilks, Javier Simón, Marcel van der Brug, Adolfo López de Munain, Silvia Aparicio, Angel Martínez Gil, Naheed Khan, Janel Johnson, Javier Ruiz Martinez, David Nicholl, Itxaso Marti Carrera, Amets Saénz Pena, Rohan de Silva, Andrew Lees, José Félix Martí-Massó, Jordi Pérez-Tur, Nick W. Wood, and Andrew B. Singleton. Cloning of the gene containing mutations that cause PARK8-linked parkinson’s disease. *Neuron*, 44(4):595–600.

[319] Alexander Zimprich, Anna Benet-Pagès, Walter Struhal, Elisabeth Graf, Sebastian H. Eck, Marc N. Offman, Dietrich Haubenberger, Sabine Spielberger, Eva C. Schulte, Peter Lichtner, Shaila C. Rossle, Norman Klopp, Elisabeth Wolf, Klaus Seppi, Walter Pirker, Stefan Presslauer, Brit Mollenhauer, Regina Katzenschlager, Thomas Foki, Christoph Hotzy, Eva Reinthaler, Ashot Harutyunyan, Robert Kralovics, Annette Peters, Fritz Zimprich, Thomas Brücke, Werner Poewe, Eduard Auff, Claudia Trenkwalder, Burkhard Rost, Gerhard Ransmayr, Juliane Winkelmann, Thomas Meitinger, and Tim M. Strom. A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset parkinson disease. *American Journal of Human Genetics*, 89(1):168–175, .

[320] Janice L. Farlow, Laurie A. Robak, Kurt Hetrick, Kevin Bowling, Eric Boerwinkle, Zeynep H. Coban-Akdemir, Tomasz Gambin, Richard A. Gibbs, Shen Gu, Preti Jain, Joseph Jankovic, Shalini Jhangiani, Kaveeta Kaw, Dongbing Lai, Hai Lin, Hua Ling, Yunlong Liu, James R. Lupski, Donna Muzny, Paula Porter, Elizabeth Pugh, Janson White, Kimberly Doheny, Richard M. Myers, Joshua M. Shulman, and Tatiana Foroud. Whole-exome sequencing in familial parkinson disease. *JAMA neurology*, 73(1):68–75.

[321] Eva C. Schulte, Immanuel Stahl, Darina Czamara, Daniel C. Ellwanger, Sebastian Eck, Elisabeth Graf, Brit Mollenhauer, Alexander Zimprich, Peter Lichtner, Dietrich Haubenberger, Walter Pirker, Thomas Brücke, Benjamin Bereznai, Maria J. Molnar, Annette Peters, Christian Gieger, Bertram Müller-Myhsok, Claudia Trenkwalder, and Juliane Winkelmann. Rare variants in PLXNA4 and parkinson's disease. *PLoS ONE*, 8(11).

[322] Julian Schubert, Aleksandra Siekierska, Mélanie Langlois, Patrick May, Clément Huneau, Felicitas Becker, Hiltrud Muhle, Arvid Suls, Johannes R Lemke, Carolien G F de Kovel, Holger Thiele, Kathryn Konrad, Amit Kawalia, Mohammad R Toliat, Thomas Sander, Franz Rüschendorf, Almuth Caliebe, Inga Nagel, Bernard Kohl, Angela Kecskés, Maxime Jacmin, Katia Hardies, Sarah Weckhuysen, Erik Riesch, Thomas Dorn, Eva H Brilstra, Stephanie Baulac, Rikke S Møller, Helle Hjalgrim, Bobby P C Koeleman, Karin Jurkat-Rott, Frank Lehman-Horn, Jared C Roach, Gustavo Glusman, Leroy Hood, David J Galas, Benoit Martin, Peter A M de Witte, Saskia Biskup, Peter De Jonghe, Ingo Helbig, Rudi Balling, Peter Nürnberg, Alexander D Crawford, Camila V Esguerra, Yvonne G Weber, and Holger Lerche. Mutations in STX1b, encoding a presynaptic protein, cause fever-associated epilepsy syndromes. *Nature genetics*, 46(12):1327–32.

[323] Jared C Roach, Gustavo Glusman, Arian F a Smit, Chad D Huff, Robert Hubley, Paul T Shannon, Lee Rowen, Krishna P Pant, Nathan Goodman, Michael Bamshad, Jay Shendure, Radoje Drmanac, Lynn B Jorde, Leroy Hood, and David J Galas. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science (New York, N.Y.)*, 328:636–639, .

[324] Jared C. Roach, Gustavo Glusman, Robert Hubley, Stephen Z. Montsaroff, Alisha K. Holloway, Denise E. Mauldin, Deepak Srivastava, Vidu Garg, Katherine S. Pollard, David J. Galas, Leroy Hood, and Arian F a Smit. Chromosomal haplotypes by genetic phasing of human families. *American Journal of Human Genetics*, 89:382–397, .

[325] Hinco J. Gierman, Kristen Fortney, Jared C. Roach, Natalie S. Coles, Hong Li, Gustavo Glusman, Glenn J. Markov, Justin D. Smith, Leroy Hood, L. Stephen Coles, and Stuart K. Kim. Whole-genome sequencing of the world's oldest people. *PLOS ONE*, 9(11):e112430.

[326] Brian L. Browning and Sharon R. Browning. A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics*, 88(2):173–182.

[327] Alexander Gusev, Jennifer K. Lowe, Markus Stoffel, Mark J. Daly, David Altshuler, Jan L. Breslow, Jeffrey M. Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2):318–326.

[328] Hui Yang, Peter N. Robinson, and Kai Wang. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature Methods*, 12(9):841–843.

[329] Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, 83(5):610–615.

[330] Chuong B. Do, Joyce Y. Tung, Elizabeth Dorfman, Amy K. Kiefer, Emily M. Drabant, Uta Francke, Joanna L. Mountain, Samuel M. Goldman, Caroline M. Tanner, J. William Langston, Anne Wojcicki, and Nicholas Eriksson. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for parkinson's disease. *PLOS Genetics*, 7(6):e1002141.

[331] Laura Dunn, George Fg Allen, Adamantios Mamais, Helen Ling, Abi Li, Kate E. Duberley, Iain P. Hargreaves, Simon Pope, Janice L. Holton, Andrew Lees, Simon J. Heales, and Rina Bandopadhyay. Dysregulation of glucose metabolism is an early event in sporadic parkinson's disease. *Neurobiology of Aging*, 35(5):1111–1115.

[332] Shanshan Gao, Chunli Duan, Ge Gao, Xiaoming Wang, and Hui Yang. Alpha-synuclein overexpression negatively regulates insulin receptor substrate 1 by activating mTORC1/s6k1 signaling. *The International Journal of Biochemistry & Cell Biology*, 64: 25–33.

[333] Hongbin Huang, Cong Peng, Yong Liu, Xu Liu, Qicong Chen, and Zunnan Huang. Genetic association of NOS1 exon18, NOS1 exon29, ABCB1 1236c/t, and ABCB1 3435c/t polymorphisms with the risk of parkinson's disease. *Medicine*, 95(40).

[334] Terrie Rife, Bareza Rasoul, Nicholas Pullen, David Mitchell, Kristen Grathwol, and Janice Kurth. The effect of a promoter polymorphism on the transcription of nitric oxide synthase 1 and its relevance to parkinson's disease. *Journal of Neuroscience Research*, 87(10):2319–2325.

[335] F. Freudenberg, A. Alttoa, and A. Reif. Neuronal nitric oxide synthase (NOS1) and its adaptor, NOS1ap, as a genetic risk factors for psychiatric disorders. *Genes, Brain and Behavior*, 14(1):46–63.

[336] Nevyana Ivanova, Valentina Peycheva, Kunka Kamenarova, Dalia Kancheva, Irina Tsekova, Iliana Aleksandrova, Dimitrina Hristova, Ivan Litvinenko, Diana Todorova, Gergana Sarailieva, Petya Dimova, Veselin Tomov, Veneta Bozhinova, Vanio Mitev, Radka Kaneva, and Albena Jordanova. Three novel SLC2a1 mutations in bulgarian patients with different forms of genetic generalized epilepsy reflecting the clinical and genetic diversity of GLUT1-deficiency syndrome. *Seizure*, 54:41–44.

[337] Dong Wang, Juan M. Pascual, and Darryl De Vivo. Glucose transporter type 1 deficiency syndrome. In Margaret P. Adam, Holly H. Ardinger, Roberta A. Pagon, Stephanie E. Wallace, Lora JH Bean, Karen Stephens, and Anne Amemiya, editors, *GeneReviews®*. University of Washington, Seattle, .

[338] Karin Writzl, Zvonka Rener Primec, Barbara Gnidovec Stražišar, Damjan Osredkar, Nuška Pečarič-Meglič, Branka Stirn Kranjc, Kiyomi Nishiyama, Naomichi Matsumoto, and Hirotomo Saitsu. Early onset west syndrome with severe hypomyelination and coloboma-like optic discs in a girl with SPTAN1 mutation. *Epilepsia*, 53(6):e106–110.

[339] Yutaka Nonoda, Yoshiaki Saito, Shigehiro Nagai, Masayuki Sasaki, Toshiyuki Iwasaki, Naomichi Matsumoto, Masahiro Ishii, and Hirotomo Saitsu. Progressive diffuse brain atrophy in west syndrome with marked hypomyelination due to SPTAN1 gene mutation. *Brain & Development*, 35(3):280–283.

[340] Carla F. Bento, Avraham Ashkenazi, Maria Jimenez-Sanchez, and David C. Rubinsztein. The parkinson's disease-associated genes ATP13a2 and SYT11 regulate autophagy via a common pathway. *Nature Communications*, 7:11803.

[341] Anamika Giri, Kin Y. Mok, Iris Jansen, Manu Sharma, Christelle Tesson, Graziella Mangone, Suzanne Lesage, José M. Bras, Joshua M. Shulman, Una-Marie Sheerin, International Parkinson's Disease Consortium (IPDGC), Mónica Díez-Fairen, Pau Pastor, María José Martí, Mario Ezquerra, Eduardo Tolosa, Leonor Correia-Guedes, Joaquim Ferreira, Najaf Amin, Cornelia M. van Duijn, Jeroen van Rooij, André G. Uitterlinden, Robert Kraaij, Michael Nalls, and Javier Simón-Sánchez. Lack of evidence for a role of genetic variation

in TMEM230 in the risk for parkinson's disease in the caucasian population. *Neurobiology of Aging*, 50:167.e11–167.e13, .

[342] Anamika Giri, Gamze Guven, Hasmet Hanagasi, Ann-Kathrin Hauser, Nihan Erginul-Unaltuna, Basar Bilgic, Hakan Gurvit, Peter Heutink, Thomas Gasser, Ebba Lohmann, and Javier Simón-Sánchez. PLA2g6 mutations related to distinct phenotypes: A new case with early-onset parkinsonism. *Tremor and Other Hyperkinetic Movements*, 6, .

[343] Cornelis Blauwendraat, Margherita Francescatto, J. Raphael Gibbs, Iris E. Jansen, Javier Simón-Sánchez, Dena G. Hernandez, Allissa A. Dillman, Andrew B. Singleton, Mark R. Cookson, Patrizia Rizzu, and Peter Heutink. Comprehensive promoter level expression quantitative trait loci analysis of the human frontal lobe. *Genome Medicine*, 8:65, .

[344] Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42:D980–D985.

[345] Gustavo Glusman, Alissa Severson, Varsha Dhankani, Max Robinson, Terry Farrah, Denise E. Mauldin, Anna B. Stittrich, Seth A. Ament, Jared C. Roach, Mary E. Brunkow, Dale L. Bodian, Joseph G. Vockley, Ilya Shmulevich, John E. Niederhuber, and Leroy Hood. Identification of copy number variants in whole-genome data using reference coverage profiles. *Frontiers in Genetics*, 6:45.

[346] Mahdi Ghani, Anthony E. Lang, Lorne Zinman, Benedetta Nacmias, Sandro Sorbi, Valentina Bessi, Andrea Tedde, Maria Carmela Tartaglia, Ezequiel I. Surace, Christine Sato, Danielle Moreno, Zhengrui Xi, Rachel Hung, Mike A. Nalls, Andrew Singleton, Peter St George-Hyslop, and Ekaterina Rogaeva. Mutation analysis of patients with neurodegenerative disorders using NeuroX array. *Neurobiology of Aging*, 36(1):545.e9–545.e14.

[347] In Joo Kim, Yeon Joo Kim, Byeong Hee Son, Sang Ook Nam, Hoon Chul Kang, Heung Dong Kim, Mi Ae Yoo, Ook Hwan Choi, and Cheol Min Kim. Diagnostic mutational analysis of MECP2 in korean patients with rett syndrome. *Experimental & Molecular Medicine*, 38(2):119–125, .

[348] Val Zvereff, Lori Carpenter, Dagny Patton, Huong Cabral, Debra Rita, Ashley Wilson, Kwame Anyane-Yeboa, Larry White, and Kenneth J. Friedman. Molecular diagnostic dilemmas in rett syndrome. *Brain & Development*, 34(9):750–755.

[349] Fabienne C. Fiesel, Thomas R. Caulfield, Elisabeth L. Moussaud-Lamodière, Kotaro Ogaki, Daniel F.A.R. Dourado, Samuel C. Flores, Owen A. Ross, and Wolfdieter Springer. Structural and functional impact of parkinson disease-associated mutations in the e3 ubiquitin ligase parkin. *Human mutation*, 36(8):774–786.

[350] Lucio Santoro, Guido J. Breedveld, Fiore Manganelli, Rosa Iodice, Chiara Pisciotta, Maria Nolano, Francesca Punzo, Mario Quarantelli, Sabina Pappatà, Alessio Di Fonzo, Ben A. Oostra, and Vincenzo Bonifati. Novel ATP13a2 (PARK9) homozygous mutation in a family with marked phenotype variability. *Neurogenetics*, 12(1):33–39.

[351] Gilad Silberberg, Ariel Darvasi, Ronit Pinkas-Kramarski, and Ruth Navon. The involvement of ErbB4 with schizophrenia: association and expression studies. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, 141B(2):142–148.

[352] Chee Yeun Chung, James B. Koprich, Hasan Siddiqi, and Ole Isacson. Dynamic changes in presynaptic and axonal transport proteins combined with striatal neuroinflammation precede dopaminergic neuronal loss in a rat model of AAV -synucleinopathy. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(11):3365–3373.

[353] Jee-Young Lee, Beom S. Jeon, Han-Joon Kim, and Sung-Sup Park. Genetic variant of HTR2a associates with risk of impulse control and repetitive behaviors in parkinson's disease. *Parkinsonism & Related Disorders*, 18(1):76–78, .

[354] Dongfeng Chen, Shuchao Pang, Xungang Feng, Wenhui Huang, Robert G. Hawley, and Bo Yan. Genetic analysis of the ATG7 gene promoter in sporadic parkinson's disease. *Neuroscience Letters*, 534:193–198, .

[355] S. Lesage, E. Lohmann, F. Tison, F. Durif, A. Dürr, A. Brice, and French Parkinson's Disease Genetics Study Group. Rare heterozygous parkin variants in french early-onset parkinson disease patients and controls. *Journal of Medical Genetics*, 45(1):43–46.

[356] Mark R. Cookson, Paul J. Lockhart, Chris McLendon, Casey O'Farrell, Michael Schlossmacher, and Matthew J. Farrer. RING finger 1 mutations in parkin produce altered localization of the protein. *Human Molecular Genetics*, 12(22):2957–2965.

[357] Cornelia Hampe, Hector Ardila-Osorio, Margot Fournier, Alexis Brice, and Olga Corti. Biochemical analysis of parkinson's disease-causing variants of parkin, an e3 ubiquitin–protein ligase with monoubiquitylation capacity. *Human Molecular Genetics*, 15(13):2059–2075.

[358] Jin-Sung Park, Brianada Koentjoro, Christine Klein, and Carolyn M. Sue. Single heterozygous ATP13a2 mutations cause cellular dysfunction associated with parkinson's disease. *Movement Disorders: Official Journal of the Movement Disorder Society*, .

[359] Teng Xie, Jie Zhang, Xianhou Yuan, Jing Yang, Wei Ding, Xin Huang, and Yong Wu. Is x-linked methyl-CpG binding protein 2 a new target for the treatment of parkinson's disease. *Neural Regeneration Research*, 8(21):1948–1957.

[360] Stephanie C. Gantz, Christopher P. Ford, Kim A. Neve, and John T. Williams. Loss of mecp2 in substantia nigra dopamine neurons compromises the nigrostriatal pathway. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(35): 12629–12637.

[361] Anas M. Alazami, Fatema Alzahrani, Saeed Bohlega, and Fowzan S. Alkuraya. SET binding factor 1 (SBF1) mutation causes charcot-marie-tooth disease type 4b3. *Neurology*, 82 (18):1665–1666.

[362] Khriezhanuo Nakhro, Jin-Mo Park, Young Bin Hong, Ji Hoon Park, Soo Hyun Nam, Bo Ram Yoon, Jeong Hyun Yoo, Heasoo Koo, Sung-Chul Jung, Hyung-Lae Kim, Ji Yon Kim, Kyoung-Gyu Choi, Byung-Ok Choi, and Ki Wha Chung. SET binding factor 1 (SBF1) mutation causes charcot-marie-tooth disease type 4b3. *Neurology*, 81(2):165–173.

[363] Andreea Manole, Alejandro Horga, Josep Gamez, Nuria Raguer, Maria Salvado, Beatriz San Millán, Carmen Navarro, Alan Pittmann, Mary M. Reilly, and Henry Houlden. SBF1 mutations associated with autosomal recessive axonal neuropathy with cranial nerve involvement. *Neurogenetics*, 18(1):63–67.

[364] Cornelis Blauwendraat, Faraz Faghri, Lasse Pihlstrom, Joshua T. Geiger, Alexis Elbaz, Suzanne Lesage, Jean-Christophe Corvol, Patrick May, Aude Nicolas, Yevgeniya Abramzon, Natalie A. Murphy, J. Raphael Gibbs, Mina Ryten, Raffaele Ferrari, Jose Bras, Rita Guerreiro, Julie Williams, Rebecca Sims, Steven Lubbe, Dena G. Hernandez, Kin Y. Mok, Laurie Robak, Roy H. Campbell, Ekaterina Rogaeva, Bryan J. Traynor, Ruth Chia, Sun Ju Chung, John A. Hardy, Alexis Brice, Nicholas W. Wood, Henry Houlden, Joshua M. Shulman, Huw R. Morris, Thomas Gasser, Rejko Krüger, Peter Heutink, Manu Sharma, Javier Simón-Sánchez, Mike A. Nalls, Andrew B. Singleton, and Sonja W. Scholz. NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. *Neurobiology of Aging*, 57:247.e9–247.e13, .

[365] Nuno André Faustino and Thomas A. Cooper. Pre-mRNA splicing and human disease. *Genes & Development*, 17(4):419–437.

[366] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

[367] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, and Nancy Young. The genotype-tissue expression (GTEx) project. *Nature genetics*, 45(6):580.

[368] Gioele La Manno, Daniel Gyllborg, Simone Codeluppi, Kaneyasu Nishimura, Carmen Salto, Amit Zeisel, Lars E. Borm, Simon R. W. Stott, Enrique M. Toledo, J. Carlos Villaescusa,

Peter Lönnerberg, Jesper Ryge, Roger A. Barker, Ernest Arenas, and Sten Linnarsson. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell*, 167 (2):566–580.e19.

LIST OF ABBREVIATIONS

aao	Age-at-onset
AD	Alzheimer's disease
ANOVA	Analysis of variance
ARE	Atypical rolandic epilepsy
AUC	Area under curve
CADD	Combined Annotation Dependent Depletion
CAGEseq	Cap analysis gene expression sequencing
CCDS	Consensus coding sequence
CI	Confidence interval
CNV	Copy number variant
CoGIE	Complex Genetics of Idiopathic generalized epilepsy
CTS	Centro-Temporal Spikes
DALY	Disability-adjusted life years
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic acid
eQTL	Expression quantitative trait locus
EVS	Exome variant server
ExAC	Exome Aggregation Consortium
GATK	Genome Analysis Tool Kit
GGE	Genetic generalized epilepsy
glm	Generalized linear model
gnomAD	Genome Aggregation Database
gVCF	Genomic variant call format

GWAS	Genome wide association study
HGMD	Human gene mutation database
HMM	Hidden Markov model
IBD	Identity by descent
IGE	Idiopathic generalized epilepsy
ILAE	International League Against Epilepsy
INDEL	Insertion/Deletion
IPD	Idiopathic PD
LoF	Loss-of-function
MAF	Minor allele frequency
MDS	Multi-dimensional scaling
MOI	Mode of Inheritance
NCBI	National Center for Biotechnology Information
NFE	Non-Finnish European
NGS	Next generation sequencing
NONSYN	Nonsynonymous variants
PCA	Principal component analysis
PD	Parkinson's disease
PET	Positron emission tomography
PPMI	Parkinson's Progression Markers Initiative
PRS	Polygenic risk score
QC	Quality control
RCPs	Reference Coverage Profiles
RE	Typical rolandic epilepsy
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
SD	Standard deviation
SKAT	Sequence kernel association test
SKAT-O	Optimized sequence kernel association test
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SV	Structural variant
SYN	Synonymous variant
Ti/Tv	Transition/Trasversion ratio
UPDRS	Unified Parkinson's disease rating scale
UPSIT	University of Pennsylvania Smell Identification Test

VCF	Variant call format
VQSR	Variant quality score recalibration
WES	Whole exome sequencing
WGS	Whole genome sequencing

Appendices

APPENDIX A

SUPPLEMENTARY MATERIAL

This chapter contains all manuscripts authored as a first author or co-author along with the supplementary material for each chapter. Journal formatted articles are provided for published manuscripts. Submitted manuscripts or manuscripts that are ready for submission are provided as the submitted versions.

A.1 Rolandic Epilepsy

`https://dropit.uni.lu/invitations?share=44877c4da9f8c3cc51b8&dl=0`

Table A.1: CADD15+LOF variants in the epilepsy associated genes that were identified in the present study. The variants are represented according to the GRCh37 human reference genome.

A.2 Excess of singleton loss-of-function variants in Parkinson's Disease

CHR	BP	A1	A2	OR	P
1	155135036	G	A	0.58	2.59e-35
3	52816840	G	A	0.68	2.25e-7
17	43994648	T	C	0.78	1.26e-68
2	169110394	C	T	0.83	5.68e-26
3	182762437	A	G	0.85	2.11e-30
6	32666660	T	C	0.85	1.26e-13
1	205723572	C	T	0.89	1.12e-2
2	135539967	T	C	0.89	8.24e-24
12	123303586	G	A	0.90	2.05e-20
4	15737101	C	A	0.90	1.22e-19
14	55348869	T	C	0.91	4.30e-16
15	61994134	G	A	0.91	3.94e-14
7	23293746	G	A	0.91	3.51e-18
8	16697091	A	G	0.91	2.38e-11
9	17579690	T	G	0.91	1.99e-12
1	226916078	C	T	0.92	2.40e-10
4	77198986	T	C	0.92	1.43e-14
10	15569598	C	A	0.93	2.37e-8
11	83544472	A	G	0.93	3.72e-9
3	48748989	G	T	0.93	6.80e-8
2	166133632	T	C	0.94	9.73e-7
8	22525980	T	C	1.06	9.06e-7
16	19279464	T	G	1.07	1.46e-9
20	3168166	A	G	1.07	1.99e-6
2	102413116	C	T	1.07	3.83e-8
14	88472612	T	C	1.08	1.20e-9
16	31121793	A	G	1.08	5.44e-12
16	52599188	T	C	1.08	8.29e-8
19	2363319	T	C	1.08	6.64e-7
11	133765367	T	C	1.09	1.11e-13
14	67984370	T	A	1.09	9.61e-11
18	40673380	G	A	1.10	5.56e-16
8	11707174	A	G	1.10	9.54e-11
3	18277488	G	T	1.11	3.02e-9
1	232664611	T	C	1.12	8.41e-13
6	27681215	A	G	1.12	3.44e-13
4	114360372	C	T	1.14	2.11e-9
12	40614434	T	C	1.15	1.21e−19
5	60273923	C	A	1.15	1.69e-11
3	87520857	C	G	1.21	1.22e-4
4	951947	C	T	1.23	1.47e-50
4	90626111	G	A	1.33	5.21e-123
10	121536327	A	G	1.65	2.23e-19

Table A.2: Summary statistics of SNVs used to generate PRS. The statistics were obtained from the study [109]. The variants are represented according to the GRCh37 human reference genome. CHR = chromosome, BP = Position of SNP on the genome, A1 = reference allele, A2 = alternate allele, OR = odds ratio and P = p-value.

A.3 CNVs in epilepsy

`https://dropit.uni.lu/invitations?share=79ef062953955ad9255f&dl=0`

Table A.3: Deletions detected in our study. RE = Rolandic epilepsy (typical and atypical), IGE = Idiopathic generalized epilepsy and Z_score = score generated by XHMM.

Gene	Brain	chr	start	end	gene__ensembl	cds__len	gene__length	del	dup	del.score	dup.score	cnv.score
USP24	+	1	55532032	55680786	ENSG00000162402.8	7740	148754	0	2	2,042565836	1,53282633	1,884072981
TJP1	+	15	29991571	30261068	ENSG00000104067.12	5441	269497	0	12	1,752442713	-0,803995937	-0,081921889
CNTN1	+	12	41086244	41466220	ENSG00000018236.10	3278	379976	0	1	1,728174309	1,425101777	1,802168696
ITPR1	+	3	4535032	4889524	ENSG00000150995.13	8519	354492	0	6	1,725139512	0,253387244	0,753565178
PCDHB6	+	5	140529683	140532868	ENSG00000113211.3	2391	3185	5	2	1,667827598	2,38494211	1,914791919
PCDHB3	+	5	140480234	140483406	ENSG00000113205.2	2397	3172	1	5	1,657469966	0,899130927	0,99912774
SF3B3	+	16	70557691	70608820	ENSG00000189091.8	3804	51129	0	2	1,529650601	0,972523491	1,375089727
TIMELESS	+	12	56810903	56843187	ENSG00000111602.7	3665	32284	0	4	1,522050316	0,446777062	0,919722383
SHANK1	+	19	51165084	51222707	ENSG00000161681.11	6323	57623	0	3	1,140663112	0,46950049	0,789092733
ZNF417	+	19	58411664	58427978	ENSG00000173480.6	1746	16314	0	6	1,088952526	-0,308218506	0,076195049
ATG16L2	+	11	72525353	72554719	ENSG00000168010.6	2355	29366	0	4	1,005027614	0,048613563	0,425402526
GABRB3	+	15	26788693	27184686	ENSG00000166206.9	1772	395993	0	9	0,988016767	-0,97958815	-0,465274823
EPG5	+	18	43427574	43547240	ENSG00000152223.8	8017	119666	1	1	0,987628271	1,386597994	1,417300608
GABRG3	+	15	27216429	27778373	ENSG00000182256.8	1428	561944	0	6	0,852648846	-0,593942874	-0,170777176
CAPN1	+	11	64948037	64979477	ENSG00000014216.11	2230	31440	0	0	0,821376542	1,16710151	1,305530475
ZNF343	+	20	2462463	2505348	ENSG00000088876.7	1824	42885	0	10	0,81624114	-1,361324352	-0,795199446
GABRA5	+	15	27111510	27194354	ENSG00000186297.7	1416	82844	0	6	0,802404149	-0,548812563	-0,157511534
LRRC4C	+	11	40135753	41481323	ENSG00000148948.3	1929	1345570	0	0	0,710652242	0,483380747	0,758829916
NDUFS3	+	11	47586888	47606114	ENSG00000213619.5	1663	19226	1	0	0,695145694	1,400181735	1,327943672
EXD3	+	9	140201348	140317714	ENSG00000187609.11	2707	116366	0	13	0,684156271	-1,725317858	-1,203058829
AGFG2	+	7	100136834	100165842	ENSG00000106351.8	1291	29008	2	3	0,6545631	0,895985923	0,849625612
ST6GALNAC3	+	1	76540404	77100286	ENSG00000184005.9	987	559882	0	1	0,635401501	0,481903843	0,71332483
CGRRF1	+	14	54976530	55005567	ENSG00000100532.7	1039	29037	0	2	0,565795517	0,150343318	0,414022933
CNTNAP2	+	7	145813453	148118090	ENSG00000174469.13	4037	2304637	5	6	0,563947177	0,204322538	0,472669463
APOC2	+	19	45449243	45452822	ENSG00000234906.4	336	3579	0	2	0,552719577	0,182312534	0,423901026
ATG14	+	14	55833110	55878576	ENSG00000126775.8	1539	45466	0	4	0,551818028	-0,357672317	-0,034249132
SAMD4A	+	14	55033815	55260033	ENSG00000020577.9	2366	226218	1	0	0,549832768	1,285731173	1,193637347
SVEP1	+	9	113127531	113342160	ENSG00000165124.13	11061	214629	2	6	0,548008965	0,157194253	0,329345173
ZNF317	+	19	9251056	9274100	ENSG00000130803.10	1758	23044	0	1	0,546852862	0,473517186	0,670057364
APBA2	+	15	29129629	29410518	ENSG00000034053.10	2647	280889	1	10	0,542886965	-1,055651771	-0,625635589
GMFB	+	14	54941202	54955914	ENSG00000197045.8	508	14712	0	0	0,526211632	0,841351244	0,980023489
GCH1	+	14	55308726	55369570	ENSG00000131979.14	892	60844	0	1	0,509578001	0,480372822	0,661531792
NDNL2	+	15	29560353	29562033	ENSG00000185115.4	921	1680	0	0	0,501962417	0,785228916	0,870913319
ATP10A	+	15	25922420	26110317	ENSG00000206190.7	4795	187897	1	5	0,501326386	-0,175277912	0,074763692
CNIH1	+	14	54893654	54908149	ENSG00000100528.7	381	14495	0	0	0,494550985	0,806566767	0,947001865
ERMAP	+	1	43282795	43310660	ENSG00000164010.9	1551	27865	1	8	0,480748455	-0,823379015	-0,425215558
KBTBD4	+	11	47599277	47599823	ENSG00000231880.1	677	546	0	0	0,477900568	0,734800622	0,865962201
SNX16	+	8	82711816	82755101	ENSG00000104497.9	1077	43285	1	2	0,456422625	0,505492858	0,576511245

KERA	+	12	91444268	91451760	ENSG00000139330.5	1071	7492	0	2	0,446874734	-0,000114033	0,250670659
SSTR4	+	20	23016057	23017314	ENSG00000132671.4	1173	1257	0	0	0,405605223	0,623262855	0,729431166
HSPA1L	+	6	31777396	31783437	ENSG00000204390.8	1932	6041	0	0	0,401632891	0,531737845	0,661692239
SOCS4	+	14	55493948	55516206	ENSG00000180008.8	1329	22258	0	1	0,392612172	0,162471626	0,368963914
GHRL	+	3	10327359	10334631	ENSG00000157017.11	483	7272	1	0	0,372479739	1,214689608	1,048091047
GRIN2A	+	16	9852376	10276611	ENSG00000183454.9	4549	424235	1	1	0,370715677	0,729710664	0,721094991
RIMBP2	+	12	130880682	131200826	ENSG00000060709.9	3261	320144	1	7	0,349561229	-0,727894178	-0,409028887
IP6K3	+	6	33689444	33714762	ENSG00000161896.6	1263	25318	0	0	0,335329745	0,724944238	0,801391166
IGSF8	+	1	160061130	160068733	ENSG00000162729.9	1808	7603	0	1	0,287817859	0,301696075	0,423784119
RAPGEFL1	+	17	38333263	38351908	ENSG00000108352.7	1540	18645	2	17	0,264748813	-1,793686773	-1,216594689
MAPK1IP1L	+	14	55518349	55536910	ENSG00000168175.10	922	18561	0	0	0,211003259	0,614292596	0,680394542
FBXO34	+	14	55738021	55828636	ENSG00000178974.5	2142	90615	0	2	0,175325399	-0,364821598	-0,160408258
C1orf54	+	1	150240600	150253327	ENSG00000118292.4	546	12727	3	12	0,124516266	-1,045379554	-0,641860823
CDKN3	+	14	54863567	54886936	ENSG00000100526.15	672	23369	1	0	0,123183899	0,934097692	0,787426186
ZNF318	+	6	43274872	43337216	ENSG00000171467.11	6798	62344	1	0	0,119892425	0,965453183	0,79947501
ATRNL1	+	10	116853124	117708503	ENSG00000107518.12	4386	855379	3	5	0,10741737	0,044128336	0,111085788
CSMD1	+	8	2792875	4852494	ENSG00000183117.13	10919	2059619	7	30	0,106786859	-1,998685059	-1,148652918
SCO1	+	17	10583654	10601692	ENSG00000133028.6	942	18038	1	8	0,084061561	-1,125325984	-0,799948209
LGALS3BP	+	17	76967320	76976191	ENSG00000108679.8	1932	8871	0	0	0,083204782	0,538583426	0,54880306
DLGAP5	+	14	55614830	55658396	ENSG00000126787.8	2649	43566	2	56	0,02444499	-2,531252393	-2,471627899
ZNF691	+	1	43312280	43318148	ENSG00000164011.13	1096	5868	0	0	0,022331951	0,493769336	0,503231742
CLPTM1	+	19	45457842	45496599	ENSG00000104853.11	2500	38757	2	5	-0,002170243	-0,219791442	-0,136935675
ZNF568	+	19	37407231	37489602	ENSG00000198453.8	3501	82371	3	3	-0,015287437	0,752761724	0,404110078
NDN	+	15	23930565	23932450	ENSG00000182636.4	972	1885	0	0	-0,052273904	0,331001753	0,330868612
IMPA1	+	8	82570196	82598928	ENSG00000133731.5	1061	28732	2	2	-0,062286695	0,494818479	0,324510263
WDHD1	+	14	55405668	55493823	ENSG00000198554.7	3540	88155	3	4	-0,075497723	0,242811576	0,174122536
PLXDC2	+	10	20105168	20578785	ENSG00000120594.12	1674	473617	2	1	-0,086459784	0,734142595	0,487533876
TMEM176A	+	7	150497491	150502208	ENSG00000002933.3	751	4717	2	0	-0,169683976	1,154291398	0,700844273
FCHSD2	+	11	72547790	72853306	ENSG00000137478.10	2271	305516	2	4	-0,198729335	-0,160404265	-0,177850106
MRPS27	+	5	71515236	71616473	ENSG00000113048.12	1642	101237	2	1	-0,277487506	0,63115712	0,340595652
LGALS3	+	14	55590828	55612126	ENSG00000131981.11	869	21298	2	1	-0,386319462	0,528835801	0,228501941
ADPRM	+	17	10600911	10614550	ENSG00000170222.11	1059	13639	2	11	-0,406990988	-1,67909556	-1,353159743
F5	+	1	169483404	169555826	ENSG00000198734.6	6838	72422	4	1	-0,449909425	1,137009467	0,5336849
SLC24A4	+	14	92788925	92962596	ENSG00000140090.13	1990	173671	3	7	-0,48267106	-0,681136634	-0,630949501
SUMF1	+	3	3742498	4508965	ENSG00000144455.9	1179	766467	4	9	-0,513018336	-0,908525108	-0,764074106
KCNQ1	+	11	2465914	2870339	ENSG00000053918.11	2168	404425	3	3	-0,518480243	0,154714284	-0,106873975
ANKRD16	+	10	5903580	5931869	ENSG00000134461.11	1128	28289	2	3	-0,532263044	-0,149995699	-0,304101007
CDH8	+	16	61681146	62070939	ENSG00000150394.9	2528	389793	3	2	-0,573570084	0,328517197	-0,008246912
STARD10	+	11	72465774	72504726	ENSG00000214530.3	853	38952	2	2	-0,614402312	0,070403966	-0,194948524

ETV1	+	7	13930853	14031050	ENSG00000006468.9	1624	100197	3	6	-0,672793719	-0,645347304	-0,682951535
ZFAND1	+	8	82613569	82645138	ENSG00000104231.6	1037	31569	3	3	-0,713229432	-0,029152124	-0,288960927
MT1F	+	16	56691606	56694610	ENSG00000198417.5	245	3004	3	75	-0,732291441	-2,531252393	-2,471627899
COL10A1	+	6	116440086	116479910	ENSG00000123500.5	2055	39824	3	1	-0,75927544	0,440689832	-0,066585525
TECR	+	19	14627897	14676792	ENSG00000099797.7	1328	48895	5	14	-0,762158576	-1,361274273	-1,144196973
TMEM220	+	17	10602332	10633633	ENSG00000187824.4	441	31301	3	13	-0,850692716	-1,910378895	-1,640173244
SGCG	+	13	23755091	23899304	ENSG00000102683.6	918	144213	7	23	-0,867575065	-1,730521173	-1,415064654
NT5DC1	+	6	116422012	116570660	ENSG00000178425.9	1605	148648	4	3	-1,027729486	0,024943674	-0,400335605
USH1C	+	11	17515442	17565963	ENSG00000006611.11	2942	50521	6	1	-1,083452944	1,162495114	0,188844589
SACS	+	13	23902965	24007841	ENSG00000151835.9	2135	104876	8	23	-1,203293261	-1,815881184	-1,563224422
ZNF790	+	19	37308330	37341689	ENSG00000197863.4	1935	33359	8	3	-1,236542756	0,955106418	-0,035262877
NCAPD2	+	12	6602522	6641121	ENSG00000010292.8	4392	38599	8	4	-1,341991076	0,505722435	-0,21514961
RIOK2	+	5	96496571	96518964	ENSG00000058729.6	1978	22393	6	18	-1,509592757	-2,256647924	-1,976393386
PTPRZ1	+	7	121513143	121702090	ENSG00000106278.7	7128	188947	8	2	-1,568023817	0,841591482	-0,198054466
CDH22	+	20	44802372	44937137	ENSG00000149654.5	2555	134765	5	0	-1,578268618	0,88145143	-0,26589405
OCA2	+	15	28000021	28344504	ENSG00000104044.11	2714	344483	9	9	-1,660380988	-0,543931674	-0,906499272
LRRK2	+	12	40590546	40763087	ENSG00000188906.9	8008	172541	11	19	-1,750477101	-1,477247598	-1,376714468
MT1A	+	16	56672578	56673999	ENSG00000205362.6	204	1421	7	73	-1,761405278	-2,531252393	-2,471627899
SCN1A	+	2	166845670	166984523	ENSG00000144285.11	6186	138853	9	5	-1,810908478	0,043445991	-0,670522876
ZNF692	+	1	249144205	249153343	ENSG00000171163.11	1540	9138	6	24	-1,868951238	-2,531252393	-2,471627899
RECQL5	+	17	73622925	73663269	ENSG00000108469.10	3188	40344	9	4	-1,875607855	0,249072756	-0,624359689
TNFRSF19	+	13	24144509	24250232	ENSG00000127863.11	1335	105723	12	26	-2,042544874	-1,987046267	-1,865746884
MTMR10	+	15	31231144	31283810	ENSG00000166912.12	2396	52666	14	28	-2,293892804	-2,030503686	-1,928391813
MT1E	+	16	56659387	56661024	ENSG00000169715.10	492	1637	10	85	-2,454760427	-2,531252393	-2,471627899
CCDC86	+	11	60609544	60618554	ENSG00000110104.7	1107	9010	11	2	-2,623544152	0,070262779	-1,498783415
ZNF276	+	16	89786808	89807311	ENSG00000158805.7	2703	20503	11	15	-2,623544152	-1,847806105	-2,06366487
NPR2	+	9	35792151	35809729	ENSG00000159899.10	3356	17578	19	13	-2,623544152	-1,04407072	-1,716095807
TOP1MT	+	8	144386554	144442149	ENSG00000184428.8	2288	55595	22	8	-2,623544152	-0,919375543	-2,360916245
SLC3A1	+	2	44502599	44548633	ENSG00000138079.9	2365	46034	23	45	-2,623544152	-2,531252393	-2,471627899
MT1M	+	16	56666145	56667898	ENSG00000205364.3	315	1753	23	78	-2,623544152	-2,531252393	-2,471627899
IQGAP2	+	5	75699074	76003957	ENSG00000145703.11	5978	304883	32	13	-2,623544152	-1,141733115	-2,435588281
ABCA7	+	19	1040102	1065571	ENSG00000064687.8	6865	25469	33	12	-2,623544152	-0,850864983	-2,33107997
PCDHB4	+	5	140501581	140505201	ENSG00000081818.1	2394	3620	40	3	-2,623544152	2,201582277	-0,988534575
IQCC	+	1	32671236	32674288	ENSG00000160051.7	1669	3052	41	1	-2,623544152	0,583195479	-2,471627899
PCDHB5	+	5	140514800	140517703	ENSG00000113209.6	2394	2903	41	4	-2,623544152	2,02513866	-0,646985259
FANCA	+	16	89803957	89883065	ENSG00000187741.10	5500	79108	67	32	-2,623544152	-2,531252393	-2,471627899
ZNF517	+	8	146024261	146036554	ENSG00000197363.5	1529	12293	81	4	-2,623544152	-0,583788753	-2,471627899
OR5A1	-	11	59210617	59211667	ENSG00000172320.2	954	1050	0	3	2,00330315	0,889007708	1,398695608
MRPS17	-	7	55954970	56022932	ENSG00000249773.3	744	67962	0	3	1,863136567	1,131022766	1,4653264

IFNA6	-	9	21349834	21351377	ENSG00000120235.3	576	1543	1	4	1,588574969	0,766991953	1,142146273
KRT222	-	17	38785049	38821393	ENSG00000264058.1	2051	36344	0	5	1,373846734	0,016502355	0,537578278
IFNA8	-	9	21409146	21410184	ENSG00000120242.3	576	1038	1	1	1,259921833	1,402169953	1,50898953
IFNA2	-	9	21384254	21385396	ENSG00000188379.5	573	1142	2	3	1,131156808	1,057697756	1,164000809
KRT1	-	12	53068520	53074191	ENSG00000167768.4	961	5671	0	1	1,079983152	0,943547389	1,203262807
NPR1	-	1	153651113	153666468	ENSG00000169418.9	3418	15355	0	17	1,025095634	-1,936764709	-1,247399084
DCAF8	-	1	160185505	160254920	ENSG00000132716.14	5519	69415	0	1	0,957237464	0,852765373	1,108323503
MC3R	-	20	54823788	54824871	ENSG00000124089.4	1089	1083	0	0	0,902668364	1,011532036	1,208983609
ZMYM6	-	1	35447134	35497342	ENSG00000271741.1	2449	50208	0	2	0,869513173	0,40436517	0,718198843
OR51A7	-	11	4928600	4929538	ENSG00000176895.8	945	938	0	4	0,804757994	-0,335079833	0,064676549
NDUFB8	-	10	102265385	102289638	ENSG00000255339.2	1649	24253	0	3	0,754240813	0,086982104	0,392793167
PSMA2	-	7	42948872	42971773	ENSG00000256646.3	2050	22901	0	3	0,712238121	-0,008386022	0,321602379
APOC4	-	19	45445495	45452820	ENSG00000267467.2	738	7325	0	4	0,704400162	-0,22857705	0,115620434
APOC4-APOC2	-	19	45445495	45452822	ENSG00000224916.4	738	7327	0	4	0,704400145	-0,228577243	0,11562026
TBPL2	-	14	55880259	55923444	ENSG00000182521.5	1170	43185	0	2	0,701161562	0,274519725	0,555543593
RNASE11	-	14	21051051	21077954	ENSG00000259060.2	1076	26903	0	2	0,698384565	0,192699302	0,488118713
ITGB3	-	17	45331212	45421658	ENSG00000259207.3	2815	90446	1	7	0,627821824	-0,513338287	-0,133480891
GPR179	-	17	36481413	36499730	ENSG00000188888.7	7170	18317	0	16	0,603751417	-2,140538837	-1,575368664
ZNF829	-	19	37379026	37407193	ENSG00000185869.9	1417	28167	2	1	0,591881786	1,521491092	1,195510211
NDUFA7	-	19	8373167	8386263	ENSG00000167774.2	514	13096	0	9	0,585504558	-1,2918201	-0,829600132
CCDC15	-	11	124824017	124911385	ENSG00000149548.10	2827	87368	0	1	0,582799287	0,530860254	0,723666949
ZNF223	-	19	44529506	44591471	ENSG00000267022.1	3615	61965	1	9	0,547366961	-0,957878289	-0,506746903
SDHD	-	11	111957497	111990353	ENSG00000204370.4	504	32856	1	1	0,508725421	0,923249361	0,892855517
SNURF	-	15	25200133	25223729	ENSG00000273173.1	999	23596	1	9	0,494625639	-0,958452948	-0,535054016
TIMM10B	-	11	6502677	6505909	ENSG00000132286.7	285	3232	0	2	0,475098274	0,128493226	0,34415333
FPR3	-	19	52298416	52329442	ENSG00000187474.4	1068	31026	0	79	0,470221145	-2,531252393	-2,471627899
UQCR11	-	19	1578338	1605444	ENSG00000267059.2	979	27106	0	3	0,469648729	-0,140174772	0,102481605
TLR9	-	3	52255096	52273183	ENSG00000239732.2	4095	18087	0	0	0,467351633	0,842277503	0,940676644
PSTPIP2	-	18	43563502	43652238	ENSG00000152229.14	1084	88736	1	4	0,461399435	-0,010655736	0,200411894
MDGA2	-	14	47308826	48144157	ENSG00000139915.14	3166	835331	1	3	0,454860699	0,114755413	0,295758017
FPR2	-	19	52255279	52273779	ENSG00000171049.8	1062	18500	0	44	0,431212051	-2,531252393	-2,471627899
PSMA1	-	11	14515329	14541890	ENSG00000256206.2	1333	26561	1	4	0,39544599	-0,090300325	0,123998029
TMEM239	-	20	2795614	2798712	ENSG00000241690.3	811	3098	0	8	0,339550937	-1,36434384	-0,965477352
F11R	-	1	160965001	160991138	ENSG00000158769.13	960	26137	1	4	0,300400902	-0,128137195	0,049469062
DDX60L	-	4	169277886	169458937	ENSG00000181381.9	5232	181051	2	3	0,294031669	0,513298286	0,489947409
FFAR2	-	19	35934809	35942669	ENSG00000126262.4	999	7860	0	0	0,284150195	0,561513087	0,647392864
C6ORF165	-	6	88117701	88174183	ENSG00000272514.1	1965	56482	1	2	0,2777637	0,349712688	0,415185071
PI4K2A	-	10	99344131	99433667	ENSG00000249967.1	1410	89536	1	1	0,271314355	0,711172925	0,66933108

HOXC4	-	12	54410715	54449813	ENSG00000198353.6	807	39098	0	0	0,262721096	0,632613766	0,703834732
TM9SF1	-	14	24658349	24682679	ENSG00000100926.10	2681	24330	2	3	0,245780253	0,420389432	0,414767658
TAP2	-	6	32781544	32806599	ENSG00000250264.1	2779	25055	1	9	0,221428691	-1,148708643	-0,784288693
OR5A2	-	11	59189416	59190449	ENSG00000172324.4	981	1033	0	3	0,20882175	-0,491335651	-0,252152641
FOXA2	-	20	22561643	22566093	ENSG00000125798.10	1402	4450	0	0	0,164701281	0,522494276	0,570225949
MFRP	-	11	119209652	119217368	ENSG00000259159.1	2342	7716	1	13	0,155107232	-1,759853362	-1,303418727
SOGA3	-	6	127759551	127840146	ENSG00000255330.4	4995	80595	1	0	0,115177061	0,934397255	0,779176604
ZNF668	-	16	31072164	31085641	ENSG00000167394.8	1942	13477	0	4	0,110829178	-0,750576	-0,525767985
NPAP1	-	15	24920541	24928593	ENSG00000185823.2	3477	8052	0	0	0,046415865	0,167295388	0,241857986
CELF6	-	15	72559087	72612287	ENSG00000273025.1	985	53200	1	0	0,03683415	0,937300396	0,732061151
MRPL30	-	2	99771461	99811761	ENSG00000273155.1	469	40300	1	97	0,028964383	-2,531252393	-2,471627899
IER3IP1	-	18	44661027	44702652	ENSG00000267228.2	633	41625	1	2	0,025075437	0,166352983	0,180038148
SLC5A3	-	21	35445870	35478559	ENSG00000198743.5	2214	32689	0	2	-0,008765236	-0,363896839	-0,220639291
ZNF747	-	16	30537244	30546173	ENSG00000261459.1	2688	8929	0	0	-0,039583786	0,365128257	0,360116263
AQP1	-	7	30893010	30963427	ENSG00000250424.3	1720	70417	3	2	-0,067720053	0,829709529	0,519393278
ALG9	-	11	111652919	111742305	ENSG00000086848.10	1799	89386	2	2	-0,102760105	0,439840777	0,292094762
ZFP41	-	8	144328991	144344875	ENSG00000181638.13	603	15884	1	0	-0,109732969	0,717745364	0,510519046
MT1B	-	16	56685811	56687116	ENSG00000169688.10	231	1305	2	74	-0,110008246	-2,531252393	-2,471627899
CRIP1	-	14	105952654	105955284	ENSG00000213145.5	212	2630	1	2	-0,136071241	0,091043029	0,045688945
SYCP2	-	20	58438618	58508710	ENSG00000196074.8	4933	70092	4	9	-0,141331208	-0,473549305	-0,316503603
CHMP4C	-	8	82644669	82671750	ENSG00000164695.4	732	27081	1	3	-0,1488854	-0,230913621	-0,192377797
DENND2C	-	1	115125469	115213043	ENSG00000175984.10	2721	87574	3	1	-0,166017758	1,016912902	0,613251714
EFNA3	-	1	155036224	155059283	ENSG00000251246.1	732	23059	1	0	-0,284459084	0,686259544	0,412188207
SLC10A5	-	8	82605842	82608409	ENSG00000253598.1	1323	2567	2	2	-0,291947566	0,104626332	-0,066575942
ZNF709	-	19	12571998	12624668	ENSG00000242852.2	1950	52670	2	2	-0,347885862	0,165404134	-0,004170961
IFNA13	-	9	21367371	21368075	ENSG00000233816.2	577	704	1	4	-0,375814849	-0,724692337	-0,663193596
LTB4R2	-	14	24774940	24781259	ENSG00000213906.5	1315	6319	1	0	-0,384529414	0,51856652	0,242996017
SOX7	-	8	10581278	10697357	ENSG00000171056.6	1618	116079	1	9	-0,385358363	-1,61821428	-1,380663381
LSP1	-	11	1874200	1913497	ENSG00000130592.9	938	39297	2	3	-0,431311613	-0,049404415	-0,199708948
ZNF8	-	19	58790317	58807254	ENSG00000083842.8	1680	16937	2	3	-0,499181617	-0,199688553	-0,334719242
ASGR2	-	17	7004641	7019019	ENSG00000161944.12	906	14378	3	4	-0,606254293	-0,156681805	-0,333131143
ACE	-	17	61554422	61599205	ENSG00000159640.10	4273	44783	4	2	-0,630531362	0,711424805	0,190219613
CCDC178	-	18	30517366	31021065	ENSG00000166960.12	2730	503699	4	1	-0,639756324	0,86839966	0,260166627
SLC25A10	-	17	79670401	79687569	ENSG00000262660.1	1404	17168	4	5	-0,692123393	-0,152211722	-0,351092297
C2ORF15	-	2	99757948	99767950	ENSG00000273045.1	390	10002	3	86	-0,751355563	-2,531252393	-2,471627899
LIPN	-	10	90521163	90537999	ENSG00000204020.5	1224	16836	3	3	-0,80807055	-0,090253703	-0,374849419
PIK3R2	-	19	18263928	18281350	ENSG00000105647.10	1678	17422	3	0	-0,818207213	0,966641614	0,234346542
MYH3	-	17	10531843	10560626	ENSG00000109063.10	6057	28783	8	9	-0,83613782	-0,059029141	-0,227385648
CFB	-	6	31895475	31919825	ENSG00000244255.1	4797	24350	8	3	-1,21655474	0,860292159	0,019899919

OC90	-	8	133036467	133071627	ENSG00000253117.4	1424	35160	4	4	-1,22297483	-0,318979538	-0,70733283
UGT2A1	-	4	70454135	70518965	ENSG00000173610.7	1620	64830	5	8	-1,452397597	-1,157114238	-1,290499802
CDRT1	-	17	15468797	15469590	ENSG00000181464.2	735	793	5	13	-1,583552773	-1,987581379	-1,933008812
ZNF763	-	19	12035890	12090105	ENSG00000267179.1	3340	54215	7	13	-1,672130671	-1,613127348	-1,620374567
PCDH15	-	10	55562531	57387702	ENSG00000150275.13	7759	1825171	15	5	-1,70907171	0,649013663	-0,270961846
TMBIM4	-	12	66517697	66563765	ENSG00000228144.2	1009	46068	6	7	-1,713035924	-0,935854517	-1,24193792
NAA60	-	16	3415099	3536960	ENSG00000262621.2	1755	121861	5	3	-1,7214776	-0,241712945	-0,899431715
KLK9	-	19	51499274	51512837	ENSG00000269741.1	1530	13563	7	2	-1,737563801	0,543418516	-0,486487867
TM4SF19	-	3	196042953	196065244	ENSG00000273331.1	1067	22291	6	8	-1,74542378	-1,052405841	-1,353431098
ITFG3	-	16	284545	319942	ENSG00000167930.11	2047	35397	8	17	-2,119119744	-2,137781451	-2,083388631
MYH1	-	17	10395624	10421860	ENSG00000109061.9	6048	26236	14	66	-2,145800473	-2,531252393	-2,471627899
KIAA0391	-	14	35591052	35743271	ENSG00000100890.11	1858	152219	8	17	-2,363937756	-2,388528159	-2,349853415
PGBD2	-	1	249200395	249214145	ENSG00000185220.7	1791	13750	10	44	-2,372034031	-2,531252393	-2,471627899
F2RL2	-	5	75911307	75919259	ENSG00000164220.6	1137	7952	9	1	-2,394665006	0,582109157	-0,860471779
MS4A10	-	11	60552821	60568778	ENSG00000172689.1	758	15957	11	2	-2,623544152	0,341150501	-1,151351654
MS4A15	-	11	60524426	60544205	ENSG00000166961.10	627	19779	11	2	-2,623544152	0,386350104	-1,100406599
FAM170A	-	5	118965254	118971517	ENSG00000164334.11	1064	6263	12	2	-2,623544152	0,470445528	-1,080167359
OR51T1	-	11	4903049	4904113	ENSG00000176900.2	1071	1064	12	3	-2,623544152	-0,312879774	-1,741515134
GALT	-	9	34638130	34651032	ENSG00000213930.7	1279	12902	18	6	-2,623544152	-0,256425303	-1,622835199
PAGR1	-	16	29827285	29841948	ENSG00000185928.7	914	14663	19	28	-2,623544152	-2,218653397	-2,356307677
MYH2	-	17	10424465	10453274	ENSG00000125414.14	6112	28809	20	20	-2,623544152	-1,479661087	-1,744933332
LCN6	-	9	139632619	139642905	ENSG00000204003.4	1113	10286	21	7	-2,623544152	-1,050552283	-2,471627899
MUSK	-	9	113431051	113563859	ENSG00000030304.8	2867	132808	24	10	-2,623544152	-0,962658632	-2,194517064
LRRC37A3	-	17	62850430	62915598	ENSG00000176809.6	4656	65168	29	32	-2,623544152	-2,531252393	-2,471627899
SHPK	-	17	3468738	3539543	ENSG00000262304.1	4111	70805	67	14	-2,623544152	-1,803423642	-2,471627899
TRPM1	-	15	31293264	31453476	ENSG00000134160.9	5046	160212	68	32	-2,623544152	-1,725519217	-2,471627899
CAPN11	-	6	44126548	44152139	ENSG00000137225.8	2274	25591	88	106	-2,623544152	-2,531252393	-2,471627899
KRT77	-	12	53083410	53097247	ENSG00000189182.5	1510	13837	165	3	-2,623544152	0,387708152	-2,471627899
AOC1	-	7	150521715	150558592	ENSG00000002726.15	2336	36877	251	1	-2,623544152	0,502388617	-2,471627899

Table A.4: Deletions common to ExAC CNVs. Data is sorted from low to high deletion score (del.score) and duplication (dup) frequencies. ”+” indicates expression in the brain. Deletion score increases with increasing intolerance.

Chr	Start	End	Sample	Length	Genes	Regions to filter
20	54823759	54824900	SN7640114_5535_ROL_0391_1	1141	MC3R	No
1	53320120	53329849	SN7640113_5560_ROL_0501_1	9729	ZYG11A	Yes
6	33693196	33703280	SN7640116_5738_ROL_0691_1	10084	IP6K3	No
5	96506883	96518935	SN7640099_3853_D202_1	12052	RIOK2	Yes
5	71519462	71533975	SN10600087_4671_E145b_1	14513	MRPS27	No
19	45447959	45465365	SN10410083_4605_EPW_10381_1	17406	APOC2, APOC4, APOC4-APOC2, CLPTM1	No
1	43296070	43317484	SN7640116_5714_ROL_0591_1	21414	ERMAP, ZNF691	No
14	77302503	77327178	SN7640113_5314_S97_1	24675	LRRC74A	No
4	169362457	169393930	SN7640116_5718_ROL_0631_1	31473	DDX60L	No
1	115137047	115168530	SN7640114_5320_E130f_1	31483	DENND2C	No
10	49383834	49420140	SN10410083_4604_EPW_10371_1	36306	FRMPD2	No
2	44502637	44539912	SN7640110_4834_EPW_10651_1	37275	SLC3A1	No
17	73623470	73661285	SN10410083_4602_EPW_10311_1	37815	RECQL5, SMIM5, SMIM6	Yes
5	140482462	140531165	SN7640097_3840_E650_1	48703	PCDHB3, PCDHB4, PCDHB5, PCDHB6	Yes
18	30873076	30928981	SN7640110_4822_EPW_10561_1	55905	CCDC178	Yes
5	75858199	75914495	SN7640110_4823_EPW_10571_1	56296	F2RL2, IQGAP2	No
3	4403776	4562816	SN7640113_5548_ROL_0451_1	159040	ITPR1, ITPR1-AS1, SUMF1	No
16	9856958	10032248	SN7640112_5141_EPW_11111_1	175290	GRIN2A	No
8	82571539	82752251	SN10410083_4581_EPW_10181_1	180712	CHMP4C, IMPA1, SLC10A5, SNX16, ZFAND1	No
17	10403892	10632442	SN7640097_3843_E91_1	228550	ADPRM, MAGOH2P, MYH1, MYH2, MYH3, MYHAS, SCO1, TMEM220	No
14	54863694	55907289	SN7640113_5558_ROL_0481_1	1043595	ATG14, CDKN3, CGRRF1, CNIH1, DLGAP5, FBXO34, GCH1, GMFB, LGALS3, MAPK1IP1L, MIR4308, SAMD4A, SOCS4, TBPL2, WDHD1	No
15	29346087	32460550	SN7640113_5549_ROL_0461_1	3114463	APBA2, ARHGAP11B, CHRFAM7A, CHRNA7, DKFZP434L187, FAM189A1, FAN1, GOLGA8H, GOLGA8J, GOLGA8R, GOLGA8T, HERC2P10, KLF13, LOC100288637, LOC283710, MIR211, MTMR10, NDNL2, OTUD7A, TJP1, TRPM1, ULK4P1, ULK4P2, ULK4P3	No
15	23811123	28525396	SN7640112_5240_EPW_11321_1	4714273	ATP10A, GABRA5, GABRB3, GABRG3, GABRG3-AS1, HERC2, IPW, LINC00929, LOC100128714, MAGEL2, MIR4715, MKRN3, NDN, NPAP1, OCA2, PWAR1, PWAR4, PWAR5, PWARSN, PWRN1, PWRN2, PWRN3, PWRN4, SNORD107, SNORD108, SNORD109A, SNORD109B, SNORD115-1, SNORD115-10, SNORD115-11, SNORD115-12, SNORD115-13	No

Table A.5: Deletions detected via WES data and validated by array data.

A.4 Familial-PD

`https://dropit.uni.lu/invitations?share=6dc8ba728b76345dd71c&dl=0`

Table A.6: Top 15 genes containing coding, non-coding and CNVs per family. When a variant is not present in a gene it is represented as ”NA”. coding_dom_gene = coding variants following autosomal dominant inheritance, coding_dom_score = phenolyzer score of coding variants following autosomal dominant inheritance, coding_dom_cand_gene = Whether coding variants following autosomal dominant inheritance are present in the candidate gene list or not, coding_dom_gene = coding variants following autosomal recessive inheritance, coding_rec_gene coding_rec_score = phenolyzer score of coding variants following autosomal recessive inheritance, coding_rec_cand_gene = Whether coding variants following autosomal recessive inheritance is present in the candidate gene list or not, noncoding_dom_score = noncoding variants following autosomal dominant inheritance, noncoding_dom_score = phenolyzer score of noncoding variants following autosomal dominant inheritance, noncoding_dom_cand_gene = Whether noncoding variants following autosomal dominant inheritance are present in the candidate gene list or not, noncoding_rec_gene = noncoding variants following autosomal recessive inheritance, noncoding_rec_score = phenolyzer score of noncoding variants following autosomal recessive inheritance, noncoding_rec_cand_gene = Whether noncoding variants following autosomal recessive inheritance are present in the candidate gene list or not, cnv_ del_gene = Genes spanning a deletion, cnv_ del_score = Phenolyzer score of genes spanning a deletion, cnv_ del_cand_gene = Whether genes spanning a deletion are present in the candidate gene list or not, cnv_ dup_gene = Genes spanning a duplication, cnv_ dup_score = Phenolyzer score of genes spanning a duplication, cnv_ dup_cand_gene = Whether genes spanning a duplication are present in the candidate gene list or not

Family	**MOI**
102	AD
104	AD
14	AD
164	AD
251	AD
252	AD
253	AD
259	AD
292	AD
3065	AD
3070	AD
315	AD
326	AD
332	AD
338	AD
3401	AD
HCB1	AD
HCB2	AD
HCB4	AD
PD290	AD
PD291	AD
PD317	AD
PD320	AD
307234	AR
3086	AR

3886	AR
Fam_034	AR
Fam_158	AR
Fam_175	AR
Fam_176	AR
HCB5	AR
PD172	AR
PD257	AR
PD296	AR
PD300	AR
PD313	AR

Table A.7: Name of the family and the mode of inheritance that was tested.

Family	MOI	Chr	Pos	Ref	Alt	Function	Gene	AA.change	HGMD disease
3065	AD	7	107350620	G	C	ex	SLC26A4	p.E737D	Hearing loss
3065	AD	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2
3065	AD	X	47315839	T	G	spl	ZNF41	.	Mental retardation
3070	AD	10	55955444	T	G	ex_spl	PCDH15	p.D398A	Usher syndrome 1
3070	AR	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
3070	AR	6	6320808	T	G	int	F13A1	.	Factor XIII deficiency
3086	AD	7	117307052	G	A	ex	CFTR	p.D1445N	Cystic fibrosis
3086	AD	8	100832259	A	G	ex	VPS13B	p.N2993S	Cohen syndrome
3086	AD	9	111911955	A	AT	ex	FRRS1L	p.I146fs	Encephalopathy%2C epileptic-dyskinetic
3086	AD	21	44480591	G	A	ex	CBS	p.R264C	Homocystinuria
3401	AD	1	76211574	C	A	ex	ACADM	p.T39N	Medium chain acyl CoA dehydrogenase deficiency
3401	AD	1	94466625	G	A	ex	ABCA4	p.R2107C	Stargardt disease
3401	AD	3	37089131	A	C	ex	MLH1	p.K377T	Colorectal cancer%2C non-polyposis
3401	AD	5	74981103	C	T	ex	POC5	p.A421T	Scoliosis
3401	AD	8	55542540	G	A	ex	RP1	p.C2033Y	Retinitis pigmentosa
3401	AD	9	136494594	G	T	intgen	FAM163B_DBH	.	Altered enzyme activity
3401	AD	12	114837349	C	A	ex	TBX5	p.D61Y	Holt-Oram syndrome
3401	AD	16	56906568	C	T	ex_spl	SLC12A3	p.A322V	Gitelman syndrome
3401	AD	16	89986130	T	C	ex	MC1R	p.I155T	Red hair%2C increased risk
3401	AD	20	33763985	C	T	ex	PROCR	p.R113C	Venous thromboembolism
3401	AD	X	153171698	GCGCCGCAGGGGA	G	ex	AVPR2	p.247_250del	Diabetes insipidus%2C nephrogenic
3886	AD	5	135391462	A	G	ex	TGFBI	p.M502V	Corneal dystrophy
3886	AR	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
GRIP_102	AD	1	22202483	G	A	ex	HSPG2	p.P1020L	Schwartz-Jampel syndrome type 1
GRIP_102	AD	1	150530505	T	TG	ex	ADAMTSL4	p.F754fs	Ectopia lentis%2C isolated form
GRIP_102	AD	1	196620941	C	T	up	CFH	.	Haemolytic uraemic syndrome%2C atypical
GRIP_102	AD	2	157369961	C	T	ex	GPD2	p.P205L	Intellectual disability
GRIP_102	AD	6	26091179	C	G	ex	HFE	p.H63D	Haemochromatosis%2C association with
GRIP_102	AD	15	28230247	C	T	ex	OCA2	p.V419I	Albinism%2C oculocutaneous II
GRIP_102	AD	17	7576841	A	G	int	TP53	.	Breast and:or ovarian cancer
GRIP_102	AR	14	95581899	G	A	int	DICER1	.	Breast cancer
GRIP_102	AR	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2
GRIP_104	AD	3	15686693	G	C	ex	BTD	p.D444H	Biotinidase deficiency%2C partial
GRIP_104	AD	3	48627789	C	A	ex_spl	COL7A1	p.G636V	Epidermolysis bullosa%2C recessive dystrophic
GRIP_104	AD	5	177638965	T	A	ex	PHYKPL	p.E396V	Phosphohydroxylysinuria
GRIP_104	AD	8	61693942	G	GAAAGCA	ex	CHD7	p.K683delinsKKA	Kallmann syndrome
GRIP_104	AD	9	99064254	G	A	ex	HSD17B3	p.R45W	Hypospadias
GRIP_104	AD	11	108124486	T	G	int	ATM	.	Breast cancer
GRIP_104	AD	13	52542680	A	G	ex	ATP7B	p.V536A	Wilson disease
GRIP_104	AD	14	94847262	T	A	ex	SERPINA1	p.E288V	Antitrypsin alpha 1 deficiency%2C partial
GRIP_104	AD	16	88904097	A	C	ex	GALNS	p.F167V	Mucopolysaccharidosis IVa
GRIP_104	AR	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
GRIP_104	AR	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
GRIP_104	AR	16	56917953	T	C	int	SLC12A3	.	Gitelman syndrome%2C without hypomagnesaemia
GRIP_104	AR	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2

GRIP_14	AD	1	53662764	C	A	ex_spl	CPT2	p.P50H	Carnitine palmitoyltransferase 2 deficiency
GRIP_14	AD	11	46747447	G	A	ex	F2	p.E200K	Dysprothrombinaemia
GRIP_14	AD	11	57365748	C	T	ex	SERPING1	p.A2V	Angioneurotic oedema
GRIP_14	AD	12	6234884	G	A	up	VWF	.	Von Willebrand disease 1
GRIP_14	AD	16	14041848	C	T	ex	ERCC4	p.R799W	Xeroderma pigmentosum (F)
GRIP_14	AR	8	55542540	G	A	ex	RP1	p.C2033Y	Retinitis pigmentosa
GRIP_164	AD	16	56917953	T	C	int	SLC12A3	.	Gitelman syndrome%2C without hypomagnesaemia
GRIP_164	AD	17	7128292	G	A	ex	ACADVL	p.R593Q	Very long chain acyl-CoA dehydrogenase deficiency
GRIP_164	AD	19	2435150	C	T	ex	LMNB2	p.R235Q	Lipodystrophy%2C partial%2C acquired%2C susceptibility to
GRIP_251	AD	1	16091760	C	T	int	FBLIM1	.	Chronic multifocal osteomyelitis
GRIP_251	AD	5	137089865	C	G	UTR5	HNRNPA0	.	Cancer%2C increased risk
GRIP_251	AD	6	26091179	C	G	ex	HFE	p.H63D	Haemochromatosis%2C association with
GRIP_251	AD	10	43598056	G	A	ex	RET	p.V202M	Hirschsprung disease
GRIP_251	AD	13	39453010	G	A	ex	FREM2	p.V2968I	Congenital high airways obstruction syndrome
GRIP_251	AD	16	56917953	T	C	int	SLC12A3	.	Gitelman syndrome%2C without hypomagnesaemia
GRIP_251	AR	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2
GRIP_252	AD	1	3329229	G	C	ex	PRDM16	p.R823P	Sudden unexpected death in infancy
GRIP_252	AD	1	16091760	C	T	int	FBLIM1	.	Chronic multifocal osteomyelitis
GRIP_252	AD	2	71896835	G	A	ex	DYSF	p.D1862N	Muscular dystrophy%2C limb girdle : Miyoshi myopathy
GRIP_252	AD	3	39226442	C	T	ex	XIRP1	p.E182K	Primary microcephaly
GRIP_252	AD	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
GRIP_252	AD	7	117199644	ATCT	A	ex	CFTR	p.507_508del	Cystic fibrosis
GRIP_252	AD	9	12694274	G	A	ex	TYRP1	p.R93H	Albinism%2C oculocutaneous 3
GRIP_252	AD	10	55955444	T	G	ex_spl	PCDH15	p.D398A	Usher syndrome 1
GRIP_252	AD	11	64577603	G	T	spl	MEN1	.	Hyperparathyroidism
GRIP_252	AD	14	51382091	C	T	ex	PYGL	p.V422M	Glycogen storage disease 6
GRIP_252	AD	15	89868870	G	A	ex	POLG	p.P587L	Progressive external ophthalmoplegia
GRIP_252	AD	15	89873415	G	A	ex	POLG	p.T251I	Progressive external ophthalmoplegia
GRIP_252	AD	20	21690094	A	C	int	PAX1	.	Klippel-Feil syndrome
GRIP_252	AD	22	36691696	A	G	ex	MYH9	p.S1114P	Alport syndrome with macrothrombocytopaenia
GRIP_252	AR	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
GRIP_252	AR	15	45408933	T	C	int	DUOXA2	.	Hypothyroidism
GRIP_252	AR	16	56917953	T	C	int	SLC12A3	.	Gitelman syndrome%2C without hypomagnesaemia
GRIP_253	AD	2	220439916	G	A	ex	INHA	p.A257T	Premature ovarian failure
GRIP_253	AD	6	26091179	C	G	ex	HFE	p.H63D	Haemochromatosis%2C association with
GRIP_253	AD	8	55542540	G	A	ex	RP1	p.C2033Y	Retinitis pigmentosa
GRIP_253	AD	8	106431420	A	G	ex	ZFPM2	p.E30G	Tetralogy of Fallot
GRIP_253	AD	10	27389395	G	C	UTR5	ANKRD26	.	Thrombocytopaenia 2
GRIP_253	AD	10	55955444	T	G	ex_spl	PCDH15	p.D398A	Usher syndrome 1
GRIP_253	AD	14	64676751	C	T	ex	SYNE2	p.T6211M	Muscular dystrophy%2C Emery-Dreifuss
GRIP_253	AD	16	3706649	G	A	ex	DNASE1	p.V111M	Autoimmune thyroid disease

GRIP_253	AD	16	81298282	C	T	ex	BCO1	p.T170M	Hypercarotenemia and hypovitaminosis A
GRIP_253	AD	17	7127894	C	T	int	ACADVL	.	Very long chain acyl-CoA dehydrogenase deficiency
GRIP_253	AD	19	11552120	G	A	ex	PRKCSH	p.R139H	Polycystic liver disease
GRIP_253	AR	11	824789	T	C	ex	PNPLA2	p.L481P	Myopathy%2C late-onset
GRIP_253	AR	15	45408933	T	C	int	DUOXA2	.	Hypothyroidism
GRIP_253	AR	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2
GRIP_259	AD	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
GRIP_259	AD	6	26091179	C	G	ex	HFE	p.H63D	Haemochromatosis%2C association with
GRIP_259	AD	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
GRIP_259	AD	16	29825015	G	GC	ex	PRRT2	p.A214fs	Paroxysmal kinesigenic dyskinesia
GRIP_259	AR	6	6320808	T	G	int	F13A1	.	Factor XIII deficiency
GRIP_259	AR	17	3561396	C	T	ex	CTNS	p.T260I	Cystinosis%2C nephropathic
GRIP_292	AD	1	211654499	T	C	ex	RD3	p.K87E	Retinitis pigmentosa%2C autosomal recessive
GRIP_292	AD	2	152385774	C	G	ex	NEB	p.D5516H	Nemaline myopathy
GRIP_292	AD	11	22296266	C	T	ex	ANO5	p.S795L	Muscular dystrophy%2C limb girdle 2L
GRIP_292	AD	17	8140757	GCTTT	G	ex	CTC1	p.K242fs	Coats plus
GRIP_292	AD	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2
GRIP_315	AD	1	27684750	G	A	ex	MAP3K6	p.P938L	Gastric cancer%2C predisposition to
GRIP_315	AD	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
GRIP_315	AD	1	161136225	G	T	UTR5	PPOX	.	Porphyria%2C variegate
GRIP_315	AD	1	161172233	C	A	ex	NDUFS2	p.P20T	Mitochondrial complex I deficiency
GRIP_315	AD	2	31805826	C	T	ex	SRD5A2	UNKNOWN	Hypospadias%2C mild
GRIP_315	AD	2	32289031	C	T	ex	SPAST	p.S44L	Spastic paraplegia
GRIP_315	AD	4	128843111	C	G	ex	MFSD8	p.E336Q	Macular dystrophy%2C nonsyndromic
GRIP_315	AD	6	166574346	G	A	ex	T	p.A280V	Vertebral malformation
GRIP_315	AD	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
GRIP_315	AD	7	117199644	ATCT	A	ex	CFTR	p.507_508del	Cystic fibrosis
GRIP_315	AD	8	55542540	G	A	ex	RP1	p.C2033Y	Retinitis pigmentosa
GRIP_315	AD	8	61693942	G	GAAAGCA	ex	CHD7	p.K683delinsKKA	Kallmann syndrome
GRIP_315	AD	9	12702410	TACAA	T	ex	TYRP1	p.T352fs	Albinism%2C oculocutaneous 3
GRIP_315	AD	9	136495229	C	G	intgen	FAM163B_DBH	.	Altered enzyme activity
GRIP_315	AD	10	13325784	C	T	ex	PHYH	p.R157Q	Phytanoyl-CoA hydroxylase deficiency%2C partial
GRIP_315	AD	10	55955444	T	G	ex_spl	PCDH15	p.D398A	Usher syndrome 1
GRIP_315	AD	16	56917953	T	C	int	SLC12A3	.	Gitelman syndrome%2C without hypomagnesaemia
GRIP_315	AD	16	86602272	A	G	ex	FOXC2	p.Q444R	Lymphoedema%2C primary
GRIP_315	AD	16	89595983	C	CT	ex_spl	SPG7	p.A286fs	Spastic paraplegia
GRIP_315	AD	18	3188778	C	T	ex	MYOM1	p.E247K	Cardiomyopathy%2C dilated
GRIP_315	AD	20	3687141	C	A	ex	SIGLEC1	p.E88X	SIGLEC1 deficiency
GRIP_315	AR	6	6320808	T	G	int	F13A1	.	Factor XIII deficiency
GRIP_315	AR	11	824789	T	C	ex	PNPLA2	p.L481P	Myopathy%2C late-onset
GRIP_315	AR	15	45408933	T	C	int	DUOXA2	.	Hypothyroidism
GRIP_315	AR	17	3561396	C	T	ex	CTNS	p.T260I	Cystinosis%2C nephropathic
GRIP_315	AR	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2
GRIP_315	AR	X	153296347	G	A	ex	MECP2	p.T323M	Rett syndrome
GRIP_326	AD	1	5927943	G	A	ex	NPHP4	p.A598V	Cardiovascular malformations
GRIP_326	AD	1	151139877	C	T	ex	SCNM1_TNFAIP8L2-SCNM1	p.R94C	Epilepsy%2C idiopathic generalized

GRIP_326	AD	1	247588858	C	A	ex	NLRP3	p.Q705K	Cryopyrin-associated periodic syndrome%2C atypical
GRIP_326	AD	2	215595159	T	C	ex	BARD1	p.R146R	Breast and:or ovarian cancer
GRIP_326	AD	3	15686693	G	C	ex	BTD	p.D444H	Biotinidase deficiency%2C partial
GRIP_326	AD	6	162206909	G	A	ex	PRKN	p.R107C	Parkinsonism%2C juvenile%2C autosomal recessive
GRIP_326	AD	7	117230454	G	C	ex	CFTR	p.G576A	Congenital absence of vas deferens
GRIP_326	AD	7	117232223	C	T	ex	CFTR	p.R668C	Cystic fibrosis
GRIP_326	AD	7	117251692	G	A	ex	CFTR	p.R1066H	Cystic fibrosis
GRIP_326	AD	9	34648167	A	G	ex_spl	GALT	p.Q79R	Galactosaemia
GRIP_326	AD	9	136494594	G	T	intgen	FAM163B_DBH	.	Altered enzyme activity
GRIP_326	AD	10	27389395	G	C	UTR5	ANKRD26	.	Thrombocytopaenia 2
GRIP_326	AD	10	55955444	T	G	ex_spl	PCDH15	p.D398A	Usher syndrome 1
GRIP_326	AD	11	67816547	G	A	spl	TCIRG1	.	Osteopetrosis%2C autosomal recessive
GRIP_326	AD	14	94847262	T	A	ex	SERPINA1	p.E288V	Antitrypsin alpha 1 deficiency%2C partial
GRIP_326	AD	15	65116390	C	A	ex	PIF1	p.E49X	Multiple sessile serrated adenoma
GRIP_326	AD	16	8905010	G	A	ex	PMM2	p.R141H	Congenital disorder of glycosylation 1a
GRIP_326	AD	16	28884858	C	T	ex	SH2B1	p.A663V	Obesity%2C severe%2C early-onset
GRIP_326	AD	16	56917953	T	C	int	SLC12A3	.	Gitelman syndrome%2C without hypomagnesaemia
GRIP_326	AD	16	86602272	A	G	ex	FOXC2	p.Q444R	Lymphoedema%2C primary
GRIP_326	AD	16	89595983	C	CT	ex_spl	SPG7	p.A286fs	Spastic paraplegia
GRIP_326	AD	17	7576841	A	G	int	TP53	.	Breast and:or ovarian cancer
GRIP_326	AD	17	18051447	C	T	ex	MYO15A	p.T2205I	Sensorineural deafness in SMS
GRIP_326	AD	17	42337247	C	T	ex	SLC4A1	p.R180H	Spherocytosis
GRIP_326	AD	22	42463140	G	C	ex	NAGA	p.S160C	N-acetylgalactosaminidase alpha deficiency
GRIP_326	AD	X	100630121	G	A	int	BTK	.	Agammaglobulinaemia
GRIP_332	AD	1	43804305	G	C	ex	MPL	p.R102P	Amegakaryocytic thrombocytopaenia%2C congenital
GRIP_332	AD	2	56145171	T	G	ex	EFEMP1	p.D49A	Cuticular drusen
GRIP_332	AD	2	166012375	C	T	ex	SCN3A	p.R357Q	Epilepsy%2C focal
GRIP_332	AD	2	179599667	G	C	ex	TTN	p.P3751R	Cardiac dysrhythmia
GRIP_332	AD	6	7586120	T	A	UTR3	DSP	.	Cardiomyopathy%2C arrhythmogenic right ventricular
GRIP_332	AD	6	66417039	G	A	UTR5	EYS	.	Retinitis pigmentosa%2C autosomal recessive
GRIP_332	AD	7	21882209	G	A	ex	DNAH11	p.R3580H	Primary ciliary dyskinesia
GRIP_332	AD	7	42007201	T	C	ex	GLI3	p.I808M	Greig cephalopolysyndactyly syndrome
GRIP_332	AD	11	22294441	C	G	ex	ANO5	p.T713S	Muscular dystrophy
GRIP_332	AD	11	118959807	C	T	ex	HMBS	p.T59I	Porphyria%2C acute intermittent
GRIP_332	AD	15	75189391	G	A	ex	MPI	p.R245H	Congenital disorder of glycosylation 1b
GRIP_332	AD	18	50848468	A	G	ex	DCC	p.N702S	Mirror movements%2C congenital
GRIP_332	AD	19	35775902	G	A	ex	HAMP	p.G71D	Haemochromatosis
GRIP_332	AD	20	5282973	G	A	ex	PROKR2	p.P290S	Kallmann syndrome
GRIP_332	AD	22	50962423	C	T	ex	SCO2	p.E140K	Cytochrome c oxidase deficiency
GRIP_332	AR	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
GRIP_338	AD	2	31805826	C	T	ex	SRD5A2	UNKNOWN	Hypospadias%2C mild
GRIP_338	AD	6	26091179	C	G	ex	HFE	p.H63D	Haemochromatosis%2C association with
GRIP_338	AD	9	136495229	C	G	intgen	FAM163B_DBH	.	Altered enzyme activity
GRIP_338	AD	14	88452941	T	C	ex	GALC	p.T89A	Krabbe disease

GRIP_338	AD	16	56917953	T	C	int	SLC12A3	.	Gitelman syndrome%2C without hypomagnesaemia
GRIP_338	AD	16	89613145	C	T	ex	SPG7	p.A510V	Upper motor neuron syndrome
GRIP_338	AD	17	46022065	G	A	ex	PNPO	p.R116Q	PNPO deficiency
GRIP_338	AD	17	47489162	C	A	ex	PHB	p.R43L	Haemolytic uraemic syndrome%2C atypical
GRIP_338	AR	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
Ital_034	AD	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
Ital_034	AD	14	95581899	G	A	int	DICER1	.	Breast cancer
Ital_034	AD	22	30642690	G	T	UTR5	LIF	.	Female infertility
Ital_034	AR	6	6320808	T	G	int	F13A1	.	Factor XIII deficiency
Ital_034	AR	15	45408933	T	C	int	DUOXA2	.	Hypothyroidism
Ital_034	AR	16	56917953	T	C	int	SLC12A3	.	Gitelman syndrome%2C without hypomagnesaemia
Ital_034	AR	17	3561396	C	T	ex	CTNS	p.T260I	Cystinosis%2C nephropathic
Ital_034	AR	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2
Ital_158	AD	10	72360577	G	A	ex	PRF1	p.R28C	Arthritis%2C juvenile
Ital_158	AD	11	824789	T	C	ex	PNPLA2	p.L481P	Myopathy%2C late-onset
Ital_175	AD	1	55509622	G	A	ex	PCSK9	p.R105Q	Hypercholesterolaemia
Ital_175	AD	2	38301847	C	T	ex	CYP1B1	p.E229K	Glaucoma%2C primary congenital
Ital_175	AD	7	84636183	G	T	ex	SEMA3D	p.P615T	Hirschsprung disease
Ital_175	AD	8	106801042	G	C	ex	ZFPM2	p.S210T	Ovotesticular disorder of sex development
Ital_175	AD	10	17113456	C	T	ex	CUBN	p.S865N	Megaloblastic anaemia
Ital_175	AD	10	27389395	G	C	UTR5	ANKRD26	.	Thrombocytopaenia 2
Ital_175	AD	10	43613907	T	A	ex	RET	p.Y791N	Hirschsprung disease
Ital_175	AD	14	95581899	G	A	int	DICER1	.	Breast cancer
Ital_175	AR	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
Ital_175	AR	6	6320808	T	G	int	F13A1	.	Factor XIII deficiency
Ital_175	AR	11	824789	T	C	ex	PNPLA2	p.L481P	Myopathy%2C late-onset
Ital_175	AR	17	3561396	C	T	ex	CTNS	p.T260I	Cystinosis%2C nephropathic
Ital_175	AR	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2
Ital_176	AD	6	161006077	C	T	spl	LPA	.	Lp(a) deficiency
Ital_176	AD	15	73615786	G	C	ex	HCN4	p.P883R	Sinus bradycardia & myocardial noncompaction
PD172	AD	1	161172233	C	A	ex	NDUFS2	p.P20T	Mitochondrial complex I deficiency
PD172	AD	11	824789	T	C	ex	PNPLA2	p.L481P	Myopathy%2C late-onset
PD172	AD	15	28326942	C	T	ex	OCA2	p.G27R	Albinism%2C oculocutaneous II
PD172	AD	16	16272711	C	T	ex	ABCC6	p.V787I	Pseudoxanthoma elasticum%2C autosomal recessive
PD172	AR	10	55955444	T	G	ex_spl	PCDH15	p.D398A	Usher syndrome 1
PD257	AD	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
PD257	AD	1	156085059	A	G	ex	LMNA	p.K117R	Cardiac disease
PD257	AD	2	21249840	A	T	spl	APOB	.	Hypobetalipoproteinaemia
PD257	AD	2	71738977	G	A	ex	DYSF	p.G129E	Muscular dystrophy%2C limb girdle : Miyoshi myopathy
PD257	AD	2	99006159	C	T	ex	CNGA3	p.P145L	Colour-blindness%2C total
PD257	AD	3	38603958	G	A	ex	SCN5A	p.T1250M	Long QT syndrome
PD257	AD	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
PD257	AD	7	117250575	G	C	ex_spl	CFTR	p.L997F	Congenital absence of vas deferens
PD257	AD	7	128043703	C	T	int	IMPDH1	.	Retinitis pigmentosa%2C autosomal dominant
PD257	AD	9	100616939	C	G	ex	FOXE1	p.A248G	Thyroid carcinoma%2C non-medullary

PD257	AD	10	55973755	G	A	ex	PCDH15	p.L310F	Hearing loss
PD257	AD	11	6631386	G	C	ex	ILK	p.Q301H	Cardiomyopathy%2C hypertrophic
PD257	AD	11	22239825	C	T	ex	ANO5	p.R57W	Myopathy : muscular dystrophy
PD257	AD	12	103306579	C	T	ex	PAH	p.R53H	Phenylketonuria
PD257	AD	13	23913549	T	C	ex	SACS	p.N1342S	Intellectual disability%2C cerebellar taxia
PD257	AD	13	28498702	C	A	ex	PDX1	p.P239Q	Diabetes mellitus%2C type 2
PD257	AD	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2
PD257	AD	18	44140279	CTCT	C	ex	LOXHD1	p.942_943del	Hearing loss
PD257	AR	6	6320808	T	G	int	F13A1	.	Factor XIII deficiency
PD257	AR	8	55542540	G	A	ex	RP1	p.C2033Y	Retinitis pigmentosa
PD257	AR	15	45408933	T	C	int	DUOXA2	.	Hypothyroidism
PD257	AR	16	56917953	T	C	int	SLC12A3	.	Gitelman syndrome%2C without hypomagnesaemia
PD257	AR	17	3561396	C	T	ex	CTNS	p.T260I	Cystinosis%2C nephropathic
PD257	AR	X	148044334	G	A	ex	AFF2	p.R568H	Autism
PD290	AD	1	216219858	C	A	ex	USH2A	p.K2080N	Retinitis pigmentosa
PD290	AD	3	121491530	C	T	ex	IQCB1	p.E348K	Leber congenital amaurosis
PD290	AD	12	6232308	C	T	ex_spl	VWF	p.G19R	Von Willebrand disease 1
PD290	AD	17	7576841	A	G	int	TP53	.	Breast and:or ovarian cancer
PD290	AD	X	30327105	C	T	ex	NR0B1	p.V126M	Adrenal hypoplasia
PD290	AD	X	55057617	G	C	up	ALAS2	.	Sideroblastic anaemia
PD290	AR	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
PD290	AR	11	824789	T	C	ex	PNPLA2	p.L481P	Myopathy%2C late-onset
PD290	AR	17	3561396	C	T	ex	CTNS	p.T260I	Cystinosis%2C nephropathic
PD291	AD	1	161180482	G	A	ex	NDUFS2	p.R323Q	Isolated Complex I deficiency
PD291	AD	2	32289197	T	TGCCTCG	ex	SPAST	p.P99delinsPAS	Amyotrophic lateral sclerosis
PD291	AD	2	128186202	C	T	ex	PROC	p.R356C	Protein C deficiency
PD291	AD	8	55542540	G	A	ex	RP1	p.C2033Y	Retinitis pigmentosa
PD291	AD	9	136494594	G	T	intgen	FAM163B_DBH	.	Altered enzyme activity
PD291	AD	20	52773755	G	A	ex	CYP24A1	p.P437L	Hypercalcaemia%2C idiopathic infantile
PD291	AD	21	45709642	C	T	ex	AIRE	p.P252L	APECED
PD291	AR	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
PD291	AR	16	56917953	T	C	int	SLC12A3	.	Gitelman syndrome%2C without hypomagnesaemia
PD296	AD	2	27726431	G	A	ex	GCKR	p.R232Q	Elevated HDL-cholesterol
PD296	AD	5	13735418	G	T	ex	DNAH5	p.S3861R	Primary ciliary dyskinesia
PD296	AD	9	126135490	G	A	ex	CRB2	p.G894S	Focal segmental glomerulosclerosis
PD296	AD	11	116701353	C	T	ex_spl	APOC3	p.R19X	Apolipoprotein C3 deficiency with apparent cardioprotection
PD296	AD	12	49430947	T	C	ex	KMT2D	p.M3398V	Kabuki syndrome
PD296	AD	17	7128292	G	A	ex	ACADVL	p.R593Q	Very long chain acyl-CoA dehydrogenase deficiency
PD296	AR	11	824789	T	C	ex	PNPLA2	p.L481P	Myopathy%2C late-onset
PD300	AD	1	5927943	G	A	ex	NPHP4	p.A598V	Cardiovascular malformations
PD300	AD	1	161172233	C	A	ex	NDUFS2	p.P20T	Mitochondrial complex I deficiency
PD300	AD	9	136494594	G	T	intgen	FAM163B_DBH	.	Altered enzyme activity
PD300	AD	11	824789	T	C	ex	PNPLA2	p.L481P	Myopathy%2C late-onset
PD300	AD	15	31329942	G	A	ex	TRPM1	p.A865V	Stationary night blindness%2C congenital
PD300	AD	15	45408933	T	C	int	DUOXA2	.	Hypothyroidism
PD300	AD	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2

PD300	AD	18	44140279	CTCT	C	ex	LOXHD1	p.942_943del	Hearing loss
PD300	AD	19	50771512	G	A	ex	MYH14	p.R933H	Hearing loss
PD300	AR	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
PD300	AR	6	6320808	T	G	int	F13A1	.	Factor XIII deficiency
PD300	AR	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
PD300	AR	17	3561396	C	T	ex	CTNS	p.T260I	Cystinosis%2C nephropathic
PD313	AD	1	114372214	C	G	ex_spl	PTPN22	p.K695N	Diabetes%2C type 1%2C increased risk
PD313	AD	2	167138296	T	C	ex	SCN9A	p.K655R	Febrile seizures
PD313	AD	5	139930460	A	G	ex	SRA1	p.I151T	Hypogonadotropic hypogonadism
PD313	AD	9	103046765	G	C	ex	INVS	p.A324P	Nephronophthisis 2
PD313	AD	9	136494594	G	T	intgen	FAM163B_DBH	.	Altered enzyme activity
PD313	AD	11	824789	T	C	ex	PNPLA2	p.L481P	Myopathy%2C late-onset
PD313	AD	15	89873364	C	G	ex	POLG	p.G268A	Progressive external ophthalmoplegia
PD313	AR	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
PD317	AD	10	13151198	C	G	ex	OPTN	p.H26D	Glaucoma 1%2C open angle
PD317	AD	11	71146886	C	G	spl	DHCR7	.	Smith-Lemli-Opitz syndrome
PD317	AD	13	52508989	G	A	ex	ATP7B	p.T1227M	Wilson disease
PD317	AD	19	15302941	T	C	ex	NOTCH3	p.H170R	CADASIL
PD317	AD	19	50370404	G	A	ex	PNKP	p.P20S	Epileptic encephalopathy
PD317	AR	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
PD317	AR	15	45408933	T	C	int	DUOXA2	.	Hypothyroidism
PD317	AR	17	3561396	C	T	ex	CTNS	p.T260I	Cystinosis%2C nephropathic
Tolosa_HCB1	AD	1	161172233	C	A	ex	NDUFS2	p.P20T	Mitochondrial complex I deficiency
Tolosa_HCB1	AD	3	37089131	A	C	ex	MLH1	p.K377T	Colorectal cancer%2C non-polyposis
Tolosa_HCB1	AD	6	129573388	AAG	A	ex	LAMA2	p.K682fs	Muscular dystrophy%2C merosin deficient
Tolosa_HCB1	AD	10	55955444	T	G	ex_spl	PCDH15	p.D398A	Usher syndrome 1
Tolosa_HCB1	AD	17	73836585	C	T	spl	UNC13D	.	Juvenile idiopathic arthritis
Tolosa_HCB1	AD	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2
Tolosa_HCB1	AD	19	34262922	C	T	ex	CHST8	p.R77W	Peeling skin syndrome%2C type A
Tolosa_HCB1	AD	22	18566288	C	G	ex	PEX26	p.L153V	Peroxisome biogenesis disorder
Tolosa_HCB1	AR	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
Tolosa_HCB1	AR	6	6320808	T	G	int	F13A1	.	Factor XIII deficiency
Tolosa_HCB1	AR	15	45408933	T	C	int	DUOXA2	.	Hypothyroidism
Tolosa_HCB1	AR	17	3561396	C	T	ex	CTNS	p.T260I	Cystinosis%2C nephropathic
Tolosa_HCB2	AD	3	193377336	C	T	ex	OPA1	p.R745W	Optic atrophy 1
Tolosa_HCB2	AD	9	139390585	C	T	ex	NOTCH1	p.V2536I	Aortic valve disease
Tolosa_HCB2	AD	10	55955444	T	G	ex_spl	PCDH15	p.D398A	Usher syndrome 1
Tolosa_HCB2	AD	20	21690094	A	C	int	PAX1	.	Klippel-Feil syndrome
Tolosa_HCB2	AR	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
Tolosa_HCB2	AR	6	6320808	T	G	int	F13A1	.	Factor XIII deficiency
Tolosa_HCB2	AR	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
Tolosa_HCB2	AR	15	45408933	T	C	int	DUOXA2	.	Hypothyroidism
Tolosa_HCB2	AR	17	3561396	C	T	ex	CTNS	p.T260I	Cystinosis%2C nephropathic
Tolosa_HCB2	AR	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2
Tolosa_HCB4	AD	3	11313449	G	A	up	ATG7	.	Parkinson disease
Tolosa_HCB4	AD	8	55542540	G	A	ex	RP1	p.C2033Y	Retinitis pigmentosa
Tolosa_HCB4	AD	10	55955444	T	G	ex_spl	PCDH15	p.D398A	Usher syndrome 1
Tolosa_HCB4	AD	10	73491873	A	G	ex	CDH23	p.N1282S	Hearing loss
Tolosa_HCB4	AD	11	824789	T	C	ex	PNPLA2	p.L481P	Myopathy%2C late-onset
Tolosa_HCB4	AD	11	64525266	C	T	ex	PYGM	p.K127K	McArdle disease
Tolosa_HCB4	AD	14	95581899	G	A	int	DICER1	.	Breast cancer
Tolosa_HCB4	AD	16	56926915	T	A	ex	SLC12A3	p.S833T	Gitelman syndrome

Tolosa_HCB4	AD	17	6331702	T	A	ex	AIPL1	p.Y74F	Leber congenital amaurosis IV
Tolosa_HCB4	AR	1	98502934	G	T	ncRNA_int	MIR137HG	.	Schizophrenia%2C increased risk
Tolosa_HCB4	AR	7	94227276	T	G	ex	SGCE	p.S432R	Myoclonus dystonia syndrome
Tolosa_HCB4	AR	15	45408933	T	C	int	DUOXA2	.	Hypothyroidism
Tolosa_HCB4	AR	17	3561396	C	T	ex	CTNS	p.T260I	Cystinosis%2C nephropathic
Tolosa_HCB5	AD	15	44944037	C	T	ex	SPG11	p.E370K	Spastic paraplegia
Tolosa_HCB5	AD	16	3714438	C	T	ex	TRAP1	p.R416H	Congenital anomalies of the kidney and urinary tract:CAKUT in VACTERL
Tolosa_HCB5	AR	17	78079509	T	G	int	GAA	.	Glycogen storage disease 2

Table A.8: Variants annotated as disease causing with high confidence by HGMD. Fam = Family name, Chr = Chromsome, Pos = Position according hg19 genome, Ref = Reference allele, Alt = Alternate allele, ex = Exonic, spl = splicing, int = Intronic, intgen = Intergenic, up = Upstream of a gene.

APPENDIX B

ARTICLES

European Journal of Human Genetics
https://doi.org/10.1038/s41431-017-0034-x

ARTICLE

Exome-wide analysis of mutational burden in patients with typical and atypical Rolandic epilepsy

Dheeraj R. Bobbili[1] · **Dennis Lal**[2,3,4,5] · **Patrick May**[1] · **Eva M. Reinthaler**[6] · **Kamel Jabbari**[7] · **Holger Thiele**[2] · **Michael Nothnagel**[2] · **Wiktor Jurkowski**[1,8] · **Martha Feucht**[9] · **Peter Nürnberg**[2] · **Holger Lerche**[10] · **Fritz Zimprich**[6] · **Roland Krause**[1] · **Bernd A. Neubauer**[11] · **Eva M. Reinthaler**[6] · **Fritz Zimprich**[6] · **Martha Feucht**[12] · **Hannelore Steinböck**[13] · **Birgit Neophytou**[14] · **Julia Geldner**[15] · **Ursula Gruber-Sedlmayr**[16] · **Edda Haberlandt**[17] · **Gabriel M. Ronen**[18] · **Janine Altmüller**[2] · **Dennis Lal**[2] · **Peter Nürnberg**[2] · **Thomas Sander**[2] · **Holger Thiele**[2] · **Roland Krause**[1] · **Patrick May**[1] · **Rudi Balling**[1] · **Holger Lerche**[10] · **Bernd A. Neubauer**[11]
EUROEPINOMICS COGIE Consortium

Received: 8 May 2017 / Revised: 27 September 2017 / Accepted: 6 October 2017
© European Society of Human Genetics 2018

Abstract

Rolandic epilepsy (RE) is the most common focal epilepsy in childhood. To date no hypothesis-free exome-wide mutational screen has been conducted for RE and atypical RE (ARE). Here we report on whole-exome sequencing of 194 unrelated patients with RE/ARE and 567 ethnically matched population controls. We identified an exome-wide significantly enriched burden for deleterious and loss-of-function variants only for the established RE/ARE gene *GRIN2A*. The statistical significance of the enrichment disappeared after removing ARE patients. For several disease-related gene-sets, an odds ratio >1 was detected for loss-of-function variants.

Introduction

Rolandic epilepsy (RE), or epilepsy with centro-temporal spikes (CTS), is one of the most common epilepsy syndromes of childhood. RE is related to rarer and less benign epilepsy syndromes, including atypical benign partial epilepsy, Landau–Kleffner syndrome and epileptic encephalopathy with continuous spike-and-waves during sleep, referred to as RE-related syndromes or atypical rolandic epilepsy (ARE) [1]. In up to 20% sib pairs or families, mutations affecting *GRIN2A*, a subunit of the excitatory glutamate receptor *N*-methyl-D-aspartate (NMDA), were found implicated as major risk factor for RE and ARE by us and others [2, 3]. Recently, the association of the genes *RBFOX1*, *RBFOX3*, *DEPDC5*, *GABRG2* and genomic duplications at 16p11.2 in 1.5–2.0% was identified in patients with RE and ARE [4–6] through candidate gene and loci screens. In the current study, an unbiased exome-wide survey was conducted in the RE/ARE cohort.

Patients and methods

Study participants

Two hundred and four unrelated European Rolandic cases (182 RE, 22 ARE) and 728 population control subjects were included [6]. Children with (typical) RE suffer from perisylvian oromotor seizures frequently starting during sleep. In adolescence, the epilepsy resolves spontaneously, frequently without any intellectual sequels. ARE share the essential electroencephalography feature with RE but show a different seizure symptomatology by their own or in addition to rolandic seizures. Seizures, like in RE, resolve

---

EuroEPINOMICS consortium contributing partners are listed above References.

D.R. Bobbili, D. Lal and P. May contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41431-017-0034-x) contains supplementary material, which is available to authorized users.

✉ Roland Krause
roland.krause@uni.lu

✉ Bernd A. Neubauer
Bernd.A.Neubauer@paediat.med.uni-giessen.de

Extended author information available on the last page of the article

Published online: 22 January 2018

spontaneously, but cognitive outcome is guarded in ARE. In detail, these epilepsies are: atypical benign partial epilepsy of childhood, with atonic seizures and atypical absences in addition to rolandic seizures; Landau–Kleffner syndrome, with loss of speech and cognitive decline; and epilepsia-aphasia syndrome with seizures and language dysfunction [1, 6]. Written informed consent was obtained from participating subjects and, if appropriate, from both patients and adolescents.

Data generation and processing

Exome sequencing of all individuals was performed with the Illumina HiSeq 2000 using the EZ Human Exome Library Kit (NimbleGen, Madison, WI). Sequencing adapters were trimmed and samples with <30× mean depth or <70% total exome coverage at 20× mean depth of coverage were excluded from further analysis. Variant calling was performed in targeted exonic intervals with 100 bp padding using the GATK best practices pipeline [7] against the GRCh37 human reference genome followed by multi-allelic variant decomposition and left normalization. Samples were excluded from further analysis if they (i) were not ethnically matched, (ii) were related, (iii) showed discrepancy with reported sex, (iv) had an excess heterozygosity >3 SD in any of the quality metrics (NALT, NMIN, NHET, NVAR, RATE and SINGLETON statistics as calculated by PLINKseq i-stats parameter [8]. The genotypes of variants with read depth <10 or genotype quality <20 were set to missing. Variants were excluded if they (i) failed variant quality score recalibration (VQSR) or GATK recommended hard filter, (ii) showed missingness >3%, (iii) were present in repeat regions or (iv) had an average read depth <10 in either cases or controls. The ExAC variants were restricted to the exonic intervals used for variant calling in this study, not present in the repeat regions and passed the VQSR threshold.

Variant annotation and filtering

Variants were annotated using ANNOVAR [9] version 2015 Mar 22 with RefSeq and Ensembl, Combined Annotation Dependent Depletion (CADD) scores [10], allele frequencies and dbNSFP (v3.0) annotations. The samples used in this study are of Non-Finnish European (NFE) ancestry, hence to investigate rare variants, we selected variants having a minor allele frequency <0.005 in the European populations of the 1000 genomes, Exome Variant Server and the NFE data from ExAC. We generated three classes of variants for further analyses: (1) deleterious variants (CADD15), which were defined as missense variants with a CADD Phred score >15 as it is the median value across all missense and canonical splice site variants [10], (2) loss-of-function (LOF) variants comprising all rare indels, stop gain, stop loss and splice site variants (2 nt plus/minus the exon boundary), (3) CADD15+LOF variants as the union of the above two datasets, and (4) rare synonymous variants.

Single variant and gene association analysis

For the statistical analysis, we employed two independent control cohorts (available in-house and ExAC) to increase reliability and power of the statistical tests. For single variant burden analysis, we applied the single score method in RVTESTS [11] to cases and in-house controls, for which individual genotypes were available. For gene burden analysis, a 2 × 2 contingency table was constructed by counting the number of alternate allele counts per gene in patients vs. controls (in-house controls and NFE ExAC controls). We then obtained a one-sided p-value, odds ratios and the 95% confidence intervals [12] by using Fisher's exact test. Resulting p-values were corrected for 18,668 RefSeq protein-coding genes [13] by Bonferroni approach. Finally, to ensure the exclusion of false positive association results and following the 'rare variant of large effect hypothesis', we selected those genes that are present in the first quartile of the Residual Variant Intolerance Score (RVIS) distribution [14].

Selection of gene-sets

We investigated the following four neuron-related gene-sets: (1) genes encoding proteins at synapses downloaded from the SynaptomeDB [15] database (“SYNAPTIC_GENES”, $N = 1887$), (2) genes of postsynaptic signalling complexes including NMDA receptors (NMDARs) and the neuronal activity-regulated cytoskeleton-associated protein (ARC) [16] (“NMDAR_ARC_COMPLEX”, $N = 80$), (3) genes encoding proteins at the inhibitory synapses (“INHIBITORY”, $N = 5941$) and excitatory synapses (“EXCITATORY”, $N = 5261$) [17], and (4) glutamate receptor subunit encoding genes (“GLUTAMATE_RECEPTORS”, $N = 18$). In addition, we included five gene-sets associated with disease and/or mutational intolerance: (1) genes encoding targets of Fragile-X-Mental-Retardation-1-Protein [18] (“FMRP_TARGETS_-DARNELL”, $N = 1772$), (2) genes intolerant for variants from ExAC (“EXAC_CONSTRAINED_GENES”, $N = 3230$), (3) genes intolerant for loss-of-function variants [19] ('constrained') (“CONSTRAINED_GENES_SAMOCHA”, $N = 1004$), (4) a curated list of dominant genes associated with developmental delay obtained from the DECIPHER database [20] (“DDG2P_MONOALLELIC”, $N = 299$), and (5) genes found related before to epileptic encephalopathies [21] (“EPILEPTIC_ENCEPHALOPATHY”, $N = 73$). As control data sets, we used (1) for each dataset the corresponding set of synonymous variants, and (2) the ‘non-constraint’ gene-set including RefSeq genes that have been found tolerant to LOF variants (“GENES_WITHOUT_CONSTRAINT”, $N =$

14,417). *GRIN2A*, as the most significant single gene from the burden analysis, was excluded from all gene-sets in order to test if other genes also contribute to the disease association.

Data availability

All the CADD15+LOF variants from our study within the "EPILEPTIC_ENCEPHALOPATHY" gene-set were deposited in the Leiden Open Variation Database (LOVD) (https://databases.lovd.nl/shared/genes). The accession numbers of the deposited variants in LOVD are 188117–188549. Also, the variants present in the cases within the "EPILEPTIC_ENCEPHALOPATHY" gene-set are available in the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) with the accession numbers SCV000588243–SCV000588353. The variants that were described in our previous studies are indicated in Supplementry Table 1.

Gene-set association analysis

The gene-set association analysis for the different types of variants was performed by using a logistic regression approach using R (version 3.2) and adjusting for the following confounding variables: the total number of called genotypes per sample, the total number of rare coding variants per sample, the total number of rare coding singletons (variants observed only once in the entire dataset) per sample, calculated sex, the first four principal components, and the total number of variants per sample for each variant class.

Results

Exome sequencing and variant filtering

We performed whole-exome sequencing on 204 patients with RE/ARE and 728 population controls. After quality control, the final dataset consisted of 19 ARE, 175 RE and 567 control samples. From the total of 761 samples, 226,521 exonic and splice site variants were called. The mean transition/transversion ratio equalled 3.39 per sample. After the final filtering 45,881 CADD15, 10,326 LOF and 38,802 synonymous variants were analysed.

Association analysis

To investigate the mutational burden within the RE spectrum, all associations were assessed for both RE and ARE

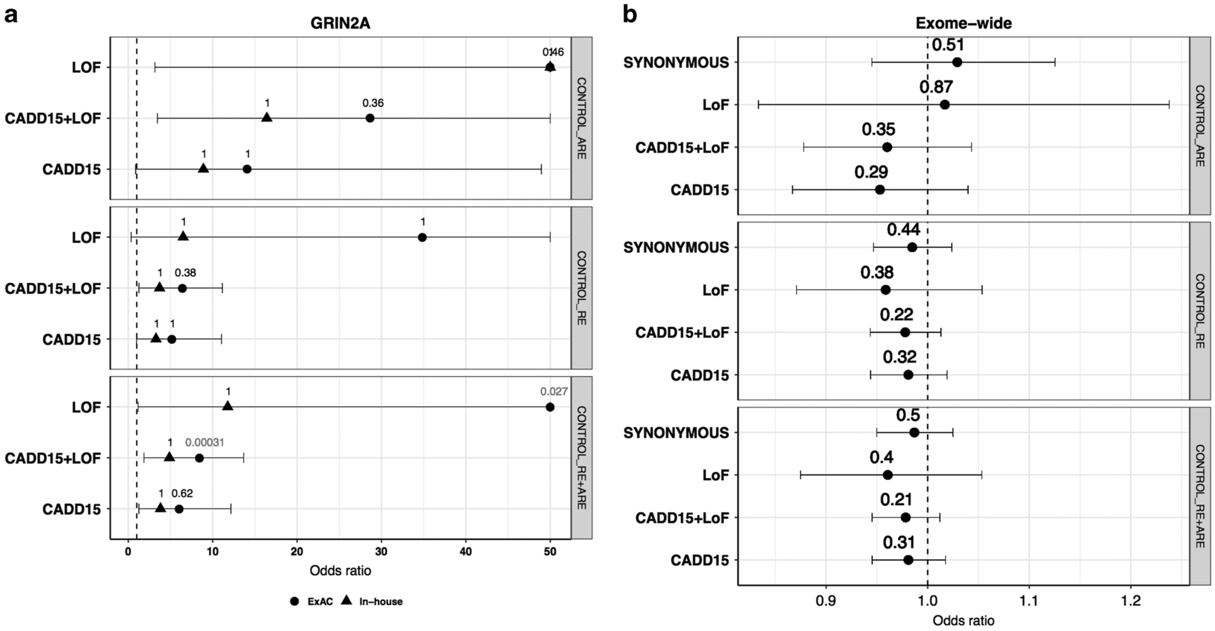


Fig. 1 Burden analysis. Typical Rolandic epilepsy is represented as RE, atypical rolandic epilepsy as ARE and RE plus ARE as ROLANDIC. On the *x* axis, the odds ratios in cases vs controls are given. The names of the variant classes are given on the *y* axis. Each panel represents a different dataset. The dashed vertical line represents the expected odds ratio of 1. The horizontal lines indicate 95% confidence intervals. **a** Assessment of risk for deleterious variants in GRIN2A against two control groups (ExAC and In-house). The values on top of each point represent multiple-testing corrected *p*-values, the ones in red are significant *p*-values and the ones in black are the *p*-values that are not significant after multiple-testing correction. The odds ratios are restricted to a maximum value of 50. **b** Exome-wide burden analysis by different variant classes. The values on top of each point represent the *p*-value. Synonymous variants serve as a control functional group (colour figure online)

separately and by combining cases from both phenotypes while assuming them to be a single disease. In comparison to 567 in-house controls, we did not observe statistically significant burden in any of the variants or genes in cases after multiple-testing correction. In order to increase the statistical power, we used the non-Finnish European (NFE) ExAC cohort as an additional control dataset. Association testing against the much larger NFE-ExAC cohort ($N = 33,370$) identified an exome-wide significant burden for CADD15, CADD15+LOF and LOF variants for *GRIN2A*

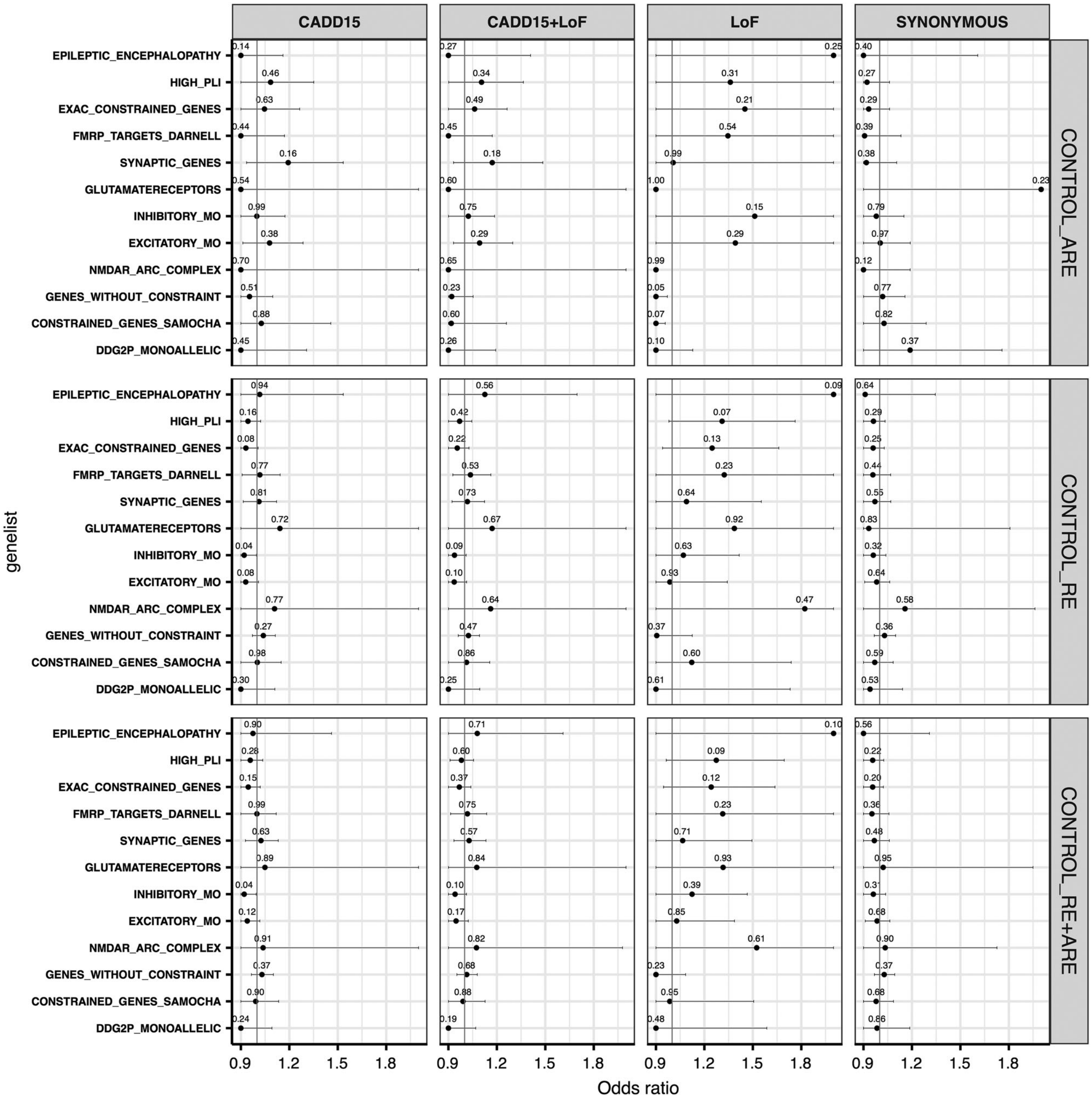


Fig. 2 Gene-set burden across different variant classes. Each panel represent a different variant class. The synonymous variants serve as a control variant class. *GRIN2A* was removed from all gene-sets to identify other contributing genes. On the *x* axis, the odds ratios in cases vs controls are given. On the *y* axis, the names of different gene-sets are given. The red vertical line represents the expected odds ratio of 1. The horizontal lines indicate 95% confidence intervals and are restricted to the maximum of odds ratios over all gene-sets. In that case, points are represented as the points without error bars to their right. The uncorrected *p*-values are shown on top of each point. CADD15 = deleterious predicted missense variants. LOF = Loss-of-function variants (colour figure online)

within the combined typical and atypical (RE+ARE) cohort. No other variant-intolerant gene (i.e., being present in the first quartile of RVIS) was significantly enriched for variants in any of the tested patient groups. Although variant enrichment for *GRIN2A* was not found to be significant after correction for RE and/or ARE separately, the odds ratio for *GRIN2A* consistently exceeded unity in all the considered datasets (Fig. 1a).

Exome-wide and gene-set burden analysis

Assuming a shared mutational burden in patients across gene-sets of convergent function and/or pathways, we performed gene-set burden analyses by using the in-house controls. A logistic regression approach was used to account for various confounding variables (see Methods section). No significant exome-wide burden was observed across the different variant classes (Fig. 1b). Despite the fact that none of the gene-sets showed a significant result after multiple-testing correction, we found several gene-sets with an odds ratio >1 for the CADD15, CADD15+LOF and LOF variant classes, especially for the LOF variants, but not for synonymous variants (Fig. 2). A similar result was seen when we performed the analysis with ARE and RE independently.

Discussion

We performed the first exome-wide association study investigating rare genetic variants of large effect in 194 patients with childhood focal epilepsies with CTS in comparison with 567 in-house and online available 33,370 population controls from the ExAC database. By performing an unbiased gene-burden analysis of patients against the in-house and ExAC controls (Fig. 1a), we show that, only for *GRIN2A* rare CADD15, CADD15+LOF and LOF variants are significantly more frequent in RE and ARE, respectively (odds ratio >1). Owing to the small sample size and genetic heterogeneity, no other gene or gene-set was significantly enriched for variants after correction for multiple-testing (Fig. 2). Since we observe a consistent trend in the odds ratios for the enrichment of LOF variants in several disease-associated gene-sets, we are optimistic that the availability of larger cohorts in the future can allow to identify other genes associated with RE/ARE.

Acknowledgements We thank all participants and their families for participating in this study. We thank the members of the Genomics Lab and the ERGO support team for their help in sampling the data and in creating the database. Parts of the analysis of this were conducted on the HPC facilities of the University of Luxembourg. The authors are grateful to the study participants, the staff of the Rotterdam Study and the participating general practitioners and pharmacists. The authors would like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found at http://exac.broadinstitute.org/about. This study makes use of data generated by the DECIPHER community. A full list of centres who contributed to the generation of the data is available from http://decipher.sanger.ac.uk and via email from decipher@sanger.ac.uk.

Author contributions Data analysis: DRB, DL, PM, EMR, HT, KJ, WJ, RK. Supervision of data analysis: PM, RK, HL, BAN. Study design: PN, HL, FZ, BAN, Writing the manuscript: DRB, DL, PM, BAN. Editing and revising the manuscript for scientific content: all authors.

Funding HL, BAN, PN, TS and FZ received grants by the Euro-EPINOMICS programme (DFG grant numbers: HL: LE1030/11-1, BAN: BN416/5-1, PN: NU50/8-1, TS: SA434/5-1, FWF grant number: FZ: I643-B09) within the EUROCORES framework of the European Science Foundation (ESF). DRB was supported by the JPND Courage-PD research grant. The generation and management of genomics data for the Rotterdam Study are supported by the Netherlands Organisation of Scientific Research NOW Investments (nr. 175.010.2005.011, 911-03-012) and the Netherlands Genomics Initiative (NGI)/NOW project nr. 050-060-810 (Netherlands Consortium for Healthy Ageing; NCHA). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII) and the Municipality of Rotterdam. Funding for the project was provided by the Wellcome Trust.

EuroEPINOMICS consortium contributing members: (contributing centers listed by country) Department of Neurology (Eva M. Reinthaler, Fritz Zimprich) and Department of Pediatrics and Adolescent Medicine, Medical University of Vienna, Vienna (Martha Feucht), 1090 Vienna, Austria; Private Practice for Pediatrics (Hannelore Steinböck), 1150 Vienna, Austria; St. Anna Children's Hospital, Department of Neuropediatrics (Birgit Neophytou), 1090 Vienna, Austria; Department of Pediatrics (Julia Geldner), Hospital SMZ Süd Kaiser-Franz-Josef, 1100 Vienna, Austria; Department of Pediatrics (Ursula Gruber-Sedlmayr), Medical University of Graz, 8036 Graz, Germany; Department of Pediatrics (Edda Haberlandt), Medical University of Innsbruck, 6020 Innsbruck, Austria; Department of Pediatrics (Gabriel M. Ronen), McMaster University, L8N3Z5 Hamilton, Ontario, Canada; Cologne Center for Genomics (Janine Altmüller, Dennis Lal, Peter Nürnberg, Thomas Sander, Holger Thiele) University of Cologne, 50931 Cologne, Germany; Luxembourg Centre for Systems Biomedicine (Roland Krause, Patrick May, Rudi Balling), University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg; Department of Neurology and Epileptology, Hertie Institute of Clinical Brain Research (Holger Lerche), University of Tübingen, 72076 Tübingen, Germany; Department of Neuropediatrics (Bernd A. Neubauer), University Medical Faculty Giessen and Marburg, 35385 Giessen, Germany.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval We confirm that we have read the Journals position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

References

1. Fejerman N. Atypical rolandic epilepsy. Epilepsia. 2009;50 Suppl 7:9–12.
2. Carvill GL, Regan BM, Yendle SC, O'Roak BJ, Lozovaya N, Bruneau N, et al. GRIN2A mutations cause epilepsy-aphasia spectrum disorders. Nat Genet. 2013;45:1073–6.
3. Lemke JR, Lal D, Reinthaler EM, Steiner I, Nothnagel M, Alber M, et al. Mutations in GRIN2A cause idiopathic focal epilepsy with rolandic spikes. Nat Genet. 2013;45:1067–72.
4. Lal D, Reinthaler EM, Altmüller J, Toliat MR, Thiele H, Nürnberg P, et al. RBFOX1 and RBFOX3 mutations in Rolandic epilepsy. PLoS ONE. 2013;8:e73323.
5. Lal D, Reinthaler EM, Schubert J, Muhle H, Riesch E, Kluger G, et al. DEPDC5 mutations in genetic focal epilepsies of childhood. Ann Neurol. 2014;75:788–92. May 1
6. Reinthaler EM, Dejanovic B, Lal D, Semtner M, Merkler Y, Reinhold A, et al. Rare variants in γ-aminobutyric acid type A receptor genes in rolandic epilepsy and related syndromes. Ann Neurol. 2015;77:972–86.
7. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
8. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Ercument Cicek A, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515:209–15.
9. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015;10:1556–66.
10. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5. Mar
11. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics. 2016;32:1423–6.
12. Syrbe S, Hedrich UBS, Riesch E, Djémié T, Müller S, Møller RS, et al. De novo loss- or gain-of-function mutations in KCNA2 cause epileptic encephalopathy. Nat Genet. 2015;47:393–9. Apr
13. Epi4K Consortium, Epilepsy Phenome/Genome Project. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. Lancet Neurol. 2017;16:135–43.
14. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013;9:e1003709–e1003709.
15. Pirooznia M, Wang T, Avramopoulos D, Valle D, Thomas G, Huganir RL, et al. SynaptomeDB: an ontology-based knowledgebase for synaptic genes. Bioinformatics. 2012;28:897–9.
16. Kirov G, Pocklington AJ, Holmans P, Ivanov D, Ikeda M, Ruderfer D, et al. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. Mol Psychiatry. 2012;17:142–53.
17. Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, et al. Epigenomic signatures of neuronal diversity in the mammalian brain. Neuron. 2015;86:1369–84.
18. Darnell JC, Van Driesche SJ, Zhang C, Hung KYS, Mele A, Fraser CE, et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell . 2011;146:247–61.
19. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46:944–50.
20. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet. 2009;84:524–33.
21. EpiPM Consortium. A roadmap for precision medicine in the epilepsies. Lancet Neurol. 2015;14:1219–28.

Affiliations

Dheeraj R. Bobbili[1] · Dennis Lal[2,3,4,5] · Patrick May[1] · Eva M. Reinthaler[6] · Kamel Jabbari[7] · Holger Thiele[2] · Michael Nothnagel[2] · Wiktor Jurkowski[1,8] · Martha Feucht[9] · Peter Nürnberg[2] · Holger Lerche[10] · Fritz Zimprich[6] · Roland Krause[1] · Bernd A. Neubauer[11] · Eva M. Reinthaler[6] · Fritz Zimprich[6] · Martha Feucht[12] · Hannelore Steinböck[13] · Birgit Neophytou[14] · Julia Geldner[15] · Ursula Gruber-Sedlmayr[16] · Edda Haberlandt[17] · Gabriel M. Ronen[18] · Janine Altmüller[2] · Dennis Lal[2] · Peter Nürnberg[2] · Thomas Sander[2] · Holger Thiele[2] · Roland Krause[1] · Patrick May[1] · Rudi Balling[1] · Holger Lerche[10] · Bernd A. Neubauer[11]
EUROEPINOMICS COGIE Consortium

[1] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

[2] Cologne Center for Genomics, University of Cologne, Cologne, Germany

[3] Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[4] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[5] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[6] Department of Neurology, Medical University of Vienna, Vienna, Austria

[7] Cologne Biocenter, Institute for Genetics, University of Cologne, Cologne, Germany

[8] The Genome Analysis Centre, Norwich, UK

[9] Department of Pediatrics, Medical University of Vienna, Vienna, Austria

[10] Department of Neurology and Epileptology, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany

[11] Department of Neuropediatrics, Medical Faculty University Giessen, Giessen, Germany

[12] Department of Pediatrics and Adolescent Medicine, Medical University of Vienna, 1090 Vienna, Austria

[13] Private Practice for Pediatrics, St. Anna Children's Hospital, 1150 Vienna, Austria

[14] Department of Neuropediatrics, 1090 Vienna, Austria

[15] Department of Pediatrics, Hospital SMZ Süd Kaiser-Franz-Josef, 1100 Vienna, Austria

[16] Department of Pediatrics, Medical University of Graz, 8036 Graz, Austria

[17] Department of Pediatrics, Medical University of Innsbruck, 6020 Innsbruck, Austria

[18] Department of Pediatrics, McMaster University, Hamilton L8N3Z5 ON, Canada

doi:10.1093/brain/awx202 BRAIN 2017: 140; 2444–2459 | 2444

BRAIN
A JOURNAL OF NEUROLOGY

Metformin reverses TRAP1 mutation-associated alterations in mitochondrial function in Parkinson's disease

Julia C. Fitzgerald,[1] Alexander Zimprich,[2] Daniel A. Carvajal Berrio,[3] Kevin M. Schindler,[1,4] Brigitte Maurer,[1] Claudia Schulte,[1] Christine Bus,[1] Anne-Kathrin Hauser,[1] Manuela Kübler,[1] Rahel Lewin,[1] Dheeraj Reddy Bobbili,[5] Lisa M. Schwarz,[1,6] Evangelia Vartholomaiou,[7] Kathrin Brockmann,[1] Richard Wüst,[1,8] Johannes Madlung,[9] Alfred Nordheim,[10] Olaf Riess,[11] L. Miguel Martins,[12] Enrico Glaab,[5] Patrick May,[5] Katja Schenke-Layland,[3,13,14] Didier Picard,[7] Manu Sharma,[15] Thomas Gasser[1] and Rejko Krüger[1,5,16]

The mitochondrial proteins TRAP1 and HTRA2 have previously been shown to be phosphorylated in the presence of the Parkinson's disease kinase PINK1 but the downstream signalling is unknown. HTRA2 and PINK1 loss of function causes parkinsonism in humans and animals. Here, we identified TRAP1 as an interactor of HTRA2 using an unbiased mass spectrometry approach. In our human cell models, TRAP1 overexpression is protective, rescuing HTRA2 and PINK1-associated mitochondrial dysfunction and suggesting that TRAP1 acts downstream of HTRA2 and PINK1. HTRA2 regulates TRAP1 protein levels, but TRAP1 is not a direct target of HTRA2 protease activity. Following genetic screening of Parkinson's disease patients and healthy controls, we also report the first *TRAP1* mutation leading to complete loss of functional protein in a patient with late onset Parkinson's disease. Analysis of fibroblasts derived from the patient reveal that oxygen consumption, ATP output and reactive oxygen species are increased compared to healthy individuals. This is coupled with an increased pool of free NADH, increased mitochondrial biogenesis, triggering of the mitochondrial unfolded protein response, loss of mitochondrial membrane potential and sensitivity to mitochondrial removal and apoptosis. These data highlight the role of TRAP1 in the regulation of energy metabolism and mitochondrial quality control. Interestingly, the diabetes drug metformin reverses mutation-associated alterations on energy metabolism, mitochondrial biogenesis and restores mitochondrial membrane potential. In summary, our data show that TRAP1 acts downstream of PINK1 and HTRA2 for mitochondrial fine tuning, whereas TRAP1 loss of function leads to reduced control of energy metabolism, ultimately impacting mitochondrial membrane potential. These findings offer new insight into mitochondrial pathologies in Parkinson's disease and provide new prospects for targeted therapies.

1 Department of Neurodegenerative Diseases, Center of Neurology and Hertie-Institute for Clinical Brain Research, University of Tübingen and German Centre for Neurodegenerative Diseases, Tübingen, Germany
2 Medical University Vienna, Department of Neurology, Vienna, Austria
3 Department of Women's Health, Research Institute for Women's Health, University of Tübingen, Tübingen, Germany
4 University of Tübingen, Interfaculty Institute of Biochemistry, Tübingen, Germany
5 Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg
6 Graduate Training Centre of Neuroscience, International Max Planck Research School, Tübingen, Germany
7 University of Geneva, Department of Cell Biology, Geneva, Switzerland
8 Department of Psychiatry and Psychotherapie, University Hospital Tübingen, Germany
9 University of Tübingen, Interfaculty Institute for Cell Biology, Proteome Center Tübingen, Tübingen, Germany
10 University of Tübingen, Interfaculty Institute of Cell Biology, Unit of Molecular Biology, Tübingen, Germany
11 University of Tübingen, Institute of Medical Genetics and Applied Genomics, Tübingen, Germany

---

Received April 18, 2017. Revised June 13, 2017. Accepted July 4, 2017.
© The Author (2017). Published by Oxford University Press on behalf of the Guarantors of Brain. All rights reserved.
For Permissions, please email: journals.permissions@oup.com

12 MRC Toxicology Unit, University of Leicester, Leicester, UK
13 Department of Cell and Tissue Engineering, Fraunhofer Institute for Interfacial Engineering and Biotechnology IGB Stuttgart, Germany
14 Department of Medicine/ Cardiology, CVRL, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA
15 Centre for Genetic Epidemiology, Institute for Clinical Epidemiology and Applied Biometry, University of Tübingen, Germany
16 Parkinson Research Clinic, Centre Hospitalier de Luxembourg (CHL), Luxembourg

Correspondence to: Julia Fitzgerald
Department of Neurodegeneration, Hertie Institute for Clinical Brain Research and DZNE, Otfried Müller Strasse 27, Tübingen, 72076, Germany
E-mail: Julia.fitzgerald@uni-tuebingen.de

Keywords: mitochondrial diseases; Parkinson's disease; metabolic disease; neuroprotection; experimental models

Abbreviations: mtUPR = mitochondrial unfolded protein response; OCR = oxygen consumption rate; PPMI = Parkinson's Progression Markers Initiative

Introduction

Parkinson's disease is an aetiologically heterogeneous syndrome caused by a combination of genetic and environmental risk factors. At least 85% of cases are sporadic, and at present, there are only symptomatic treatments available, but the advancement of genetic testing and identification of patient endophenotypes has given hope for the emerging field of individualized medicine. Mitochondrial dysfunction, ensuing cellular energy failure and oxidative stress may be one important disease pathway in a subgroup of Parkinson's disease patients (Kruger *et al*., 2017). The aim is that these patients can be therapeutically targeted or serve as an entry point for precision medicine.

TRAP1 (tumour necrosis factor type 1 receptor associated protein, also known as HSP 75) is a chaperone that resides in the mitochondrial matrix (Altieri *et al*., 2012). It has a regulatory role in stress sensing in mitochondria allowing cellular adaption to the environment. TRAP1 is recognized as a potential effector protein in Parkinson's disease signalling, since it was found to be phosphorylated by the Parkinson's disease kinase PINK1 (Pridgeon *et al*., 2007). Loss-of-function mutations in *PINK1* and *PARK2* (encoding parkin) cause familial Parkinson's disease (Kitada *et al*., 1998; Valente *et al*., 2004) and impair the elimination of damaged mitochondria (Geisler *et al*., 2010; Narendra *et al*., 2010). However, beyond mitophagy, there is relatively little known about the mitochondrial quality control pathways in Parkinson's disease.

Chaperones and proteases maintain mitochondrial proteostasis. Tight control of protein quality and turnover inside mitochondria is essential for the function of electron transport complexes, which provide energy through oxidative phosphorylation. PINK1 has previously been shown to be required for the phosphorylation of the mitochondrial protease and Parkinson's disease-associated protein HTRA2 (Plun-Favreau *et al*., 2007). Here we highlight a signalling pathway involving PINK1, HTRA2 and TRAP1, where TRAP1 is the effector modulating mitochondrial chaperone activities and metabolic homeostasis.

The hypothesis that TRAP1 is an important downstream effector in mitochondrial signalling is underscored by reports that TRAP1 rescues mitochondrial dysfunction in neuronal models where PINK1 is silenced (Costa *et al*., 2013; Zhang *et al*., 2013). TRAP1 also protects cells from oxidative toxicity caused by respiratory complex I inhibition via an α-synuclein variant known to induce a genetic form of Parkinson's disease (Butler *et al*., 2012). TRAP1 protects mitochondria via its chaperone function (Altieri *et al*., 2012; Rasola *et al*., 2014) and by reducing reactive oxygen species (Masuda *et al*., 2004; Hua *et al*., 2007; Im *et al*., 2007).

TRAP1 also acts as a metabolic switch controlling the tumour cell's preference for aerobic glycolysis (Yoshida *et al*., 2013). ERK1/2 orchestrates the phosphorylation of TRAP1 controlling the metabolic switch (Masgras *et al*., 2017), which is reportedly via TRAP1 inhibition of succinate dehydrogenase (Sciacovelli *et al*., 2013; Masgras *et al*., 2017), although this remains controversial (Rasola *et al*., 2014). TRAP1 deficiency promotes mitochondrial respiration, accumulation of tricarboxylic acid cycle intermediates, ATP and reactive oxygen species (Yoshida *et al*., 2013). *TRAP1* deletion in mice does not affect viability and delays the appearance of tumours in a breast cancer model (Vartholomaiou *et al*., 2017).

Therefore, the identification of TRAP1 as a novel HTRA2 interactor prompted us to further explore the PINK1-HTRA2-TRAP1 pathway related to neurodegeneration in Parkinson's disease. Here we show that TRAP1 takes an important role as downstream effector in this pathway and therefore provides an interface between Parkinson's disease and energy metabolism.

Materials and methods

Cell culture

Fibroblast culture from skin biopsies has been previously described by our laboratory (Burbulla and Kruger, 2012). All biopsies and DNA samples were obtained with patient's consent and approval of the local ethics committee and according to the Declaration of Helsinki. HeLa, SH-SY5Y, HEK293 cell culture has been described previously (Burbulla *et al.*, 2014). TRAP1 knockout mouse adult fibroblasts and HTRA2 knockout mouse embryonic fibroblasts have been described by Kieper *et al.* (2010) and Yoshida *et al.*, (2013), respectively. Human induced pluripotent stem cells from a PINK1 knockout line generated in our laboratory and its isogenic control were used to generate small molecule neuronal precursor cells (smNPCs) according to Reinhardt *et al.* (2013). smNPCs were cultivated in 1:1 Dulbecco's modified Eagle medium (DMEM)/Ham's F12 (Biochrom, Harvard Bioscience) and Neurobasal® (Gibco, Thermo Fisher) medium supplemented with 1% Pen/Strep, 1% GlutaMAX™ (Gibco, Thermo Fisher), B-27 Supplement (Gibco, Thermo Fisher), N2 (Gibco, Thermo Fisher), 200 μM ascorbic acid (Sigma-Aldrich), 3 μM CHIR 99021 (Axon Medchem) and 0.5 μM purmorphamine (Calbiochem, Merck Millipore) on Matrigel® (Corning) coated cell culture dishes.

DNA constructs and RNAs

Human *TRAP1* cDNA was cloned into the pIRES vector (Clontech, Takara). GST-coupled wild-type *HTRA2*, A141S *HTRA2*, and G399S *HTRA2* have been previously described (Martins, 2002). Cloning of wild-type *HTRA2* and S306A *HTRA2* cDNA into the pcDNA3.1 vector was previously described (Strauss *et al.*, 2005). Short interfering (si)RNAs targeting *HTRA2* were purchased from Sigma Aldrich (Fitzgerald *et al.*, 2012) and targeting *TRAP1* and non-targeting controls from Dharmacon (siGENOME SMARTpool #D001206-13-05 POOL#1, non-targeting siGENOME SMARTpool).

Mass spectrometry

We used recombinant, mature GST-HTRA2 (wild-type HTRA2, HTRA2-A141S, and HTRA2-G399S) as baits and lysates from SH-SY5Y cells. The supernatant contained the fusion proteins that were then bound to glutathione agarose (Molecular Probes, Thermo Scientific) and eluted with imidazole. Analyses were performed on 1D gel pieces of the eluates. The measurements of the peptides derived from tryptic in-gel digest were performed using a nano-HPLC-ESI-MS/MS system [Ultimate (LC Packings/Dionex, Germany)/QStar Pulsar i (Applied Biosystems/Sciex)], described by Sauer *et al.* (2006). Mass spectrometry data were processed against the National Center for Biotechnology Information (NCBI) protein sequence database with the search engine MASCOT (Matrix Science, UK) (Perkins *et al.*, 1999).

Co-immunoprecipitation

HeLa and HEK293 cell lysates were prepared using a lysis buffer [1 % (v/v) Triton™ X-100, 1 × protease inhibitor cocktail (Roche Complete, Roche), 1 × phosphatase inhibitor (Roche PhosStop, Roche)] and the nuclear material removed following homogenization. Where wild-type HTRA2 was overexpressed, HTRA2 was transiently transfected (48 h) using Effectine transfection reagent (Qiagen, according to the manufacturer's instructions). Mitochondrial enrichment was previously described (Fitzgerald *et al.*, 2012). Brain tissue from TRAP1 knockout mice previously described (Vartholomaiou *et al.*, 2017) was prepared by separating the cortices from the basal ganglia (mid-brain) and cerebellum/brainstem (hind-brain) on ice. Brain tissue lysates were prepared according to Casadei *et al.* (2016). Immunoprecipitation was carried out using HTRA2 (R and D Biosciences) or TRAP1 (BD biosciences) antibodies or bovine IgG coupled to protein A Sepharose beads (Sigma Aldrich P9424), according to Fitzgerald *et al.* (2012).

SDS-PAGE and western blotting

Cell lysates were prepared as described for co-immunoprecipitation and proteins electrophoresed on acrylamide gels and transferred to membranes, as previously described (Fitzgerald *et al.*, 2012). Brain tissue lysates from non-transgenic and HTRA2 overexpressing mice previously characterized and described (Casadei *et al.*, 2016) were prepared from whole brain and the total extracts (nuclear material removed) were prepared according to Casadei and colleagues (2016). Total protein stain (copper pthalocyanine-3, 4', 4' 4'-tetra-sulphonic acid tetra sodium salt in 12 mM HCl) and destain (12 mM NaCl). Antibodies against TRAP1 (BD Biosciences), β-actin (Sigma Aldrich), GAPDH (Invitrogen, Thermo Scientific) Tom20 (Santa Cruz Biotechnology), Hsp60 (Bio-Rad), α-tubulin (Sigma Aldrich), rodent OXPHOS (#MS604 Mitosciences, AbCam), Hsp70 (Santa Cruz Biotechnology), Hsp90 (BD Biosciences), Human Total OXPHOS (all nuclear encoded subunits from Mitosciences, Abcam), ERK1/2 and P-ERK1/2 (Cell Signaling Technolgy) and mitobiogenesis antibody (containing SDH, GAPDH and COX, Abcam) were used. Secondary antibodies were purchased from GE Healthcare. Densitometry from western blot was carried out using the ImageJ 1.41o software (Wayne Rasband; National Institutes of Health, USA).

Live cell imaging

Mitochondrial morphology, mass and colocalization studies were visualized using 100 nM MitoTracker® Green FM (Thermo Scientific), lysosomes by 100 nM Lyostracker® Red DND-99 (Thermo Scientific) as previously described (Burbulla *et al.*, 2014). Analyses were performed as previously described (Burbulla *et al.*, 2014). The series of images were saved uncompressed and analysed with AxioVision software (Zeiss) and ImageJ 1.41o software.

Fluorescence-activated cell sorting

Cells were trypsinized and centrifuged at 300*g* for 5 min and the cells incubated in dye, buffer only or dye plus a control. For early apoptosis, Annexin V-Pacific Blue™ in Annexin V binding buffer (both from BioLegend) or Annexin V-Pacific Blue™ plus staurosporine was used. For mitochondrial membrane potential, 200 nM tetramethylrhodamine, ethyl ester, perchlorate (TMRE, from Thermo Scientific) in Hanks buffer or TMRE plus carbonyl cyanide-p-trifluoromethoxyphenylhydrazone (CCCP) 10 μM was used. For mitochondrial reactive oxygen species, 2 μM MitoSox™ (Thermo Scientific) in Hanks buffer or MitoSox™ plus 10 μM rotenone was used. Cells were sorted using a MACSQuant® automated flow cytometer (Mitenyi Biotechnology) according to their mean average fluorescence signal. All mean average fluorescence values were divided by the background fluorescence in the same channel in the same unstained cells to account for autofluorescence.

Live cell kinetic measurement of mitochondrial membrane potential

Cells were seeded in Ibidi® dishes and the media exchanged for Hank's balanced salt solution (HBSS) containing 200 nM TMRE stain (Thermo Scientific) for 15 min at 37°C with CO_2. The TMRE was removed and replaced with 360 µl Hanks buffer. The cells were imaged using a Zeiss inverted confocal microscope at excitation HeNe1, 543 nm and emission LP 560 nm and brightfield for 20 × 4 s cycles. Followed by the addition of 360 µl (0.25 mg/ml oligomycin), measured for 20 × 4 s cycles, 180 µl (10 µM rotenone), measured for 20 × 4 s cycles and 100 µl (10 µM FCCP) and measured for 20–40 × 4 s cycles. Using ImageJ, each transfected cell (detected using ZsGreen-TRAP1) in each frame was analysed for TMRE fluorescence intensity, mean fluorescence and total area. The corrected total cell fluorescence (CTCF) over time was calculated using the formula: CTCF = fluorescence intensity – (cell area × mean background fluorescence).

Genetic screening by high resolution melting analysis

Both polymerase chain reaction (PCR) and high resolution melting analysis were performed in the presence of a saturating DNA binding dye. Mutations were detectable because heterozygote DNA forms heteroduplices that begin to separate in single strands at a lower temperature and with a different curve shape than homozygote DNA, as described previously (Wust *et al*., 2016).

Whole exome sequencing and consanguinity analysis

Whole exome data were generated from 200 Parkinson's disease patients from Vienna ($n = 100$) and Tübingen ($n = 100$). Genomic DNA (3 µg) was fragmented into ~250 bp fragments, end-repaired, adaptor-ligated and sample index barcodes were included. Pooled libraries were enriched with SureSelect Human All Exon 50 Mb kit (AgilentTechnologies) to capture 50 Mb of exonic and flanking intronic regions. Sequencing of post-enrichment libraries was carried out on the Illumina HiSeq 2000 sequencing instrument (Illumina) as 2 × 100 bp paired-end runs. On average, this yielded ~10 Gb of mapped sequences and a $>100\times$ average coverage for 90% of the targeted sequence per individual. Raw image files were processed by the Illumina pipeline. Reads were aligned to the human reference genome hg19 with the Burrows-Wheeler Aligner. SAM tools were used to identify single nucleotide variants and small insertions and deletions. Patients were screened for consanguinity using an implemented algorithm of an analysing tool of the Helmholtz Zentrum, München. Patients with homozygous regions encompassing a total of more than 20 Mb were considered as likely consanguineous. Particularly, stretches of >2 Mb were surveyed for rare homozygous variants (missense, nonsense, frameshift and splice-site). Variants were further filtered for a minor allele frequency smaller than 1% in the in-house dataset of ~10 000 control exomes from patients with other unrelated diseases and exomes and in public available databases (ExAC database and 1000 Genomes).

Computational analysis of *TRAP1* genetic variants

Computational analysis of *TRAP1* variants can be found in the Supplementary material.

Quantitative RT-PCR

Quantitative PCR reactions were performed using FastStart SYBR® green Master mix (Roche) to amplify 1 µl of the 1:10 diluted cDNA using 5 µm of each primer h_TRAP1 5' UTR Forward: TTCCCATCGTGTACGGTCCCGC, h_TRAP1 Exon2 Reverse: GGCCCAACTGGGCTGTGGTCC, h_TRAP1 Spanning Exon2-3 Reverse: TGTTTGGAAGTGGAACCCTGC. Housekeeping gene *GAPDH* primers: Forward: CCATCACCATCTTCCAGGAGCGA, Reverse: GGATGACCTTGCCCACAGCCTTG. Standard curves of each amplified gene were created to calculate the PCR efficiency and relative expression using the efficiency corrected delta–delta Ct method. RNA was prepared from human fibroblasts using Qiashredder® and RNAEasy® preparation kits (Qiagen). RNA (1 µg) was reverse transcribed to cDNA using QuantiTECT® (Qiagen).

Oxygen consumption and extracellular acidification rate

Oxygen consumption rates (OCR) were measured in whole cells using a Seahorse™ XF96 Extracellular Flux Analyzer (Agilent) according to Rogers *et al*. (2011). The concentrations of mitochondrial toxins used were optimized by titration in human fibroblasts according to the manufacturer's recommendations. The final concentration of all toxins used was 1 µM and the volume of the toxin injected in each port was sequentially increased by several microlitres to maintain the correct final concentration. Human fibroblasts were plated in Seahorse™ XF96 well plates 24 h prior to measuring at a density of ~15 000 cells per well. The OCR for each well was corrected for cell number. Stained nuclei were counted using high content image capture and analysis using the BD Pathway 855 (BD Biosciences). Extracellular acidification rates (ECAR) from the same experiments provide an indication of glycolytic activity and were normalized to OCR/cell to account for the cell numbers in each well in each experiment.

Complex I activity

Following isolation of crude mitochondria from approximately five million cells, described previously by Burte *et al*. (2011), complex I activity was measured according to Hargreaves and colleagues (2007). The activity of complex I was normalized to citrate synthase activity, also according to Hargreaves *et al*. (2007) and data expressed as a ratio of complex I/citrate synthase.

ATP

The concentration of ATP per microgram of total protein was measured using the ATPLite^{TM} Kit from Perkin Elmer. ATP standards are used to determine the concentration of ATP in a cell lysate replicated a minimum of three times in each experiment. Concentration (μM) of ATP is expressed per microgram of total protein in each well as measured by protein assay (Bio-Rad).

Fluorescence lifetime imaging microscopy

A detailed description of the fluorescence lifetime imaging microscopy (FLIM) method can be found in the Supplementary material and is described in Lakner *et al.* (2017).

Measurement of NAD$^+$/NADH levels

NAD$^+$ and total NAD$^+$ and NADH levels were measured using a fluorometric assay kit (Abcam). The levels of NAD$^+$ and NADH were quantified using standards and normalized to total protein in each sample according to protein assay (Bio-Rad).

Statistics

Analyses of statistical significance were performed using GraphPad Prism 6.0 and the relevant statistical test. The statistical test used and the *P*-values are shown in the figure legends. All cell culture experiments [including all imaging and fluorescence-activated cell sorting (FACS) experiments] were performed a minimum of three times, using a different cell passage and on different days. In the genetic studies, the initial screening by high temperature melt analysis was performed on 280 German Parkinson's disease patients and a group of 192 healthy individuals. The exome sequencing was performed on the DNA from 200 Parkinson's disease patients collected in Tübingen, Germany and Vienna, Austria.

Results

TRAP1 interacts with HTRA2

We have previously reported loss-of-function mutations in *HTRA2* in Parkinson's disease patients and therefore performed unbiased mass spectrometry on GST-HTRA2 baited SH-SY5Y lysates to identify novel interaction proteins (Fig. 1A). We identified TRAP1 as an interactor of HTRA2 with the relevant controls.

To confirm the physical interaction, HTRA2 immunoprecipitations were performed in HeLa cells overexpressing HTRA2 or not. Immunoblotting revealed the presence of TRAP1 in the HTRA2 immunoprecipitation (endogenous and overexpressed HTRA2) but not in the IgG control (Fig. 1B), enriched in the mitochondrial fraction. Knockdown of TRAP1 using siRNA reduced the amount of TRAP1 interacting with HTRA2, confirming the specificity of the immunoprecipitation (Fig. 1C). The interaction between HTRA2 and TRAP1 occurs in mouse brain (cortex, midbrain and hindbrain) as demonstrated by the immunoprecipitation of TRAP1 with HTRA2 in extracts from wild-type mice and not TRAP1 knockout mice (Fig. 1D).

To investigate the relevance of the HTRA2-TRAP1 interaction we monitored the amount of TRAP1 immunoprecipitated with HTRA2 under several stress conditions. We found that acute treatment with the mitochondrial toxins rotenone and antimycin A abolished the interaction in human HEK293 cells and this was due to reduced TRAP1 and not a global reduction of total protein (Fig. 1E). We then assessed the influence of several other stressors, this time the concentrations of the toxins were titrated for HeLa cells and for the cells to survive a chronic treatment over a 24 h period. We found that dopamine treatment had no effect on the interaction of HTRA2 and TRAP1, whereas, the TRAP1 inhibitor 17-AAG, hydrogen peroxide, the ionophore valinomycin and mitochondrial respiratory inhibitors oligomycin, antimycin A, and rotenone all largely reduced or abolished the interaction (Fig. 1F). These data from two different human cell lines suggest that the interaction of HTRA2 and TRAP1 serves the mitochondria under normal physiological conditions, under starvation and dopamine toxicity, but not respiratory inhibition.

TRAP1 rescues HTRA2 and PINK1 loss-of-function phenotypes but is not a proteolytic substrate of HTRA2

We hypothesized that HTRA2 interacted with TRAP1 to degrade it since HTRA2 is a key mitochondrial protease and the levels of TRAP1 appear to be a key factor in mitochondrial control (Kang *et al.*, 2007; Zhang *et al.*, 2015; Amoroso *et al.*, 2016; Lv *et al.*, 2016). Using PhosTag^{TM} SDS-PAGE, we found a significant increase in the levels of phosphorylated and non-phosphorylated TRAP1 when we immunoprecipitated endogenous TRAP1 in the absence of HTRA2 (Fig. 2A). We also found that stimulation of PINK1 kinase with the ionophore valinomycin (at concentrations known to induce accumulation of PINK1) (Rakovic *et al.*, 2013), increased the amount of phosphorylated TRAP1 in wild-type HTRA2 mouse adult fibroblasts (Fig. 2A). Phosphorylated TRAP1 levels were increased to the same extent in HTRA2 knockout mouse adult fibroblasts, whether treated with valinomycin or not (Fig. 2A). However, there was no significant effect of PINK1 knockout on TRAP1 phosphorylation status in neuronal progenitor cells (Fig. 2A).

Overexpression of wild-type HTRA2 in human cells from four independent experiments (Fig. 2B) or in mice (Fig. 2C) results in reduced TRAP1 protein levels. However, overexpression of a protease dead form of HTRA2 (S306A), which is catalytically inactive but still targeted to the mitochondria (Martins *et al.*, 2002) in human cells has the same effect on TRAP1 levels as the wild-type, indicating that the protease activity of HTRA2 is not important for the interaction between TRAP1 and HTRA2 (Fig. 2D).

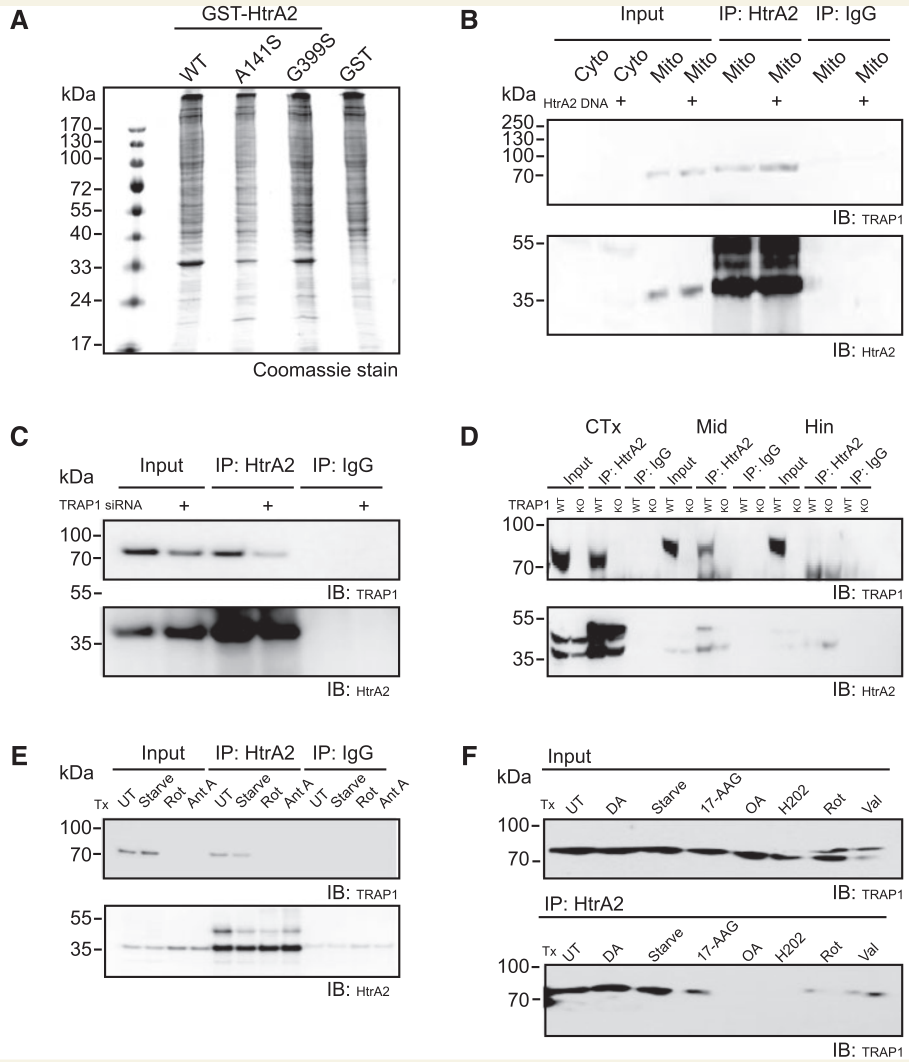


Figure 1 TRAP1 interacts with HTRA2. (**A**) A Coomassie-stained gel of GST-HTRA2 eluates from SH-SY5Y cells for unbiased mass spectrometry. (**B**) Immunoblot (IB) of TRAP1 and HTRA2 in cytosolic (Cyto) mitochondrial (Mito) fractions from HeLa cells overexpressing wild-type (WT) HTRA2 or an empty vector control. Input lysates (input), HTRA2 immunoprecipitates (IP HtrA2) and control immunoprecipitates using bovine IgG (IP IgG). (**C**) Immunoblot of TRAP1 and HTRA2 in lysates in HeLa cells transfected with TRAP1 siRNA or a non-targeting siRNA control. Input lysates (input), HTRA2 immunoprecipitates (IP HtrA2) and control immunoprecipitates using bovine IgG (IP IgG). (**D**) Immunoblot of TRAP1 and HTRA2 in wild-type and TRAP1 knockout mouse brain lysates from cortex (CTx), basal ganglia/midbrain (Mid) and hindbrain (Hin). Input lysates (input), HTRA2 immunoprecipitates (IP HtrA2) and control immunoprecipitates using bovine IgG (IP IgG). (**E**) Immunoblot of TRAP1 and HTRA2 in total cell lysates from untreated HEK293 cells (UT) or HEK293 cells treated with serum-free medium (starve), 1 µM rotenone (Rot) and 25 nM antimycin A (Ant A) for 24 h. Input lysates (input), HTRA2 immunoprecipitates (IP HtrA2) and control immunoprecipitates using bovine IgG (IP IgG). (**F**) Immunoblot of TRAP1 in HeLa cell extracts either untreated (UT) or treated with 200 µM dopamine (DA), serum-free media (starve), 2 µM Hsp90/TRAP1 inhibitor (17-AAG), 1 µM oligomycin and 0.4 µM antimycin A (OA), 40 µM hydrogen peroxide (H_2O_2), 5 µM rotenone (Rot) or 100 nM valinomycin (Val) for 24 h. Input lysates (input) and HTRA2 immunoprecipitates (IP HtrA2) are shown.

The HTRA2-TRAP1 interaction is not a protease-substrate interaction, yet TRAP1 is likely downstream of HTRA2 since the overexpression of TRAP1 rescues the HTRA2 knock-down-induced loss of mitochondrial membrane potential (Fig. 2E), reduced basal oxygen consumption (Fig. 2F), increased mitochondrial reactive oxygen species (Fig. 2G) and sensitivity towards serum starvation-induced apoptosis (Fig. 2H). TRAP1 overexpression also rescues the reduced mitochondrial membrane potential observed in PINK1-deficient neuroprogenitor cells measured over a time course inclusive of mitochondrial toxin controls (Fig. 2I).

TRAP1 loss-of-function in Parkinson's disease

Mutations in PINK1 cause early onset Parkinson's disease (Valente *et al.*, 2004) and HTRA2 risk variants have been reported in German (Strauss *et al.*, 2005) and Belgian

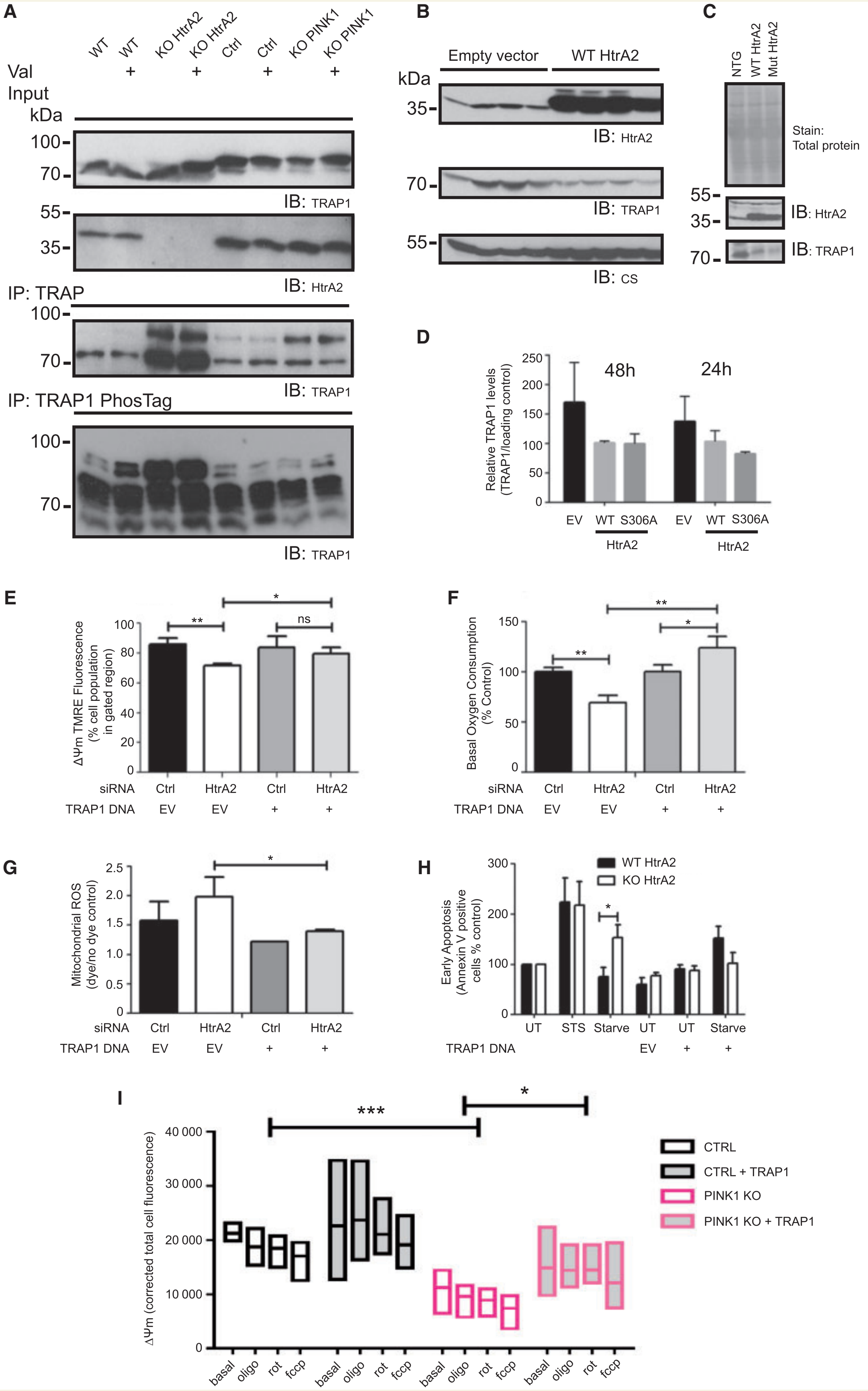


Figure 2 TRAP1 rescues HTRA2 and PINK1 loss of function phenotypes but is not a proteolytic substrate of HTRA2. (**A**) Immunoprecipitation (IP) of TRAP1 from wild-type (WT) and knockout (KO) HTRA2 mouse embryonic fibroblasts and PINK1 knockout and isogenic control (Ctrl) neuroprogenitor cells treated with or without 1 µM valinomycin (Val) for 24 h to activate PINK1. Input lysates (input) and TRAP1 immunoprecipitates were run on SDS-PAGE (**A–C**) and PhosTag^{TM} SDS-PAGE gels (IP: TRAP1 PhosTag) to visualize all phosphorylated TRAP1 enriched by the TRAP1 pulldown. (**B**) Immunoblot (IB) of TRAP1, HTRA2 and mitochondrial marker citrate synthase (CS) in SH-SY5Y extracts from four experiments where wild-type HTRA2 or an empty vector (EV) control is overexpressed. (**C**) Immunoblot of TRAP1 and HTRA2 in brain extracts from non-transgenic (NTG), overexpressing wild-type HTRA2 and overexpressing G399S mutant HTRA2 (Mut HtrA2) mice. (**D**) Quantification of TRAP1 protein levels (normalized to GAPDH loading control) at 24 h and 48 h after transfection with wild-type

(continued)

(Bogaerts *et al.*, 2008) Parkinson's disease cohorts. Therefore, we used the exome sequencing data from the Parkinson's Progression Markers Initiative (PPMI), to determine all non-synonymous *TRAP1* single nucleotide variants with a minor allele frequency $<1\%$ in the European non-Finnish population (ExAC) (Lek *et al.*, 2016) and an odds ratio >1 in Parkinson's disease patients compared to controls. Variants (Supplementary Fig. 1C) were further filtered by selecting those 'damaging', 'probably damaging' or 'likely damaging' and where crystal structures are available. All three resulting variants are located within known functional domains: S221P falls within a histidine kinase-like ATPase domain (HATPase_c, InterPro: IPR003594), and H311Q and R469C are positioned within an HSP90 domain (Pfam database: PF00183). However, none of these variants overlapped with known ubiquitination, acetylation or phosphorylation sites (Supplementary Fig. 1A). The residues for variants S221P and R469C are largely buried (5% and 13% solvent accessibility), whereas variant H311Q affects a residue that is partially accessible (25%) and could alter protein-protein interactions via residue size and charge alterations. In a multiple sequence alignment, high sequence conservation was observed for residues H311 and R469, but not for S211 (Supplementary Fig. 1B). Correspondingly, a destabilizing effect was predicted by the majority of algorithms for H331Q and R469C, while the S221P variant was estimated to be neutral. Notably, R469C was also predicted to decrease the chaperone binding function of TRAP1 (LIMBO software).

We used high resolution melting to screen for sequence variations in the *TRAP1* gene in the genomic DNA from German Parkinson's disease patients and a group of healthy individuals. We detected several genetic variants, further identified as single nucleotide polymorphisms (listed in Supplementary Fig. 1C). Burden analysis of *TRAP1* was performed using the PPMI dataset (summarized in Supplementary Fig. 1D). Truncating variants predicted to cause loss-of-function of *TRAP1* are very rare and were only observed in Parkinson's disease patients and not in controls. Interestingly, rare missense *TRAP1* mutations were found to have significantly different burden (P-values < 0.05) between patients and controls (Supplementary Fig. 1D). In parallel, we analysed 200 exomes of Parkinson's disease patients from Austria and Germany for consanguinity. In addition to TRAP1 variants (listed in Supplementary Fig. 1C), we found a moderate, but significant consanguinity of ~20 Mb in a German Parkinson's disease patient. One homozygous stretch encompasses 5 Mb at Chr.16, including the *TRAP1* gene. Here we found a homozygous c.C158 > T (R47X) mutation (Fig. 3A). This mutation is not present in 10 000 control exomes of the Helmholtz database; however, it occurs 12 times heterozygously in the ExAC database (60 000 controls). The R47X TRAP1 Parkinson's disease patient has no rare variant in any of the other established Parkinson's disease genes (see Supplementary material for the full list of genes).

The homozygous p.Arg47Ter single nucleotide exchange (R47X) in exon 2 of *TRAP1* leads to a premature stop codon and truncation at the transit sequence of *TRAP1* in a late-onset Parkinson's disease patient (Fig. 3A). A TRAP1 antibody that binds at a region of TRAP1 encompassing amino acids 253–464 (shown in Fig. 3B) was, as expected, unable to detect TRAP1 protein in fibroblasts biopsied from the R47X patient (Fig. 3C). Using PCR primers upstream and downstream of the mutation to amplify patient cDNA, we found that *TRAP1* RNA is present, suggesting no nonsense-mediated RNA decay (Fig. 3D). The R47X TRAP1 patient was diagnosed with Parkinson's disease at age 70 years. There is no family history of Parkinson's disease but the mother of the index patient had dementia. The R47X patient has also been diagnosed with dilated cardiomyopathy, chronic pancreatitis, polyneuropathy and chronic kidney insufficiency (Table 1).

TRAP1 R47X Parkinson's disease patient mitochondria meet ATP demand but have reduced membrane potential

To understand the relevance of the R47X TRAP1 mutation, we assessed several readouts of mitochondrial form and function in patient-derived fibroblasts. There were no obvious differences in mitochondrial morphology between controls and the index patient under basal or serum starvation conditions (binary *z*-stack images shown in Fig. 4A).

Figure 2 Continued

HTRA2 or HTRA2 protease dead mutant (S306A). (**E**) Mitochondrial membrane potential ($\Delta\Psi$m) in HeLa cells transfected with HTRA2 siRNA (HtrA2) or a non-targeting control (Ctrl) and overexpressing an empty vector or wild-type TRAP1 DNA construct. (**F**) Basal oxygen consumption in HeLa cells transfected with HTRA2 siRNA (HtrA2) or a non-targeting control (Ctrl) and overexpressing an empty vector or wild-type TRAP1 DNA construct. (**G**) Mitochondrial reactive oxygen species (ROS) using MitoSox^{TM} in HeLa cells transfected with HTRA2 siRNA (HtrA2) or a non-targeting control (Ctrl) and overexpressing an empty vector or wild-type TRAP1 DNA construct. (**H**) Early apoptosis measured by annexin V in HeLa cells either untreated (UT) or treated with 1 μM staurosporine (STS), serum-free media (starve) for 24 h and overexpressing an empty vector or wild-type TRAP1 DNA (+) construct. (**I**) Reduced $\Delta\Psi$m in PINK1 knockout neuroprogenitor cells is rescued by overexpression of wild-type TRAP1. TRAP1 (or empty vector control) transfected neuroprogenitor cells were identified using a ZsGreen tag. Confocal images were taken every 4 s following incubation with TMRE, followed by washing (basal), oligomycin (oligo), rotenone (rot) and FCCP (fccp). All statistical tests were the Student's *t*-test assuming different standard deviation, except 2H, where two-way ANNOVA was used to compare groups and condition. Error bars show standard deviation and $^{*}P < 0.05$; $^{**}P < 0.01$. TMRE = tetramethylrhodamine, ethyl ester, perchlorate.

Computational analysis of *z*-stack images revealed no difference in average mitochondrial size (Fig. 4B) or mitochondrial branching (Fig. 4C). However, after serum starvation there was a significant fragmentation of mitochondria in the R47X fibroblasts compared to controls (Fig. 4D).

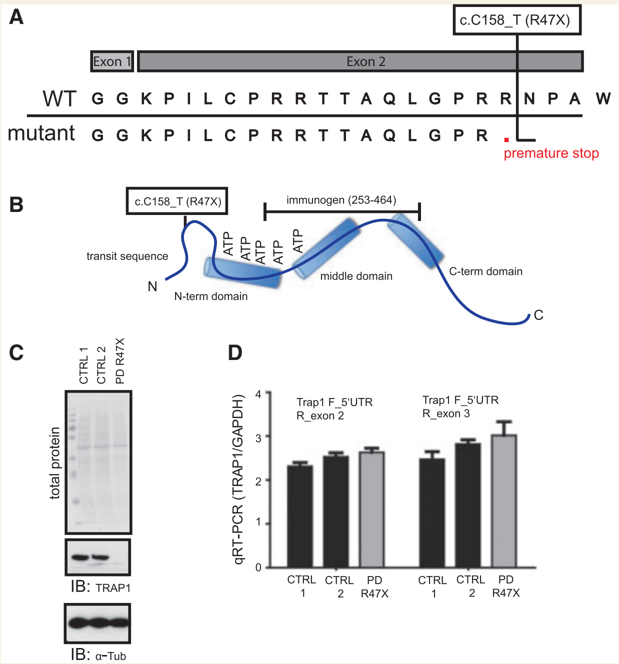


Figure 3 TRAP1 loss-of-function in Parkinson's disease. (**A**) A diagram showing the position of the c.C158_T (R47X) mutation within the amino acid sequence of TRAP1. (**B**) Diagram showing the position of the c.C158_T (R47X) mutation in the protein structure of TRAP1. The binding site of the TRAP1 antibody (amino acid sequence 253–464) is also shown. (**C**) Copper stain showing total protein loading of an immunoblot (IB) (*top*) probed for TRAP1 (*middle*) and the loading control α-tubulin (*bottom*). (**D**) Real-time PCR amplified *TRAP1* transcripts normalized to housekeeping gene *GAPDH* in healthy controls and in the TRAP1 R47X patient using primers amplifying a region starting at the 5' UTR and spanning until exon 2 of *TRAP1* upstream of the R47X mutation and exon 3 downstream of the R47X mutation. PD = Parkinson's disease; WT = wild-type.

We measured the co-localization of mitochondria and lysosomes in patient and control fibroblasts and the results show similar co-localization of mitochondria with lysosomes under normal physiological conditions. Following mild induction of autophagy by serum withdrawal, we found that mitochondria to lysosome translocation was more pronounced in TRAP1 R47X cells (Fig. 4E), suggesting increased mitochondrial turnover.

Respiratory analysis of patient cells and controls, recorded oxygen consumption during a mitochondrial stress test, where minimal, maximal and inhibited respiration is induced by oligomycin, the uncoupler FCCP and antimycin A, respectively (Fig. 4F). Basal respiration was significantly increased in R47X fibroblasts (Fig. 4F). The extracellular acidification rate (an indicator of glycolysis) normalized to the rate of oxygen consumption per cell in R47X fibroblasts was generally higher than that of healthy controls (Fig. 4G). However, the calculated glycolytic shift after the addition of oligomycin is similar between patient and controls (Fig. 4G). Complex I enriched in mitochondrial extracts of R47X cells was significantly more active in oxidizing NADH given as a substrate along with decylubiquinone *in vitro* than healthy individuals (Fig. 4H). Furthermore, significantly more ATP was produced in R47X fibroblasts compared to controls (Fig. 4I), indicating that the complexes of the respiratory chain are not damaged and suggesting that mitochondria in R47X patient cells have increased respiratory activity.

TRAP1 deficiency is reported to promote increases in mitochondrial respiration, ATP levels and reactive oxygen species in mice (Yoshida *et al.*, 2013). Therefore, we identified the index patient in a previous study measuring mitochondrial reactive oxygen species in sporadic Parkinson's disease patient fibroblasts. The index patient has above average levels of mitochondrial reactive oxygen species, but are only mildly elevated in comparison to several other sporadic Parkinson's disease patients, genetic Parkinson's disease (PINK1, parkin and DJ-1 patients)

Table 1 R47X patient information

Patient Information	Details	Comments
Age of onset	2004 (71 years of age)	Presented with pain in right shoulder and aching affecter (smaller) with slowing down in general; later mild tremor when tired
Diagnosis	2004	Idiopathic Parkinson's disease
Family history	No Parkinson's disease in family	Mother with dementia
Other	Dilatative cardiomyopathy	No medication
	Benign prostata hyperplasia	
	Pancreatitis	Chronic
	Polyneuropathy	
	Kidney insufficiency	Chronic
	Sleep apnoea	
	Cataracts	

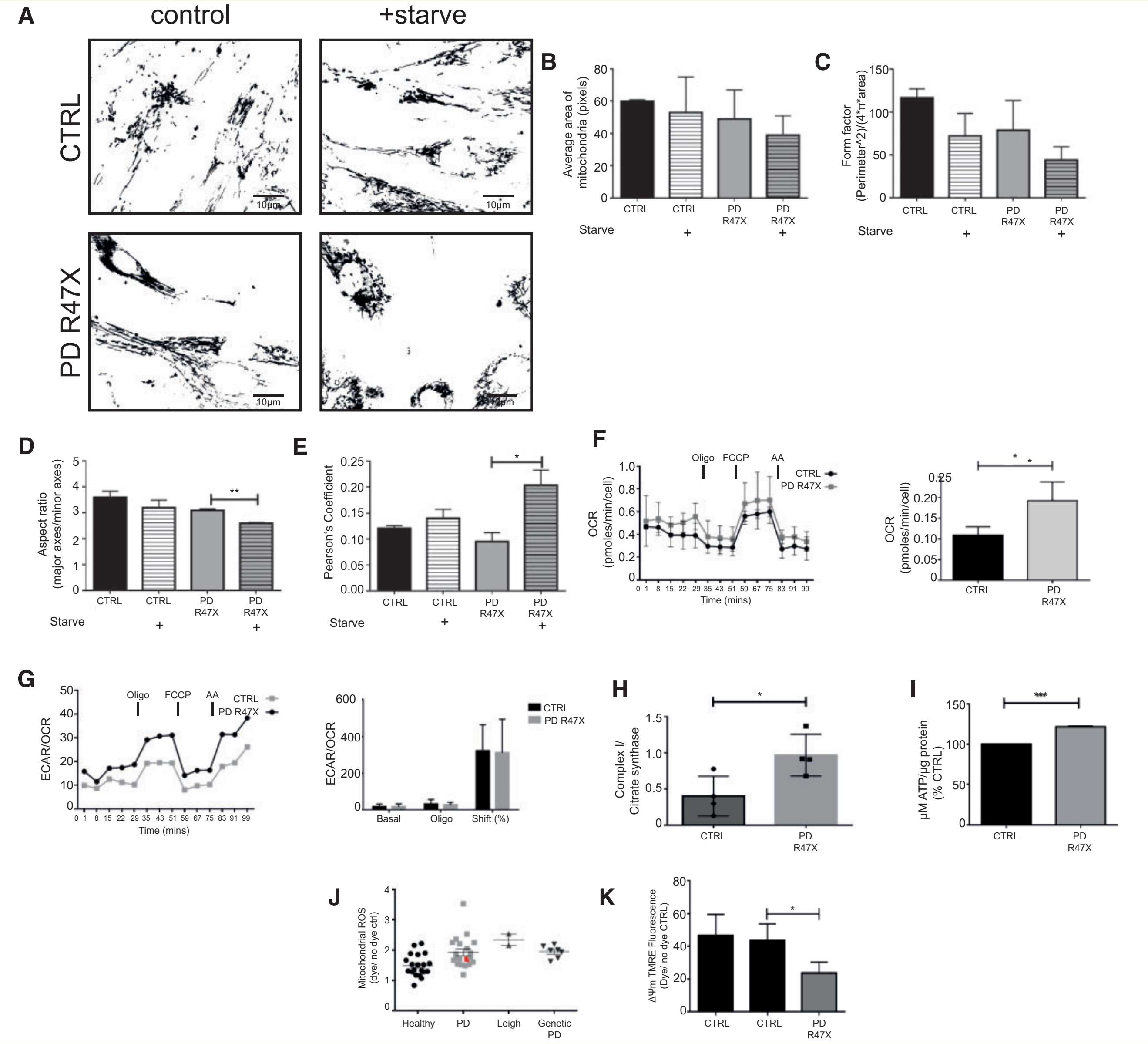


Figure 4 TRAP1 R47X Parkinson's disease patient mitochondria meet ATP demand but have reduced membrane potential. (**A**) Representative binary images of mitochondria and the calculated (**B**) mitochondrial area, (**C**) form factor (mitochondrial branching) and (**D**) aspect ratio (mitochondrial length) in fibroblasts from healthy individuals (CTRL) and a Parkinson's disease patient carrying the R47X mutation (PD R47X) untreated or serum starved (Starve) for 24 h. (**E**) Mitochondrial-lysosomal co-localization expressed as Pearson's coefficient in fibroblasts from healthy individuals (CTRL) and a Parkinson's disease patient carrying the R47X mutation (PD R47X) untreated or deprived of serum (Starve) for 24 h. (**F**, *left*) Mean average OCR of two healthy control fibroblasts and Parkinson's disease patient carrying the R47X mutation over a time course. Measurement of basal OCR is followed by the addition of oligomycin (oligo) 1 µM final concentration, FCCP 1 µM final concentration and antimycin A (Ant A, 1 µM final concentration) and rotenone (AA) 1 µM final concentration. (*Right*) Statistical analysis showing increased mean average basal OCR in Parkinson's disease patient carrying the R47X mutation compared to the mean average OCR of two healthy control fibroblasts. (**G**, *left*) Extracellular acidification rate (ECAR) normalized to OCR/cell to account for cell numbers of two healthy control fibroblasts and Parkinson's disease patient carrying the R47X mutation over a time course. (*Right*) Statistical analysis showing no changes in ECAR/OCR under basal conditions, minimal OXPHOS (oligo) and the per cent shift from basal condition to minimal OXPHOS (glycolytic shift) in Parkinson's disease patient carrying the R47X mutation compared to the mean average OCR of two healthy control fibroblasts. (**H**) Complex I enzyme activity (normalized to citrate synthase enzyme activity) in isolated mitochondria is increased in Parkinson's disease patient carrying the R47X TRAP1 mutation compared to two healthy control fibroblasts lines. (**I**) ATP levels (normalized to total protein) are increased in Parkinson's disease patient carrying the R47X mutation compared to two healthy control fibroblasts lines. (**J**) Mitochondrial reactive oxygen species (ROS) levels in a Parkinson's disease patient carrying the R47X TRAP1 mutation (highlighted in red), compared to the mean average mitochondrial ROS measured in healthy controls (healthy), sporadic Parkinson's disease patients (PD), Leigh syndrome patients (Leigh) and familial Parkinson's disease patients (genetic Parkinson's disease, including PINK1, Parkin and DJ-1) in fibroblasts. (**K**) Mitochondrial membrane potential (ΔΨm) is significantly reduced in Parkinson's disease patient carrying the R47X mutation compared to two healthy control fibroblast lines. All statistical tests were the Student's *t*-test assuming different standard deviation. Error bars show standard deviation and $*P < 0.05$; $**P < 0.01$ and $***P < 0.001$. TMRE = tetramethylrhodamine, ethyl ester, perchlorate.

and Leigh syndrome patient fibroblast lines (Fig. 4J). Finally, mitochondrial membrane potential was significantly reduced in TRAP1 R47X fibroblasts compared to controls (Fig. 4K).

Metformin rescues the R47X phenotype via a mechanism involving mitochondrial biogenesis

In ovarian cancer, TRAP1 silencing causes resistance to chemotherapy drugs because oxidative phosphorylation is increased. Interestingly, the resistance to chemotherapy could be reversed by mild inhibition of mitochondrial respiration by the diabetes drug metformin or oligomycin (Matassa *et al.*, 2016). Therefore, we treated TRAP1 R47X patient fibroblasts with 10 mM metformin, sublethal concentrations of oligomycin and the antioxidant *N*-acetyl cysteine (NAC) to see whether we could rescue the patient phenotype. Metformin and oligomycin treatment restored the mitochondrial membrane potential observed in the patient, whereas the antioxidant NAC could not (Fig. 5A).

We subjected human cancer cells to a range of toxins and stressors and measured their effect on mitochondrial membrane potential. Dopamine, hydrogen peroxide and the ionophore valinomycin greatly reduced mitochondrial membrane potential and this could not be rescued by metformin. Dopamine toxicity was protected by addition of the antioxidant NAC and metformin treatment alone does not reduce mitochondrial membrane potential (Fig. 5B). Finally, reduced mitochondrial membrane potential induced by the Hsp90 family/TRAP1 inhibitor 17-AAG could be reversed by metformin (Fig. 5B), suggesting there is a specific effect of metformin in paradigms related to TRAP1.

To investigate further the mechanism by which metformin is protective in our model, we measured the fluorescence lifetime of NADH in living cells from the TRAP1 R47X patient and healthy individuals with and without treatment with metformin. Bound NADH indicates usage in mitochondrial respiration, whereas free NADH is associated with glycolysis (Bird *et al.*, 2005; Blacker *et al.*, 2014). We found significantly reduced bound NADH and increased free NADH following metformin treatment in all cell types (Fig. 5C). This finding supports the observation that metformin suppresses gluconeogenesis (Kim *et al.*, 2008), inhibits complex I (Owen *et al.*, 2000) and shifts the balance between coupling and uncoupling reactions via the TCA cycle (Andrzejewski *et al.*, 2014). We found a similar bound/unbound NADH ratio in the untreated R47X patient fibroblasts as in the metformin treated controls and the addition of metformin in the patient did not reverse the bound/unbound NADH ratio (Fig. 5C). These data, although highly significant, represent overall a very small shift in the total levels of bound versus unbound NADH levels (Fig. 5C). The data suggest that the protective mechanism of metformin in the R47X patient is not via the metabolic switch between oxidative phosphorylation (OXPHOS) and glycolysis. These data could mean that either glycolysis is favoured in the R47X patient or mitochondrial turnover and/or the NAD^+/NADH pool are altered.

NAD^+ and combined NAD^+ and NADH levels are significantly increased in TRAP1 R47X patient cells compared to controls (Fig. 5D). Metformin treatment in healthy individuals and in TRAP1 R47X patient cells lowers both NAD^+ and total NAD^+ and NADH levels in one control and the patient, but not significantly (Fig. 5D).

We observe a significantly reduced ratio of succinate dehydrogenase (SDH) to cytochrome *c* oxidase (COX mtDNA-encoded subunit) in both R47X patient fibroblasts and TRAP1 knockout mouse adult fibroblasts compared to controls (Fig. 5E), which indicates an imbalance between nuclear and mitochondrially encoded mitochondrial proteins (termed mitonuclear imbalance), likely induced by the increased NAD^+ and NADH pool and in agreement with the effect of NAD^+ boosters on the age-associated metabolic decline and promotion of longevity in worms (Mouchiroud *et al.*, 2013).

Metformin is able to reverse the mitonuclear imbalance in the TRAP1 R47X patient fibroblasts (Fig. 5E), indicating that the mitonuclear imbalance is the converging step in the survival pathway that can be targeted pharmacologically. Mitonuclear protein imbalance controls longevity in mammals via induction of the mitochondrial unfolded protein response (mtUPR) (Houtkooper *et al.*, 2013; Mouchiroud *et al.*, 2013). Therefore, we monitored the levels of Hsp60, Hsp70 and mitochondrial Hsp90, three markers of the mtUPR. We found that on average both Hsp60 and mtHsp70 levels were higher in R47X TRAP1 patient fibroblasts compared to two healthy controls in three independent experiments (Fig. 5F). Hsp90 levels were also elevated but the difference was not significant (Supplementary Fig. 2C). These data suggest that in TRAP1 loss-of-function models, the mtUPR is upregulated. This is associated with increased turnover of mitochondria and the significant elevation of subunits of mitochondrial respiratory complexes I, II, III and IV, which is also rescued by metformin (Supplementary Fig. 2B). Phosphorylated ERK1/2 orchestrates metabolic switching via TRAP1 (Masgras *et al.*, 2017). Here we found that the levels of phosphorylated ERK1/2 are increased in the index patient fibroblasts and can be reversed by metformin (Supplementary Fig. 2D).

Discussion

TRAP1 and HTRA2 are targets of the Parkinson's disease kinase PINK1 (Plun-Favreau *et al.*, 2007; Pridgeon *et al.*, 2007). However, how these three proteins act together in Parkinson's disease signalling still remains to be elucidated. One of the barriers to dissecting a pathway involving HTRA2 and TRAP1 was the lack of mechanistic evidence for the downstream mitochondrial function observed.

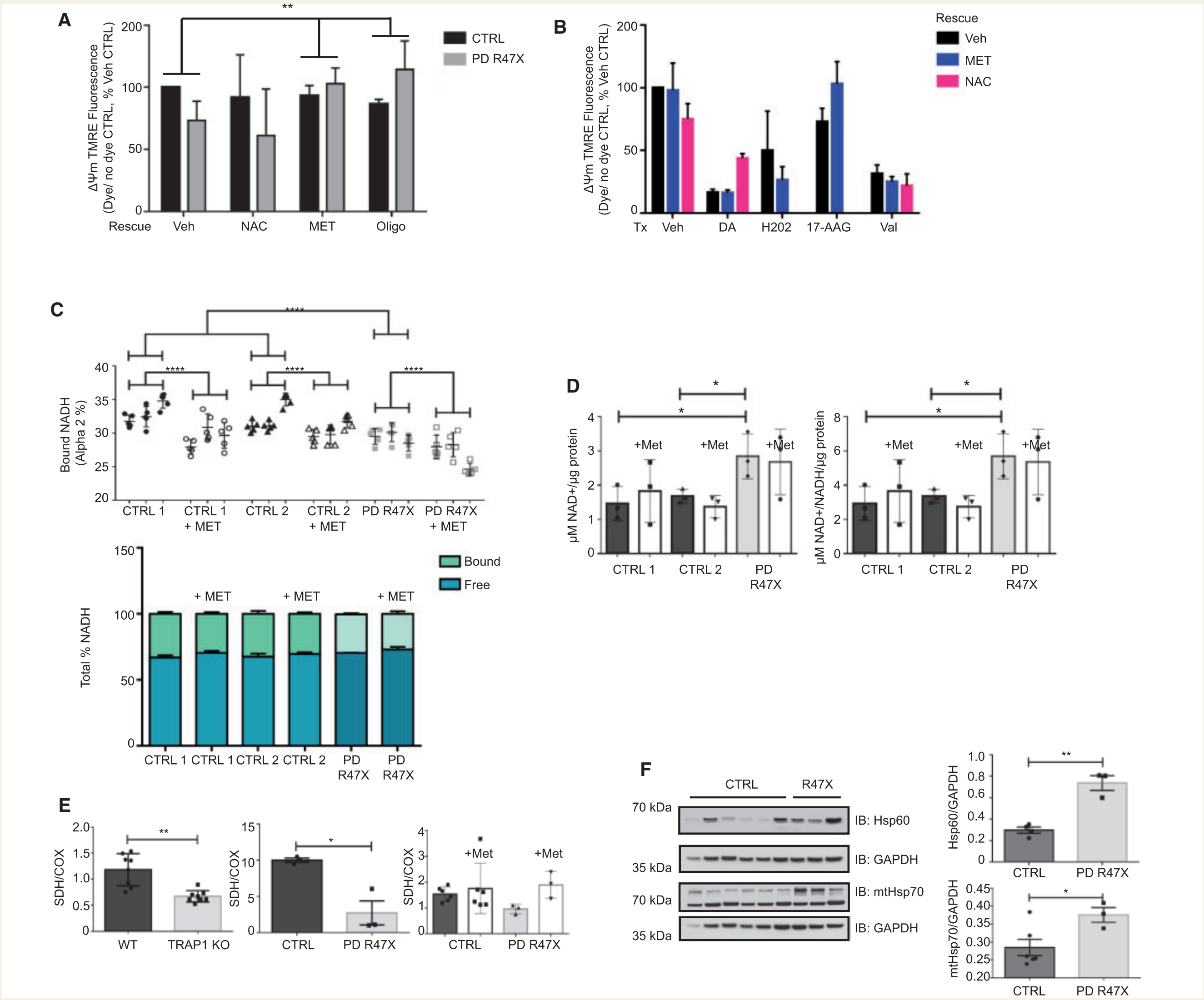


Figure 5 Metformin rescues the R47X phenotype via a mechanism involving mitochondrial biogenesis. (**A**) Mitochondrial membrane potential (ΔΨm) is reduced in Parkinson's disease patient carrying the R47X mutation (PD R47X) compared to the mean average of two healthy control (CTRL) fibroblast lines. All fibroblasts were treated with DMSO vehicle (Veh), 0.5 mM antioxidant *N*-acetyl cysteine (NAC), 10 mM Metformin hydrochloride (MET) or 250 nM oligomycin (oligo) for 24 h, of which metformin and oligomycin reverted the phenotype. (**B**) Mitochondrial membrane potential (ΔΨm) is reduced in HeLa cells treated (Tx) with 500 μM dopamine (DA), 400 μM hydrogen peroxide (H_2O_2), 2 μM 17-AAG, and 100 nM valinomycin (Val) but not the DMSO vehicle control (Veh) for 24 h. Antioxidant *N*-acetyl cysteine (NAC, 0.5 mM) rescues the dopamine toxicity to some extent, whereas 10 mM metformin hydrochloride (MET) rescued the inhibition of TRAP1 by 17-AAG. (**C**, *top*) The percentage of bound NADH in two healthy control fibroblast lines (CTRL) and in the Parkinson's disease patient carrying the R47X mutation (PD R47X), with or without treatment with 10 mM metformin for 24 h. (*Bottom*) The overall percentage of bound (green) and free (blue) NADH in two healthy control fibroblast lines (CTRL) and in the Parkinson's disease patient carrying the R47X mutation (PD R47X), with or without treatment with 10 mM metformin for 24 h. (**D**) Levels of NAD^+ (*left*) and total NAD^+/NADH (*right*) measured in two healthy fibroblast lines and the Parkinsons' disease patient carrying the R47X mutation with or without treatment with 10 mM metformin for 24 h. (**E**) Succinate dehydrogenase (nuclear encoded) to mt COX (mitochondrial encoded) protein ratio in TRAP1 knockout mouse embryonic fibroblasts (*left*), R47X Parkinson's disease patient (*middle*) and R47X patient cells treated with metformin (*right*). (**F**) Immunoblots (IB) of Hsp60 and mtHsp70 in three independent extractions from two healthy fibroblast lines (CTRL) and the R47X Parkinson's disease patient (*left*), quantified for statistical analyses (*right*). The Student's *t*-test was used assuming different standard deviation to compare patient and control group. Two-way ANOVA was used to compare cell types and treatments. Error bars show standard deviation and $^*P < 0.05$; $^{**}P < 0.01$ and $^{***}P < 0.001$.

Here we have shown that HTRA2 and TRAP1 physically interact and regulate each other. The biochemistry of the interaction is non-canonical and does not involve the protease activity of HTRA2, leaving us to speculate that HTRA2 and TRAP1 perform in a common intra-mitochondrial chaperoning or quality control system. In this study, overexpression of the catalytically inactive HTRA2 S306A reduces TRAP1 protein levels to the same extent as wild-type HTRA2. As in mnd2 mice carrying the S276C HTRA2 mutation, HTRA2 is catalytically inactive and as the mice phenocopy the mitochondrial dysfunction and neurodegeneration seen in HTRA2 knockout mice (Martins *et al.*, 2004), S306A is also unlikely to rescue HTRA2 loss of mitochondrial function. However, HTRA2 possesses chaperone activity in its basal state (Li *et al.*, 2002). Protease dead HTRA2 could still bind TRAP1 via its PDZ domain. The PDZ domain of HTRA2 has a 'YIGV' recognition pattern but also detects long hydrophobic stretches (Zhang *et al.*, 2007), preferentially C-terminal peptides (Clausen *et al.*, 2002). Interestingly, analysis of TRAP1 hydrophobicity shows a hydrophobic stretch at the C-terminal (Supplementary Fig. 2A) and therefore an alternative mode of interaction by association should not be ruled out. Other mitochondrial proteases could also be influencing TRAP1 and loss of HTRA2 could trigger other proteases in order to maintain proteostasis, which displays some redundancy.

One concept that links HTRA2 and TRAP1 in the context of neurodegeneration is the mtUPR. The mtUPR is a highly conserved cellular response activated when the accumulation of unfolded or misfolded proteins goes beyond the chaperone capacity of the mitochondria (Pellegrino *et al.*, 2013). The mtUPR activates transcription of nuclear-encoded mitochondrial chaperone genes to promote protein homeostasis within mitochondria. HTRA2 levels are increased during mtUPR (Spiess *et al.*, 1999) and loss of HTRA2 contributes to transcriptional stress response (Moisoi *et al.*, 2009). Overexpression of TRAP1 activates mtUPR and extends lifespan in *Drosophila* (Baqri *et al.*, 2014) and TRAP1 inhibition promotes the mtUPR response in *Caenorhabditis elegans* (Munch and Harper, 2016).

TRAP1 gain-of-function rescues PINK1 (Zhang *et al.*, 2013) and PINK1/parkin loss-of-function in *Drosophila* (Costa *et al.*, 2013) and here we can show that TRAP1 rescues HTRA2 and PINK1 loss-of-function in human cells. In addition to its role as a chaperone, TRAP1 is also involved in metabolic switching (Yoshida *et al.*, 2013; Sciacovelli *et al.*, 2013; Rasola *et al.*, 2014; Masgras *et al.*, 2017) and therefore through the identification of a sporadic Parkinson's disease patient homozygous for a premature stop mutation in TRAP1 and data from the patient fibroblasts, we have uncovered a mechanism involving mitochondrial metabolism.

TRAP1 mutations could be important for our understanding of the underlying biological mechanisms that lead to Parkinson's disease and although the role and influence of rare variants in complex diseases is a debated subject, data generated so far indicate that common and rare variants are not mutually exclusive. We used the PPMI repository (with 380 Parkinson's disease cases and 162 controls) to perform a comprehensive burden analysis. Truncating variants predicted to cause loss-of-function of TRAP1 are very rare and were only observed in Parkinson's disease patients and not in controls. For rare missense *TRAP1* mutations, we found a significantly different burden (P-values < 0.05) between patients and controls. We also investigated whether there are healthy individuals who have both alleles of the *TRAP1* gene inactivated. Using our in-house Helmholtz database and several available large datasets, we found no such *TRAP1* mutation, showing that biallelic loss-of-function mutations are not well tolerated in healthy individuals. Overall, the result of the burden analysis points to an association of *TRAP1* rare, missense variants in controls that may be protective for Parkinson's disease. To further validate the findings on low frequency variants in Parkinson's disease, we would need independent, larger sample sets.

In 2014, Luykx *et al.* (2014) hinted that TRAP1 variants are associated with neurotransmitter metabolism and Parkinson's disease. The authors performed a genome-wide association study (GWAS) analysis and found a significant association of the ratio of HVA/5-HIAA, indicating enhanced monoamine turnover in variants of six genes, among them were *PINK1* and *TRAP1*, further supporting the genetic contribution of TRAP1 to Parkinson's disease.

In the case reported here, a homozygous stop mutation in *TRAP1* in a Parkinson's disease patient leads to complete loss of the TRAP1 protein. *TRAP1* mutations have previously been associated with chronic pain, fatigue and gastrointestinal dysmotility (Boles *et al.*, 2015), a recognized common dysfunction in Parkinson's disease (Pfeiffer, 2003). One highly conserved variant in this study (p.Ile253Val) was also identified in both German and Austrian Parkinson's disease patients. Furthermore, recessive mutations in TRAP1 were identified in two families with congenital abnormalities of the kidney and urinary tract (CAKUT) and VACTERL association (congenital abnormalities in multiple organs) (Saisawat *et al.*, 2014). Interestingly, the late-onset Parkinson's disease patient with a homozygous stop mutation (R47X) in *TRAP1* described here was also diagnosed with chronic pancreatitis and, chronic kidney insufficiency but not diabetes. The R47X patient also shows other symptoms related to mitochondriopathies such as cardiomyopathy, polyneuropathy, sleep apnoea and cataracts. Studies in mice have shown that TRAP1 overexpression protects against cardiac hypertrophy (Zhang *et al.*, 2011) and underscores the link between TRAP1 defects and mitochondriopathy.

In line with previous work performed in TRAP1 knockout mice (Yoshida *et al.*, 2013), TRAP1-deficient patient fibroblasts show increased respiration, complex I activity and ATP output. We also found more unbound NADH, which indicates favouring of glycolysis. However, these changes, although highly significant, are overall very small, which might reflect the low metabolic demand in fibroblasts compared to neurons. Unbound NADH could also come from the NAD + /NADH pool, which is increased in mitochondrial biogenesis. NAD^+ metabolism engages key effectors of longevity, and interestingly modulating NAD^+ levels has become a focus for intervention in age-related diseases (Karpac and Jasper, 2013). NAD^+ signals mitochondrial biogenesis via the sirtuin pathway, this impacts mitonuclear protein balance and initiates the mtUPR, promoting longevity (Mouchiroud *et al.*, 2013).

Altered stoichiometry between nuclear and mtDNA encoded proteins (mitonuclear protein balance) is a conserved longevity mechanism across many species. Mitonuclear protein imbalance is coupled with the activation of the mtUPR, activation of mitochondrial chaperones and longevity (Houtkooper *et al.*, 2013). Mitochondrial biogenesis, normal ageing, mitochondrial transcription and translation all influence the balance of nuclear and mtDNA encoded mitochondrial proteins (Houtkooper *et al.*, 2013).

The diabetes mellitus type 2 drug metformin was investigated in this study because of its ability to reverse TRAP1-dependent chemotherapy resistance in ovarian cancer (Matassa *et al.*, 2016). The ability of metformin (and not an antioxidant) to rescue the reduced mitochondrial membrane potential phenotype is of particular interest as metformin has previously been shown to be protective in Parkinson's disease models (Patil *et al.*, 2014; Perez-Revuelta *et al.*, 2014) and there are significantly fewer cases of Parkinson's disease in diabetes mellitus type 2 patients taking metformin (Wahlqvist *et al.*, 2012). We propose that loss of TRAP1 hinders the fine tuning of energy metabolism, proteostasis and the mtUPR response. It is this fine tuning that over time, when not available, pushes the cell in favour of meeting immediate energy needs, rather than energy restriction. Further work to generate induced pluripotent stem cells from the TRAP1 R47X patient fibroblasts and gene correct the mutation would confirm cause of disease. In conclusion, loss-of-function mutations in TRAP1 are rare, yet analyses of the biological pathway involving TRAP1, show that TRAP1 is important for mitochondrial signalling in Parkinson's disease. These data also underscore the role of rare variants in the pathogenesis of Parkinson's disease and suggest that treatments other than antioxidants should also be considered for individualized therapies in aetiologically heterogeneous syndromes such as Parkinsonism.

Acknowledgements

The authors are grateful to Tim M. Strom of the Institut für Humangenetik, Helmholtz Zentrum München, Munich, Germany and gratefully acknowledge the access to the Helmholtz Zentrum data base. Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org.

Funding

The project underlying this publication was funded by the German Federal Ministry of Education and Research (BMBF) under the support code 031 A 430 A. This work was further supported by grants from the Luxembourg National Research Fund (FNR) within the National Centre of Excellence in Research on Parkinson's disease (NCER-PD), PEARL programme (FNR; FNR/P13/6682797 to R.K.), the German Research Council (DFG; KR2119/8-1 to R.K. and T.G.), the EU Joint Program-Neurodegenerative Diseases (JPND; COURAGE-PD) and by the European Union's Horizon2020 research and innovation program under grant agreement No 692320 (WIDESPREAD; CENTRE-PD). T.G. and C.B. gratefully acknowledge the support of the Eva Thies Stiftung. E.G. acknowledges FNR support by grant C13/BM/5782168. K.S.L. acknowledges support by the Ministry of Science, Research and the Arts of Baden-Württemberg (33-729.55-3/214 and SI-BW 01222-91), and the Deutsche Forschungsgemeinschaft (INST 2388/30-1). E.V. and D.P. were supported by the Swiss National Science Foundation. Bioinformatics analyses presented in this paper were carried out in part using the HPC facilities of the University of Luxembourg (see http://hpc.uni.lu). PPMI, a public-private partnership, is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including Abbvie, Avid, Biogen, Bristol-Myers Squibb, Covance, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeek, Merck, Meso Scale Discovery, Pfizer, Piramal, Roche, Servier, Teva, UCB, and Golub Capital.

Supplementary material

Supplementary material is available at *Brain* online.

References

Altieri DC, Stein GS, Lian JB, Languino, LR. TRAP-1, the mitochondrial Hsp90. Biochim Biophys Acta 2012; 1823: 767–73.

Amoroso MR, Matassa DS, Agliarulo I, Avolio R, Lu H, Sisinni L, et al. TRAP1 downregulation in human ovarian cancer enhances invasion and epithelial-mesenchymal transition. Cell Death Dis 2016; 7: e2522.

Andrzejewski S, Gravel SP, Pollak M, St-Pierre J. Metformin directly acts on mitochondria to alter cellular bioenergetics. Cancer Metab 2014; 2: 12.

Baqri RM, Pietron AV, Gokhale RH, Turner BA, Kaguni LS, Shingleton AW, et al. Mitochondrial chaperone TRAP1 activates the mitochondrial UPR and extends healthspan in Drosophila. Mech Ageing Dev 2014; 141–2: 35–45.

Bird DK, Yan L, Vrotsos KM, Eliceiri KW, Vaughan EM, Keely, PJ, et al. Metabolic mapping of MCF10A human breast cells via multiphoton fluorescence lifetime imaging of the coenzyme NADH. Cancer Res 2005; 65: 8766–73.

Blacker TS, Mann ZF, Gale JE, Ziegler M, Bain AJ, Szabadkai G, et al. Separating NADH and NADPH fluorescence in live cells and tissues using FLIM. Nat Commun 2014; 5: 3936.

Bogaerts V, Nuytemans K, Reumers J, Pals P, Engelborghs S, Pickut B, et al. Genetic variability in the mitochondrial serine protease HTRA2 contributes to risk for Parkinson disease. Hum Mutat 2008; 29: 832–40.

Boles RG, Hornung HA, Moody AE, Ortiz TB, Wong SA, Eggington JM, et al. Hurt, tired and queasy: specific variants in the ATPase domain of the TRAP1 mitochondrial chaperone are associated with common, chronic "functional" symptomatology including pain, fatigue and gastrointestinal dysmotility. Mitochondrion 2015; 23: 64–70.

Burbulla LF, Fitzgerald JC, Stegen K, Westermeier J, Thost AK, Kato H, et al. Mitochondrial proteolytic stress induced by loss of mortalin function is rescued by Parkin and PINK1. Cell Death Dis 2014; 5: e1180.

Burbulla LF, Kruger R. The use of primary human fibroblasts for monitoring mitochondrial phenotypes in the field of Parkinson's disease. J Vis Exp 2012; 68: 4228.

Burte F, De Girolamo LA, Hargreaves AJ, Billett EE. Alterations in the mitochondrial proteome of neuroblastoma cells in response to complex 1 inhibition. J Proteome Res 2011; 10: 1974–86.

Butler EK, Voigt A, Lutz AK, Toegel JP, Gerhardt E, Karsten P, et al. The mitochondrial chaperone protein TRAP1 mitigates alpha-Synuclein toxicity. PLoS Genet 2012; 8: e1002488.

Casadei N, Sood P, Ulrich T, Fallier-Becker P, Kieper N, Helling S, et al. Mitochondrial defects and neurodegeneration in mice overexpressing wild-type or G399S mutant HtrA2. Hum Mol Genet 2016; 25: 459–71.

Clausen T, Southan C, Ehrmann M. The HtrA family of proteases: implications for protein composition and cell fate. Mol Cell 2002; 10: 443–55.

Costa AC, Loh SH, Martins LM. Drosophila Trap1 protects against mitochondrial dysfunction in a PINK1/parkin model of Parkinson's disease. Cell Death Dis 2013; 4: e467.

Fitzgerald JC, Camprubi MD, Dunn L, Wu HC, Ip NY, Kruger R, et al. Phosphorylation of HtrA2 by cyclin-dependent kinase-5 is important for mitochondrial function. Cell Death Differ 2012; 19: 257–66.

Geisler S, Holmstrom KM, Skujat D, Fiesel FC, Rothfuss OC, Kahle PJ, et al. PINK1/Parkin-mediated mitophagy is dependent on VDAC1 and p62/SQSTM1. Nat Cell Biol 2010; 12: 119–31.

Hargreaves IP, Duncan AJ, Wu L, Agrawal A, Land JM, Heales SJ. Inhibition of mitochondrial complex IV leads to secondary loss complex II-III activity: implications for the pathogenesis and treatment of mitochondrial encephalomyopathies. Mitochondrion 2007; 7: 284–7.

Houtkooper RH, Mouchiroud L, Ryu D, Moullan N, Katsyuba E, Knott G, et al. Mitonuclear protein imbalance as a conserved longevity mechanism. Nature 2013; 497: 451–7.

Hua G, Zhang Q, Fan Z. Heat shock protein 75 (TRAP1) antagonizes reactive oxygen species generation and protects cells from granzyme M-mediated apoptosis. J Biol Chem 2007; 282: 20553–60.

Im CN, Lee JS, Zheng Y, Seo JS. Iron chelation study in a normal human hepatocyte cell line suggests that tumor necrosis factor receptor-associated protein 1 (TRAP1) regulates production of reactive oxygen species. J Cell Biochem 2007; 100: 474–86.

Kang BH, Plescia J, Dohi T, Rosa J, Doxsey SJ, Altieri DC. Regulation of tumor cell mitochondrial homeostasis by an organelle-specific Hsp90 chaperone network. Cell 2007; 131: 257–70.

Karpac J, Jasper H. Aging: seeking mitonuclear balance. Cell 2013; 154: 271–3.

Kieper N, Holmstrom KM, Ciceri D, Fiesel FC, Wolburg H, Ziviani, E, et al. Modulation of mitochondrial function and morphology by interaction of Omi/HtrA2 with the mitochondrial fusion factor OPA1. Exp Cell Res 2010; 316: 1213–24.

Kim YD, Park KG, Lee YS, Park YY, Kim DK, Nedumaran B, et al. Metformin inhibits hepatic gluconeogenesis through AMP-activated protein kinase-dependent regulation of the orphan nuclear receptor SHP. Diabetes 2008; 57: 306–14.

Kitada T, Asakawa S, Hattori N, Matsumine H, Yamamura Y, Minoshima S, et al. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. Nature 1998; 392: 605–8.

Kruger R, Klucken J, Weiss D, Tonges L, Kolber P, Unterecker S, et al. Classification of advanced stages of Parkinson's disease: translation into stratified treatments. J Neural Transm (Vienna) 2017; 124: 1015–27.

Lakner PH, Monaghan MG, Moller Y, Olayioye MA, Schenke-Layland K. Applying phasor approach analysis of multiphoton FLIM measurements to probe the metabolic activity of three-dimensional in vitro cell culture models. Sci Rep 2017; 7: 42730.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016; 536: 285–91.

Li W, Srinivasula SM, Chai J, Li P, Wu JW, Zhang Z, et al. Structural insights into the pro-apoptotic function of mitochondrial serine protease HtrA2/Omi. Nat Struct Biol 2002; 9: 436–41.

Luykx JJ, Bakker SC, Lentjes E, Neeleman M, Strengman E, Mentink L, et al. Genome-wide association study of monoamine metabolite levels in human cerebrospinal fluid. Mol Psychiatry 2014; 19: 228–34.

Lv Q, Sun H, Cao C, Gao B, Qi Y. Overexpression of tumor necrosis factor receptor-associated protein 1 (TRAP1) are associated with poor prognosis of epithelial ovarian cancer. Tumour Biol 2016; 37: 2721–7.

Martins LM. The serine protease Omi/HtrA2: a second mammalian protein with a Reaper-like function. Cell Death Differ 2002; 9: 699–701.

Martins LM, Iaccarino I, Tenev T, Gschmeissner S, Totty NF, Lemoine NR, et al. The serine protease Omi/HtrA2 regulates apoptosis by binding XIAP through a reaper-like motif. J Biol Chem 2002; 277: 439–44.

Martins LM, Morrison A, Klupsch K, Fedele V, Moisoi N, Teismann P, et al. Neuroprotective role of the Reaper-related serine protease HtrA2/Omi revealed by targeted deletion in mice. Mol Cell Biol 2004; 24: 9848–62.

Masgras I, Ciscato F, Brunati AM, Tibaldi E, Indraccolo S, Curtarello M, et al. Absence of neurofibromin induces an oncogenic metabolic switch via mitochondrial ERK-mediated phosphorylation of the chaperone TRAP1. Cell Rep 2017; 18: 659–72.

Masuda Y, Shima G, Aiuchi T, Horie M, Hori K, Nakajo S, et al. Involvement of tumor necrosis factor receptor-associated protein 1 (TRAP1) in apoptosis induced by beta-hydroxyisovalerylshikonin. J Biol Chem 2004; 279: 42503–15.

Matassa DS, Amoroso MR, Lu H, Avolio R, Arzeni D, Procaccini C, et al. Oxidative metabolism drives inflammation-induced platinum resistance in human ovarian cancer. Cell Death Differ 2016; 23: 1542–54.

Moisoi N, Klupsch K, Fedele V, East P, Sharma S, Renton A, et al. Mitochondrial dysfunction triggered by loss of HtrA2 results in the activation of a brain-specific transcriptional stress response. Cell Death Differ 2009; 16: 449–64.

Mouchiroud L, Houtkooper RH, Moullan N, Katsyuba E, Ryu D, Canto C, et al. The NAD(+)/sirtuin pathway modulates longevity through activation of mitochondrial UPR and FOXO signaling. Cell 2013; 154: 430–41.

Munch C, Harper JW. Mitochondrial unfolded protein response controls matrix pre-RNA processing and translation. Nature 2016; 534: 710–3.

Narendra DP, Jin SM, Tanaka A, Suen DF, Gautier CA, Shen J, et al. PINK1 is selectively stabilized on impaired mitochondria to activate Parkin. PLoS Biol 2010; 8: e1000298.

Owen MR, Doran E, Halestrap AP. Evidence that metformin exerts its anti-diabetic effects through inhibition of complex 1 of the mitochondrial respiratory chain. Biochem J 2000; 348 (Pt 3): 607–14.

Patil SP, Jain PD, Ghumatkar PJ, Tambe R, Sathaye S. Neuroprotective effect of metformin in MPTP-induced Parkinson's disease in mice. Neuroscience 2014; 277: 747–54.

Pellegrino MW, Nargund AM, Haynes CM. Signaling the mitochondrial unfolded protein response. Biochim Biophys Acta 2013; 1833: 410–6.

Perez-Revuelta BI, Hettich MM, Ciociaro A, Rotermund C, Kahle PJ, Krauss S, et al. Metformin lowers Ser-129 phosphorylated alpha-synuclein levels via mTOR-dependent protein phosphatase 2A activation. Cell Death Dis 2014; 5: e1209.

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999; 20: 3551–67.

Pfeiffer RF. Gastrointestinal dysfunction in Parkinson's disease. Lancet Neurol 2003; 2: 107–16.

Plun-Favreau H, Klupsch K, Moisoi N, Gandhi S, Kjaer S, Frith D, et al. The mitochondrial protease HtrA2 is regulated by Parkinson's disease-associated kinase PINK1. Nat Cell Biol 2007; 9: 1243–52.

Pridgeon JW, Olzmann JA, Chin LS, Li L. PINK1 protects against oxidative stress by phosphorylating mitochondrial chaperone TRAP1. PLoS Biol 2007; 5: e172.

Rakovic A, Shurkewitsch K, Seibler P, Grunewald A, Zanon A, Hagenah J, et al. Phosphatase and tensin homolog (PTEN)-induced putative kinase 1 (PINK1)-dependent ubiquitination of endogenous Parkin attenuates mitophagy: study in human primary fibroblasts and induced pluripotent stem cell-derived neurons. J Biol Chem 2013; 288: 2223–37.

Rasola A, Neckers L, Picard D. Mitochondrial oxidative phosphorylation TRAP(1)ped in tumor cells. Trends Cell Biol 2014; 24: 455–63.

Reinhardt P, Glatza M, Hemmer K, Tsytsyura Y, Thiel CS, Hoing S, et al. Derivation and expansion using only small molecules of human neural progenitors for neurodegenerative disease modeling. PLoS One 2013; 8: e59252.

Rogers GW, Brand MD, Petrosyan S, Ashok D, Elorza AA, Ferrick DA, et al. High throughput microplate respiratory measurements using minimal quantities of isolated mitochondria. PLoS One 2011; 6: e21746.

Saisawat P, Kohl S, Hilger AC, Hwang DY, Yung Gee H, Dworschak GC, et al. Whole-exome resequencing reveals recessive mutations in TRAP1 in individuals with CAKUT and VACTERL association. Kidney Int 2014; 85: 1310–7.

Sauer M, Jakob A, Nordheim A, Hochholdinger F. Proteomic analysis of shoot-borne root initiation in maize (Zea mays L.). Proteomics 2006; 6: 2530–41.

Sciacovelli M, Guzzo G, Morello V, Frezza C, Zheng L, Nannini N, et al. The mitochondrial chaperone TRAP1 promotes neoplastic growth by inhibiting succinate dehydrogenase. Cell Metab 2013; 17: 988–99.

Spiess C, Beil A, Ehrmann M. A temperature-dependent switch from chaperone to protease in a widely conserved heat shock protein. Cell 1999; 97: 339–47.

Strauss KM, Martins LM, Plun-Favreau H, Marx FP, Kautzmann S, Berg D, et al. Loss of function mutations in the gene encoding Omi/HtrA2 in Parkinson's disease. Hum Mol Genet 2005; 14: 2099–111.

Valente EM, Abou-Sleiman PM, Caputo V, Muqit MM, Harvey K, Gispert S, et al. Hereditary early-onset Parkinson's disease caused by mutations in PINK1. Science 2004; 304: 1158–60.

Vartholomaiou E, Madon-Simon M, Hagmann S, Muhlebach G, Wurst W, Floss T, et al. Cytosolic Hsp90 alpha and its mitochondrial isoform Trap1 are differentially required in a breast cancer model. Oncotarget 2017; 8: 17428–42.

Wahlqvist ML, Lee MS, Hsu CC, Chuang SY, Lee JT, Tsai HN. Metformin-inclusive sulfonylurea therapy reduces the risk of Parkinson's disease occurring with Type 2 diabetes in a Taiwanese population cohort. Parkinsonism Relat Disord 2012; 18: 753–8.

Wust R, Maurer B, Hauser K, Woitalla D, Sharma M, Kruger R. Mutation analyses and association studies to assess the role of the presenilin-associated rhomboid-like gene in Parkinson's disease. Neurobiol Aging 2016; 39: 217.e13–5.

Yoshida S, Tsutsumi S, Muhlebach G, Sourbier C, Lee MJ, Lee S, et al. Molecular chaperone TRAP1 regulates a metabolic switch between mitochondrial respiration and aerobic glycolysis. Proc Natl Acad Sci USA 2013; 110: E1604–12.

Zhang B, Wang J, Huang Z, Wei P, Liu Y, Hao J, et al. Aberrantly upregulated TRAP1 is required for tumorigenesis of breast cancer. Oncotarget 2015; 6: 44495–508.

Zhang L, Karsten P, Hamm S, Pogson JH, Muller-Rischart AK, Exner N, et al. TRAP1 rescues PINK1 loss-of-function phenotypes. Hum Mol Genet 2013; 22: 2829–41.

Zhang Y, Appleton BA, Wu P, Wiesmann C, Sidhu SS. Structural and functional analysis of the ligand specificity of the HtrA2/Omi PDZ domain. Protein Sci 2007; 16: 1738–50.

Zhang Y, Jiang DS, Yan L, Cheng KJ, Bian ZY, Lin GS. HSP75 protects against cardiac hypertrophy and fibrosis. J Cell Biochem 2011; 112: 1787–94.

ARTICLE OPEN ACCESS

Rare *ABCA7* variants in 2 German families with Alzheimer disease

Patrick May, PhD,* Sabrina Pichler, PhD,* Daniela Hartl, PhD, Dheeraj R. Bobbili, MSc, Manuel Mayhaus, PhD, Christian Spaniol, PhD, Alexander Kurz, Prof, Rudi Balling, Prof, Jochen G. Schneider, Prof, and Matthias Riemenschneider, MD, PhD, Prof

Correspondence
Dr. Pichler
sabrina.pichler@uks.eu

Neurol Genet 2018;4:e224. doi:10.1212/NXG.0000000000000224

Abstract

Objective

The aim of this study was to identify variants associated with familial late-onset Alzheimer disease (AD) using whole-genome sequencing.

Methods

Several families with an autosomal dominant inheritance pattern of AD were analyzed by whole-genome sequencing. Variants were prioritized for rare, likely pathogenic variants in genes already known to be associated with AD and confirmed by Sanger sequencing using standard protocols.

Results

We identified 2 rare *ABCA7* variants (rs143718918 and rs538591288) with varying penetrance in 2 independent German AD families, respectively. The single nucleotide variant (SNV) rs143718918 causes a missense mutation, and the deletion rs538591288 causes a frameshift mutation of *ABCA7*. Both variants have previously been reported in larger cohorts but with incomplete segregation information. *ABCA7* is one of more than 20 AD risk loci that have so far been identified by genome-wide association studies, and both common and rare variants of *ABCA7* have previously been described in different populations with higher frequencies in AD cases than in controls and varying penetrance. Furthermore, ABCA7 is known to be involved in several AD-relevant pathways.

Conclusions

We conclude that both SNVs might contribute to the development of AD in the examined family members. Together with previous findings, our data confirm *ABCA7* as one of the most relevant AD risk genes.

---

*These authors contributed equally to this work.

From the Luxembourg Centre for Systems Biomedicine (LCSB) (P.M., D.R.B., R.B., J.G.S.), University of Luxembourg, Esch-sur-Alzette; Department of Psychiatry and Psychotherapy (S.P., D.H., M.M., C.S., M.R.), Saarland University Hospital, Saarland University, Homburg; and Department of Psychiatry and Psychotherapy (A.K.), Klinikum Rechts der Isar, TU-Muenchen, Munich, Germany.

Funding information and disclosures are provided at the end of the article. Full disclosure form information provided by the authors is available with the full text of this article at Neurology.org/NG.

The Article Processing Charge was funded by Prof Dr Riemenschneider, Klinik für Psychiatrie und Psychotherapie, Universität des Saarlandes.

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND), which permits downloading and sharing the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Copyright © 2018 The Author(s). Published by Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology.

Glossary

AD = Alzheimer disease; **GWAS** = genome-wide association study; **MAF** = minor allele frequency; **WGS** = whole-genome sequencing; **SNV** = single nucleotide variant; **LD** = linkage disequilibrium.

Several genome-wide association studies (GWASs) have identified *ABCA7* (ATP-binding cassette transporter A7) as a risk factor for sporadic late-onset Alzheimer disease (AD).[1–3] *ABCA7* encodes a protein with major function in lipid transport.[4] The protein is involved in AD pathology, as it was demonstrated to play a role in formation, clearance, and aggregation of amyloid beta, the etiologic agent in AD.[5,6] Recently, multiple rare loss-of-function variants in *ABCA7* associated with AD risk and possible causal variants in familial cases and pedigrees have been identified through sequencing efforts.[7–11] Alterations in *ABCA7* have not only been observed in European but also in African American[12] and Asian[13,14] populations either by GWASs or targeting sequencing with varying minor allele frequencies (MAFs). In addition, a protective *ABCA7* variant has also been described, emphasizing the role of this gene in AD.[15] We now present the data of 2 rare variants of *ABCA7* in 2 German families.

Methods

Standard protocol approvals, registrations, and patient consents

All individuals provided written informed consent before their participation in this study for the clinical evaluation and genetic analysis of leukocyte DNA. Clinical phenotyping, whole-genome sequencing (WGS), and genetic analysis were approved by the Central Ethics Committee of the Bavarian Medical Association and the Ethics Review Panel of the University of Luxembourg.

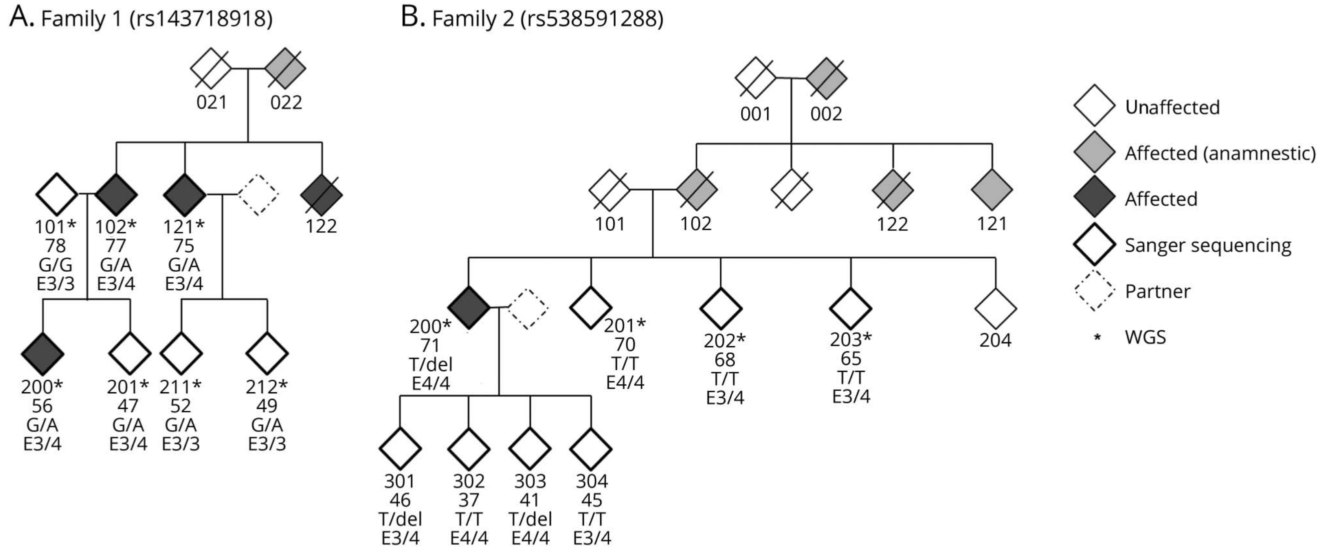
Patient information

Two families with an autosomal dominant inheritance pattern of AD were analyzed, and a pedigree chart is shown in figure 1. The 3 patients with AD sequenced in family 1 had reported ages at onset of <56, 70–75, and 71–77 years. The *APOE* status for all 3 patients was ε3/4. Family members 021, 022, and 122 died at the age of 47, 56, and 75 years, respectively. The patient with AD in family 2 had an age at onset of 66 years; the *APOE* status was ε4/4. Family members 101 and 102 died at the age of 74 and 73, respectively. For 001, 002, and 122, the age at death is unknown. Blood samples were taken from 7 (family 1) and 8 (family 2) family members, respectively, and DNA was extracted from leukocytes using standard procedures.

WGS and analysis

WGS was performed by Complete Genomics Inc. (CG, Mountain View, CA) using their proprietary paired-end, nanoarray-based sequencing-by-ligation technology.[16] Sequencing, quality control, mapping, and variant calling for the sequencing data were performed by CG as part of their sequencing service using the Standard Sequencing Service pipeline version 2.0. Sequencing reads were mapped against

Figure 1 Pedigree charts



The age at examination of each individual sequenced in this study is given beneath the identifier number. Individuals diagnosed with AD are indicated as affected (dark gray), individuals with AD-like symptoms reported by their family members are indicated in light gray. (A) Pedigree of family 1. The genotypes are wild type (G/G) or the alteration (G/A) that causes the *ABCA7* missense mutation. (B) Pedigree of family 2. The genotypes are wild type (T/T) or the alteration (T/del) that causes the *ABCA7* frameshift mutation.

NCBI Build 37. For further analysis, only single nucleotide variants (SNVs) and small insertions, deletions, and block substitutions up to a size of about 50 nt (indels) were used.

Variant prioritization

Variants were annotated by ANNOVAR[17] (version 2015 March 12) using the NCBI RefSeq release 60 and the Ensembl release 74 genome. As input for our family WGS analysis pipeline,[18] we first combined all variants from all genomes of every sequenced family member into the union of variants using CG analysis software (CGATOOLS, version 1.5) listvariant tool and the CG "var" files of all individuals per family as input. We used CGATOOLS testvariant to test each sample for the presence of each allele at every variant position from the union set of variants. We removed variants that were not called as high-quality calls (VQHIGH) in at least 1 individual. For both families, we used ISCA version 0.1.9[19] to search for shared haplotype blocks between pairs of samples and determined the number of shared alleles per block. For family 1, we filtered for haplotype blocks that shared 1 allele between the cases that was not shared with the unaffected individual 101 (figure 1A). For family 2, we excluded blocks where any pair of unaffected siblings shared 2 alleles (figure 1B). For each family, we applied an autosomal dominant inheritance model and filtered for exonic variants excluding synonymous variants and for variants in essential splice sites (±2 nucleotides from the exon boundary).

Variants within regions that are known to show very high mutation rates, like in mucins and olfactory receptors, were excluded (commonly mutated region).[20] We filtered for rare variants having an MAF of less than 5% in the European American population of the 1000 Genomes Project, the European NHLBI ESP exomes, and the Non-Finnish European population from the ExAC project as well as in the control data set CG69 provided by CG. We annotated the remaining variants for pathogenicity by considering either loss-of-function variants (indels, stop-gain, stop-loss, and splice-site variants) or missense mutations predicted to be deleterious by SIFT, PolyPhen-2_HDIV, LRT, and MutationTaster or mutated at highly conserved positions (GERP_RS>3). All annotations were derived from dbNSFP3.0a.[21] We further used a list of AD candidate genes that was collected from various GWAS in the dbGAP, the Alzgene database,[22] and the Genotator tool[23] to prioritize variants.

Population stratification

We performed population stratification by using EIGENSTRAT[24] with default parameters. First, we merged our data with the 1000 Genomes data. We chose only the autosomal SNVs concordant with hapmap[25] that were biallelic and not in linkage disequilibrium (LD) with each other by using PLINK (version 1.9)[26] with the parameters—indep 50 5 2, MAF of at least 0.1, and minimum call rate of 0.99 to perform the population stratification. To identify the ethnicity of samples in the current study, the first and the second principal components were visualized.

Genetic and linkage analysis

For linkage analysis, high-quality SNV positions (complete call rate over all individuals from VQHIGH status in CG var files) were extracted from the WGS data. Variants with high LD were removed using PLINK 1.9[26]; further thinning of variants was performed using mapthin.[27] A set of 2,000 variants per chromosome along with the identified variants segregating with the disease through the pedigree were used to check for genotype errors and mendelian inconsistencies using MERLIN[28] and were subsequently removed if they were identified as errors. The remaining variants were used for linkage analysis and their genomic positions were linearly interpolated based on the hapmap genetic map (2011-01_phaseII_B37). MERLIN was used to perform both haplotyping and multipoint parametric linkage analysis with a rare autosomal dominant disease model with a disease frequency of 0.0001 and penetrance of 0.0001, 1.0, and 1.0. Haplotyping results were visualized using HaploPainter.[29] Using the R package "paramlink,"[30] we calculated the power of each pedigree given as the maximal LOD scores for each family under an autosomal dominant inheritance model and 10,000 simulated markers. Relationship detection between all individuals was performed using software GRAB.[20]

Validation by sanger sequencing

The presence of both variants identified by WGS were validated and replicated by Sanger sequencing in each family member of both pedigrees using standard protocols with the following oligonucleotide sets: ABCA7_delT_1055908_FWD: 3′-TTGTCCACCCTTGACTCTGTGC-5′; ABCA7_delT_1055908_REV: 3′-CTTGAGACTGTCCTGAGCATCC-5′; ABCA7_rs143718918_FWD: 3′-ACAGGTCCATCTTGAGTGGC-5′; ABCA7_rs143718918_REV: 3′-GAGACCAGCCCCACATCC-5′.

Results

We used WGS to identify the genetic cause of AD in 10 families with an autosomal dominant pattern of inheritance. Among these families, variants in *ABCA7* were identified in 2 families (family 1 and family 2). In family 1, we sequenced the genomes of 7 family members; 3 of them were diagnosed with AD (figure 1A). In the second family (family 2), we sequenced genomes of 4 family members, 1 affected index patient with AD (age at diagnosis 66 years) and 3 unaffected siblings (figure 1B). Relationship estimation[20] confirmed all relationships in both families, given the original pedigree information. All families were self-reported of German ethnicity. European ancestry could be confirmed using EIGENSTRAT[24] analysis (figure e-1, links.lww.com/NXG/A38).

WGS fully called on average 97% of genome and 98% of exonic regions. Seventy-seven percent of the genome and 86% of the exome were covered with at least 30X. We detected on average over all samples from both families 3,415,106 SNVs and 577,534 indels and substitutions per genome (table e-1, links.lww.com/NXG/A39).

In total, 7,516,717 and 6,139,540 variants (SNVs and indels) different from the reference genome were identified in at least 1 family member for family 1 and 2, respectively. Disease-associated variants were searched using an autosomal dominant inheritance model. In family 2, only variants present in the affected and not present in the unaffected individuals were considered. After strict quality, mode of inheritance, shared haplotype and MAF < 0.05 filtering, 51,269 and 56,962 variants remained (table e-2, links.lww.com/NXG/A39). After annotation using RefSeq, we screened for exonic splice-site affecting variants. After excluding variants within commonly mutated and brain-expressed genes, we prioritized the remaining variants according to their predicted pathogenicity and conservation. In total, only 11 (family 1, tables e-3 and e-5) and 8 (family 2, tables e-4 and e-6) variants were found in AD-related genes (table e-7) and were therefore considered to be relevant to AD. Strict variant filtering revealed for each family rare *ABCA7* variants, rs143718918 (family 1) and rs538591288 (family 2) as best candidate variants.

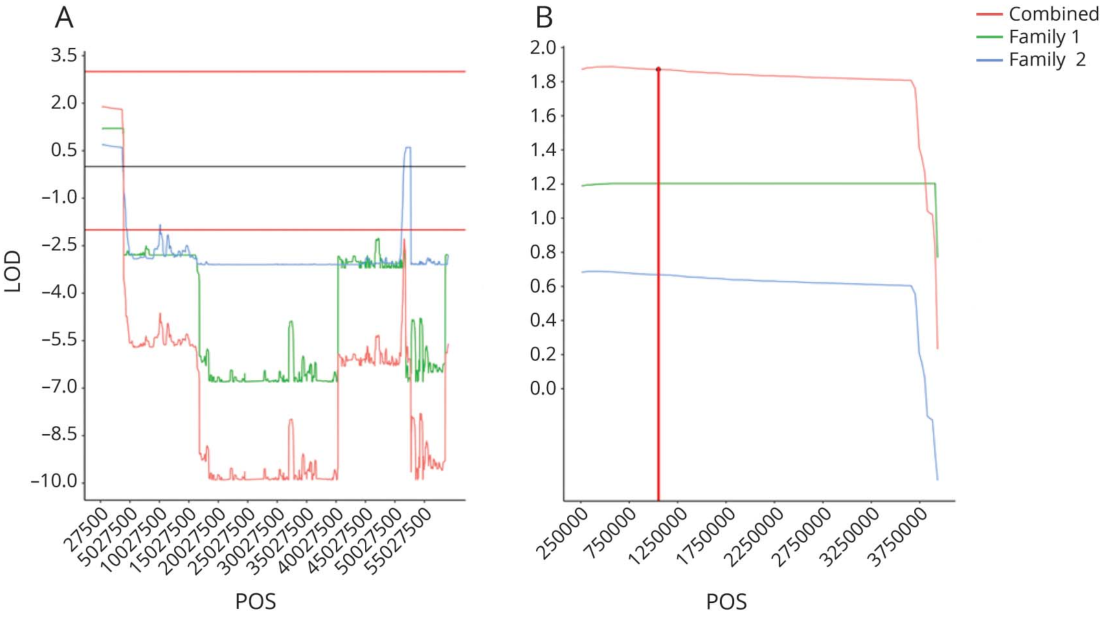
European (Non-Finnish) population allele frequencies for rs143718918 and for rs538591288 add up to 0.0021 and 0.0016, respectively. Both variants are very rare (MAF < 0.01) in the European population according to the ExAC[31] and 1000 Genomes (tables e-3 and e-4, links.lww.com/NXG/A39).

In addition, the performed linkage (figure 2, A and B) and haplotype block analysis (figure 3, A and B) show cosegregation and association of both variants with AD in both German families. We confirmed the presence of both variants in the initially screened and the additional family members by Sanger Sequencing (figure 1).

The SNV rs143718918 identified in family 1 causes a missense mutation of *ABCA7* (c.2693G>A) that affects the ABC1 domain of the protein (p.R880Q). This variant was previously identified in patients with AD and controls of a larger Belgian cohort in a French as well as in an European cohort with early-onset patients and in patients with AD of a Caucasian cohort.[7,32,33] We identified the SNV in all sequenced family members except for 1 healthy member (figure 1A). Three family members (201, 211, and 212) also carrying the risk variant were not affected and/or did not report cognitive deficits at the time of the last consultation, but were considerably younger than the affected family members and therefore possibly presymptomatic at the time of examination. As such, genetic counseling and clinical follow-up examinations will be conducted.

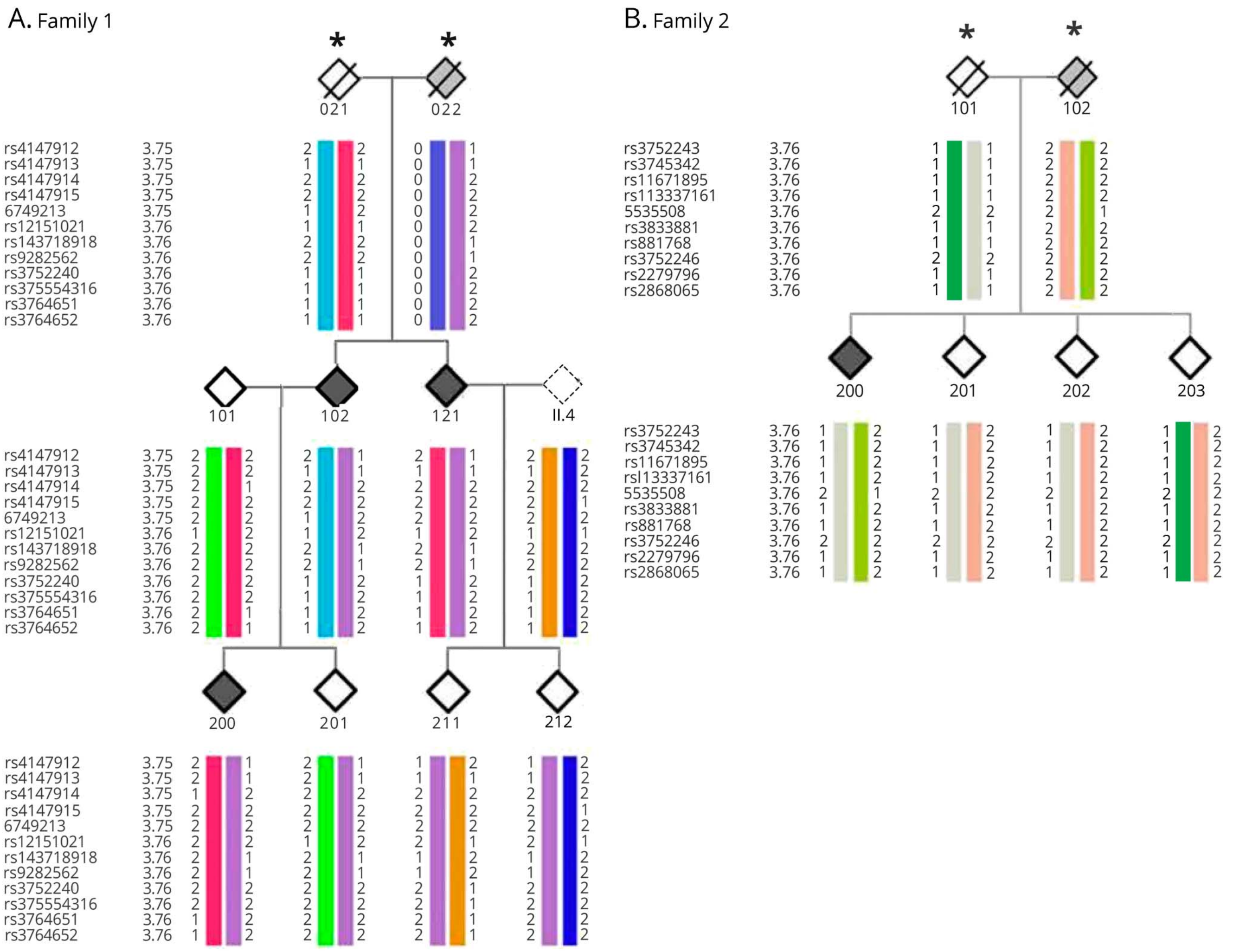
The second variant (rs538591288) identified in family 2 causes a frameshift deletion in exon 31 of *ABCA7* (c.4208delT; p.L1403fs). This variant was also previously identified in patients with AD and controls of a larger Belgian cohort as well as in a French and in a European Cohort with early-onset AD patients.[7,32,33] Of interest, in one of these studies, additional Italian relatives with EOAD carrying the deletion were reported.[32] Furthermore, 2 groups have recently shown that p.L1403fs variant carriers

Figure 2 Linkage analysis of chromosome 19



(A) The maximum LOD score (1.8) over the whole chromosome is seen in the region containing *ABCA7*. (B) Linkage analysis of the *ABCA7* region on chromosome 19. The maximal LOD score (1.8) could be found on chromosome 19 in the region from 257,507 to 3,909,104 suggesting linkage, the region in red spans the gene *ABCA7*. Of the combined LOD score of 1.8 in the region spanning *ABCA7*, family 1 and family 2 contributed LOD scores of 1.2 and 0.6, respectively.

Figure 3 Segregating haplotype blocks



The affected, unaffected, and disease status of unknown individuals are filled in black, white, and gray, respectively. An asterisk indicates the individuals who were not sequenced and their haplotypes were inferred. (A) The disease haplotype is indicated in purple. (B) The disease haplotype is indicated in light green. In both families, cosegregation of the disease haplotype including the corresponding *ABCA7* variant can be seen in all affected individuals.

had decreased ABCA7 protein levels but unchanged mRNA levels.[32,34]

We have identified the SNV (rs538591288) in 3 family members, including 2 children of the index patient (301 and 303, figure 1B), which were not diagnosed with AD but due to young age possibly presymptomatic at the time of examination. Of interest, both so far unaffected carriers reported already having occasional memory problems.

Discussion

We conducted a whole-genome sequencing (WGS) study to search for SNVs cosegregating with Alzheimer disease (AD) cases in German families. Of interest, we identified 2 rare variants of *ABCA7* possibly contributing to AD pathogenesis in 2 families, respectively. *ABCA7* is one of more than 20 AD risk loci that have so far been identified by GWASs and sequencing studies. ABCA7 is also involved in AD-relevant pathways (lipid metabolism, microglial phagocytosis, and altered amyloid-beta processing) and abundantly expressed in the brain.

We identified the rs143718918 to cosegregate with AD in family 1. The SNV causes a missense mutation of *ABCA7* in exon 19 (c.2693G>A) that affects the ABC1 domain of the protein (p.R880Q) and is probably damaging. This variant has previously been identified by GWASs in Caucasians with late-onset AD.[8] Furthermore, several studies reported the presence of this variant in AD and in control subjects of (1) a Belgian cohort,[7] (2) a French EOAD cohort,[33] and (3) an EOAD cohort including samples of diverse origin.[32] Overall, the variant was present with higher frequency in AD cases compared with controls.

The rs538591288 cosegregated with AD in family 2 and causes a frameshift mutation in exon 31 of *ABCA7* (c.4208delT; p.L1403fs). Initially, this SNV has been reported in an Icelandic cohort[9] and was later also identified with higher frequency in cases than controls of German, Swedish, Italian,[32] French,[33] and Belgian[7] cohorts. Mutation carriers express lower levels of full-length ABCA7 protein with unchanged mRNA expression levels.[32,34] However, it has been reported that by in-frame exon skipping of the premature termination codon bearing exon 31, the transcript escapes nonsense-mediated mRNA decay.[7,32] Exon skipping leads to the production of a shorter version of ABCA7 protein, which might partly compensate for the reduced full-length protein levels and might cause incomplete penetrance of rs538591288.

It has to be mentioned that we cannot exclude that other variations might cause additive effects on the development of AD in both families. Because of the previously shown involvement of ABCA7 in AD, the presented variants represent the most promising candidates. Together, our results support the notion that rare variants of *ABCA7* exert considerable risk to the development of AD.

Author contributions

Study concept and design: P. May, J.G. Schneider, and M. Riemenschneider. Acquisition and analysis of the data: all authors. Collection, analysis, and interpretation the data: P. May, S. Pichler, D. Hartl, D.R. Bobbili, and C. Spaniol. Drafting the manuscript and/or figure: all authors.

Acknowledgment

The authors are grateful to the probands and their families for contributing to this study. They thank the ISB Family Genomics Group for their support and project management of WGS data. They also thank Katrin Williams for excellent technical assistance.

Study funding

Financial support was provided by the University of Luxembourg (UL)–Institute for Systems Biology (ISB) Strategic Partnership by "le plan Technologies de la Santé par le Gouvernement du Grand-Duché de Luxembourg" through the Luxembourg Centre for Systems Biomedicine and the UL to P. May, from the Saarland University (Homfor 2014 to D. Hartl), the German Federal Ministry of Education and Research (BMBF) National Genome Research Network (NGFN) grant No. 01GS08125 to M. Riemenschneider, and through the Helmholtz Alliance for Mental Health in an Aging Society (HELMA) Grant No. Ha-15 to M. Riemenschneider. Computational results presented in this article were carried out using the HPC facility of the UL (hpc.uni.lu).

Disclosure

P. May, S. Pichler, D. Hartl, and D.R. Bobbili report no disclosures. M. Mayhaus has been employed by Evotec AG, Hamburg, Germany. C. Spaniol reports no disclosures. A. Kurz has served on the scientific advisory board of MSD Germany; has served on the editorial board of *Der Nervenarzt*; and has received research support from the Federal Ministry of Health, the Federal Ministry of Education and Research, and the German Academic Exchange Service. R. Balling has served on the scientific advisory board of the Novo Nordisk Foundation; has received funding for travel and/or speaker honoraria from Bayer; has served on the editorial board of *Annual Review Nutrition*; and has held stock, stock options, and/or board of director's compensation from Megeno SARL and ITTM. J.G. Schneider has received research support from the FNR. M. Riemenschneider has received research support from the BMBF, NGFN, and the Helmholtz Alliance for Mental Health in an Aging Society. Full disclosure form information provided by the authors is available with the full text of this article at Neurology.org/NG.

Received July 31, 2017. Accepted in final form February 13, 2018.

References

1. Hollingworth P, Harold D, Sims R, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet 2011;43:429–435.
2. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 2013;45:1452–1458.
3. Logue MW. A comprehensive genetic association study of Alzheimer disease in African Americans. Arch Neurol 2011;68:1569.
4. Takahashi K, Kimura Y, Nagata K, Yamamoto A, Matsuo M, Ueda K. ABC proteins: key molecules for lipid homeostasis. Med Mol Morphol 2005;38:2–12.
5. Sakae N, Liu CC, Shinohara M, et al. ABCA7 deficiency accelerates amyloid-β generation and Alzheimer's neuronal pathology. J Neurosci 2016;36:3848–3859.
6. Li H, Karl T, Garner B. Understanding the function of ABCA7 in Alzheimer's disease. Biochem Soc Trans 2015;43:920–923.
7. Cuyvers E, De Roeck A, Van den Bossche T, et al. Mutations in ABCA7 in a Belgian cohort of Alzheimer's disease patients: a targeted resequencing study. Lancet Neurol 2015;14:814–822.
8. Vardarajan BN, Ghani M, Kahn A, et al. Rare coding mutations identified by sequencing of Alzheimer disease genome-wide association studies loci. Ann Neurol 2015;78:487–498.
9. Steinberg S, Stefansson H, Jonsson T, et al. Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. Nat Genet 2015;47:445–447.
10. Bellenguez C, Charbonnier C, Grenier-Boley B, et al. Contribution to Alzheimer's disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779 cases and 1273 controls. Neurobiol Aging 2017;59:220.e1–220.e9.
11. Kunkle BW, Carney RM, Kohli MA, et al. Targeted sequencing of ABCA7 identifies splicing, stop-gain and intronic risk variants for Alzheimer disease. Neurosci Lett 2017;649:124–129.
12. Reitz C, Jun G, Naj A, et al. Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E ε4, and the risk of late-onset Alzheimer disease in african americans. JAMA 2013;309:1483.
13. Yang P, Sun YM, Liu ZJ, et al. Association study of ABCA7 and NPC1 polymorphisms with Alzheimer's disease in Chinese Han ethnic population. Psychiatr Genet 2013;23:268.
14. Chung SJ, Lee JH, Kim SY, et al. Association of GWAS top hits with late-onset Alzheimer disease in Korean population. Alzheimer Dis Assoc Disord 2012;27:1.
15. Sassi C, Nalls MA, Ridge PG, et al. Neurobiology of aging ABCA7 p.G215S as potential protective factor for Alzheimer' s disease. Neurobiol Aging 2016;46:235.e1–235.e9.
16. Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 2010;327:78–81.
17. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.
18. Schubert J, Siekierska A, Langlois M, et al. Mutations in STX1B, encoding a presynaptic protein, cause fever-associated epilepsy syndromes. Nat Genet 2014;46:1327–1332.
19. Roach JC, Glusman G, Smit AF, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 2010;328:636–639.
20. Li H, Glusman G, Huff C, Caballero J, Roach JC. Accurate and robust prediction of genetic relationship from whole-genome sequences. PLoS One 2014;9:e85437.
21. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 2011;32:894–899.
22. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet 2007;39:17–23.
23. Wall DP, Pivovarov R, Tong M, et al. Genotator: a disease-agnostic tool for genetic annotation of disease. BMC Med Genomics 2010;3:50.
24. Price AL, Patterson NJ, Plenge RM. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909.
25. International HapMap Consortium. The international HapMap project. Nature 2003;426:789–796.

26. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2014;4:7.
27. Howey R, Cordell HJ. MapThin: thinning your map files for linkage analyses! 2011. Available at: staff.ncl.ac.uk/richard.howey/mapthin/mapthin.pdf. Accessed July 28, 2017.
28. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002;30:97–101.
29. Thiele H, Nürnberg P. HaploPainter: a tool for drawing pedigrees with complex haplotypes. Bioinformatics 2005;21:1730–1732.
30. Egeland T, Pinto N, Vigeland MD. A general approach to power calculation for relationship testing. Forensic Sci Int Genet 2014;9:186–190.
31. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285–291.
32. De Roeck A, Van den Bossche T, van der Zee J, et al. Deleterious ABCA7 mutations and transcript rescue mechanisms in early onset Alzheimer's disease. Acta Neuropathol 2017;134:475–487.
33. Le Guennec K, Nicolas G, Quenez O, et al. ABCA7 rare variants and Alzheimer disease risk. Neurology 2016;86:2134–2137.
34. Allen M, Lincoln SJ, Corda M, et al. *ABCA7* loss-of-function variants, expression, and neurologic disease risk. Neurol Genet 2017;3:e126.

Neurology® Genetics

Rare *ABCA7* variants in 2 German families with Alzheimer disease

Patrick May, Sabrina Pichler, Daniela Hartl, et al.
Neurol Genet 2018;4;
DOI 10.1212/NXG.0000000000000224

This information is current as of March 21, 2018

Updated Information & Services	including high resolution figures, can be found at: http://ng.neurology.org/content/4/2/e224.full.html
References	This article cites 33 articles, 5 of which you can access for free at: http://ng.neurology.org/content/4/2/e224.full.html##ref-list-1
Subspecialty Collections	This article, along with others on similar topics, appears in the following collection(s): **Alzheimer's disease** http://ng.neurology.org//cgi/collection/alzheimers_disease
Permissions & Licensing	Information about reproducing this article in parts (figures,tables) or in its entirety can be found online at: http://ng.neurology.org/misc/about.xhtml#permissions
Reprints	Information about ordering reprints can be found online: http://ng.neurology.org/misc/addir.xhtml#reprintsus

Neurol Genet is an official journal of the American Academy of Neurology. Published since April 2015, it is an open-access, online-only, continuous publication journal. Copyright Copyright © 2018 The Author(s). Published by Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology.. All rights reserved. Online ISSN: 2376-7839.

AMERICAN ACADEMY OF NEUROLOGY®

Community-Reviewed Biological Network Models for Toxicology and Drug Discovery Applications

The sbv IMPROVER project team and challenge best performers: Aishwarya Alex Namasivayam[1], Alejandro Ferreiro Morales[2], Ángela María Fajardo Lacave[2], Aravind Tallam[11], Borislav Simovic[10], David Garrido Alfaro[2], Dheeraj Reddy Bobbili[1], Florian Martin[3], Ganna Androsova[1], Irina Shvydchenko[9], Jennifer Park[7], Jorge Val Calvo[12], Julia Hoeng[3], Manuel C. Peitsch[3], Manuel González Vélez Racero[2], Maria Biryukov[1], Marja Talikka[3], Modesto Berraquero Pérez[2], Neha Rohatgi[8], Noberto Díaz-Díaz[2], Rajesh Mandarapu[5], Rubén Amián Ruiz[2], Sergey Davidyan[13], Shaman Narayanasamy[1], Stéphanie Boué[3], Svetlana Guryanova[4], Susana Martínez Arbas[1], Swapna Menon[6], and Yang Xiang[3]

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Campus Belval, Esch-sur-Alzette, Luxembourg, [2]Pablo de Olavide University, Ctra. de Utrera, Seville, Spain. [3]Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud, Neuchâtel, Switzerland (part of Philip Morris International group of companies). [4]Institute of Bioorganic Chemistry Russian Academy of Sciences, Moscow, Russia. [5]Prakhya Research Laboratories, Lakshminagar, Selaiyur, Chennai, Tamil Nadu, India. [6]AnalyzeDat Consulting Services, Edapally Byepass Junction, Kochi, Kerala, India. [7]Selventa, Alewife Center, Cambridge, MA, USA. [8]Center for Systems Biology, University of Iceland, Reykjavik, Iceland. [9]Kuban State University of Physical Education, Sport and Tourism, Krasnodar, Russia. [10]FM Pharm, Ltd., Sencanski put bb, Subotica, Serbia. [11]TWINCORE, Zentrum für Experimentelle und Klinische Infektionsforschung, Hannover, Germany. [12]Center for Molecular Biology, "Severo Ochoa" – Spanish National Research Council, Madrid, Spain. [13]Institute of Biochemical Physics Russian Academy of Sciences, Moscow, Russia.

ABSTRACT: Biological network models offer a framework for understanding disease by describing the relationships between the mechanisms involved in the regulation of biological processes. Crowdsourcing can efficiently gather feedback from a wide audience with varying expertise. In the Network Verification Challenge, scientists verified and enhanced a set of 46 biological networks relevant to lung and chronic obstructive pulmonary disease. The networks were built using Biological Expression Language and contain detailed information for each node and edge, including supporting evidence from the literature. Network scoring of public transcriptomics data inferred perturbation of a subset of mechanisms and networks that matched the measured outcomes. These results, based on a computable network approach, can be used to identify novel mechanisms activated in disease, quantitatively compare different treatments and time points, and allow for assessment of data with low signal. These networks are periodically verified by the crowd to maintain an up-to-date suite of networks for toxicology and drug discovery applications.

KEYWORDS: biological network, crowdsourcing, COPD, drug discovery, toxicology

CITATION: Namasivayam et al. Community-Reviewed Biological Network Models for Toxicology and Drug Discovery Applications. *Gene Regulation and Systems Biology* 2016:10 51–66 doi: 10.4137/GRSB.S39076.

TYPE: Original Research

RECEIVED: February 04, 2016. **RESUBMITTED:** March 31, 2016. **ACCEPTED FOR PUBLICATION:** April 12, 2016.

ACADEMIC EDITOR: Li-Na Wei, Editorial Board member

PEER REVIEW: Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,029 words, excluding any confidential comments to the academic editor.

FUNDING: The Network Verification Challenge was funded by Philip Morris International. The research described in this article was funded by Philip Morris International in a collaborative project with Selventa. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Selventa and Philip Morris International authors performed this work under a joint research collaboration. SB, JH, FM, MCP, MT, and YX are employees of Philip Morris International. Philip Morris International is the sole source of funding and the sponsor of this project. DGA could not be contacted to provide a statement of his potential conflicts of interest.

CORRESPONDENCE: Julia.Hoeng@pmi.com

COPYRIGHT: © the authors, publisher and licensee Libertas Academica Limited. This is an open-access article distributed under the terms of the Creative Commons CC-BY-NC 3.0 License.

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Chronic obstructive pulmonary disease (COPD) is a progressive chronic inflammatory lung disease characterized by persistent limited airflow caused by various environmental exposures such as cigarette smoke (CS), occupational hazards, and air pollution.[1] Mechanisms underlying the disease include a complex interplay of inflammation, proliferation, oxidative stress, tissue repair, and other processes driven by various immune, epithelial, and airway cell types.[2,3] Understanding the molecular mechanisms associated with COPD is important for preventing disease onset, slowing down disease progression, and managing treatment. Biological network models offer a framework for understanding disease by describing the relationships between the molecular mechanisms involved in the regulation of a particular biological process. Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome are open access pathway databases widely used by the scientific community.[4–7] They describe signaling in various areas of biology and can be used to interpret large-scale molecular data through integration and overlay on pathways to assess pathway overrepresentation. In contrast to these general pathway databases, we have developed a set of networks within defined boundaries relevant to COPD that are available to the public on the Bionet website at https://bionet.sbvimprover.com, where they can not only be viewed and downloaded but can also be actively commented on and edited.[8,9] These networks can also be used to interpret large-scale molecular data to a fine-grained degree, due to their construction in Biological Expression Language (BEL), a human-readable computable language with the ability to capture precise biological information and associated context (www.openbel.org). The networks were based on a set of previously published lung-relevant healthy biological networks, which along with the most current network versions are available for download at http://www.causalbionet.com/.[10–14]

To ensure a comprehensive and up-to-date set of biological network models that cover a wide range of biological signaling, crowdsourcing can be used to gather input from the scientific community. Crowdsourcing is a powerful tool to efficiently gather feedback from a wide audience that covers expertise in many biological areas. Crowdsourcing efforts in biology are useful in the collection of creative solutions to challenging problems in various fields of biology such as signaling networks, protein folding, RNA design, and sequence alignment.[15–18] Crowdsourcing has also been harnessed to accomplish a large amount of manual work in annotation projects including disease-related genes, interactome pathways, and PubMed abstracts.[19–21] We have previously reported the creation of a set of biological networks describing COPD processes that were improved by the scientific community during the first Network Verification Challenge (NVC).[8,9] In this study, we show that the networks were further improved during a second NVC (NVC2), in which the crowd added mechanistic details in the form of new nodes and edges.

We illustrate possible network applications for the crowd-improved set of networks using network scoring by TopoNPA, a method to infer mechanism and network perturbation based on transcriptomics data and known activators and inhibitors of gene expression reported in the literature.[22] Quantitative scoring of networks is enabled by BEL, an open platform technology, where cause and effect relationships from the literature are described and annotated using a precise language and collected in a knowledgebase. This knowledgebase is used to predict upstream regulators of measured transcriptomics data.[23] This backward reasoning approach differs from other gene set enrichment approaches using gene sets defined as KEGG pathways or Gene Ontology (GO) classes for example,[24] which make the assumption that RNA expression is equivalent to protein activity. Another limitation of methods such as gene set enrichment analysis (GSEA)[25] is that they do not take direction into account for each gene within the gene set. TopoNPA scoring of networks allows for quantitative scoring of inferred mechanisms and networks based on signed fold changes in the dataset. Using TopoNPA on a set of networks enables quantitative comparison between different compounds, disease subtypes, or other perturbations of interest.[22] We describe here one application for how the improved set of 46 computable BEL-encoded NVC network models can be used by the scientific community for toxicology and drug discovery applications.

Materials and Methods

Biological expression language. BEL is a triple-based language, where statements consist of two biological entities connected by a relationship (for causal statements: cause, relationship, and effect). The BEL framework, including a database of BEL statements and other tools to be used with BEL, is an open-platform technology available for download at http://www.openbel.org/. BEL captures specific entities from chemicals to proteins to biological processes and relationship links that are directional, providing information on activation or inhibition. Statements within BEL are derived from the published literature and are compiled together to express the existing causal knowledge in a graph-based, computable format. These entities connected by relationships are represented as nodes and edges within a BEL graph network and are linked to metadata such as literature support, which contains PubMed ID, tissue, disease, cell type, and species. A BEL node consists of a function, namespace, and entity. The function gives information about the type of entity (eg, abundance and activity), and the namespace is a standardized ontology that defines the entity that each node represents (eg, MeSH, ChEBI, GO, and HGNC). See Supplementary File 1 for a list of BEL functions and namespaces. Just as the networks are continuously improved by the crowd, the BEL language evolves based on suggestions made by the OpenBEL community. Namespaces in the NVC networks version 2.0 reported here were updated from v1.0 BEL Namespaces to the most recent version (v20150611), which includes additional and refined namespaces.

Network Building

Networks were constructed in a three-phase process, as described previously.[8] Briefly, networks were built using data and literature during Phase 1, enhanced with lung- and COPD-relevant mechanisms (represented by nodes in the networks) by the crowd during Phase 2 on the Bionet website (https://bionet.sbvimprover.com/), and discussed during a jamboree meeting during Phase 3 in which the best performers were invited based on their point totals from the online phase. Networks with high crowd activity or interest were selected for discussion during the jamboree. Phases 2 and 3 were repeated in NVC2. Fifteen networks were discussed during the NVC1 jamboree (apoptosis, cell cycle, dendritic cell signaling, growth factor, hypoxic stress, macrophage signaling, neutrophil signaling, NFE2L2 signaling, nuclear receptors, oxidative stress, response to DNA damage, mechanisms of cellular senescence, Th1 signaling, Th2 signaling [Th1–Th2 signaling were merged as a result of the jamboree], and xenobiotic metabolism response) and nine networks were discussed during the NVC2 jamboree (calcium, epigenetics, macrophage signaling, necroptosis, neutrophil signaling, oxidative stress, senescence, Th1–Th2 signaling, and xenobiotic metabolism response). After the NVC2, it was decided to merge the four senescence-related models (mechanisms of cellular senescence, regulation of CDKN2A expression, regulation by tumor suppressors, and transcriptional regulation of the SASP) into one model called senescence. In both NVC1 and NVC2, changes were implemented by the organizers and new versions were uploaded to the Bionet website. The latest versions edited after the NVC2 jamboree are the version 2.0 networks.

Network Statistics

Network statistics and metrics were calculated on the networks presented to the crowd at the start of the NVC (v1.1) and on the most recent networks containing the outcomes of NVC1 and NVC2 (v2.0). Basic network metrics such as number of nodes, edges, activation edges, inhibition edges, and the proportion of inhibition edges were calculated. In addition to these basic network characteristics, the following metrics were computed:

- Mean degree: the average of node degrees. This metric informs the overall topology of the network. A low average degree (<2) is typically observed in linear networks.
- Max degree: the maximum degree in the network, representing the size of the largest hub.
- Mean node betweenness (MNB) or betweenness centrality: the number of shortest paths between pairs of other nodes that go through that node. Nodes with high betweenness centrality are considered as high trafficking nodes. This metric characterizes the centrality of the nodes and hence the topology of the networks (for example, bottlenecks for the paths in the network). A complete graph has a vanishing ($=0$) MNB.
- Largest clique size: the number of nodes in the largest complete undirected subgraph in a network. This number is expected to be low because complete subgraphs that are not triangles are not expected to be biologically meaningful.
- Mean path length (MPL): the average of the shortest path length between all pairs of nodes. This metric gives an indication of the density of the network. A low MPL characterizes networks for which the shortest path of causal statements, from one node to another, are made of few edges; for example, in a complete graph, this equals 1. It does not necessarily imply that the mean degree is high. A typical cascading signaling pathway with little feedback would be expected to have a high MPL.
- Frustration: the minimum number of edges that should be removed to make the network balanced. Balance in a signed graph is characterized by the property that every path between two nodes has the same sign (the sign of a path is the product of its edge signs). Equivalently, a graph is balanced if and only if every cycle is positive. A negative feedback loop contributes to the network frustration. For example, tightly regulated processes such as cell cycle or apoptosis are expected to have a high frustration metric.
- # connected components: number of connected components, that is, the number of disjoint (ie, not sharing any edge) subnetworks within the network.

For all of these network metrics, the differences between the pre-NVC networks (v1.1) and post-NVC2 networks (v2.0) were calculated to understand crowd contribution effects on the networks. For the Th1–Th2 signaling and senescence networks, both of which were integrated from separate networks following jamboree discussions, the individual pre-NVC networks (v1.1) were combined for comparison with the already combined post-NVC2 networks (v2.0).

Datasets Analysis

The three datasets that were analyzed are shown in Table 1.

Network perturbation amplitude. The Network Perturbation Amplitude (NPA) methodology aims at contextualizing high-dimensional transcriptomics data by combining gene expression ($\log_2$) fold-changes into fewer differential node values (one value for each node of the network), representing a biological entity (mechanism, chemical, biological process).[22,26,27] A node can be inferred as increased or decreased based on gene expression data, because there are signed relationships (increase or decrease) between the node and downstream mRNA abundance entities.[23,27] The differential node values are determined by a fitting procedure that infers values that best satisfy the directionality of the causal relationships (positive or negative signs) contained in the network model, while being constrained by the experimental data (the gene $\log_2$-fold-changes, which are described as downstream effects of the network itself).

The differential values of the network are then used to calculate a score for the network as a whole, called the TopoNPA score.[22] For these network scores, a confidence interval accounting for the experimental variation and the associated *P*-value are computed. In addition, companion statistics are derived to inform the specificity of the TopoNPA score with respect to the biology described in the network model. These are depicted as *O and K* if their *P*-values are below the significance level (0.05). A network is considered to be significantly impacted if all three values (the *P*-value for experimental variation, *O, and K* statistics) are below 0.05.[22]

Leading nodes are the main contributors to the network score, making up 80% of the TopoNPA score. These nodes can be useful for interpreting the data to predict mechanisms that might be driving the biological process that the network represents.[22]

Table 1. Dataset overview.

DATA ID[a]	TISSUE	TREATMENT	ENDPOINT
GSE28464	Human fibroblasts	Oncogenic Ras (H-RasV12) expression 4 days	Model of senescence; autophagic markers
E-MTAB-3150	Mouse lung	Reference cigarette (3R4F) smoke, prototype modified risk tobacco product (pMRTP), switch, cessation for 7 months	Lung function; Immune cell numbers and inflammatory markers in bronchoalveolar lavage fluid (BALF); lung macrophage counts; pulmonary morphometry
GSE52509	Mouse lung	Reference cigarette (3R4F) smoke for 4, 6 months	B and T-cell counts and histology in lung; immune markers in bronchoalveolar lavage (BAL) and lung

Notes: [a]The GSE datasets are from the NCBI GEO database and the E-MTAB dataset is from the EMBL-EBI ArrayExpress database.

To increase the specificity and relevance of node scores and network scores, we consider only the nodes in the network that are bounded by experimental evidence in the following sense: for any given node, at least one ancestor node (ie, a node from which a directed path to the node under consideration exists) and at least one child node (ie, a node to which a directed path from the node under consideration exists) in the directed graph must have downstream RNA abundance nodes: their values can be directly inferred based on experimental mRNA data. After removing the nodes that do not satisfy the above criteria, the largest connected component is kept (if the resulting network is not connected). Finally, the "causeNoChange" edges are disregarded for scoring. Selections of these simplified networks that have been scored using these criteria are shown in the results.

Results

Network resource comparison. We previously described novel aspects of the NVC networks compared with other network resources.[8,9] Herein, we select a particular network, calcium signaling, to further illustrate the differences between the NVC networks constructed using BEL (https://bionet.sbvimprover.com) and the pathways available in the KEGG (http://www.genome.jp/kegg/pathway.html) and Reactome Pathway Databases (http://www.reactome.org) (Fig. 1).

Network boundaries. The NVC Calcium Network (v2.0) is an example of a network with similar content and size as the KEGG Calcium Signaling pathway map (map04020) and Reactome Calmodulin pathway (R-HSA–111997.1). All three networks describe the increase of calcium as a result of inositol 1,4,5-triphosphate activation (Fig. 1, box 1 highlighted in yellow) and the role of calcium in activating calmodulin kinase (CAMK) (Fig. 1, box 2 highlighted in yellow). However, the BEL network was constructed specifically to describe calcium signaling that leads to cell proliferation in the lung, while the KEGG and Reactome pathways describe calcium signaling in a more general manner that is tissue agnostic and can lead to proliferation as well as, for example, contraction, metabolism, apoptosis, and exocytosis in the KEGG pathway.

Network resource comparison. The NVC Calcium Network (v2.0) contains 47 nodes (35 unique concepts when genes, proteins, and activity nodes are flattened together) and 52 edges, the KEGG pathway map contains 48 nodes/unique concepts and 60 edges, and the Reactome pathway contains 46 nodes (34 unique concepts) and 49 edges (Table 2). The NVC2 network is supported by 38 unique literature references for specific edges, while there are 20 references for the KEGG pathway and 28 references for the Reactome pathways. There is no overlap in references between the three resources and the average date of publication for the NVC2 references is 2006, whereas the KEGG and Reactome average dates are 2002 and 2000, respectively. The NVC2 and Reactome references support a particular edge, whereas the KEGG references are not specific to a particular edge. The NVC2 network contains multiple node functions such as abundance, activities, and phosphorylations that have been specifically tested in the literature, while the KEGG pathway depicts a single layer of gene symbol nodes that could represent RNAs, proteins, modified proteins, or protein activities. Reactome contains nodes that reflect activities and phosphorylations that can be repeated throughout the diagram to indicate location.

The cellular localization graphics in KEGG and Reactome give a second layer of information, with inositol 1,4,5-triphosphate (IP3 in KEGG, I(1,4,5)P3) in Reactome activating inositol 1,4,5-trisphosphate receptor (IP3R) depicted on the endoplasmic reticulum (ER) membrane, increasing calcium in the cytoplasm (Fig. 1, box 1 highlighted in yellow). From the KEGG and Reactome diagrams, IP3R/IP3 receptor can be inferred to be a calcium channel transporting calcium across the ER, although it is not explicitly stated. In BEL, this relationship is described explicitly in the NVC network as three different family members defined by the HUGO Gene Nomenclature Committee (HGNC) database (http://www.genenames.org/) with transporter activities (tport): tport(p(HGNC:ITPR1)), tport(p(HGNC:ITPR2)), and tport(p(HGNC:ITPR3)) that activate the bp(GOBP:"store-operated calcium entry") node defined by the GO biological process database.[28] The nodes in the NVC network have more granularity than the Reactome and KEGG networks, specifying the type of activity and particular residues that are phosphorylated.

Along with the IP3 receptor, another process that is described by all three network resources is CAMK activation by calcium (Fig. 1, box 2 highlighted in yellow), although the NVC2 network describes CAMK2 while KEGG and Reactome pathways describe CAMK4 (only obvious for the

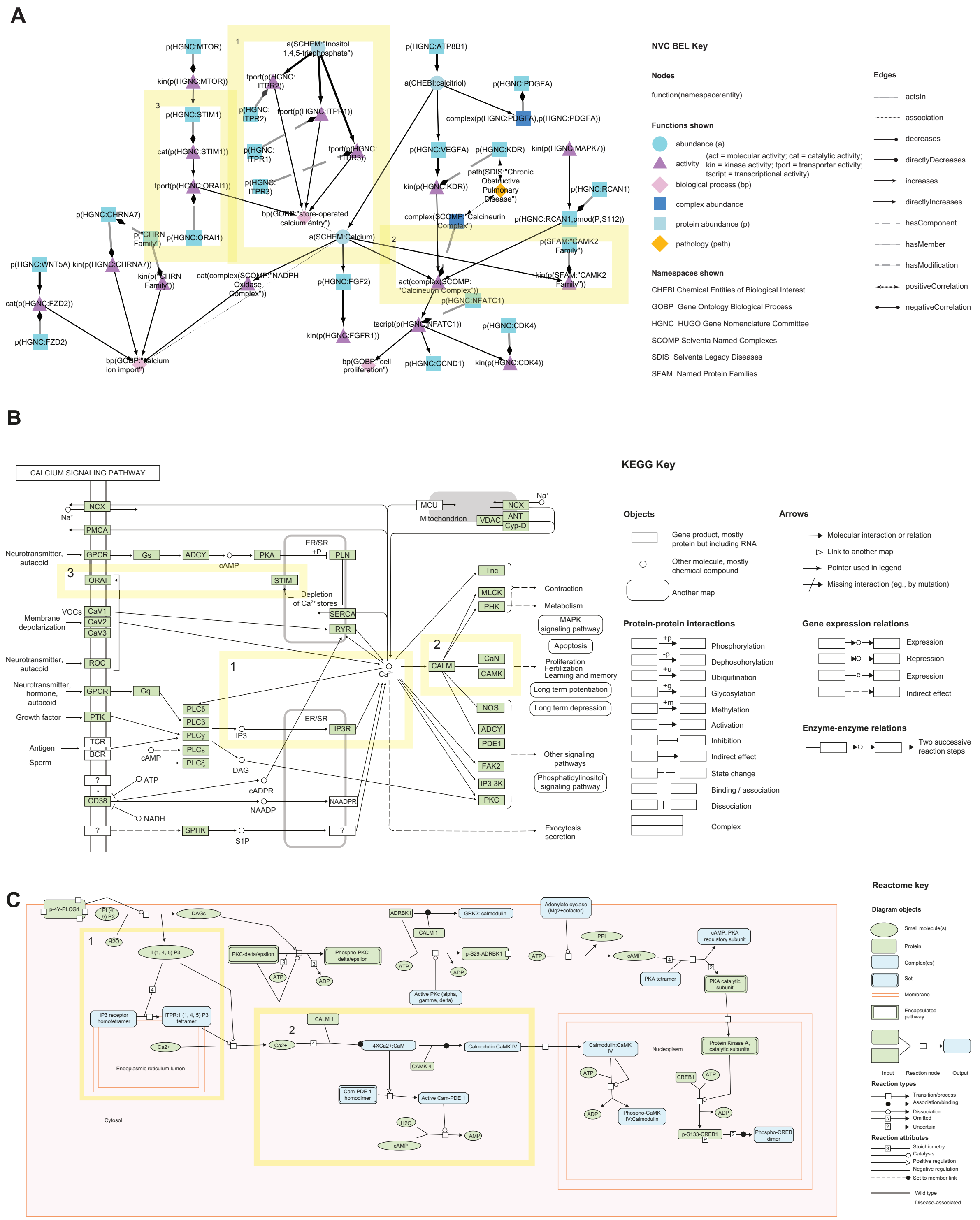


Figure 1. Comparison of the NVC (**A**), KEGG (**B**), and Reactome (**C**) calcium/calmodulin signaling pathways. Shared portions highlighted in yellow with corresponding numbers.

KEGG pathway after clicking on the node within the online pathway). The final group of overlapping nodes between NVC and KEGG networks include stromal interaction molecular 1 (STIM1) and calcium release-activated calcium channel protein 1 (ORAI1), describing store-operated calcium entry (Fig. 1, box 3 highlighted in yellow), a concept that the Reactome network does not cover due to its focus on calmodulin signaling. Despite the differences in biological content, these

Table 2. Network resource comparison.

ATTRIBUTE	NVC	KEGG	REACTOME
# Nodes	47	48	46
# Unique concepts	35	48	34
# Edges	52	60	49
# References	38	20	28
Average date of references	2006	2002	2000

networks illustrate the similarities in causal, computational formats and differences in detail and visualization features in the NVC, KEGG, and Reactome networks. The edges in the NVC, KEGG, and Reactome networks are similar in that they can represent causal increase or decrease relationships and can be downloaded for computational use. However, the NVC networks contain more layers of information, with direct causal, indirect causal, correlative, and other noncausal relationships (eg, member, biomarker, and component).

Network crowd verification. *Participant feedback.* Scientists had many options for engagement during the NVC, including commenting on networks, voting for or against the validity of evidence for specific edges, adding evidence to existing edges, or adding new edges (in order of easiest to most challenging according to a participant survey). The most impactful, but most challenging (and highest point value), action was to add new edges that represented missing biology in the networks. This action required participants to perform a sophisticated set of tasks beyond identifying relevant papers, namely, identify the correct network to include the paper and translate the biology to correct BEL statements in a format that contained direct, mechanistic biology relevant to the boundaries of the particular network. Most participants had expertise in identifying relevant papers that included biology that was missing in the network and overall, participants were able to easily learn BEL and construct correct statements that depicted the biology from the papers they identified. The most challenging task was assembling these statements into direct, mechanistic edges to integrate into the boundaries of a particular network. Participant feedback indicated that improved ways were desired to view networks, particularly to highlight areas of the networks that needed more development. Feedback also indicated that clearer network boundaries were necessary, highlighting the challenges that working with networks entails. With regard to participant engagement, feedback showed that participants were motivated by learning about biology in the networks, the BEL language, and about biological networks in general.

Network changes. The latest version of the NVC networks edited by the crowd during the NVC2 is available as version 2.0 at www.bionet.sbvimprover.com. These networks were changed in various ways throughout the two NVC challenges, as summarized in Figure 2. Networks before the NVC (v1.1) were compared with networks changed at the end of NVC2 (v2.0). Network statistics for each network version are available in Supplementary File 2. The largest amount of new biology in terms of new nodes that was added during NVC2 by the crowd and resulting from the jamboree was to the epigenetics, xenobiotic metabolism response, and calcium networks (Fig. 2). COPD- and lung-relevant contexts were added to the epigenetics and xenobiotic metabolism response networks, and cancer- and liver-related contexts, respectively, were removed. In the calcium network, growth factors and smoke-relevant mechanisms that lead to calcium signaling were added, as well as mechanisms of store-operated calcium entry.

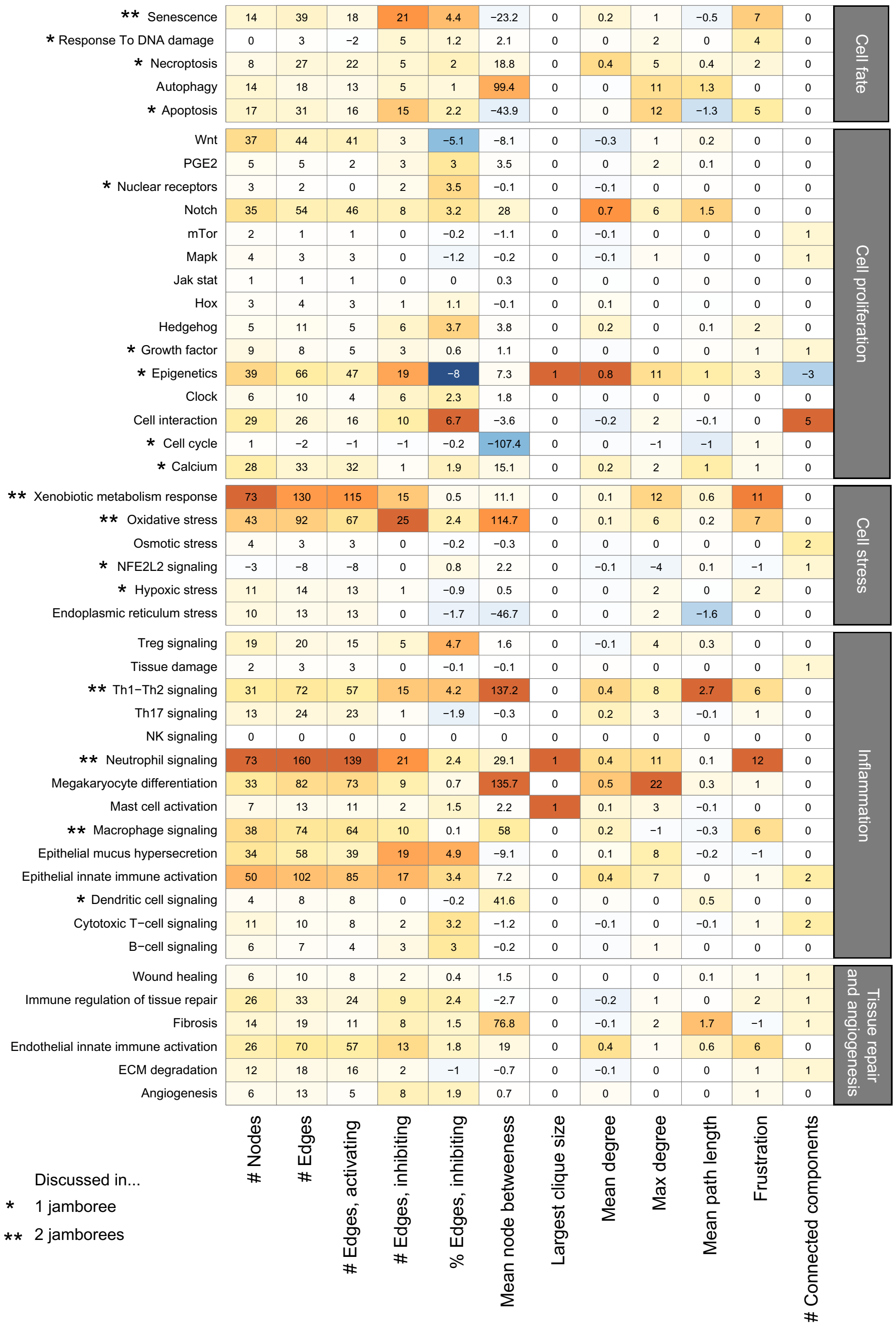
Overall during the NVC1 and NVC2, the size of the networks (number of nodes and edges) grew, as seen in the four left columns of the heat map (Fig. 2). While the total number of edges increased, the proportion of negative edges also increased slightly, with a few exceptions such as Wnt and epigenetics signaling. This increase may reflect the addition of regulatory mechanisms to the networks.

Mean node betweenness (MNB) did not change substantially, with noticeable exceptions for the cell cycle, autophagy, and Th1–Th2 signaling networks. For both cell cycle and autophagy, the number of nodes and edges stayed relatively constant. A difference in MNB may be indicative of a reorganization of the network topology. These networks were all discussed during the jamborees where network topologies could more easily be changed than on a per user basis during the open phase. For Th1–Th2 signaling, MNB went up tenfold from 15 to 152. This may be because these networks were originally two separate networks with linear (tree-like) structures that were then integrated after the jamboree.

The sizes of the largest cliques did not change, which suggests that the crowd did not add feedback loops. A clique of size 3 is a triangle that may be a simple positive or negative feedback of the form A→B→C→A (A→B→C-|A, respectively). Most of the networks exhibit this property, while only eight networks have a clique of size 4 or more, the maximum being 5 (neutrophil signaling, after verification). A clique between four nodes implies that the set of nodes all regulate each other; for example, in the epithelial mucus hypersecretion network, the nodes A = cat(p(HGNC:ADAM17)), B = kin(p(HGNC:EGFR)), C = p(HGNC:MUC5AC), and D = bp(GOBP:mucus secretion)) are all related to each other as A→B,C,D; B→C,D; C→D.

The mean degree stayed stable while some maximum node degrees increased (ie, some nodes are stronger *hubs*). As a case in point, for the megakaryocyte differentiation network, the maximum degree went from 12 to 34. The MPL stayed stable for all networks, meaning that, on average, the shortest path between two nodes did not change (eg, no long *hanging* linear paths).

The frustration, representing the complexity of autoregulation of a network, increased in half of the networks. After verification, only eight networks have a decreased frustration.



	# Nodes	# Edges	# Edges, activating	# Edges, inhibiting	% Edges, inhibiting	Mean node betweeness	Largest clique size	Mean degree	Max degree	Mean path length	Frustration	# Connected components	
** Senescence	14	39	18	21	4.4	−23.2	0	0.2	1	−0.5	7	0	Cell fate
* Response To DNA damage	0	3	−2	5	1.2	2.1	0	0	2	0	4	0	
* Necroptosis	8	27	22	5	2	18.8	0	0.4	5	0.4	2	0	
Autophagy	14	18	13	5	1	99.4	0	0	11	1.3	0	0	
* Apoptosis	17	31	16	15	2.2	−43.9	0	0	12	−1.3	5	0	
Wnt	37	44	41	3	−5.1	−8.1	0	−0.3	1	0.2	0	0	Cell proliferation
PGE2	5	5	2	3	3	3.5	0	0	2	0.1	0	0	
* Nuclear receptors	3	2	0	2	3.5	−0.1	0	−0.1	0	0	0	0	
Notch	35	54	46	8	3.2	28	0	0.7	6	1.5	0	0	
mTor	2	1	1	0	−0.2	−1.1	0	−0.1	0	0	0	1	
Mapk	4	3	3	0	−1.2	−0.2	0	−0.1	1	0	0	1	
Jak stat	1	1	1	0	0	0.3	0	0	0	0	0	0	
Hox	3	4	3	1	1.1	−0.1	0	0.1	0	0	0	0	
Hedgehog	5	11	5	6	3.7	3.8	0	0.2	0	0.1	2	0	
* Growth factor	9	8	5	3	0.6	1.1	0	0	0	0	1	1	
* Epigenetics	39	66	47	19	−8	7.3	1	0.8	11	1	3	−3	
Clock	6	10	4	6	2.3	1.8	0	0	0	0	0	0	
Cell interaction	29	26	16	10	6.7	−3.6	0	−0.2	2	−0.1	0	5	
* Cell cycle	1	−2	−1	−1	−0.2	−107.4	0	0	−1	−1	1	0	
* Calcium	28	33	32	1	1.9	15.1	0	0.2	2	1	1	0	
** Xenobiotic metabolism response	73	130	115	15	0.5	11.1	0	0.1	12	0.6	11	0	Cell stress
** Oxidative stress	43	92	67	25	2.4	114.7	0	0.1	6	0.2	7	0	
Osmotic stress	4	3	3	0	−0.2	−0.3	0	0	0	0	0	2	
* NFE2L2 signaling	−3	−8	−8	0	0.8	2.2	0	−0.1	−4	0.1	−1	1	
* Hypoxic stress	11	14	13	1	−0.9	0.5	0	0	2	0	2	0	
Endoplasmic reticulum stress	10	13	13	0	−1.7	−46.7	0	0	2	−1.6	0	0	
Treg signaling	19	20	15	5	4.7	1.6	0	−0.1	4	0.3	0	0	Inflammation
Tissue damage	2	3	3	0	−0.1	−0.1	0	0	0	0	0	1	
** Th1−Th2 signaling	31	72	57	15	4.2	137.2	0	0.4	8	2.7	6	0	
Th17 signaling	13	24	23	1	−1.9	−0.3	0	0.2	3	−0.1	1	0	
NK signaling	0	0	0	0	0	0	0	0	0	0	0	0	
** Neutrophil signaling	73	160	139	21	2.4	29.1	1	0.4	11	0.1	12	0	
Megakaryocyte differentiation	33	82	73	9	0.7	135.7	0	0.5	22	0.3	1	0	
Mast cell activation	7	13	11	2	1.5	2.2	1	0.1	3	−0.1	0	0	
** Macrophage signaling	38	74	64	10	0.1	58	0	0.2	−1	−0.3	6	0	
Epithelial mucus hypersecretion	34	58	39	19	4.9	−9.1	0	0.1	8	−0.2	−1	0	
Epithelial innate immune activation	50	102	85	17	3.4	7.2	0	0.4	7	0	1	2	
* Dendritic cell signaling	4	8	8	0	−0.2	41.6	0	0	0	0.5	0	0	
Cytotoxic T−cell signaling	11	10	8	2	3.2	−1.2	0	−0.1	0	−0.1	1	2	
B−cell signaling	6	7	4	3	3	−0.2	0	0	1	0	0	0	
Wound healing	6	10	8	2	0.4	1.5	0	0	0	0.1	1	1	Tissue repair and angiogenesis
Immune regulation of tissue repair	26	33	24	9	2.4	−2.7	0	−0.2	1	0	2	1	
Fibrosis	14	19	11	8	1.5	76.8	0	−0.1	2	1.7	−1	1	
Endothelial innate immune activation	26	70	57	13	1.8	19	0	0.4	1	0.6	6	0	
ECM degradation	12	18	16	2	−1	−0.7	0	−0.1	0	0	1	1	
Angiogenesis	6	13	5	8	1.9	0.7	0	0	0	0	1	0	

Discussed in...
* 1 jamboree
** 2 jamborees

Figure 2. Changes in network statistics as a result of NVC activity. Differences between the latest version of the networks and the original networks have been posted to the Bionet website.

Notes: *Discussed in one jamboree. **Discussed in two jamborees. Networks are organized in the following biological categories: cell fate, cell proliferation, cell stress, inflammation, and tissue repair and angiogenesis. The details of the analysis and the description of the different statistics are described in the "Materials and methods" section.

The number of connected components increased in the following networks (usually from one to two components): mTor, Mapk, Hox, growth factor, cell interaction, osmotic stress, NFE2L2 signaling, epithelial innate immune activation, wound healing, fibrosis, and ECM degradation. However, the ratio of the size of the second largest component to the size of the largest is less than 5% (except for cell interaction 12%, cytotoxic T-cell signaling 15%, and Hox 66%), meaning that, except for the Hox network, the largest components comprise almost all the nodes. The extra components added during network verification may be a starting point for further extending the biggest component. However, in the case of the Hox network, two components describing separated processes are described in the context of this network. Besides the metrics discussed above, a scale-free property (ie, the degree distribution follows an exponential distribution) was tested. None of the networks (v1.1. and v2.0) exhibit a significant scale-free property (Supplementary File 2).

Network applications. Because the networks were constructed in BEL, they can be shared within the scientific community and used to understand data through overlay on to specific pathways of interest or implementing a more global process overview using computational inference approaches. We illustrate a few cases of how the networks could be used in toxicity assessment and drug discovery for network computation using the TopoNPA approach. This approach employs the two-layer network model to infer the activation or inhibition of model backbone nodes based on gene expression data.[22] Using these inferences and the network model topology, TopoNPA computes the perturbation of the network as a whole. The approach differs from traditional pathway analyses, because it is quantitative and it uses backward reasoning instead of assuming that changes in gene expression directly imply changes in protein activity. The comparison of TopoNPA with other methods was described in detail by Martin et al.[22]

In vitro treatment effects on transcriptomics data are reflected in TopoNPA network scores. The NVC2 networks were scored on the *in vitro* dataset GSE28464 from the NCBI GEO database to illustrate that expected pathway activation can be inferred from transcriptomics data using network scoring.[29] In this dataset, HRASV12 was expressed in fibroblasts, as a model for oncogene-induced senescence and cell cycle arrest. Consistent with the expectations, the senescence and cell cycle networks scored significantly in the HRASV12 dataset (Fig. 3). Within the senescence network, leading nodes that contribute to 80% of the senescence network score were predicted to be increased, including bp(GOBP:oncogene-induced cell senescence), representing oncogene-induced cell senescence, and p(HGNC:HRAS sub(G, 12, V)), representing HRASV12 mutation, ranking first and eighth in their contribution to the significant senescence network score (Fig. 3A, boxed in yellow). Many nodes representing RAS, RAF, and MAPK mechanisms also scored highly and/or were high contributors to the network score as leading nodes. The relationship from angiotensin II activating CDKN1A protein is an example of an edge added to the senescence network during the NVC process.

The cell cycle network also had a significant network score with cell cyclins and E2Fs inferred as decreased leading nodes (Fig. 3B, highlighted in yellow), while inhibitors of cyclins and E2Fs (CDKN1A and RB1) were inferred as increased leading nodes (Fig. 3B, highlighted in blue). NVC contributions include RRM1, MAD2L1, SIRT1, and TP53 acetylation, which adds more detail to the role of THAP1 and TP53 in regulating cell cycle. The nodes predicted in the senescence and cell cycle networks are consistent with an expected decrease in cell cycle due to HRASV12 signaling.

Quantification/comparison of toxicity in two related datasets using the network suite. Networks were used to evaluate and compare two recently published mouse lung datasets (E-MTAB-3150 and GSE52509), in order to quantify the effects of different exposures on biological processes at different time points.[30] In the first study (E-MTAB-3150), mice were exposed to CS or aerosol from a prototype modified risk tobacco product (pMRTP). After two months, mice were switched from CS exposure to pMRTP or fresh air (cessation) for an additional five months and compared with mice subjected to CS for the whole duration (seven months). In the study reported in the GSE52509 dataset, mice were exposed to smoke for four or six months.[31]

Macrophage signaling is of particular interest in the first study (E-MTAB-3150). The NPA score for the macrophage signaling network significantly increased with smoke exposure for all time points and decreased with switch and cessation (Fig. 4A). This trend matched the measured end points of macrophage count in bronchoalveolar lavage fluid (BALF) and pigmented macrophages in lung tissue (Fig. 4B).[30] Leading nodes within the macrophage signaling network that contributed most to the score are depicted by relative contribution to network scores in Figure 5. The Il1r1 protein and activity were top contributors to the network score for the first four months of smoke exposure, after which Irak4 and Myd88 activity were top scoring contributors. These nodes also contributed most to the five-month pMRTP, switch to pMRTP, and cessation scores. Irak4 and Myd88 act in the TLR pathway that leads to macrophage activation induced by smoke for six months (Fig. 6, boxed in yellow). A number of new nodes were added during the NVC2 process, including detail around the TLR pathway and effects of macrophage activation. Two of these new nodes, prostaglandin E2 and nitric oxide, were leading nodes that contributed highly to the macrophage signaling network score.

NPA scores can be calculated for the whole suite of networks and also allow to compare different datasets, as the relative signal compared with a control is used. Figure 7 shows that, as expected, most of the networks were predicted to be significantly impacted with CS exposure in the E-MTAB-3150 dataset, with an increasing impact over time. In contrast, most of the networks were predicted to be not impacted significantly with pMRTP exposure. Upon cessation or switch to pMRTP from smoke exposure, the network scores decreased. Interestingly, this approach also proves powerful when applied to a dataset with fainter signal, as judged by the number of differentially expressed genes. Indeed, the number of differentially expressed genes in GSE25209 is low (hundreds) compared with those in the E-MTAB-3150 dataset (thousands) for smoke-exposed mice (Supplementary File 3). Despite the low signal, TopoNPA still detected a signal and predicted activation of key networks known to be involved in smoking,

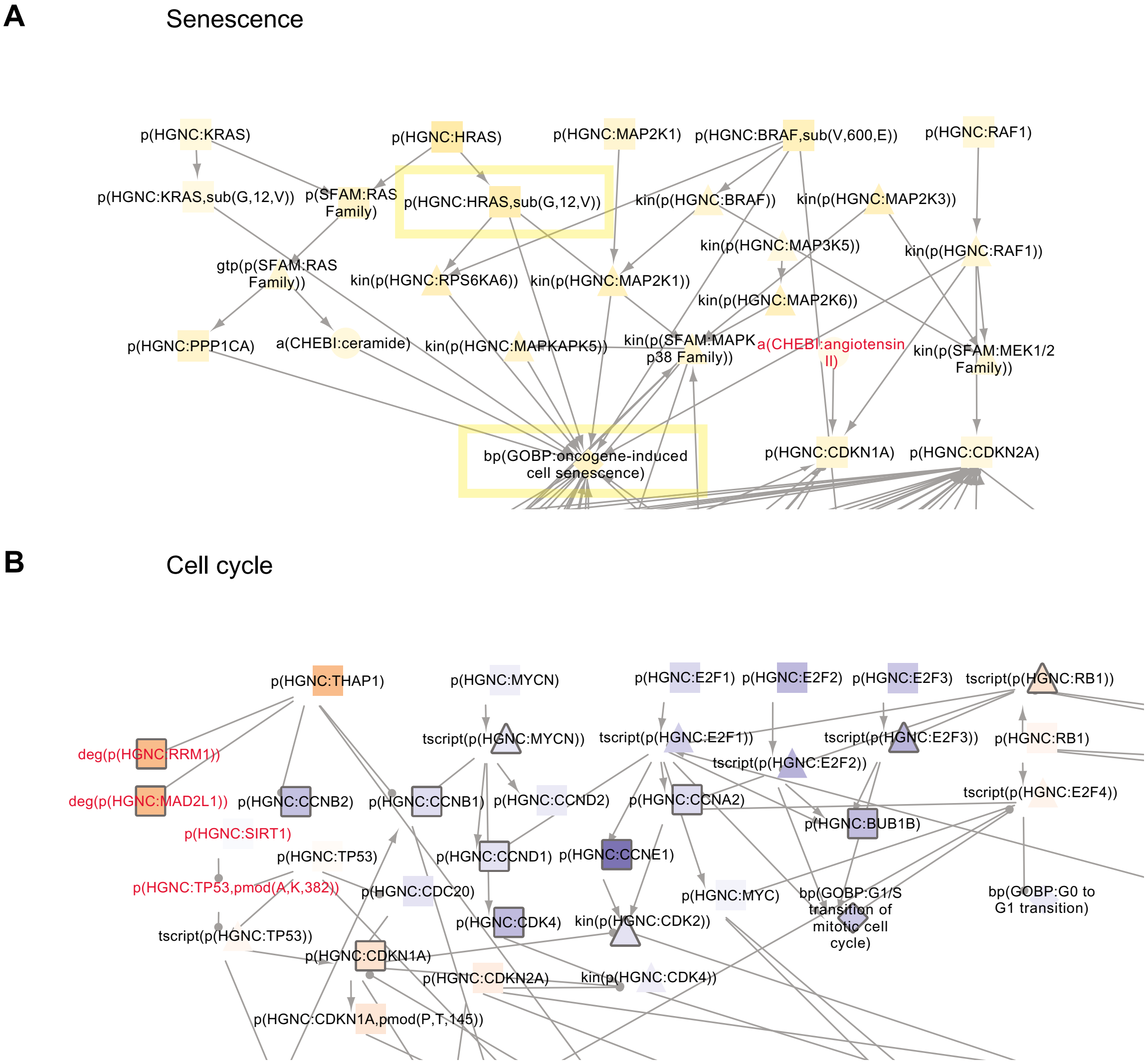


Figure 3. Senescence (**A**) and cell cycle (**B**) networks scored with GSE28464 HRASV12 data from the NCBI GEO database. A selection from the TopoNPA-scored version is shown. Arrow edge indicates a positive relationship while ball and stick edge indicates a negative relationship (includes causal and correlative statements). Nodes are colored by their NPA score; yellow/orange indicates inferred increase and blue indicates inferred decrease in activity or abundance. Darker colors denote higher magnitude scores. Leading nodes contribute to 80% of the network score and are denoted by their shapes outlined in gray. Nodes added within this section of the network during the NVC are labeled in red. (**A**) Senescence network. Nodes boxed in yellow reflect experimental HRASV12 mutation, resulting in oncogene-induced senescence. (**B**) Cell cycle network. Predicted upregulated nodes (yellow) contain cell cycle inhibitors RB1, E2F4, and CDKN1A predicted increased. Predicted decreased nodes (blue) contain cell cyclins and E2Fs predicted decreased.

including the inflammatory, cell stress, cell proliferation, and tissue repair networks (Fig. 7). The networks that score significantly in GSE52509 were similar to those in the C57BL6-pMRTP-SW dataset, sharing 24 significant and 11 nonsignificant networks out of the 46 total networks. Note that scores cannot be compared across datasets.

One of the networks that scored significantly for the impact of six-month smoke was the Th17 signaling network. The network shows mechanisms that can contribute to Th17 signaling and were predicted to be increased or decreased. Il17 differential gene expression was not statistically significant based on the microarray data; however, evidence of Il17a and Il17f activation from the overall transcriptomics signal was inferred and contributed to the significant Th17 signaling network score (Fig. 8, boxed in yellow). These network inferences match measurements from the study, reporting a higher number of Th17 cells and IL17-positive cells in the six-month smoke-exposed lung tissue.[31] Additionally, the study reported enrichment of innate and adaptive immune cell communication pathways by Ingenuity Pathway Analysis of transcriptomics data, which matches the significant network scores in T-cell and other immune networks (Fig. 7).

Discussion

Network resources have different strengths. Many different network resources are available online, with different

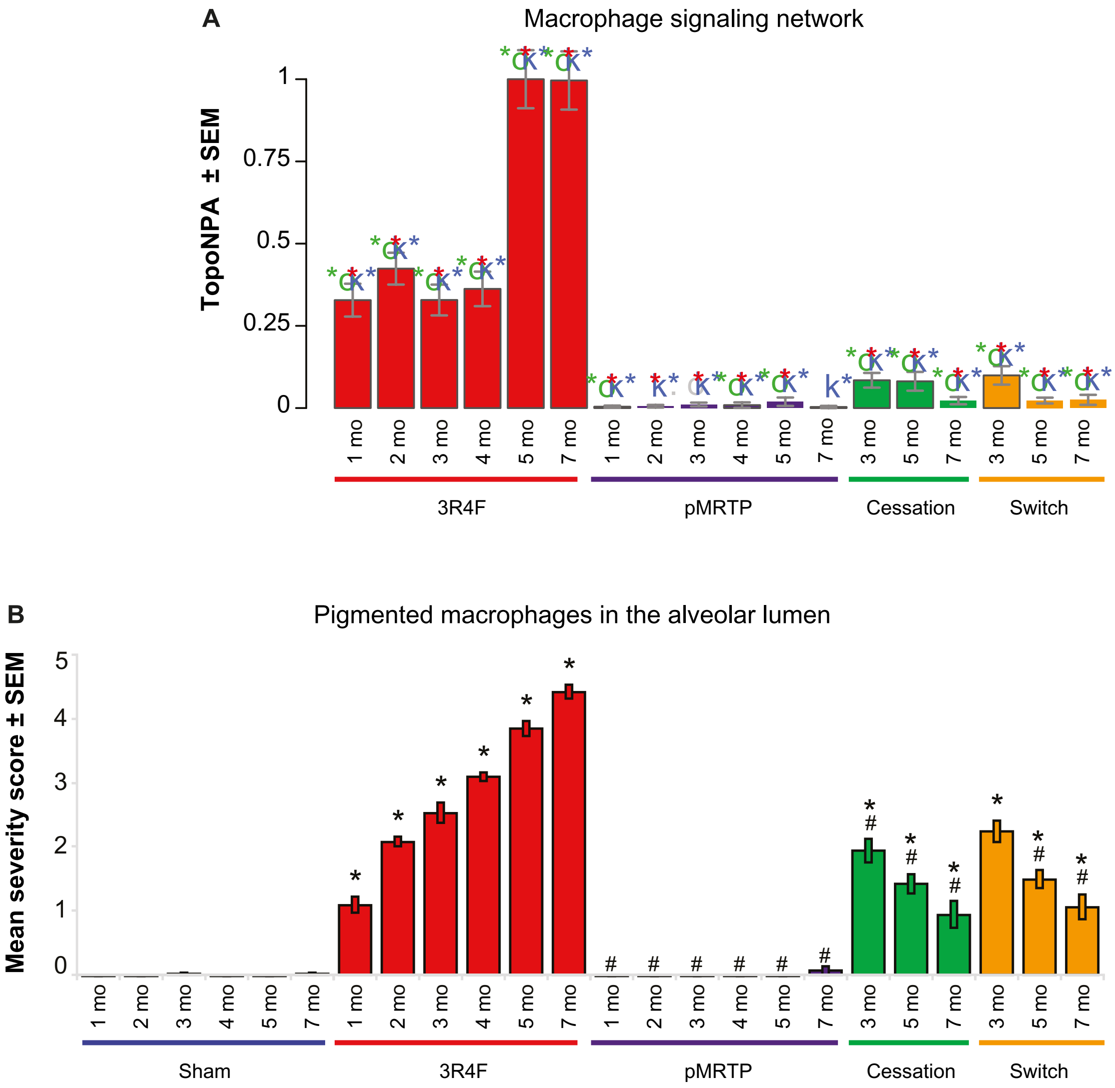


Figure 4. Macrophage signaling network scores in the E-MTAB-3150 dataset and pigmented macrophage counts in the same study. (**A**) Macrophage signaling network score increased with time with smoke exposure and decreased with switch or cessation. pMRTP did not have significant macrophage signaling network scores at any time point. Green, blue, and red asterisks indicate significant O, K, and experimental *P*-values, respectively. (**B**) Pigmented macrophage in the alveolar lumen increased with smoke exposure over time and decreased with switch or cessation. pMRTP did not induce an increase in pigmented macrophages.
Notes: $^{*}P < 0.05$ compared with sham. $^{\#}P < 0.05$ compared with smoke exposure.

language formats, visualization, and download application capabilities.[32,33] Out of these, we chose to compare two of the most widely used network resources, KEGG and Reactome, to the NVC networks focusing on the calcium signaling network as a point of comparison. BEL networks enhanced in the NVC cover 46 different COPD-relevant processes. The KEGG pathway database is a well-known resource in the scientific community that can be used to interpret data.[4,5] Created by a select team of biologists, KEGG contains hundreds of pathways covering a wide variety of processes including metabolism, cellular processes, diseases, and more. Reactome is an open-source, open-access collection of manually curated and peer-reviewed pathways and suite of data analysis tools to support pathway-based analysis.[6,7] Similarly, the NVC networks are manually curated by a team of scientists and organized into discrete subject areas. However, unlike the KEGG and Reactome pathways, these network graphs are open to the crowd for editing and each of the edges that make up the network is supported by literature source(s) along with a quotation from the paper that supports the edge and experimental context. The ability for the crowd to edit the networks facilitates a peer-review process, which ensures comprehensive and current networks.

The NVC networks have different edge and node types that describe the relationships between nodes in great detail to reflect exactly what was proven in the experiment the

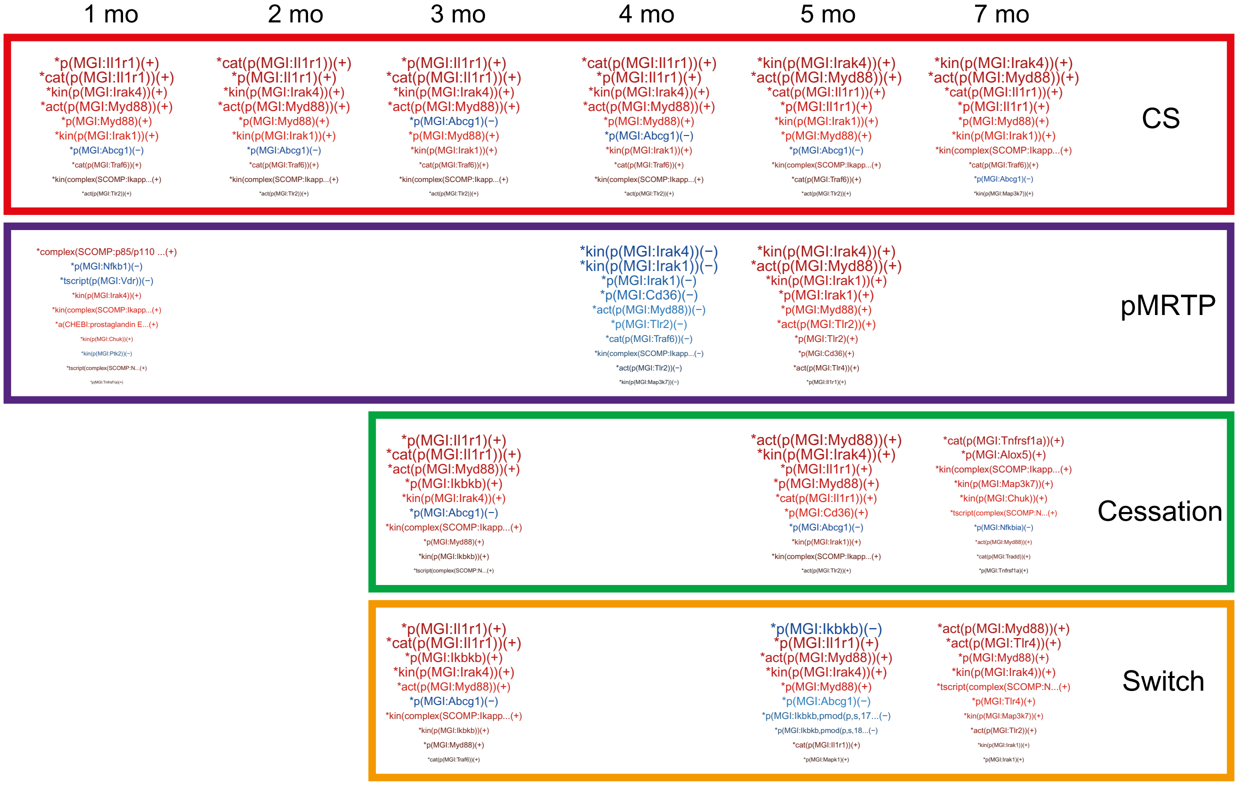


Figure 5. Leading node contribution for macrophage signaling network in the E-MTAB-3150 dataset. Word size indicates relative contribution to network score.

Notes: *significant score; (+) inferred increase; (–) inferred decrease.

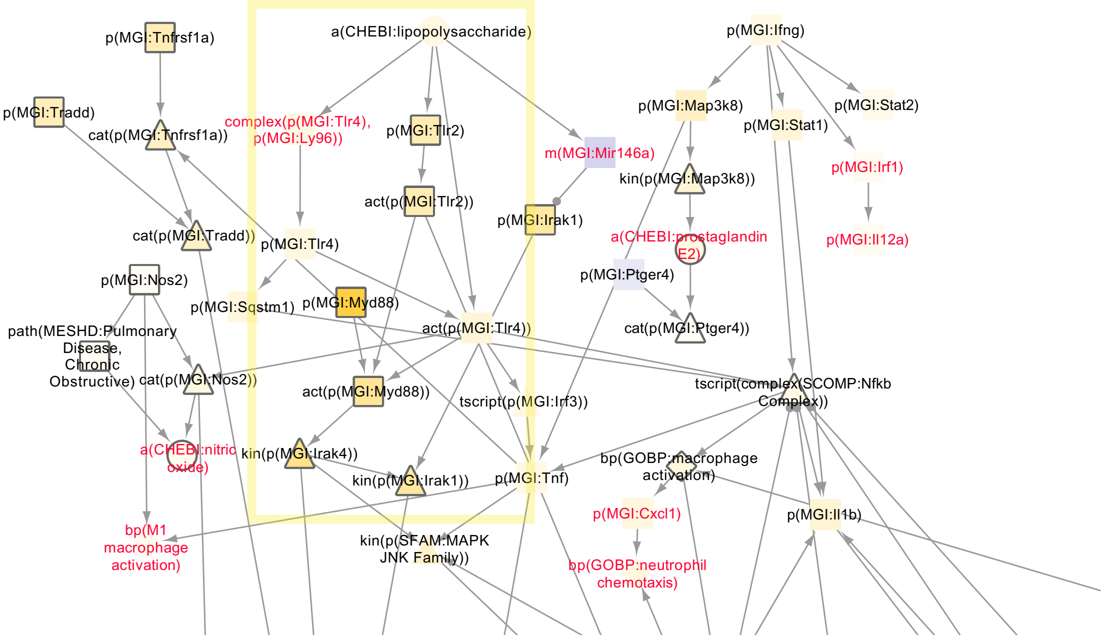


Figure 6. Macrophage signaling network scores for seven-month smoke vs seven-month fresh air using the E-MTAB-3150 dataset. A selection from the TopoNPA-scored version is shown. Arrow edge indicates a positive relationship, while ball and stick edge indicates a negative relationship (includes causal and correlative statements). Nodes are colored by NPA score; yellow indicates inferred increase and blue indicates inferred decrease. Darker colors denote higher magnitude scores. Leading nodes contribute to 80% of the network score and are denoted by their shapes outlined in gray. Nodes added within this section of the network during the NVC process are labeled in red. Nodes boxed in yellow reflect prediction of TLR pathway.

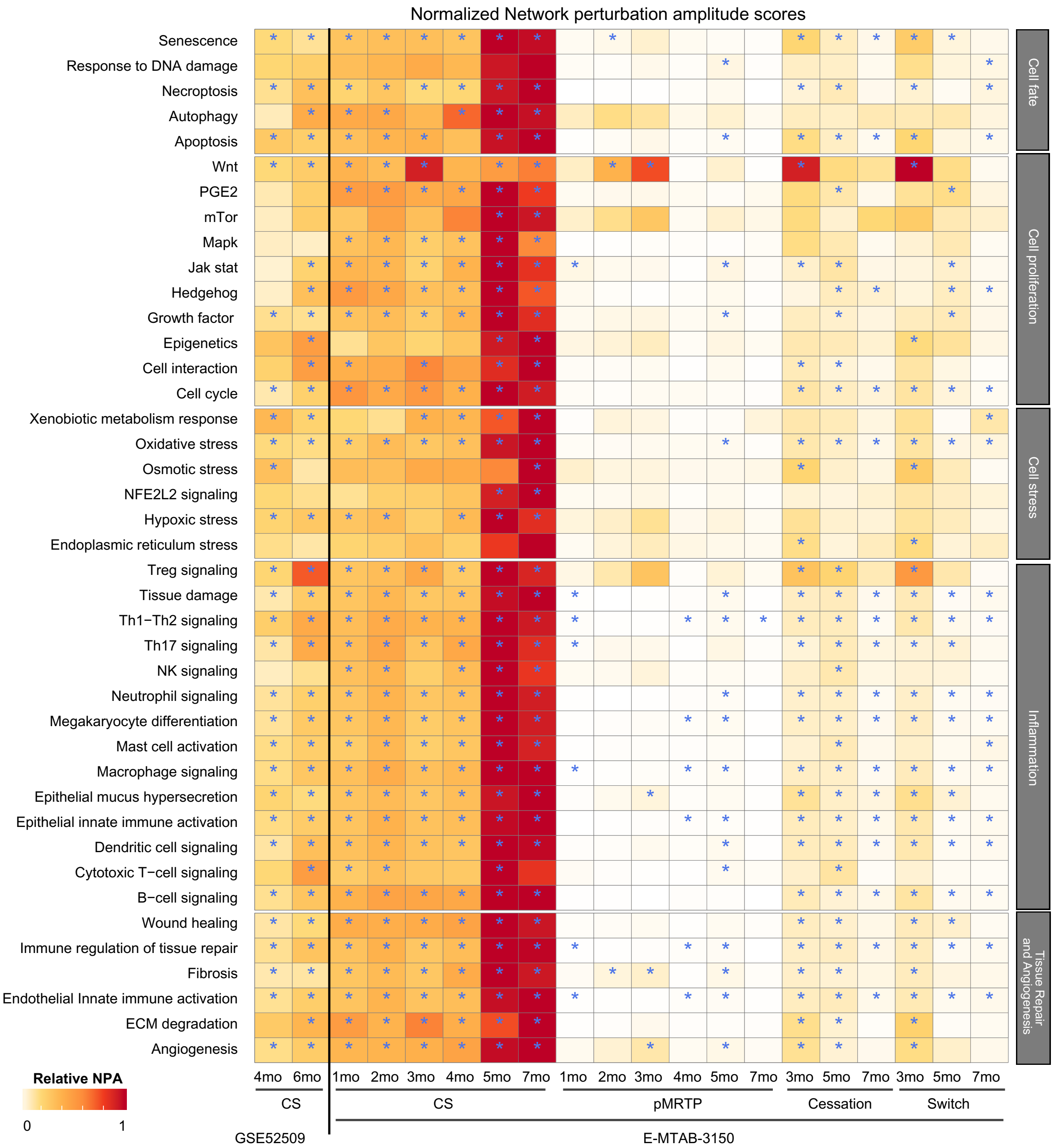


Figure 7. Heat map of network scores comparing the impact of CS exposure, pMRTP, and cessation in the E-MTAB-3150 and GSE52509 datasets. Each treatment is compared to fresh air at the same time point. Scores are normalized to the maximum scores for each network. A network is considered impacted if, in addition to the significance of the score with respect to the experimental variation, the two companion statistics (O and K) derived to inform the specificity of the score with respect to the biology described in the network, are significant.

Note: *O and K statistic *P*-values below 0.05 and NPA significantly nonzero.

annotated reference describes. Nodes defined by a namespace serve to standardize the language and multiple functions such as abundance, activity, modifications (ie, phosphorylation), biological process, and pathology to describe the biology in a fine-grained manner. Edges are defined by causal, correlative, and other numerous noncausal relationships and each causal/correlative edge is based on a literature reference containing tissue, species, disease, and experimental metadata. Like the NVC networks, KEGG and Reactome describe biological processes in a causal manner, though they have less granular information about the nodes and edges and, for the case of KEGG, no specific literature reference was found for each relationship. Reactome has references by edge in the network downloads but not in an easily viewable format on the

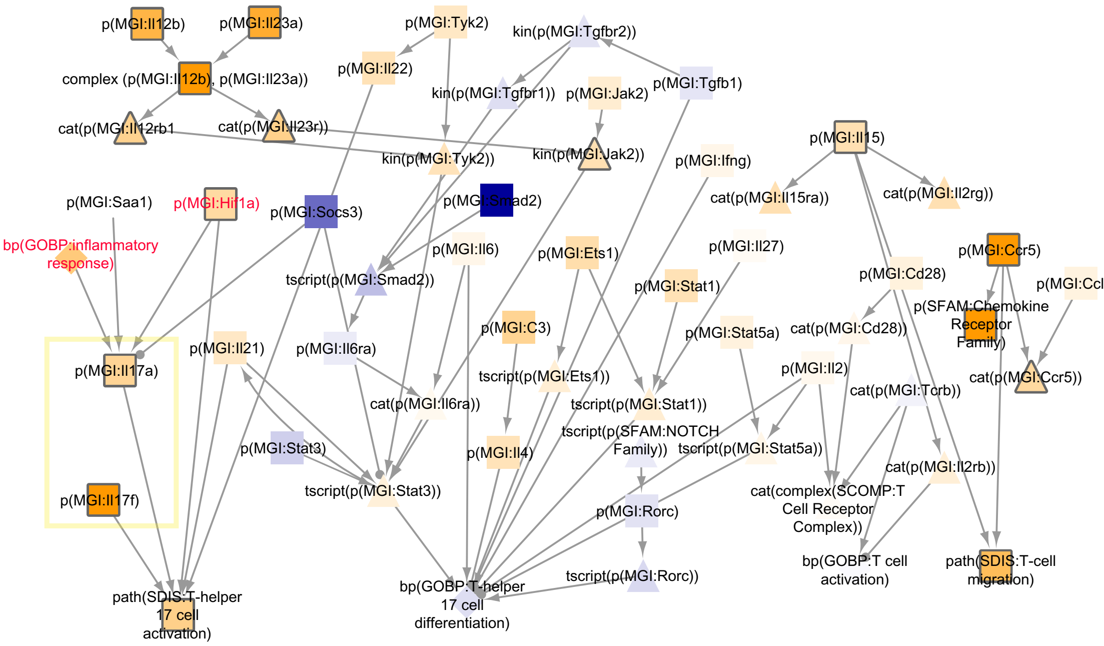


Figure 8. Th17 signaling network scored with GSE52509 mouse lung exposed to 6 month smoke. The whole TopoNPA-scored version is shown. Arrow edge indicates a positive relationship, while ball and stick edge indicates a negative relationship (includes causal and correlative statements). Nodes are colored by NPA score; yellow indicates inferred increase and blue indicates inferred decrease. Darker colors denote higher magnitude scores. Leading nodes contribute to 80% of the network score and are denoted by their shapes outlined in gray. Nodes added within this section of the network during the NVC process are labeled in red. Nodes boxed in yellow reflect prediction of Il17 cytokines.

graph itself. References for the NVC calcium network were, on average, more recent than the KEGG and Reactome networks, implying that the NVC network contains more up-to-date information, most likely because of the crowdsourcing effort. Among the 86 references used to support the calcium pathways across all three resources, all references were unique. This illustrates the range of literature and boundaries that were used to build the calcium pathways across the three network formats. The visualization of the KEGG and Reactome pathways allows the viewer to easily traverse the networks within a graphical representation that includes cellular localization of the nodes. KEGG and Reactome pathway diagrams have detailed cellular localization information that the BEL networks do not show graphically. However, this information can be described in the edge annotation or the node label.

Many analysis tools are available to use with the KEGG and Reactome pathways to interpret data. NVC networks also support analytics for mapping nodes in a dataset as well as taking into account the relationships between the nodes with the exact edge data. NVC networks can be downloaded in JSON graph format (JGF) and viewed and applied to data using Cytoscape or other JGF-compatible network visualization software. Edge information can be used to filter and compute on the networks.

Other network resources that are geared toward a community-driven approach include WikiPathways[34] and the Cell Collective.[35] These resources do not have a calcium pathway appropriate for comparison, but like KEGG and Reactome, they are limited by less granular information about the nodes and edges compared with NVC networks and, like KEGG, no specific literature reference is given for each relationship. However, they do benefit from the contribution of information from the scientific crowd, where WikiPathway users can edit and contribute to existing pathways and Cell Collective users can contribute information to the Knowledge Base, collaboratively build models and simulate and analyze them in real time. Like KEGG and Reactome, WikiPathways provides a graphical representation, containing cellular localization information.

Each of these network resources offers advantages for viewing and interpreting biology. The NVC networks cover lung- and COPD-relevant processes in a very detailed and granular manner and are open to public feedback, and the data can be computed at the node and edge level. The KEGG and Reactome pathways cover a wide range of biology with many widely used node-centric analysis tools, the Cell Collective allows for quantitative computation of networks, and KEGG, Reactome, and WikiPathways provide a simplified representation for easy visualization.

NVC crowd excels at identifying and encoding literature. A review of the crowd changes and participant survey feedback after two iterations of the NVC allowed for an understanding of aspects that worked well and aspects that can be improved for subsequent challenges. One important finding was that the crowd was able to identify relevant literature that contained COPD mechanisms missing from the networks. Keeping networks up-to-date with the constant stream of published literature is difficult for the small team of scientists who created the networks. Crowdsourcing this effort through the Bionet website allows for a diverse group of international scientists to share in this effort to collect relevant literature and note missing areas in a network using each individual's expertise and biological perspective. This process allows the community as a whole to benefit from up-to-date networks.

The main incentive for participants, according to a survey, was the learning process, and although educating the community about BEL and network biology is an excellent outcome of the NVC, there were many challenges associated with this large, crowdsourced effort to edit the networks. These challenges included clearly defining and communicating rules and boundaries up front in a way that everyone can consistently follow, the follow-up effort required to edit the changes made to the networks to ensure consistency and adherence to the network framework rules, and the creation of accurate BEL statements capturing the biology stated in a publication.

An idea for future challenges is to separate knowledge creation from network construction. Adding new and relevant edges to a network was a heavily incentivized portion of the challenge and is an important mechanism for filling knowledge gaps in the network and maintaining the networks with newer information from the literature. While the crowd participants performed well at identifying relevant literature and representing key ideas in BEL, it was challenging for participants to select and add mechanistic, nonredundant paths that were well integrated with the rest of the network, especially for the larger networks. As seen from the network statistics, the crowd contributed to the number of nodes and edges but not necessarily to changing the topology of the network. Separating the curation and network building portions of the task could provide several advantages. For example, BEL evidences could be voted on by the crowd for accuracy and relevance and refined prior to incorporation into a network. It is difficult to edit evidences and statements once they are connected into a network, as all neighboring edges and all individual evidences supporting the same edge are affected. Moreover, evidences could be more readily shared across networks where applicable, and evidences that are highly relevant, but not the most streamlined, direct connection within a given network, could be omitted from the network but retained for other applications. Making the challenge tasks more manageable and narrowly defined in this manner could potentially attract more participants as well as increase the quality and value of the resulting networks and associated knowledge. Every year, as more biological experts participate in the challenge and more literature is published, the networks can be kept up-to-date with the current understanding of the biology contained in these networks.

Networks can be used in toxicity and drug discovery applications. In addition to application as a tool to understand signaling pathways regulating a disease process, biological networks can be used to predict active mechanisms driving measured biological changes based on a knowledgebase of known regulators of these measured changes. In this study, we use network scoring to infer upstream mechanisms known to regulate measured gene changes applied to three datasets. Networks that contain these mechanisms can then be scored to infer perturbation of biological processes represented by the networks in a quantitative manner. In the GSE28464 study, mutated HRASV12 was expressed in fibroblasts and activation of senescence and cell cycle was inferred by network scoring of the transcriptomics data. These results were consistent with experimental expectations of HRASV12, inducing senescence and cell cycle arrest.[36] This example illustrates the ability of the network scoring approach to infer known active mechanisms using transcriptomics data. Novel mechanisms predicted to be active from transcriptomics data as a result of a treatment could also be identified in biological networks using this approach.

A major advantage of this network-based transcriptomics data scoring approach is the ability to quantitatively compare treatments and time points within a dataset within discrete biological processes. In the E-MTAB-3150 dataset, the effects of smoke, pMRTP, switch to pMRTP, and cessation were quantified on the biological process and mechanistic level through network and mechanism scores. Network scoring indicated that smoke impacted lung biology captured by networks more than pMRTP, switch to pMRTP, or cessation and with a greater magnitude over time. pMRTP appeared to impact lung biology less than smoke, based on the lower pMRTP vs sham network scores and fewer networks scoring significantly. Switching from smoke to pMRTP or cessation showed a decrease in network perturbation compared with sham group over time. Additionally, scoring mechanisms within the network gives insights on which mechanisms are predicted to induce gene expression changes observed in the dataset. Il1 receptor signaling was predicted to impact macrophage activation the most in early time points with smoke treatment, followed by an increased impact of Irak4 and Myd88 activity on macrophage activation in later time points (Fig. 5). Il1r1/MyD88 signaling has been shown to contribute to elastase-induced lung inflammation and emphysema,[37] and although there are no publications implicating Irak4 in emphysema or COPD, a recent conference poster reported MyD88/Irak4 promotion of lung fibrosis in a mouse model of COPD.[38] This network approach can potentially highlight novel mechanisms such as Irak4 that drive disease and increase our understanding of COPD progression. Findings such as these could lead to a list of potential biomarkers or novel targets that could then be confirmed in multiple datasets in the primary disease tissue and narrowed down by aspects of ease of targetability and low off-target effects to identify ideal targets. Additionally, the quantitative aspect to network scoring can be used in toxicity testing to rank the impact of different treatments and study dosing and time effects for a particular perturbation.

Another advantage of the network approach is the ability to glean information from a dataset with a low transcriptomics signal. Similar to the E-MTAB-3150 dataset, GSE52509 contained data from smoke-exposed mouse lungs for four and six months; however, this dataset had a much lower transcriptomics signal. This difference in signal could be attributed to a larger variation in the data, or potentially the lower dosage and duration per day of smoke exposure in GSE52509 compared with the E-MTAB-3150 dataset. In the E-MTAB-3150 study, mice were exposed to smoke 2.4 times longer per day at 1.5 times higher concentration. Similar types of networks and leading nodes were inferred in both studies to be activated in processes relevant to CS exposure, and they matched experimental end points of pigmented macrophage and Th17 counts in E-MTAB-3150 and GSE52509 studies, respectively.

Although the networks focus on lung- and COPD-relevant context and were scored on lung datasets, these networks can apply to other diseases and tissues. The networks include edges that are based on literature from lung-relevant cell types such as fibroblasts, smooth muscle, and immune cells; these cell types are not specific to lung but can apply to many other tissues and disease contexts. The networks to be scored should be evaluated based on the context of the dataset. For the GSE28464 dataset, only the senescence and cell cycle networks were scored, while the immune networks were not scored since the experiment was performed in fibroblasts and not immune cells. Since many of the pathways that the networks describe such as canonical MAPK and NFKB signaling are conserved across tissues, these networks provide an important resource that can be built on to include context-specific mechanisms according to scientists' needs.

Conclusion

The computable biological language BEL allows for encoding of scientific literature with high granularity and is well suited for sharing mechanistic biology in a network context. The NVC takes advantage of the well-defined nature and ease of use of BEL to allow the scientific community to verify, enhance, and use these networks. These networks can then be used for toxicological and drug discovery applications. We illustrated one way to use these networks through quantitative network scoring based on transcriptomics data. Mechanisms were inferred from the data and could be quantitatively compared within a dataset, leading to insights in disease-driving mechanisms and toxicity assessment.

Acknowledgments

The authors thank Anouk Ertan, Laure Cannesson, and David Page for their help in organizing the Network Verification Challenge and jamboree, and Michael Maria and David Page for their help in project management and preparation of this manuscript. The project team expresses their gratitude to the subject matter experts and moderators who actively participated in the jamboree: Natalia Boukharov, Norberto Diaz-Diaz, Larisa Federova, Ignacio Gonzalez, Svetlana Guryanova, Anita Iskander, Ulrike Kogel, Marek Ostaszewski, Carine Poussin, Walter Schlage, Justyna Szostak, and Aravind Tallam.

Author Contributions

Conceived and designed the experiments: JH, MCP. Analyzed the data: JP, SB, MT, YX, AAN, RAR, GA, MCP, MB, DRB, SD, ND-D, ÁMFL, AFM, DGA, SG, RM, FM, SMA, SM, SN, NR, IS, BS, AT, JVC, MGVR, MBP. Wrote the first draft of the manuscript: JP, SB, FM, MT. Contributed to the writing of the manuscript: JP, SB, FM, MT, GA. Agreed with manuscript results and conclusions: JP, SB, MT, YX, AAN, RAR, GA, MB, DRB, SD, ND-D, ÁMFL, AFM, DGA, SG, RM, FM, SMA, SM, SN, NR, IS, BS, AT, JVC, MR, JH, MCP, MBP. Jointly developed the structure and arguments for the paper: JP, SB, FM, MT. Made critical revisions and approved the final version: JP, SB, FM, MT. All the authors reviewed and approved the final manuscript. DGA could not be contacted to approve the final proofs.

Supplementary Material

Supplementary File 1. Biological Expression Language (BEL) functions and namespaces.

Supplementary File 2. Network statistics for the Network Verification Challenge (NVC) v1.1 and v2.0 Bionet networks.

Supplementary File 3. Network scores for the GSE28464 dataset from the NCBI GEO database.

REFERENCES

1. (GOLD) GIfCOLD. *From the Global Strategy for the Diagnosis, Management and Prevention of COPD*. Global Initiative for Chronic Obstructive Lung Disease (GOLD); 2014. Available at: http://www.goldcopd.org/.
2. King PT. Inflammation in chronic obstructive pulmonary disease and its role in cardiovascular disease and lung cancer. *Clin Transl Med*. 2015;4(1):68.
3. Thorley AJ, Tetley TD. Pulmonary epithelium, cigarette smoke, and chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis*. 2007;2(4):409–28.
4. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28(1):27–30.
5. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*. 2014;42(Database issue):D199–205.
6. D'Eustachio P. Reactome knowledgebase of human biological pathways and processes. *Methods Mol Biol*. 2011;694:49–61.
7. Croft D, O'Kelly G, Wu G, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. 2011;39(Database issue):D691–7.
8. sbv IMPROVER Project Team, Boue S, Fields B, et al. Enhancement of COPD biological networks using a web-based collaboration interface. *F*1000*Res*. 2015;4:32.
9. sbv IMPROVER Project Team, Binder J, Boue S, et al. Reputation-based collaborative network biology. *Pac Symp Biocomput*. 2015:270–81.
10. Westra JW, Schlage WK, Frushour BP, et al. Construction of a computable cell proliferation network focused on non-diseased lung cells. *BMC Syst Biol*. 2011;5:105.
11. Schlage WK, Westra JW, Gebel S, et al. A computable cellular stress network model for non-diseased pulmonary and cardiovascular tissue. *BMC Syst Biol*. 2011;5:168.

12. Park JS, Schlage WK, Frushour BP, et al. Construction of a Computable Network Model of Tissue Repair and Angiogenesis in the Lung. *J Clinic Toxicol*. 2013:S:12.http://dx.doi.org/10.4172/2161-0495.S12-002.
13. Gebel S, Lichtner RB, Frushour B, et al. Construction of a computable network model for DNA damage, autophagy, cell death, and senescence. *Bioinform Biol Insights*. 2013;7:97–117.
14. Westra JW, Schlage WK, Hengstermann A, et al. A modular cell-type focused inflammatory process network model for non-diseased pulmonary tissue. *Bioinform Biol Insights*. 2013;7:167–92.
15. Prill RJ, Saez-Rodriguez J, Alexopoulos LG, Sorger PK, Stolovitzky G. Crowdsourcing network inference: the DREAM predictive signaling network challenge. *Sci Signal*. 2011;4(189):mr7.
16. Eiben CB, Siegel JB, Bale JB, et al. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat Biotechnol*. 2012;30(2):190–2.
17. Lee J, Kladwang W, Lee M, et al. RNA design rules from a massive open laboratory. *Proc Natl Acad Sci U S A*. 2014;111(6):2122–7.
18. Kawrykow A, Roumanis G, Kam A, et al. Phylo: a citizen science approach for improving multiple sequence alignment. *PLoS One*. 2012;7(3):e31362.
19. Loguercio S, Good BM, Su AI. Dizeez: an online game for human gene-disease annotation. *PLoS One*. 2013;8(8):e71171.
20. Vashisht R, Mondal AK, Jain A, et al. Crowd sourcing a new paradigm for interactome driven drug target identification in *Mycobacterium tuberculosis*. *PLoS One*. 2012;7(7):e39808.
21. Good BM, Nanis M, Wu C, Su AI. Microtask crowdsourcing for disease mention annotation in PubMed abstracts. *Pac Symp Biocomput*. 2015:282–93.
22. Martin F, Sewer A, Talikka M, Xiang Y, Hoeng J, Peitsch MC. Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models. *BMC Bioinformatics*. 2014;15:238.
23. Catlett NL, Bargnesi AJ, Ungerer S, et al. Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinformatics*. 2013;14(1):340.
24. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*. 2009;10(1):47.
25. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–50.
26. Hoeng J, Deehan R, Pratt D, et al. A network-based approach to quantifying the impact of biologically active substances. *Drug Discov Today*. 2012;17(9–10):413–8.
27. Martin F, Thomson TM, Sewer A, et al. Assessment of network perturbation amplitude by applying high-throughput data to causal biological networks. *BMC Syst Biol*. 2012;6(1):54.
28. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25–9.
29. Narita M, Young AR, Arakawa S, et al. Spatial coupling of mTOR and autophagy augments secretory phenotypes. *Science*. 2011;332(6032):966–70.
30. Phillips B, Veljkovic E, Peck MJ, et al. A 7-month cigarette smoke inhalation study in C57BL/6 mice demonstrates reduced lung inflammation and emphysema following smoking cessation or aerosol exposure from a prototypic modified risk tobacco product. *Food Chem Toxicol*. 2015;80:328–45.
31. John-Schuster G, Hager K, Conlon TM, et al. Cigarette smoke-induced iBALT mediates macrophage activation in a B cell-dependent manner in COPD. *Am J Physiol Lung Cell Mol Physiol*. 2014;307(9):L692–706.
32. Boué S, Talikka M, Westra JW, et al. Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database*. 2015;2015:bav030.
33. Talikka M, Boue S, Schlage WK. Causal Biological Network Database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Comput Syst Toxicol*. 2015;54:65–93.
34. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. *PLoS Biol*. 2008;6(7):e184.
35. Helikar T, Kowal B, McClenathan S, et al. The cell collective: toward an open and collaborative approach to systems biology. *BMC Syst Biol*. 2012;6:96.
36. Serrano M, Lin AW, McCurrach ME, Beach D, Lowe SW. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16INK4a. *Cell*. 1997;88(5):593–602.
37. Couillin I, Vasseur V, Charron S, et al. IL-1R1/MyD88 signaling is critical for elastase-induced lung inflammation and emphysema. *J Immunol*. 2009;183(12):8195–202.
38. Daliri S, Del Bosque D, Umer M, et al. A promoting role for MyD88/IRAK4 signaling in lung fibrosis during COPD progression. *B37. Tell Me Why: COPD Pathogenesis*. American Thoracic Society, Denver, Colorado; 2015:A2905–A2905. http://www.atsjournals.org/doi/book/10.1164/ajrccm-conference.2015.