



PhD-FSTC-2018-51
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defence held on 03/07/2018 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN *BIOLOGIE*

by

AISHWARYA ALEX NAMASIVAYAM

Born on 12 November 1987 in Bharananganam (India)

STANDARDISATION AND ORGANISATION OF CLINICAL DATA AND DISEASE MECHANISMS FOR COMPARISON OVER HETEROGENEOUS SYSTEMS IN THE CONTEXT OF NEURODEGENERATIVE DISEASES

Dissertation defence committee

Prof. Dr Reinhard Schneider, dissertation supervisor
Professor, Université du Luxembourg

Dr Inna Kuperstein
Researcher and scientific coordinator, Institut Curie, Paris

Dr Enrico Glaab, Chairman
Senior research scientist, Université du Luxembourg

Prof. Dr Karsten Hiller
Professor, Technische Universität Braunschweig

Dr Marek Ostaszewski, Vice Chairman
Research associate, Université du Luxembourg

Affidavit

I hereby confirm that the PhD thesis entitled "*Standardisation and Organisation of Clinical Data and Disease Mechanisms for Comparison Over Heterogeneous Systems in the Context of Neurodegenerative Diseases*" has been written independently and without any other sources than cited.

Luxembourg, July 26, 2018

Aishwarya Alex Namasivayam

Acknowledgements

First and foremost, I would like to thank Dr. Reinhard Schneider, my supervisor for giving me the opportunity and support to pursue my PhD in the group. Biocore is a very wonderful working environment. I couldn't ask for a better boss! I would like to thank all my colleagues for their support and making this a memorable journey. Special thanks to Marek, Venkata, Wei and Piotr for their valuable suggestions and feedbacks. My sincere gratitude to Dr. Jochen Schneider and Dr. Karsten Hiller for agreeing to be part of the CET committee and the constructive criticism during the PhD. I would also like to thank Dr. Inna Kuperstein, for accepting to be on the defense committee. I am sincerely very happy to have a woman on the committee! A very special thank you to Adriano for the AETIONOMY times, without your support this wouldn't have been difficult. Thanks to Marie-Laurie for being the one stop person for everything at LCSB.

These four years would not be possible without family and friends. Thank you Susana, for being my first friend in Luxembourg. I am glad we met and you are truly Smartinez. Shaman and Susi, thank you for all the times you welcomed me with open arms, provided me with shelter and food when I wanted to be lazy for the weekend (or any day). Thanks to Gaia, Dheeraj, Maharshi, Kavita, and Anshika for their friendship and interesting conversation during lunch time and beyond. Bruna, thank you for bringing a little bit of Brazil to us. A special thanks to Marouen, Gaia, Zuogong, Berto, Dheeraj and Shaman for RSG Luxembourg! It was an amazing experience, it wouldn't be possible without you guys. I would like to thank my friends at the ISCB student council, for a beautiful volunteering experience. Thank you Jakob, for believing I can deal with finances and bringing me on board. Farzana, thank you for a friendship beyond the Student Council. My sincere gratitude to my parents for always being supportive and believing in me, no questions asked. To my sister, for checking on me when I go into hibernate mode! To my in-laws, Achan, Amma and Krishnan, thank you for making me feel at home. Last but not the least, Raman, for realising that the wife will be 318km away but still marrying me!

Dedication

To my sister, Theresa, for never ever asking me “*When do you finish the PhD?*” and my husband, Raman, for asking me only five times (ok, maybe fifteen).

Abstract

With increased interest in studying neurodegenerative diseases, data generated is growing exponentially. The sheer amount of patient data being collected gives rise to the problem of how it can be stored, represented and classified. The representation of collected data varies from one centre to other, based on several factors such as language, region, standards adapted, study design. The results of this variation makes it difficult to study different cohorts by combining or comparing them. Therefore, variables that are collected in all these cohorts need to be standardised and harmonised for further re-use and analysis.

Disease maps are another form of knowledge resources collecting existing biological facts in a single resource. Disease mechanisms are represented visually in the form of models or maps, capturing the knowledge extracted from the literature. There are several modelling languages which serve this purpose. Disease maps capture knowledge about disease related mechanisms at different molecular levels. Comparison of different disease maps can support co-morbidity studies to identify common disease mechanisms or drug targets. To this end, we developed a method to compare disease maps. We then compared Parkinson's and Alzheimer's disease maps using this approach.

However, there are several modelling formats available. Therefore, here arises a need to harmonise and standardise the representation and make the different disease modelling formats convertible and comparable. For the course of the project, we focus on Open Biological Expression Language (OpenBEL) and Systems Biology Markup Language (SBML) modelling languages. A semi-automated convertor from OpenBEL format to SBML was developed to compare knowledge over heterogeneous systems his was then used to convert an Alzheimer's OpenBEL model to SBML format for better visualization and hierarchical representation and to enable comparison against other SBML models.

In conclusion, the work presented in the thesis, emphasises the importance of standards in the representation and modelling of clinical and biological information to

ensure interoperability between tools and models and facilitate data sharing, reusability and reproducibility.

Contents

Acknowledgements	iii
Abstract	v
1 Introduction	1
1.1 Trends in biomedical research and their impact on bioinformatics	3
1.2 Standards in biomedical research	3
1.2.1 Need for standards	4
1.2.2 Standardisation efforts	6
1.3 Disease maps as knowledge resources	9
1.3.1 Complexities of diseases and their comorbidities	10
1.3.2 Representing diseases as a map	11
1.3.3 Common modelling formats	11
1.3.4 Available data	18
1.4 Scope and Aim	20
1.5 Thesis Overview	21

2	Integrating heterogeneous data	23
2.1	Integrative platforms	23
2.2	Data acquisition	24
2.3	Data harmonisation	25
2.4	Extraction, Transformation and Loading (ETL)	26
2.5	From unstructured to structured data	28
2.6	Results	31
2.6.1	Use case 1: Alzheimer’s disease cytokine study	33
2.6.2	Use case 2: Expression data of substantia nigra from postmortem human brain of PD patients	40
2.7	Summary	42
3	Comparison of disease maps	44
3.1	Overview of existing comparison methods	45
3.2	Methods for comparison	46
3.3	Results	51
3.3.1	Comparison of AlzPathway and PD map	51
3.3.2	<i>AKT1</i> Activity	59
3.3.3	TAU (<i>MAPT</i>) hyper-phosphorylation	62
3.3.4	MAPK signalling	64
3.3.5	Endoplasmic Reticulum Stress	67

3.3.6	Inflammation	70
3.3.7	Wnt signalling	72
3.3.8	Synaptic area	73
3.4	Summary	74
4	Comparison of different disease models	76
4.1	Interoperability between models	77
4.2	Results	84
4.3	Comparison of APP map to PD map	88
4.4	Comparison of APP map to AlzPathway	90
4.5	Summary	96
5	Discussion	98
5.1	Integrating Heterogeneous Data	99
5.2	Comparison and Conversion of Maps	101
6	Summary and Outlook	104
6.1	Summary	104
6.2	Outlook	105
A	Appendix A : Supplementary Materials	108
A.1	Alzheimer's and Parkinson's Disease datasets integrated in AETIONOMY109	
A.2	AlzPathway overlayed on PD Map	111

A.3	Semantic Mapping: CellDesigner (SBGN)-BEL	119
A.4	Nodes extracted from APP BEL model	121
A.5	Reactions extracted from APP BEL model	122
B	Publications	123
	Bibliography	166

List of Tables

1.1	Features of SBML, OpenBEL, BioPAX	15
3.1	Number of elements and reactions in the AlzPathway Map and PD Map and submaps	52
3.2	Summary of model sizes	54
3.3	Number of elements and reactions identified from AlzPathway in PD Map	55
3.4	Number of elements and reactions identified from PD Map in AlzPath- way Map	57
4.1	OpenBEL predicates and corresponding representation in CellDesigner	82
4.2	OpenBEL activity terms and corresponding GO annotation	83
4.3	OpenBEL statements lost in conversion	85
4.4	Submaps components based on cell type	87
4.5	Reactions in the APP Map	90
A.1	Summary of performance with different annotators	111
A.2	Submap PD 180412 2 alzpath 8APR Element	117

A.3	Snapshot of reaction matches in PD Map and Alzpathway	118
A.4	Example of nodes extracted from APP BEL model	121
A.5	Example of reactions extracted from APP BEL model	122

List of Figures

1.1	Complexity of diseases and their comorbidities	10
1.2	BEL Statement structure	14
1.3	Disease models in SBML and OpenBEL formats	16
1.4	Overview of the project	20
2.1	Curation and harmonisation overview	26
2.2	Unstructured or semi-structured to structured data	28
2.3	Harmonised and structured datasets via tranSMART standard format files	29
2.4	Extraction Transformation and Loading of datasets	30
2.5	Linking heterogeneous data in tranSMART	31
2.6	Curated studies loaded in AETIONOMY tranSMART instance	32
2.7	Overview of the AD Cytokine dataset loaded in tranSMART	34
2.8	Distribution of <i>MCP-1</i> and <i>MIF</i> levels in AD subjects compared to controls	35
2.9	Correlation of cognitive functions with cytokine levels in AD subjects .	35

2.10	TNF- α was elevated in MCI compared to controls and AD subjects . . .	36
2.11	TNF- α in MCI subjects was reported to be negatively correlated with cognitive scores	37
2.12	Role of inflammation in neurodegeneration	39
2.13	Heatmap generated from curated data (GSE7621) loaded in tranSMART	40
2.14	Overlaying differentially expressed genes on the Parkinson's disease map	41
3.1	Element Match Decision Diagram	48
3.2	Examples of reaction match	49
3.3	Representing complexes in disease maps	50
3.4	Increase in matched elements and reaction by updating model annotation	53
3.5	Time taken per comparison	54
3.6	Reactions and elements found in AlzPathway highlighted on the PD Map	56
3.7	Reactions and elements found in PD Map highlighted on the AlzPathway	58
3.8	Perfect match in <i>AKT1</i> activity in PD Map and AlzPathway Map . . .	60
3.9	<i>AKT1</i> activity in PD Map	61
3.10	<i>TSC1:TSC2</i> activity in AlzPathway and PD Map	62
3.11	<i>MAPT</i> activity in AD and PD Map	63
3.12	PD Map: <i>MAPK</i> cascade triggered by ROS and α synuclein fibrils . . .	65
3.13	<i>MAPK8</i> signalling in AD and PD	66
3.14	ER stress signalling in AD and PD	69

3.15	Inflammation triggering transport of <i>TNF</i> and <i>IL1B</i> from astrocyte . . .	71
3.16	Reactions identified from AlzPathway highlighted in PD Map	73
4.1	Nested statements in OpenBEL	81
4.2	Example of a lossless statement conversion	84
4.3	A version of the APP OpenBEL as a map	86
4.4	APP OpenBEL model converted to CellDesigner, visualised in MINERVA	88
4.5	Elements and reaction from APP model on PD map	89
4.6	Elements from APP model highlighted on AlzPathway	91
4.7	APP and A β highlighted on AlzPathway	92
4.8	Reactions from APP model highlighted on AlzPathway	93
4.9	<i>BACE</i> and <i>APP</i> activity in Alzheimer's Disease	94
4.10	<i>GSK3β</i> as a functional link between Amyloid β and Tau pathology . .	95
4.11	<i>IL6</i> , <i>IL1B</i> and <i>TNFα</i> transported to microglia	95
4.12	<i>IL6</i> and <i>IL1B</i> transported to astrocyte	96

Chapter 1

Introduction

A large number of clinical cohorts are set up to study diseases and their mechanisms. As of May 2018, `Clinicaltrials.gov` records 450 clinical studies planned or currently recruiting subjects with Parkinson's Disease. Another 100 have completed recruitment and are still active. During a clinical study, researchers collect diverse data over several visits, e.g.: clinical assessments, biomarker tests, imaging, genetic tests, all in various formats. The study design, ontologies used and even the language of the study may differ. As a support to clinical studies being setup to study diseases, disease maps offer an approach to collect and integrate existing knowledge about the disease mechanisms, providing context to the hypotheses about the disease. Disease maps integrate multiple knowledge resources at different molecular levels and also enable visual exploration.

Current disease classifications rely mainly on the phenotypes of the diseases, especially in the case of neurodegenerative diseases like Parkinsons and Alzheimers. This is primarily due to the inherent complexity of the biological systems and partly because the aetiology of the disease is still unknown to us. Recently projects like AETIONOMY (<https://www.aetionomy.eu>) and SYSCID (<https://syscid.eu/>) focus on capitalising on the knowledge about the underlying mechanisms of the disease to

better explain the pathology of the disease [Hofmann-Apitius et al., 2015a, Schultze and Rosenstiel, 2018]. Integration of these heterogeneous resources requires a harmonised format and standards. This also helps to link different tools and subsequent analysis.

Today, the data generated by research grows both in volume and variety. This data is definitely valuable but mostly unstructured or partially structured. Making data useful requires cleaning and organising, which is both expensive and time consuming and adds to the cost at every step of data processing, from analysis to decision making. Data collected from disparate sources require harmonization for them to provide a single view, otherwise they will remain separate pieces.

Harmonisation transforms datasets such that different pieces fit together both in terms of semantics and information. Thus, it improves the power of large scale studies by facilitating integration of different analyses. Harmonising collected data increases the quality of data and the precision of resulting analysis. Utilising standards to harmonise data and knowledge bridges the gap between their representations and makes the data easily identifiable, sharable, useful and interoperable. Data sharing is one of the key components of reproducible and efficient research, to maximize the value of research. To promote good data stewardship, a community of international stakeholders have developed a set of guidelines to make data Findable, Accessible, Interoperable and Reusable, and have been widely accepted by the scientific community, and various institutions, projects and initiatives to share data and maximise its use and reuse [Wilkinson et al., 2016].

1.1 Trends in biomedical research and their impact on bioinformatics

With the advent of high-throughput technologies, the rate at which data is generated to support translational research and personalised medicine is growing. These high-throughput technologies are used to investigate distinct aspects of the cellular processes at several levels such as genome, transcriptome, proteome and metabolome. To understand complex biological systems, we need to study the effect of alterations on the genome, transcriptome, proteome and metabolome simultaneously [Horgan and Kenny, 2011]. Data integration has been reported to be an effective strategy to extract meaningful biological data from heterogeneous data sets in several fields [Xie et al., 2017, Huang et al., 2017]. It has been employed to identify candidate genes for further investigation, thereby scaling down the translation of genome-wide data into smaller list [Zhong and Sternberg, 2007]

The heterogeneity and inaccessibility across data sources are the major factors hindering formalized integration. Mandates on data sharing, considerations of standardized data collection, and mechanisms to integrate heterogeneous data are necessary to address these issues [Allen et al., 2016]. Today several efforts are in place in the scientific community to enforce and promote the use of a uniform standard, on data sharing [Wilkinson et al., 2016, McQuilton et al., 2016, Auffray et al., 2016, Wolstencroft et al., 2017].

1.2 Standards in biomedical research

Standards are an agreed or compliant term or form of representation. In other words, standards are essentially a set of rules and definitions that specify how to name or describe any entity or process. In a data-driven field like biomedicine, stan-

dards play a major role. While some standards evolve over time, it is also essential to develop them deliberately. Standards enable diverse research groups to communicate and work in co-operative and collaborative environments, especially in healthcare domain, where different groups of people work towards a common goal. For instance, in a healthcare environment, diverse groups such as patients, doctors, hospitals, biologists, statisticians, bioinformaticians, patient organisations need to work together. This requires coordination, communication and transfer of knowledge and data from one group to another. In addition, medical knowledge is complex. Thus, encoding knowledge and data using accepted standards and ontologies can reduce both ambiguity and technical challenges for data exchange and interoperability arising due to the heterogeneity of the data generated [Oemig and Snelick, 2016, Bodenreider, 2008, Smith and Brochhausen, 2010].

1.2.1 Need for standards

Most biomedical systems and resources are designed and developed independently of each other. Therefore, they do not share a common format or structure. This makes it time consuming and complex to determine the correspondences between these heterogeneous sources. Challenges arising due to the heterogeneity in the design of various databases have been previously reported in several biological research domains such as genome wide studies and gene expression studies [Zhong and Sternberg, 2007]. The process is considered both time and computationally exhaustive since different databases use different identifiers, formats and access methods. Therefore, to overcome such computational challenges researchers have preferred to obtain data from least number of sources possible [Zhong and Sternberg, 2007].

Standards facilitate re-use of data by enabling easier data sharing and reproducibility. In addition, standards also promotes interoperability across different data formats and analysis tools. In an ideal biological research environment, standards form

the base to all the higher layers of the infrastructure [Lapatas et al., 2015]. Therefore, a strong foundation is necessary to build subsequent integration and analysis tools. For data driven research goals, integration of results from different bioinformatic tools or software are required. The first step is to find the specific service or tools which are necessary to attain this goal. This may be one or more tools required in consecutive or parallel steps. The results may be generated through a web service or locally. However, in several domains including biological research, it is a well accepted fact that these tools and web services are not expected to be designed or represented using the same schema or ontologies, since they are most likely developed for different purposes or target users [Ethier et al., 2018, Shvaiko, 2005]. Henceforth, both for finding the adequate services and linking them, it will be necessary to establish the correspondences between the interpretation of the inputs and generated results [Wilkinson et al., 2016, Hammond et al., 2014]. Following standards in the design of the tools, representation and naming of the inputs, process and outputs play an important role in this process. For instance, if a service provides its result and description in an ontology and the next required service or tool uses a different ontology for its input, matching both ontologies and formats are essential for

- a) ensuring what is delivered by the first service, matches what is required by the second
- b) verifying prerequisites of the second service or tool, and
- c) creating a middle layer which acts as an interface to transform the output of the first service such that it can be the input expected by the second service or tool.

Most scientific studies are built on previous findings. The scientific process therefore depends heavily on the reproducibility and interoperability of results. In 2011, a team at Bayer Health Care in Germany investigated 67 in-house projects. They reported that only about 25% of published preclinical studies could be validated for

further investigation [Prinz et al., 2011]. Ioannidis et al., in 2009 reported the reanalysis of 18 articles published in *Nature Genetics* on comparative analysis of microarray experiments. Only two analyses could be reproduced in principle and six partially or with some discrepancies. The main reason for failure, was reported as data unavailability, incomplete data annotation or specification of data processing and analysis [Ioannidis et al., 2009]. A lack of transparency and standards leads to loss in resource and time to replicate these results [Baker, 2016, Begley and Ellis, 2012].

1.2.2 Standardisation efforts

Several initiatives coordinate and collaborate on the consolidation and creation of standards. They also play a major role to advocate the use of standards in biological research at all levels. The Findability, Accessibility, Interoperability, and Reusability (FAIR) principles for scientific data management and stewardship were established to enhance the ability of machines to automatically find, access, exchange and use data. FAIR stands for the four foundational principles Findability, Accessibility, Interoperability, and Reusability [Wilkinson et al., 2016].

Clinical data interchange standards consortium (CDISC)

The Clinical Data Interchange Standards Consortium (CDISC) works on developing standards in clinical trials data and metadata (<https://www.cdisc.org/>). The CDISC mission is "*to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare*". These standards facilitate the acquisition, submission, exchange, and archiving of clinical trial data. For instance, the CDISC standard for acquisition, Clinical Data Acquisition Standards Harmonization (CDASH), aims to improve interoperability in clinical research and drug development processes. The Study Data

Tabulation Model (SDTM), provides a standard for organising and formatting data to streamline processes in collection, management, analysis and reporting. Currently, SDTM is one of the required standards for data submission to FDA (U.S.) and PMDA (Japan).

MIRIAM and Identifiers.org

In order to ensure re-usability of biological models the computational biology community proposed a set of guidelines, the Minimum Information Required in the Annotation of Models (MIRIAM) [Le Novère et al., 2005]. These guidelines describe not only the need to unambiguously and perennially identify components in the model, but also required meta-information such as provenance and development. The MIRIAM Registry (currently Identifiers.org), available at <http://identifiers.org/registry> provides such a centralised, unique, perennial and location independent identifiers for use in the biomedical domain. The registry is catalogue of data collections. Each data catalogue is associated with a unique namespace and extensive metadata. This namespace then allows the generation of Uniform Resource Identifiers (URIs) to uniquely identify any record in the collection. To increase usability, Identifiers.org [Juty et al., 2012] provides a service which provides directly resolvable identifiers, in the form of Uniform Resource Locators (URLs). The flexibility of the identification scheme and resolving system allows its use in many different fields, where unambiguous and perennial identification of data entities is necessary. Many ontologies and databases currently use these URIs, including Reactome [Croft et al., 2011], BioModels Database [Li et al., 2010], OpenPHACTS [Williams et al., 2012] and Bio2RDF [Belleau et al., 2008].

Computational Modelling in Biology Network (COMBINE)

The growing model sizes and their complexities make it necessary to standardise forms of representations [Waltemath and Wolkenhauer, 2016a]. Standardising modelling formats is essential for largescale modelling, eases sharing results and permits other researchers to use, re-use them [Hucka and Finney, 2005]. The Computational Modelling in Biology Network (COMBINE, <http://co.mbine.org>), guides the development of standards for modelling in computational biology. COMBINE helps to coordinate common activities and to establish a common infrastructure by fostering communication between the various standardization efforts [Waltemath et al., 2015]. COMBINE supports both, mature standards and emerging efforts, in covering the current needs in the interoperability landscape. The network identifies missing standards and promotes further developments for the exchange of modelling and results [Hucka et al., 2015]. The COMBINE, currently covers standards for CellML [Lloyd et al., 2008], SBML [Hucka et al., 2003], SBOL [Galdzicki et al., 2014], BioPAX [Demir et al., 2010], SEDML [Waltemath et al., 2011], NeuroML [Gleeson et al., 2010] and SBGN [Novère et al., 2009]. In addition, projects such as FAIRDOM (<http://fair-dom.org>) develop management guidelines and infrastructure for collaborative modelling. They also offer curation, training, and run workshops and summer schools to promote these standard settings within the systems biology community.

Disease Maps Community (DMC)

Disease maps are emerging concept, providing computationally readable yet comprehensive knowledge-based resource of disease mechanisms. Disease maps visually represent hallmark pathways and biological processes associated with the disease [Mizuno et al., 2012, Fujita et al., 2014, Kuperstein et al., 2015]. Disease maps bring together domain experts from bioinformatics, molecular biology and clinical research.

To ensure the interoperability of disease maps, it is essential to adopt relevant standards for knowledge encoding and annotation [Mazein et al., 2018]. Also, appropriate tools are needed to support creation and use of the maps. The DMC (<http://disease-maps.org/>) brings together developers and users of disease maps. The community was formed to identify challenges by exchanging experiences from the different disease maps' projects [Ostaszewski et al., 2018]. The community aims to establish best practices for creation, maintenance and application of disease maps.

1.3 Disease maps as knowledge resources

Systems biology is a data driven domain, with rapid generation of data about the individual components such as genes, proteins, chemicals, diseases, cell types and organs [Greene and Troyanskaya, 2010]. To understand complex biological systems and diseases, we need to bring into context available data to detect relations, pattern and links between the individual components, allowing us to formulate and validate scientific hypotheses [Kitano, 2002]. The existing knowledge is distributed over different databases. Disease maps are one such method to integrate knowledge about the disease mechanisms from literature and different databases into a single resource and to add context to the knowledge, by organising it into a structured and organised network. Disease maps integrate and annotate knowledge from different molecular mechanisms and biological pathway relevant to the disease into a computer-readable format and enable visual exploration.

1.3.1 Complexities of diseases and their comorbidities

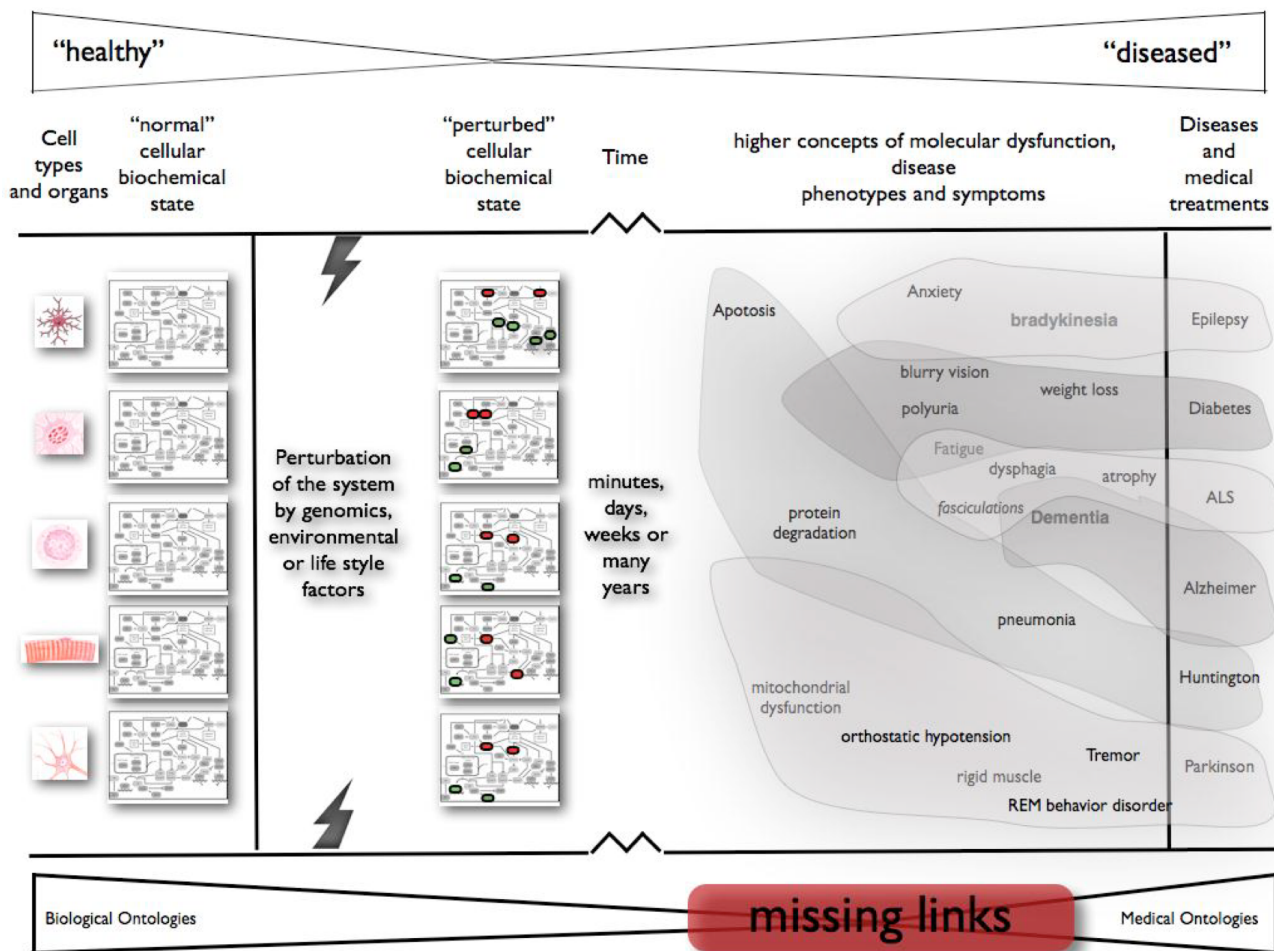


Figure 1.1: Complexity of diseases and their comorbidities
 Source: Dr. Reinhard Schneider

Complex diseases like neurodegenerative diseases are affected by several factors. The Figure 1.1 illustrates the knowledge gap between the high-level medical ontologies, describing the disease state, and biological models and ontologies, encoding molecular processes, and their perturbation by genomic, environmental or lifestyle factors. There is a missing link between the way we represent the biological knowledge, diseases and their treatments. To bridge this link, disease maps integrate the knowledge about disease mechanisms from literature in a single resource.

1.3.2 Representing diseases as a map

Disease mechanisms can be modelled as a molecular interaction graph, i.e. nodes that are connected by edges. Each component of the biological system and other factors contributing to the disease pathology are represented as a node. Each node is annotated by a unique identifier. Localisation of the interaction and nodes are represented as clusters or compartments providing the map a hierarchical organisation. The relationship or interaction between these "factors" are modelled as edges connecting the node. Curating a disease-related pathway comprises of identifying and structuring content, mining for information either manually or computationally, or both, and building a knowledge base using appropriate software [Viswanathan et al., 2008].

Representation of molecular pathways requires a format for modelling that is computable and allows for exchange, integration. Several such formats exist, varying in representation depending on their purpose.

1.3.3 Common modelling formats

Most widely used pathway-related formats [Strömbäck et al., 2006], are XML-based. We would like to focus on three important notations, namely Systems Biology Markup Language (SBML) [Hucka et al., 2003], Biological Pathways eXchange (BioPAX)[Demir et al., 2010] and Open Biological Expression Language (OpenBEL) (<http://openbel.org/>).

BIOlogical PATHway eXchange language (BioPAX)

BioPAX (Biological Pathway Exchange) is a standard language to facilitate exchange biological pathway data at the molecular and cellular level. It is defined in Web Ontology Language (OWL) and represented in XML [Demir et al., 2010].

BioPAX has a large user base and is supported by many pathway databases such as Reactome [Fabregat et al., 2018], Panther [Mi et al., 2017] and network visualisation and analysis tools such as cytoscape [Shannon et al., 2003]. BioPAX was created through a community process and continues to be an open and collaborative effort.

Systems Biology Markup Language(SBML)

SBML [Hucka et al., 2003] is a software-independent language used to build models in the computational biology domain. SBML is used mainly for modelling, it can also be used for pathways representations including metabolic pathways, gene regulation, and cell signalling pathways [Caron et al., 2010]. As of May 2018, SBML is supported by over 280 software systems (http://sbml.org/SBML_Software_Guide). With greater support and interaction between tools, and a common format like SBML, users would be better able to spend more time on actual research rather than on complying with data format issues. As of May 2018, the BioModels (<https://www.ebi.ac.uk/biomodels-main/>) lists 8428 SBML models.

BioPAX and SBML are two of the most commonly used format in the systems biology modelling domain and are supported by a wide base of user and developer community to make them interoperable [Büchel et al., 2012, Rodriguez et al., 2016]

Open Biological Expression Language (OpenBEL)

There exists numerous modelling languages and formats for modelling biological knowledge as networks. However many require at least a basic understanding of programming knowledge to use and are generally not adopted by biologists and clinicians. In recent years, with the explosion of data and knowledge in the biomedical domain, it is important to develop tools that can foster collaboration between experts in different domains. One of the most important features of OpenBEL is that

it is both human readable and computable. The subject and object are annotated by namespaces, in this case MeSH Disease and Gene Ontology respectively. BEL focuses on representing the causal and correlative relationship between entities. These entities can be biological entities such as proteins, genes, RNA, etc or chemicals, complexes or phenotypes. The relationships represent primarily cause-effect events between these entities. BEL also captures the provenance of the relationships, at the statement level.

BEL statements are modelled as a semantic triple. The subject and object are connected by the predicate which describes their relationship. In this example (Listing 1.1 and Figure 1.2), we can see that from the statement the subject *Atherosclerosis* has a *positive correlation* with the object *lipid oxidation*.

```
SET Disease = "Atherosclerosis"
SET CardiovascularSystem = "Arteries"
SET TextLocation = "Review"
SET Evidence = "Oxidation and nitration of macromolecules, such as
  proteins, DNA and lipids, are prominent
  in atherosclerotic arteries."
SET Citation = {"PubMed", "Trends in molecular medicine", "12928037",
  "", "de Nigris F, Lerman A, Ignarro LJ, Williams-Ignarro S, Sica
  V, Baker AH, Lerman LO, Geng YJ, Napoli C", ""}
pathology(MESHD:Atherosclerosis) positiveCorrelation
  biologicalProcess(GO:"lipid oxidation")
```

Listing 1.1: BEL Statement Example

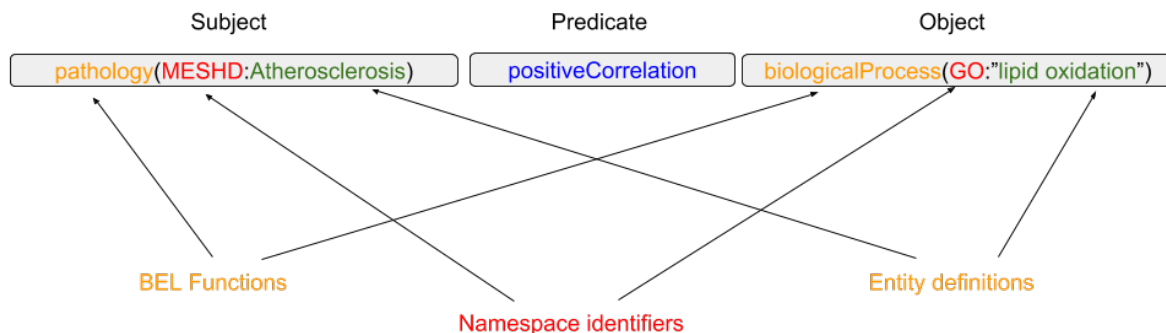


Figure 1.2: BEL Statement structure

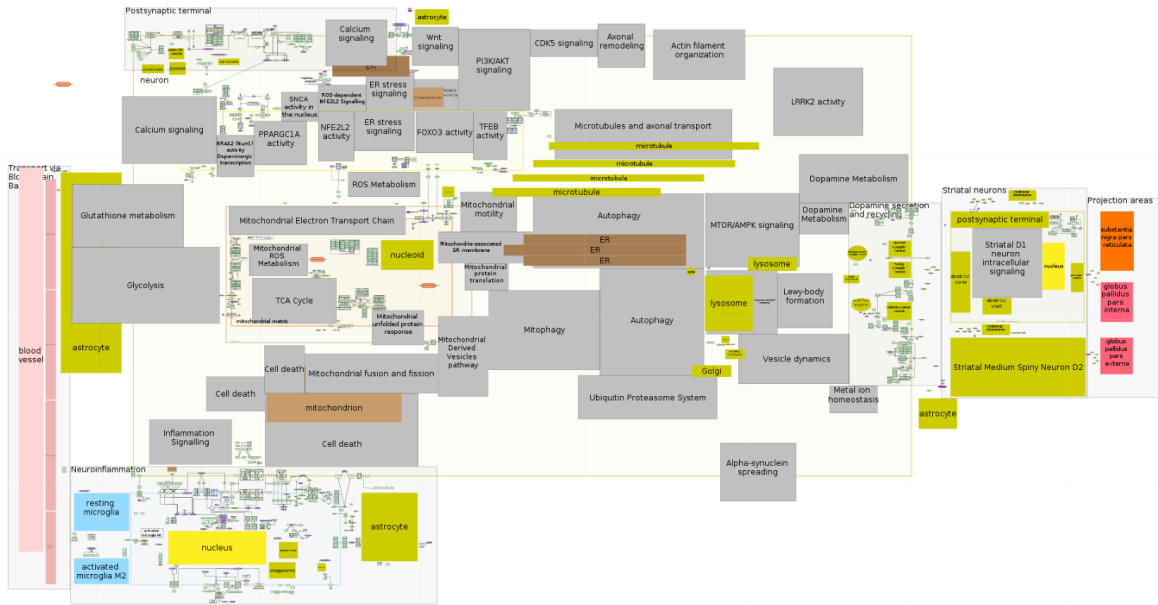
Representation of BEL terms as functional expressions, helps to make the language concise. The concept of abundance provides a systematic way to represent an unknown quantity of biological material or activity and how its activity can increase or decrease in the system. This allows a qualitative representation of biological knowledge.

OpenBEL is one of the popular modelling languages among biologists primarily due to its simplicity and resemblance to natural language. It has been widely accepted by the research community especially because of its simplicity and short learning curve. Although it is close to natural language it is still a computable model. As a result, it has been adopted by various crowd sourcing challenges to build and generate networks collaboratively [Namasivayam et al., 2016]. Additionally, in recent years several text mining and information extraction challenges and tasks have also adopted OpenBEL [Fluck et al., 2015, Fluck et al., 2016] [Lai et al., 2016].

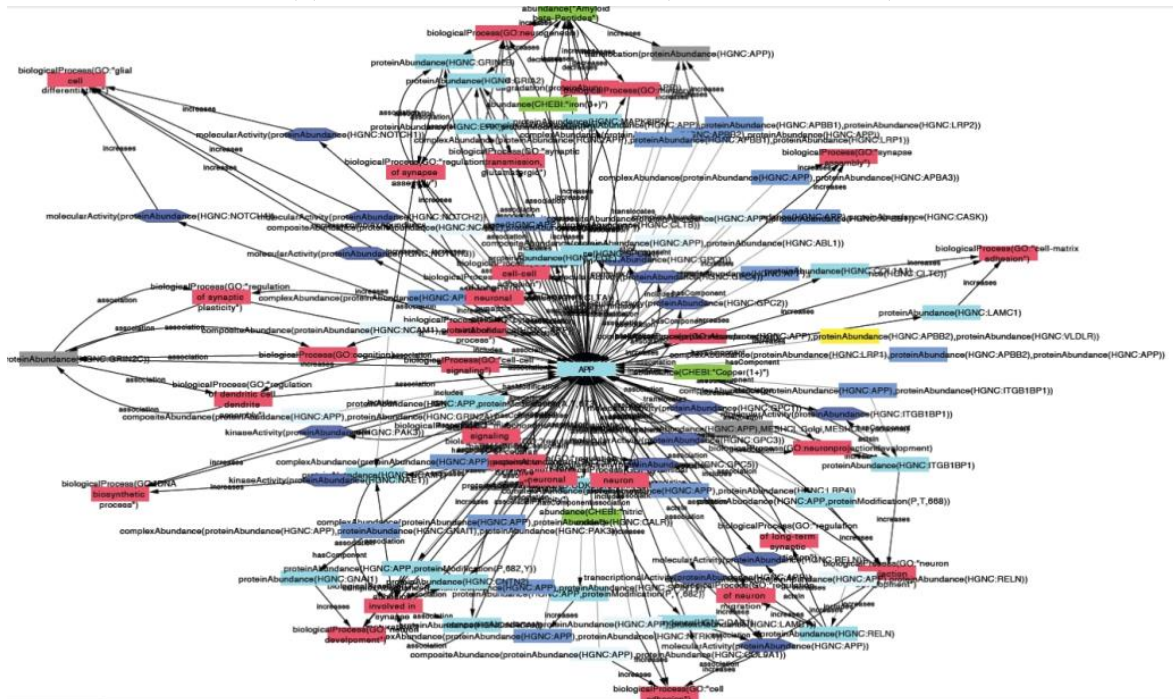
	SBML	BEL	BioPAX
Inventors	Systems Biology Workbench	Selventa: OpenBEL	BioPAX group
Focus	Process description	Entities and causal relationship	More general representation, Focus on reaction/interaction
Tools	Validation, visualisation, conversion and modelling e.g. CellDesigner, Cytoscape (BiNoM plugin)	BEL Framework: converter to XML, Validation, Visualisation by Cytoscape, PyBEL	Validation, visualisation, conversion and modelling e.g. Protege, ChiBE, BioLayout, Cytoscape (BiNoM plugin)
Interactors	Species	Subject/Object	PhysEntity
Interactions	Reactions	Relationship	Reaction
Role of Interactors	Reactants, Products or Modifiers	Subject, Object	Pathway representing set of interactions
Mathematical relations	Yes	No	No
Inheritance	Yes	No	Yes
New entities	Unknown type	abundance	Possible to make application specific additions

Table 1.1: Features of SBML, OpenBEL, BioPAX

Table 1.1 gives a summary of the features of SBML, OpenBEL and BioPAX (Source: [Strömbäck et al., 2007], <http://sbml.org/>, <https://binom.curie.fr/>, <http://openbel.org/>, <http://www.biopax.org/>).



(a) Parkinson's Disease Map (SBML compliant)



(b) Amyloid Precursor Protein normal physiology model (BEL)

Figure 1.3: Disease models in SBML and OpenBEL formats

Although BEL models, capture the context information, this information is not utilised for its visualisation via Cytoscape (Figure 1.3b) [Shannon et al., 2003]. SBML, on the other hand can graphically represent cellular location and pathways with the help of diagram editors like CellDesigner (Figure 1.3a) [Funahashi et al.,

2003, Funahashi et al., 2008]. Visualising context or location information is essential to the concept of disease maps for navigational capabilities similar to geographical maps. Currently, the concept of disease maps is implemented in domains such as cancer [Kuperstein et al., 2015], influenza [Matsuoka et al., 2013] and neurodegenerative diseases [Fujita et al., 2014] [Mizuno et al., 2012].

So far all the publicly available maps are constructed using CellDesigner in SBML and notations based on the process description of Systems Biology Graphical Notation (SBGN) [Novère et al., 2009]. CellDesigner is a diagram editor for drawing gene-regulatory and biochemical networks. Diagrams are drawn based on the process diagram, with graphical notation system proposed by Kitano et. al. [Kitano et al., 2005], and are stored in an SBML-compliant format. To support efficient navigation and management of community driven curation of these maps platforms such as the NaviCell [Kuperstein et al., 2013, Bonnet et al., 2015] and MINERVA [Gawron et al., 2016] are available. NaviCell is a platform for exploring large maps of molecular interactions built in CellDesigner. NaviCell features efficient navigation, semantic zooming of the map for viewing different levels of details. Additionally, it also provides support for collecting curation feedbacks from the community. MINERVA [Gawron et al., 2016] (Molecular Interaction NETwoRks VisuAlization) platform is a web service supporting curation, annotation and visualization of molecular interaction networks in Systems Biology Graphical Notation (SBGN)-compliant format. MINERVA also supports automated content annotation and verification, thereby improving the quality of the maps. Both these platforms use the Google Maps API for semantic zooming and navigation of the maps.

All the elements (proteins, genes, RNA, chemicals, metabolites, etc.) in a map should ideally be annotated by publicly available databases such as UniProt [Magrane and Consortium, 2011], HGNC [Gray et al., 2015], Ensembl [Yates et al., 2016], Entrez Gene [Maglott et al., 2011], KEGG [Kanehisa et al., 2012], Reactome [Croft

et al., 2011], Gene Ontology [Gene Ontology Consortium, 2000], ChEBI [Hastings et al., 2013]. Annotation of the contents of a map facilitates the knowledge exploration by providing additional information about the elements and their interactions.

1.3.4 Available data

While several such disease modelling formats are available, for the scope of this project we focus on the SBML and BEL modelling formats. As an use case for neurodegenerative diseases we use the Parkinson’s disease (PD) Map [Fujita et al., 2014] (1.3a), Alzheimer’s disease map (AlzPathway) [Mizuno et al., 2012] and Amyloid Precursor Protein (APP) BEL model [Kodamullil et al., 2015]).

The AlzPathway and PD Map are both built in CellDesigner and hosted on MINERVA platform. All the reactions in these maps have evidences referenced by PubMed identifiers using the MIRIAM uri.

The PD map integrates and visualises molecular interactions within a cellular context with a focus on processes associated in PD pathology such as synaptic and mitochondrial dysfunction, α -synuclein pathology, impaired protein degradation, and neuroinflammation. It is also the first freely accessible and manually curated knowledge repository of Parkinson’s Disease.

AlzPathway is the first comprehensive and manually curated map of intra, inter and extra cellular signaling pathways of AD. It is also available as the web service (online map) implemented on Payao [Matsuoka et al., 2010], a community-based, collaborative web service platform for pathway model curation.

The APP model is built in BEL to systematically model causal and correlative relationships between bio-entities. The model was built around knowledge about physiological functions and pathological responses of amyloid precursor protein (*APP*). BEL disease models were also used to perform comorbidity analysis between

Alzheimer's disease and Type 2 Diabetes Mellitus based on shared pathways and role of drugs [Kodamullil et al., 2015, Karki et al., 2017].

1.4 Scope and Aim

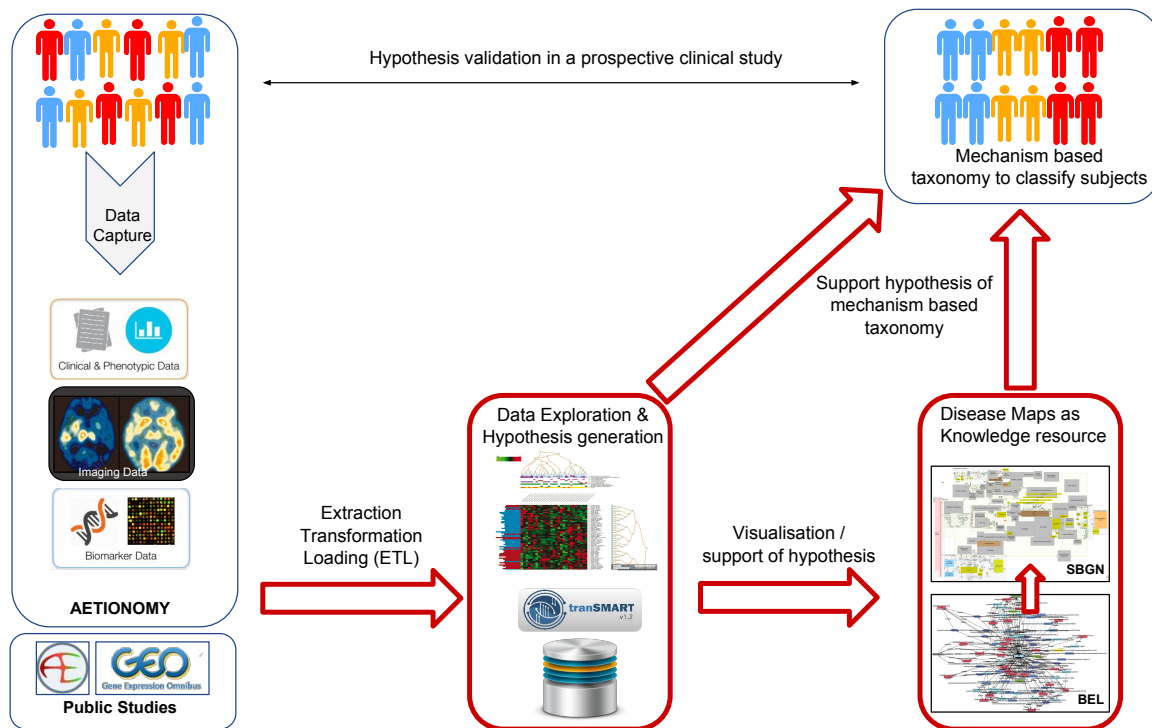


Figure 1.4: Overview of the project

The AETIONOMY project, integrates publicly available data and knowledge with proprietary data across several scales such as clinical, omics, imaging to identify candidate mechanisms and propose a mechanism based taxonomy for Alzheimer’s and Parkinson’s disease, which will be validated in a prospective clinical study (Figure 1.4). As part of the PhD, I will focus on the data harmonisation and linking this heterogeneous data. This includes the publicly available and clinical data from the AETIONOMY project with associated transcriptomics data and support the hypotheses generation. In addition to the clinical data, I will also focus on interoperability and comparison of maps and their application.

The project described in this thesis was built on three main objectives. First, the integration and harmonisation of heterogeneous publicly available and consortium data to support hypothesis generation. Second, to implement comparison of two SBML

maps to enable co-morbidity studies. Third, to implement a convertor from BEL model to SBML models to take advantage of the hierarchical organisation of SBML maps. Overall, the thesis will highlight the need to make data generation and sharing standard and harmonised to facilitate integrative and data-driven research.

1.5 Thesis Overview

Several data and knowledge repositories are constructed to study disease progression and mechanisms. Such datasets are available in public repositories or as proprietary datasets. Since, the data collected are from different sources and in different formats, the data is either unstructured or semi structured.

Chapter 2, describes how publicly available and consortium datasets were integrated. An ETL (extraction, transformation and loading) pipeline is utilised to harmonise and integrate into the translational medicine platform, in this case transMART. The chapter describes how using transMART enables to explore and analyse integrated clinical and associated molecular (-omics) data and facilitates hypothesis generation.

Contributions: Aishwarya Alex Namasivayam (AAN), curated and integrated all the datasets, except the proprietary datasets (PPMI and ADNI), and public PD studies (inkind from eTRIKS). The AETIONOMY consortium provided all other project datasets. AAN and Adriano Barbosa da Silva (ABS), supported the project with data acquisition and management and analytical tools integration. Reinhard Schneider (RS) supervised the project.

Chapter 3, describes the comparison of SBML maps using the PD Map and AlzPathway. Similarities are discussed with several examples. Also discussed are the challenges due to differences in annotations and advantages of a harmonisation.

Contributions: AAN, Piotr Gawron (PG), Marek Ostaszewski (MO) and Reinhard

Schneider (RS) planned and designed the project. The comparison was implemented by AAN and supported by the MINERVA platform (PG, MO). Stephan Gebel (SG) contributed to interpreting the biological relevance of the detected similarities.

Chapter 4, describes the conversion of BEL model to SBML maps using the APP BEL model as a use case. The methods and challenges are discussed. The converted map was compared to the AlzPathway and PD map using the methods discussed in Chapter 2.

Contributions: AAN and MO designed and planned the project. The convertor was implemented by AAN and supported by MO to convert to the Cell Designer format and hosted on the MINERVA platform.

Chapter 5, discusses the results and lessons learned. **Chapter 6** summarises the thesis and provides future directions.

Chapter 2

Integrating heterogeneous data

Biological systems are complex, with many levels of regulation and interaction [Conesa and Mortazavi, 2014]. Large amounts of biological data are generated and collected to investigate these individual levels, but a more comprehensive understanding of the system requires the integration of these data, allowing analytical approaches to describe relationships between the components [Gomez-Cabrero et al., 2014].

2.1 Integrative platforms

A large amount of omics data is generated by high-throughput technologies from a broad spectrum of domains. These omics data need to be considered in the context of the phenotype and diseases to achieve their full potential.

Translational platforms enable efficient data sharing and integration. Moreover, it increases the quantity of data available in a common format for research facilitating interoperability and comparability of the data. Efficient data integration also requires that translational research platforms can utilise existing data collection processes within the institutions. Platforms should provide reusable ETL pipelines to

handle not only research data (e.g. text or spreadsheets) but also standard omics and clinical data formats.

Currently several translational research platforms are available to integrate clinical and omics data. [Canuel et al., 2015]. The cBio Cancer Genomics Portal (cBioPortal) [Gao et al., 2013], is an open-source platform facilitating the access to data sets generated by large-scale cancer genomics projects, like International Cancer Genome Consortium (<http://icgc.org/>) and The Cancer Genome Atlas (<http://cancergenome.nih.gov/>). It integrates pseudonymised clinical data with genomics, transcriptomics and proteomics [Cerami et al., 2012]. Integrative analysis approaches have utilised such platforms in several cancer studies [Gao et al., 2013, Rance et al., 2016].

TranSMART [Szalma et al., 2010], is another translational research platform that integrates powerful visualization and interoperability functionalities of Informatics for Integrating Biology and the Bedside (i2b2) platform [Murphy et al., 2006]. tranSMART is a well-accepted platform in translational medicine research [Athey et al., 2013, Schumacher et al., 2014, Bauer et al., 2016]. It facilitates integration of low-dimensional clinical data and high-dimensional transcriptomics data.

2.2 Data acquisition

To address the challenge of ensuring smooth and efficient entry of datasets into the AETIONOMY knowledgebase (<http://aetionomy.scai.fhg.de/>, <https://aetionomy.uni.lu/transmart>), the AETIONOMY consortium uses a study request system (<https://aetionomy.uni.lu/StudyRequest/>). A user interested in bringing a specific study into the knowledgebase starts by sending a request to the system, which then follows several stages. First, the project office approves the inclusion of the dataset, followed by the legal team's review of legal and ethical principles concerning the usage of the dataset. Once the dataset is reviewed for inclusion. The data acquisition

and curation process is initiated. Finally, the dataset is loaded into the Data Cube and then the study requester is informed of its availability.

2.3 Data harmonisation

Unstructured data predominantly contains free text and are difficult to analyse, whereas structured data can be easily extracted for analysis and research because the data elements are comparable. Data can be harmonised using a controlled vocabulary such as CDISC, SNOMED-CT. Lack of harmonised naming conventions and structured meta-information are the main reasons for the lack of semantic integration in the life sciences. Data cleansing is often necessary to bring consistency to different sets of data that have been merged from separate clinical sites or databases. Cleansing data involves consolidating data within a database by identifying and correcting inconsistent data and removing duplicates in order to achieve concise and accurate data resource.

The data sources available to the consortium were systematically explored and approved for inclusion to the knowledge base. The project data to be integrated into the platform included CSV (comma-separated values), excel worksheets and additional non-standardized study data. The raw files retrieved from public databases and received from data providers undergo a curation and harmonization phase. They were first converted to a tab separated format. Each dataset, then undergoes a curation and harmonisation process which involves several steps (Figure 2.1)

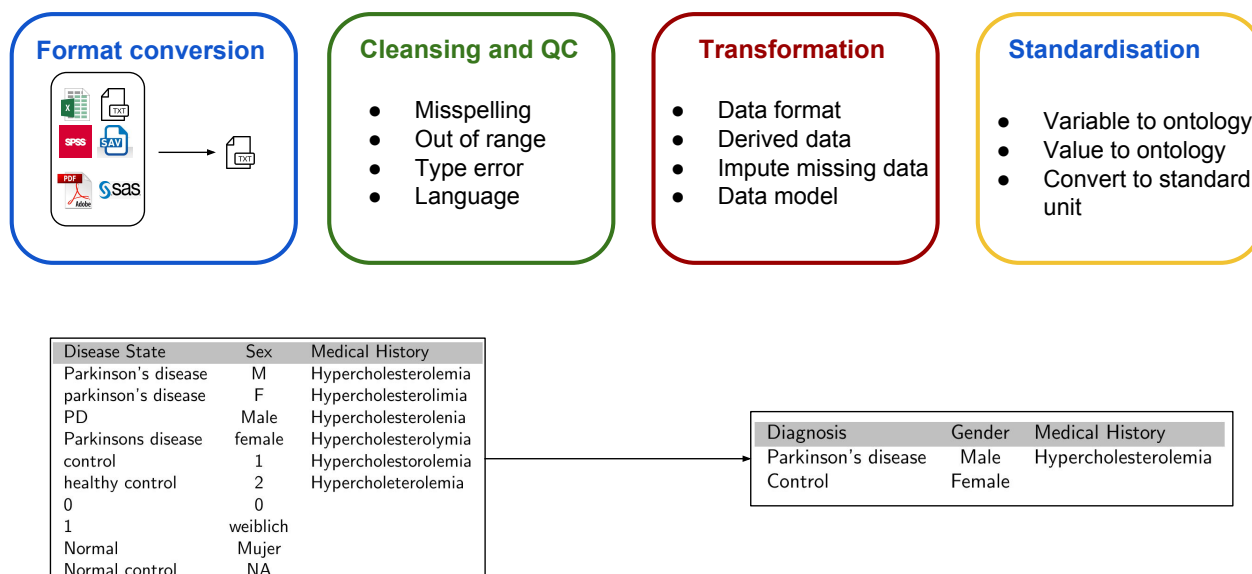


Figure 2.1: Curation and harmonisation overview

Integration of omics and clinical research data is not straightforward specifically because clinical research data collection is often non-standardised. Clinical research and cohort data are difficult to handle, mainly due to their non standardised parameters and varying representations over time and study centres. Careful data cleaning and preparation are necessary prerequisites to any process involving integrated biological data. Often the clinical data was not accompanied by a data dictionary and therefore required constant follow up with the collaborators and additional manual effort to map the variables and values to a standard ontology.

2.4 Extraction, Transformation and Loading (ETL)

The AETIONOMY project, uses tranSMART to integrate publicly available omics datasets on Alzheimer's and Parkinson's disease from Gene Expression Omnibus (GEO) and Array Express and consortium datasets [Hofmann-Apitius et al., 2015b]. tranSMART offers visually aided data exploration and drag and drop enabled cohort selection.

Integration of heterogeneous datasets require extraction, transformation, and loading (ETL) processes to harmonise the representation. Data can be added to the tranSMART database by mapping the variables to a data-scheme via standard templates or mapping files. The mapping files for the curated data files are then generated. The mapping files are generated to follow the tranSMART standard files for the ETL scripts. Additional data can be associated on the subject level data and linked via these mapping files. For instance, for datasets which include expression data, additional files for the platforms used for the experiment have to be generated. These platform mapping files enable the mapping of probe ids from the platform to its corresponding GeneID and Gene Symbol.

The ETL process to load data in tranSMART ensures that all integrated data makes use of unique identifiers and provides a uniform structure. In addition to the benefits of integrating heterogeneous data it also enables easy sharing of data in the future. This structured and standardised structure fosters data exchange in the scientific community, which is also a pre-requisite for many translational medicine projects and multi-subject expert teams [Maier et al., 2011].

In the example (Figure 2.2), the data collected were in different formats over different files and languages, etc. Though valuable, they are disparate and provide no structure. These have to be transformed for further analysis as a single dataset. After data curation and harmonisation we load them in tranSMART. This then gives a structure to the dataset, allowing it to be explored and shared easily.

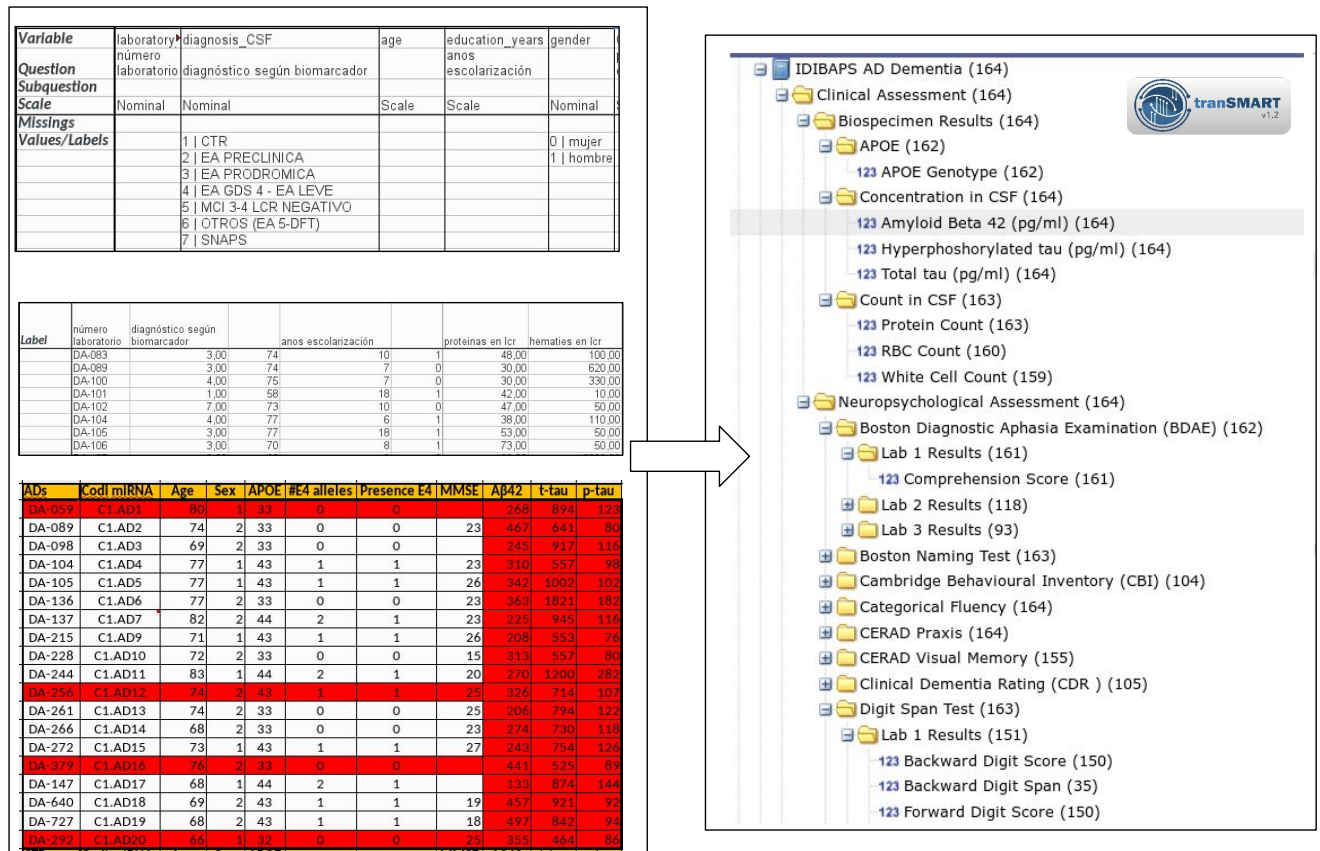


Figure 2.2: Unstructured or semi-structured to structured data

2.5 From unstructured to structured data

Structured data is data that can be easily stored, queried, exported, and analysed by computational methods. Clinical research data is often unstructured or semi-structured data like medical records, handwritten notes. Healthcare applications require efficient ways to integrate and convert a variety of data including automating conversion from structured to unstructured data. To take advantage of the various functionalities of tranSMART, the data sets have to go through extraction, harmonization, curation, and quality checking. Data acquired from publicly available databases or resources are usually structured or semi-structured. However, they may require transformation to retrieve relevant information to be integrated into tranSMART. Next, a set of standard format files need to be generated to map each subject to sample level

data. In addition to harmonising the raw data, metadata is also annotated by relevant ontologies.

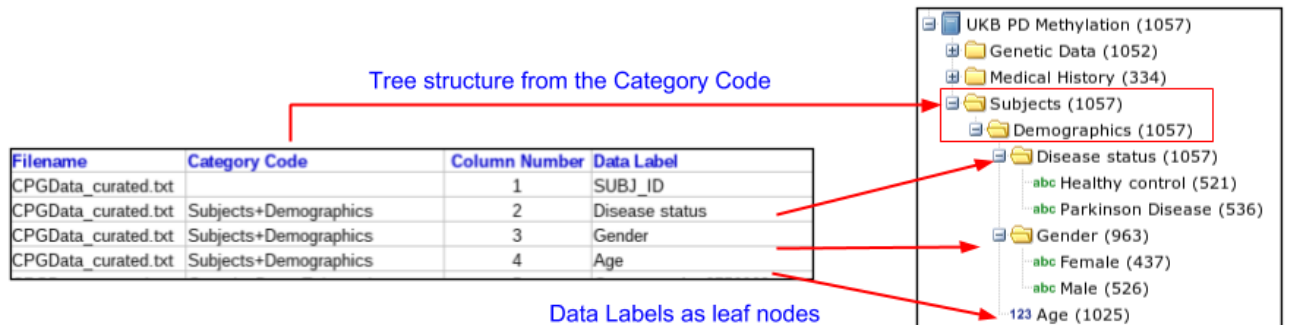


Figure 2.3: Harmonised and structured datasets via tranSMART standard format files

Figures 2.3 and 2.4 show the how the ETL process controls the structure of the variables loaded. These variables, are represented as a hierarchical parent-child tree (Figure 2.3). This tree structure allows efficient data sets exploration and also the selection of variables from the hierarchy to build customised patient cohorts by visual exploration and for further analysis or export. Variables in the dataset, such as age, gender, or measure of a blood marker could be used as filters to build a sub cohort.

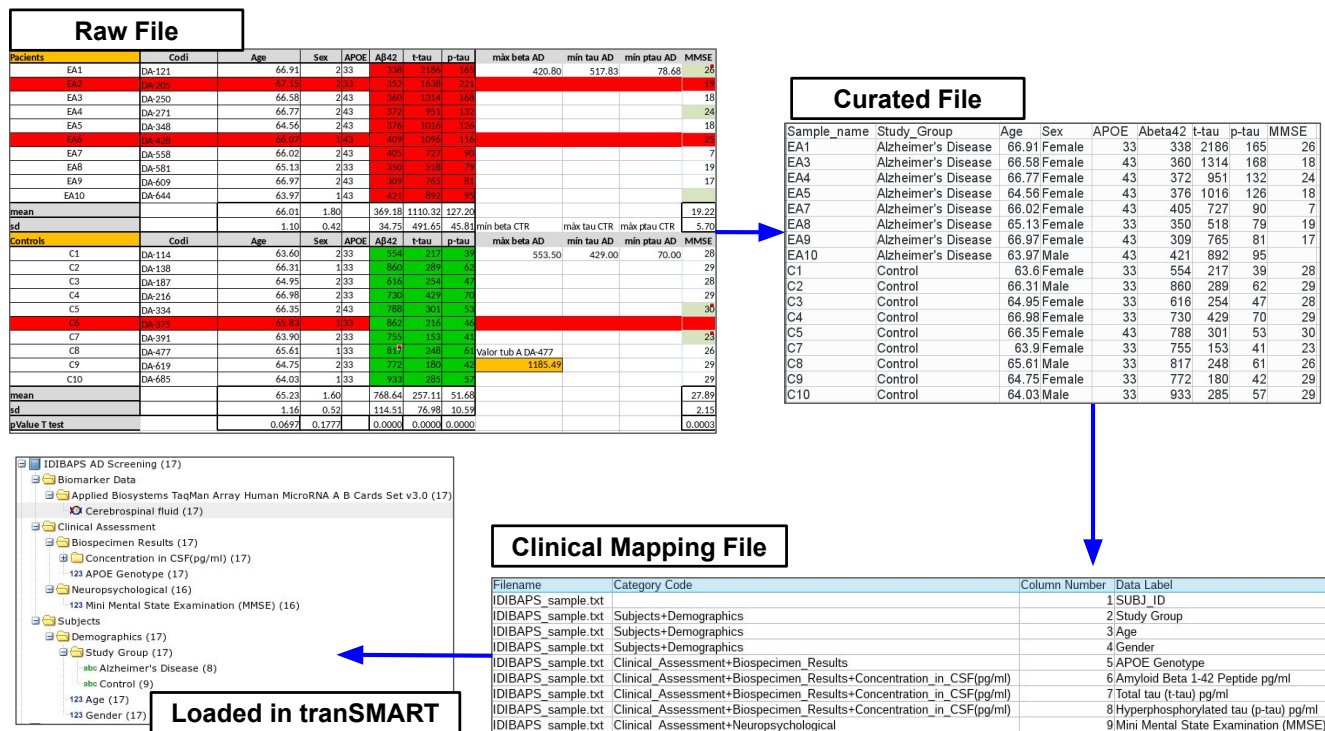


Figure 2.4: Extraction Transformation and Loading of datasets

The mapping file generated to load the datasets into transMART also generate the i2b2 tree structure for the study. Therefore, the category for each variable and the naming of the features are assigned at this stage. The mapping files are hence responsible for structuring the different studies in the AETIONOMY knowledge base. Each feature collected across studies should eventually be assigned to the same category and leaf node for every study loaded. Features specific to a single study, however will have a new leaf node, nevertheless the structure of the tree (in terms of category and branching) can be harmonised to the extent possible.

2.6 Results

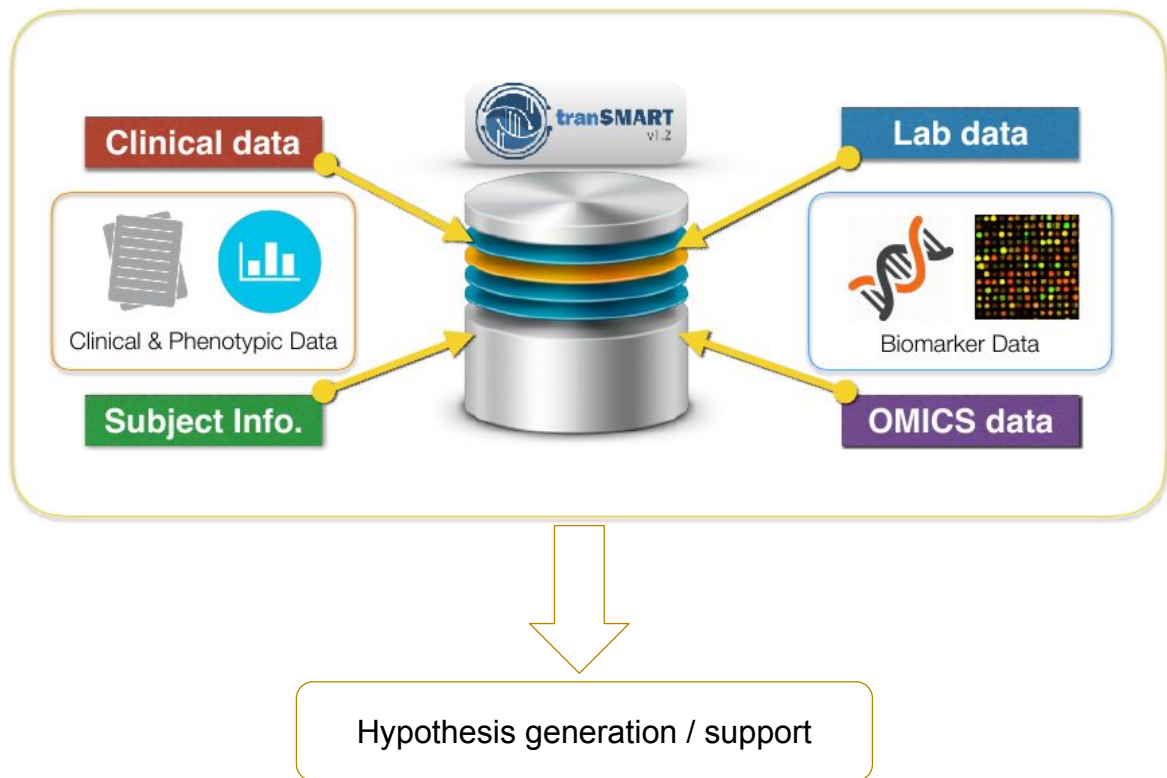


Figure 2.5: Linking heterogeneous data in transSMART

Platforms like transSMART helps to integrate disparate datasets to analyze them for support of research hypotheses [Hofmann-Apitius et al., 2015b] (Figure 2.5). transSMART serves as a collaboration platform by integrating data from heterogeneous sources. It enables code free data exploration and interactive visual analytics, and thus brings together researchers from different areas of expertise (biologists or clinicians and bioinformaticians or statisticians) [Satagopam et al., 2016]. The data can also be easily exported for further in depth analysis. This chapter demonstrates with two examples, how the curated and loaded data on transSMART enables hypothesis generation and support.

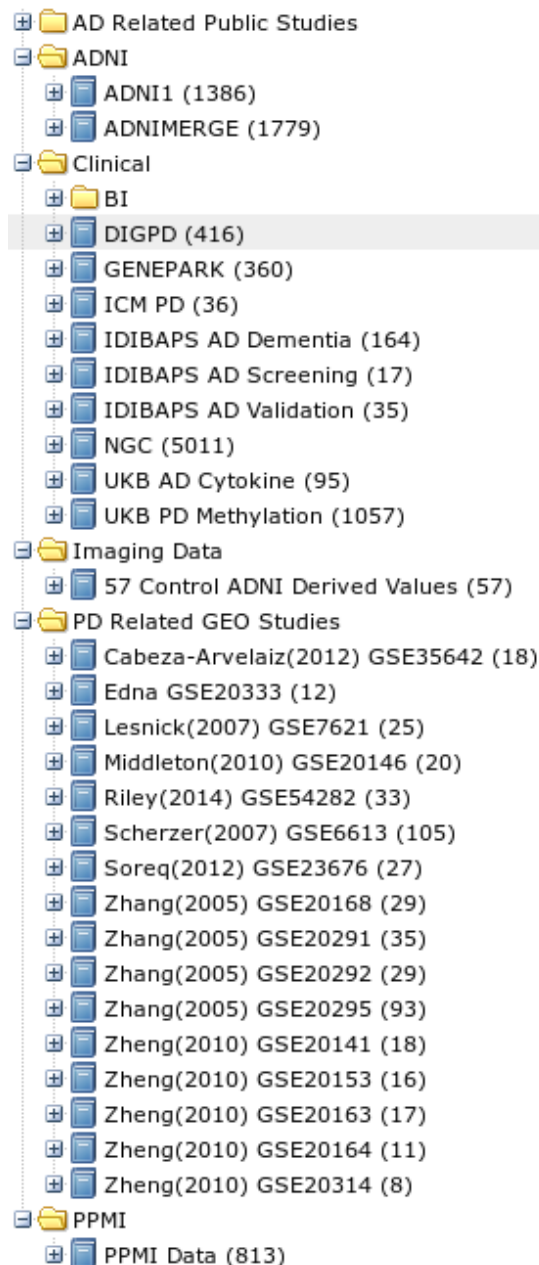


Figure 2.6: Curated studies loaded in AETIONOMY tranSMART instance

Figure 2.6, shows a number of Parkinson’s disease and Alzheimer’s disease datasets were curated and loaded into the tranSMART system for the AETIONOMY project, this includes publicly available studies from Gene Expression Omnibus (GEO) and Array Express, proprietary datasets such as PPMI and ADNI and studies from the consortium. The variables include clinical and neuropsychological assessments, biospecimen analysis results, imaging and transcriptomics data. tranSMART enables

to bring together these different sources in a common format for exploration, basic analysis and sharing of data.

2.6.1 Use case 1: Alzheimer's disease cytokine study

Alzheimer's disease (AD) is the most prevalent cause of dementia and characterized by cognitive deficits, neuronal death and, ultimately, severe brain atrophy. At the molecular level, the hallmarks of AD are extracellular plaques of amyloid β ($A\beta$) and intracellular tangles of tau protein. Neuroinflammation represents a further characteristic feature of neurodegenerative diseases. To limit further $A\beta$ accumulation, the microglia and astroglia are reported to react to $A\beta$ exposure by phagocytosis, and also by prolonged release of inflammatory mediators creating a neurotoxic environment. Presuming $A\beta$ aggregation precedes the onset of clinical symptoms by decades, the innate immune activation may be an early and contributing progress in AD pathogenesis.

The AD cytokine dataset was collected to determine cerebrospinal fluid (CSF) levels of interleukin 6 (IL-6), tumor necrosis factor α (TNF- α), monocyte chemoattractant protein 1 (MCP-1 / CCL2), and macrophage migration inhibitory factor (MIF) as biomarkers of neuroinflammation in mild cognitive impairment (MCI) and Alzheimer's disease (AD) and to evaluate their diagnostic utility [Brosseron et al., 2014]. The dataset was curated and loaded in tranSMART and analysed using the visual analytical plugin smartR [Herzinger et al., 2017].

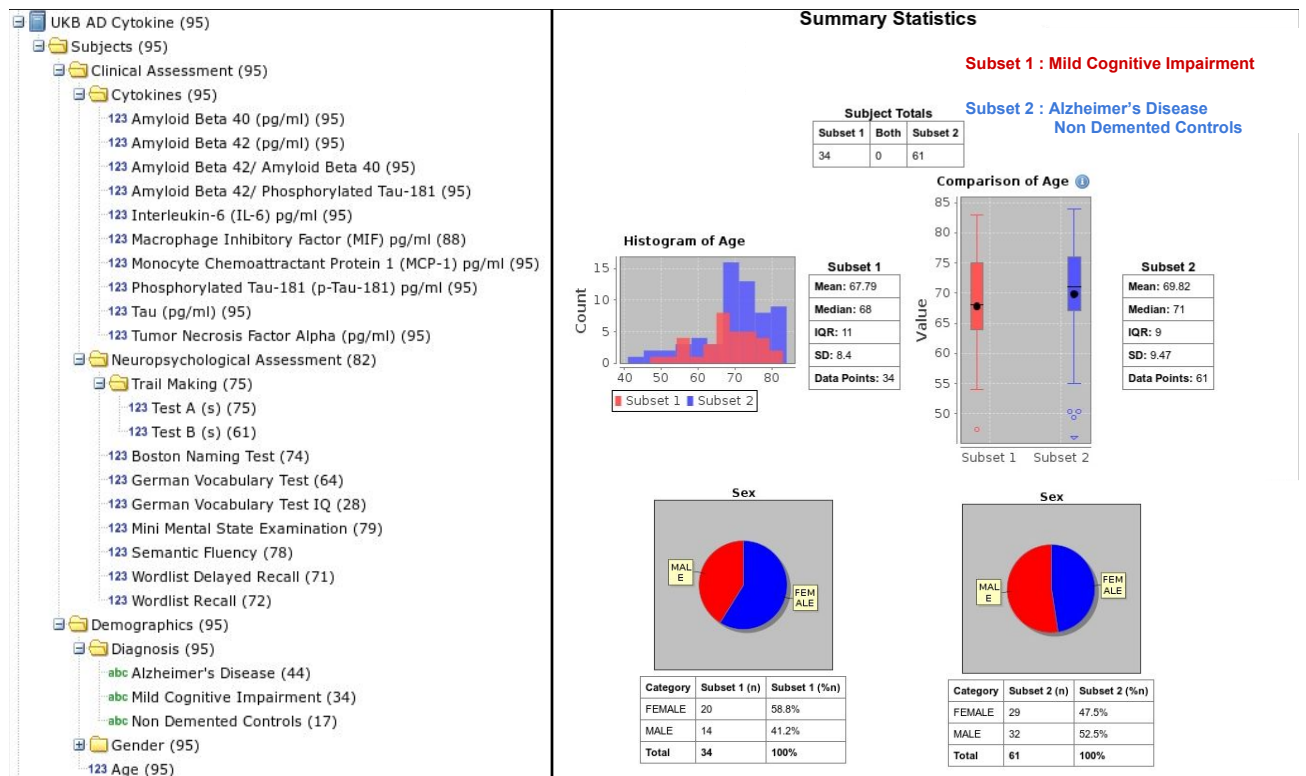


Figure 2.7: Overview of the AD Cytokine dataset loaded in transSMART

The dataset includes 95 subjects, consisting of 44 subjects diagnosed with Alzheimer's Disease, 34 as Mild Cognitive Impaired (MCI) and 17 Non Demented Controls.

First we compares the distribution of cytokine levels in controls against AD subjects. Figure 2.8 shows the distribution of *MCP-1* and *MIF* between the subsets.

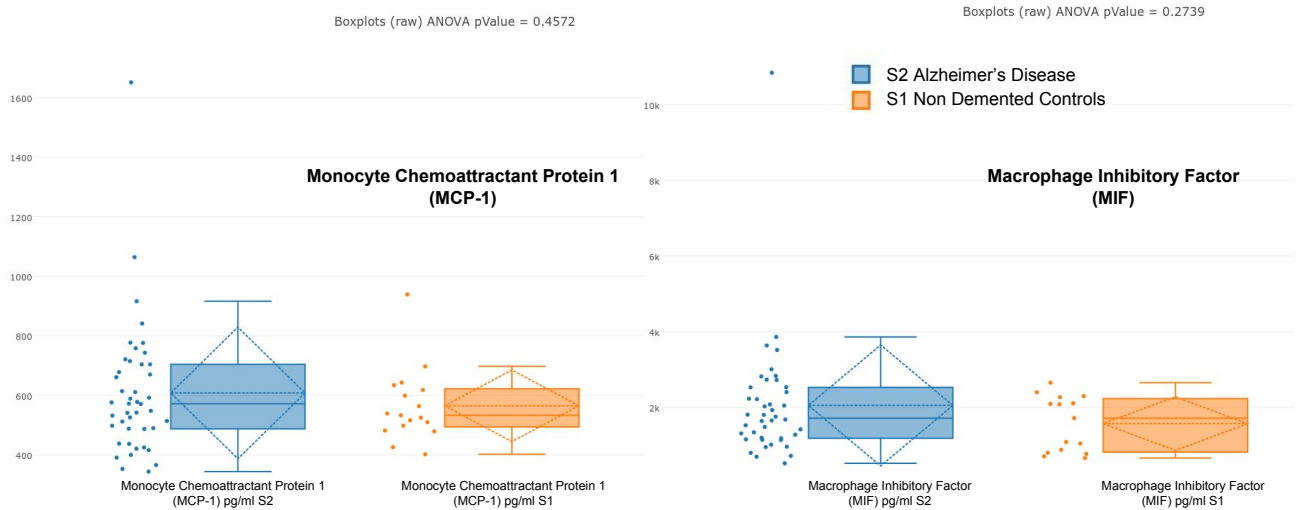


Figure 2.8: Distribution of *MCP-1* and *MIF* levels in AD subjects compared to controls

There were no elevated cytokine levels (*MCP-1* and *MIF*) in AD compared to controls. Further we then tested for correlations between cognitive decline and cytokine levels in AD. The MMSE, Wordlist Recall, Boston Naming Test, Semantic Fluency and Trail making test were used to score the cognitive decline. Figure 2.9 the cytokines *MCP-1*, *MIF* and *IL6* had no significant correlations to cognitive function in the AD group compared (MMSE shown in Figure 2.9).

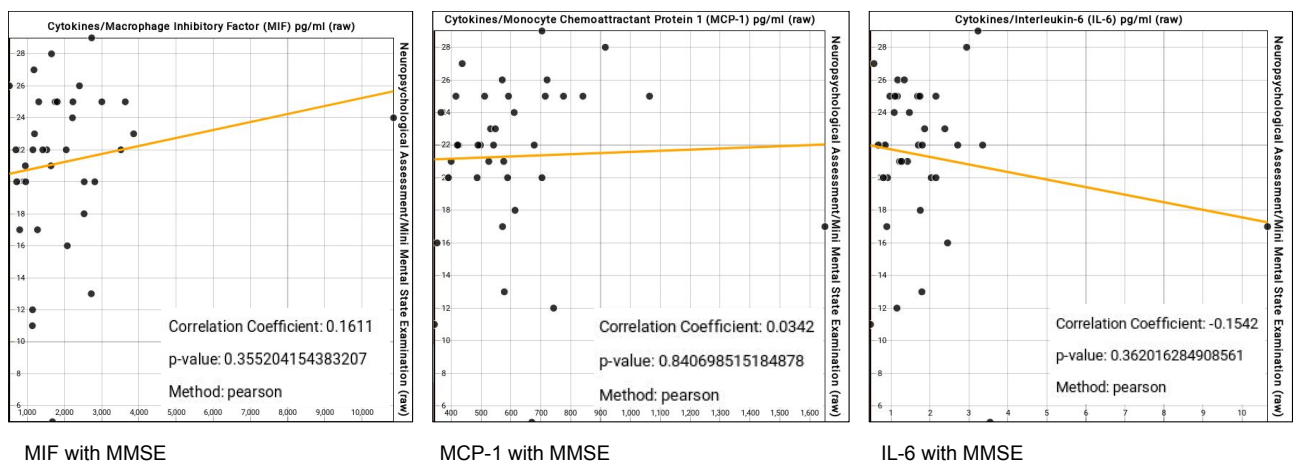


Figure 2.9: Correlation of cognitive functions with cytokine levels in AD subjects

To investigate the difference in cytokine level of MCI subjects against both control and AD. Two subsets were created, subset1 comprising of MCI subjects and

subset2 comprising of AD and Non demented controls (Figure 2.7). Next we tested for changes in $TNF\alpha$ levels between subsets 1 and 2 (Figure 2.10)

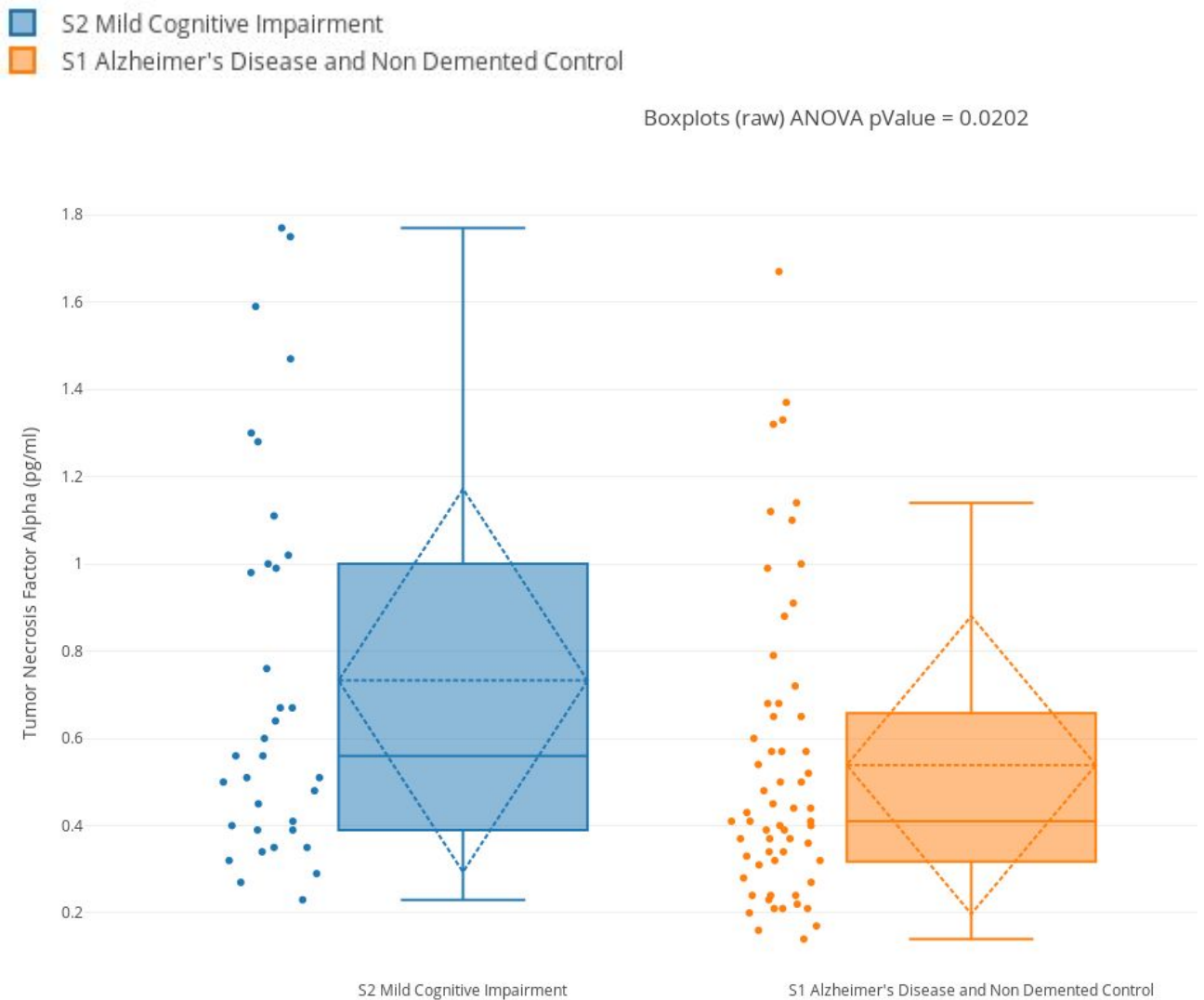


Figure 2.10: $TNF-\alpha$ was elevated in MCI compared to controls and AD subjects

$TNF-\alpha$ was elevated in MCI compared to controls and AD pValue=0.0202. Finally we test for correlation between cognitive decline and these elevated levels of $TNF\alpha$ in MCI subjects.

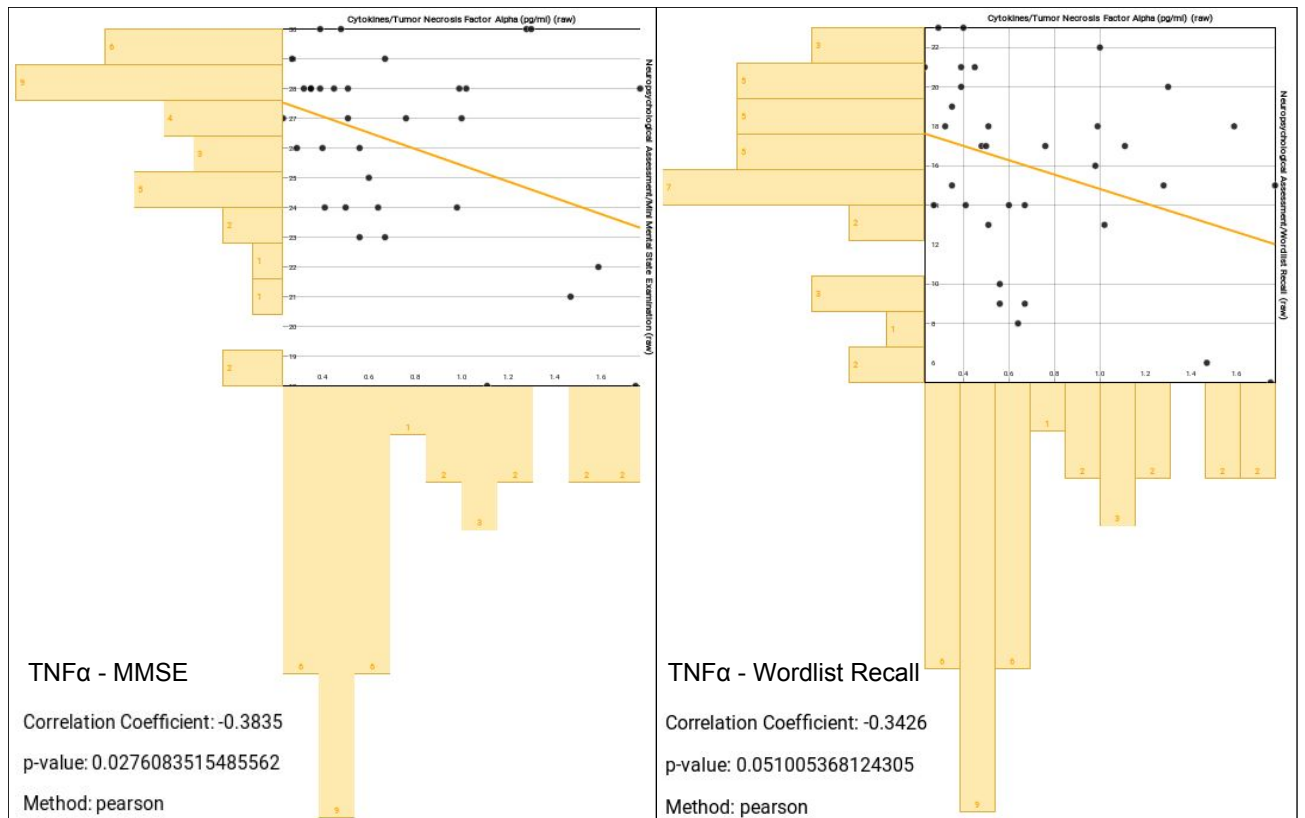


Figure 2.11: TNF- α in MCI subjects was reported to be negatively correlated with cognitive scores

TNF- α was reported to be negatively correlated with MMSE with a $p=0.03$ and $r=0.383$ and for word list recall word-list recall ($p = 0.05$, $r = -0.343$) in MCI (Figure 2.11). The analysis was performed by Pearson correlation. Levels of TNF- α has negative correlation with the cognitive functions in MCI.

Common markers of neurodegeneration provided better discriminative power, as demonstrated by A β 42, A β 42/p-tau-181, MMSE or semantic fluency. However, all standard markers differentiated clinical controls and MCI from AD, but not controls from MCI, which was in return only achieved by TNF- α . In summary, the data reports elevation of CSF TNF- α levels in MCI accompanied by correlations of TNF- α , MCP-1 and MIF to cognitive decline in this group of patients. Noteworthy, there were no elevated cytokine levels and few correlations to cognitive function in the AD group. Therefore, signaling of MCP-1, MIF and especially TNF- α might be involved

in pathological inflammatory processes during MCI that impact negatively on cognition. However, the study may have been biased by a certain degree of inhomogeneity in patient group size, age and gender. Yet, the collective reflected typical results for standard AD protein markers from CSF analysis and neuropsychological tests, therefore providing a reliable source for comparisons. Furthermore, age or gender driven effects on cytokine levels were ruled out by ANCOVA and Pearson correlation. Assessing neuroinflammation in MCI could therefore be of clinical importance in diagnostic procedures. CSF cytokines may reflect processes of disease progression and indicate an impact of innate immune activation on cognitive function in early disease stages. It is therefore desirable to monitor neuroinflammation, which can be used with the same routine as the established markers of other key pathological processes of MCI and AD.

Although, Amyloid β pathology is considered the primary hallmark of AD, recent studies suggest that the disease has a multifactorial origin [Llorens-Marín et al., 2014, Medina et al., 2017, Gong et al., 2018]. Currently, several reports support neuroinflammation as a significant contributor to the pathogenesis of Alzheimer's Disease [Hong et al., 2016]. As a result of brain damage (e.g., brain trauma, ischemia, $A\beta$ accumulation, etc.) microglia and astrocytes acquire reactive phenotype losing their physiological functions [Karve et al., 2016]. The persistent microglial activation stimulated by $A\beta$ via Toll Like Receptors (TLR) creates a vicious circle between microglia activation, neuroinflammation, and $A\beta$ accumulation. A crucial role on pathogenesis of AD is an absolute culprit for both amyloid plaque and other pathologic change such as the neuronal damage. Moreover, after activation, these cells produce a wide range of cytokines and proinflammatory mediators, leading to chronic inflammation [Heppner et al., 2015]. Even if the initial intent of these modifications is reparative, such long-lasting and uncontrolled activation causes further neurodegeneration (Figure 2.12) [Heneka et al., 2015].

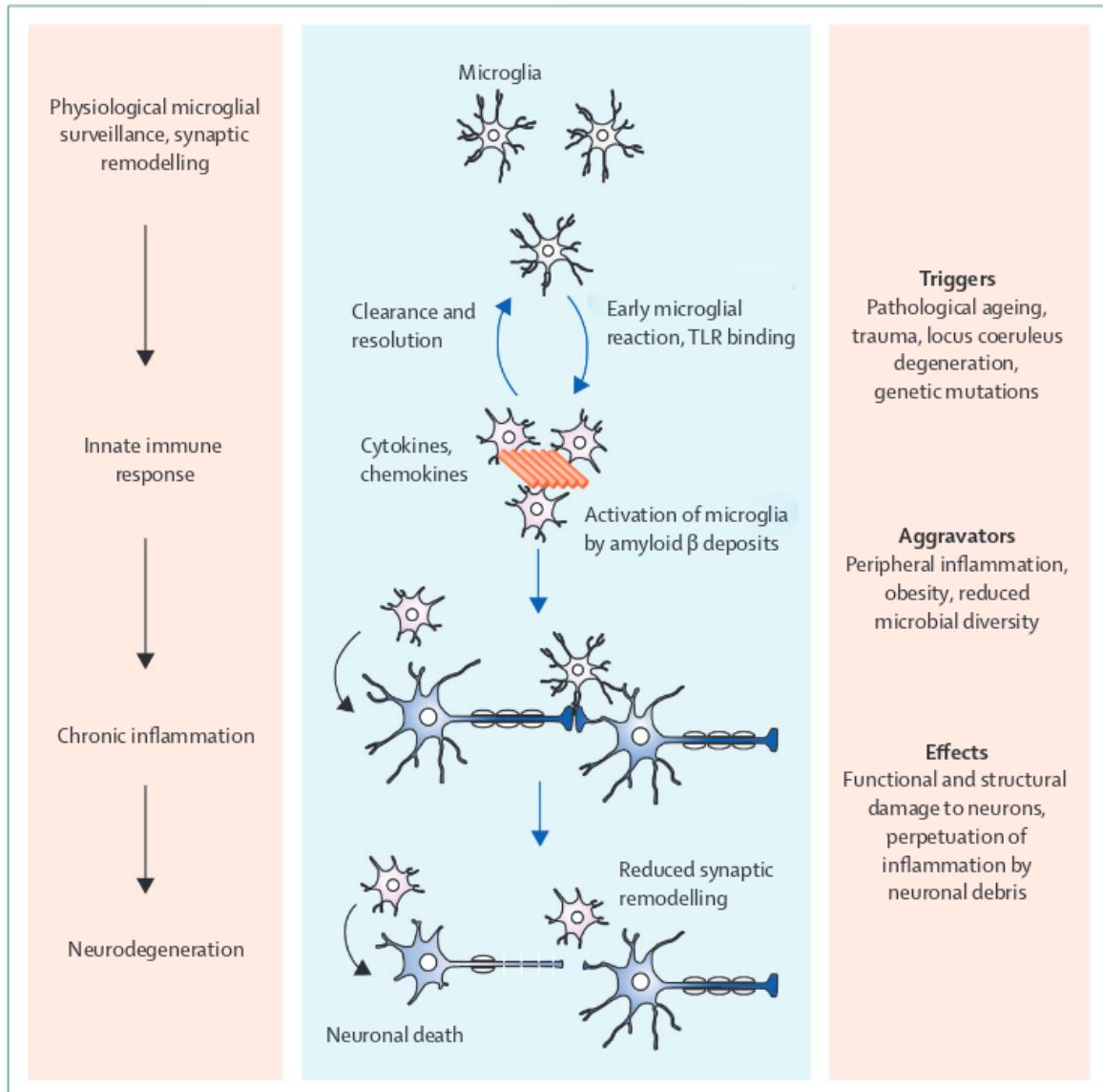


Figure 2.12: Role of inflammation in neurodegeneration [Heneka et al., 2015]

Physiological functions of microglia, including tissue surveillance and synaptic remodeling, are compromised when microglia sense pathological amyloid β accumulations. Initially, the acute inflammatory response is thought to aid clearance and restore tissue homeostasis. Triggers and aggravators promote sustained exposure and immune activation, which ultimately leads to chronic neuroinflammation. Perpetuation of microglia activation, persistent exposure to proinflammatory cytokines, and microglial process retraction cause functional and structural changes that result in neuronal degeneration [Heneka et al., 2015, Brosseron et al., 2014].

In a follow up study, the AETIONOMY collaborators investigated the utility of inflammatory biomarkers in diagnostic procedures of AD. This was designed in three steps: (1) to screen for proteins that are robustly detectable in cerebrospinal fluid; (2) to explore associations between the analytically robust markers and pathological features of AD; (3) to determine the discriminative power of these markers in the clinical diagnosis of AD. 46 proteins were screened, out of which 14 met the criteria for robust detectability. A subsequent analysis of these markers in a cohort of 399 patients (non demented subjects, patients with mild cognitive impairment, and patients with AD, supplemented by smaller cohorts of other diseases) was conducted. Although a large number of significant associations between clinical cohorts or AD pathology markers and inflammatory markers were observed, currently, none of the tested proteins reached a discriminative power as the existing pathological markers for clinical practice. Implementation of assays with higher sensitivity or investigation of signalling mediators from alternative pathways could lead to discovery of candidates with higher potential for use in clinical diagnostic procedures.

2.6.2 Use case 2: Expression data of substantia nigra from postmortem human brain of PD patients

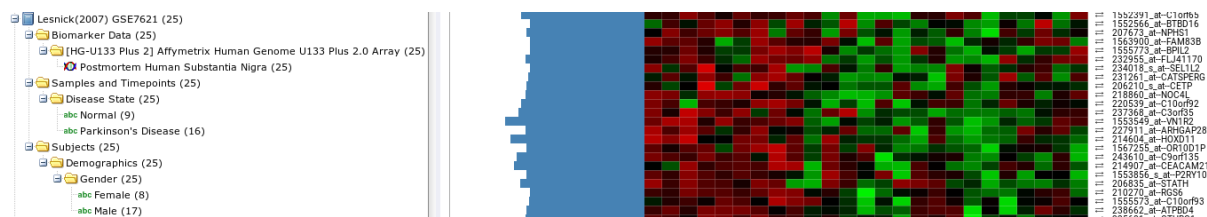


Figure 2.13: Heatmap generated from curated data (GSE7621) loaded in tranSMART

Here we use a GEO public study GSE7621 [Lesnick et al., 2007]. The study used microarrays to detail the global program of gene expression underlying Parkinson’s disease. Substantia nigra tissue from postmortem brain of normal and Parkinson

disease patients were used for RNA extraction and hybridization on Affymetrix microarrays: 9 replicates for the controls and 16 replicates for the Parkinson's disease patients were used. Both cohorts included males and females. The heatmap workflow was used to retrieve the differentially expressed genes between control and diseased.

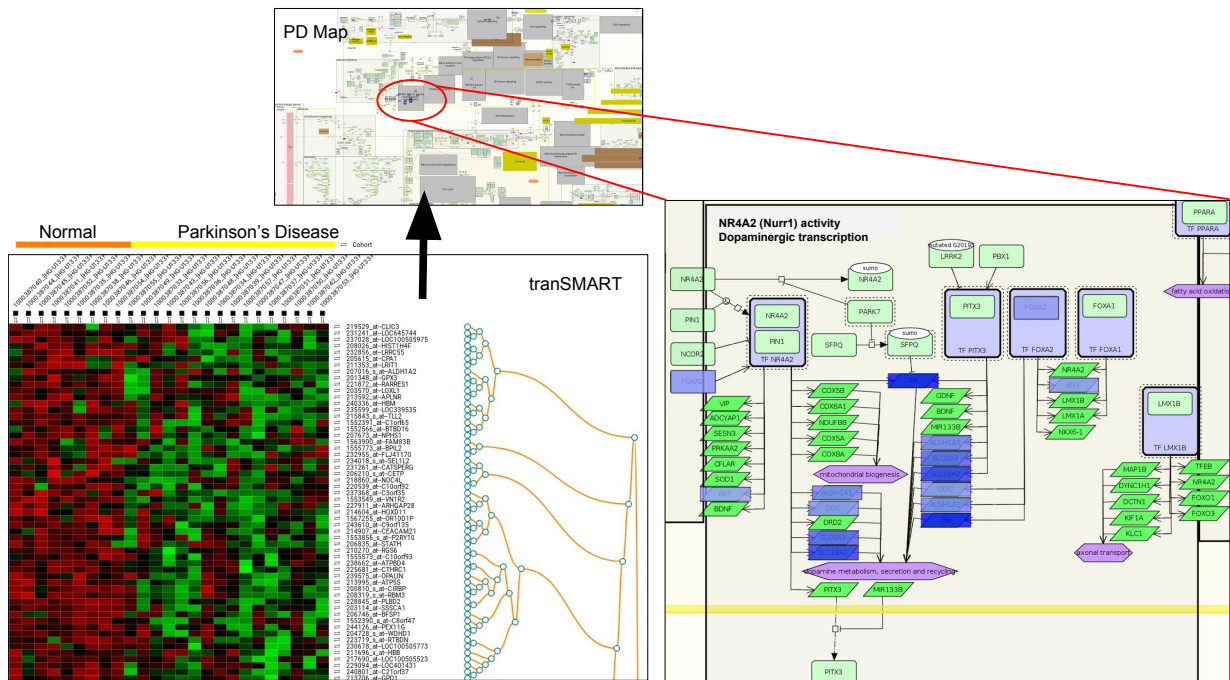


Figure 2.14: Overlaying differentially expressed genes on the Parkinson's disease map

Differential gene expression data comparing post mortem brain tissues from male PD patients versus controls are overlaid on the PD map blue representing down-regulated and red representing up-regulated genes (See Figure 2.14). Overlaying the differentially expressed genes on the PD Map provides context about the pathways and mechanisms these genes are involved in. These may suggest new targets for further investigation towards potential treatments. Overlaying the differentially expressed genes on the PD Map show perturbations in:

- Dopamine secretion and recycling : down-regulation of SLC18A2, RIMS1, SLC6A3
- Dopaminergic transcription : down-regulation of RET, TH, ALDH1A1, DDC,

SLC6A3, SLC18A2, FOXA2, EN1

- Dopamine metabolism : down-regulation of TH, DDC, ALDH1A1 and SLC6A3
- Post synaptic terminal processes : up-regulation of GRIA4, down-regulation of SLC6A3, RGS4, ALDH1A1
- Autophagy: up-regulation of AMBRA1
- Calcium signaling and NEF2L2 Activity : up-regulation of CREBBP
- Neuroinflammation : up-regulation of PTGS2, SOCS3 and NCF4

TH, ALDH1A1, SLC6A3, SLC18A2, DDC, RET, EN1, FOXA2 are down-regulated, all are involved in dopaminergic transcription. Down-regulated genes (SLC18A2, RIMS1, SLC6A3) are involved in dopamine secretion and recycling. Down-regulated genes (TH, DDC, ALDH1A1) are involved in dopamine metabolism.

2.7 Summary

A number of public and consortium studies relevant to Alzheimer's and Parkinson's Disease was curated, structured and loaded into tranSMART. Refer Appendix A.1, for a complete list of studies and data types. This included transcriptomics, clinical and imaging datasets. The process included several steps from data acquisition, curation and harmonisation, structuring and loading. The harmonised and structured data, can then be used for further analysis and sharing. We demonstrate with two examples how integration of data from heterogeneous sources can support hypothesis generation. Translational research platforms like tranSMART provide visual and exploratory analysis facilitating the identification of patterns in data and the subsequent hypothesis generation, hypothesis validation. However a major challenge in this process is the heterogeneity in data representation and formats. The major effort spent

in the whole process is to curate and map data to standard ontologies. This process can be streamlined if data is exchanged in well accepted standards and with adequate meta-data.

Chapter 3

Comparison of disease maps

There are many disease maps and pathways publicly available [Kuperstein et al., 2015, Matsuoka et al., 2013, Fujita et al., 2014, Mizuno et al., 2012, Oda and Kitano, 2006] and several others currently in development (<http://disease-maps.org/projects>). In addition to serving as a single point to access the existing knowledge, comparison of disease maps have a number of possible benefits. Comparison of two disease maps shed light on the underlying mechanisms which are common or specific to each disease. Several diseases may have common mechanisms that affect the pathology and progression of each other and therefore is important to study the comorbidity of diseases. Apart from this, comparison of a disease models against normal biological state models can help understand the disease pathology and progression. Another potential use case of comparing maps and pathways is to enrich the knowledge in disease maps and understand how the pathway plays a role in the aetiology of diseases.

Pathguide [Bader et al., 2006] is a meta-database that provides an overview of web-accessible biological pathway and network databases. These include databases on metabolic pathways, signalling pathways, transcription factor targets, gene regulatory networks, genetic interactions, protein-compound interactions, and protein-protein interactions. As of May 2018, Pathguide (<http://pathguide.org>) contains 702 biolog-

ical pathway and molecular reaction related resources, this includes several organism models in standards such as BioPAX, CellML, PSI-MI or SBML. Since many different models in different formats exist, the ability to compare these models is important, both to compare models of different systems and to compare different versions of the same model. Comparison of maps highlights the similarities and differences in the context of the diseases, how the identified elements relate to the interacting elements and their role in the disease mechanism. This chapter describes the comparison of two disease models and how such methods can add context to the comorbidity of diseases and their pathology. We use the MINERVA platform to visualise the results using its graphical layout and the PD Map Spring 2018 edition and AlzPathway April 2015 edition as use cases.

3.1 Overview of existing comparison methods

The comparison of systems biology models gained interest in recent years. In 2017, Scott-Brown and Papachristodoulou presented a tool *sbml-diff*, that is able to read synthetic biology models in SBML format and produce a range of diagrams showing different levels of detail [Scott-Brown and Papachristodoulou, 2017]. Each of these diagram type can then be used to visualize a single model or to visually compare two or more models. However, in addition to their focus on mathematical models, the web service could not handle large models like the AlzPathway and PD Map. Moreover, species and reactions are compared based on their identifiers (ID) and two elements are treated as the same only if both share the exact same set of MIRIAM identifiers.

BiVeS [Scharm et al., 2015], another tool for comparing SBML models can track changes in a model over time. Although it produces outputs in a different formats, the visualisation abilities are limited. The main focus of the *BiVeS* is to provide version control for model repositories to accurately detect and describe differences

between versions of model depending on the encoding, mathematical expressions and the structure of the networks.

Another approach by Calderone et al., to compare Alzheimer and Parkinson's disease networks, uses AlzPathway and PD Map [Calderone et al., 2016]. The authors considered direct comparison of SBML models not feasible due to level of details and differences in entities and annotations. Nevertheless, they extracted the genes and proteins from the two SBML models and complemented the lists with genes and proteins from the KEGG database. The AD and PD lists were then used as seeds to extract two subnetworks from Mentha [Calderone et al., 2013] a human interactome database. In order to generate the networks, all the genes and proteins were translated to UniProt identifiers, since the Mentha uses UniProt accession numbers. A graph-communities-based similarity matrix method was implemented to cross-compare two networks to highlight differences and similarities in terms of network topology and functions. Entities that were detected as similar in both network and clustered according to their Gene Ontology overlap form a community. Each Communities present in both networks signify common biological processes and on the other hand the communities unique to each network may signify characteristics of the specific pathology.

3.2 Methods for comparison

Although several methods attempt to compare disease maps or models, no direct comparison of such models were previously reported especially to visualise the comparison. Therefore, in order to compare two maps directly, an algorithm was implemented that parses two models and identify similar reactions and entities with respect to the context in terms of cellular localisation or mechanism. For each map pair, lists of node and reaction identifiers detected in both maps, as well as a list of reaction pair, describing their similarity are generated.

The reaction and node lists can then be visualised on the corresponding map using the MINERVA platform. The PD Map and AlzPathway hosted on the MINERVA platform were used as use-cases. The comparison required mapping of annotation for entities, this was complemented by the annotators in the MINERVA platform.

To compare two maps, we need to compare both reactions and their interacting elements. In order to compare reactions, we take into account the neighbouring participants of the reactions i.e. the reactant(s), product(s), and modifier(s). Additionally, to compare the elements we require measures to uniquely identify the elements irrespective of how they may be named or represented. For this, we use MIRIAM identifiers, with which the elements are annotated. The algorithm parses the models, and updates the annotation if required. For instance, if the models use different databases to annotate elements, the annotator using the name or MIRIAM identifiers to extract other annotation from corresponding databases.

Figure 3.1 summarises the process to identify similar elements. To match entities, the MIRIAM annotations were used. The MIRIAM Registry and Identifiers.org system are a set of services and resources that provide support for generating, interpreting and resolving MIRIAM URIs. The annotation will help to identify the same elements even if they may be named differently. If the type of the element/node, e.g.: protein, gene, RNA, etc. matches, then the annotations are checked for a match. Once the annotation also match, the localization information i.e. the compartment the element is in is matched. If the entity is a complex itself, or part of the complex is also taken into account i.e. if the entity is annotated as the same complex or if it is part of the same complex in both maps.

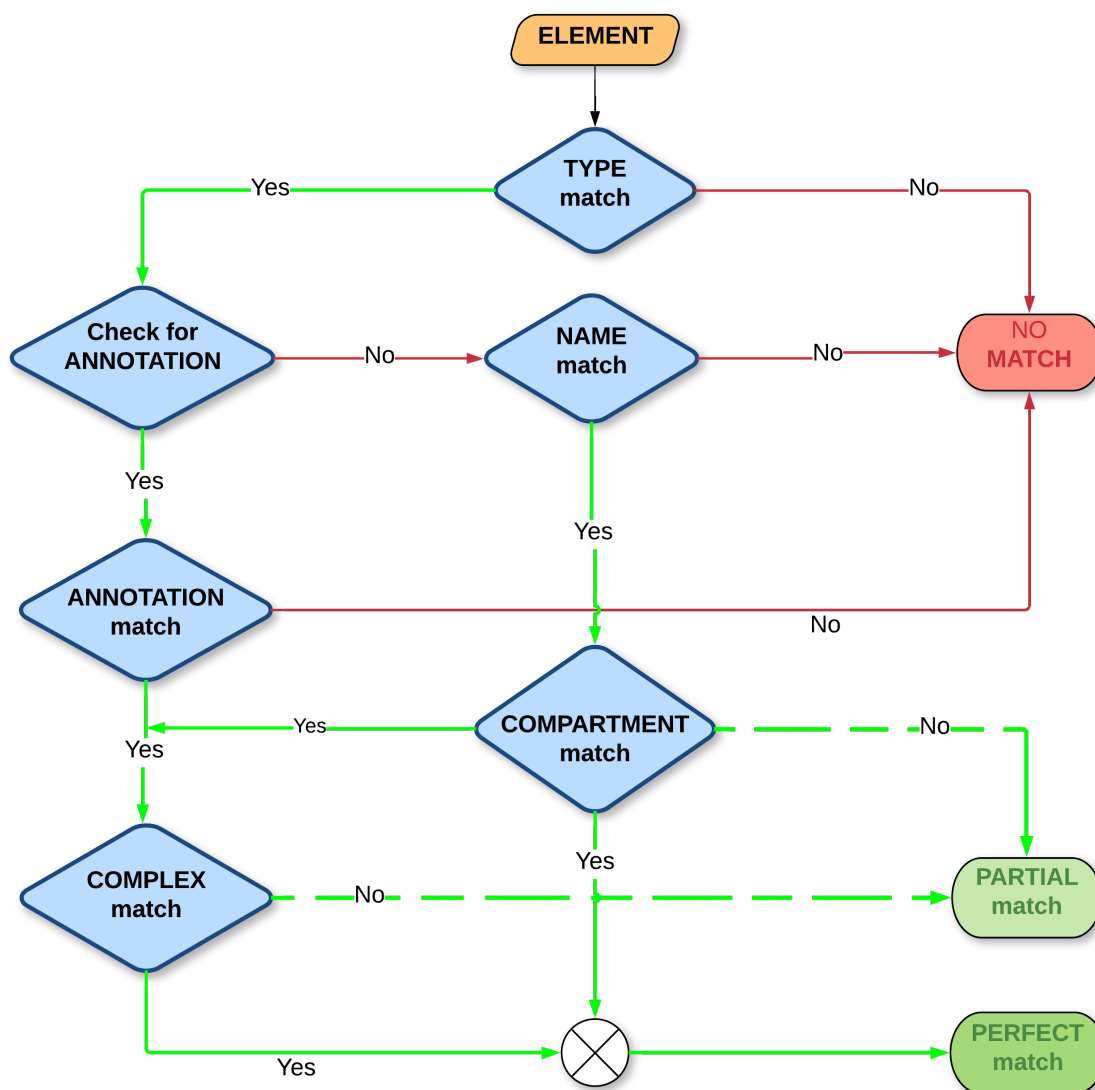


Figure 3.1: Element Match Decision Diagram

The reactions are matched by checking for i) the reaction type ii) the entities involved in the reaction. The reactions have more than one reactant and product. Additionally, reactions may involve one or more modifiers as well. The matches for reactions were categorised into the following categories:

1. All Products, Reactants, Modifiers, annotation and compartment match
2. All Products, Reactants, Modifiers match
3. All Products and Reactants match

4. All Products and Reactants match and model1 has no Modifiers
5. All Products and Reactants from model1 match and at least 1 Modifier match
6. All Products and Reactants from model1 match and no modifiers match

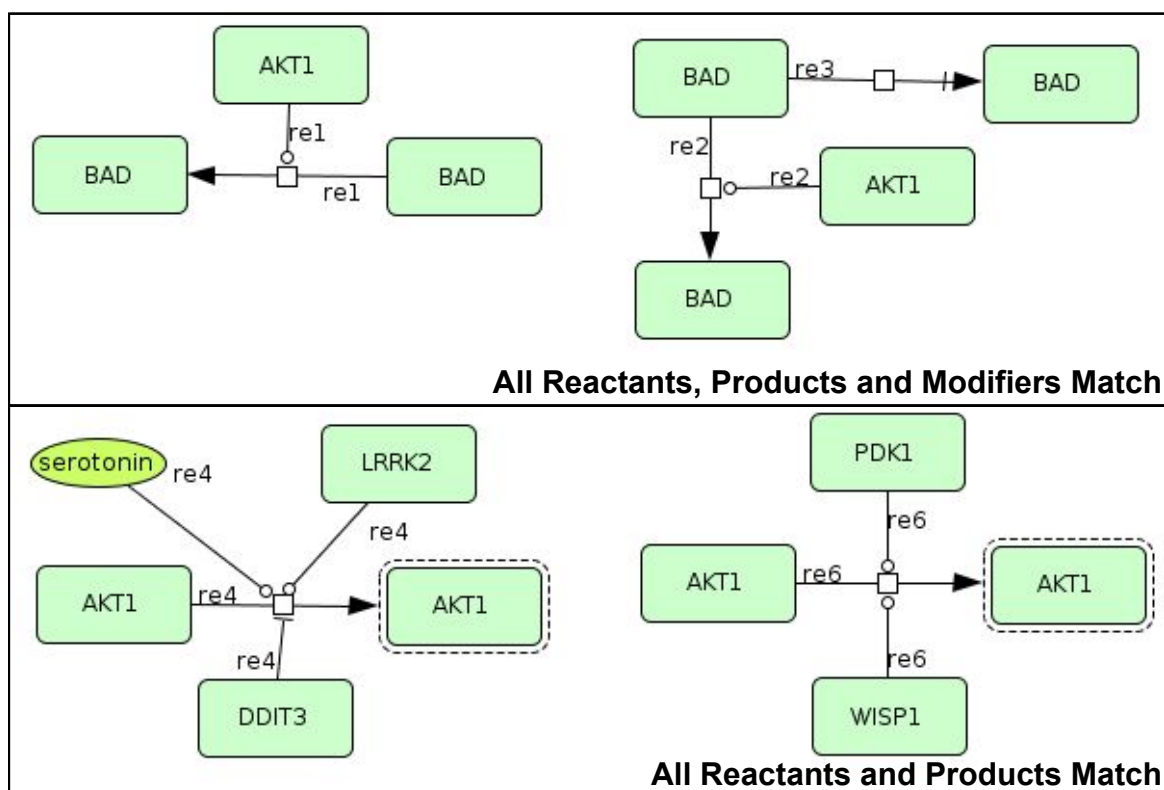


Figure 3.2: Examples of reaction match

The reaction pair i.e. the corresponding reaction identifiers that matched in both models, and the above categories are saved in a ReactionPair table. This table provides additional information that can supplement a closer examination of the relevance of the match in the disease mechanism context. Figure 3.2 shows two example of reaction pairs. Each reaction involves at least one reactant and one product. However, reactions may have more than two participating elements.

Another challenge in the comparison is the representation of complexes. Often in cases when the relevant scientific literature does not provide enough information to annotate the complex itself, curators have to resort to create complexes which are

not annotated by a unique identifier. Therefore complexes cannot be compared directly if they are not annotated. We need to approach the comparison of complexes at a content level i.e. two unannotated complexes are identical if they have exactly the same content.

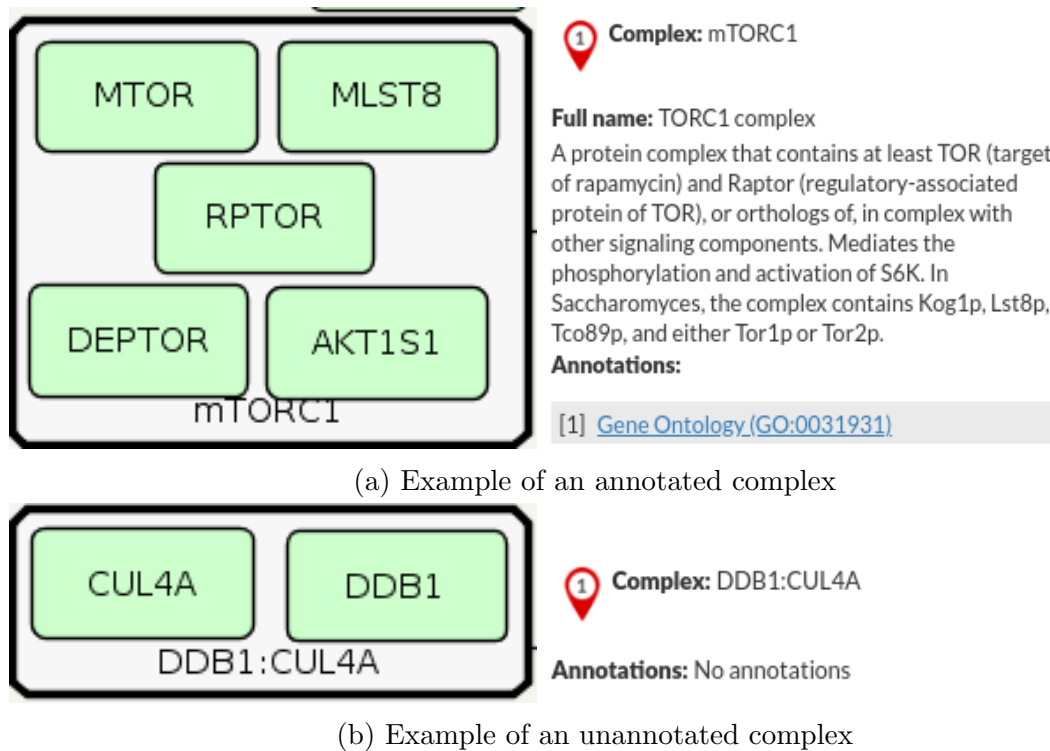


Figure 3.3: Representing complexes in disease maps

For instance in Figure 3.3a is annotated by the GO identifier *GO:0031931*. *TORC1 complex* is a protein complex that contains at least *TOR* (target of rapamycin) and *Raptor* (regulatory-associated protein of TOR), or orthologs of, in complex with other signalling components. Whereas the *DDB1:CUL4A* complex in Figure 3.3b is not annotated. This may be due to several reasons such as the relevant article did not provide enough information about all the contents of the complex, or such a complex could not be mapped to a unique entry in corresponding databases like Gene Ontology.

3.3 Results

For each map pair that is compared, three result tables are generated. Two lists for overlay on the map i) List of element identifiers in model1 which have a counterpart that was matched in model2 ii) List of reaction identifiers in model1 which have a counterpart that was matched in model2. The third table ReactionPair, lists each pair of reaction, reaction identifiers from model1 and model2 and the match type as discussed before. The result tables for AlzPathway and PD Map are also included in the Appendix A.2.

3.3.1 Comparison of AlzPathway and PD map

The AlzPathway and PD Map were compared using this method. Because PD map has a number of submaps, their contents were compared, one by one, with the contents of AlzPathway. The supplementary ReactionPair table, makes it easier to compare the reactions side by side on both maps, to give a context to the identified reaction in both diseases.

MINERVA enables updating the model with additional annotations which are used for comparison. Currently MINERVA supports HGNC and BioCompendium to extract by name and Uniprot, Gene Ontology, Ensembl, Entrez Gene, and ChEBI for update by MIRIAM identifiers [Gawron et al., 2016]. Different maps may use different namespaces to annotate their entities with a unique identifier. The AlzPathway uses Uniprot to annotate the proteins, genes and RNA. Whereas the PD Map uses HGNC to annotate these species. Therefore, it requires a conversion between Uniprot and HGNC. The annotator extracts the HGNC symbol and identifier from the corresponding Uniprot identifier. Drugs and chemicals are annotated by ChEBI in the PD Map, while AlzPathway uses PubChem. Since MINERVA does not currently support PubChem and this mapping has not been implemented, drugs and chemicals comparisons

are not considered unless they are annotated by the same database in both models. The following database identifiers are used to annotate the entities in the PD Map.

- Protein, Gene and RNA: HGNC, HGNC symbol, Uniprot, InterPro, Entrez Gene, Ensembl, GO
- Drugs, chemical, ion, simple molecule : ChEBI, Reactome
- Phenotype : GO, MeSH
- Complex : GO, Reactome, MeSH, InterPro
- Compartment: GO, MeSH

If one or more of the MIRIAM annotation of the elements match, the elements are considered a match.

Map/submap	Elements	Reactions
AlzPathway	2464	979
PD map - main diagram	5444	2416
Ubiquitin-proteasome system	100	18
PRKN substrates	75	64
Fatty acid and ketone body metabolism	194	75
Iron metabolism	557	183

Table 3.1: Number of elements and reactions in the AlzPathway Map and PD Map and submaps

Table 3.1, summarises the number of elements and reactions in each map. To retrieve elements and reactions in AlzPathway Map that are present in the PD Map, the models were updated using the MINERVA automatic annotators. Since the AlzPathway and PD Map utilise different namespace to annotate the entities, updating

the annotation was essential to identify similar entities which would otherwise not be detected.

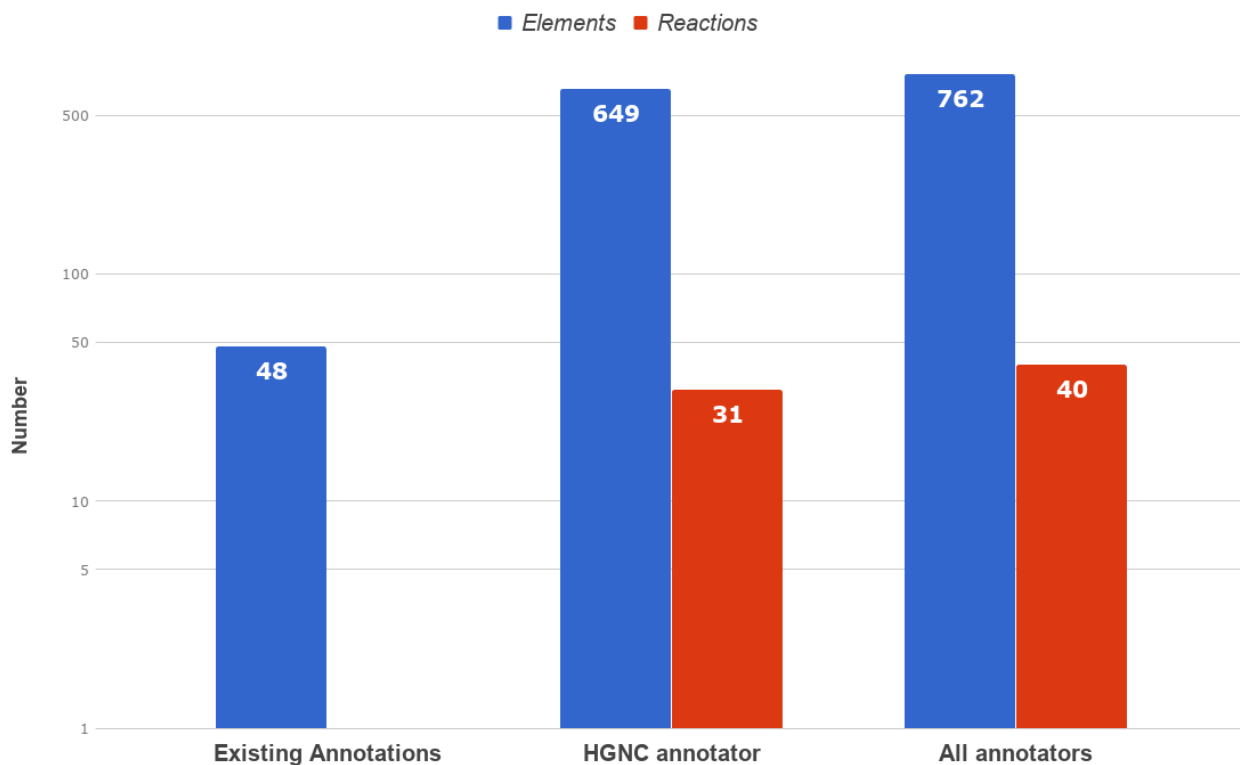


Figure 3.4: Increase in matched elements and reaction by updating model annotation

Figure 3.4, depicts the increase in identified elements and reactions when the models annotations were updated by additional identifiers. Comparing models without updating the annotation returns 48 elements, from 2464 elements in AlzPathway (Table 3.1). No reactions were identified. Since the PD Map uses HGNC to annotate the Protein, Gene and RNA, the HGNC annotator was used to fetch the Uniprot identifiers (used by AlzPathway). This increased the number of elements identified to 649. Although 31 Reactions were also identified, none of these reactions were a perfect match, signifying that there were changes in number of reactants, products or modifiers. This may be due to the fact that similar entities were still not detected due to no namespaces or different database used for annotation. Using all the annotators (HGNC, Uniprot, Entrez, Ensembl, GO, and bioCompendium) increased the number

of entities identified to 762 and reactions to 40. Six of these reaction were a perfect match. The low number of perfect matches could also be a result of different literature source used for curation, different scope (disease) of the maps or expertise of the curator.

	No. of species	No. of reactions
PD_neuroinflammation	310	209
AlzPathway	1312	979
PD Map-main	2606	2416

Table 3.2: Summary of model sizes

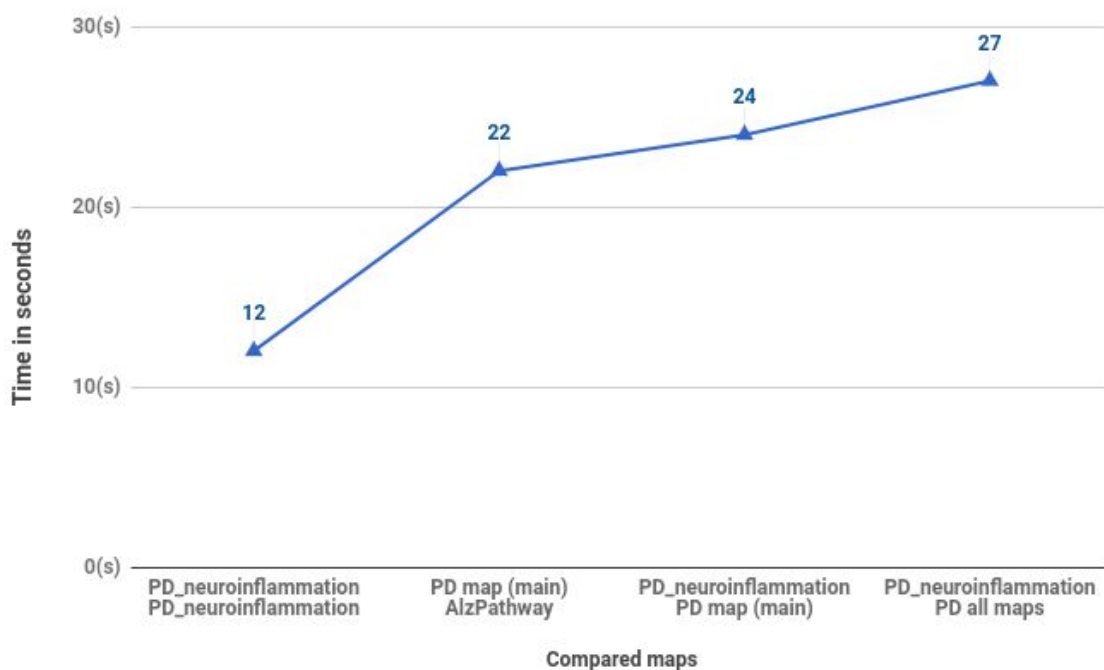


Figure 3.5: Time taken per comparison

To estimate the time taken for comparing smaller models, a smaller section of the PD map was used, PD_neuroinflammation (Table 3.2). The comparison of models takes about 10-30 seconds depending on the size of the model (Figure 3.5).

Highlights on PD Map, similarities from AlzPathway

PD Map Submap	Elements from AlzPathway	Reactions from AlzPathway	Perfect matches: reaction	Possible reaction pairs
PD map - main diagram	163	39	5	61
Iron metabolism	20	2	0	2
Ubiquitin-proteasome system	3	0	0	0
Fatty acid and ketone body metabolism	6	0	0	0
PRKN substrates	1	0	0	0

Table 3.3: Number of elements and reactions identified from AlzPathway in PD Map

Figure 3.6 shows the reactions and elements found in AlzPathway and highlighted on the PD Map. Table 3.3 shows the number of reactions and elements detected in the PD Map submaps. The matches can be further investigated to understand the mechanisms that may result in the comorbidity of these diseases.



Figure 3.6: Reactions and elements found in AlzPathway highlighted on the PD Map

Highlights on AlzPathway, similarities from PD Map

PD Map Submaps	Elements from Submap	Reactions from Submap	Perfect matches: reaction	Possible reaction pairs
PD map - main diagram	181	31	5	57
Iron metabolism	21	2	0	2
Ubiquitin-proteasome system	3	1	0	2
Fatty acid and ketone body metabolism	9	0	0	0
PRKN substrates	1	0	0	0

Table 3.4: Number of elements and reactions identified from PD Map in AlzPathway Map

Figure 3.7 shows the reactions and elements found in PD Map and highlighted on AlzPathway. Table 3.4 shows the number of reactions and elements detected in the PD Map submaps. From Figure 3.7 and 3.6, we can see that in general the reactions identified on the AlzPathway Map are more centralised around a specific node, whereas in the PD Map they are more distributed in terms of different processes they are involved in.

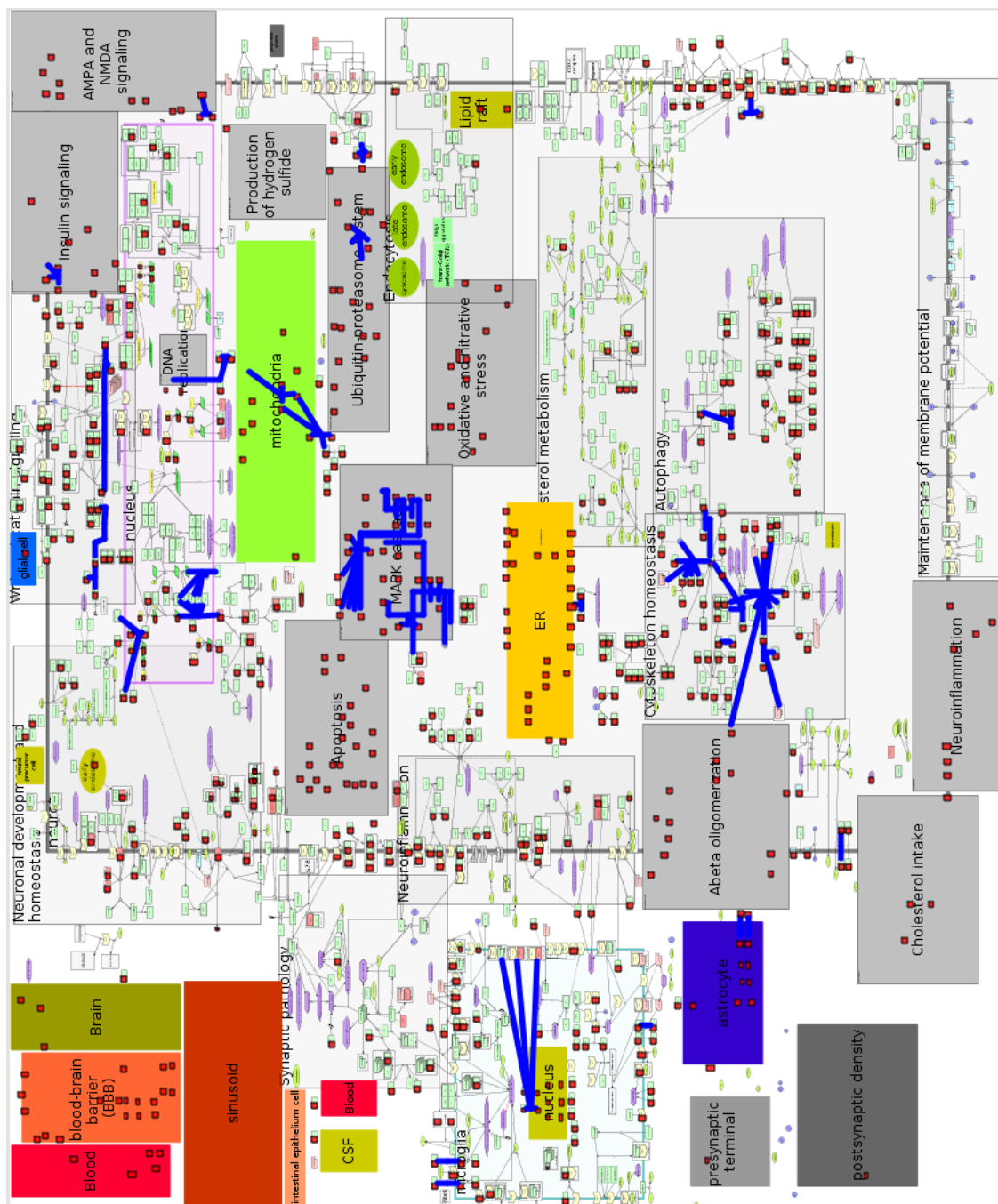
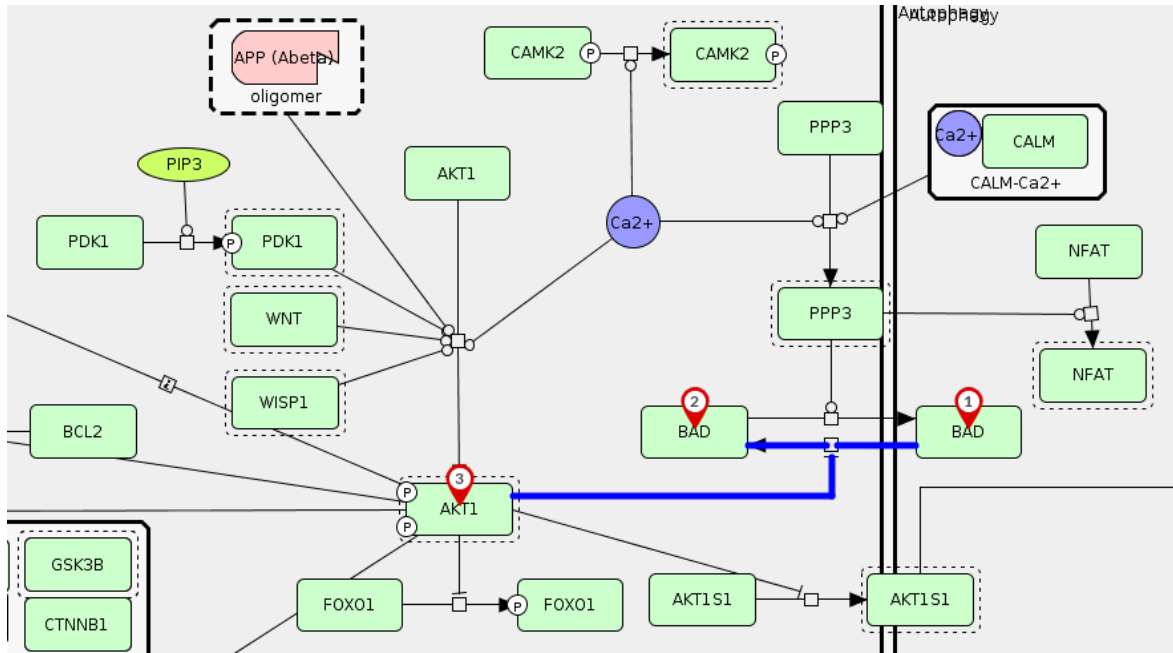


Figure 3.7: Reactions and elements found in PD Map highlighted on the AlzPathway

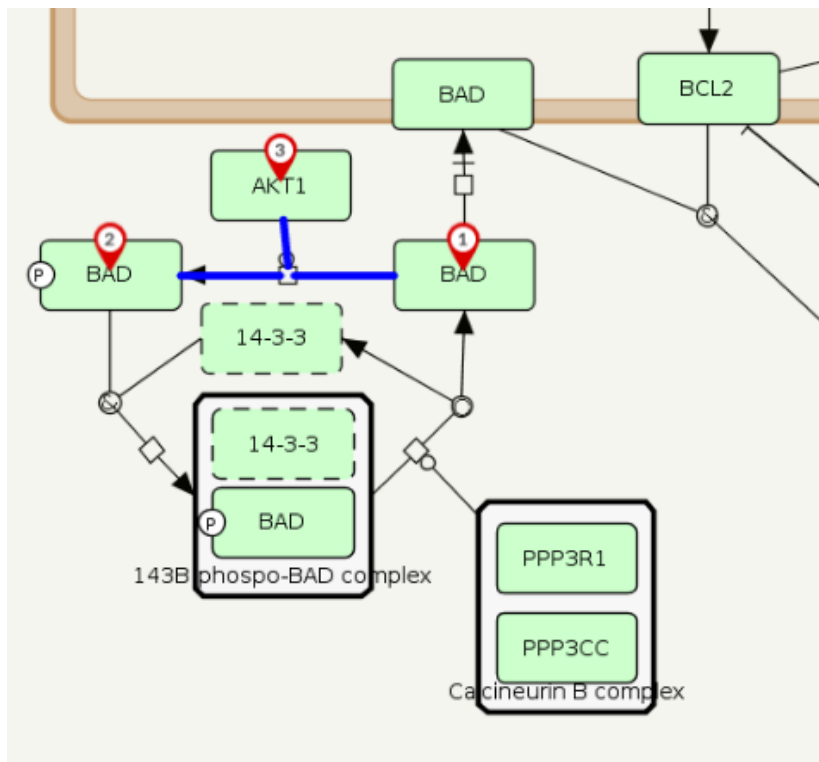
3.3.2 *AKT1* Activity

The serine/threonine kinase (Akt), has been widely researched for their involvement in several cellular processes, including insulin metabolism and diseases like PD, AD and cancer. The isoforms of *Akt- Akt1*, *Akt2*, and *Akt3* are reported to be involved in regulation of the apoptotic machinery [Vivanco and Sawyers, 2002, Reddy, 2013, Greene et al., 2011]. In the developing nervous system *AKT* is reported to be an important mediator of growth factor-induced neuronal survival. Activation of *AKT1* can suppress apoptosis and oxidative stress by inactivation and phosphorylation of pro-apoptotic targets, including *BAD* and *FOXO1* and *FOXO3* [Zhang et al., 2011, Hers et al., 2011].

One of the matches detected in AlzPathway and PD Map is shown in Figure 3.8. While in the PD Map (Figure 3.8b) *AKT1* moderates the phosphorylation of *BAD*, Phosphorylated *AKT1* inhibits the transition of *BAD* in AlzPathway. Other similar reactions found in the PD Map are shown in Figure 3.9. In Figure 3.8a, *PDK1*, *WNT* and *WISP1* are shown mediating the phosphorylation of *AKT1* in the AlzPathway. Whereas in Figure 3.9a *ROCK2*, *mTORC2* and *PDPK1* mediate the phosphorylation in PD. It is also interesting to note that *PDK1* in AlzPathway is *PDPK1* in PD Map, but they were not detected as identical since the *PDK1* in AlzPathway was annotated as *pyruvate dehydrogenase kinase 1(PDK1)* instead of *3-phosphoinositide dependent protein kinase 1 (PDPK1)*. Another downstream target of *AKT1* is the *TSC1*, *TSC2* complex which is reported to inhibit mTOR activity (seen also in Figure 3.10b) which in turn regulates cell growth and protein degradation [Olney et al., 2017]. *TSC1:TSC2* (in PD Map) and *TSC* in AlzPathway are seen to be mediated by several entities including *AKT1* and *MAPK1/3* in both AD and PD, but they were not identified since *TSC* was not annotated in AlzPathway, additionally *MAPK1/3* was annotated as *MAPK3* in AlzPathway. However, downstream reactions of *TSC* mediating *RHEB* were identified in both AlzPathway and PD Map. To summarise, several of the downstream targets of *AKT1* are similar, but the modifiers of the phosphorylation are different.

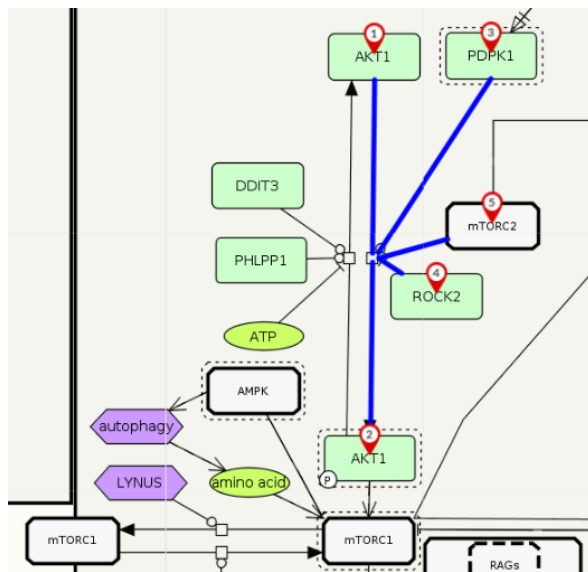


(a) AlzPathway Map: *BAD* modulation by *AKT1*

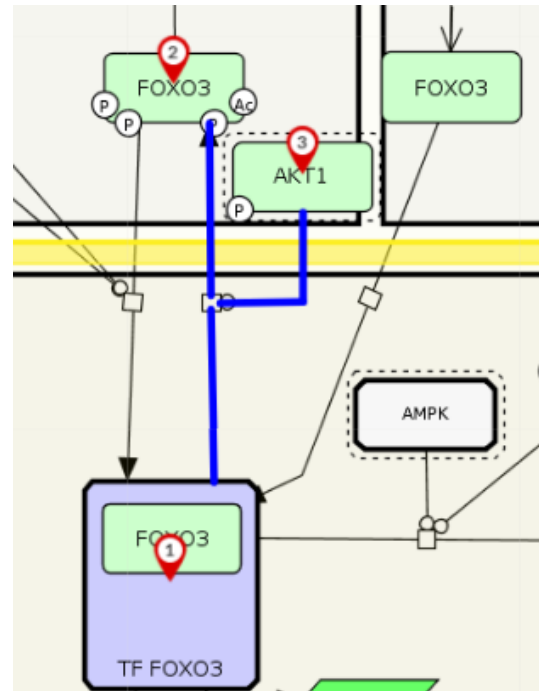


(b) PD Map: *BAD* modulation by *AKT1*

Figure 3.8: Perfect match in *AKT1* activity in PD Map and AlzPathway Map

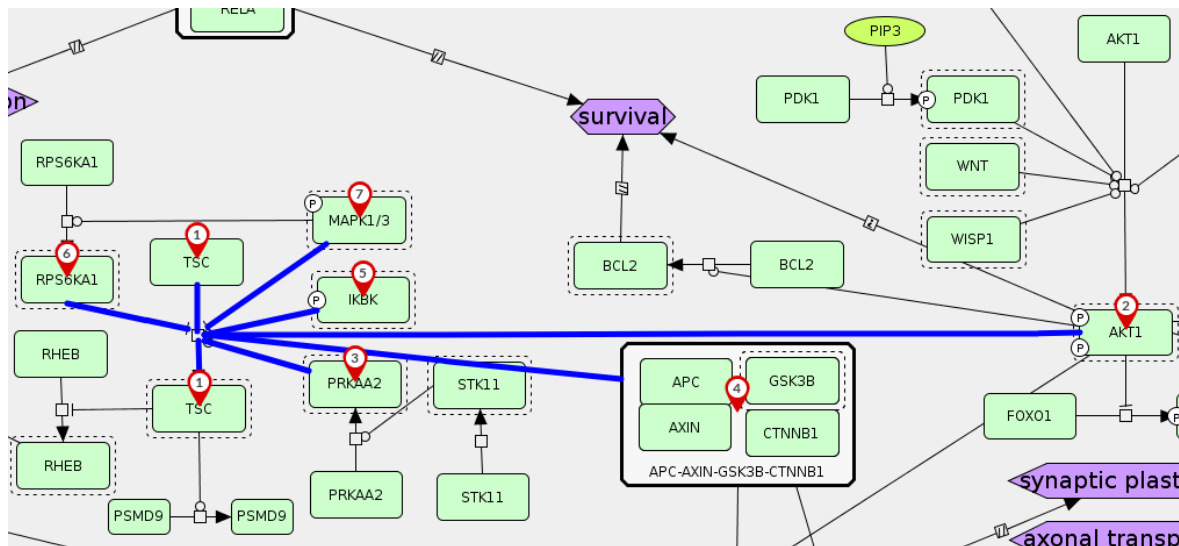


(a) PD Map: Modulators of *AKT1* Phosphorylation

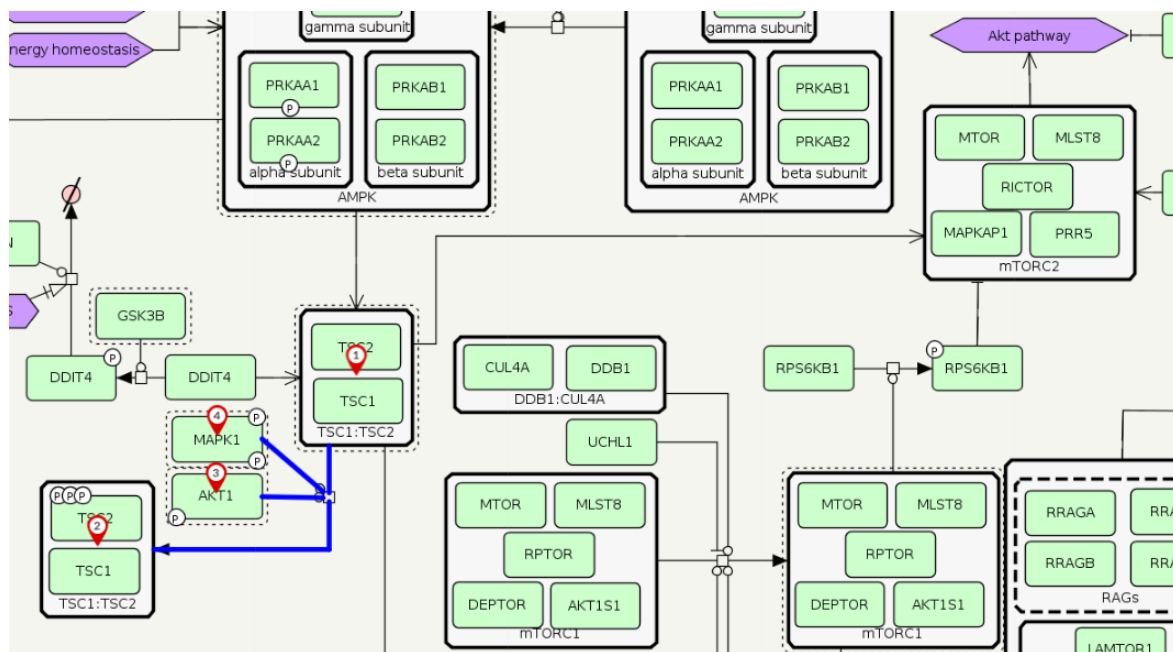


(b) PD Map: Phosphorylated *AKT1* mediating *FOXO3*

Figure 3.9: *AKT1* activity in PD Map



(a) AlzPathway: Downstream targets of *AKT1*



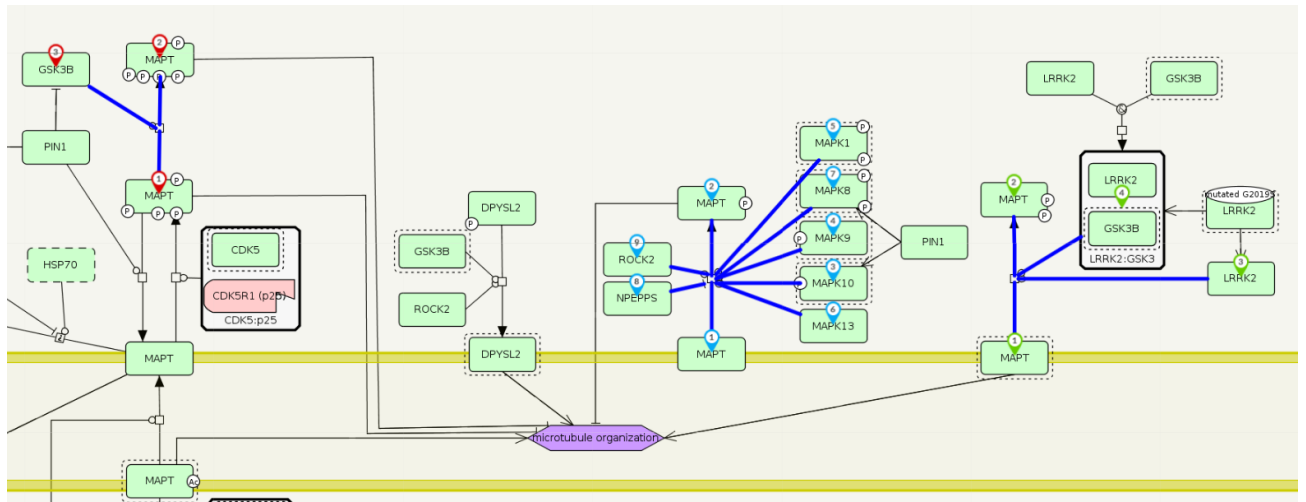
(b) PD Map: Downstream targets of AKT1

Figure 3.10: *TSC1:TSC2* activity in AlzPathway and PD Map

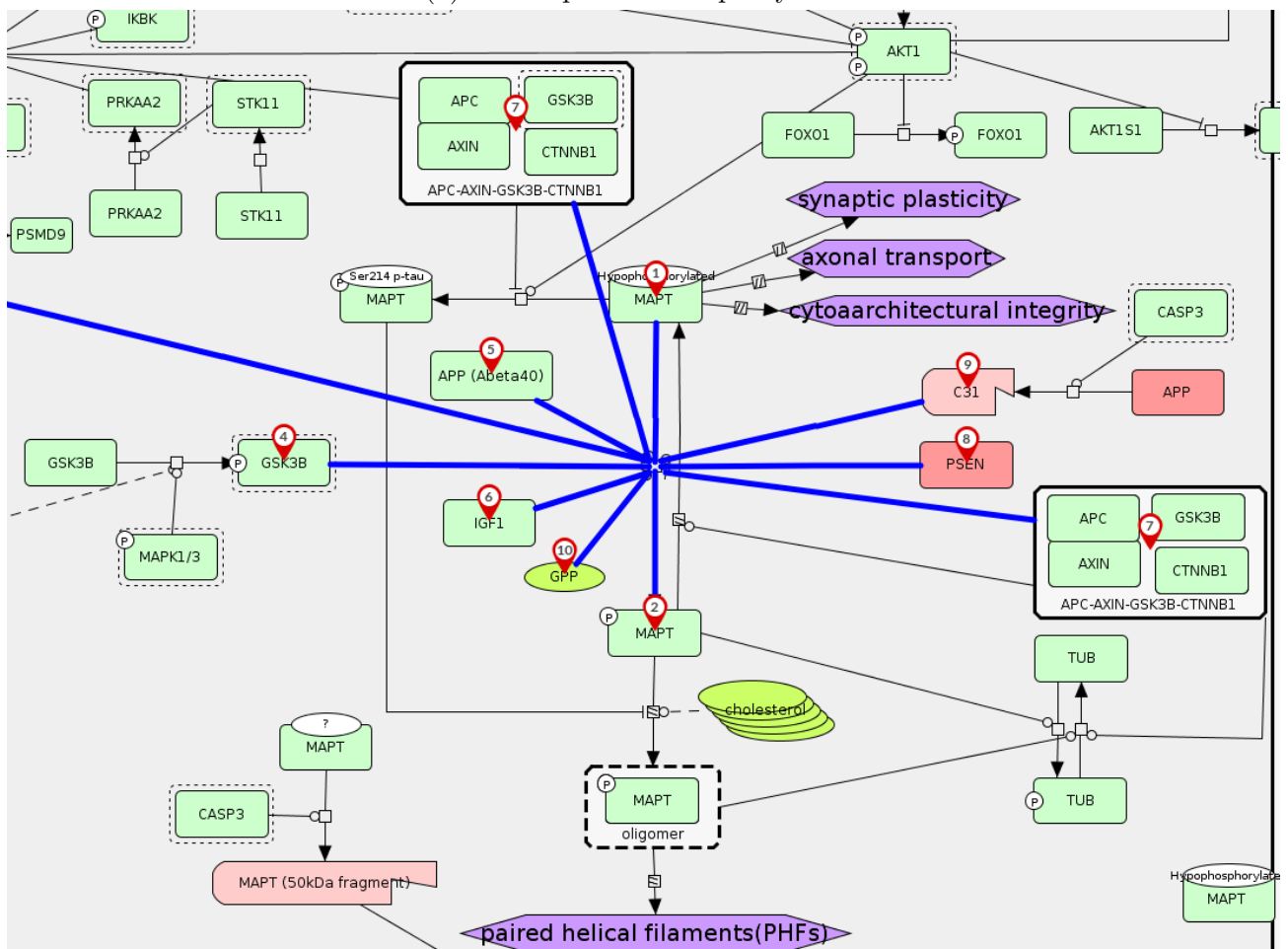
3.3.3 TAU (*MAPT*) hyper-phosphorylation

Microtubule-associated protein tau (*MAPT*), promotes microtubule assembly and stability, is supposed to be involved in the establishment and maintenance of neuronal polarity [Gendron, 2009]. Tau hyper-phosphorylation is one of the pathological hallmarks of Alzheimer's Disease. *MAPT* is also a risk factor in PD [Noble et al., 2013, Lei et al., 2010]

Although there were several reactions that were detected similar in PD and AlzPathway, they all differed in the modifiers involved (Figure 3.11). The only common modifier between all three reaction identified in the PD Map and the reaction in AlzPathway is *GSK3*. To summarise, although the Tau pathology plays a role in both diseases, the post translational modifications may be induced by different factors.



(a) PD Map: Tau Phosphorylation



(b) AlzPathway: Tau Phosphorylation

Figure 3.11: *MAPT* activity in AD and PD Map

3.3.4 MAPK signalling

Mitogen-activated protein kinases (*MAPKs*) are serine-threonine kinases that mediate intracellular signalling associated with cellular activities including cell survival, death, proliferation, and differentiation. MAPK signalling is reported to be involved in neuronal apoptosis in both AD and PD [Kim and Choi, 2010].

In PD, oxidative stress is a prominent cause of neuronal death. Studies have shown that ROS production induced by the toxins results in the activation of microglial cells, which subsequently attack neighbouring dopaminergic neurons. Amplified levels of α -Synuclein activates the *MAPK* pathway, resulting in subsequent inflammatory response [Fadaka et al., 2017]. This can also be observed in the PD Map in neuroinflammation, as seen in Figure 3.12)

In AD, Amyloid β aggregation triggers the activation of microglial macrophages, which then produce reactive oxygen species (ROS) and pro-inflammatory cytokines such as *TNF- α* and *IL-1 β* . These cytokines then stimulate the *MAPK* signalling pathway [Corrêa and Eales, 2012]. This was also observed in AlzPathway, as shown in Figure 3.13a.

MAPK8 in the AlzPathway was incorrectly annotated by mitogen-activated protein kinase 8 interacting protein 1 (*MAPK8IP1*) in addition to *MAPK8*. However, the *MAPK8* was detected as a match by updating the model annotations. The *MAPK8* transition in both AlzPathway and PD Map are modified by *MAP2K7* and *MAP2K4*. While in the PD Map, the reaction has only two modifiers, the AlzPathway has several other modifiers mediating the activation of *MAPK8*. Additionally, the reactions highlighted in Figure 3.13b were all in Apoptosis in the PD Map and mediated by activation of entities in neuroinflammation.

These observations support the current knowledge that *MAPK* signalling pathway contribute to neuroinflammatory responses and neuronal death and functional deficiencies in neurodegenerative diseases. Recently, many studies investigate possible role of *MAPK* as an attractive therapeutic target against neuroinflammation in AD, PD

and several chronic inflammatory diseases [Yarza et al., 2016, Fadaka et al., 2017, Lee and Kim, 2017]. Moreover, in recent years efforts have been made towards targeting the inhibition of *MAPK* pathways in AD and PD [Munoz and Ammit, 2010, Gehringer et al., 2015, Leonoudakis et al., 2017, Shah et al., 2017]

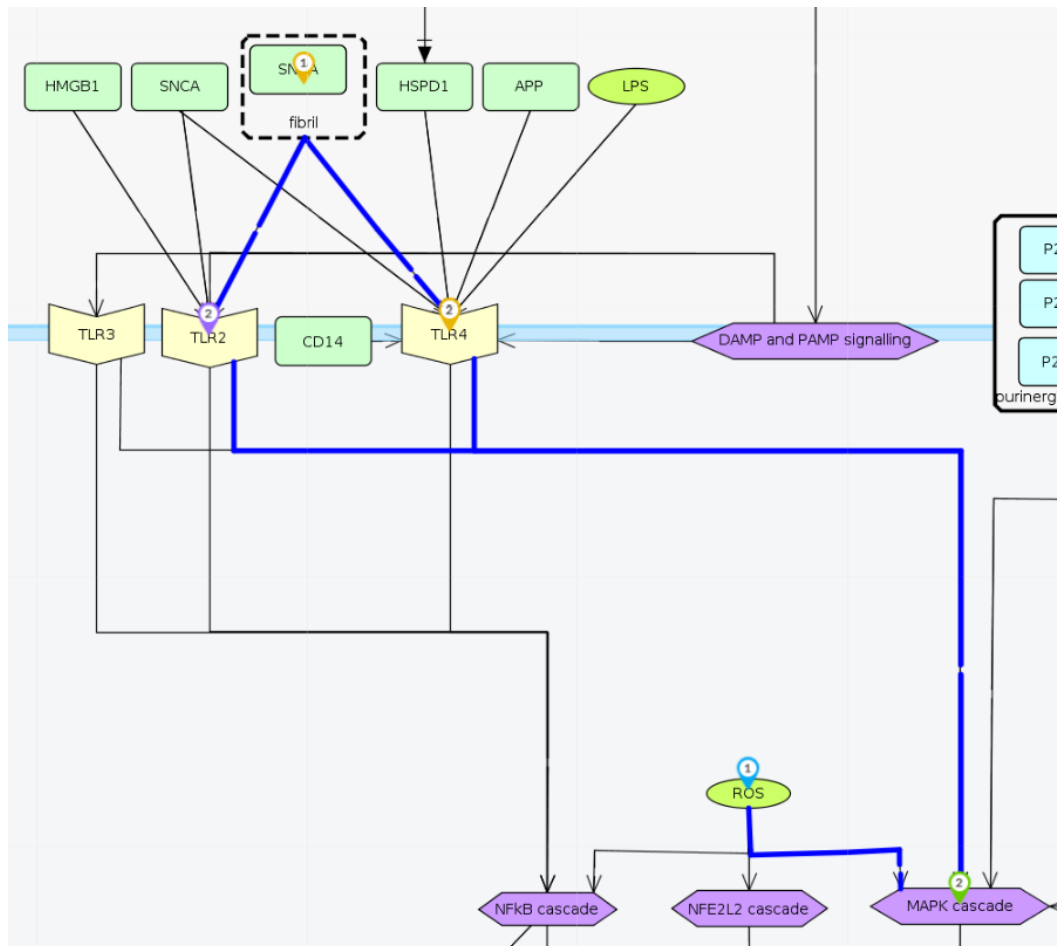
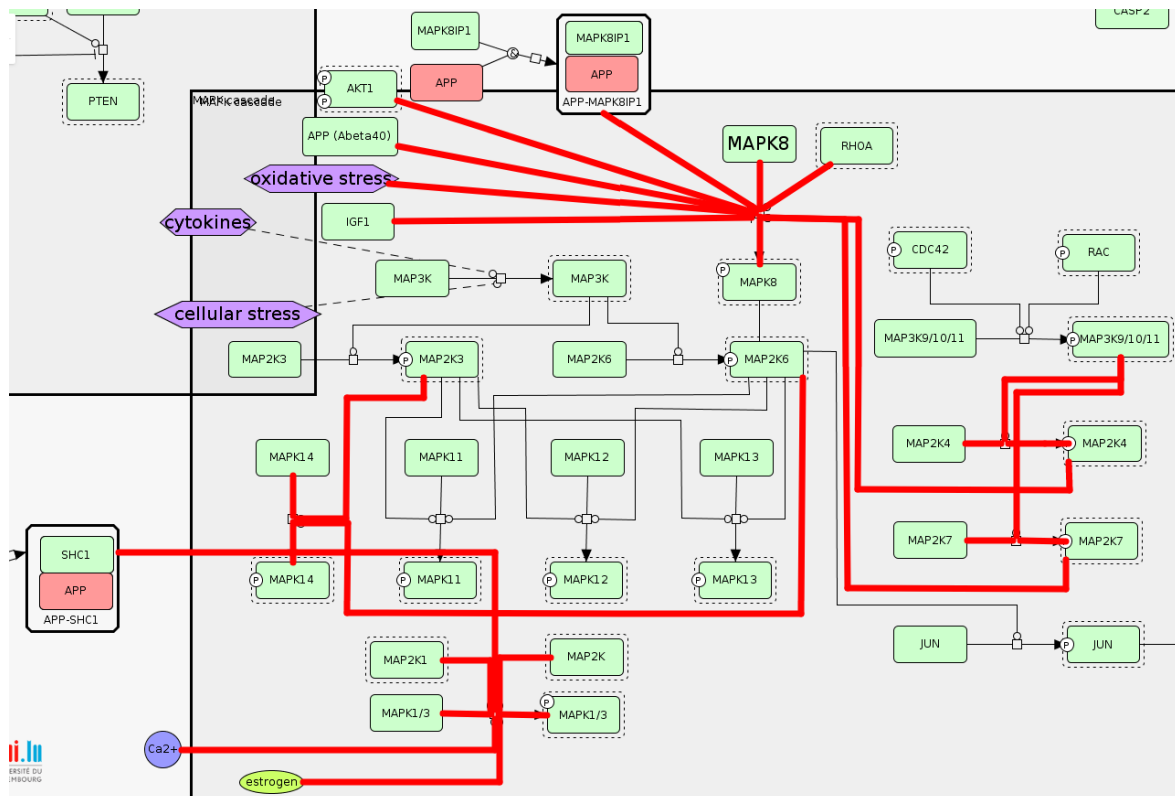
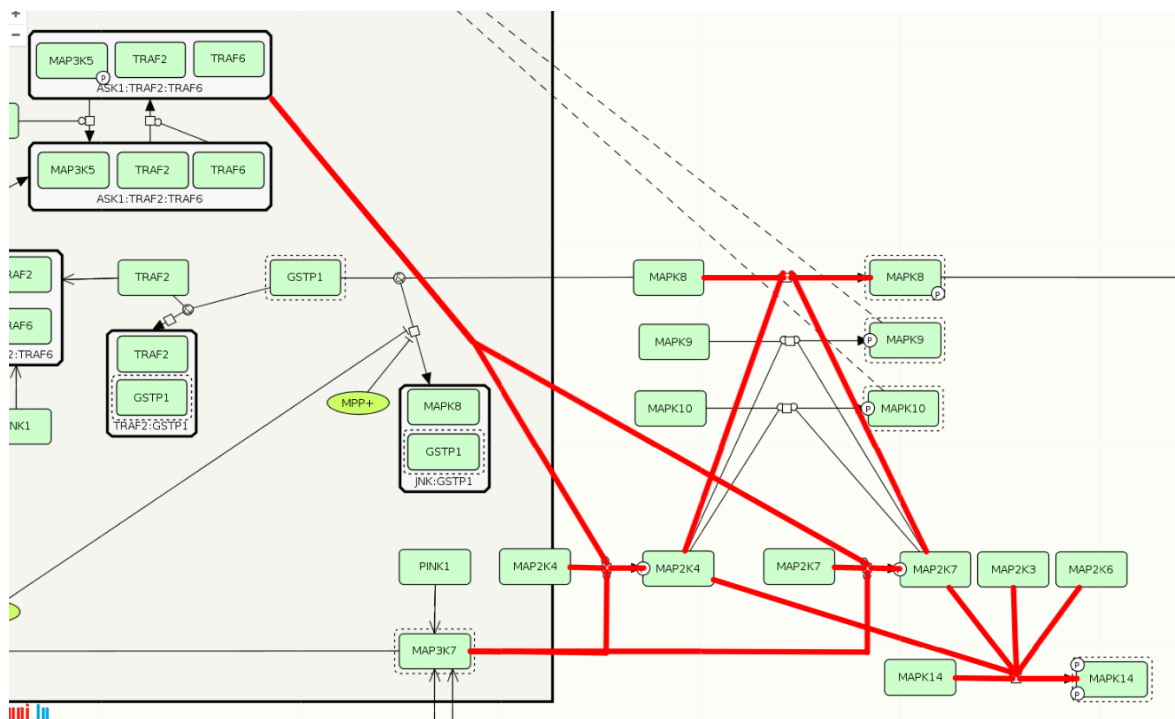


Figure 3.12: PD Map: *MAPK* cascade triggered by ROS and α synuclein fibrils



(a) *MAPK* cascade in AlzPathway highlighted in red are reactions found similar to PD Map



(b) PD Map: *MAPK* cascade in autophagy. Highlighted in red are reactions found similar to *MAPK* cascade in AlzPathway Map.

Figure 3.13: *MAPK8* signalling in AD and PD

3.3.5 Endoplasmic Reticulum Stress

Endoplasmic reticulum is responsible for the synthesis and folding of transmembrane and secretory proteins. In neurodegenerative disorders like AD and PD progressive loss of neuronal functions leads to accumulation of damaged proteins, which results in Endoplasmic Reticulum (ER) Stress. To restore homeostasis by clearance of misfolded proteins the unfolded protein response (UPR) pathway is activated. The UPR signalling mechanism improves the efficiency of protein folding and clearance of the abnormally folded proteins. However, under chronic ER stress, UPR fails to restore homeostasis and triggers apoptotic processes through alternate pathways to eliminate permanently damaged cells [Mercado et al., 2016, Urrea et al., 2013].

ER stress is modelled in both diseases maps. Several overlapping elements were found including *EIF2AK3 (PERK)*, *ERN1 (IRE1)*, *ATF6*, *EIF2A*, *XBP1*, and *HSPA5*. Moreover, one of the perfect matches in the comparison was involved in ER stress mediated transcription/translation Figure 3.14. Activated *EIF2AK3 (PERK)* mediates the activation of *EIF2A* in both AD and PD.

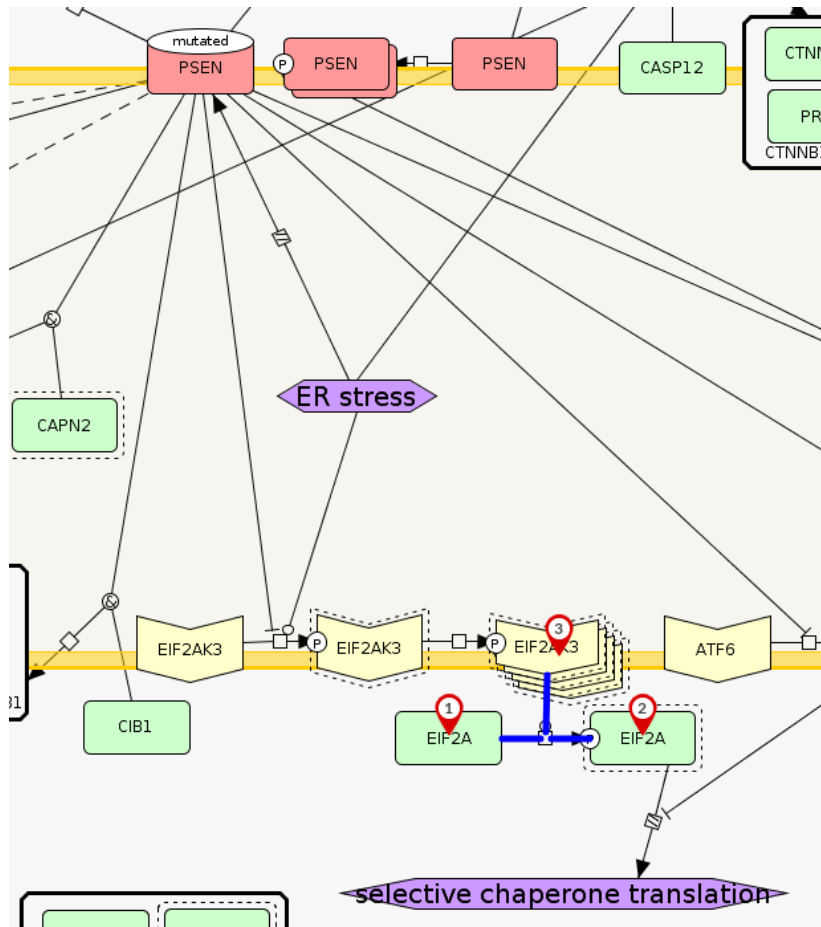
While ER stress in PD is mediated by *SNCA* fibril, in AD it is mediated by mutated presenilin. Although presenilin 1 was incorrectly annotated as *PSEN*, the algorithm could identify the entity as *PSEN1*. UPR is regulated by the three master regulators, *IRE1*, *PERK*, and *ATF6*. All the three proteins were found in both AlzPathway and PD Map. Further downstream, *ERN1 (IRE1)* activates *XBP1* which then induces ER-associated degradation (ERAD) [Cai et al., 2016], this can be seen in the PD Map (not shown here).

The second regulator, *PERK*, on activation induces phosphorylation of *EIF2A* causing translational arrest. Phosphorylated *EIF2A* activates *ATF4* increasing the levels of the transcription factor *CHOP*. *CHOP* then triggers the expression of several pro-apoptotic proteins [Urrea et al., 2013]. This can also be identified in the PD Map (Figure 3.14b, and downstream reactions) and selective chaperone translation in AlzPathway (Figure 3.14a)

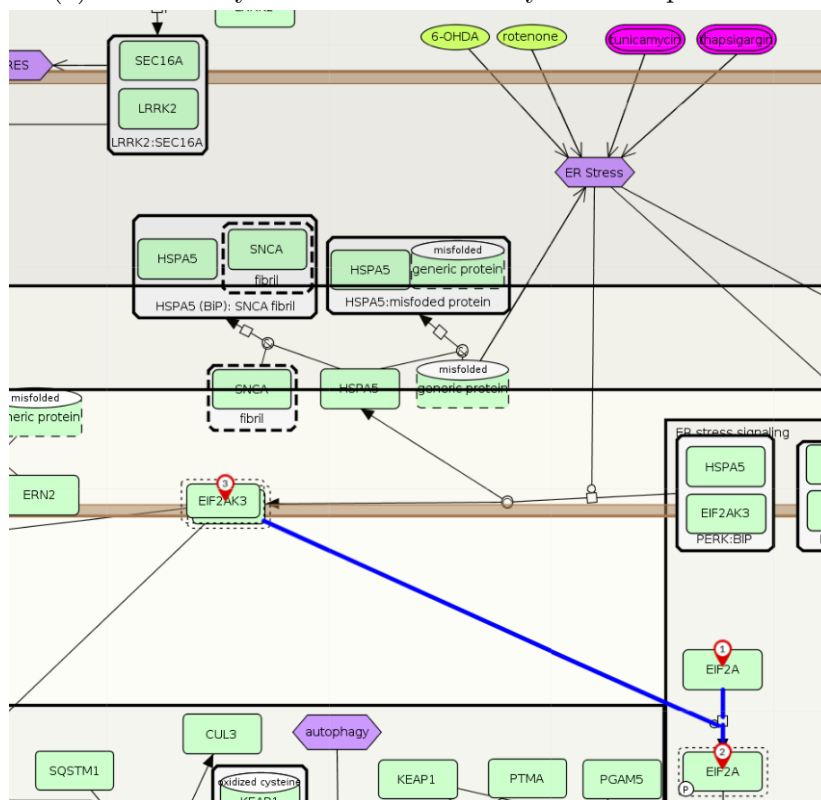
ATF6 traffics to Golgi apparatus where it is cleaved and fragment named *ATF6*. *ATF6* then translocates to the nucleus regulating the transcription of genes involved in ER homeostasis [Urrea et al., 2013]. This is modelled in the PD Map (re 4212 and downstream, not shown here) and *ATF6* mediating selective chaperone translation in AlzPathway (Figure 3.14a).

These similarities identified, reveal a complex scenario in which the ER stress response can have distinct downstream effects in both diseases and when different signalling components are manipulated. Moreover, as discussed earlier, the UPR can have different effects depending on the stage of the disease, initially working towards a pro-survival factor to restore homeostasis, but later triggering apoptosis to clear irreversibly damaged proteins.

Although response to ER stress were similar in both AD and PD leading to apoptosis in distinct regions of the brain, however the upstream triggers were different and alternate pathways leading to apoptosis exist in both diseases. Moreover, alterations in the function of the ER also play a major role in the aetiology of diseases like diabetes, cancer, heart diseases, inflammation and several other neurodegenerative [Hetz and Saxena, 2017, Osowski and Urano, 2011, Sano and Reed, 2013]. Together, studies of inflammatory disease and neuronal injury also support that persistent ER stress represents a more general mechanism of neurodegeneration that is triggered not only by the accumulation of disease-related damaged proteins but also by the pro-inflammatory environment that is observed in neurodegenerative diseases. Therefore, further efforts are needed to define the components of the ER stress response that could be specifically targeted for optimal disease interventions for distinct diseases.



(a) AlzPathway: ER stress induced by mutated presenilin



(b) PD Map: ER stress mediated by α synuclein toxicity

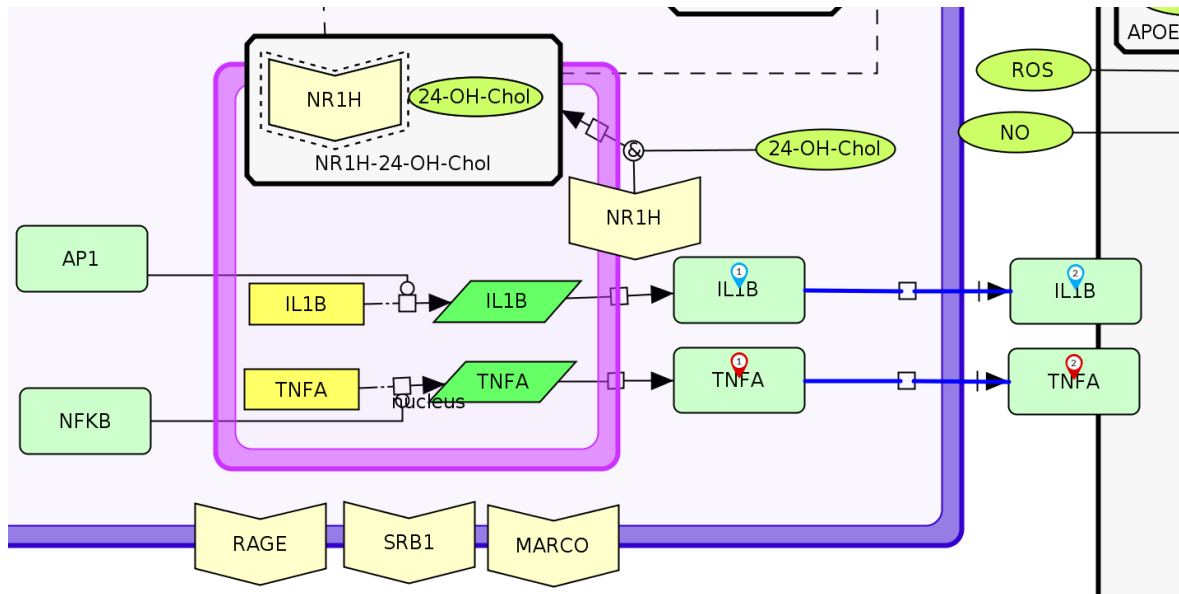
Figure 3.14: ER stress signalling in AD and PD

3.3.6 Inflammation

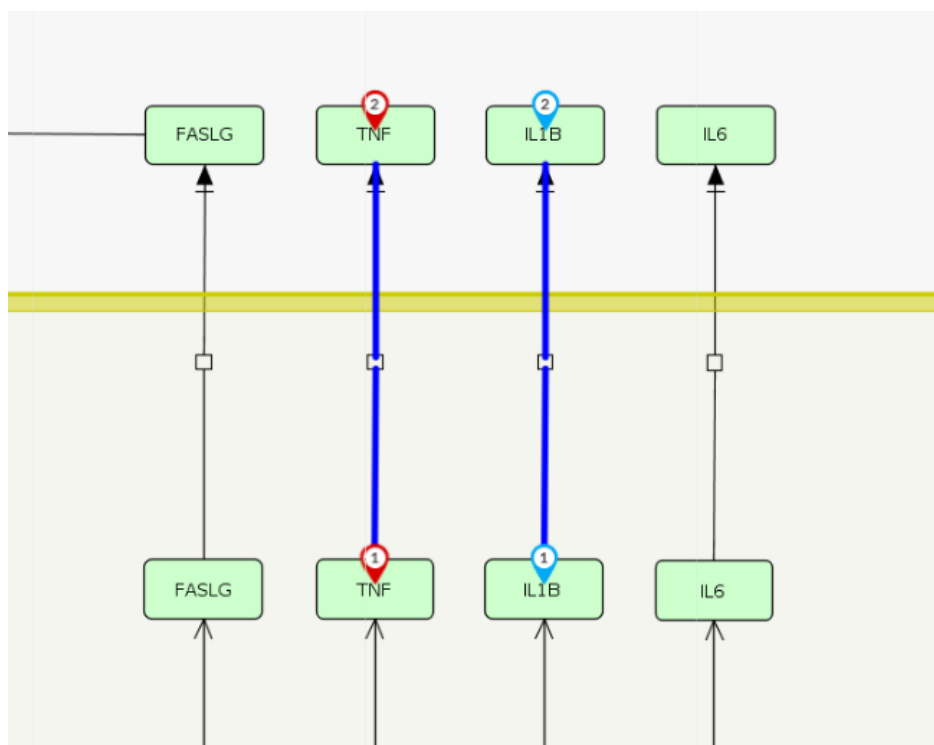
Inflammation is a natural response of the immune system to restore tissue homeostasis. Inflammation is triggered by stressful stimuli such as tissue damage and infections. The inflammatory response kills pathogens, repairs injured tissue and removes abnormal metabolite deposits. Neuroinflammation is reported in many neurodegenerative diseases and in some of them not only considered to be a consequence but also could trigger the pathology. Moreover, in many neurodegenerative diseases, inflammation markers are being investigated as diagnostic measures [Andreasson et al., 2016, Wang et al., 2015] (also discussed in section 2.6.1).

Three of the matched reaction were involved in neuroinflammation in both AD and PD. Figure 3.15, shows the transport of *TNF* and *IL1B* from the astrocyte in both AD and PD Maps. Also detected as similar was the transport of *IL6* from the microglia triggered by their inflammatory response (reaction 779 in AlzPathway, reaction 5241 in PD Map (not shown here)).

In many cases, AlzPathway entities were identified only after updating the annotations. For example, nuclear factor kappa B subunit 1 was annotated as *NFKB* instead of *NFKB1*, although named as subunit1. Similarly *IL6* in many instances were annotated as *IL6R* and annotated by two Uniprot identifiers. Several other cytokines including *AP1*, *TNF*, *IL6*, and *IRF* were also incorrectly annotated. No other reactions, other than discussed above were identified as similar in the neuroinflammation pathway, suggesting that the consequence of neuroinflammatory responses may be similar in both disease but the trigger could be distinct.



(a) AlzPathway: Transport of *TNF* and *IL1B* from astrocyte



(b) PD Map: Transport of *TNF* and *IL1B* from astrocyte

Figure 3.15: Inflammation triggering transport of *TNF* and *IL1B* from astrocyte

3.3.7 Wnt signalling

Figure 3.16 shows common reactions detected in Wnt signalling in AlzPathway highlighted on PD Map. Wnt signalling regulates several aspects of development including organogenesis, mid brain development especially in dopaminergic neurons and stem cell proliferation [Berwick and Harvey, 2014]. In PD, Wnt and β catenin signalling serves as the common final pathway for neuroprotection and self repair [Marchetti et al., 2013].

On the other hand, a variant of Wnt signalling pathway co-receptor *LRP6* is associated with late-onset of AD and presents low level of Wnt signalling activation. Wnt signalling is a neuroprotective mechanism against Amyloid β toxicity. With the increase in $A\beta$ aggregates, levels of *Dkk1* increases. *Dkk1* is a negative regulator of Wnt signalling, resulting in higher GSK-3 β activity. GSK-3 β activity is reported to be involved in several hallmark signatures of AD like the hyperphosphorylation of tau, increased memory impairment and increased production of Amyloid β [Inestrosa and Varela-Nallar, 2014, De Ferrari et al., 2007]. Also as discussed section 3.3.3, we identified *GSK3* as the only common modifier between all reactions matches in context to tau hyper-phosphorylation.

Similar to several other mechanisms discussed here, the reaction identified as similar are mediated by different modifiers signifying the upstream triggers in both disease are distinct.

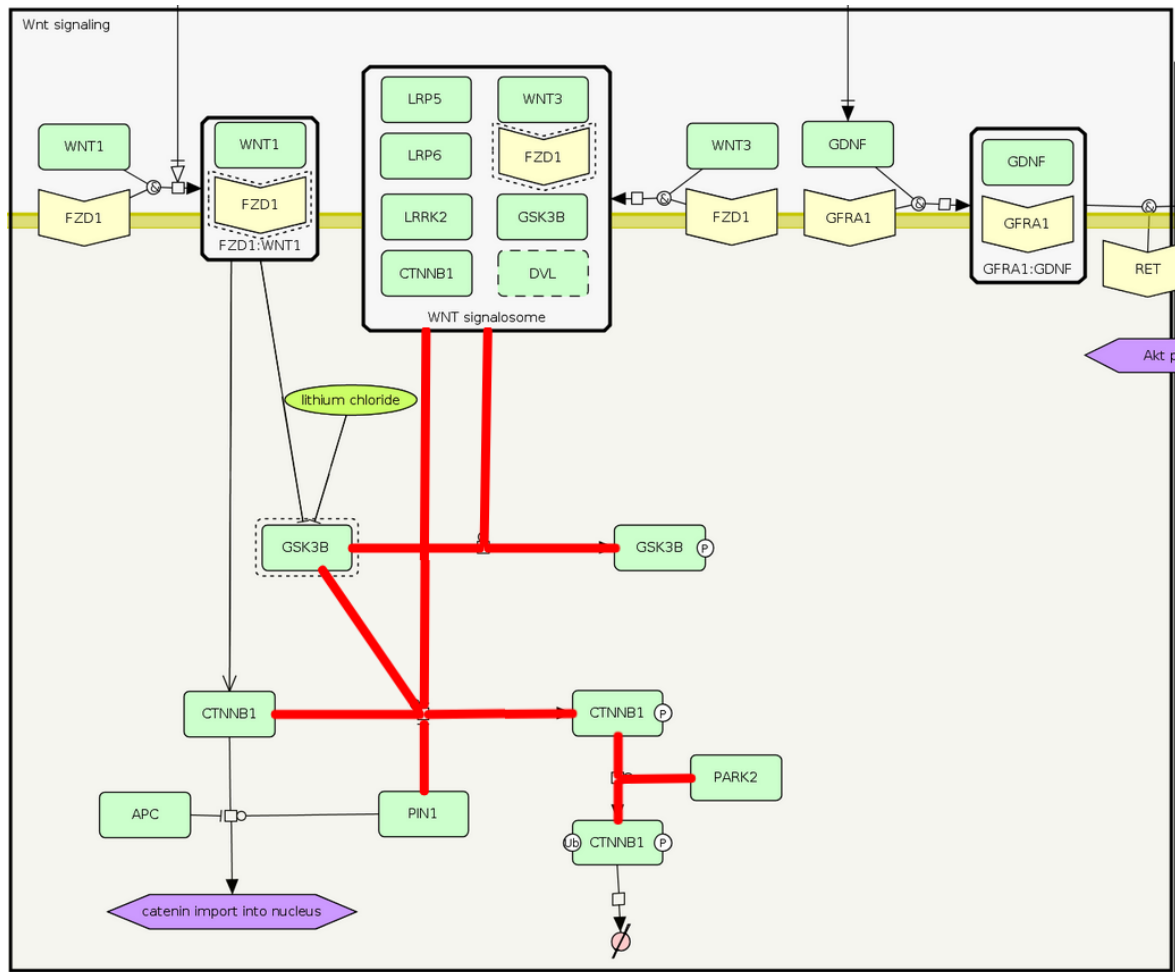


Figure 3.16: Reactions identified from AlzPathway highlighted in PD Map

3.3.8 Synaptic area

The synaptic area does not show much similarities. This is primarily due to the fact that the common receptors are represented as complexes without single proteins. Since the complex comparison is not robust, these entities could not be compared. Additionally, several entities in the AlzPathway were annotated. For instance, Calmodulin in the AlzPathway was annotated as "PICALM" (Uniprot identifier:Q13492), which is Phosphatidylinositol-binding clathrin assembly protein instead of "CALM1". Glutamate was annotated as "glutamate(2-)" in the AlzPathway and as "L-glutamic acid" in the PD Map.

3.4 Summary

Comparison of disease maps to detect similarities and differences was implemented and demonstrated using PD map and AlzPathway. Although many elements and reactions were found similar in many hallmark neurodegenerative pathways they were potentially triggered by different upstream events and cause different downstream effects.

In *AKT1* activity, *TSC* complex mediating activation of *RHEB* was detected as a similar downstream reactions of phosphorylated *AKT1* in both AD and PD.

Although *TSC* was not annotated in AlzPathway, the *RHEB* activation by *TSC* was identified as a match. *GSK3* was identified as the only common modifier for Tau hyperphosphorylation in both AD and PD maps. Additionally, in AlzPathway *Dkk1* negatively regulates Wnt signalling resulting in higher *GSK3* activity.

Although there were incorrect annotations for *MAPK* cascade elements in AlzPathway, several reactions were identified as similar with the PD Map. In the PD Map *SNCA* triggers the *MAPK* signalling pathway in the PD map, whereas Amyloid β aggregates trigger the *MAPK* cascade.

Similarly, while ER stress response has several alternate paths leading to apoptosis in both AlzPathway and PD map, it is mediated by *SNCA* in PD and mutated *PSEN1* in AlzPathway. Presenilin 1 was incorrectly annotated as *PSEN* in AlzPathway instead of *PSEN1*.

NFKB1 was annotated by two Uniprot identifiers and *IL6*, *TNF*, etc., were not annotated. However, the annotated based comparison could identify several reactions involving these cytokines as similar in inflammatory responses in both AD and PD.

Several entities in the AlzPathway were annotated by two Uniprot identifiers, leading to ambiguity in the identity of the elements. The annotation-based comparison could overcome the issues of multiple annotation, missing annotations and different

encodings (HGNC and Uniprot) of various proteins between PD and AD maps. However, incorrect annotation of entities, for instance, *PICALM* for Calmodulin, could not be detected automatically. Such errors in the model generation will limit the interoperability of models. The results of comparison between AlzPathway and PD Map emphasise the need for curation standards to ensure higher quality and interoperability of disease maps.

Chapter 4

Comparison of different disease models

As discussed earlier, all the published disease maps are built in CellDesigner. However, such maps are only a part of the knowledge landscape. Other formats like BioPAX and OpenBEL are also used to built disease models (Section 1.3.3) While SBML and BioPAX have several software packages and packages for visualisation and analysis, the framework and environment supporting OpenBEL is much more limited. For instance, while OpenBEL captures the context of the relationship, this information is not used for visualisation in cytoscape. This limits the re-usability and interoperability of OpenBEL models [Hoyt et al., 2017].

System Biology Format Converter (SBFC) provides a generic framework to includes several converters translating between several formats including SBML, BioPAX, and SBGN-ML [Rodriguez et al., 2016]. Knowledge assemblers, like INDRA, provide support for import of many formats and PyBEL [Hoyt et al., 2017] enables the import of OpenBEL documents into a common format in INDRA. However, there exists no converter from OpenBEL to SBML. Therefore, a converter from OpenBEL to CellDesigner format was implemented. The converter was used to generate a node file and reaction file from an XBEL file. The converter is available at <https://git->

r3lab.uni.lu/aishwarya.alex/xBELtoCellD. The list of nodes and reactions can then be used to generate an CellDesigner model. The generated CellDesigner model can then be visualised using the MINERVA platform.

4.1 Interoperability between models

The previous chapter details the comparison between two CellDesigner maps. The next step was to convert OpenBEL models to the CellDesigner format. As a use case, we convert the APP OpenBEL model. The species and reactions in CellDesigner and their corresponding representation in OpenBEL are represented in the Appendix A.2. The namespaces listed in the table are limited to the ones used in the APP model. Since both modelling language have different purposes the conversion is lossful.

Each OpenBEL statement records a biological fact and can be annotated with references, typically a PubMed ID. Additionally, each statement can also be associated with a set of annotation that describes the context in which the statement was observed. This adds to the knowledge associated with the statement being captured. The additional information about the encoded statement such as the tissue, species, and cell line can then be used for hierarchical organisation of the map. However, these annotations are optional.

```
SET Disease = "Atherosclerosis"
SET CardiovascularSystem = "Arteries"
SET TextLocation = "Review"
SET Evidence = "Oxidation and nitration of macromolecules, such
               as proteins, DNA and lipids, are prominent
               in atherosclerotic arteries."
SET Citation = {"PubMed", "Trends in molecular medicine", "
               12928037", "", "de Nigris F, Lerman A, Ignarro LJ, Williams-
               Ignarro S, Sica V, Baker AH, Lerman LO, Geng YJ, Napoli C", ""
               }
```



```
pathology(MESH: Atherosclerosis) positiveCorrelation
    biologicalProcess(GO: "lipid oxidation")
```

Listing 4.1: BEL Statement Example

An example of a OpenBEL statement and associated annotation is shown in Listing 4.1. The statement with the biological fact is provided as the evidence. The statement is accompanied by the citation and additional information e.g. the disease and tissue. These information provided is essential to annotate the entities and reactions in the context in which they should be added to the model. The subject (*pathology(MESH:Atherosclerosis)*) and object (*biologicalProcess(GO:"lipid oxidation")*) in OpenBEL, are translated to reactants and products (or modifiers) in CellDesigner. Relationships in OpenBEL are converted to reactions in CellDesigner.

To identify each element (subject or object) is annotated by a namespace. While the OpenBEL Language Documentation [BEL v2.0 Language Documentation, <https://github.com/OpenBEL/language>, Date Accessed: 30 May 2018], recommends as best practice the use of well defined domains, external vocabularies and public ontologies to define entities, it must be noted that the users are free to define and use their own vocabularies to refer to entities.

Legacy and custom namespaces

OpenBEL also provides a list of legacy namespaces for domains such as chemical, protein families, etc. These are lists of accepted names for chemicals, protein families, etc., and also allows entities with no namespace. The legacy namespaces cover each of the OpenBEL function types. These namespaces are available through their resource framework at <http://resources.openbel.org/belframework/>. However, this has not been updated to reflect the changes in the underlying resources. Additionally, most of these resources identify entries by name and not a unique identifier. Due to the large number of terms across many namespaces, the main challenge was to extract the namespace and corresponding identifiers to generate a MIRIAM uris.

For entities that are not annotated or use the OpenBEL legacy namespace, we use the MeSH dictionary and reflect-client [O'Donoghue et al., 2010] to query GO and Ensembl to try to extract the identifiers. This approach was also used to extract the entities which were annotated by namespace and name alone, to extract the corresponding identifier. E.g.: abundance (amyloid beta peptides). Such terms were looked up in GO, MeSH and Ensembl. If not retrieved, they were set to unknown species in CellDesigner.

Missing Namespaces and Improper Names

OpenBEL facilitates the use of openly shared controlled vocabularies (namespaces) to promote exchange and consistency of information. Finding an appropriate namespace-identifier pair is often an essential part of the curation process. An important point to be noted is the element "abundance" which can represent any entity which does not fit any of defined species like protein, gene etc.. OpenBEL uses this element to encode elements of unknown quantity like chemicals, metabolites, ion, and peptides.

An overview of the conversion process is given below. The algorithm takes as input the OpenBEL model in .XBEL format. The OpenBEL framework provides tools to convert the OpenBEL document into .XBEL, a parsable xml format. The converter extracts the relationships and involved subject and object to return two lists, a node list and a reaction list. The node list contains the elements, their unique identifier as miriam uris, and location information. The reaction list contains reactants, products and modifiers (if any) and the corresponding citation from which the reaction was extracted. The reaction lists refer to the elements involved by the identifier used in the node list. The elements in the node list can be reactants, products or modifiers in one or more reactions. Therefore, we create a list of nodes with unique identifiers which is used to refer to the same element as a participant of multiple reactions. An extract of the result tables generated during the conversion of the APP OpenBEL model are attached in the Appendix.

Algorithm 1: Converting XBEL to nodes and reactions

Data: OpenBEL model in .XBEL**Result:**

nodes list (nodes with annotation and location information)

reactions list (reactions, with nodes referring to nodes list and annotation)

for *Each statement* **in** *file* **do**

Extract annotations : citation, cell, cell line, tissue, disease, etc., ;

for *Each relationship* **in** *statement* **do** **if** *valid relationship* **then**

get subject ;

if *object is nested* **then** modifier \leftarrow subject ; modifying relationship \leftarrow relationship ;

repeat section for relationship in object ;

else

get object ;

end **if** *subject, object and modifier exists in nodes* **then**

retrieve identifiers ;

else

create nodes for subject, object and modifiers ;

retrieve identifiers ;

end

add new relationship with subject, object and modifier to reaction ;

else

skip to next relationship ;

end **end****end**

While CellDesigner is a process focused representation, the information captured is detailed based on interaction between entities rather than directly on entities. OpenBEL on the other has its focus in relationship between entities and has specific representation for such relationships which are missing in CellDesigner.

OpenBEL allows nested statements. They are translated as reactions with modifiers in CellDesigner. Figure 4.1, the subject of reaction1 acts as a modifier to the nested reaction in CellDesigner.

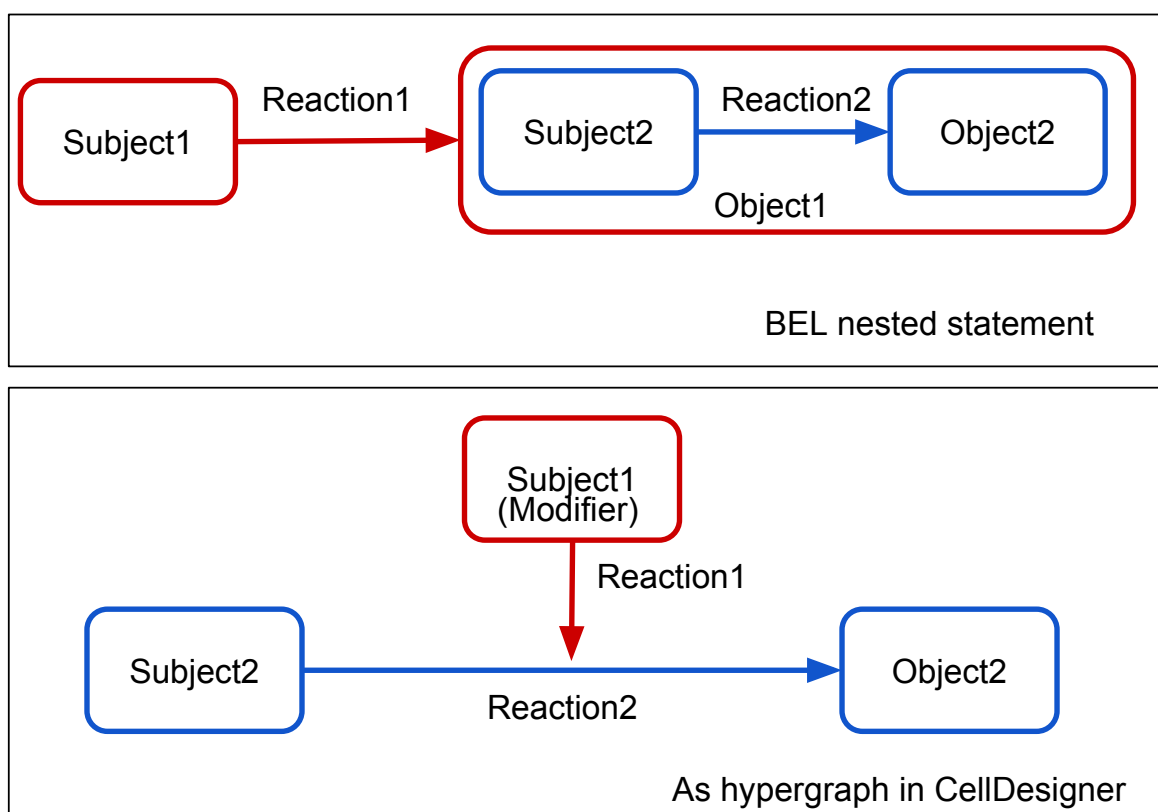


Figure 4.1: Nested statements in OpenBEL

However, this is only done for reactions which allow modifiers like state transition and transport. For reaction types that do not permit modifiers, the reactions are converted to two separate reactions:

1. modifier(subject1) -modifying relationship (reaction1)- reactant (subject2)
2. reactant(subject2) -relationship (reaction2)- product(object2)

Currently the converter handles the OpenBEL predicates following the mapping shown in Table 4.1.

OpenBEL predicates	CellDesigner reaction
increases	UNKNOWN_POSITIVE_INFLUENCE
directlyIncreases	POSITIVE_INFLUENCE
directlyDecreases	NEGATIVE_INFLUENCE
decreases	UNKNOWN_NEGATIVE_INFLUENCE
positiveCorrelation	UNKNOWN_REDUCED_MODULATION
negativeCorrelation	UNKNOWN_REDUCED_MODULATION
translocation	STATE_TRANSITION
MODIFIER_increases	UNKNOWN_CATALYSIS
MODIFIER_directlyIncreases	CATALYSIS
MODIFIER_directlyDecreases	INHIBITION
MODIFIER_decreases	UNKNOWN_INHIBITION

Table 4.1: OpenBEL predicates and corresponding representation in CellDesigner

As mentioned earlier, due to the differences in the language structure we expect a loss of information upon conversion. Some of the predicates in OpenBEL do not have a corresponding equivalent representation in CellDesigner, e.g. negativeCorrelation, positiveCorrelation, association, biomarkerFor, hasComponents, prognosticBiomarkerFor, complexAbundance. These relationships are not represented in CellDesigner since they do not directly describe a process, and add no mechanistic value to the model. Such relationships in OpenBEL are ignored by the converter. OpenBEL captures the context of the reaction using the "SET" statements (See listing 4.1). The converter extracts this information to organise the model hierarchically into compartments for better visualisation and context. The cell and cell line information provides the context in which the reaction occurs and are represented as compartments in CellDesigner.

OpenBEL activity term	GO annotation
catalyticActivity	urn:miriam:obo.go:GO:0003824
chaperoneActivity	urn:miriam:obo.go:GO:1903332
gtpBoundActivity	urn:miriam:obo.go:GO:0008277
chaperoneActivity	urn:miriam:obo.go:GO:1903332
kinaseActivity	urn:miriam:obo.go:GO:0016301
peptidaseActivity	urn:miriam:obo.go:GO:0008233
phosphataseActivity	urn:miriam:obo.go:GO:0016791
ribosylationActivity	urn:miriam:obo.go:GO:1990404
transcriptionalActivity	urn:miriam:obo.go:GO:0006355
transportActivity	urn:miriam:obo.go:GO:0005215
degradation	urn:miriam:obo.go:GO:0009056

Table 4.2: OpenBEL activity terms and corresponding GO annotation

To specify distinct molecular activity of protein, complex, and RNA, OpenBEL uses the "activity" functions providing distinct terms that differentiate these activity from the abundance.

For example, `kinaseActivity(proteinAbundance(HGNC:AKT1)) directlyDecreases transcriptionalActivity(proteinAbundance (HGNC:FOXO1))`, indicates that the kinase activity of AKT1 directly decreases the transcriptional activity of *FOXO1*. These functions are annotated as GO terms (Table 4.2) and added as additional information to the specific node (Protein, Gene, etc.) by the converter.

4.2 Results

The APP OpenBEL model was converted in to an CellDesigner format using the converter and visualised using the MINERVA platform. The conversion process consists of several steps. First, the OpenBEL model in .XBEL format was parsed by the converter and the entities and their annotations were extracted to generate a nodes list. At this stage , the converter also extracts annotation for unannotated elements and legacy namespaces. Next, the relationships in OpenBEL are translated to reaction types in CellDesigner and a reaction list was generated referring to elements from the node list and also the cellular location extracted from the OpenBEL model. The node and reaction list were then converted to an SBML format. Finally, to provide a layout to the model, the SBML was converted to CellDesigner format.

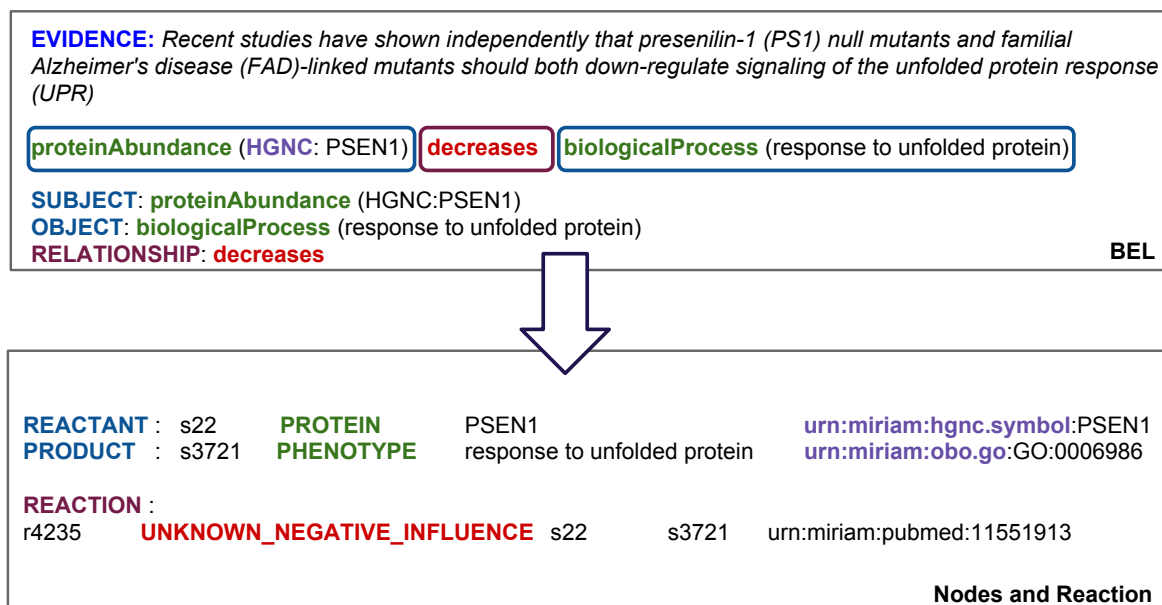


Figure 4.2: Example of a lossless statement conversion

Figure 4.2 shows an example of a statement that could be translated without any loss of knowledge. Moreover, the object in the OpenBEL statement *biologicalProcess (response to unfolded protein)* had no annotation associated with it. This term was retrieved as a GO term. From a total of 7083 statements in the APP OpenBEL model 4347 were partially or completely converted to the CellDesigner model. Table

4.3 gives a summary of the relationships that were ignored or partially translated.

Relationship	Count	Partially converted
Association	1445	9
Decreases	17	17
Increases	86	86
directlyIncreases	1	1
biomarkerFor	19	0
causesNoChange	54	0
ComplexAbundance	1014	0
hasComponents	2	0
hasMember	1	0
hasMembers	54	0
isA	56	0
negativeCorrelation	1	1
rateLimitingStepOf	1	0
translocation	52	4

Table 4.3: OpenBEL statements lost in conversion

Some statement with the following relationships: decreases, increases, directlyIncreases, translocation and negativeCorrelation were partially converted because they were a nested statements with a modifying relationship that did not have any equivalent. In such cases, only the nested statement was translated. In the case of translocation, statements which were missing either a from or to location were also ignored.

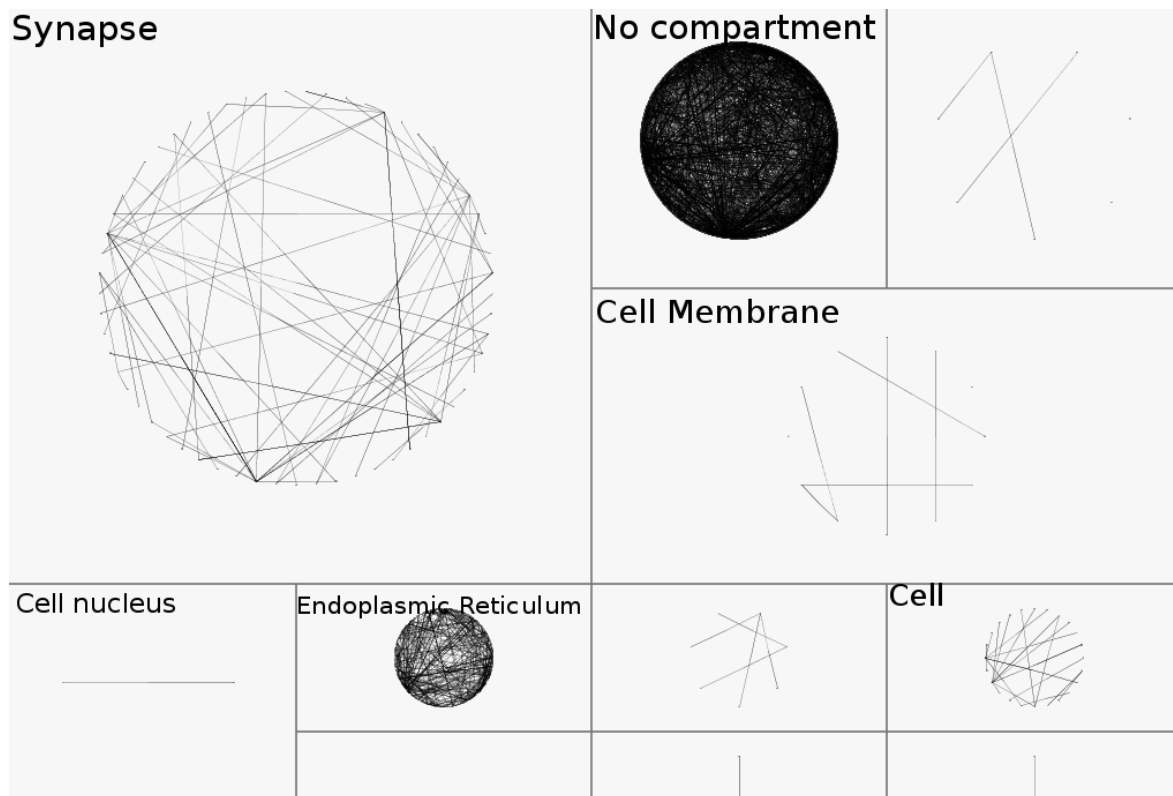


Figure 4.3: A version of the APP OpenBEL model converted to CellDesigner visualised using the MINERVA platform.

In Figure 4.3, each box represents the context (cellular location or mechanism) where the reactions occur. The top left box represents the synapse. The *No compartment* box had no context (no tissue, cell type or compartment) associated with it. This is an example of how a model built without required standardized notation for context would be visualised. There is a very dense network, but offers very little use in any knowledge exploration, or data interpretation.

The generated model that had proper context annotation was quite large and therefore was split into five maps depending on the cell type of the reactions involved. Table 4.4 shows components of each of the smaller maps.

Celltype group	Components
Blood	Blood Cells, Blood Platelets, Lymphocytes, Erythrocytes, Leukocytes
Neurons	Neurons, primary neuron, Dopaminergic Neurons, Pyramidal Cells, Motor Neurons, primary cortical neuron, Interneurons
Only tissue	Bone Marrow Cells, Beta cell, INS-1 cells, Myocytes, Smooth Muscle, Endothelial Cells, Human Umbilical Vein Endothelial Cells, Fibroblasts, Neural Stem Cells
Glial cells	Astrocytes, Microglia
Others (Model systems)	Neuroblastoma cell, CHOAPPsw, N2a695 cell, 293, NT2N cells, PC-12

Table 4.4: Submaps components based on cell type

The only tissue map comprises of reaction and nodes which had no precise inter cellular location information associated to them, but only tissue information. Many statements tagged only with tissue information, were pertaining to literature about cerebrovascular diseases, Diabetes Mellitus Type 2, etc., investigating cross-talk with significant pathways of AD [Ubeda et al., 2004, Freude et al., 2009, Jung et al., 2003]. These interactions may not be specific to AD, but are interesting for co-morbidity studies. Additionally, several statements were extracted from literature reporting results from model systems and cell lines, such reactions were included in the model system map.

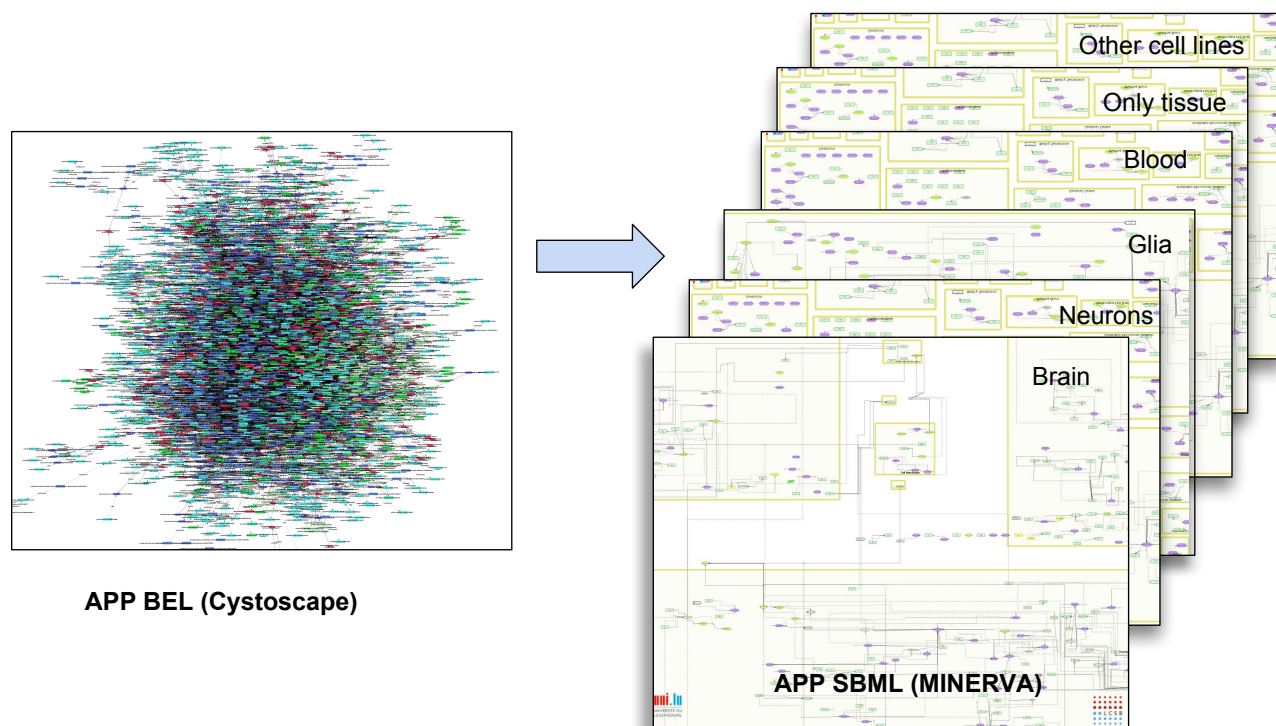


Figure 4.4: APP OpenBEL model converted to CellDesigner, visualised in MINERVA

Figure 4.4 highlights the difference in visualisation using contextual information. The map provides an easily navigable, hierarchical and compartmentalised structure as opposed to the dense "hairball" of the OpenBEL model. Moreover, the community support for SBML compatible software is much larger than the OpenBEL community, providing better analytical and exploratory tools for the SBML network.

4.3 Comparison of APP map to PD map

The converted APP submodels were compared to the PD Map with methods described in Chapter 3. Figure 4.5 shows the matches highlighted on the PD map. Although all the submaps identified many elements in common with the PD Map only one reaction was identified as similar; the transition of *APP* to Amyloid β 42. This reaction was identified both in the APP blood and APP neuron maps.

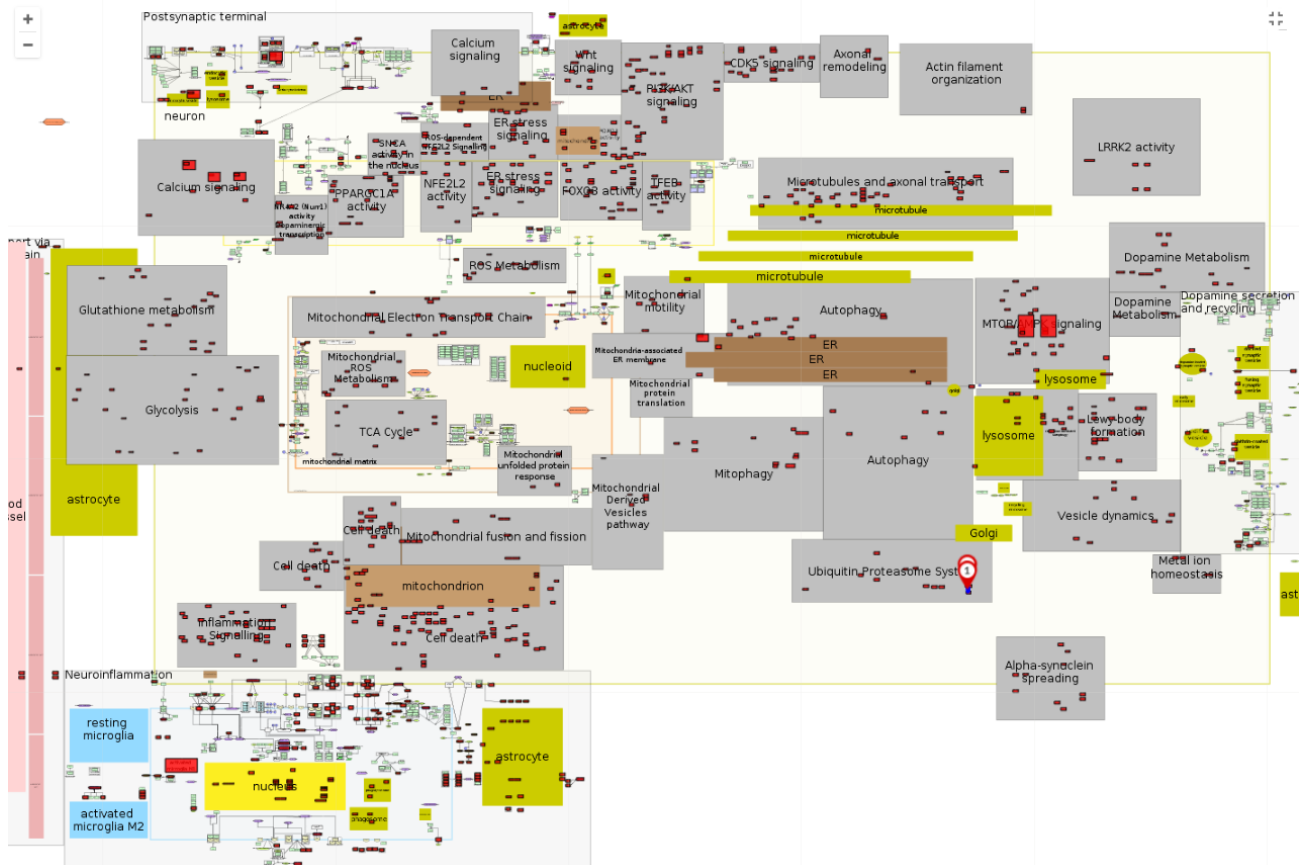


Figure 4.5: Elements and reaction from APP model on PD map

The identified elements were mainly concentrated around the neuroinflammation, cell death and ER stress signalling neighbourhood on the PD Map. ROS activity and the iron metabolism submap also detected matches in the PD Map. The Alzpathway also detected similarities in these regions.

However, the absence of similar reactions is evident. This may be primarily due to the fact that the distribution of reaction types in both the maps are considerably different. Table 4.5, shows all the reaction types in the APP map and the number of such reactions in the PD Map. It is evident that the APP model has larger number of reaction of type modulation and influence, arising from the "increases" or "decreases" relationships in OpenBEL. The greater number of such interactions is because OpenBEL aims to capture causal and correlative relationships between entities. Additionally, unlike the AlzPathway, the APP model is primarily focused around the Amyloid β Pathology, therefore it is not surprising that there were not many reaction

similarities with the PD Map.

Reaction type	Count (APP Map)	Count in PD Map	Count in AlzPathway
NEGATIVE_INFLUENCE	25	206	2
STATE_TRANSITION	49	678	380
POSITIVE_INFLUENCE	76	1060	9
UNKNOWN_REDUCED_MODULATION	525	13	1
UNKNOWN_NEGATIVE_INFLUENCE	1059	26	10
UNKNOWN_POSITIVE_INFLUENCE	2498	2	14

Table 4.5: Reactions in the APP Map

4.4 Comparison of APP map to AlzPathway

On comparing the APP maps with the AlzPathway, 308 elements were found in common, out of which only 125 unique elements. The blood and brain submap had the maximum number of matches. Figure 4.6 show the elements from APP maps highlighted on AlzPathway.

The matched elements from APP are spread out on the Alzpathway, except around the region of cholesterol metabolism. Interestingly, none of the elements of type "phenotype" were matched, although both APP and AlzPathway have a large number of elements of type "phenotype". This could be due to the fact that phenotypes elements were not annotated in AlzPathway. On the other hand, phenotypes like "long-term memory", "astrocyte activation", "neuroinflammation", "ER stress response", etc., were detected as matches between the PD Map and APP maps. Therefore, we could expect an increase in the number of matches, if phenotypes were accounted for.

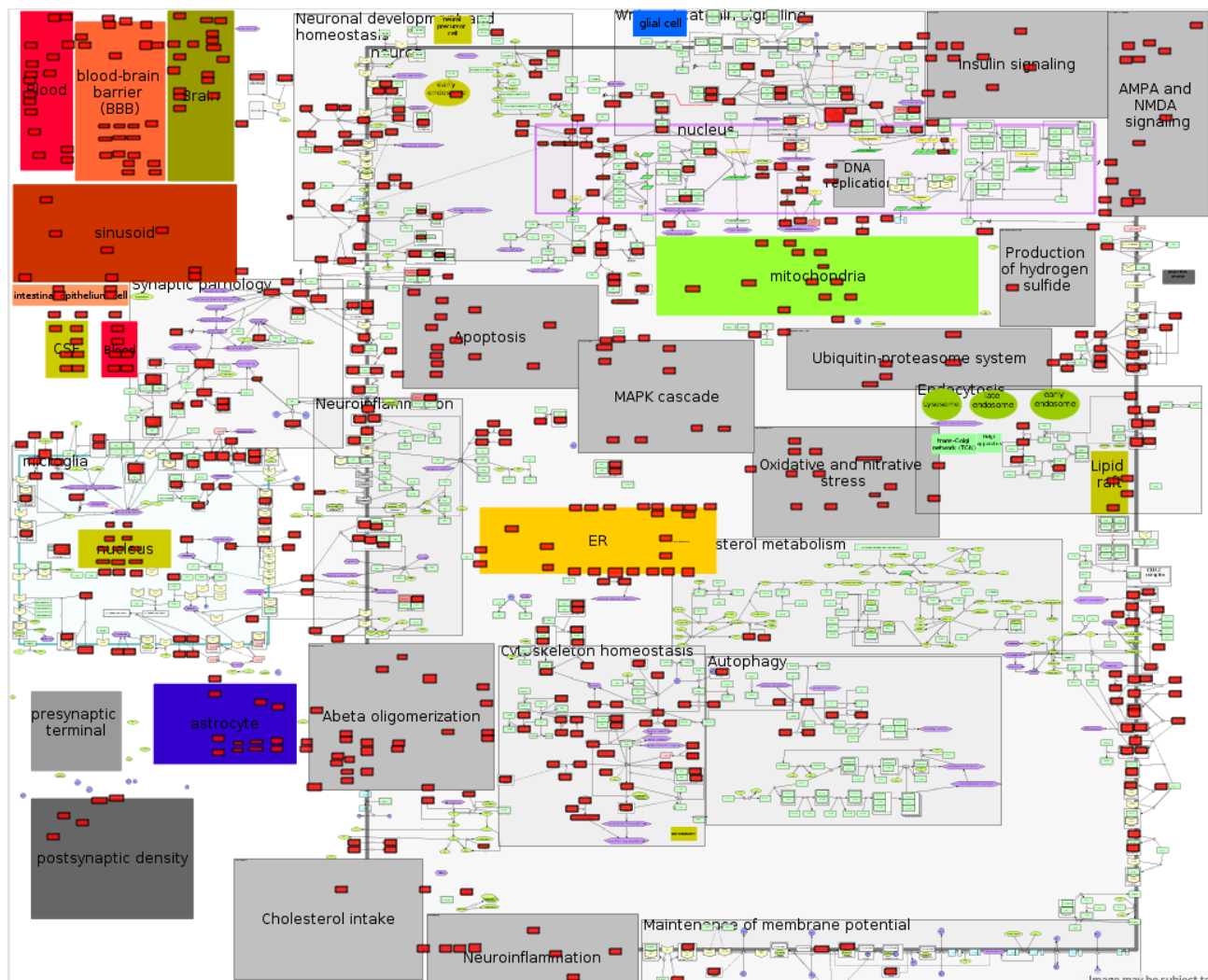


Figure 4.6: Elements from APP model highlighted on AlzPathway

Figure 4.7 shows the results for a search for "APP" on the AlzPathway, the results include "APP, APP (Abeta), APP (Abeta 40), APP (Abeta 40, APP (c99)". The distribution is similar to all matches detected from the APP maps, since the APP map was built around the Amyloid Beta pathology.

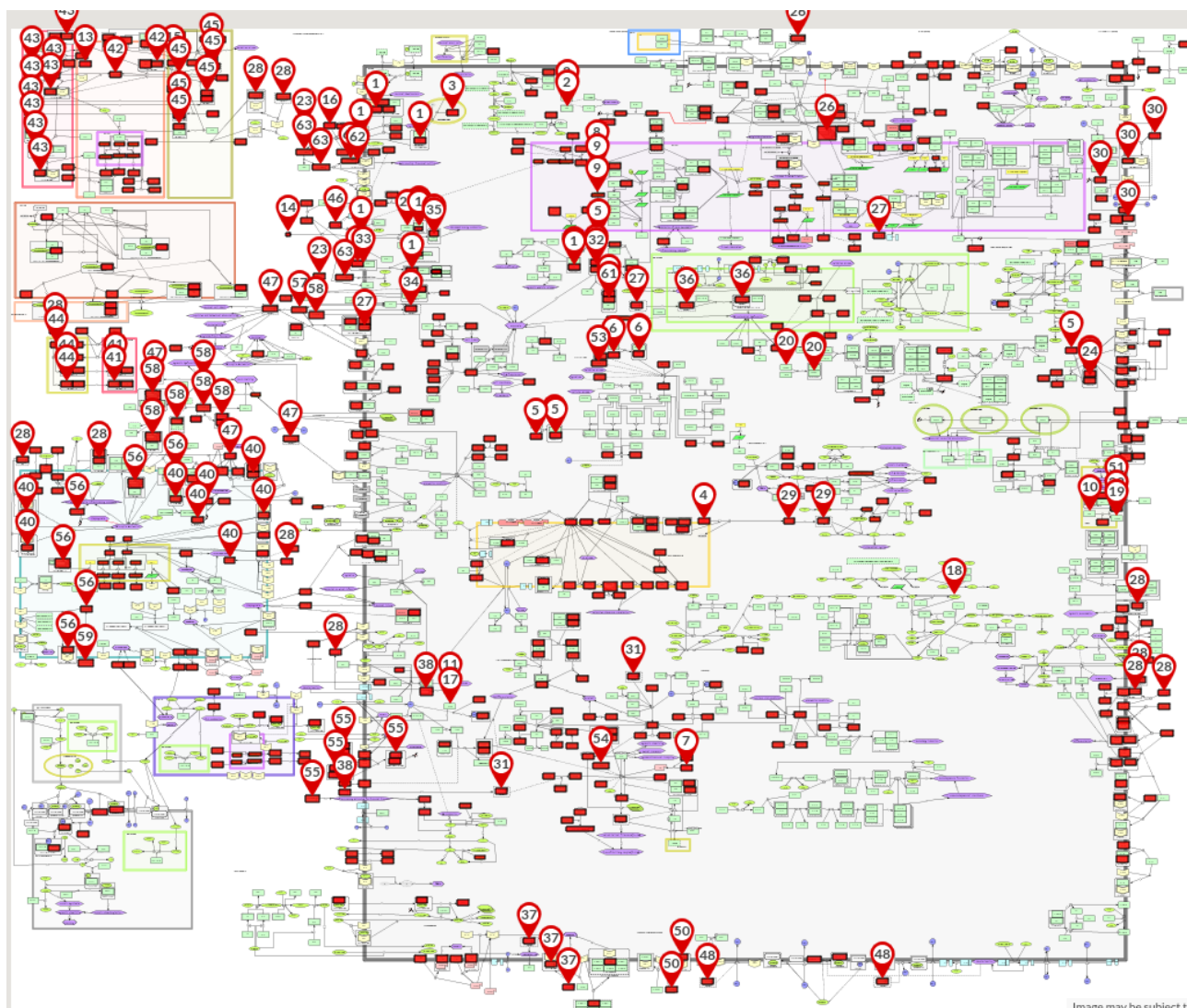


Figure 4.7: APP and A β highlighted on AlzPathway

Similar to the reactions detected in comparison to the PD Map, number of reactions detected as matches was lower than expected based on the fact, that both resources are models of Alzheimer's disease pathology. Only 5 reactions, were detected as similar in the AlzPathway. All five reactions were involving *APP*. As discussed earlier, this was primarily due to the reaction type, the APP map has greater reactions of *unknown influence or modulation* (Refer Table 4.5). Next, we identified the reactions that have the same interacting elements i.e. reactants and products but not necessarily the same reaction type. This approach detected 31 reactions. Figure 4.8 shows these reactions highlighted on the AlzPathway.

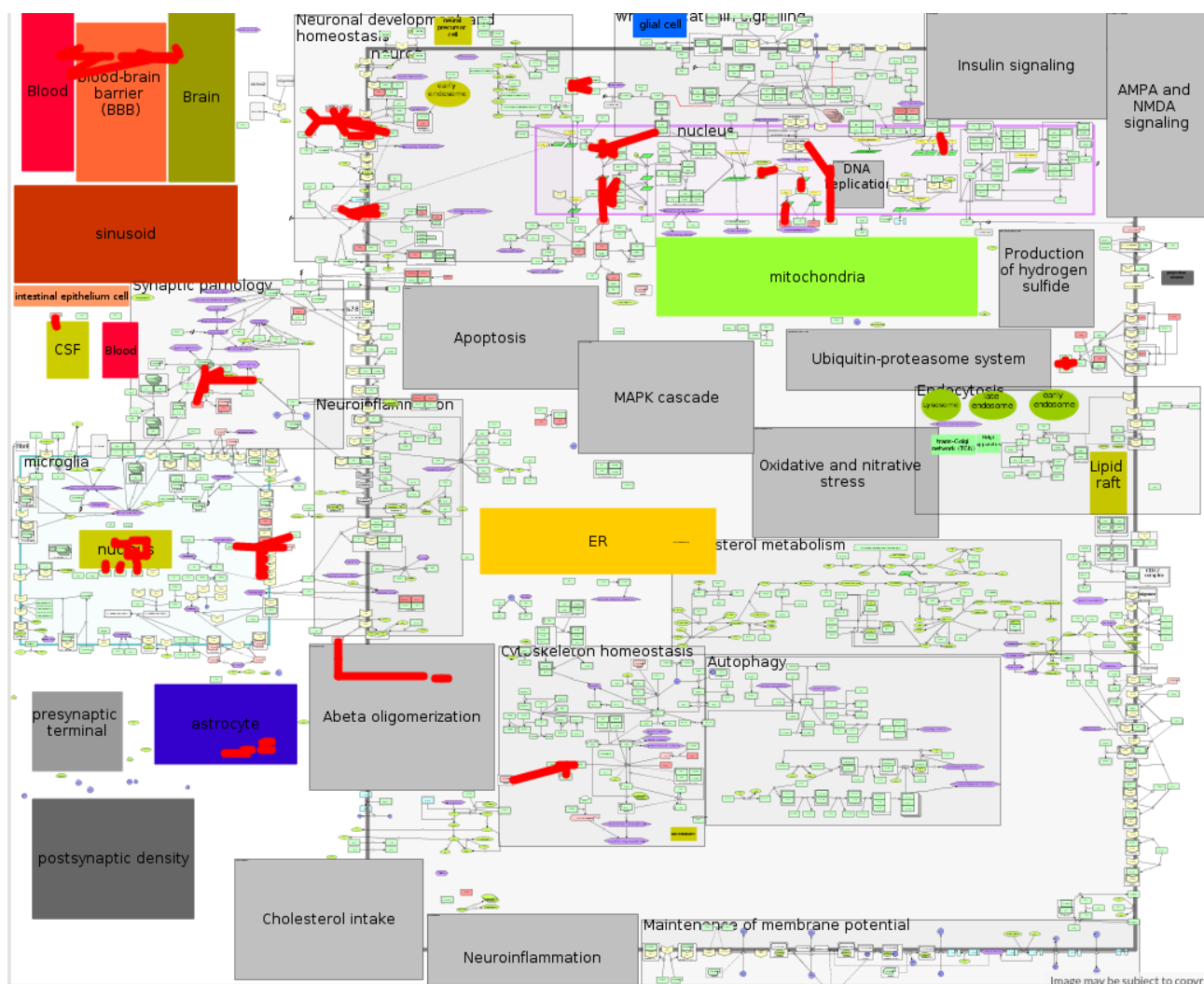


Figure 4.8: Reactions from APP model highlighted on AlzPathway

However, the majority of the reactions were still involved with APP, Abeta peptides and *BACE*. Also detected was the activation of *GSK3B* and transport of inflammation markers *TNF- α* , *IL6*, *IL1B*.

Figure 4.9, shows reactions found similar in APP and AlzPathway. *BACE* (*BACE1*) is reported to cut *APP* to generate the N terminus of A β producing a C-terminal fragment called C99 [Tanzi and Bertram, 2005]. *BACE* is also a known drug target for the therapeutic inhibition of A β production in AD [Vassar et al., 2009].

In AD, *GSK3 β* serves as a functional link between Amyloid β and Tau pathology [Llorens-Marín et al., 2014](also discussed in section 3.3.3 and 3.3.7). This was also detected as common in the comparison between APP and Alzpathway map. Shown

in Figure 4.10, activated *GSK3 β* , modulates the phosphorylation of *MAPT*, which further downstream results in production of Neurofibrillary Tangles.

Although in recent decades, Amyloid β was considered the primary hallmark of AD pathology, the present consensus is that the disease has a multifactorial origin [Llorens-Marítin et al., 2014, Medina et al., 2017, Gong et al., 2018]. Currently, several reports support neuroinflammation as a significant contributor to several neurodegenerative pathogenesis including Alzheimer’s Disease [Hong et al., 2016] (also refer section 3.3.6).

GSK3 β is also known to play regulatory role in the inflammatory response [Sudduth et al., 2013, Llorens-Marítin et al., 2014]. Reactions involving inflammatory markers *IL6*, *IL1B* and *TNF α* were detected in the astrocyte and microglia in both APP and AlzPathway (See Figures 4.11 and 4.12, also discussed in section 3.3.6)

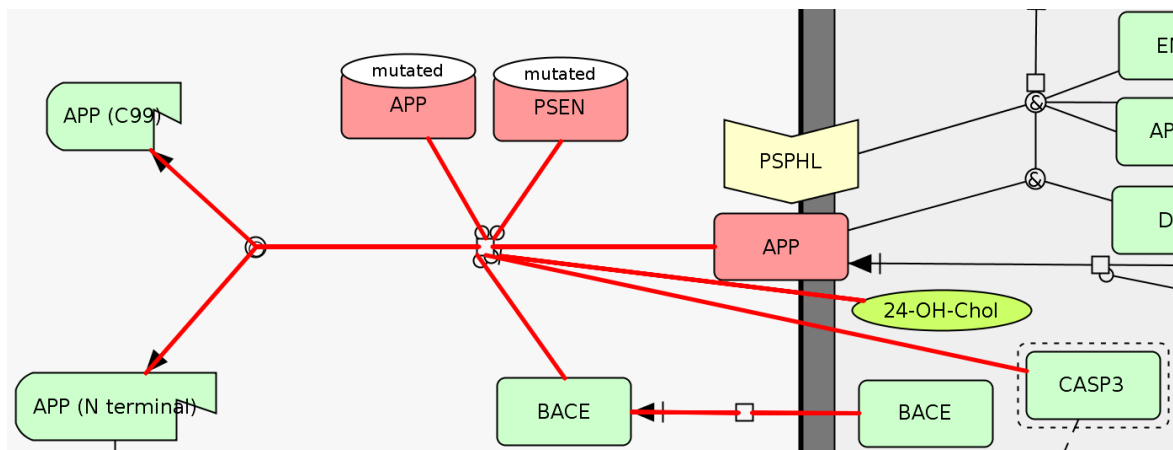


Figure 4.9: *BACE* and *APP* activity in Alzheimer’s Disease

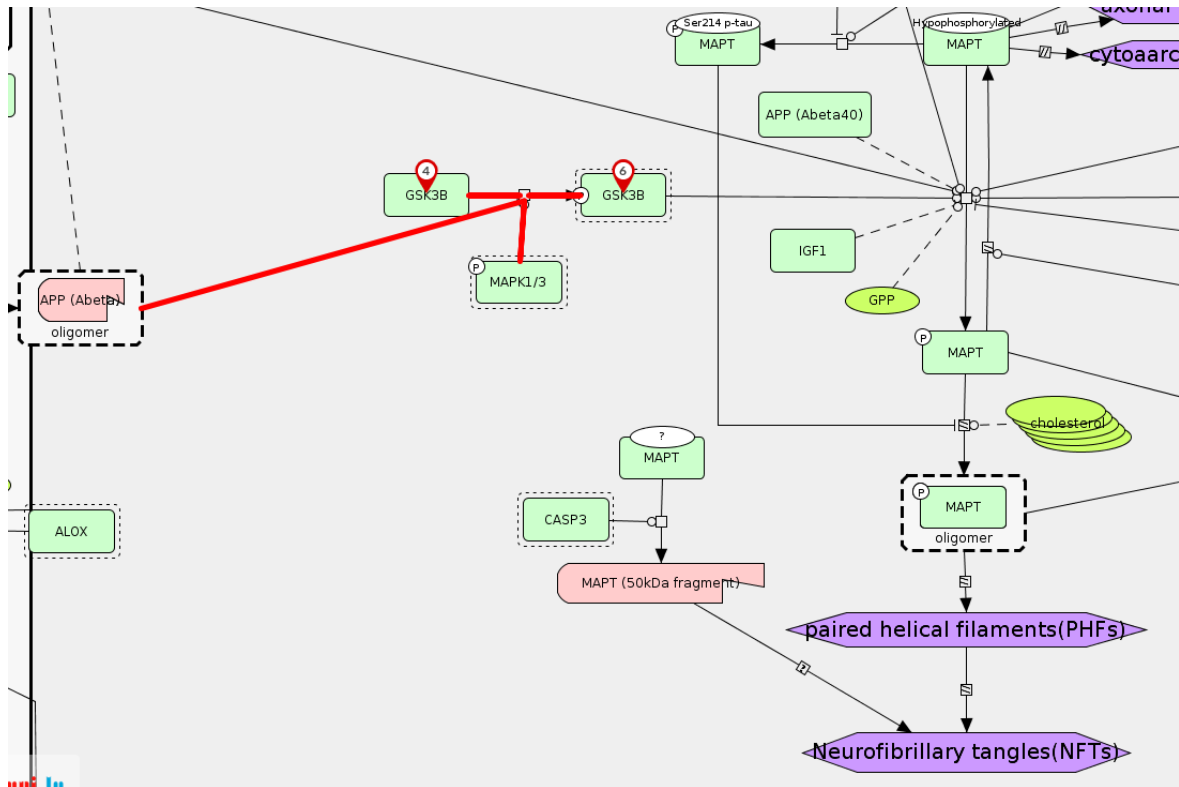


Figure 4.10: *GSK3β* as a functional link between Amyloid β and Tau pathology

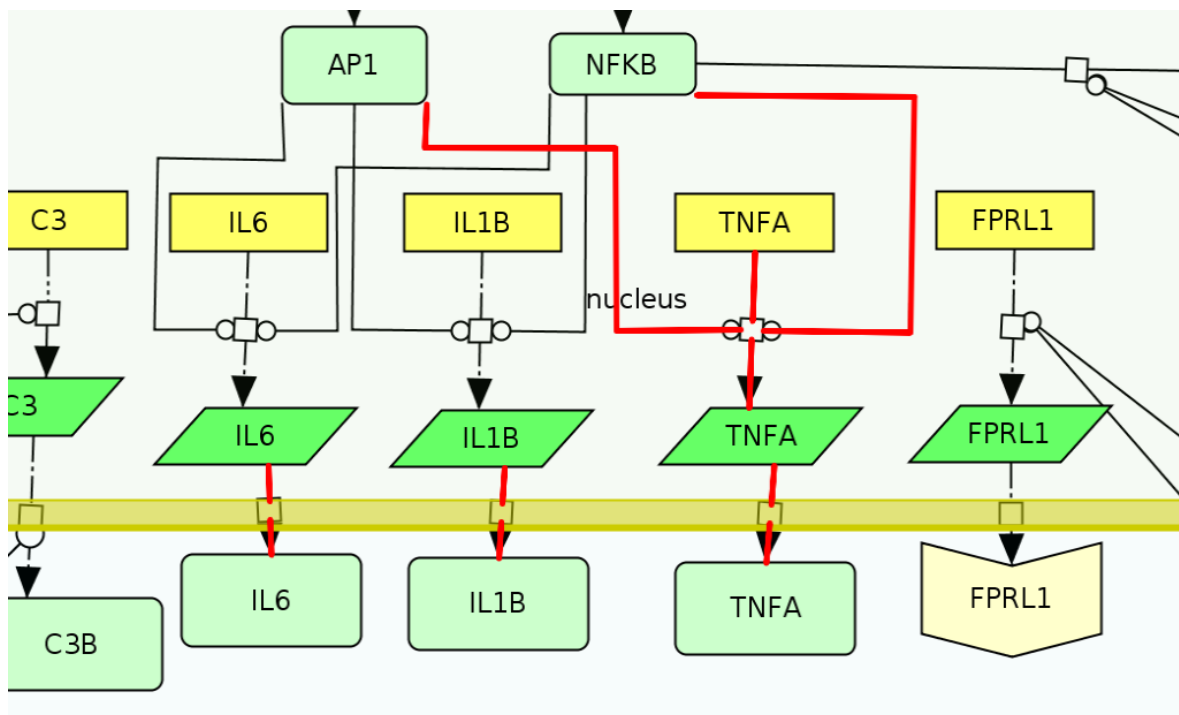


Figure 4.11: *IL6*, *IL1B* and *TNF α* transported to microglia

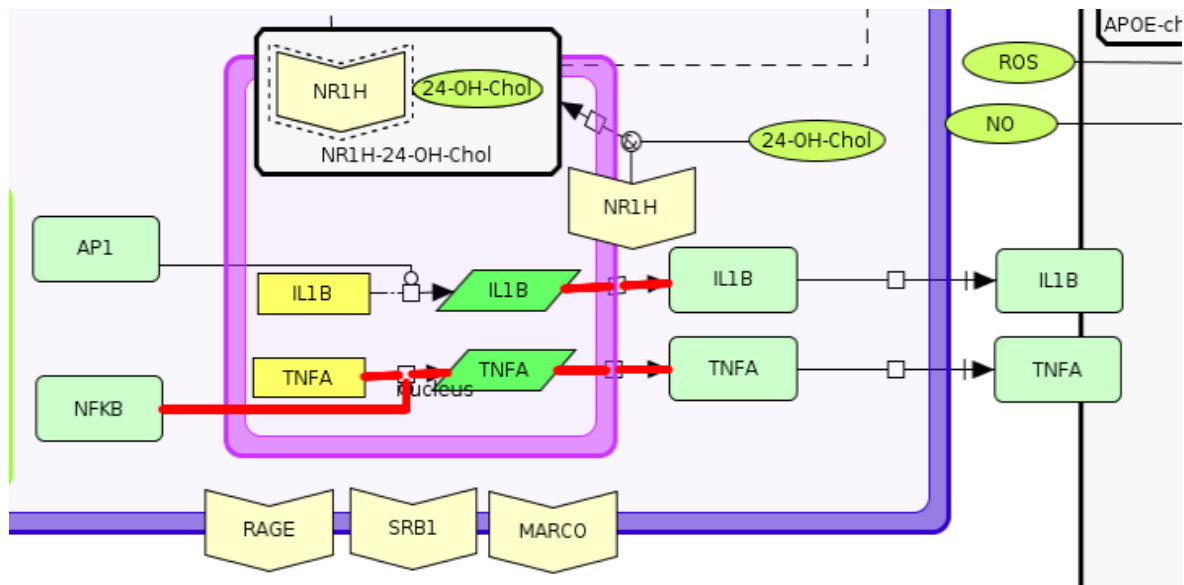


Figure 4.12: *IL6* and *IL1B* transported to astrocyte

4.5 Summary

The converter extracted the nodes and reactions from the OpenBEL model. This was then converted to the CellDesigner format for layout. The converted model were divided into five smaller maps based on cell type or tissue context. The converted maps were then used to compare against the AlzPathway. Although many elements were identified as similar between the maps, due to the differences in reaction types, only five reactions were identified as similar. It was evident that identified reactions were involved in the Amyloid β pathology. Next, we performed a comparison not strict on the reaction type, but considering the reactants and products involved in the reaction. 31 reactions were identified as similar. While several additional reaction involving APP or $A\beta$ were identified, similarities in inflammation and Tau pathology were also identified. However, as discussed earlier both maps have several elements as phenotypes, and several of them were not annotated in AlzPathway. Although unannotated genes or proteins are handled by the comparison, the current version of does not handle unannotated phenotypes. Overall, we demonstrate that annotations and modelling standards are essential for interpretability of the models. Moreover,

these standards can facilitate re-use and easier interoperability of different modelling formats for further investigation.

Chapter 5

Discussion

Data heterogeneity is one of the biggest challenge in data integration . This particularly dominant in data driven domains like systems biology and translational medicine since many researchers with different areas of expertise work together. The difficulties in data integration have only increased with the advent of high throughput technologies. Moreover, interdisciplinary research brings together different experts such as clinicians, biologists, bioinformaticians, and software developers together. Therefore, it is particularly important for these diverse groups to use, share and exchange their data and results. This is only possible if they speak the "same language".

Today, an increasing amount of data is available in databases that are maintained by different organizations for different purposes and therefore are often designed independent of each other. The increase in amount of data does not necessarily signify increase in the amount of knowledge. Data and resources are useful when they are understandable and interpretable [Panahiazar et al., 2014, Wang et al., 2018]. Applications can then integrate, search and extract information from interpretable data to support clinical decision making. Due to the volume and complexity of data available today, traditional methods of analysing such a large amount of data are not feasible [Raghupathi and Raghupathi, 2014].

With increasing power of computers resources we can now look for "infor-

mation” and ”relationships” that were not obvious. In 2013, healthcare expense of the United States was estimated at 17.6 percent of GDP (nearly \$600 billion) [Groves et al., 2013]. Despite the significant funding spent on healthcare, the main challenge of interoperability still remains [Kruse et al., 2016]. In 2011, dirty data was estimated to cost the US healthcare industry over \$300 billion every year and 60% of estimated time spent on cleaning and organising data [Redman, 2016, Tibbetts, 2011]. Over the last decades, the failure to organise and standardise the rapidly generated data makes it an increasingly costly effort to make use of this data today [Attwood et al., 2009]. In order to use and reuse the data, its storage and communication should be in a structured and standardized format.

5.1 Integrating Heterogeneous Data

Research approaches today rely heavily on information available through public databases. These datasets are often inconsistent, not standardized, or properly annotated. Moreover, the quality of the data is also uncertain [Wolstencroft et al., 2015, Comber et al., 2006]. Despite several normalisation efforts and initiatives, data standardization still remains an open issue. There is a need to define standard formats for every data type. Several domains already have such efforts which has been at least partially successful (e.g. MIAME). However, the metadata, i.e. information describing the provenance and structure of the data, design of experiment, is still neglected in many cases and are not comprehensive enough to support large scale data integration approaches [Bagewadi et al., 2015].

Chapter 2 details, the curation, harmonisation and integration of publicly available studies from GEO and Array Express and datasets generated within the AETIONOMY consortium. Data access, acquisition, curation and integration is tracked through a study request system to address the challenge of ensuring smooth and efficient entry of datasets into the AETIONOMY knowledgebase. Each dataset undergoes a curation and harmonization process. The standard format files are generated for each curated dataset. Following the ETL process, the datasets are loaded in transSMART

to enable further visual and exploratory analysis. Section 2.6, discuss how these harmonised and integrated datasets support hypothesis generation in AD and PD. While datasets from public repositories such as GEO and Array Express were standardised to large extent, clinical data from collaborators was far more challenging.

When it comes to clinical data collection, data harmonisation and standardisation is not a trivial task [Hudson et al., 2018]. Standards harmonise meaning across different studies and even sites within studies. This will enable individual elements to be aggregated into a larger picture. Often data collection involves more than one person and is carried out over a considerable period of time, without guidelines and agreed upon standards, the process will vary from person to person and over time [Dickersin and Mayo-Wilson, 2018, Leroux et al., 2017]. As discussed in Chapter 2, this variance in data representation and sharing is the biggest challenge to the integration process. Representations varying between study centre and often within the same centre. Therefore, standards will ensure the uniformity and additionally facilitate the compatibility between different systems. Collaborating with the CDISC, the Consortium for Prevention of Alzheimer’s Disease (CPAD, <https://c-path.org/programs/cpad/>) mapped the data from nine different organizations to create an openly available database containing individual records of 6,500 Alzheimer patients from 24 clinical studies [Neville et al., 2015].

Traditionally, clinical research studies relied on collecting data with case report forms. These were subsequently entered into a database by a double data entry to generate electronic records. This method is time-consuming and error-prone. Today, electronic case report forms (eCRFs) and electronic data capture (EDC) solutions are available to reduce the duration of data capture and increase accuracy. [Rorie et al., 2017, Walther et al., 2011]

Creating and revising the EDC and accepted set of arguments along with the different target groups is crucial to accommodate the requirements and potential use cases of the data generated. Minimising free text fields for data input or provide free text fields in addition to mandatory fields with select or multiple select variables will accommodate the needs of both data collector and processor. For instance, although a

data processor would prefer strict data representations to reduce ambiguity and errors that can be caused due to uncontrolled data entry, a medical researcher or data collector may find it difficult to restrict certain inputs to limited set of values. Therefore, it is particularly important to consider the needs of both the data collector and processor.

Following standards in data collection and data representation at early stages of the data collection will ensure harmonised data capture and reduce errors and ambiguity that may arise if there are no strict rules [Bellary et al., 2014, Cowie et al., 2017]. Therefore, it is necessary to bring together several groups of users when creating and designing the study. Data integration should rely on bioinformaticians and software engineers, but it also needs to be driven by the people involved directly or indirectly with the data i.e. research communities, clinicians, informaticians and analysts.

5.2 Comparison and Conversion of Maps

Comparing of disease maps is a crucial element in biomedical research. Disease maps integrate current knowledge about disease mechanisms in a context of hierarchical organisation representing the different layers of biological complexity. Disease maps help to visually represent extensive knowledge about a disease integrated in a single resource. The usefulness of disease maps, largely depend on their quality. As discussed in Chapter 3, the results of comparison between AlzPathway and PD Map emphasise the need for curation standards. For instance, several entities in the AlzPathway were annotated by two Uniprot identifier, leading to ambiguity in the identity of the elements. However, many elements were identified by extending the model annotation to additional namespaces and detecting unannotated genes and proteins. This improved interoperability between two maps and facilitated their comparison. Initiatives like COMBINE and FAIRDOM coordinate and promote the adoption of standards and ontologies in disease modelling through the experience of the community [Wolstencroft et al., 2017, Waltemath and Wolkenhauer, 2016b, Stanford et al., 2015]. Such initiatives

Although there has been considerable efforts in ensuring model sharing and reusability in disease maps using widely used languages like SBML and BioPAX, OpenBEL still remains a challenge [Hoyt et al., 2017]. It is important for languages like BEL to adopt and promote standards and ensure best practices in modelling particularly because it has been adopted by communities for crowd sourcing challenges and community built maps. Moreover, from Chapter 4 we see that while BEL is very flexible and allows to generate large networks, the quality and consistency of the network may vary depending on the curator. Since, there are no strict rules for the adoption of standards, over time the curator could potentially capture the same knowledge in different representations. This could also rise if more than one curator is involved in adding knowledge to the model, since they bring different perspective of modelling the disease. While OpenBEL suggests best practices to follow, it does not impose any strict consistency checks to ensure them. Moreover, OpenBEL is a relatively new and evolving language, hence it could be potentially useful to coordinate modelling and annotation standards in OpenBEL to ensure better model and data exchange and reuse in the community.

While standards in modelling may differ with different formats, annotation standards could be unified. However, development model annotation standards and advocating the benefits is not a trivial task. The absence of such annotation standards, makes linking data and knowledge from different model and interoperability of models a difficult task. Although it is understandable that making different modelling languages interoperable requires considerable effort, a unique format for representation of annotation could facilitate the process. Such an approach could address the challenges discussed in Chapter 4 arising due to differences in element annotation in OpenBEL and CellDesigner. Currently, no such standard protocols for model annotation exists in biological modelling community [Neal et al., 2018]. [Neal et al., 2018] propose an interesting approach to address this issue by storing annotations in a separate file, thereby the possibility to harmonise annotation standards across different formats. Recent works, like the SemanticSBML [Liebermeister et al., 2009] allow merging two SBML models. However, a computer software can only support, but not replace the modeller in building biochemically meaningful models, because it cannot handle the

assumptions and intentions on which all models are based.

Models may naturally differ in assumptions and intentions when built by different researchers. Therefore comparison or merging them can be tedious process. However, the standards both in modelling and annotation can reduce ambiguity and bridge this gap. A major challenge in extracting information from publications is the identification of entities within the article [Mons, 2005]. The usage of unique identifiers and standards in the form of controlled vocabularies and ontologies is essential for a unambiguously identifying and annotated entities. As a solution, journals and databases should not only encourage but also mandate the use of complete, standardized and structured data in their submissions.

Using integrative systems biology approaches, we can leverage the existing knowledge and large-scale data to add to our limited understanding of unknown factors and disease mechanisms [Greene et al., 2011]. Hypotheses generated from these approaches can support clinical decision making and targeted approaches in a cost-effective manner [Castaneda et al., 2015, Auffray et al., 2016, Wang et al., 2018]. Although, integrative approaches are limited by the semantic disparity between components standardising data and results will help us achieving this goal. However, it is not straightforward, but it is not impossible. Several initiatives and consortia advocating the FAIR principle is a step forward in this direction [Sansone et al., 2018].

Chapter 6

Summary and Outlook

6.1 Summary

In the last year alone, 88 new biomolecular-related databases were listed in the latest Nucleic Acids Research database issue [Rigden and Fernández, 2018]. Researchers require tools to identify relevant information in the maze of biological data. Several systems have been developed to address this need and help scientists work with omics data, e.g. Gene Expression Omnibus, Array Express, etc,. However, omics data have to be analysed together with clinical data to be useful for translational research. We use tranSMART, a translational medicine platform to curated, harmonise and integrate publicly available datasets and datasets from the AETIONOMY project. These datasets were then used to support hypotheses generation, demonstrated with examples.

Disease maps support hypothesis generation by providing context to the disease mechanism. The concept of shared mechanisms and underlying co-morbidity is common in complex disease like Parkinson's and Alzheimer's. This can be investigated by comparing disease maps. To this extent, comparison of CellDesigner maps was implemented. The comparison takes into account the annotation, translation of namespaces, and localisation of the entities and reactions. This was demonstrated us-

ing AlzPathway and PD Map. The elements and reactions found in both maps were highlighted on the PD Map, detecting similarities in both disease mechanisms, but potentially triggered by different upstream events. Comparison of maps could therefore support identification of disease specific drug targets and support clinical decision making.

Several disease modelling formats are available today. To enable comparison of disease models, they should first be interoperable. In the scope of this project, we implemented a converter from OpenBEL to CellDesigner. The converter takes into account mapping of the representation and tries to extract unique identifiers for all the entities and the context of the reaction. As a use case, the APP BEL model was converted to the CellDesigner format. The converted APP map was then used to compare against the PD Map and AlzPathway. Standardising semantic annotations in models eliminates the bottlenecks and helps researchers to easily locate models, automate and translate between modelling formats. This supports the integration of biological knowledge encoded in different models and resources.

Overall, we demonstrate how harmonised and curated data makes heterogeneous datasets and formats interoperable, bridging the gap between data and knowledge.

6.2 Outlook

Although we have successfully integrated several heterogeneous sources and making them interoperable, the process could be further stream lined by reducing manual efforts if the input data formats are harmonised. Furthermore, to maximize data sharing, the use standards for data collection and modelling should be mandatory for funding support. A prerequisite for data-driven analytics is a data sharing culture. To address this issue, we require significant efforts in adopting standards at all stages of healthcare data life cycle. This will significantly improve the quality of data and the accuracy of analysis and prediction. CDISC being widely accepted by the several

consortia and funding agencies is a big step forward to facilitate efficient data sharing and interoperability in the clinical data domain.

In addition to comparison of maps, visualising the results are of great importance. Additional scoring mechanism to the matches and intuitive visualisation methods need to be developed for e.g. a colouring scheme, which signifies the confidence on the match. Currently, we have been successful in highlighting the identified similarities. In addition to highlighting the similarities, more intuitive would be to superimpose both the compared maps and visualising both similarities and differences.

The conversion from OpenBEL to CellDesigner loses some information, but this is expected due to the difference in the objectives of the languages. However, there are many tasks that can improve this converter, but this largely depends on the different models. For instance, in the APP model SNP were encoded as genes with the dbSNP identifiers, the current version of the converter ignores these entities. This could be converted as gene with the corresponding HGNC identifier and additional annotation of the mutation. Expanding to additional namespaces, could improve entity recognition for unannotated elements.

One of the challenges in front of the scientific community today is to change how knowledge is organized and communicated. Several of the knowledge resources overlap in content [Perez-Riverol et al., 2018, Masseroli et al., 2014, Williams et al., 2012]. New resources are being created in parallel, whether they should be a novel resource or could be integrate into an existing resource is debatable. Moreover, there is no broad consensus about which resources should be used for an annotation in an ideal scenario. Therefore, the same concept could potentially be represented in two different scenarios might be annotated by different knowledge resources. This adds additional efforts to the community to compare and compose models or integrate resources in an automated fashion, as well as convert between standard formats. In an ideal case, the content of the models should be annotated using the same set of reference terms and qualifiers. However, the choice of knowledge resources for annotation may vary from group to group. Therefore, it cannot be defined by strict rules. On the other hand, making these specifications publicly available with along with the model will make the

model much more interoperable and re-usable.

Data is an essential part of research today, therefore it is also necessary to maintain standards for data annotation. With initiatives like the FAIR and COMBINE, linking annotations in models and data sources is foreseeable in the future. In this era of big data, data-driven analytics has the potential to transform the technologies used by healthcare providers by gaining insight from clinical and public data repositories and supporting decision making. The rapid, widespread implementation and use of big data analytics in the healthcare industry is challenged by the volume and heterogeneity of relevant data. Harmonised and standardised data and representation can accelerate the development of such analytical tools to support clinical decision making.

Appendix A

Appendix A : Supplementary Materials

A.1 Alzheimer's and Parkinson's Disease datasets integrated in AETIONOMY

Partner	Name of Cohort	Subjects	Biospecimen	Type	Total Received	Comments	Data access after AETIONOMY (+5 years): Public, closed, deleted, data access committee
UL	ADNI	ADNI1	Plasma CSF	Demographic Clinical Neuropsychological	1386	Completed	Closed
	ADNIMERGE	ADNIMERGE: Alzheimer's Disease Neuroimaging Initiative. R package version 0.0.1.	Plasma CSF	Demographic Clinical Neuropsychological	1779	Completed	Public
FhG SCAI	AD Public studies	4 GEO Studies 39 AE Studies		Demographic+Clinical+mRNA/miRNA Expression	43 Studies	Completed	Public
EMC	Imaging ADNI	BIGR Connectome Freesurfer		BIGR Connectome Freesurfer	57	Completed	
IDIBAPS	Screening	Control 9 AD Moderate Cognitive Decline 6 Prodormal AD 2	CSF	Demographic Clinical Neuropsychological MRNA/miRNA Expression	17	Completed	Restricted (Data Access committee)
	Validation	Control 20 AD Moderate Cognitive Decline 11 Prodormal AD 4	CSF	Demographic Clinical Neuropsychological MRNA/miRNA Expression	34	Completed	Restricted (Data Access committee)
	AD Dementia Stage	Control 68 AD Moderate Cognitive Decline 26 Preclinical AD 23 Prodormal AD 39 SNAPS 8	CSF	Demographic Clinical Neuropsychological	164	Completed	Restricted (Data Access committee)
UKB (AD)	AD Inflammation	Consecutive Preclinical AD 30	CSF	Demographic Clinical Neuropsychological	399	Completed	Restricted (Data Access committee)
	AD Cytokine	AD 44 MCI 34 Control 17	CSF	Demographic Clinical Neuropsychological	95	Completed	Closed

Partner	Name of Cohort	Subjects	Biospecimen	Type	Total Received	Comments	Data access during AETIONOMY: Public, closed, deleted, data access committee	Data access after AETIONOMY (+5 years): Public, closed, deleted, data access committee
UL	PPMI	eTRIKS (in Kind)		Demographic Clinical Neuropsychological	813	Completed	Closed	Closed
	PD Public studies	eTRIKS (in Kind)		Demographic Clinical mRNA/miRNA Expression	16 Studies	Completed	Public	Public
FhG SCAI	AD Public studies	4 GEO Studies 39 AE Studies		Demographic Clinical MRNA/miRNA Expression	43 Studies	Completed	Public	Public
ICM	GenoPark	LRRK2 Relative (1) Parkin Relative(4) Parkinson Disease Asymptomatic Carrier(1) Genetic(27) MSA (10) PSP (11)	Blood	Demographic Clinical Neuropsychological	360	Completed	Data Access Committee	Closed
	NGC/ PD Repository	At Risk 1082 Control 578 Diseased 3331 Unknown 20		Demographic Clinical Neuropsychological	5011	Completed	Data Access Committee	Closed
	DIGPD	Consecutive idiopathic PD	DNA Plasma	Demographic Clinical Neuropsychological Genetic Data Results	416	Completed	Data Access Committee	Closed
	PD Transcriptomic data	Control 10 LRRK2 12 Parkin 14	Fibroblasts	Demographic mRNA/miRNA Expression	36	Completed.	Data Access Committee	Closed
UKB (PD)	Epigenetics Data PD (CpG data)	Control 521 PD 536	DNA (Blood)	Demographic Clinical Neuropsychological	1057	Completed.	Data Access Committee	Public
UCB	Kings College London	Control 161 Parkinson's Disease 40	DNA CSF	Demographic		Completed	Data Access Committee	Data Access Committee
	D13B Tubingen plasma and CSF	Proteomics and methylation studies	CSF Plasma			Completed	Data Access Committee	Data Access Committee
	EFPIA PD Clinical Studies	14 STUDIES				Curation	Data Access Committee	
BI	BI001	76 IDS		Aggregate Datasets		Completed	Data Access Committee	
	BI002	7 IDS		Aggregate Datasets		Completed	Data Access Committee	
	BI003	6 IDS		Aggregate Datasets		Completed.	Data Access Committee	
KI		PD 100 Control 50	DNA Plasma CSF					
	AETIONOMY	PD 240 Genetics 65 Control 90 IRBD 70						
	ICEBERG	PD 53 Control 10 IRBD 17						

A.2 AlzPathway overlayed on PD Map

Summary of performance with different annotators

	Existing Annotations	HGNC annotator	All annotators
Elements (Annotation)	48	649	762
Elements (Annotation+Compartment)	25	430	528
Perfect Element (Annotation+Compartment+Complex)	25	291	325
Reactions	0	31	40
Perfect Reaction (Reactants+Products+Modifiers)	0	5	6

Table A.1: Summary of performance with different annotators

Elements from AlzPathway found in the PD Map

name	lineWidth	Color
ATG10	10	#FF0000
IL1B	10	#FF0000
SNCA	10	#FF0000
VPS35	10	#FF0000
BCL2	10	#FF0000
ATG13	10	#FF0000
BAX	10	#FF0000
PPARG	10	#FF0000
GSK3B	10	#FF0000
TXN	10	#FF0000

XBP1	10	#FF0000
NFKB1	10	#FF0000
BAD	10	#FF0000
TLR2	10	#FF0000
multivesicular body	10	#FF0000
NGFR	10	#FF0000
PPP2CA	10	#FF0000
ULK1	10	#FF0000
FAS	10	#FF0000
P2RX7	10	#FF0000
CXCL1	10	#FF0000
CYCS	10	#FF0000
NFKBIA	10	#FF0000
glutathione disulfide	10	#FF0000
EIF4EBP1	10	#FF0000
CAMK2B	10	#FF0000
TNF	10	#FF0000
ATG12	10	#FF0000
CTNNB1	10	#FF0000
RELA	10	#FF0000
mitochondrion	10	#FF0000
HSPA5	10	#FF0000
CASP8 (p10)	10	#FF0000
CASP9	10	#FF0000
PRKAA2	10	#FF0000
CAT	10	#FF0000
ubiquitin	10	#FF0000

NGF	10	#FF0000
MAPK8	10	#FF0000
PIP2	10	#FF0000
astrocyte	10	#FF0000
APC	10	#FF0000
TNFRSF1A	10	#FF0000
STXBP1	10	#FF0000
CCL2	10	#FF0000
TLR3	10	#FF0000
ATG3	10	#FF0000
FYN	10	#FF0000
peroxynitrite	10	#FF0000
RPS6KA1	10	#FF0000
BCL2L1	10	#FF0000
APAF1	10	#FF0000
AKT1	10	#FF0000
ATG16L1	10	#FF0000
AKT1S1	10	#FF0000
RAC1	10	#FF0000
CASP3	10	#FF0000
AGER	10	#FF0000
VPS26A	10	#FF0000
RYR3	10	#FF0000
FADD	10	#FF0000
ATG7	10	#FF0000
JUN	10	#FF0000
PDP1	10	#FF0000

MAPK1	10	#FF0000
RHOA	10	#FF0000
EIF2AK3	10	#FF0000
GSR	10	#FF0000
PTGS2	10	#FF0000
WNT1	10	#FF0000
NEDD8	10	#FF0000
CD36	10	#FF0000
EIF2A	10	#FF0000
ERN1	10	#FF0000
early endosome	10	#FF0000
DLG4	10	#FF0000
CSNK1A1	10	#FF0000
CASP6	10	#FF0000
IKBKB	10	#FF0000
CASP8 (p18)	10	#FF0000
CDK5	10	#FF0000
GAPDH	10	#FF0000
TRAF2	10	#FF0000
Golgi	10	#FF0000
RHEB	10	#FF0000
late endosome	10	#FF0000
HSPD1	10	#FF0000
IGF1R	10	#FF0000
IRS1	10	#FF0000
IL4	10	#FF0000
BID	10	#FF0000

BACE1	10	#FF0000
PPARA	10	#FF0000
NTRK2	10	#FF0000
HMGB1	10	#FF0000
MAPK12	10	#FF0000
trans-Golgi network	10	#FF0000
ER	10	#FF0000
BDNF	10	#FF0000
MAPT	10	#FF0000
IFNG	10	#FF0000
MAP2K7	10	#FF0000
MTOR	10	#FF0000
PGK1	10	#FF0000
APP	10	#FF0000
TRADD	10	#FF0000
KCNIP3	10	#FF0000
IL6	10	#FF0000
TP53	10	#FF0000
VPS29	10	#FF0000
BTRC	10	#FF0000
TH	10	#FF0000
WISP1	10	#FF0000
NLRP3	10	#FF0000
GTP	10	#FF0000
MAPK14	10	#FF0000
FOXO1	10	#FF0000
TLR4	10	#FF0000

MMP3	10	#FF0000
advanced glycation end-products	10	#FF0000
LRP6	10	#FF0000
GNAQ	10	#FF0000
NTRK1	10	#FF0000
C3	10	#FF0000
SLC2A1	10	#FF0000
exosome	10	#FF0000
PTEN	10	#FF0000
CASP7	10	#FF0000
MAPK cascade	10	#FF0000
mitochondrial matrix	10	#FF0000
VCAM1	10	#FF0000
ATG5	10	#FF0000
CASP2	10	#FF0000
MAPK pathway	10	#FF0000
CASP8	10	#FF0000
MAP2K3	10	#FF0000
JAK2	10	#FF0000
PRKCI	10	#FF0000
CHRM1	10	#FF0000
LAG3	10	#FF0000
STK11	10	#FF0000
NTF3	10	#FF0000
MAPK13	10	#FF0000
MMP9	10	#FF0000
CD14	10	#FF0000

PDK1	10	#FF0000
MAP2K4	10	#FF0000
CDC42	10	#FF0000
PSEN1	10	#FF0000
resting microglia	10	#FF0000
ATF6	10	#FF0000
NCOR2	10	#FF0000
CAPN2	10	#FF0000
LRP5	10	#FF0000
TREM2	10	#FF0000
SKP1	10	#FF0000
S100B	10	#FF0000
GABARAPL2	10	#FF0000
CASP1	10	#FF0000
ICAM1	10	#FF0000
blood vessel	10	#FF0000
MAPK3	10	#FF0000
MAP2K6	10	#FF0000

Table A.2: Submap PD 180412 2 alzpath 8APR Element

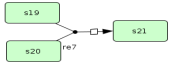
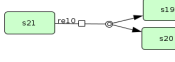

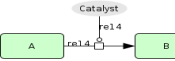
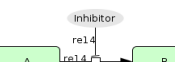
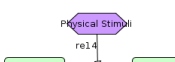



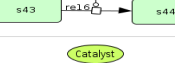






reactionIdentifier	lineWidth	Color
re629	7	#FF0000
re5103	7	#FF0000
re4885	7	#FF0000
re3346	7	#FF0000
re3352	7	#FF0000
re5314	7	#FF0000

re5188	7	#FF0000
re5153	7	#FF0000
re4341	7	#FF0000
re5044	7	#FF0000
re5336	7	#FF0000
re5221	7	#FF0000
re2858	7	#FF0000
re624	7	#FF0000
re4330	7	#FF0000
re3344	7	#FF0000
re2869	7	#FF0000
re5241	7	#FF0000
re478	7	#FF0000
re5415	7	#FF0000
re919	7	#FF0000
re2850	7	#FF0000
re5305	7	#FF0000
re5077	7	#FF0000
re4343	7	#FF0000
re505	7	#FF0000
re5342	7	#FF0000
re2865	7	#FF0000

Table A.3: Snapshot of reaction matches in PD Map and Alzpathway

A.3 Semantic Mapping: CellDesigner (SBGN)-BEL

CELLDESIGNER			BEL			
Species	Notation	Namespaces	BEL species	Notation/Function	Namespaces	BEL Example
Protein		HGNC	Protein	proteinAbundance, p	HGNC, SPAC, MGI, RGD	proteinAbundance(HGNC:AKT1)
Receptor		HGNC	Receptor	proteinAbundance, p	HGNC, SPAC, MGI, RGD	
Ion Channel		HGNC	Ion Channel	proteinAbundance, p	HGNC, SPAC, MGI, RGD	
Truncated Protein		HGNC	Peptides	abundance, a	no namespace	a("Amyloid beta-Peptides")
Gene		HGNC	Gene	geneAbundance, g	HGNC, SPAC, EGID, MGI, RGD	geneAbundance(HGNC:AKT1)
RNA		HGNC	RNA	maAbundance, r	HGNC, SPAC, MGI, RGD	maAbundance(HGNC:AKT1)
Anti Sense RNA		HGNC	RNA	maAbundance, r	HGNC, SPAC, MGI, RGD	
miRNA		HGNC	miRNA	microRNAAbundance, m	HGNC, SPAC, MGI, RGD	microRNAAbundance(HGNC:MIR21)
Phenotype		GO:Biological Process	Disease Bioprocesses	pathology, path biologicalProcess, bp	MESH GO, MESHPP	
Ion		CHEBI	Ion	abundance, a	CHEBI, CHEBIID, SCHEM	
Simple Molecule		CHEBI	Simple Molecule	abundance, a	CHEBI, CHEBIID, SCHEM	
Unknown		GO, CHEBI, Interpro	Unknown	abundance, a	no namespace	
Drug		CHEBI, PubChem	Drug	abundance, a	CHEBI, CHEBIID, SCHEM, or custom namespaces like DrugBank	
Hypothetical Protein		Protein family: Interpro	ProteinFamily	proteinAbundance, p	PFR, PFH, PFM	
Complex		GO: Cellular Component	Complex	complex()		
Compartment		GO:Cellular components, MeSH	Cellular Component	abundance, a	GOCCACC, GOCCTERM, MESHCL	
Reaction	Notation	Additional notes	BEL reaction	Notation		
BASIC						
State Transition		Reaction+Translocation, Refer supplementary document				
Known Transition Omitted						
Unknown Transition						
Transport			Translocation	tlloc		
Transcription			Transcription	transcription(A)		
Translation			Translation	translation(A)		
Add reactant			reactants()			
Add product			products()			

Heterodimer association					
Dissociation					
Truncation				peptidaseActivity()	pep()
Reaction	Notation	Additional notes	BEL reaction	Notation	
MODIFIERS					
Catalysis			Catalysis	cat(example)	
Inhibition			directlyIncreases()	=	
Physical Stimulation			directlyIncreases()	=>	
Modulation			increases()	->	
Trigger			directlyIncreases()	=>	
Unknown Catalysis		extended by CellDesigner	increases()	->	
Unknown Inhibition		extended by CellDesigner	decreases()	-	
Reaction	Notation	Additional notes	BEL reaction	Notation	
REDUCED					
Positive Influence			directlyIncreases()	=>	
Unknown positive influence			increases()	->	
Negative Influence			directlyDecreases()	=	
Unknown negative influence			decreases()	-	
Reduced Trigger			increases()	->	
Unknown Reduced Trigger			increases()	->	

A.4 Nodes extracted from APP BEL model

Table A.4: Example of nodes extracted from APP BEL model

IDENTIFIER	TYPE	NAME	URN	ANNOTATIONS	CELL	TISSUE	COMPARTMENT	ORGANISM
s1	PROTEIN	GCG	urn:miriam:hgnc.symbol:GCG					
s2	PHENOTYPE	Diabetes Mellitus, Type 2	urn:miriam:mesh.2012:D003924					
s3	PROTEIN	MAPT	urn:miriam:hgnc.symbol:MAPT					
s4	PROTEIN	Amyloid beta-peptides	urn:miriam:ensemble:ENSP00000453144					
s12	PROTEIN	PSENEN	urn:miriam:hgnc.symbol:PSENEN					
s13	PHENOTYPE	Polymorphism, Single Nucleotide	urn:miriam:mesh.2012:D020641					
s14	GENE	PSENEN	urn:miriam:hgnc.symbol:PSENEN	urn:miriam:obo.go:GO:0006355				
s15	COMPLEX	gamma Secretase Complex	urn:miriam:obo.go:GO:0070765					
s25	PHENOTYPE	acetylcholine secretion	urn:miriam:obo.go:GO:0061526					
s27	PHENOTYPE	learning or memory	urn:miriam:obo.go:GO:0007611					
s29	PROTEIN	GALR1	urn:miriam:hgnc.symbol:GALR1					
s30	PROTEIN	GALR2	urn:miriam:hgnc.symbol:GALR2					
s31	PHENOTYPE	neuroprotection						
s32	PROTEIN	GAL	urn:miriam:hgnc.symbol:GAL			Cerebrospinal Fluid		
s33	PROTEIN	MAPT	urn:miriam:hgnc.symbol:MAPT			Cerebrospinal Fluid		
s34	UNKNOWN	-MSH				Cerebrospinal Fluid		
s35	PROTEIN	DHCR24	urn:miriam:hgnc.symbol:DHCR24					
s37	PHENOTYPE	apoptotic process	urn:miriam:obo.go:GO:0006915					
s38	PROTEIN	DHCR24	urn:miriam:hgnc.symbol:DHCR24		Neuroblastoma cell			
s39	PHENOTYPE	Cholesterol	urn:miriam:mesh.2012:D002784		Neuroblastoma cell			
s40	PROTEIN	Amyloid beta-peptides	urn:miriam:ensemble:ENSP00000453144		Neuroblastoma cell			
s41	PROTEIN	Amyloid beta-peptides	urn:miriam:ensemble:ENSP00000453144	urn:miriam:obo.go:GO:0009056	Neuroblastoma cell			
s42	PHENOTYPE	Estrogens	urn:miriam:mesh.2012:D004967		Neuroblastoma cell			
s43	PHENOTYPE	neuroprotection			Neuroblastoma cell			
s44	PROTEIN	BLMH	urn:miriam:hgnc.symbol:BLMH		CHOAPPsw			
s45	PROTEIN	APP	urn:miriam:hgnc.symbol:APP		CHOAPPsw			

A.5 Reactions extracted from APP BEL model

Table A.5: Example of reactions extracted from APP BEL model

IDENTIFIER	TYPE	REACTANTS	MODIFIERS	PRODUCTS	MODIFIER_TYPE	REFERENCE	SPECIES	DISEASE	CELL	TISSUE	COMPARTMENT
r1	UNKNOWN_NEGATIVE_INFLUENCE	s1		s2		urn:nir:iam:pubmed:23603201					
r2	UNKNOWN_NEGATIVE_INFLUENCE	s1		s3		urn:nir:iam:pubmed:23603201		Alzheimer Disease			
r3	UNKNOWN_NEGATIVE_INFLUENCE	s1		s4		urn:nir:iam:pubmed:23603201		Alzheimer Disease			
r4	UNKNOWN_NEGATIVE_INFLUENCE	s5		s7		urn:nir:iam:pubmed:23603201	9006				
r5	UNKNOWN_POSITIVE_INFLUENCE	s8		s6		urn:nir:iam:pubmed:23603201	10116				
r133	STATE_TRANSITION	s185	s163	s186	UNKNOWN_INHIBITION	urn:nir:iam:pubmed:16511867	10090				Endosome
r3388	UNKNOWN_POSITIVE_INFLUENCE	s293		s295		urn:nir:iam:pubmed:22046282		Inflammation		Hippocampus	Endoplasmic Reticulum
r3389	UNKNOWN_NEGATIVE_INFLUENCE	s293		s296		urn:nir:iam:pubmed:22046282		Inflammation		Hippocampus	Endoplasmic Reticulum
r3390	UNKNOWN_POSITIVE_INFLUENCE	s293		s297		urn:nir:iam:pubmed:22046282		Inflammation		Hippocampus	Endoplasmic Reticulum
r3391	UNKNOWN_POSITIVE_INFLUENCE	s293		s298		urn:nir:iam:pubmed:22046282		Inflammation		Hippocampus	Endoplasmic Reticulum
r3392	UNKNOWN_REDUCED_MODULATION	s6		s299		urn:nir:iam:pubmed:17712163	9006	Alzheimer Disease			
r3393	UNKNOWN_REDUCED_MODULATION	s6		s2574		urn:nir:iam:pubmed:17712163	9006	Alzheimer Disease			
r3394	UNKNOWN_REDUCED_MODULATION	s6		s3300		urn:nir:iam:pubmed:17712163	9006	Alzheimer Disease			
r3395	UNKNOWN_REDUCED_MODULATION	s6		s282		urn:nir:iam:pubmed:17712163	9006	Alzheimer Disease			
r3396	UNKNOWN_REDUCED_MODULATION	s6		s3301		urn:nir:iam:pubmed:17712163	9006	Alzheimer Disease			
r3397	UNKNOWN_REDUCED_MODULATION	s6		s2652		urn:nir:iam:pubmed:17712163	9006	Alzheimer Disease			
r3398	UNKNOWN_POSITIVE_INFLUENCE	s302		s3304		urn:nir:iam:pubmed:12300529		Alzheimer Disease			Endoplasmic Reticulum
r3399	UNKNOWN_NEGATIVE_INFLUENCE	s302		s3305		urn:nir:iam:pubmed:12300529		Alzheimer Disease			Endoplasmic Reticulum
r3400	UNKNOWN_POSITIVE_INFLUENCE	s306		s31		urn:nir:iam:pubmed:23073831					
r3401	UNKNOWN_NEGATIVE_INFLUENCE	s6		s2476		urn:nir:iam:pubmed:22142155		Insulin Resistance			
r3402	UNKNOWN_NEGATIVE_INFLUENCE	s6		s2531		urn:nir:iam:pubmed:22142155		Insulin Resistance			
r3403	UNKNOWN_POSITIVE_INFLUENCE	s307		s2823		urn:nir:iam:pubmed:21158163	10116	Alzheimer Disease		Brain	
r3404	UNKNOWN_POSITIVE_INFLUENCE	s307		s2822		urn:nir:iam:pubmed:21158163	10116	Alzheimer Disease		Brain	
r3405	UNKNOWN_POSITIVE_INFLUENCE	s307		s3308		urn:nir:iam:pubmed:21158163	10116	Alzheimer Disease		Brain	

Appendix B

Publications

Featured Article

Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease

Genevera I. Allen^a, Nicola Amoroso^{b,c}, Catalina Anghel^d, Venkat Balagurusamy^e, Christopher J. Bare^f, Derek Beaton^g, Roberto Bellotti^{b,c}, David A. Bennett^h, Kevin L. Boehmeⁱ, Paul C. Boutros^{d,j,k}, Laura Caberlotto^l, Cristian Caloian^d, Frederick Campbell^a, Elias Chaibub Neto^f, Yu-Chuan Chang^m, Beibei Chenⁿ, Chien-Yu Chen^o, Ting-Ying Chien^p, Tim Clark^{q,r}, Sudeshna Das^{q,r}, Christos Davatzikos^s, Jieyao Deng^{t,u}, Donna Dillenberger^e, Richard J. B. Dobson^{v,w,ccc}, Qilin Dong^{t,u}, Jimit Doshi^s, Denise Duma^x, Rosangela Errico^y, Guray Erus^s, Evan Everett^a, David W. Fardo^{z,aa}, Stephen H. Friend^f, Holger Fröhlich^{bb}, Jessica Gan^a, Peter St George-Hyslop^{cc}, Satrajit S. Ghosh^{dd,ee}, Enrico Glaab^{ff}, Robert C. Green^{gg}, Yuanfang Guan^{hh,ii,jj}, Ming-Yi Hong^o, Chao Huang^{kk}, Jinseub Hwang^{ll}, Joseph Ibrahim^{kk}, Paolo Inglese^{mm}, Anandhi Iyappan^{oo,bb}, Qijia Jiang^a, Yuriko Katsumata^{aa}, John S. K. Kauwe^{i,*}, Arno Klein^{f,**}, Dehan Kong^{kk}, Roland Krause^{ff}, Emilie Lalonde^d, Mario Lauria^l, Eunjee Lee^{kk}, Xihui Lin^d, Zhandong Liu^a, Julie Livingstone^d, Benjamin A. Logsdon^f, Simon Lovestoneⁿⁿ, Tsung-wei Maⁿ, Ashutosh Malhotra^{oo,bb}, Lara M. Mangravite^{i,**}, Taylor J. Maxwell^{pp}, Emily Merrill^q, John Nagorski^a, Aishwarya Namasivayam^{ff}, Manjari Narayan^a, Mufassra Naz^{oo,bb}, Stephen J. Newhouse^{v,qq}, Thea C. Norman^f, Ramil N. Nurtdinov^{rr}, Yen-Jen Oyang^m, Yudi Pawitan^{ss}, Shengwen Peng^{t,u}, Mette A. Peters^{f,**}, Stephen R. Piccoloⁱ, Paurush Praveen^{l,bb}, Corrado Priami^l, Veronica Y. Sabelnykova^d, Philipp Senger^{oo}, Xia Shen^{ss,aaa,bbb}, Andrew Simmons^v, Aristeidis Sotiras^s, Gustavo Stolovitzky^{tt,e}, Sabina Tangaro^c, Andrea Tateo^b, Yi-An Tung^{uu}, Nicholas J. Tustison^{vv}, Erdem Varol^s, George Vradenburg^{ww}, Michael W. Weiner^{xx}, Guanghua Xiaoⁿ, Lei Xie^{yy}, Yang Xieⁿ, Jia Xuⁿ, Hojin Yang^{kk}, Xiaowei Zhanⁿ, Yunyun Zhouⁿ, Fan Zhu^{hh}, Hongtu Zhu^{kk}, Shanfeng Zhu^{t,u,zz}, for the Alzheimer's Disease Neuroimaging Initiative

^aDepartment of Statistics and Electrical and Computer Engineering, Rice University, Houston, TX, USA

^bDipartimento di Fisica "M. Merlin", Università degli studi di Bari "A. Moro", Bari, Italy

^cSezione di Bari, Istituto Nazionale di Fisica Nucleare, Bari, Italy

^dOntario Institute for Cancer Research, Informatics and Bio-computing Program, MaRS Centre, Toronto, ON, Canada

^eIBM Computational Biology Center, IBM Research, NY, USA

^fSage Bionetworks, Seattle, WA, USA

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

*Corresponding author. Tel.: +1 801 422 2993; Fax: +1 801 422 0004.

**Lara M Mangravite: Tel.: +1 206 667 6044; Fax: +1 206 667 2062
Mette A. Peters: Tel.: +1 206 667 2113; Fax: +1 206 667 2062
Arno Klein: Tel.: +1 917 512 5627

E-mail address: kauwe@byu.edu (J.S.K.K.), arno@binarybottle.com (A.K.), lara.mangravite@sagebase.org (L.M.M.), mette.peters@sagebase.org (M.A.P.)

<http://dx.doi.org/10.1016/j.jalz.2016.02.006>

1552-5260/© 2016 The Authors. Published by Elsevier Inc. on behalf of the Alzheimer's Association. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

- ^gSchool of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA
^hRush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA
ⁱDepartment of Biology, Brigham Young University, Provo, UT, USA
^jDepartment of Medical Biophysics, University of Toronto, Toronto, Canada
^kDepartment of Pharmacology & Toxicology, University of Toronto, Toronto, Canada
^lThe Microsoft Research, University of Trento Centre for Computational and Systems Biology (COSBI), Rovereto, Italy
^mGraduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan
ⁿQuantitative Biomedical Research Center, The University of Texas Southwestern Medical Center, Dallas, TX, USA
^oDepartment of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei, Taiwan
^pInnovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan, Taiwan
^qDepartment of Neurology, Massachusetts General Hospital, Cambridge, MA, USA
^rDepartment of Neurology, Harvard Medical School, Boston, MA, USA
^sCenter for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA, USA
^tSchool of Computer Science, Fudan University, Shanghai, Shanghai, China
^uShanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, Shanghai, China
^vNIHR Biomedical Research Centre for Mental Health, Kings College London, London, UK
^wInstitute of Psychiatry, Psychology and Neuroscience, MRC Social, Genetic and Developmental Psychiatry Centre, Kings College London, London, UK
^xDepartment of Pediatrics-Neurology, Baylor College of Medicine, Houston, TX, USA
^yUniversità degli Studi di Genova, Genova, Italy
^zSanders-Brown Center on Aging, University of Kentucky, Lexington, KY, USA
^{aa}Department of Biostatistics, University of Kentucky, Lexington, KY, USA
^{bb}Bonn-Aachen International Center for IT, University of Bonn, Bonn, Germany
^{cc}Cambridge Institute for Medical Research, University of Cambridge and University of Toronto, Cambridge, CB2, UK
^{dd}McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA
^{ee}Department of Otolaryngology, Harvard Medical School, Boston, MA, USA
^{ff}Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg
^{gg}Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Broad Institute and Harvard Medical School, Boston, MA, USA
^{hh}Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA
ⁱⁱDepartment of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA
^{jj}Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA
^{kk}Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
^{ll}Department of Computer science and Statistics, Daegu University, Gyeongsan-si, Gyeongsangbuk-do, Republic of Korea
^{mm}Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK
ⁿⁿDepartment of Psychiatry, University of Oxford, Warneford Hospital, Oxford, UK
^{oo}Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Department for Bioinformatics, Schloss Birlinghoven, Sankt Augustin, Germany
^{pp}Computational Biology Institute, The George Washington University, Ashburn, VA, USA
^{qq}Department of Biostatistics, Kings College London, London, UK
^{rr}Department of Neuroimmunology, Foundation Institut de Recerca, Hospital Universitari Vall d'Hebron, Barcelona, Spain
^{ss}Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
^{tt}Genetics and Genomics Sciences Department, Icahn School of Medicine at Mount Sinai, New York, NY, USA
^{uu}Genome and systems biology degree program, National Taiwan University, Taipei, Taiwan
^{vv}Department of Radiology and Medical Imaging, The University of Virginia, Charlottesville, VA, USA
^{ww}Global CEO Initiative on Alzheimer's disease, Washington, DC, USA
^{xx}Radiology, Medicine, Psychiatry, and Neurology, UCSF, SFVAMC, San Francisco, CA, USA
^{yy}Department of Computer Science, Hunter College, The City University of New York, New York, NY, USA
^{zz}Centre for Computational Systems Biology, Fudan University, Shanghai, China
^{aaa}Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK
^{bbb}MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
^{ccc}Farr Institute of Health Informatics Research, UCL Institute of Health Informatics, University College London, London WC1E 6BT, UK

Abstract

Identifying accurate biomarkers of cognitive decline is essential for advancing early diagnosis and prevention therapies in Alzheimer's disease. The Alzheimer's disease DREAM Challenge was designed as a computational crowdsourced project to benchmark the current state-of-the-art in predicting cognitive outcomes in Alzheimer's disease based on high dimensional, publicly available genetic and structural imaging data. This meta-analysis failed to identify a meaningful predictor developed from either data modality, suggesting that alternate approaches should be considered for prediction of cognitive performance.

© 2016 The Authors. Published by Elsevier Inc. on behalf of the Alzheimer's Association. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Keywords:

Alzheimer's disease; Biomarkers; Crowdsourcing; Big data; Bioinformatics; Cognitive decline; Imaging; Genetics

1. Introduction

The Alzheimer's disease DREAM challenge (<http://dx.doi.org/10.7303/syn2290704>) was designed to provide an unbiased assessment of current capabilities for estimation of cognition and prediction of cognitive decline using genetic and imaging data from public data resources using a crowdsourced approach. The ability to predict rate of cognitive decline—both before and after diagnosis—is essential to effective trial design for the development of therapies for Alzheimer's disease (AD) prevention and treatment. Major collaborative efforts in the field are assessing the association of genetic loci with AD diagnosis and the application of structural imaging for development of early biomarkers of diagnosis, but the utility of these approaches to estimate cognition or predict cognitive decline is not well established. This project was designed under the advisement of a panel of experts in the field to evaluate whether these questions could be meaningfully addressed with current methods given existing public data sources. To ensure that these questions were tested across a broad spectrum of the latest analytical approaches, the study was designed as a crowdsourced, community-based challenge in which participants were invited to address one or more of the following three questions [1]: The prediction of cognitive decline over time based on genetic data [2]. The prediction of resilience to cognitive decline in individuals with elevated amyloid burden based on genetic data [3]. The estimation of cognitive state based on structural magnetic resonance (MR) imaging data.

2. Results

2.1. Study design and data harmonization

To ensure that predictors were detecting true biological variation rather than study-specific technical variation, this project required inclusion of data from multiple study sources. Although genetic and imaging data have been generated within many rich longitudinal cohorts across the field, the procurement and harmonization of these data sets were a nontrivial problem that required solutions to overcome political, ethical, and technical barriers. For example, the generation of whole genome sequencing data across multiple AD cohorts within the NIH-funded AD sequencing project has resulted in a powerful resource for genetic analysis in the field but longitudinal information on cognitive traits is not readily available in those data sets. Despite limitations on data accessibility, multiple relevant data sources were identified and used in this project including the Alzheimer's Disease Neuroimaging Initiative (ADNI) [1], the Rush Alzheimer's Disease Center Religious Orders Study [2], Memory and Aging Project (ROS/MAP) [3], and the European AddNeuroMed [4] study, which is part of InnoMed, a precursor to the innovative medicines initiative. Data selection and processing were performed based on data availability across these three data sets. As such, cognition was defined using mini mental state examination (MMSE) scores [5], genetic data were provided based

on imputation across array-based genotype data, and structural MR imaging data were reprocessed in each cohort using a common processing pipeline. Genetic and imaging data were supplemented with a limited set of covariates including diagnosis, initial MMSE score, age at the initial examination, years of education, gender, and *APOE* haplotype. Participants were provided with data from ADNI to train algorithms over a 4-month period and to ensure that participation was not limited by access to compute resources, they were offered use of the IBM zEnterprise cloud to perform analyses. The challenge generated significant interest with 527 individuals from around the world registered to participate. A leaderboard displayed accuracy of submissions throughout the duration of the challenge: 1157 submissions were made for question 1, 478 submissions for question 2, and 434 submissions for question 3. Thirty-two teams submitted final results that were scored based on prediction and/or estimation of blinded outcomes within ROS/MAP for genetic predictions and AddNeuroMed for imaging-based estimations (Fig. 1).

2.2. Genetic prediction of cognitive decline

The first challenge question assessed the ability of current methods to predict change in cognitive examination performance based on genetic data. High prediction accuracy would signal the potential for noninvasive biomarkers of cognition to have a major clinical impact on early AD diagnosis and prevention. Previous efforts to develop predictors of change in cognitive function have not succeeded in providing robust and replicable models [6–8]. Genetic variation has been demonstrated to influence AD status: rare genetic mutations at several loci are implicated in familial forms of early-onset disease [9], whereas common variation contributes 33% to variance in sporadic AD, and 22 loci have been implicated by large-scale genetic association analyses [10,11]. However, with the exception of the *APOE* $\epsilon 4$ haplotype, there has been little success in transforming these genetic associations into meaningful clinical predictions of cognitive decline. For this purpose, participants were challenged to predict 2-year changes in MMSE scores based on genotypes imputed from SNP array data. Participants trained their algorithms with 767 ADNI samples, and the algorithms' predictions were evaluated on a test set of 1175 ROS/MAP samples with blinded outcome measures. The algorithm with the best predictive performance at the midpoint of the challenge did not contain any genetic features beyond *APOE* haplotype. As the goal of this question was to assess genetic contribution to prediction of cognitive decline, this top-ranked algorithm was openly shared across teams as an interim baseline on which to incorporate additional genetic predictors (<http://dx.doi.org/10.7303/syn2838779>). Eighteen teams submitted final predictions. Most methods performed significantly better than a permutation-based random model prediction (Fig. 2A). A cluster of six methods performed significantly better than the others (including the interim baseline model) but were

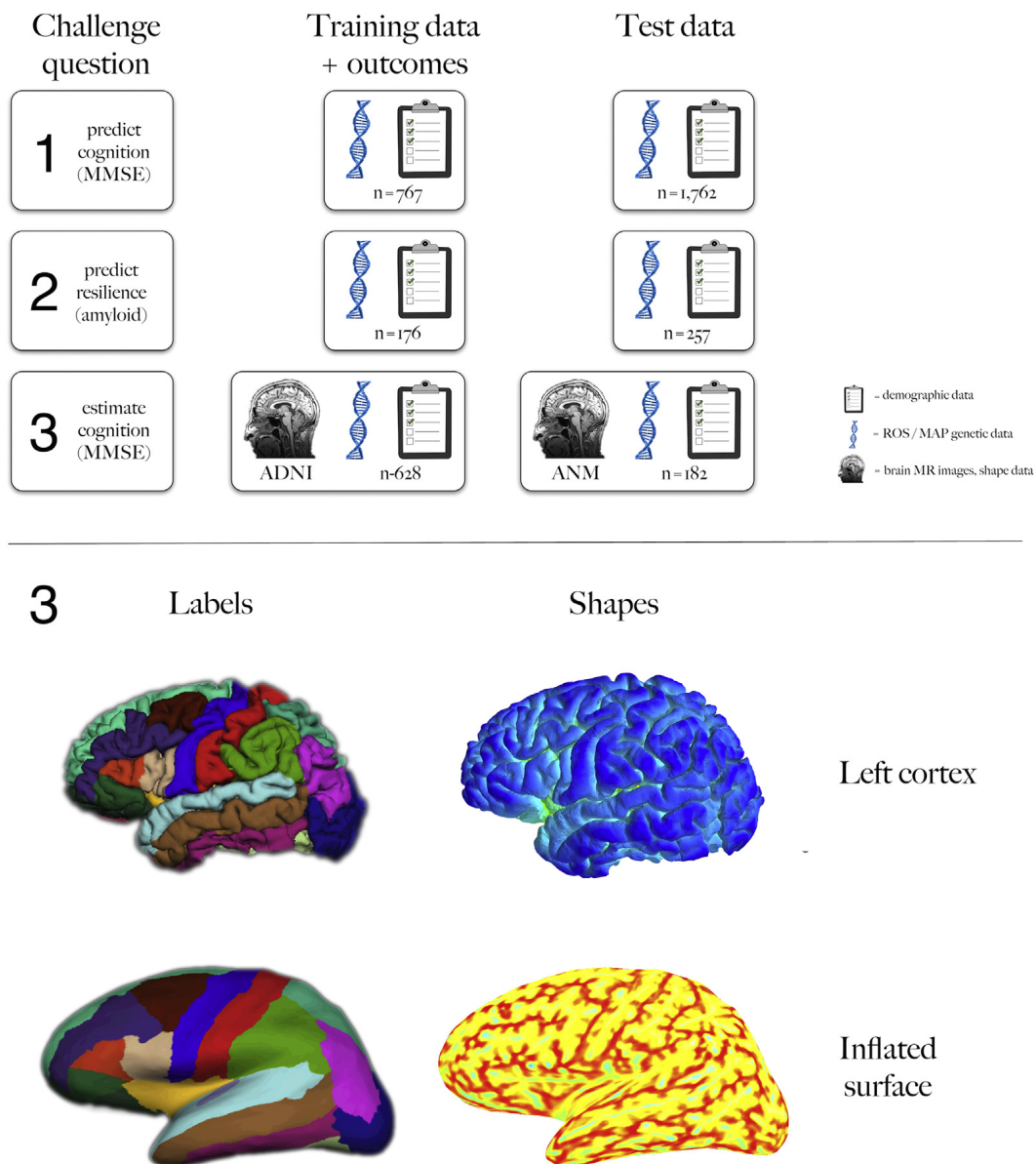


Fig. 1. Challenge overview. The top schematic summarizes the three challenge questions on the left column, the training data in the middle, and the test data on the right, including numbers of subjects. The symbols represent sources of data (demographic, ROS/MAP genetic, and ADNI or ANM brain images and shape information). The bottom panel provides example brain image labels and shape information derived from the Mindboggle software (<http://mindboggle.info>) provided to the participants for question 3. Anatomic labels for left cortical regions are shown on the left and just a couple of the cortical surface shape measures are shown on the right (travel depth on top and mean curvature below), for both uninflated and inflated surfaces (top and bottom rows, respectively).

statistically indistinguishable among themselves (Fig. 2D). Of these, the prediction with the best overall score (team GuanLab_umich from the University of Michigan) achieved a Pearson correlation of 0.382 and a Spearman correlation of 0.433 (the overall score was a rank-based combination of these two measures of performance; see online Supplement and Supplementary Methods: <http://dx.doi.org/10.7303/syn3383106>). However, no significant contribution of genetics beyond *APOE* haplotype to predictive performance was observed across any of the submissions. Given the small sample size, no conclusions can be inferred from this analysis regarding the existence of genetic loci associated with cognitive decline. Rather, these observations suggest that predic-

tors of cognitive decline developed based on genetic data will not be useful within the clinical setting.

2.3. Genetic prediction of cognitive resilience

The second question challenged participants to identify genetic predictors that could distinguish individuals who exhibit resilience to AD pathology as defined by minimal change in cognitive function despite evidence of amyloid deposition [12,13]. Identification of genetic signatures predictive of cognitive resilience would aid in the elucidation of mechanisms that may confer resilience, providing a powerful tool to help advance AD prevention

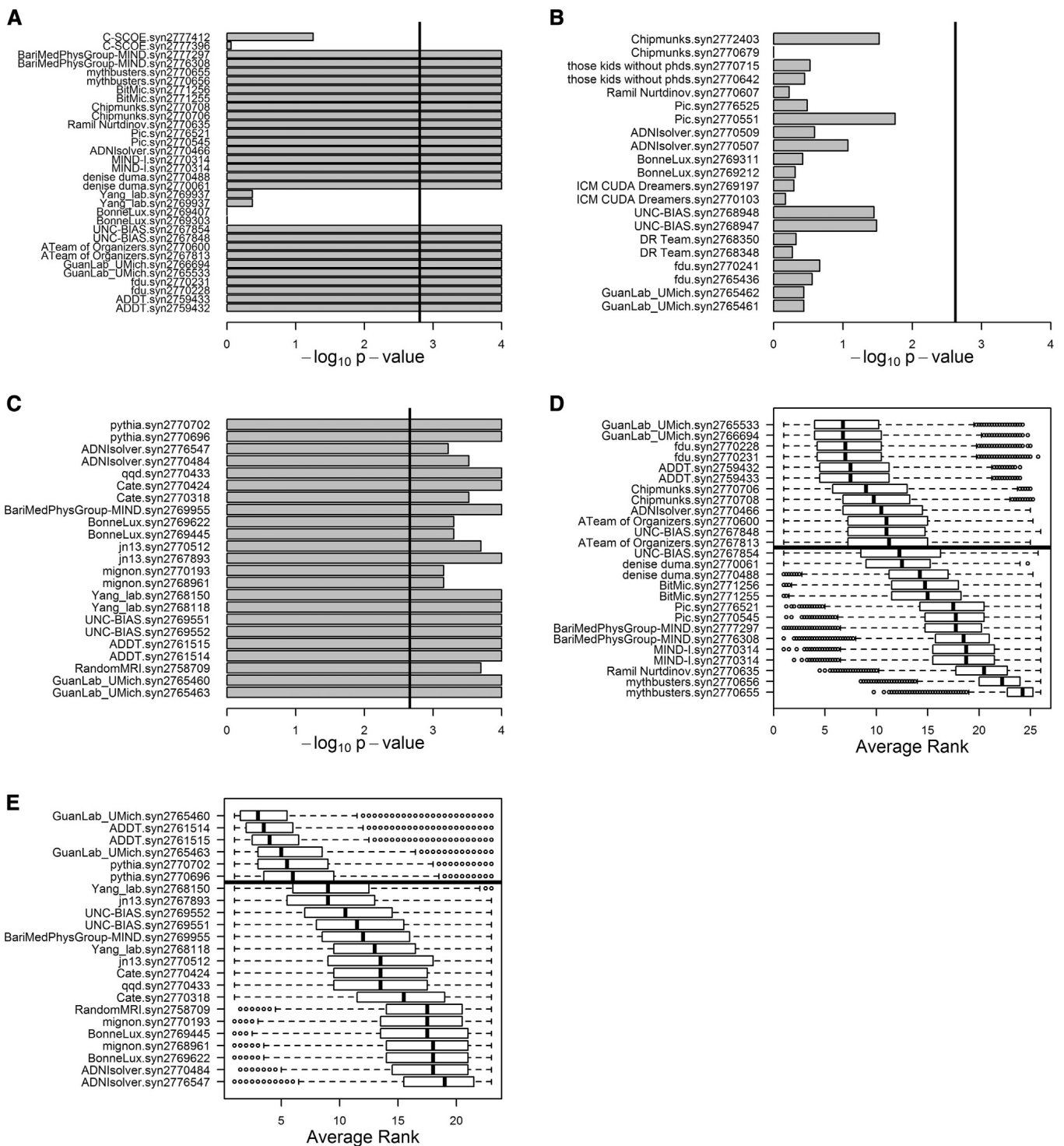


Fig. 2. Performance evaluation results. (A), (B), and (C) report the P values (in negative log 10 scale) for intersection union tests investigating which teams performed better than random for questions 1, 2, and 3, respectively. Explicitly, for question 1 (A), we tested the null hypothesis that at least one of the four correlation coefficients (namely Pearson/clinical, Pearson/clinical + genetics, Spearman/clinical, and Spearman/clinical + genetics) is equal to zero, against the alternative that all four correlation coefficients are larger than zero. Adopting a 0.05 significance level, 26 of the 32 submissions were statistically better than random, after Bonferroni multiple testing correction for 32 tests (submissions crossing the black vertical line). For question 2 (B), we tested the null hypothesis that balanced accuracy = 0.5 or AUC = 0.5, against the alternative that balanced accuracy > 0.5 and AUC > 0.5. In this case, no model performed significantly better than random, and, therefore, no best performer was declared. For question 3 (C), we tested the null hypothesis that Pearson's correlation (COR) or Lin's concordance correlation coefficient (CCC) are equal to zero, against the alternative that both COR and CCC are larger than zero. Adopting a 0.05 significance level, all 23 submissions were statistically better than random, after Bonferroni correction. For all three questions, the P values were computed from an empirical null distribution based on 10,000 permutations. (D) and (E) report the bootstrapped assessment of ranks for questions 1 and 3, respectively. Samples were resampled with replacement from the original data (true outcome and team's predictions), and the ranks of the different teams were reassessed in each of 100,000 resamplings. Submissions were sorted according to the median of their bootstrapped average ranking distributions. The black horizontal line represents the posterior odds cutoff from the Bayesian analysis. Teams above the black line are statistically tied to the top-ranked model, according to a posterior odds threshold of 3.

strategies and treatment development. Eleven teams submitted predictions of resilience based on a training set derived from 176 ADNI subjects. Evaluations were made using data derived from 257 individuals from the ROS/MAP data. Despite using the largest such public data set assembled to date, participants were unable to develop algorithms with predictive performances significantly better than random (see Fig. 2B, online Supplement and Supplementary Methods in Synapse: <http://dx.doi.org/10.7303/syn3383106>). Although it is likely that the study was underpowered due to small sample size and trait heterogeneity, this result suggests that information about cognitive resilience is not easily discoverable from SNP analysis.

2.4. Structural imaging-based estimation of cognition

The third question challenged participants to estimate cognitive state using structural brain image data (Fig. 1, lower panel). Brain imaging has emerged as a powerful method for monitoring neurodegeneration, and there is a great enthusiasm in the field to make use of images for diagnosis and prediction. There have been many attempts in the past to correlate changes in brain shape with disease progression and/or diagnosis, conventionally using measures of volume for a given brain region [14,15]. More detailed shape measures of image features including cortical thickness, curvature, and depth have also been found to be relevant to a variety of neurologic conditions [16]. Participants were challenged to estimate MMSE scores based on structural brain images, or shape information derived from these images. Participants trained algorithms using ADNI data ($N = 628$) and were evaluated using AddNeuromed data ($N = 182$) for which they were blind to outcome measures. To engage as many participants as possible from both within and beyond the neuroimaging community, the data were provided both as raw MR images and as tables containing shape measures (volume, thickness, area, curvature, depth, and so forth) for every labeled brain region. Thirteen teams submitted estimates for final evaluation, and all teams performed better than a random model (see online Supplement and Supplementary Methods in Synapse: <http://dx.doi.org/10.7303/syn3383106>). Three teams performed significantly better than the others (teams GuanLab_umich from the University of Michigan, ADDT from the Karolinska Institute and Pythia from the University of Pennsylvania; Fig. 2C) but were statistically indistinguishable from one another and tied for top average rank (Fig. 2E). The algorithm that generated the best absolute mean combined rank (Team GuanLab_umich) achieved a concordance correlation coefficient of 0.569 and Pearson's correlation of 0.573 (the overall score was a rank-based combination of these two measures of performance). The most common features that contributed heavily to the MMSE estimates across the algorithms were hippocampal volume and entorhinal thickness, corroborating prior work [17–19]. The top three teams also found that inclusion

of shape measures of the entorhinal cortex (volume, curvature, surface area, travel, and geodesic depth) improved overall estimation. Other features that contributed to predictions within the top three teams' results included volume of inferior lateral ventricle and amygdala (see online Supplement and Supplementary Methods in Synapse: <http://dx.doi.org/10.7303/syn3383106>). These results validate an established relationship between structural imaging data and cognition. However, the correlative performance of these estimators was low suggesting that their application in the clinical setting may not be sufficient to inform patient care.

3. Discussion

The AD DREAM challenge provided a formalized assessment of the ability to develop meaningful predictions of cognitive performance from public genetic or imaging data using contemporary state-of-the-art predictive algorithms. Predictive performance across all three of the questions was modest, and most methods performed roughly equivalently. Given this uniform performance, we do not expect that the presented results are a failure of current modeling methods. A more likely explanation is that the data used to address these questions were inadequate to support these tasks. We also note that most research teams that participated in this challenge did not have expertise in the field of AD. Although the few teams that did possess this knowledge did not do better than the others, there remains the possibility that performance would have been improved by the inclusion of more domain experts.

3.1. Use of genetic information for cognitive prediction

The modest performance observed in the 3 questions focused on genetic analysis demonstrated that contemporary algorithms were not able to leverage genetic signal to make useful predictions for cognition. These results support the prevailing expectation that genetic variants of moderate to high frequency will not support viable biomarker development in AD [9–11]. Although heritability estimates and linkage studies have demonstrated that there is a considerable estimated genetic contribution to AD onset and progression [11,20,21], evidence both within the AD field and across other complex disease [22] traits has indicated that this overall genetic contribution is the aggregated compilation of a large number of loci with small— independent or epistatic—effects. Historically, this type of signal is difficult to capture in predictive models and unlikely to be useful in a diagnostic setting [23]. Furthermore, cognition is highly influenced by a host of nongenetic factors relating to lifestyle choices and accumulated exposures that were not represented across all these data sets and, in fact, are not fully captured in most cohorts [24–27]. Nongenetic contributions to cognitive performance may themselves provide an important base for successful predictions.

Within the context of genetic analysis, the absence of these factors from models confounds the ability to detect real genetic signal and impacts the ability to accurately model state-specific genetic contributions. As such, future inquiry into the use of genetic testing for prediction of cognitive performance and AD risk assessment may be better served by focusing on the contribution of rare genetic variation. Recently discovered disease-associated rare variants have larger effect sizes than common variants and confer 2- to 5-fold greater risk or protection in carriers relative to the general population [28–30]. Ongoing large-scale sequencing analyses will identify additional associated rare risk variants. In sufficient numbers, the aggregate prevalence would support the development of a genetic diagnostic containing a library of rare variants.

3.2. Use of structural imaging data for cognitive estimation

Although the inexpensive and noninvasive nature of genetic testing make this approach amenable to population-level disease screening, the resource-intensive nature of image-based testing is better positioned for careful evaluation of high-risk individuals. As such, these approaches are needed to provide a higher confidence estimate of cognitive performance. Although a variety of methods developed within the context of this challenge were able to successfully estimate cognition, none of these methods were sufficiently accurate to merit clinical consideration. These observations support previous work in the field [17,19] and highlight the imperfect relationship between brain structure and function. Newer imaging modalities that focus on brain function and/or pathology—such as FDG-PET [31] or tau imaging [32]—may prove more successful for assessing cognitive dysfunction.

3.3. Effective performance of meta-analysis across diverse cohorts

A major consideration for any meta-analysis is the issue of appropriate harmonization of data across disparate sources. Despite leveraging several of the most deeply phenotyped cohorts in the field, this challenge limited analysis to those traits that were in common across cohorts. Although this approach to data harmonization is standard practice for meta-analyses [10], it greatly reduced the depth of the information available for modeling and influenced the selection of cognitive measures for use as prediction outcomes. Because each cohort had performed a battery of study-specific tests, this greatly limited the ability for finer grained assessment across cognitive processes. A more sensible approach for future analyses may be to focus effort on more sophisticated methods to calibrate disparate cognitive phenotypes across cohorts [33]. Another undesirable consequence of the focus on traits measured in common was the inability to incorporate into model development the full

spectrum of nongenetic and nonimaging factors that are known to influence cognitive performance [24–27]. This suggests the need for development of different approaches for integrating heterogeneous data and/or assessing replication across cohorts. Alternatively, smaller scale analyses that prioritize phenotypic depth over sample size may afford a more refined view of disease.

In summary, this challenge demonstrated that predictions of cognitive performance developed from genetic or structural imaging data were modest across a diverse set of contemporary modeling methods. Future efforts to identify clinically relevant predictors of cognition will benefit from a focus on alternative data sources and methods that work to incorporate greater phenotypic complexity.

Acknowledgments

Author contributions: C.J.B., E.C.N., D.W.F., S.H.F., S.S.G., A.K., J.S.K.K., Y.K., B.A.L., L.M.M., T.J.M., T.C.N., M.A.P., G.S., G.V., and N.J.T. contributed to the challenge organization. V.B., D.D., P.S.H., and R.C.G. contributed with compute resources and scientific advice. G.I.A., N.A., C.A., D.B., R.B., K.L.B., P.C.B., L.C., C.C., F.C., Y.C.C., B.C., C.Y.C., T.Y.C., T.C., S.D., C.D., J.D., Q.D., J.D., D.D., R.E., G.E., E.E., H.F., J.G., E.G., Y.G., M.Y.H., C.H., J.H., J.I., P.I., A.I., Q.J., D.K., R.K., E.L., M.L., E.L., X.L., Z.L., J.L., T-w.M., A.M., E.M., J.N., A.N., M.N., M.N., R.N.N., Y.J.O., Y.P., S.P., S.R.P., P.P., C.P., V.Y.S., P.S., X.S., A.S., S.T., A.T., Y.A.T., E.V., G.X., L.X., Y.X., J.X., H.Y., X.Z., Y.Z., F.Z., H.Z., and S.Z. participated in the challenge community phase. D.A.B., R.J.B.D., S.L., S.J.N., A.S., and M.W.W. contributed with data used in the challenge.

The following people provided final submissions but did not participate in the community phase of the challenge: Lorna Barron (GIDAS, miRcore, 2929 Plymouth Rd. Suite 207, Ann Arbor, MI, USA), Oliver Barron (GIDAS, miRcore, 2929 Plymouth Rd. Suite 207, Ann Arbor, MI, USA), Riccardo Bellazzi (Electrical, Computer and Biomedical Engineering Department, Via Ferrata, 1, Pavia, Italy), Jungwoon Chang (GIDAS, miRcore, 2929 Plymouth Rd. Suite 207, Ann Arbor, MI, USA), Marianne H Cowherd (GIDAS, miRcore, 2929 Plymouth Rd. Suite 207, Ann Arbor, MI, USA), Grace Ganzel (GIDAS, miRcore, 2929 Plymouth Rd. Suite 207, Ann Arbor, MI, USA), Łukasz Grad (Interdisciplinary Centre for Mathematical and Computational Modelling, Pawińskiego 5A, Warsaw, Poland), Inhan Lee (GIDAS, miRcore, 2929 Plymouth Rd. Suite 207, Ann Arbor, MI, USA), Ivan Limongelli (Electrical, Computer and Biomedical Engineering Department, Via Ferrata, 1, Pavia, Italy), Simone Marini (Electrical, Computer and Biomedical Engineering Department, Via Ferrata, 1, Pavia, Italy), Szymon Migacz (Interdisciplinary Centre for Mathematical and Computational Modelling, Pawińskiego 5A, Warsaw, Poland), Ettore Rizzo (Electrical, Computer and Biomedical Engineering Department, Via Ferrata, 1, Pavia, Italy),

Witold R Rudnicki (Interdisciplinary Centre for Mathematical and Computational Modelling, Pawińskiego 5A, Warsaw, Poland; Department of Bioinformatics, University of Białystok, Ciołkowskiego 1M, Białystok, Poland), Andrzej Sułeczki (Interdisciplinary Centre for Mathematical and Computational Modelling, Pawińskiego 5A, Warsaw, Poland), Leo Tunkle (GIDAS, miRcore, 2929 Plymouth Rd. Suite 207, Ann Arbor, MI, USA), Francesca Vitali (Electrical, Computer and Biomedical Engineering Department, Via Ferrata, 1, Pavia, Italy)

This study was supported by the following individuals and organizations: Alan Evans (McGill University), Gaurav Pandey (MSSM), Gil Rabinovici (UCSF), Kaj Blennow (Göteborg University), Kristine Yaffe (UCSF), Maria Isaac (EMA), Nolan Nichols (University of Washington), Paul Thompson (UCLA), Reisa Sperling (Harvard), Scott Small (Columbia), Guy Eakin (BrightFocus Foundation), Maria Carillo (Alzheimer's Association), Neil Buckholz (NIA), Alzheimer's Research UK, European Medicines Agency, Global CEO Initiative on Alzheimer's Disease, Pfizer, Inc, Ray and Dagmar Dolby Family Fund, Rosenberg Alzheimer's Project, Sanofi S.A, and Takeda Pharmaceutical Company Ltd, USAgainstAlzheimer's.

Study data were provided by the following groups: (1) The Alzheimer's Disease Neuroimaging Initiative (ADNI)—ADNI is funded by the National Institutes of Health (U01 AG024904), the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This research was also supported by NIH grants P30 AG010129 and K01 AG030514. (2) The Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago—Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, and U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute. (3) Euro-

pean AddNeuroMed study—The AddNeuroMed data are from a public-private partnership supported by EFPIA companies, SMEs and the EU under the FP6 programme. Clinical leads responsible for data collection are Iwona Kłoszewska (Lodz), Simon Lovestone (London), Patrizia Mecocci (Perugia), Hilikka Soininen (Kuopio), Magda Tsolaki (Thessaloniki), and Bruno Vellas (Toulouse).

RESEARCH IN CONTEXT

1. Systematic review: Extensive literature searches using PubMed establish this as the largest study to date using demographic, clinical, imaging, and genetic data to predict cognitive decline and the first major instance of crowdsourcing analysis in AD.
2. Interpretation: Over 500 scientists worldwide in the analytical portion of the challenge, demonstrating the viability of crowdsourced approaches in AD research. Unfortunately, we were unable to detect meaningful predictors of either cognitive decline or resilience through this effort.
3. Future directions: This experiment in crowdsourcing AD analyses is an invaluable first-of-its-kind contribution that provides a snapshot of both the strengths and limitations in big data analytics in AD research. The relative inaccessibility and heterogeneity across data sources severely limits formalized integration. Mandates on data sharing, considerations of standardized data collection, and mechanisms to integrate heterogeneous data are necessary to address these issues. We anticipate that this work will initiate those discussions across the community.

References

- [1] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, et al. The Alzheimer's disease neuroimaging initiative. *Neuroimaging clinics of North America* 2005;15:869–77. xi-xii.
- [2] Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS. Overview and findings from the religious orders study. *Current Alzheimer research* 2012;9:628–45.
- [3] Bennett DA, Schneider JA, Buchman AS, Barnes LL, Boyle PA, Wilson RS. Overview and findings from the rush Memory and Aging Project. *Current Alzheimer research* 2012;9:646–63.
- [4] Lovestone S, Francis P, Kłoszewska I, Mecocci P, Simmons A, Soininen H, et al. AddNeuroMed—the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Annals of the New York Academy of Sciences* 2009;1180:36–46.
- [5] Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research* 1975;12:189–98.

- [6] Ercoli LM, Siddarth P, Dunkin JJ, Bramen J, Small GW. MMSE items predict cognitive decline in persons with genetic risk for Alzheimer's disease. *Journal of geriatric psychiatry and neurology* 2003;16:67-73.
- [7] Hsiung GY, Alipour S, Jacova C, Grand J, Gauthier S, Black SE, et al. Transition from cognitively impaired not demented to Alzheimer's disease: an analysis of changes in functional abilities in a dementia clinic cohort. *Dementia and geriatric cognitive disorders* 2008; 25:483-90.
- [8] Vemuri P, Wiste HJ, Weigand SD, Shaw LM, Trojanowski JQ, Weiner MW, et al. MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. *Neurology* 2009; 73:294-301.
- [9] Ridge PG, Ebbert MT, Kauwe JS. Genetics of Alzheimer's disease. *BioMed research international* 2013;2013:254954.
- [10] Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics* 2013;45:1452-8.
- [11] Ridge PG, Mukherjee S, Crane PK, Kauwe JS, Alzheimer's Disease Genetics Consortium. Alzheimer's disease: analyzing the missing heritability. *PloS one* 2013;8:e79771.
- [12] Bennett DA, Schneider JA, Arvanitakis Z, Kelly JF, Aggarwal NT, Shah RC, et al. Neuropathology of older persons without cognitive impairment from two community-based studies. *Neurology* 2006; 66:1837-44.
- [13] Price JL, Morris JC. Tangles and plaques in nondemented aging and "preclinical" Alzheimer's disease. *Annals of neurology* 1999; 45:358-68.
- [14] Davatzikos C, Xu F, An Y, Fan Y, Resnick SM. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. *Brain : a journal of neurology* 2009;132(Pt 8):2026-35.
- [15] Misra C, Fan Y, Davatzikos C. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. *NeuroImage* 2009; 44:1415-22.
- [16] Im K, Lee JM, Seo SW, Hyung Kim S, Kim SI, Na DL. Sulcal morphology changes and their relationship with cortical thickness and gyral white matter volume in mild cognitive impairment and Alzheimer's disease. *NeuroImage* 2008;43:103-13.
- [17] Haight TJ, Jagust WJ, Alzheimer's Disease Neuroimaging Initiative. Relative contributions of biomarkers in Alzheimer's disease. *Annals of epidemiology* 2012 Dec;22:868-75.
- [18] Nho K, Risacher SL, Crane PK, DeCarli C, Glymour MM, Habeck C, et al. Voxel and surface-based topography of memory and executive deficits in mild cognitive impairment and Alzheimer's disease. *Brain imaging and behavior* 2012;6:551-67.
- [19] Thung KH, Wee CY, Yap PT, Shen D, Alzheimer's Disease Neuroimaging Initiative. Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. *NeuroImage* 2014;91:386-400.
- [20] Escott-Price V, Sims R, Bannister C, Harold D, Vronskaya M, Majounie E, et al. Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain : a journal of neurology* 2015; 138(Pt 12):3673-84.
- [21] Lee SH, Harold D, Nyholt DR, ANZGene Consortium, International Endogene Consortium, Genetic and Environmental Risk for Alzheimer's disease Consortium, et al. Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum Mol Genet* 2013; 22:832-41.
- [22] Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature genetics* 2013; 45:400-5. 5e1-3.
- [23] Manolio TA. Bringing genome-wide association findings into clinical use. *Nat Rev Genet* 2013;14:549-58.
- [24] Scarmeas N, Stern Y, Mayeux R, Luchsinger JA. Mediterranean diet, Alzheimer disease, and vascular mediation. *Arch Neurol* 2006; 63:1709-17.
- [25] Podewils LJ, Guallar E, Kuller LH, Fried LP, Lopez OL, Carlson M, et al. Physical activity, APOE genotype, and dementia risk: findings from the Cardiovascular Health Cognition Study. *Am J Epidemiol* 2005;161:639-51.
- [26] Lindsay J, Laurin D, Verreault R, Hebert R, Helliwell B, Hill GB, et al. Risk factors for Alzheimer's disease: a prospective analysis from the Canadian Study of Health and Aging. *Am J Epidemiol* 2002; 156:445-53.
- [27] Wang HX, Karp A, Winblad B, Fratiglioni L. Late-life engagement in social and leisure activities is associated with a decreased risk of dementia: a longitudinal study from the Kungsholmen project. *Am J Epidemiol* 2002;155:1081-7.
- [28] Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogava E, Majounie E, et al. TREM2 variants in Alzheimer's disease. *N Engl J Med* 2013;368:117-27.
- [29] Jonsson T, Stefansson H, Steinberg S, Jonsdottir I, Jonsson PV, Snaedal J, et al. Variant of TREM2 associated with the risk of Alzheimer's disease. *N Engl J Med* 2013;368:107-16.
- [30] Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, Bjornsson S, et al. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* 2012;488:96-9.
- [31] Gray KR, Wolz R, Heckemann RA, Aljabar P, Hammers A, Rueckert D, et al. Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease. *NeuroImage* 2012 Mar; 60:221-9.
- [32] James OG, Doraiswamy PM, Borges-Neto S. PET Imaging of Tau Pathology in Alzheimer's Disease and Tauopathies. *Front Neurol* 2015; 6:38.
- [33] Gross AL, Sherva R, Mukherjee S, Newhouse S, Kauwe JS, Munsie LM, et al. Calibrating longitudinal cognition in Alzheimer's disease across diverse test batteries and datasets. *Neuroepidemiology* 2014;43:194-205.

Community-Reviewed Biological Network Models for Toxicology and Drug Discovery Applications

The sbv IMPROVER project team and challenge best performers: Aishwarya Alex Namasivayam¹, Alejandro Ferreiro Morales², Ángela María Fajardo Lacave², Aravind Tallam¹¹, Borislav Simovic¹⁰, David Garrido Alfaro², Dheeraj Reddy Bobbili¹, Florian Martin³, Ganna Androsova¹, Irina Shvydchenko⁹, Jennifer Park⁷, Jorge Val Calvo¹², Julia Hoeng³, Manuel C. Peitsch³, Manuel González Vélez Racero², Maria Biryukov¹, Marja Talikka³, Modesto Berraquero Pérez², Neha Rohatgi⁸, Noberto Díaz-Díaz², Rajesh Mandarapu⁵, Rubén Amián Ruiz², Sergey Davidyan¹³, Shaman Narayanasamy¹, Stéphanie Boué³, Svetlana Guryanova⁴, Susana Martínez Arbas¹, Swapna Menon⁶, and Yang Xiang³

¹Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Campus Belval, Esch-sur-Alzette, Luxembourg, ²Pablo de Olavide University, Ctra. de Utrera, Seville, Spain. ³Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud, Neuchâtel, Switzerland (part of Philip Morris International group of companies). ⁴Institute of Bioorganic Chemistry Russian Academy of Sciences, Moscow, Russia. ⁵Prakhya Research Laboratories, Lakshminagar, Selaiyur, Chennai, Tamil Nadu, India. ⁶AnalyzeDat Consulting Services, Edapally Bypass Junction, Kochi, Kerala, India. ⁷Selventa, Alewife Center, Cambridge, MA, USA. ⁸Center for Systems Biology, University of Iceland, Reykjavik, Iceland. ⁹Kuban State University of Physical Education, Sport and Tourism, Krasnodar, Russia. ¹⁰FM Pharm, Ltd., Sencanski put bb, Subotica, Serbia. ¹¹TWINCORE, Zentrum für Experimentelle und Klinische Infektionsforschung, Hannover, Germany. ¹²Center for Molecular Biology, "Severo Ochoa" – Spanish National Research Council, Madrid, Spain. ¹³Institute 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: ^aThe 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

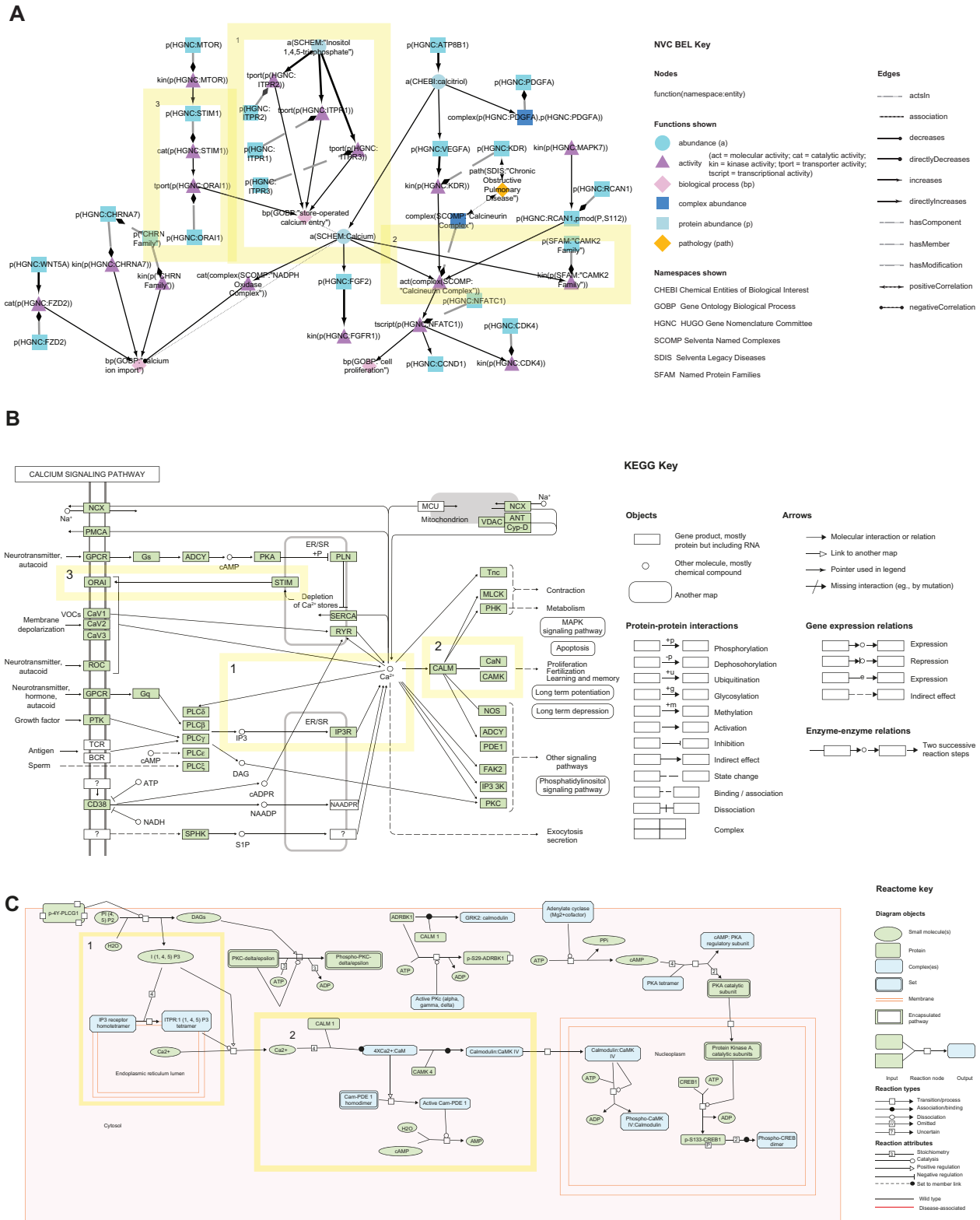


Figure 1. Comparison of the NVC (A), KEGG (B), and Reactome (C) calcium/calmodulin signaling pathways. Shared portions highlighted in yellow with corresponding numbers.

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 $A \rightarrow B \rightarrow C \rightarrow A$ ($A \rightarrow B \rightarrow C \rightarrow A$, 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 $A = \text{cat}(p(\text{HGNC:ADAM17}))$, $B = \text{kin}(p(\text{HGNC:EGFR}))$, $C = p(\text{HGNC:MUC5AC})$, and $D = \text{bp}(\text{GOBP:mucus secretion})$ are all related to each other as $A \rightarrow B, C, D$; $B \rightarrow C, D$; $C \rightarrow D$.

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 betweenness	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	
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 I11r1 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,

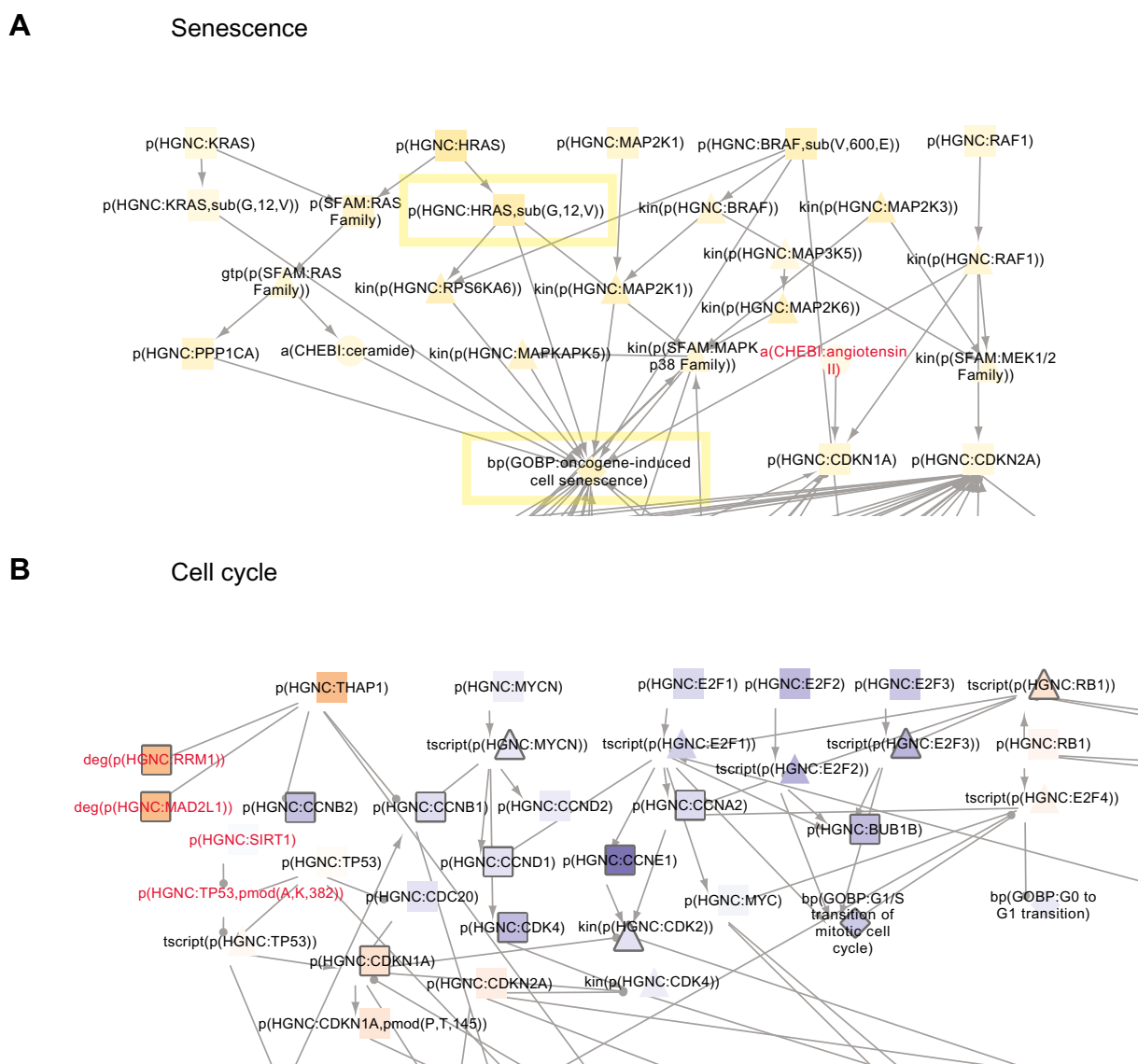


Figure 3. Senescence (A) and cell cycle (B) networks scored with GSE28464 HRASV12 data from the NCBI GEO database. A selection from the TopoNPA-scored version is shown. Arrow edge indicates a positive relationship while ball and stick edge indicates a negative relationship (includes causal and correlative 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

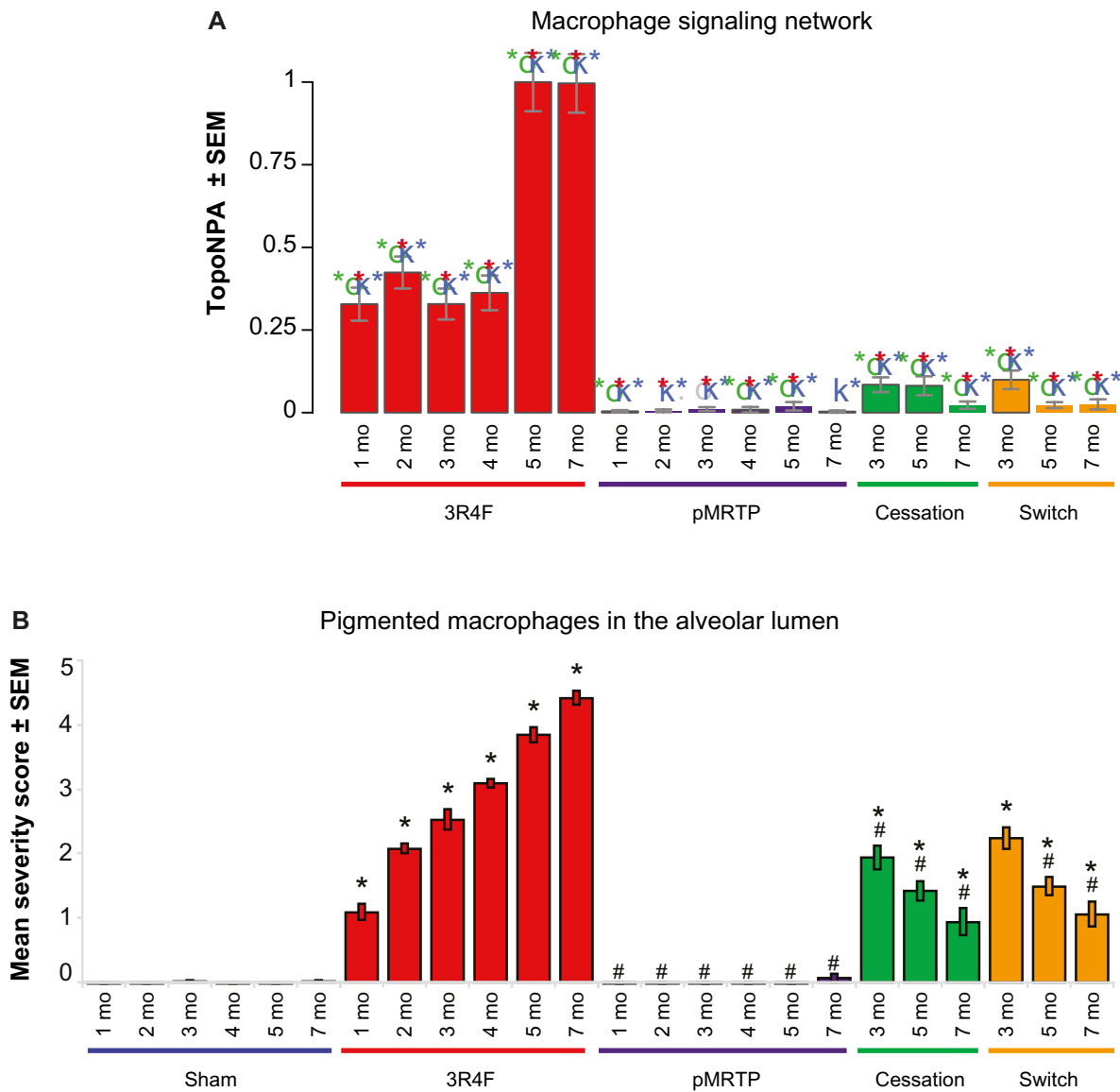


Figure 4. Macrophage signaling network scores in the E-MTAB-3150 dataset and pigmented macrophage counts in the same study. (A) Macrophage signaling network score increased with time with smoke exposure and decreased with switch or cessation. pMRTP did not have significant macrophage signaling network 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

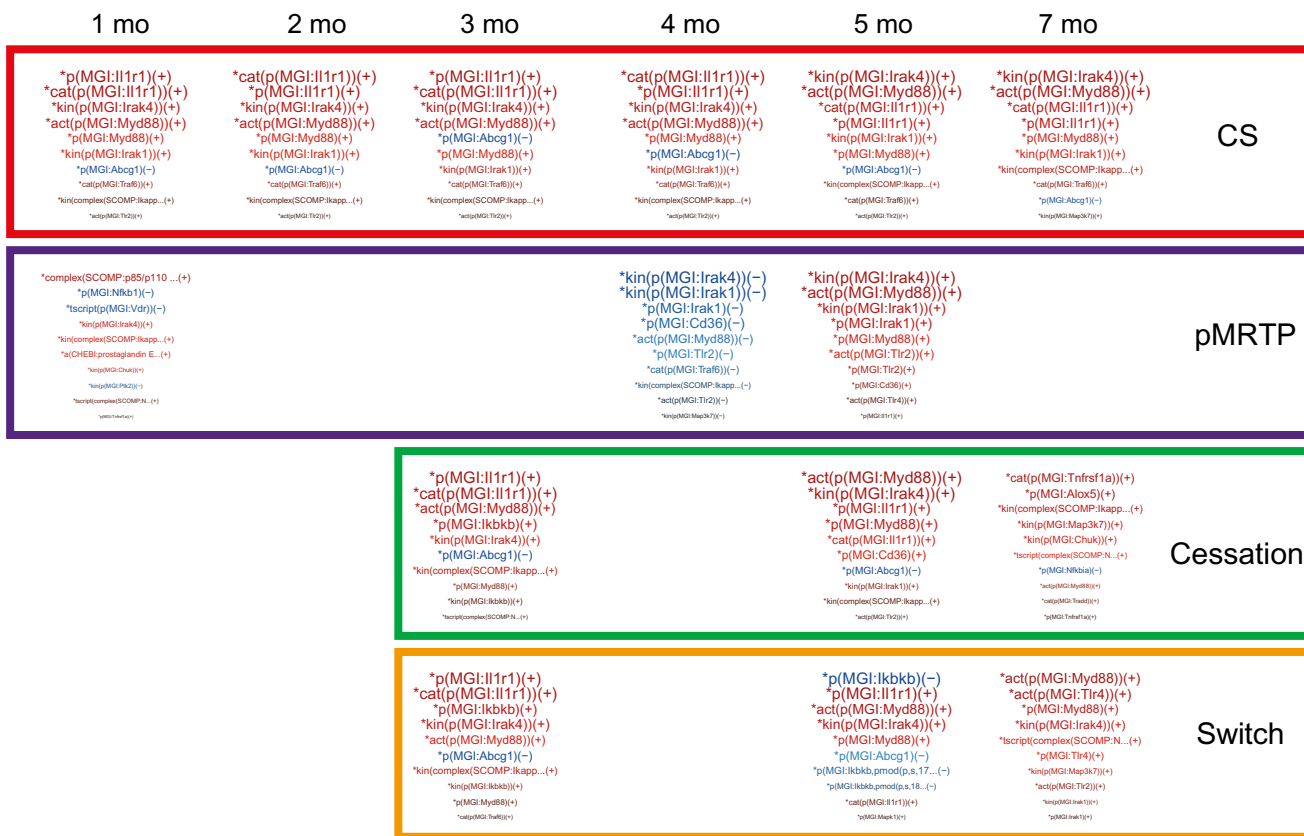


Figure 5. Leading node contribution for macrophage signaling network in the E-MTAB-3150 dataset. Word size indicates relative contribution to network score.

Notes: *significant score; (+) inferred increase; (-) inferred decrease.

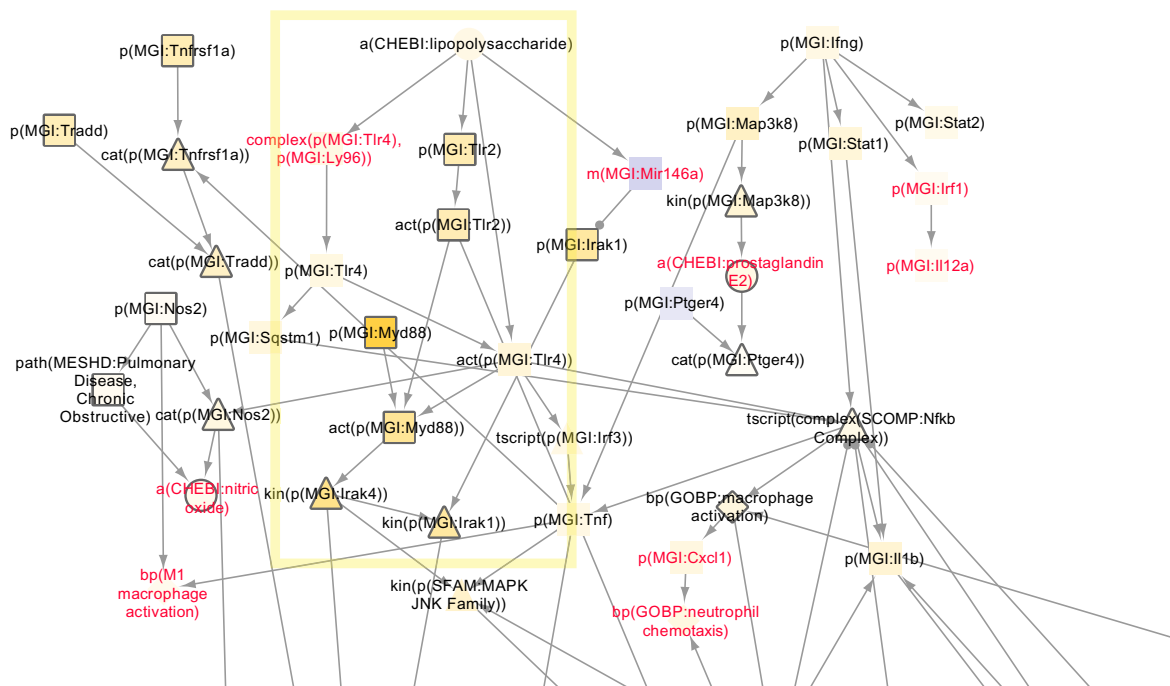


Figure 6. Macrophage signaling network scores for seven-month smoke vs seven-month fresh air using the E-MTAB-3150 dataset. A selection from the TopoNPA-scored version is shown. Arrow edge indicates a positive relationship, while ball and stick edge indicates a negative relationship (includes causal 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, pMRTTP, 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

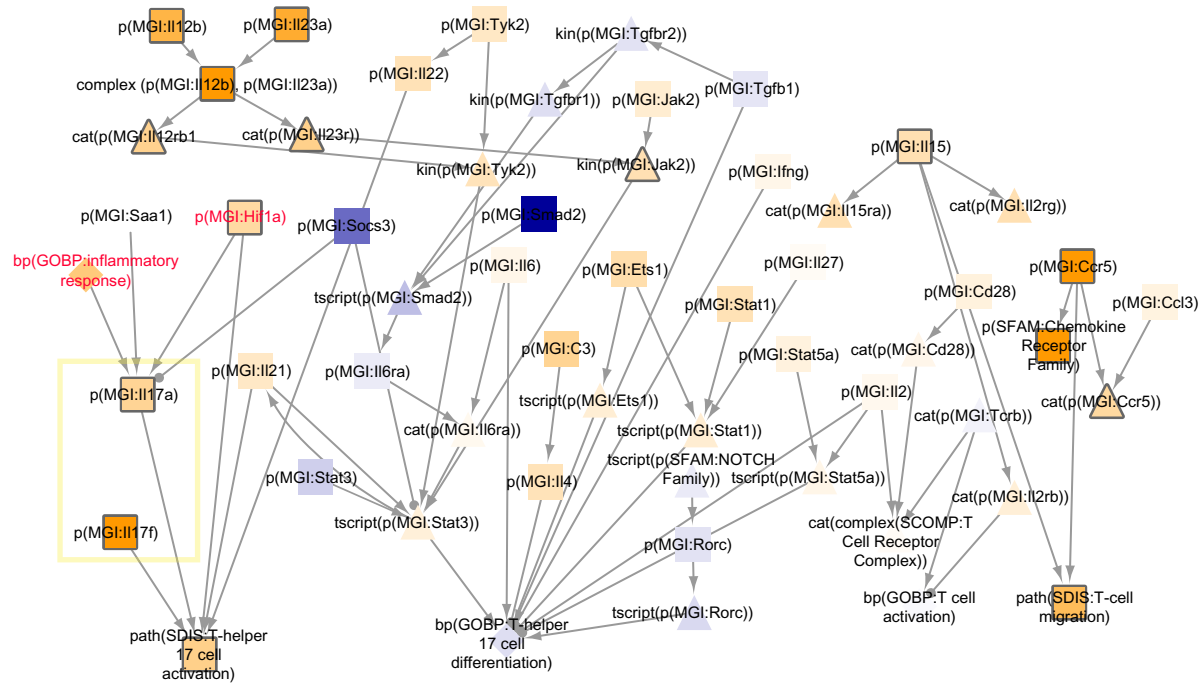


Figure 8. Th17 signaling network scored with GSE52509 mouse lung exposed to 6 month smoke. The whole TopoNPA-scored version is shown. Arrow edge indicates a positive relationship, while ball and stick edge indicates a negative relationship (includes causal and correlative statements). Nodes are colored 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 WikiPathways³⁴ 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 NF κ B 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. *F1000Res*. 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 Clin 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, Hengsternmann 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. Crowd-sourcing 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. Dizceez: 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. Wiki-Pathways: 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>.



Original Article

NeuroTransDB: highly curated and structured transcriptomic metadata for neurodegenerative diseases

Shweta Bagewadi^{1,2,*}, Subash Adhikari³, Anjani Dhrangadhariya^{1,2,†}, Afroza Khanam Irin^{1,2,†}, Christian Ebeling¹, Aishwarya Alex Namasivayam⁴, Matthew Page⁵, Martin Hofmann-Apitius^{1,2} and Philipp Senger¹

¹Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany, ²Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn-Aachen International Center for Information Technology, 53113, Bonn, Germany, ³Department of Chemistry, South University of Science and Technology of China, No 1088, Xueyuan Road, Xili, Shenzhen, China, ⁴Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg and ⁵Translational Bioinformatics, UCB Pharma, 216 Bath Rd, Slough SL1 3WE, United Kingdom

*Corresponding author: Tel: +49-2241-14-2360, Fax: +49-2241-14-2656, Email: shweta.bagewadi@scai.fraunhofer.de

Correspondence may also be addressed to Philipp Senger. Tel: +49-2241-14-2280, Fax: +49-2241-14-2656, Email: philipp.senger@scai.fraunhofer.de

[†]These authors contributed equally to this work.

Citation details: Bagewadi, S., Adhikari, S., Dhrangadhariya, A. *et al.* NeuroTransDB: highly curated and structured transcriptomic metadata for neurodegenerative diseases. *Database* (2015) Vol. 2015: article ID bav099; doi:10.1093/database/bav099

Received 2 April 2015; Revised 7 September 2015; Accepted 10 September 2015

Abstract

Neurodegenerative diseases are chronic debilitating conditions, characterized by progressive loss of neurons that represent a significant health care burden as the global elderly population continues to grow. Over the past decade, high-throughput technologies such as the Affymetrix GeneChip microarrays have provided new perspectives into the pathomechanisms underlying neurodegeneration. Public transcriptomic data repositories, namely Gene Expression Omnibus and curated ArrayExpress, enable researchers to conduct integrative meta-analysis; increasing the power to detect differentially regulated genes in disease and explore patterns of gene dysregulation across biologically related studies. The reliability of retrospective, large-scale integrative analyses depends on an appropriate combination of related datasets, in turn requiring detailed meta-annotations capturing the experimental setup. In most cases, we observe huge variation in compliance to defined standards for submitted metadata in public databases. Much of the information to complete, or refine meta-annotations are distributed in the associated publications. For example, tissue preparation or comorbidity information is frequently described in an article's supplementary

tables. Several value-added databases have employed additional manual efforts to overcome this limitation. However, none of these databases explicate annotations that distinguish human and animal models in neurodegeneration context. Therefore, adopting a more specific disease focus, in combination with dedicated disease ontologies, will better empower the selection of comparable studies with refined annotations to address the research question at hand. In this article, we describe the detailed development of *NeuroTransDB*, a manually curated database containing metadata annotations for neurodegenerative studies. The database contains more than 20 dimensions of metadata annotations within 31 mouse, 5 rat and 45 human studies, defined in collaboration with domain disease experts. We elucidate the step-by-step guidelines used to critically prioritize studies from public archives and their metadata curation and discuss the key challenges encountered. Curated metadata for Alzheimer's disease gene expression studies are available for download.

Database URL: www.scai.fraunhofer.de/NeuroTransDB.html

Background

Considerable effort by the global research community has been dedicated to addressing a limited understanding of the pathogenic events underlying neurodegenerative disease (NDD) (1, 2). The cumulative output of these efforts has established an increased amount of deposited molecular data and published knowledge. As life expectancy continues to rise and treatment options for NDD remain limited, there is an increasing urgency to translate this amassed molecular data into biomarker tools for early diagnosis; to open the possibility of disease altering and preventative therapy (3, 4). Furthermore, biomarkers aiding the decision-making process for therapies targeting specific pathophysiological mechanisms will help to address the high drug attrition rate in the NDD pharmaceutical industry. Informatic efforts to facilitate the integration and interrogation of the distributed molecular data legacy for NDD can enable a systematic and objective prioritization of molecular protagonists and therefore potential biomarkers in NDD (5–8).

In this direction, we have previously developed a semantic framework, called *NeuroRDF* (9), for integration of heterogeneous molecular data types, extracted from biomedical literature, transcriptomic repositories and bespoke databases. *NeuroRDF* enables researchers to formulate biological questions that relate to the interplay of different facets of molecular biology as a formalized query. Even today, the most abundant source of quantitative molecular data remains transcriptomic data, which can support hypothesis-free, elucidation of biological function (10). When the same biological function is replicated in additional expression data sets, it increases the plausibility of the derived hypothesis (11).

The inaccessibility of the brain is a significant barrier to molecular analysis of NDD and this frequently limits the availability of samples from post-mortem tissue (12,

13). This is evident when simply comparing the availability of NDD studies to other disease domains, like cancer (14), in public archives such as Gene Expression Omnibus (GEO) (15) and ArrayExpress (16) (see [Supplementary Figure S1](#)). For instance, GEO contains 157 NDD studies in contrast to 16,910 cancer studies. Therefore, animal models are an important complement to human-derived samples but are at best an incomplete reflection of the human conditions. Assessing the biological complementarity of studies is important when considering a meta-analysis. Such an assessment can be a cumbersome process as searching in these public repositories is principally based on free text. Additionally, limited adoption of controlled vocabularies, such as the Experimental Factor Ontology (EFO) (17), to describe the metadata fields and lack of compliance to defined standards (18) has contributed to the dilemma. This has resulted in metadata being scattered as unstructured prose in public databases and as additional annotations, widely distributed in originating publications. Moreover, applying automated methods to retrieve information from these databases could compromise on the accuracy. On the other hand, capturing missing annotations through the manual curation can incur huge costs of trained labour.

Capturing the associated metadata in a standardized and precise fashion will empower integrative analysis by helping to control sources of variability that do not relate to the hypothesis under investigation (11, 19–21). Ober *et al.* (22) have reported on differing gene-expression patterns related to gender and suggest gender-specific gene architectures that underlay pathological phenotypes. Li *et al.* (23) observed distinct expression patterns, strongly correlated with tissue pH of the studied subjects; these patterns are not random but dependent on the cause of death: brief or prolonged agonal states. Thus, studies enriched

with metadata annotations provide the power to obtain more precise differential estimates.

Related work

Numerous approaches have been proposed to tackle the problem of identifying relevant gene-expression studies and annotating metadata information resulting in several databases, web servers and data exploration tools. These (value added) databases differ from one another based on their objectives, information content and mode of query.

AnnotCompute (24) is an information discovery platform that allows effective querying and grouping of similar experiments from ArrayExpress, based on conceptual dissimilarity. The dissimilarity measure used, Jaccard distance, which is derived from the MAGE-TAB fields submitted by the data owners. Another tool, Microarray Retriever (MaRe) (25) enables simultaneous querying and batch retrieval from both GEO and ArrayExpress for a range of common attributes (e.g. authors, species) (MAGE-TAB is a submission template, tab-delimited, for loading functional genomics data into ArrayExpress. <https://www.ebi.ac.uk/fgpt/magetab/help/>). GEOMETADB (26) is a downloadable database of structured GEO metadata with programmatic querying libraries in both R and MATLAB. However, all the above-mentioned resources suffer from a common limitation: they rely completely on the submitted data and do not provide solutions for missing metadata information.

Several value-added databases invest manual curation effort to enrich metadata information for gene-expression studies. Many Microbe Microarrays Database (M³D) (27) contains manually curated metadata, retrieved from the originating publications, for three microbial species, conducted on Affymetrix platforms. Similarly, the OncoPrint database (28) contains extensive, standardized and curated human cancer microarray data. A-MADMAN (19); an open source web application, mediates batch retrieval and reannotation of Affymetrix experiments contained in GEO for integrative analyses. Microarray meta-analysis database (M²DB) (11) contains curated single-channel human Affymetrix experiments (from GEO, ArrayExpress and literature); categorized into five clinical characteristics, representing disease state and sample origin. However, experiments with missing link to the published paper in GEO and ArrayExpress were excluded. A substantial paucity of sample associated gender information in GEO and ArrayExpress motivated Buckberry *et al.* (29) to develop a R package, *massiR* (MicroArray Sample Sex Identifier) to label the missing and mislabelled samples retrospectively with gender information, based on data from Y chromosome probes. Apart from publicly available resources,

there are various commercial products that contain manually curated transcriptomic metadata: NextBio, Genevestigator and InSilicoDB (30) (<http://www.nextbio.com/b/nextbioCorp.nb> and <https://genevestigator.com/gv/>). However, none of the above databases are optimized to capture detailed metadata specific to neurodegenerative disease. In addition, these databases fail to handle species-specific annotations; especially treatments applied on animal models to partially explicate or treat human-related NDD mechanisms, which may strongly contribute to increase the predictive power of translating preclinical results in NDD drug trials.

Here, we describe the detailed development of *NeuroTransDB*, a manually curated database containing metadata annotations for neurodegenerative studies and an enabling resource for supporting integrative studies across human, mouse and rat species. The participation of our group, at Fraunhofer Institute SCAI, in projects funded by the Neuroallianz Consortium (a part of the BioPharma initiative of the German Ministry of Education and Research) and the evident lack of a comprehensive NDD specific metadata archive has motivated us to develop **Neurodegenerative Transcriptomic DataBase** (*NeuroTransDB*) (<http://www.neuroallianz.de/en/mission.html>). This database now contains more than 20 dimensions of metadata annotations for human studies, as well as mouse and rat models, defined in agreement with disease experts. To demonstrate our approach, we chose to highlight Alzheimer's disease for this publication because it depicts a wide spectrum of the possible annotations across different types of metadata in neurodegeneration. Additionally, we have applied the same approach to all publicly available Parkinson's and Epilepsy studies, which shows that the overall approach is unspecific to the disease. However, the curated data for these two diseases will be released in the future under the terms of a Neuroallianz agreement. The database is updated every six months using highly trained curators. An interactive graphical user interface to access this data is currently being developed as part of the AETIONOMY IMI project (<http://www.aetionomy.eu>).

Curation of gene-expression studies: prerequisites, key issues and solutions

This section discusses the workflow we followed to retrieve relevant gene-expression datasets and to generate detailed metadata annotations for each study (Figure 1). First, we retrieved all functional genomics studies from GEO and ArrayExpress that reference Alzheimer's disease (AD) or a set of AD synonyms, along with the provided metadata (*cf.* Data Retrieval section). Each study

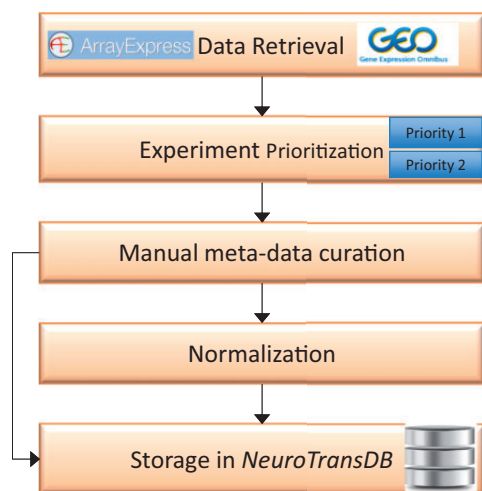


Figure 1. Overall workflow for curation of gene expression studies related to neurodegeneration from public archives. The first step involves automated retrieval of gene expression studies (along with metadata) from public archives such as GEO, and ArrayExpress. The related studies were further assigned to one of the two prioritization classes (priority 1 or priority 2), based on the specific experimental variables. Next, manual curation was applied to capture missing metadata information on priority 1 studies. All the harvested metadata was normalized using standard vocabularies. Both raw and normalized data are stored in *NeuroTransDB*.

was then prioritised (*cf.* Experiment Prioritization section) based on the disease relevancy, experimental type and sample source. Only studies in the top prioritization category were subjected to rigorous, semiautomated metadata curation (*cf.* Metadata Curation section). Annotations are standardized by reference to controlled vocabularies for each extracted metadata dimension (*cf.* Normalization of Metadata Annotations section). The curated Alzheimer's data is stored in *NeuroTransDB*, but in principle the proposed workflow can be applied with little adaptation to any disease indication, especially NDD.

Primary data resources

Together the GEO and ArrayExpress databases constitute a wealth of gene expression studies and are commonly reused for validating new hypotheses and identifying novel signatures through meta-analysis by multi-data integration (11). GEO is the largest public repository of functional genomic data; maintained by the National Center for Biotechnology Information (NCBI) in the USA. ArrayExpress is the European counterpart of GEO and consists of manually curated experimental information imported from GEO, in addition to the data that are directly submitted by the researchers. To support reuse of the deposited studies, each repository adheres to annotation

standards for submission of transcriptomic data: 'Minimum Information about a Microarray Experiment' (MIAME) and 'Minimum Information about a high-throughput nucleotide SEQuencing Experiment' (MINSEQE) (<http://fged.org/projects/miame/> and <http://www.fged.org/projects/minseqe/>). GEO allows data submission in Excel, SOFT or MINiML format and ArrayExpress as MAGE-TAB through Annotare webform tool (<http://www.ncbi.nlm.nih.gov/geo/info/submission.html> and <http://www.ebi.ac.uk/arrayexpress/submit/overview.html>).

Curation team

An obvious prerequisite for any curation process is to have access to specially trained personnel, who understand the key attributes required to adequately describe an expression experiment and are able to complete these attributes by reference to appropriate resources (31). Such individuals are known as biocurators. We assembled a team of candidate biocurators who have adequate biological experience. Each biocurator underwent extensive training in the fundamentals of curation, including the basics of gene expression study design, outlined by experts, scientists and disease experts. Clear curation guidelines (see Experiment Prioritization and Metadata Curation section) and a weekly meeting of the biocurators with one of the experts ensured good quality, consistency, and uniformity in curation procedure. In addition, this provided an opportunity to get feedback from the biocurators for improving and updating the defined guidelines. To keep abreast and eliminate any bias, the curated data was regularly exchanged between them for good interannotator agreement. The experts resolve any disagreement that may arise between the curators.

Data retrieval

Putative AD studies were programmatically retrieved from GEO and ArrayExpress by applying a recall-optimized keyword search approach, *cf.* Figure 2. The keywords include a set of AD synonyms such as 'Alzheimer', 'Alzheimer's' or 'AD' in combination with a species filter. Since ArrayExpress imports and curates the majority of GEO experiments, we firstly queried the former through its REST service (http://www.ebi.ac.uk/arrayexpress/help/programmatic_access.html). Conjointly, we further queried GEO using the *eSearch* Entrez Programming Utilities (E-utils) service to fetch additional identifiers (IDs), which were not picked up by the previous query (http://www.ncbi.nlm.nih.gov/geo/info/geo_paccess.html). The final list of unified experiment IDs was downloaded

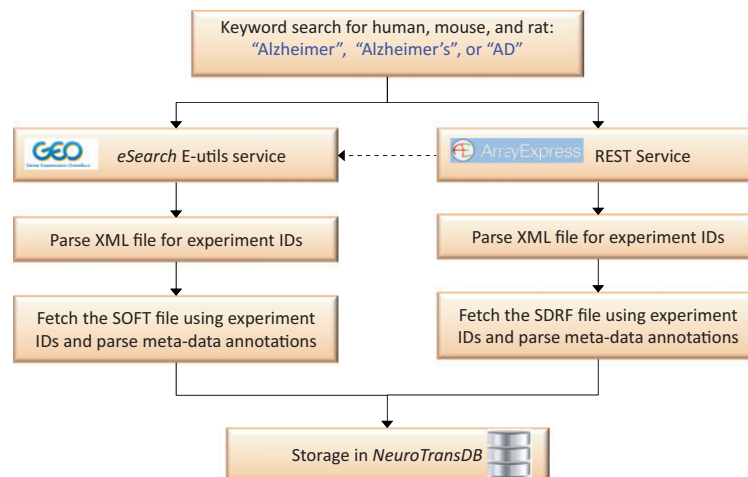


Figure 2. Automated data retrieval of Alzheimer's Disease specific gene expression studies from ArrayExpress and GEO. Here, the dotted line represents the sequence of query performed. Alzheimer's disease specific experiment IDs were automatically retrieved from GEO and ArrayExpress, using keywords, through *eSearch* and REST service respectively. Metadata information was extracted by automatically parsing sample information files (SDRF and SOFT) of these experiment IDs.

(along with their metadata) and stored in *NeuroTransDB*. Metadata information was captured from Sample and Data Relationship Format (SDRF) file of ArrayExpress and SOFT file of GEO (https://www.ebi.ac.uk/fgpt/magetag/help/creating_a_sdrf.html and <http://www.ncbi.nlm.nih.gov/geo/info/soft.html>). The above-described steps are fully automated; enabling an automatic update procedure we run every 6 months to obtain new published studies.

Experiment prioritization

For integrative meta-analysis, combining studies that address the same objectives could minimize biases from cohort selection (inclusion and exclusion criteria) and other design effects. Anatomical and functional heterogeneity arising from experimental sample source, imposes yet another challenge for integrative analysis. Moreover, keyword-based, recall optimized retrieval of experiments does not guarantee its clinical relevancy to the queried indication or organism. Thus, we propose a straightforward binning approach to select potentially eligible studies for AD as illustrated in [Figure 3](#).

Firstly, we identified experiments relevant to AD indication, if not relevant we mark them as unrelated (referred as AD3 in the database). Relevancy is defined on the basis of the experiment's characteristics: investigation on AD mechanism, AD associated mechanism, AD genes or contains samples that belong to direct or implicated effects of or on AD. For example, GSE4757 is relevant to AD since it investigates the role of neurofibrillary tangle formation in Alzheimer patients between normal and affected neurons.

The retained AD-related experiment IDs were manually classified by biocurators into one of the two-prioritization categories (*cf.* [Figure 3](#)). To support this process, a set of classification rules were devised that capture two important considerations: organism specificity and source of the samples used in the study. Although curation with regards to these considerations is of obvious importance, no previously published guidelines were available for reference. To our knowledge, this is the first work where such a guideline has been explicitly detailed. A simplified description of the classification rules adopted for AD disease prioritization is provided below:

Priority 1

- Experiments that study AD pathophysiology in *in vivo* systems
- Studies containing samples from:
 - Human AD patients such as blood, brain tissue, serum, etc.
 - Animal model samples such as mouse brain tissue or rat brain, e.g. C57BL/6 mice, Sprague-Dawley rat, etc.
 - Animal models modified to study the role of an AD gene (knock-out models), or AD mechanism (transfected models), or diet/drug treatments (treated models), such as TgAPP23, APLP2-KO mice, etc.
- Experiments containing only healthy/normal samples from human/mouse/rat that are a part of a bigger study investigating AD

Priority 2

- Experiments that study AD pathophysiology in *in vitro* systems

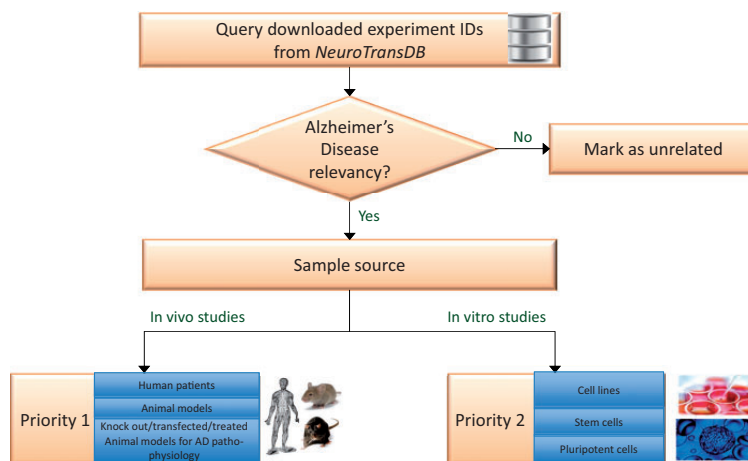


Figure 3. Experiment prioritization for metadata curation in *NeuroTransDB*. All the downloaded Alzheimer's Disease experiments were first checked for their disease relevancy. Those experiments which were falsely retrieved, are marked as unrelated. The remaining experiments were classified into one of two priority classes based on the experiment type: In vivo or In vitro studies. For priority 1, we considered direct/primary samples from human or animal models such as brain tissue, blood, etc. Experiments that were conducted on derived sample sources such as cell lines, were put into priority 2 class.

- Studies containing samples from derived or cultures sources:
 - Cell lines
 - Pluripotent cells
 - Stem cells

Incorrect organism or disease specificity

Although the experiment retrieval step was restricted to a specific organism and disease conditions, we observed differing levels of specificity. For example, some mouse studies were retrieved when querying for human studies. Similarly, we obtained experiments for related diseases such as Parkinson's disease, or diabetes, when querying for AD. Therefore, during study prioritization it was important to confirm the species of origin and relevancy of the study to AD. It's also possible that keyword-based retrieval may miss AD studies due to incorrect disease or organism tagging. However, we did not perform an exhaustive search for such falsely ignored studies, since it would require immense human effort.

Ambiguous species designation

In some studies, human cells such as embryonic stem cells are injected into animal models and post-mortem samples from these animal models are extracted for transcriptomic analysis (e.g. GSE32658 experiment in GEO). Such a study could arguably be classified as either human priority 2 or mouse priority 1. After several discussions, we concluded to prioritize such experiments based on the organism from which the final sample was extracted. In this case,

although the mouse was grafted with human tissue, we prioritized it to mouse priority 1.

Superseries redundancy

During prioritization, we retrieved several superseries experiments from GEO. Manual inspection revealed that not all the subseries IDs of these superseries experiments were retrieved (see Data Retrieval section) (A SuperSeries is simply a wrapper to group of related Series (typically described in a single publication). It facilitates access to the entire dataset, and establishes a convenient reference entry that can be quoted in the publication (definition provided by the GEO team, as of 27 October 2014) and a subseries is an experiment that is a part of superseries.). With careful manual inspection, we included missing subseries, further subjected to prioritization. Conversely, if the inclusion of superseries resulted in the duplication of experiments, we removed the duplicates. Having assigned priority categories to all retrieved AD studies, further metadata curation was focused on the priority 1 studies. Metadata curation steps are described below.

Metadata curation

Precisely and comprehensively capturing the accessory information for a transcriptomic study as meta-annotations, is an important precursor to identification of comparable experiments that address the biological question at hand. Unfortunately, the current, general, submission standards do not cater to the needs of metadata annotations, specific to a disease domain, during submission. In subsequent

sections, we discuss the metadata curation for NDD and key issues faced during the process.

Metadata annotation fields

We assembled a list of metadata annotations determined to be important for evaluating NDD studies in a process involving consultation by NDD domain experts. All the metadata fields were categorized as organism attributes and sample annotations, based on their relevancy to organism or sample source. Table 1 provides detailed descriptions of curated metadata fields including examples for human, mouse and rat.

Several animal models and *in vitro* systems have been defined that partially mimic the human diseased conditions. Animal models provide experimentally tractable systems for interrogating NDD, however, not all animal models faithfully mimic human pathophysiology. A dedicated set of meta-annotation was defined for NDD animal models to support assessments of inter-study comparability and translatability to human disease, cf. Table 2. These fields were defined with assistance from biologists and disease experts from industry.

Metadata curation workflow

To capture all the relevant meta-annotations, we designed a semiautomated curation workflow, illustrated in Figure 4. Firstly, we automatically retrieved all the available meta-annotations from GEO and ArrayExpress (cf. Figure 2). Annotations were captured in an Excel template as shown in Supplementary Figure S2 (A) and confirmed by our trained curators to rectify any inaccuracies.

To capture incomplete and newly defined meta-annotations, we followed a two-step approach. First, we check if the required meta-annotation entries are directly available in GEO, GEO2R or ArrayExpress (<http://www.ncbi.nlm.nih.gov/geo/geo2r/>). Where the required information is complete, we directly update *NeuroTransDB*, otherwise we move to a second step to manually harvest information for missing annotations. Missing information is retrieved from the originating publications and associated Supplementary files. When necessary, corresponding authors were contacted to request missing entries. The list of experiment IDs where we contacted the authors for further information, along with reason of contact (priority 1 experiments only) are provided in Supplementary Table S1. In most cases, the corresponding author or one of the coauthors responded to our queries; whereas, in few other cases the email addresses no longer remained valid. In the event that the authors do not respond or we were unable to contact them, information in primarily deposited database is

used as the final authoritative source. Once all the relevant data was captured, we updated the annotations in *NeuroTransDB*. If needed, we updated our automated retrieval iteratively.

To demonstrate the metadata curation process, here we relate our experience with study GSE36980 that includes a total of 79 samples. Common MIAME annotations such as gender, age and sample tissue were automatically captured from ArrayExpress and GEO. The associated publication contained further useful information on the enrolled patients, namely: disease stage, post mortem interval before sample extraction and preservation, pathological diagnosis and whether the patient suffered from comorbidities such as diabetes. This information was located in Supplementary File S2 of the associated publication (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4128707/bin/supp_bht101_bht101supp_table2.xls). However, lack of a common ID to enable mapping between the sample entries in GEO and the associated Supplementary File S2 impeded curation. For example, sample GSM907797 in GEO is annotated as being derived from a 95-year-old female patient. However, in their Supplementary file, there are two entries that contain information for patients with same age and gender. The ‘No.’ column, assumed to be patient ID, in the Supplementary file was not helpful for mapping, since it was not mentioned in GEO. Thus, we contacted the authors for the missing link. They provided us an additional Excel sheet where the GEO sample ID was mapped to the ‘No.’ column in the Supplementary file (cf. Supplementary Figure S2 (B) and (C)). As a consequence, we achieved a 28.5% increase in the missing metadata information (cf. Table 1 for total number of fields) after contacting the authors.

Automated meta-annotation retrieval challenges

During automated retrieval of metadata fields, we observed several alternate representations of information for certain annotation types in the archives. For example, age information can be provided in the Characteristics section of GEO or ArrayExpress as ‘age: 57 years’ or ‘Stage IV, male, 57 years’ and so on. We attempted to prenormalize these diverse representations and automatically extract the correct information, however, due to the heterogeneity in data representation, manual curation was still required.

Although ArrayExpress and GEO provide programmatic access to their meta-annotations, much essential information appears in fields meant for general categories. For example, information about the sample source and clinical disease presentation appear in the sample title ‘PBMC mRNA from Alzheimer’s disease patient 2’. Adhering to the standard submission protocol for data

Table 1. Detailed description of Neurodegenerative disease metadata fields outlined for human, mouse and rat

Annotation type	Metadata fields	Description of the annotation	Relevancy for NDD	Examples	References
Organism attributes	Age	Age of the organism	Main factor for predisposition to disease	84 years, 9 months	(32–35)
	Gender	Gender of the organism	Possible disproportionate effect arising from difference in anatomy and hormonal composition	Male, female	(36, 37)
	Phenotype	Clinical phenotypes of the organism from which the sample was extracted	Supports comparative analysis for underlying pathomechanisms based on the observable/measurable characteristics	Healthy control, early incipient	(38)
	Behavioural Effect	Description of behavioural changes occurring in organism due to treatment or other effects	Impact of developed drug or other environmental factors to treat or reduce the disease/disease symptoms	Reduced agitation/aggression	(39, 40)
	Disease type	The disease occurrence is due to hereditary or effect of environmental factors	To distinguish the genetic variability and complexity between the two types during analysis	Sporadic, familial	(41)
	Stage	Disease stage of the organism from which the sample was extracted	Capability to distinguish severity of the affected disease	Incipient, severe, BRAAK II	(42)
	Cause of death	Reason for the organism's death	To determine if Alzheimer's disease or its associated comorbidities are major contributors to death rate	Respiratory disorder	(43)
	Comorbidity	Existence of another disease other than Alzheimer's	To determine the impact of another disease on Alzheimer's disease aetiology and progression	Type 2 diabetes	(44, 45)
Sample annotations	Post mortem duration (PMD)	Duration from death till the sample extraction from the dead organism	To assess quality and reliability of the sample obtained by measuring RNA integrity that is influenced by natural degradation of the sample after death	2.5 hours	(46, 47)
	pH	pH value of the extracted post-mortem sample	Indicator of agonal status and RNA integrity	6	(48–50)
	Functional effect	Description of functional effects observed	Observed changes such as gene expression, post-translation, or pathway due to external effects	Decreased expression of BDNF gene, reduced A β toxicity	(51, 52)
	Brain region	Brain region of the extracted sample	Provides information of pathogenesis and disease progression, as AD does not affect all the brain regions simultaneously	Hippocampus	(53, 54)
	Cell and cell parts	Type of cells or cell parts extracted from the sample for analysis (if any)	To determine cell type specific expression influencing pathogenesis and regional vulnerability	Synaptoneurosome, neurons and astrocyte	(55, 56)
	Body Fluid	Type of body fluid used for analysis	Could serve as biomarkers for early diagnosis and therapy monitoring	CSF, blood	(57–59)

The table provides a list of metadata fields, confirmed by disease experts, critical for NDD meta-analysis. The selected fields are classified as organism attributes and sample annotations based on their relevancy to organism or sample source.

Table 2. Detailed description of additional metadata fields, defined specifically for mouse and rat models

Annotation type	Metadata fields	Description of the annotation	Relevancy for NDD	Examples	References
Organism attributes	Physical injury	Method used to cause brain injury in animal models	Consideration for analysing plaque formation in animal models to mimic disease symptoms in human	Traumatic brain injury, ischemia reperfusion injury	(60, 61)
	Type of treatment	Description of chemical, drug, genetic or diet treatment	Consideration for determining the effect of treatment on animal models either to mimic or treat the disease/symptoms	Long-term pioglitazone, BDNF treated	(62, 63)
	Dosage	Detailed description of the dosage associated with “type of treatment” description	Consideration of the right quantity of the substance for determining the effect on animal models either to mimic or treat the disease/symptoms	Total polyphenol 6mg/kg/day, received drinking water without ACE inhibitor	(64, 65)
	Mouse/rat strain name	Mouse model official or author given name	To determine the effect of different manipulated animal models in recapitulating key AD features capable of extrapolating to human studies	C57BL/6-129 hybrid, Sprague–Dawley rat	(66, 67)
	Mouse/rat weight	Weight of the animal model during analysis	Establishing a causative link to metabolic disruption	100–150 g	(68)

These additional metadata fields are defined by disease experts as critical for translating mouse/rat model outcomes to human, in the field of neurodegenerative diseases.

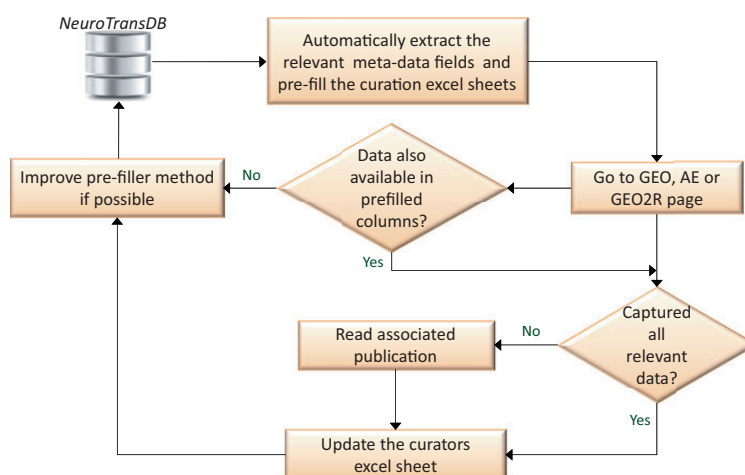


Figure 4. Semi-automated workflow for metadata curation. Automatically extracted metadata fields are rechecked by the curators. To capture the missing fields, curators browse through GEO, ArrayExpress (AE) or GEO2R experiment’s description pages. For cases where the information is still incomplete, associated fulltext publications and their associated supplementary material are read. All the extracted metadata annotations are stored in *NeuroTransDB*. Intermediately, if feasible, automated extraction leverages on curator’s experience for improvement. This process is carried out half yearly.

entry, this information should appear in the ‘Characteristics column’ of ArrayExpress and GEO. Again inconsistent adherence to annotation standards means that manual inspection is needed to capture correct and complete information from these archives.

Accessing linked publications

For annotation information that is not directly available from the source repositories, we refer to the associated full text publications. However, not all deposited studies link to an associated publication in PubMed, contributing to a

significant loss of information while curating. We attempted to overcome this by searching for an associated article using the study title with search engines such as SCAIView and/or Google (<http://www.scaiview.com> and <https://www.google.com>). [Supplementary Figure S3](#) shows the percentage of articles that were retrieved with different search strategies. We are aware that not all the experiments in these databases are associated with published article (14%), but for 9% of the experiments (prioritized as 1) we were able to link them to publications through a title search. We strongly encourage study depositors to provide PubMed annotation whenever available to allow enhanced meta-annotation. Additionally, database owners should find a more robust way to update their resources.

Duplication and inconsistent sample counts

We observed differences in sample counts for some experiments between ArrayExpress and GEO, when downloaded automatically. For example, GSE49160 contained 36 samples in GEO and 72 samples in ArrayExpress. Following closer inspection at several similar experiments, we found that ArrayExpress duplicates sample IDs to provide separate links to different raw file formats or large raw files split into smaller ones (57%), processed raw files (17%), separate entry for each channel in double channel arrays (14%) and replicates (12%) (*cf.* [Supplementary Figure S4](#)); moreover, the duplicated samples mostly represented the same annotation information. Since, we used sample IDs as a unique entry in our database, the duplicated IDs were replaced with the last entry from the archive, in *NeuroTransDB*, as read by our algorithm; thus a risk of losing the raw file or other non-duplicated annotation information.

Apart from duplication, occasionally some samples were missing in one archive relative to the other. For example, GSE47038 had some additional samples in ArrayExpress, which were not present in GEO. When we contacted the ArrayExpress team, they suggested that the experiment entry could be out of sync, since each entry from GEO is uploaded into ArrayExpress only once and is not updated if GEO deletes some samples later. However, they have now corrected the entry. This demonstrates a need for periodic review of study records in each database.

Missing RAW filenames

Public transcriptomic archives provide a gateway for the search and retrieval of studies for subsequent analysis outside of the platform. Therefore, one has to obtain the link between a sample's raw file name and corresponding phenotype. However, this is not the case when applying automated downloads. The majority of the raw file names present in public archives contain syntactical errors such as

surrounded by brackets or separated by comma; moreover, such entries could be normalized through a simple script. In cases where no information about sample's raw file name is provided, manual intervention is required to link sample's raw file to its respective sample. This clearly indicates the need for standardization of the database entries for automation and to prevent loss of information.

Incorrect and incomplete metadata information

We also observed inconsistent meta-annotations between a study deposited in an archive and the information in the linked publication. In GEO for experiment GSE2880, the sample description page states that male Wistar rats have been used for the study. However, when we looked into the associated full text article, in the Methods section, the authors clearly mention using female Wistar rats (69). We are still waiting for the author's reply to correct the gender information for this entry. Another example is GSE18838, we observe that the ratio of male to female patients provided in GEO (male/female: 19/9) is different from that reported in the Supplementary file (male/female: 18/10); additionally, [Supplementary Table S2](#) provides detailed challenges faced during mapping of age and gender information to samples. When searched in ArrayExpress, this experiment has been removed from the database, for unknown reasons. In yet another example, GSE36980, the age information for sample GSM907823 and GSM907823 vary between GEO (84 and 81 years, respectively) and ArrayExpress (74 and 86 years, respectively). From these anecdotal experiences, it is evident that one has to spend immense effort to obtain correct metadata information. Database owners and the submitters have to take utmost care to provide the correct data for reproducibility.

Information extraction from chained references

One further time consuming task included looking following chains of references to previous publications for human and animal model information such as mouse name, cross breeding steps applied and human subject information. In some cases, we had to tediously trace back 5–6 cross-referred publications to obtain the original source of information.

Normalization of metadata annotations

Meta-annotation involved the curation team extracting the original text as provided in GEO/ArrayExpress or in the published literature. We observed many different ways to express information for each annotation field, with obviously ramifications for accurate and efficient querying of *NeuroTransDB*. In an effort to standardize entries for different annotation fields specific controlled vocabularies were adopted during curation.

Age and gender normalization

We observed several different ways of representing age such as ‘24 yrs’, ‘25 yo’ and ‘23 ± 2 years old’. All age values were standardised by converting to simple decimal numbers, e.g. 24.00 for 24 years. Similarly for gender, we used a consistent representation of ‘M’ and ‘F’. As an example, gender information for GSE33528 samples were reported in the associated article (40) as ‘70% of the participants were women’. Here, we annotated the information as ‘70% female’. Although the annotations such as ranges (e.g. ‘23 ± 2 years old’), ratios (male/female: 19/9), or percentages (70% female) (40) are study-level annotations, they were provided as sample level annotations; as they do not contribute to reasonable cohort selection we did not normalize them.

Phenotype, brain region and stage normalization

Disease phenotype and stage information contributes to specific details of clinical manifestations whereas the tissue source (hereafter brain region) caters to the sample origin. For all the curated phenotype mentions (human), we generated a binning scheme: diseased, control or treated. These binned terms were further mapped to controlled vocabularies provided by Alzheimer’s Disease Ontology (ADO) (32). Other annotated terms that are not specific to AD were mapped to the Human Disease Ontology (33), Medical Subject Headings (MESH), Medicinal Dictionary for Regulatory Activities (MEDDRA) and Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) (34) ontologies (<http://bioportal.bioontology.org/ontologies/MESH> and <http://bioportal.bioontology.org/ontologies/MEDDRA>). This caters the need to query samples at a more abstract level, for downstream analysis. In total, for AD, we obtained 481 phenotype mentions assigned to at least one entry in the bins generated. Similarly, all the stage mentions (117 terms) were mapped to ADO, and ONTOAD (35). Mentions of brain region (41 unique terms) were tagged to Brain Region and Cell Type Terminology (BRCT) (<http://bioportal.bioontology.org/ontologies/BRCT?p=summary>). Please refer to [Supplementary File S2](#) for detailed mapping of human annotation terms to controlled vocabularies.

Normalization of animal models

Similar to human phenotype normalization, we have normalized mouse and rat phenotype terms to EFO and SNOMED-CT. Different treatment procedures have been used to generate animal models that capture specific aspects of human diseases. At times, the incomplete nature of the models could lead to inadequate or misinterpretation of results. Thus, it is necessary to know the experimental procedures used on these animal models. To enhance this

interpretation, we have binned all the captured animal model information, during the metadata curation, to a higher level of abstraction, further mapped to EFO, the National Cancer Institute Thesaurus (36), and the BioAssay Ontology (37). In addition, we mapped mouse and rat names to EFO, Jackson Laboratory database identifiers, and Sage Bionetworks Synapse Ontology (<http://jax-mice.jax.org/query/f?p=205:1:0> and <http://bioportal.bioontology.org/ontologies/SYN>). This provides more flexibility during querying of samples from specifically treated animal models. Please refer to [Supplementary Files S3](#) and [S4](#) for mapping of mouse and rat-related terms to controlled vocabularies.

For some of the metadata terms, there were no controlled vocabularies available, e.g. ‘Vehicle #1:non-transgenic’ or ‘BDNF-treated’, describes that the mouse is non-transgenic and a vehicle in the former case, while in the second case it is specific gene treatment. Such terms were mapped to either of the phenotype’s controlled vocabulary. In case of human stage mentions, specific stages such as Braak II or cognitive scores, such as CERAD, MMSE, etc. could not be mapped to any staging controlled vocabulary as most of the ontologies used higher level of staging, namely Braak. Moreover, in most of the ontologies cognitive tests are not classified under staging, but rather as cognitive tests. This has prompted us to generate a more detailed hierarchical representation of the above-mentioned binning schemes, which will be published separately as ontology, specifically for neurodegenerative gene expression studies. However, for current version, we stick to the already available controlled vocabularies, in addition to our internal classification.

Curation results and discussion

Compliance to standards

Authors tend to provide minimum information as required by the guidelines in archives; publishing major part of the experimental metadata annotations in associated publication. To test, whether the authors adhere to the minimum compliant standards, we performed an assessment of the complaint scores provided by ArrayExpress, the highest score being 5, for Alzheimer’s studies. [Figure 5](#) shows the trend in distribution of retrieved AD experiments (see Data Retrieval section) in ArrayExpress, based on the published MIAME and MINSEQE scores (for human, mouse and rat experiments). We observe the trend of submission is concentrated around the score of 4, showing that the submitted data are not fully MIAME or MINSEQE compliant; leading to variable levels of information stored in these archives.

To conclude that not all the submitters' abide 100% by the compliant standards, we investigated if this trend is same for all other disease domains; we chose one among the most studied cancer disease, Lung Cancer, and generated similar results to AD. [Supplementary Figure S5](#) shows the distribution of compliant standards across Lung Cancer studies. From this observation, we show that the loss of information follow the same pattern across all submissions (varying mostly around score of 4). As a result, automated retrieval and meta-analysis is impeded, due to lack of information availability. Details of the experiment IDs investigated for AD and Lung Cancer, along with compliant scores is provided in [Supplementary File S1](#).

Retrieval and prioritization of indication specific studies from GEO and ArrayExpress

Retrieval of experiment IDs using a keyword search (*cf.* Data Retrieval section) also acquires false positive experiments. Any non-disease specific experiment performed by an author named 'Alzheimer' is also retrieved when searching for AD specific experiments. For example, E-MTAB-2584 aims to investigate neuronal gp130 regulation in mechanonociception but was retrieved for AD since one of its author's name is Alzheimer. Moreover, we also obtained experiments for related diseases such as Epilepsy, or Breast Cancer, when querying for AD. For example, GSE6771, and GSE6773 are Epilepsy studies; GSE33500 belongs to Nasu Hakola Disease; all these studies were retrieved when queried for Alzheimer. Incorrect organism specificity was also noticed during prioritization. For example, GSE5281 was retrieved as rat study although it

belonged to human. Similarly, GSE2866 was retrieved as mouse study but it belonged to zebra fish. Although incorrectly identified studies are not too high, this still indicates the need to include organism and disease specificity filter during prioritization. Additionally, we manually identified a few experiments that were not retrieved using these keywords, which were also included in the database.

Further on, just by applying these two filter criteria does not assure that all retained experiments were specific to AD. For example, there could be some experiments that aim at a certain pathway that are also relevant in the area of neurodegeneration, but the experiment submitted to the repository does not deal with AD pathology. As a consequence, additional disease relevancy conditions were included before prioritization (*cf.* Experiment Prioritization section). An overview of all the retrieved AD experiments, categorized to one of the priority classes is shown in [Figure 6](#). In addition, a list of priority 1 experiments (for human, mouse and rat) is provided in [Supplementary file S5](#). This figure indicates that nearly 20% of the retrieved studies are in any case not related to AD. On the other hand, to identify the remaining 80% of the experiments (prioritized as 1 and 2) we need massive manual filtering by trained personnel. Only if the archives take an initiative to apply such a structured classification for all uploaded experiments, individual time-cost can be reduced to a greater extent.

Some experiments contain cell lines or other disease samples in addition to Alzheimer's patient samples. Experiment GSE26927 additionally contain samples from patients suffering from Parkinson's disease, multiple sclerosis, etc. To be able to query only AD related samples for integrative

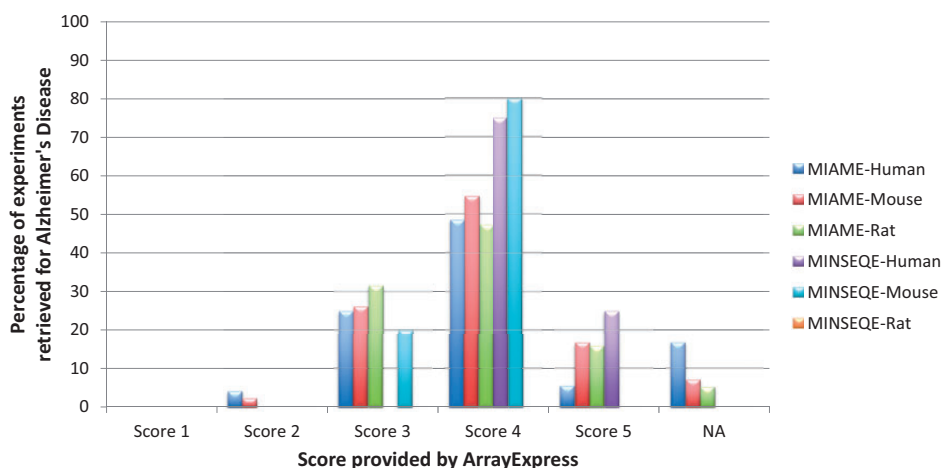


Figure 5. Distribution of MIAME and MINSEQE scores for all automatically retrieved Alzheimer's Disease gene expression experiments in ArrayExpress Database (for human, mouse and rat), as of December 2014. Percentage is calculated as (total number of AD experiments with a certain score)/(total number of AD experiments). 'NA' are the experiments which were not present in ArrayExpress. These scores reflect adherence to compliance standards by the data submitters, needed for re-investigation and reproducibility. It is observed that large percentage of experiments fall under score 4, shows that the required minimum information is still incomplete. The list of experiment IDs along with their associated scores, used for generating this statistics are provided in [Supplementary File S1](#).

analysis, we additionally included priority information at sample level. For example, we tagged Alzheimer's disease samples to AD1 whereas multiple sclerosis samples to MS1. Please refer to the README.txt file for various priority notations used.

Metadata curation

The underlying metadata information for any gene expression study has been underrepresented and thus is largely under-utilized. To perform large-scale analysis, associated annotations are of utmost importance. With the availability of detailed annotation information, one is capable of selecting studies that focus on a particular attribute, such as stage or gender. Each priority class has a specific set of fields for curation; some fields are organism dependent. After prioritization of experiments (*cf.* Experiment Prioritization section), we expect to have ~100% coverage of essential clinical and relational parameters during manual metadata curation for priority 1 studies. For example, age, gender, phenotype and stage are basic experimental variables for human studies. Additionally, in case of animal models, mouse and rat strain names are important for translational pipelines, as some strains are highly specific models for human NDD while others not (38). Irrespective of the organisms, samples mapped to their corresponding raw file identifiers are vital for running large-scale analysis. However, as shown in Figure 7, this does not hold true for human studies. From Figure 7, it is evident that even after performing thorough curation, we cannot achieve 100% in capturing information for these five basic metadata fields, a fact that is largely due to patient data privacy regulations. Similar is the case with mouse and rat information, see Supplementary Figure S6. Moreover, information related to animal models are much more scarce, obstructing

automated retrieval. Hence, manual curation accuracy is highly dependent on information availability, as curators cannot harvest information for annotation fields that are not available. On the contrary, the level of detail also depends on the type or aim of the experiment carried out. The authors and database owners obviously need to focus on the qualitative aspect of the experimental information, especially the phenotype of the sample, to allow normalized access for beginners, with standard prose, in order to support a robust computational analysis across all studies in ArrayExpress and GEO.

We selected five of the most common metadata fields (common to any disease domain such as age, gender, phenotype, stage and raw filename) and carried out a trend analysis of information availability versus time. Figure 8 (A) shows the trend over time for the metadata information provided in the archives versus the number of annotation fields that can be harvested after manual curation for human AD priority 1 experiments. Although a bit obscure, we can observe that the level of information submitted to the databases remains almost stable in the last decade (between 2 and 4 metadata fields). Moreover, with manual curation support, we were able to capture the majority of the remaining metadata from associated publications, Figure 8 (B) shows the shift in the mean value of the metadata availability. However, the trend is recently declining since the authors submit relatively lesser level of detailed information than in former times in the associated publications.

The incompleteness of metadata annotations contributed to a substantial increase in curation workload through an increased need for publication reading. This leads to a steep increase of the cost of the trained personal for curation. Overall, for the prioritization and metadata curation of AD gene expression studies, we spent about 1 year of four biocurators effort (working 10 h/week). This does not

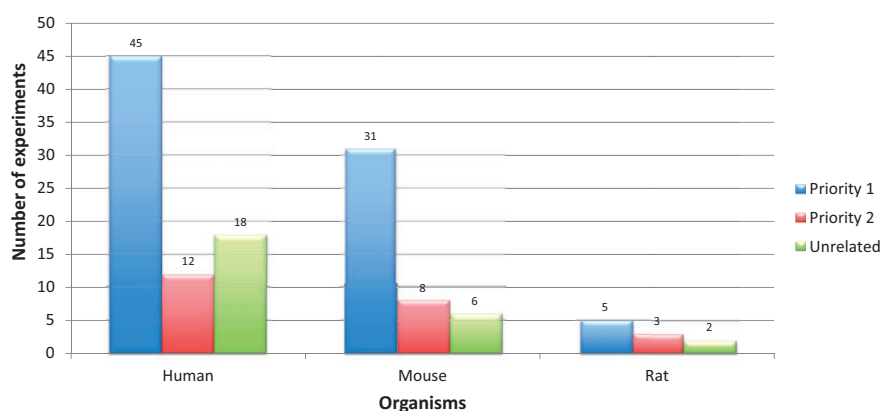


Figure 6. Priority classification statistics for Alzheimer's disease gene expression experiments retrieved from ArrayExpress and GEO (for human, mouse and rat). Alzheimer's disease experiments were retrieved using keywords. Applying the Experiment Prioritization guidelines, they were manually classified to one of the priority classes. Among them, 20% of the experiments were not related to Alzheimer's disease. The digits on the bars represent number of experiments.

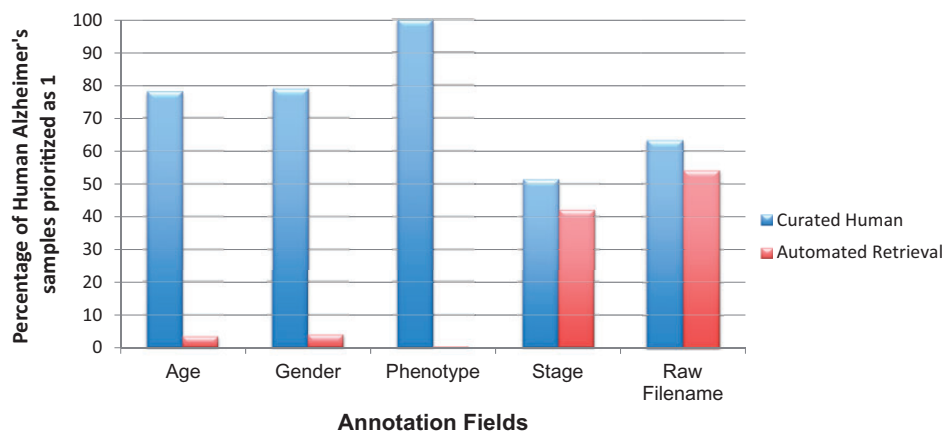


Figure 7. Coverage of basic metadata annotation fields for human AD priority 1 samples with automated retrieval and manual curation. Automated retrieval involved downloading the metadata information from ArrayExpress and GEO, programmatically. For missing meta-annotations, we applied manual curation step to harvest information from the published articles and their associated Supplementary materials. It is clear from the above statistics that manual curation accuracy for basic annotations, such as patient's clinical manifestations, and raw file information, is highly dependent on data availability.

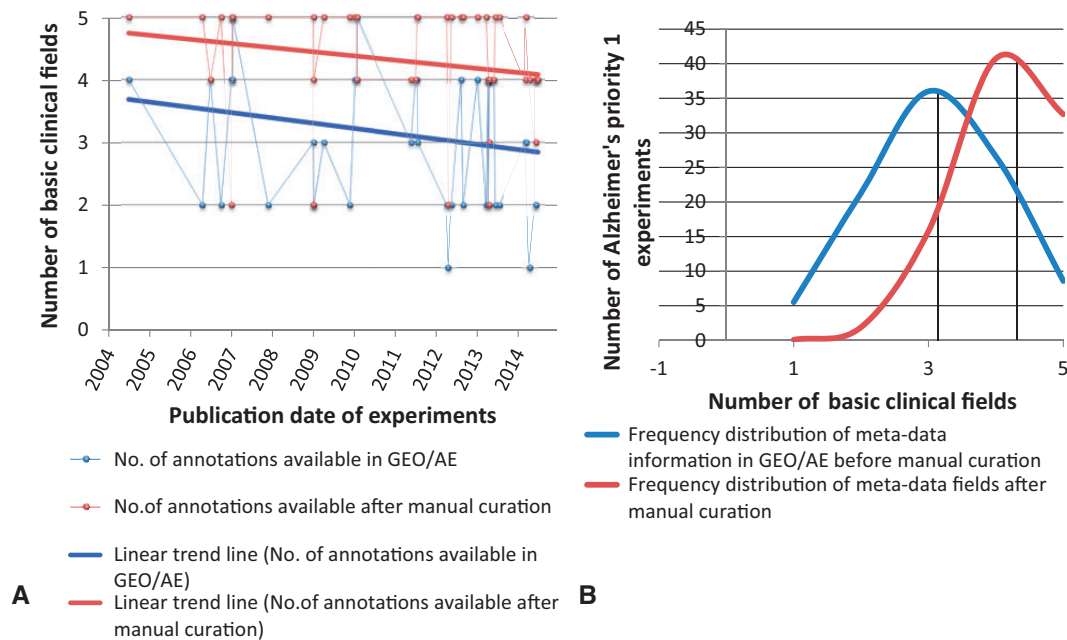


Figure 8. Frequency distribution and Trend Analysis of human priority 1 Alzheimer's disease gene-expression experiments for availability of five basic annotation fields in GEO/ArrayExpress sample page versus manual curation. The five basic annotations considered here are age, gender, stage, phenotype and raw filename. (A) Red and blue line represents the linear trend analysis of the availability of meta-annotations for experiments (represented as dots) over years, which has declined. (B) The black line represents mean value of the number of annotation fields filled. It is evident from the shift in mean of the distribution analysis that manual curation plays a very important role in capturing the missing metadata information.

include the expert's effort, who constantly provided guidance and monitored the curation work during the same duration.

Accessing *NeuroTransDB*

Metadata annotations for priority 1 AD gene expression studies for human, mouse and rat organisms, from GEO

and ArrayExpress, are stored as MySQL tables separately; downloadable as dump files at Fraunhofer SCAI File Transfer Protocol (FTP) website: <http://www.scai.fraunhofer.de/NeuroTransDB.html>. Please refer to the README.txt for details of how to install and use MySQL dumps. Additionally, these tables are provided as Excel files to allow users to use the curated information in their preferred tools/interface. Currently, the data is in its non-

normalized form. Normalized data, tagged with standard ontologies (*cf.* Normalization of Metadata Annotations section), will be made available through the AETIONOMY Knowledge Base. Currently, we have provided human priority 1 studies normalized using our internal binning scheme. Half yearly updates are planned. Our ultimate goal is to make *NeuroTransDB* a comprehensive resource for researchers working on large-scale meta-analysis in the field of neurodegenerative diseases.

Conclusion and future directions

NeuroTransDB fills the gap for large-scale meta-analysis on publicly available gene-expression studies in the field of neurodegeneration. It joins bits of missing metadata information, scattered in public archives and associated publications, into a consistent, easily accessible and regularly updated data resource. Additionally, in this paper, we have systematically specified key issues encountered during selection of relevant gene expression studies from public archives, along with their associated metadata information. We observed a huge lack of structured metadata in these archives, hampering automated large-scale reusability on a usable level of abstraction. We present here recommendations, as guidelines, for prioritizing relevant studies and a step-by-step protocol for metadata curation. The challenges faced in the course of the development of these guidelines have been pointed out, and the huge manual effort has been made explicit.

The work presented here has listed metadata fields, which have been generated based on disease expert consultation. They are highly important for choosing the right subsets of expression studies to answer complex biological questions underlying a diseased pathology. Some additional fields are included for animal models studies to allow maximal use for translational research. For all the manually curated fields, we describe normalization strategies in an attempt to provide standards for more robust automated querying and interoperability. Our results show the amount of information that is scattered in various resources, requiring extensive manual effort to capture the same. Additionally, we report that even with comprehensive manual harvesting, we were not able to capture 100% of information to fill for the basic annotation fields. We demonstrate convincingly that data availability depends largely on the meticulousness of the submitters. Additionally, it also depends on the aim of the experiment carried out. On an average, considering all the retrieved AD experiments, the submitters provide about 60% of the most basic metadata information. The outlined guidelines could be of significant value to other researchers working on gene-expression studies. The described key issues we faced during

such a curation work could influence the data submission and data storage architecture of public repositories.

Subsequently, we plan to extend the curation pipeline to other NDD diseases namely, Huntington's disease. A more gene-expression specific ontology will be built based on the curated annotations for selecting a subset of studies for meta-analyses. Although, microarray studies are the major contributors to the public repositories, RNA-Seq data are rapidly growing. We comprehend that it will be necessary for us to identify all the relevant RNA-Seq studies, since their large storage space has contributed to disperse nature of the available raw data.

Supplementary Data

Supplementary data are available at *Database* Online.

Acknowledgements

We thank Dieter Scheller, whilst at UCB Pharma, for contributions as a disease expert. We are also grateful to Jonathan van Eyll for his inputs. We acknowledge Nidhi Singh for her contribution as a biocurator during the early stages of the project. The authors thank Dr. Erfan Younesi for suggesting various ontologies used during the normalization process. We are indebted to Jasmin Zohren, Anandhi Iyappan and Jiali Wang for their initial work on curation in gene-expression studies, the outcome of which guided us to develop the more robust workflow described in this manuscript. We additionally thank the researchers, who deposit experimental data for public use. The authors are grateful for the comments and valuable recommendations from two anonymous reviewers.

Funding

German Federal Ministry for Education and Research (BMBF) within the BioPharma initiative 'Neuroallianz', project D10 'In Silico Discovery for putative Biomarkers' (grant number: 1616060B); UCB Pharma GmbH (Monheim, Germany).

Conflict of interest. None declared.

References

1. Johnson,R., Noble,W., Tartaglia,G.G. *et al.* (2012) Neurodegeneration as an RNA disorder. *Prog. Neurobiol.*, **99**, 293–315.
2. Alzheimer's Association. (2014) Alzheimer's disease facts and figures. *Alzheimer's Dement*, **10**, e47–e92.
3. Herrup,K., Carrillo,M.C., Schenk,D. *et al.* (2013) Beyond amyloid: Getting real about nonamyloid targets in Alzheimer's disease. *Alzheimers Dement*, **9**, 452–458.e1.
4. Leidinger,P., Backes,C., Deutscher,S. *et al.* (2013) A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.*, **14**, R78.
5. Greco,I., Day,N., Riddoch-Contreras,J. *et al.* (2012) Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation. *J. Transl. Med.*, **10**, 217.
6. Jedynak,B.M., Lang,A., Liu,B. *et al.* (2012) A computational neurodegenerative disease progression score: method and results

- with the Alzheimer's disease Neuroimaging Initiative cohort. *Neuroimage*, **63**, 1478–1486.
7. Mayburd, A. and Baranova, A. (2013) Knowledge-based compact disease models identify new molecular players contributing to early-stage Alzheimer's disease. *BMC Syst. Biol.*, **7**, 121.
 8. Stokes, M.E., Barmada, M.M., Kamboh, M.I. *et al.* (2014) The application of network label propagation to rank biomarkers in genome-wide Alzheimer's data. *BMC Genomics*, **15**, 282.
 9. Iyappan, A., Bagewadi, S., Page, M. *et al.* (2014) NeuroRDF: semantic data integration strategies for modeling neurodegenerative diseases. In: *Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM2014)*. Aveiro, Portugal. pp. 11–8.
 10. Lappalainen, T. and Sammeth, M. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*.
 11. Cheng, W.-C., Tsai, M.-L., Chang, C.-W. *et al.* (2010) Microarray meta-analysis database (M(2)DB): a uniformly pre-processed, quality controlled, and manually curated human clinical microarray database. *BMC Bioinformatics*, **11**, 421.
 12. Atz, M., Walsh, D., Cartagena, P. *et al.* (2007) Methodological considerations for gene expression profiling of human brain. *J. Neurosci. Methods*, **163**, 295–309.
 13. Monoranu, C.M., Apfelbacher, M., Grünblatt, E. *et al.* (2009) measurement as quality control on human post mortem brain tissue: a study of the BrainNet Europe consortium. *Neuropathol. Appl. Neurobiol.*, **35**, 329–337.
 14. American Cancer Society (2014). Cancer Facts and Figures.
 15. Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
 16. Rocca-Serra, P., Brazma, A., Parkinson, H. *et al.* (2003) ArrayExpress: a public database of gene expression data at EBI. *C. R. Biol.*, **326**:1075–1078.
 17. Malone, J., Holloway, E., Adamusiak, T. *et al.* (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**:1112–1118.
 18. Brazma, A., Hingamp, P., Quackenbush, J. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
 19. Bisognin, A., Coppe, A., Ferrari, F. *et al.* (2009) A-MADMAN: annotation-based microarray data meta-analysis tool. *BMC Bioinformatics*, **10**, 201.
 20. Piwowar, H. and Chapman, W. (2010) Recall and bias of retrieving gene expression microarray datasets through PubMed identifiers. *J. Biomed. Discov. Collab.*, **5**, 7–20.
 21. Ramasamy, A., Mondry, A., Holmes, C.C. *et al.* (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.*, **5**, 1320–1332.
 22. Ober, C., Loisel, D.A. and Gilad, Y. (2008) Sex-specific genetic architecture of human disease. *Nat. Rev. Genet.*, **9**, 911–922.
 23. Li, J.Z., Vawter, M.P., Walsh, D.M. *et al.* (2004) Systematic changes in gene expression in postmortem human brains associated with tissue pH and terminal medical conditions. *Hum. Mol. Genet.*, **13**, 609–616.
 24. Zheng, J., Stoyanovich, J., Manduchi, E. *et al.* (2011) AnnotCompute: annotation-based exploration and meta-analysis of genomics experiments. *Database*, **2011**, 1–14.
 25. Ivliev, A.E., 't Hoen, P.C., Villerius, M.P. *et al.* (2008) Microarray retriever: a web-based tool for searching and large scale retrieval of public microarray data. *Nucleic Acids Res.*, **36**, 327–331.
 26. Zhu, Y., Davis, S., Stephens, R. *et al.* (2008) GEOmetadb: Powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*, **24**, 2798–800.
 27. Faith, J.J., Driscoll, M.E., Fusaro, V. *et al.* (2008) Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, 866–870.
 28. Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V. *et al.* (2007) OncoPrint 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180.
 29. Buckberry, S., Bent, S.J., Bianco-miotto, T. *et al.* (2014) massiR: a method for predicting the sex of samples in gene expression microarray datasets. *Bioinformatics.*, **30**, 2084–5. doi:10.1093/bioinformatics/btu161.
 30. Coletta, A., Molter, C., Duqué, R. *et al.* (2012) InSilico DB genomic datasets hub: an efficient starting point for analyzing genome-wide studies in GenePattern, Integrative Genomics Viewer, and R/Bioconductor. *Genome Biol.*, **13**, R104.
 31. Burge, S., Attwood, T.K., Bateman, A. *et al.* (2012) Biocurators and biocuration: surveying the 21st century challenges. *Database (Oxford)*, **2012**, bar059.
 32. Glass D, Viñuela A, Davies MN, *et al.* Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome biology* 2013;**14**:R75. doi:10.1186/gb-2013-14-7-r75
 33. Holland D, Desikan RS, Dale AM, *et al.* Rates of decline in Alzheimer disease decrease with age. *PLoS one* 2012;**7**:e42325. doi:10.1371/journal.pone.0042325
 34. Bernick C, Cummings J, Raman R, *et al.* Age and rate of cognitive decline in Alzheimer disease: implications for clinical trials. *Archives of neurology* 2012;**69**:901–5. doi:10.1001/archneurol.2011.3758
 35. Mattsson N, Rosén E, Hansson O, *et al.* Age and diagnostic performance of Alzheimer disease CSF biomarkers. *Neurology* 2012;**78**:468–76. doi:10.1212/WNL.0b013e3182477eed
 36. Carter CL, Resnick EM, Mallampalli M, *et al.* Sex and gender differences in Alzheimer's disease: recommendations for future research. *Journal of women's health (2002)* 2012;**21**:1018–23. doi:10.1089/jwh.2012.3789
 37. Vest RS, Pike CJ. Gender, sex steroid hormones, and Alzheimer's disease. *Hormones and behavior* 2013;**63**:301–7. doi:10.1016/j.yhbeh.2012.04.006
 38. Mirnics K, Pevsner J. Progress in the use of microarray technology to study the neurobiology of disease. *Nature neuroscience* 2004;**7**:434–9. doi:10.1038/nn1230
 39. Knöchel C, Oertel-Knöchel V, O'Dwyer L, *et al.* Cognitive and behavioural effects of physical exercise in psychiatric patients. *Progress in neurobiology* 2012;**96**:46–68. doi:10.1016/j.pneurobio.2011.11.007
 40. Cummings JL, Schneider E, Tariot PN, *et al.* Behavioral effects of memantine in Alzheimer disease patients receiving donepezil treatment. *Neurology* 2006;**67**:57–63. doi:10.1212/01.wnl.0000223333.42368.f1
 41. Nacmias B. Genetics of familial and sporadic Alzheimer's disease. *Frontiers in Bioscience* 2013;**E5**:167. doi:10.2741/E605

42. Braak H, Braak E. Neuropathological staging of Alzheimer-related changes. *Acta Neuropathologica* 1991;82:239–59. doi:10.1007/BF00308809
43. Helmer C. Mortality with Dementia: Results from a French Prospective Community-based Cohort. *American Journal of Epidemiology* 2001;154:642–8. doi:10.1093/aje/154.7.642
44. Solomon A, Dobranici L, Kåreholt I, et al. Comorbidity and the rate of cognitive decline in patients with Alzheimer dementia. *International journal of geriatric psychiatry* 2011;26:1244–51. doi:10.1002/gps.2670
45. Hiltunen M, Bertram L, Saunders AJ. Genetic risk factors: their function and comorbidities in Alzheimer's disease. *International journal of Alzheimer's disease* 2011;2011:925362. doi:10.4061/2011/925362
46. Sherwood KR, Head MW, Walker R, et al. RNA integrity in post mortem human variant Creutzfeldt-Jakob disease (vCJD) and control brain tissue. *Neuropathology and applied neurobiology* 2011;37:633–42. doi:10.1111/j.1365-2990.2011.01162.x
47. Durrenberger PF, Fernando S, Kashefi SN, et al. Effects of ante-mortem and postmortem variables on human brain mRNA quality: a BrainNet Europe study. *Journal of neuropathology and experimental neurology* 2010;69:70–81. doi:10.1097/NEN.0b013e3181c7e32f
48. Koppelkamm A, Vennemann B, Lutz-Bonengel S, et al. RNA integrity in post-mortem samples: influencing parameters and implications on RT-qPCR assays. *International journal of legal medicine* 2011;125:573–80. doi:10.1007/s00414-011-0578-1
49. Stan AD, Ghose S, Gao X-M, et al. Human postmortem tissue: what quality markers matter? *Brain research* 2006;1123:1–11. doi:10.1016/j.brainres.2006.09.025
50. Li JZ, Vawter MP, Walsh DM, et al. Systematic changes in gene expression in postmortem human brains associated with tissue pH and terminal medical conditions. *Human molecular genetics* 2004;13:609–16. doi:10.1093/hmg/ddh065
51. Long JM, Lahiri DK. MicroRNA-101 downregulates Alzheimer's amyloid- β precursor protein levels in human cell cultures and is differentially expressed. *Biochemical and biophysical research communications* 2011;404:889–95. doi:10.1016/j.bbrc.2010.12.053
52. Ly PTT, Wu Y, Zou H, et al. Inhibition of GSK3 β -mediated BACE1 expression reduces Alzheimer-associated phenotypes. *The Journal of clinical investigation* 2013;123:224–35. doi:10.1172/JCI64516
53. Blalock EM, Buechel HM, Popovic J, et al. Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease. *Journal of chemical neuroanatomy* 2011;42:118–26. doi:10.1016/j.jchemneu.2011.06.007
54. Ray M, Zhang W. Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks. *BMC systems biology* 2010;4:136. doi:10.1186/1752-0509-4-136
55. Grolla AA, Sim JA, Lim D, et al. Amyloid- β and Alzheimer's disease type pathology differentially affects the calcium signalling toolkit in astrocytes from different brain regions. *Cell death & disease* 2013;4:e623. doi:10.1038/cddis.2013.145
56. Miller JA, Woltjer RL, Goodenbour JM, et al. Genes and pathways underlying regional and cell type changes in Alzheimer's disease. *Genome medicine* 2013;5:48. doi:10.1186/gm452
57. Roed L, Grave G, Lindahl T, et al. Prediction of mild cognitive impairment that evolves into Alzheimer's disease dementia within two years using a gene expression signature in blood: a pilot study. *Journal of Alzheimer's disease: JAD* 2013;35:611–21. doi:10.3233/JAD-122404
58. Kiko T, Nakagawa K, Tsuduki T, et al. MicroRNAs in plasma and cerebrospinal fluid as potential markers for Alzheimer's disease. *Journal of Alzheimer's disease: JAD* 2014;39:253–9. doi:10.3233/JAD-130932
59. Koehler NKU, Stransky E, Shing M, et al. Altered serum IgG levels to α -synuclein in dementia with Lewy bodies and Alzheimer's disease. *PLoS one* 2013;8:e64649. doi:10.1371/journal.pone.0064649
60. Bachstetter A, Webster S, Van Eldik L. Traumatic brain injury in a mouse model of Alzheimer's disease leads to persistent glial activation, chronic proinflammatory phenotype, and increased cognitive deficits. *Alzheimer's & Dementia* 2014;10:P337. doi:10.1016/j.jalz.2014.05.333
61. Washington PM, Morffy N, Parsadian M, et al. Experimental traumatic brain injury induces rapid aggregation and oligomerization of amyloid-beta in an Alzheimer's disease mouse model. *Journal of neurotrauma* 2014;31:125–34. doi:10.1089/neu.2013.3017
62. Szczodry O, Van der Staay FJ, Arndt SS. Modelling Alzheimer-like cognitive deficits in rats using biperiden as putative cognition impairer. *Behavioural brain research* 2014;274:307–11. doi:10.1016/j.bbr.2014.08.036
63. Marlatt MW, Potter MC, Bayer TA, et al. Prolonged running, not fluoxetine treatment, increases neurogenesis, but does not alter neuropathology, in the 3xTg mouse model of Alzheimer's disease. *Current topics in behavioral neurosciences* 2013;15:313–40. doi:10.1007/7854_2012_237
64. Claxton A, Baker LD, Wilkinson CW, et al. Sex and ApoE genotype differences in treatment response to two doses of intranasal insulin in adults with mild cognitive impairment or Alzheimer's disease. *Journal of Alzheimer's disease: JAD* 2013;35:789–97. doi:10.3233/JAD-122308
65. Ho SW, Tsui YTC, Wong TT, et al. Effects of 17-allylamino-17-demethoxygeldanamycin (17-AAG) in transgenic mouse models of frontotemporal lobar degeneration and Alzheimer's disease. *Translational neurodegeneration* 2013;2:24. doi:10.1186/2047-9158-2-24
66. Chin J. Selecting a mouse model of Alzheimer's disease. *Methods in molecular biology (Clifton, NJ)* 2011;670:169–89. doi:10.1007/978-1-60761-744-0_13
67. James D, Kang S, Park S. Injection of β -amyloid into the hippocampus induces metabolic disturbances and involuntary weight loss which may be early indicators of Alzheimer's disease. *Aging clinical and experimental research* 2014;26:93–8. doi:10.1007/s40520-013-0181-z
68. Hassel B, Taubøll E, Shaw R, et al. Region-specific changes in gene expression in rat brain after chronic treatment with levetiracetam or phenytoin. *Epilepsia* 2010;51:1714–20. doi:10.1111/j.1528-1167.2010.02545.x

Bibliography

[Allen et al., 2016] Allen, G. I., Amoroso, N., Anghel, C., Balagurusamy, V., Bare, C. J., Beaton, D., Bellotti, R., Bennett, D. A., Boehme, K. L., Boutros, P. C., Caberlotto, L., Caloian, C., Campbell, F., Chaibub Neto, E., Chang, Y. C., Chen, B., Chen, C. Y., Chien, T. Y., Clark, T., Das, S., Davatzikos, C., Deng, J., Dillenberger, D., Dobson, R. J., Dong, Q., Doshi, J., Duma, D., Errico, R., Erus, G., Everett, E., Fardo, D. W., Friend, S. H., Fröhlich, H., Gan, J., St George-Hyslop, P., Ghosh, S. S., Glaab, E., Green, R. C., Guan, Y., Hong, M. Y., Huang, C., Hwang, J., Ibrahim, J., Inglese, P., Iyappan, A., Jiang, Q., Katsumata, Y., Kauwe, J. S., Klein, A., Kong, D., Krause, R., Lalonde, E., Lauria, M., Lee, E., Lin, X., Liu, Z., Livingstone, J., Logsdon, B. A., Lovestone, S., Ma, T. W., Malhotra, A., Mangravite, L. M., Maxwell, T. J., Merrill, E., Nagorski, J., Namasivayam, A., Narayan, M., Naz, M., Newhouse, S. J., Norman, T. C., Nurtdinov, R. N., Oyang, Y. J., Pawitan, Y., Peng, S., Peters, M. A., Piccolo, S. R., Praveen, P., Priami, C., Sabelnykova, V. Y., Senger, P., Shen, X., Simmons, A., Sotiras, A., Stolovitzky, G., Tangaro, S., Tateo, A., Tung, Y. A., Tustison, N. J., Varol, E., Vradenburg, G., Weiner, M. W., Xiao, G., Xie, L., Xie, Y., Xu, J., Yang, H., Zhan, X., Zhou, Y., Zhu, F., Zhu, H., and Zhu, S. (2016). Crowdsourced estimation of cognitive decline and resilience in Alzheimer’s disease. *Alzheimer’s and Dementia*, 12(6):645–653.

[Andreasson et al., 2016] Andreasson, K. I., Bachstetter, A. D., Colonna, M., Ginhoux, F., Holmes, C., Lamb, B., Landreth, G., Lee, D. C., Low, D., Lynch, M. A., Monsonogo, A., O’Banion, M. K., Pekny, M., Puschmann, T., Russek-Blum, N., Sandusky, L. A., Selenica, M. L. B., Takata, K., Teeling, J., Town, T., and Van El-

- dik, L. J. (2016). Targeting innate immunity for neurodegenerative disorders of the central nervous system.
- [Athey et al., 2013] Athey, B. D., Braxenthaler, M., Haas, M., and Guo, Y. (2013). tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2013:6–8.
- [Attwood et al., 2009] Attwood, T., Kell, D., McDermott, P., Marsh, J., Pettifer, S., and Thorne, D. (2009). Calling International Rescue: knowledge lost in literature and data landslide! *Biochemical Journal*, 424(3):317–333.
- [Auffray et al., 2016] Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., Bernal-Delgado, E., Blomberg, N., Bock, C., Conesa, A., Del Signore, S., Delogne, C., Devilee, P., Di Meglio, A., Eijkemans, M., Flicek, P., Graf, N., Grimm, V., Guchelaar, H. J., Guo, Y. K., Gut, I. G., Hanbury, A., Hanif, S., Hilgers, R. D., Honrado, ., Hose, D. R., Houwing-Duistermaat, J., Hubbard, T., Janacek, S. H., Karanikas, H., Kievits, T., Kohler, M., Kremer, A., Lanfear, J., Lengauer, T., Maes, E., Meert, T., Müller, W., Nickel, D., Oledzki, P., Pedersen, B., Petkovic, M., Pliakos, K., Rattray, M., i Màs, J. R., Schneider, R., Sengstag, T., Serra-Picamal, X., Spek, W., Vaas, L. A., van Batenburg, O., Vandelaer, M., Varnai, P., Villoslada, P., Vizcaíno, J. A., Wubbe, J. P. M., and Zanetti, G. (2016). Making sense of big data in health research: Towards an EU action plan. *Genome Medicine*, 8(1).
- [Bader et al., 2006] Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic acids research*, 34(Database issue):504–6.
- [Bagewadi et al., 2015] Bagewadi, S., Adhikari, S., Dhrangadhariya, A., Irin, A. K., Ebeling, C., Namasivayam, A. A., Page, M., Hofmann-Apitius, M., and Senger, P. (2015). NeuroTransDB : highly curated and structured transcriptomic metadata for neurodegenerative diseases. *Database*, 2015:bav099.
- [Baker, 2016] Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454.

- [Bauer et al., 2016] Bauer, C. R., Knecht, C., Fretter, C., Baum, B., Jendrossek, S., Rühlemann, M., Heinsen, F.-A., Umbach, N., Grimbacher, B., Franke, A., Lieb, W., Krawczak, M., Hütt, M.-T., and Sax, U. (2016). Interdisciplinary approach towards a systems medicine toolbox using the example of inflammatory diseases. *Briefings in Bioinformatics*, page bbw024.
- [Begley and Ellis, 2012] Begley, C. G. and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533.
- [Bellary et al., 2014] Bellary, S., Krishnankutty, B., and Latha, M. S. (2014). Basics of case report form designing in clinical research. *Perspectives in clinical research*, 5(4):159–66.
- [Belleau et al., 2008] Belleau, F., Nolin, M. A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706–716.
- [Berwick and Harvey, 2014] Berwick, D. C. and Harvey, K. (2014). The regulation and deregulation of Wnt signaling by PARK genes in health and disease. *Journal of Molecular Cell Biology*, 6(1):3–12.
- [Bodenreider, 2008] Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of medical informatics*, 3841:67–79.
- [Bonnet et al., 2015] Bonnet, E., Viara, E., Kuperstein, I., Calzone, L., Cohen, D. P., Barillot, E., and Zinovyev, A. (2015). NaviCell Web Service for network-based data visualization. *Nucleic Acids Research*.
- [Brosseron et al., 2014] Brosseron, F., Krauthausen, M., Kummer, M., and Heneka, M. T. (2014). Body Fluid Cytokine Levels in Mild Cognitive Impairment and Alzheimers Disease: a Comparative Overview.
- [Büchel et al., 2012] Büchel, F., Wrzodek, C., Mittag, F., Dräger, A., Eichner, J., Rodriguez, N., Le Novère, N., and Zell, A. (2012). Qualitative translation of relations from BioPAX to SBML qual. *Bioinformatics*, 28(20):2648–2653.

- [Cai et al., 2016] Cai, Y., Arikath, J., Yang, L., Guo, M. L., Periyasamy, P., and Buch, S. (2016). Interplay of endoplasmic reticulum stress and autophagy in neurodegenerative disorders.
- [Calderone et al., 2013] Calderone, A., Castagnoli, L., and Cesareni, G. (2013). Menta: A resource for browsing integrated protein-interaction networks.
- [Calderone et al., 2016] Calderone, A., Formenti, M., Aprea, F., Papa, M., Alberghina, L., Colangelo, A. M., and Bertolazzi, P. (2016). Comparing Alzheimers and Parkinsons diseases networks using graph communities structure. *BMC Systems Biology*, 10(1):25.
- [Canuel et al., 2015] Canuel, V., Rance, B., Avillach, P., Degoulet, P., and Burgun, A. (2015). Translational research platforms integrating clinical and omics data: A review of publicly available solutions. *Briefings in Bioinformatics*, 16(2):280–290.
- [Caron et al., 2010] Caron, E., Ghosh, S., Matsuoka, Y., Ashton-Beaucage, D., Therrien, M., Lemieux, S., Perreault, C., Roux, P. P., and Kitano, H. (2010). A comprehensive map of the mTOR signaling network.
- [Castaneda et al., 2015] Castaneda, C., Nalley, K., Mannion, C., Bhattacharyya, P., Blake, P., Pecora, A., Goy, A., and Suh, K. S. (2015). Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of Clinical Bioinformatics*, 5(1):4.
- [Cerami et al., 2012] Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., and Schultz, N. (2012). The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404.
- [Comber et al., 2006] Comber, A. J., Fisher, P. F., Harvey, F., Gahegan, M., and Wadsworth, R. (2006). Using metadata to link uncertainty and data quality assessments. In *Progress in Spatial Data Handling - 12th International Symposium on Spatial Data Handling, SDH 2006*, pages 279–292.

- [Conesa and Mortazavi, 2014] Conesa, A. and Mortazavi, A. (2014). The common ground of genomics and systems biology.
- [Corrêa and Eales, 2012] Corrêa, S. A. L. and Eales, K. L. (2012). The Role of p38 MAPK and Its Substrates in Neuronal Plasticity and Neurodegenerative Disease. *Journal of Signal Transduction*, 2012:1–12.
- [Cowie et al., 2017] Cowie, M. R., Blomster, J. I., Curtis, L. H., Duclaux, S., Ford, I., Fritz, F., Goldman, S., Janmohamed, S., Kreuzer, J., Leenay, M., Michel, A., Ong, S., Pell, J. P., Southworth, M. R., Stough, W. G., Thoenes, M., Zannad, F., and Zalewski, A. (2017). Electronic health records to facilitate clinical research.
- [Croft et al., 2011] Croft, D., O’Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., MayMahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D’Eustachio, P., and Stein, L. (2011). Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(SUPPL. 1).
- [De Ferrari et al., 2007] De Ferrari, G. V., Papassotiropoulos, A., Biechele, T., Wavrant De-Vrieze, F., Avila, M. E., Major, M. B., Myers, A., Sáez, K., Henríquez, J. P., Zhao, A., Wollmer, M. A., Nitsch, R. M., Hock, C., Morris, C. M., Hardy, J., and Moon, R. T. (2007). Common genetic variation within the low-density lipoprotein receptor-related protein 6 and late-onset Alzheimer’s disease. *Proceedings of the National Academy of Sciences of the United States of America*, 104(22):9434–9.
- [Demir et al., 2010] Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D’Eustachio, P., Schaefer, C., Luciano, J., Schacherer, F., Martinez-Flores, I., Hu, Z., Jimenez-Jacinto, V., Joshi-Tope, G., Kandasamy, K., Lopez-Fuentes, A. C., Mi, H., Pichler, E., Rodchenkov, I., Splendiani, A., Tkachev, S., Zucker, J., Gopinath, G., Rajasimha, H., Ramakrishnan, R., Shah, I., Syed, M., Anwar, N., Babur, ., Blinov, M., Brauner, E., Corwin, D., Donaldson, S., Gibbons, F., Goldberg, R., Hornbeck, P., Luna, A., Murray-Rust, P., Neumann, E., Reubenacker, O., Samwald, M., Van Iersel, M., Wimalaratne, S., Allen, K., Braun, B.,

- Whirl-Carrillo, M., Cheung, K. H., Dahlquist, K., Finney, A., Gillespie, M., Glass, E., Gong, L., Haw, R., Honig, M., Hubaut, O., Kane, D., Krupa, S., Kutmon, M., Leonard, J., Marks, D., Merberg, D., Petri, V., Pico, A., Ravenscroft, D., Ren, L., Shah, N., Sunshine, M., Tang, R., Whaley, R., Letovksy, S., Buetow, K. H., Rzhetsky, A., Schachter, V., Sobral, B. S., Dogrusoz, U., McWeeney, S., Aladjem, M., Birney, E., Collado-Vides, J., Goto, S., Hucka, M., Novère, N. L., Maltsev, N., Pandey, A., Thomas, P., Wingender, E., Karp, P. D., Sander, C., and Bader, G. D. (2010). The BioPAX community standard for pathway data sharing.
- [Dickersin and Mayo-Wilson, 2018] Dickersin, K. and Mayo-Wilson, E. (2018). Standards for design and measurement would make clinical research reproducible and usable. *Proceedings of the National Academy of Sciences*, 115(11):2590–2594.
- [Ethier et al., 2018] Ethier, J.-F., McGilchrist, M., Barton, A., Cloutier, A.-M., Curcin, V., Delaney, B. C., and Burgun, A. (2018). The TRANSFoRm project: Experience and lessons learned regarding functional and interoperability requirements to support primary care. *Learning Health Systems*, 2(2):e10037.
- [Fabregat et al., 2018] Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorsler, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D’Eustachio, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*.
- [Fadaka et al., 2017] Fadaka, A., Ojo, O., Osukoya, O., Akuboh, O., and Ajiboye (2017). Role of p38 MAPK Signaling in Neurodegenerative Diseases: A Mechanistic Perspective. *Ann Neurodegener Dis*, 2(1).
- [Fluck et al., 2016] Fluck, J., Madan, S., Ansari, S., Kodamullil, A. T., Karki, R., Rastegar-Mojarad, M., Catlett, N. L., Hayes, W., Szostak, J., Hoeng, J., and Peitsch, M. (2016). Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (BEL). *Database*, 2016:baw113.
- [Fluck et al., 2015] Fluck, J., Madan, S., Effendorf, T., Mevissen, H.-T., Clematide, S., van der Lek, A., and Rinaldi, F. (2015). Track 4 Overview: Extraction of Causal

Network Information in Biological Expression Language (BEL). *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, 1:333–346.

[Freude et al., 2009] Freude, S., Schilbach, K., and Schubert, M. (2009). The role of IGF-1 receptor and insulin receptor signaling for the pathogenesis of Alzheimer’s disease: from model organisms to human disease. *Current Alzheimer research*, 6:213–223.

[Fujita et al., 2014] Fujita, K. A., Ostaszewski, M., Matsuoka, Y., Ghosh, S., Glaab, E., Trefois, C., Crespo, I., Perumal, T. M., Jurkowski, W., Antony, P. M., Diederich, N., Buttini, M., Kodama, A., Satagopam, V. P., Eifes, S., Del Sol, A., Schneider, R., Kitano, H., and Balling, R. (2014). Integrating pathways of parkinson’s disease in a molecular interaction map.

[Funahashi et al., 2008] Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., and Kitano, H. (2008). CellDesigner 3.5: A versatile modeling tool for biochemical networks. *Proceedings of the IEEE*.

[Funahashi et al., 2003] Funahashi, A., Morohashi, M., Kitano, H., and Tanimura, N. (2003). CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*.

[Galdzicki et al., 2014] Galdzicki, M., Clancy, K. P., Oberortner, E., Pocock, M., Quinn, J. Y., Rodriguez, C. A., Roehner, N., Wilson, M. L., Adam, L., Anderson, J. C., Bartley, B. A., Beal, J., Chandran, D., Chen, J., Densmore, D., Endy, D., Grünberg, R., Hallinan, J., Hillson, N. J., Johnson, J. D., Kuchinsky, A., Lux, M., Misirli, G., Peccoud, J., Plahar, H. A., Sirin, E., Stan, G. B., Villalobos, A., Wipat, A., Gennari, J. H., Myers, C. J., and Sauro, H. M. (2014). The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology.

[Gao et al., 2013] Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., and Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269).

- [Gawron et al., 2016] Gawron, P., Ostaszewski, M., Satagopam, V., Gebel, S., Mazein, A., Kuzma, M., Zorzan, S., McGee, F., Otjacques, B., Balling, R., and Schneider, R. (2016). MINERVAa platform for visualization and curation of molecular interaction networks. *npj Systems Biology and Applications*, 2(1):16020.
- [Gehring et al., 2015] Gehring, M., Muth, F., Koch, P., and Laufer, S. A. (2015). c-Jun N-terminal kinase inhibitors: a patent review (2010 - 2014). *Expert opinion on therapeutic patents*, 25(8):849–72.
- [Gendron, 2009] Gendron, T. F. (2009). The role of tau in neurodegeneration. *Molecular Neurodegeneration*, 4(1).
- [Gene Ontology Consortium, 2000] Gene Ontology Consortium (2000). The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*.
- [Gleeson et al., 2010] Gleeson, P., Crook, S., Cannon, R. C., Hines, M. L., Billings, G. O., Farinella, M., Morse, T. M., Davison, A. P., Ray, S., Bhalla, U. S., Barnes, S. R., Dimitrova, Y. D., and Silver, R. A. (2010). NeuroML: A language for describing data driven models of neurons and networks with a high degree of biological detail. *PLoS Computational Biology*, 6(6):1–19.
- [Gomez-Cabrero et al., 2014] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschen-dorff, A., Merckenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). Data integration in the era of omics: current and future challenges.
- [Gong et al., 2018] Gong, C.-X., Liu, F., and Iqbal, K. (2018). Multifactorial Hypothesis and Multi-Targets for Alzheimers Disease. *Journal of Alzheimer's Disease*, Preprint(Preprint):1–11.
- [Gray et al., 2015] Gray, K. A., Yates, B., Seal, R. L., Wright, M. W., and Bruford, E. A. (2015). Genenames.org: The HGNC resources in 2015. *Nucleic Acids Research*.

- [Greene and Troyanskaya, 2010] Greene, C. S. and Troyanskaya, O. G. (2010). Integrative systems biology for data-driven knowledge discovery. *Seminars in Nephrology*, 30(5):443–454.
- [Greene et al., 2011] Greene, L. A., Levy, O., and Malagelada, C. (2011). Akt as a victim, villain and potential hero in Parkinson’s disease pathophysiology and treatment. *Cellular and molecular neurobiology*, 31(7):969–978.
- [Groves et al., 2013] Groves, P., Kayyali, B., Knott, D., and Van Kuiken, S. (2013). The ”big data” revolution in healthcare: accelerating value and innovation. *McKinsey Global Institute*, (January):1– 22.
- [Hammond et al., 2014] Hammond, W. E., Jaffe, C., Cimino, J. J., and Huff, S. M. (2014). Standards in biomedical informatics. In *Biomedical Informatics: Computer Applications in Health Care and Biomedicine: Fourth Edition*, pages 211–253.
- [Hastings et al., 2013] Hastings, J., De Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., and Steinbeck, C. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Research*.
- [Heneka et al., 2015] Heneka, M. T., Carson, M. J., Khoury, J. E., Landreth, G. E., Brosseron, F., Feinstein, D. L., Jacobs, A. H., Wyss-Coray, T., Vitorica, J., Ransohoff, R. M., Herrup, K., Frautschy, S. A., Finsen, B., Brown, G. C., Verkhratsky, A., Yamanaka, K., Koistinaho, J., Latz, E., Halle, A., Petzold, G. C., Town, T., Morgan, D., Shinohara, M. L., Perry, V. H., Holmes, C., Bazan, N. G., Brooks, D. J., Hunot, S., Joseph, B., Deigendesch, N., Garaschuk, O., Boddeke, E., Dinarello, C. A., Breitner, J. C., Cole, G. M., Golenbock, D. T., and Kummer, M. P. (2015). Neuroinflammation in Alzheimer’s disease. *The Lancet Neurology*, 14(4):388–405.
- [Heppner et al., 2015] Heppner, F. L., Ransohoff, R. M., and Becher, B. (2015). Immune attack: The role of inflammation in Alzheimer disease.
- [Hers et al., 2011] Hers, I., Vincent, E. E., and Tavaré, J. M. (2011). Akt signalling in health and disease.

- [Herzinger et al., 2017] Herzinger, S., Gu, W., Satagopam, V., Eifes, S., Rege, K., Barbosa-Silva, A., and Schneider, R. (2017). SmartR: An open-source platform for interactive visual analytics for translational research data. In *Bioinformatics*.
- [Hetz and Saxena, 2017] Hetz, C. and Saxena, S. (2017). ER stress and the unfolded protein response in neurodegeneration.
- [Hofmann-Apitius et al., 2015a] Hofmann-Apitius, M., Alarcón-Riquelme, M. E., Chamberlain, C., and McHale, D. (2015a). Towards the taxonomy of human disease. *Nature Reviews Drug Discovery*.
- [Hofmann-Apitius et al., 2015b] Hofmann-Apitius, M., Ball, G., Gebel, S., Bagewadi, S., De Bono, B., Schneider, R., Page, M., Kodamullil, A. T., Younesi, E., Ebeling, C., Tegnér, J., and Canard, L. (2015b). Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders.
- [Hong et al., 2016] Hong, H., Kim, B. S., and Im, H. I. (2016). Pathophysiological role of neuroinflammation in neurodegenerative diseases and psychiatric disorders.
- [Horgan and Kenny, 2011] Horgan, R. P. and Kenny, L. C. (2011). Omic technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, 13(3):189–195.
- [Hoyt et al., 2017] Hoyt, C. T., Konotopez, A., and Ebeling, C. (2017). PyBEL: a Computational Framework for Biological Expression Language. *Bioinformatics*.
- [Huang et al., 2017] Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: Recent progress in multi-omics data integration methods.
- [Hucka and Finney, 2005] Hucka, M. and Finney, A. (2005). Escalating model sizes and complexities call for standardized forms of representation. *Molecular Systems Biology*, 1(1):2005.0011.
- [Hucka et al., 2003] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J.,

- Hodgman, T. C., Hofmeyr, J. H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Krelling, A., Kummer, U., Le Novère, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and Wang, J. (2003). The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.
- [Hucka et al., 2015] Hucka, M., Nickerson, D. P., Bader, G. D., Bergmann, F. T., Cooper, J., Demir, E., Garny, A., Golebiewski, M., Myers, C. J., Schreiber, F., Waltemath, D., and Le Novère, N. (2015). Promoting Coordinated Development of Community-Based Information Standards for Modeling in Biology: The COMBINE Initiative. *Frontiers in Bioengineering and Biotechnology*.
- [Hudson et al., 2018] Hudson, L. D., Kush, R. D., Navarro Almario, E., Seigneuret, N., Jackson, T., Jauregui, B., Jordan, D., Fitzmartin, R., Zhou, F. L., Malone, J. K., Galvez, J., and Becnel, L. B. (2018). Global Standards to Expedite Learning From Medical Research Data. *Clinical and Translational Science*.
- [Inestrosa and Varela-Nallar, 2014] Inestrosa, N. C. and Varela-Nallar, L. (2014). Wnt signaling in the nervous system and in Alzheimer’s disease. *Journal of Molecular Cell Biology*, 6(1):64–74.
- [Ioannidis et al., 2009] Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G. P., Petretto, E., and van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41(2):149–155.
- [Jung et al., 2003] Jung, S. S., Zhang, W., and Van Nostrand, W. E. (2003). Pathogenic A beta induces the expression and activation of matrix metalloproteinase-2 in human cerebrovascular smooth muscle cells. *Journal of neurochemistry*, 85(5):1208–15.

- [Juty et al., 2012] Juty, N., Le Novere, N., and Laibe, C. (2012). Identifiers.org and MIRIAM Registry: Community resources to provide persistent identification. *Nucleic Acids Research*, 40(D1).
- [Kanehisa et al., 2012] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*.
- [Karki et al., 2017] Karki, R., Kodamullil, A. T., and Hofmann-Apitius, M. (2017). Comorbidity Analysis between Alzheimer’s Disease and Type 2 Diabetes Mellitus (T2DM) Based on Shared Pathways and the Role of T2DM Drugs. *Journal of Alzheimer’s Disease*, 60(2):721–731.
- [Karve et al., 2016] Karve, I. P., Taylor, J. M., and Crack, P. J. (2016). The contribution of astrocytes and microglia to traumatic brain injury. *British Journal of Pharmacology*, 173(4):692–702.
- [Kim and Choi, 2010] Kim, E. K. and Choi, E.-J. (2010). Pathological roles of MAPK signaling pathways in human diseases. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1802(4):396–405.
- [Kitano, 2002] Kitano, H. (2002). Systems biology: A brief overview.
- [Kitano et al., 2005] Kitano, H., Funahashi, A., Matsuoka, Y., and Oda, K. (2005). Using process diagrams for the graphical representation of biological networks.
- [Kodamullil et al., 2015] Kodamullil, A. T., Younesi, E., Naz, M., Bagewadi, S., and Hofmann-Apitius, M. (2015). Computable cause-and-effect models of healthy and Alzheimer’s disease states and their mechanistic differential analysis. *Alzheimer’s and Dementia*.
- [Kruse et al., 2016] Kruse, C. S., Goswamy, R., Raval, Y., and Marawi, S. (2016). Challenges and Opportunities of Big Data in Health Care: A Systematic Review. *JMIR Medical Informatics*, 4(4):e38.

- [Kuperstein et al., 2015] Kuperstein, I., Bonnet, E., Nguyen, H. A., Cohen, D., Viara, E., Grieco, L., Fourquet, S., Calzone, L., Russo, C., Kondratova, M., Dutreix, M., Barillot, E., and Zinovyev, A. (2015). Atlas of Cancer Signalling Network: A systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*.
- [Kuperstein et al., 2013] Kuperstein, I., Cohen, D. P., Pook, S., Viara, E., Calzone, L., Barillot, E., and Zinovyev, A. (2013). NaviCell: A web-based environment for navigation, curation and maintenance of large molecular interaction maps. *BMC Systems Biology*.
- [Lai et al., 2016] Lai, P. T., Lo, Y. Y., Huang, M. S., Hsiao, Y. C., and Tsai, R. T. H. (2016). BelSmile: a biomedical semantic role labeling approach for extracting biological expression language from text. *Database : the journal of biological databases and curation*, 2016.
- [Lapatas et al., 2015] Lapatas, V., Stefanidakis, M., Jimenez, R. C., Via, A., and Schneider, M. V. (2015). Data integration in biological research: an overview. *Journal of Biological Research-Thessaloniki*, 22(1):9.
- [Le Novère et al., 2005] Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., Crampin, E. J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J. L., Spence, H. D., and Wanner, B. L. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM).
- [Lee and Kim, 2017] Lee, J. K. and Kim, N.-J. (2017). Recent Advances in the Inhibition of p38 MAPK as a Potential Strategy for the Treatment of Alzheimers Disease. *Molecules*, 22(8):1287.
- [Lei et al., 2010] Lei, P., Ayton, S., Finkelstein, D. I., Adlard, P. A., Masters, C. L., and Bush, A. I. (2010). Tau protein: Relevance to Parkinson's disease.
- [Leonoudakis et al., 2017] Leonoudakis, D., Rane, A., Angeli, S., Lithgow, G. J., Andersen, J. K., and Chinta, S. J. (2017). Anti-Inflammatory and Neuroprotective Role

- of Natural Product Securinine in Activated Glial Cells: Implications for Parkinson's Disease. *Mediators of Inflammation*, 2017.
- [Leroux et al., 2017] Leroux, H., Metke-Jimenez, A., and Lawley, M. J. (2017). Towards achieving semantic interoperability of clinical study data with FHIR. *Journal of Biomedical Semantics*, 8(1).
- [Lesnick et al., 2007] Lesnick, T. G., Papapetropoulos, S., Mash, D. C., Ffrench-Mullen, J., Shehadeh, L., De Andrade, M., Henley, J. R., Rocca, W. A., Ahlskog, J. E., and Maraganore, D. M. (2007). A genomic pathway approach to a complex disease: Axon guidance and Parkinson disease. *PLoS Genetics*.
- [Li et al., 2010] Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M. I., Snoep, J. L., Hucka, M., Le Novère, N., and Laibe, C. (2010). BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 4.
- [Liebermeister et al., 2009] Liebermeister, W., Krause, F., Uhlenendorf, J., Lubitz, T., and Klipp, E. (2009). SemanticSBML: a tool for annotating, checking, and merging of biochemical models in SBML format. *Nature Precedings*.
- [Llorens-Marítin et al., 2014] Llorens-Marítin, M., Jurado, J., Hernández, F., and Ávila, J. (2014). GSK-3 β , a pivotal kinase in Alzheimer disease. *Frontiers in Molecular Neuroscience*, 7.
- [Lloyd et al., 2008] Lloyd, C. M., Lawson, J. R., Hunter, P. J., and Nielsen, P. F. (2008). The CellML Model Repository. *Bioinformatics*, 24(18):2122–2123.
- [Maglott et al., 2011] Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2011). Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Research*.
- [Magrane and Consortium, 2011] Magrane, M. and Consortium, U. P. (2011). UniProt Knowledgebase: A hub of integrated protein data. *Database*.
- [Maier et al., 2011] Maier, D., Kalus, W., Wolff, M., Kalko, S. G., Roca, J., Marin de Mas, I., Turan, N., Cascante, M., Falciani, F., Hernandez, M., Villà-Freixa, J., and

- Losko, S. (2011). Knowledge management for systems biology a general and visually driven framework applied to translational medicine. *BMC Systems Biology*, 5.
- [Marchetti et al., 2013] Marchetti, B., L'Episcopo, F., Morale, M. C., Tirolo, C., Testa, N., Caniglia, S., Serapide, M. F., and Pluchino, S. (2013). Uncovering novel actors in astrocyte-neuron crosstalk in Parkinson's disease: The Wnt/ β -catenin signaling cascade as the common final pathway for neuroprotection and self-repair. *European Journal of Neuroscience*, 37(10):1550–1563.
- [Masseroli et al., 2014] Masseroli, M., Mons, B., Bongcam-Rudloff, E., Ceri, S., Kel, A., Rechenmann, F., Lisacek, F., and Romano, P. (2014). Integrated Bio-Search: Challenges and trends for the integration, search and comprehensive processing of biological information.
- [Matsuoka et al., 2010] Matsuoka, Y., Ghosh, S., Kikuchi, N., and Kitano, H. (2010). Payao: A community platform for SBML pathway model curation. *Bioinformatics*.
- [Matsuoka et al., 2013] Matsuoka, Y., Matsumae, H., Katoh, M., Einfeld, A. J., Neumann, G., Hase, T., Ghosh, S., Shoemaker, J. E., Lopes, T. J., Watanabe, T., Watanabe, S., Fukuyama, S., Kitano, H., and Kawaoka, Y. (2013). A comprehensive map of the influenza A virus replication cycle. *BMC Systems Biology*.
- [Mazein et al., 2018] Mazein, A., Ostaszewski, M., Kuperstein, I., Watterson, S., Le Novère, N., Lefaudeux, D., De Meulder, B., Pellet, J., Balaur, I., Saqi, M., Nogueira, M. M., He, F., Parton, A., Lemonnier, N., Gawron, P., Gebel, S., Hainaut, P., Ollert, M., Dogrusoz, U., Barillot, E., Zinovyev, A., Schneider, R., Balling, R., and Auffray, C. (2018). Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *npj Systems Biology and Applications*, 4(1):21.
- [McQuilton et al., 2016] McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., and Sansone, S. A. (2016). BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database : the journal of biological databases and curation*, 2016.

- [Medina et al., 2017] Medina, Y. I., Carbonell, F., Sotero, R. C., and Evans, A. C. (2017). IDENTIFYING AND REVERSING MULTIFACTORIAL INTERACTIVE MECHANISMS IN ALZHEIMERS DISEASE. *Alzheimer's & Dementia*, 13(7):P124.
- [Mercado et al., 2016] Mercado, G., Castillo, V., Soto, P., and Sidhu, A. (2016). ER stress and Parkinson's disease: Pathological inputs that converge into the secretory pathway.
- [Mi et al., 2017] Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. (2017). PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*.
- [Mizuno et al., 2012] Mizuno, S., Iijima, R., Ogishima, S., Kikuchi, M., Matsuoka, Y., Ghosh, S., Miyamoto, T., Miyashita, A., Kuwano, R., and Tanaka, H. (2012). AlzPathway: A comprehensive map of signaling pathways of Alzheimer's disease. *BMC Systems Biology*.
- [Mons, 2005] Mons, B. (2005). Which gene did you mean? *BMC Bioinformatics*, 6.
- [Munoz and Ammit, 2010] Munoz, L. and Ammit, A. J. (2010). Targeting p38 MAPK pathway for the treatment of Alzheimer's disease.
- [Murphy et al., 2006] Murphy, S. N., Mendis, M. E., Berkowitz, D. a., Kohane, I., and Chueh, H. C. (2006). Integration of clinical and genetic data in the i2b2 architecture. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*.
- [Namasivayam et al., 2016] Namasivayam, A. A., Morales, A. F., Lacave, . M. F., Tallam, A., Simovic, B., Alfaro, D. G., Bobbili, D. R., Martin, F., Androsova, G., Shvydchenko, I., Park, J., Val Calvo, J., Hoeng, J., Peitsch, M. C., Racero, M. G. V., Biryukov, M., Talikka, M., Pérez, M. B., Rohatgi, N., Díaz-Díaz, N., Mandarapu, R., Ruiz, R. A., Davidyan, S., Narayanasamy, S., Boué, S., Guryanova, S., Arbas, S. M., Menon, S., and Xiang, Y. (2016). Community-reviewed biological network models for toxicology and drug discovery applications. *Gene Regulation and Systems Biology*, 10:51–66.

- [Neal et al., 2018] Neal, M. L., König, M., Nickerson, D., Msrl, G., Kalbasi, R., Dräger, A., Atalag, K., Chelliah, V., Cooling, M., Cook, D. L., Crook, S., Alba, M. d., Friedman, S. H., Garny, A., Gennari, J. H., Gleeson, P., Golebiewski, M., Hucka, M., Juty, N., Novère, N. L., Myers, C., Olivier, B. G., Sauro, H. M., Scharm, M., Snoep, J. L., Touré, V., Wipat, A., Wolkenhauer, O., and Waltemath, D. (2018). Harmonizing semantic annotations for computational models in biology. *bioRxiv*.
- [Neville et al., 2015] Neville, J., Kopko, S., Broadbent, S., Avilés, E., Stafford, R., Solinsky, C. M., Bain, L. J., Cisneroz, M., Romero, K., and Stephenson, D. (2015). Development of a unified clinical trial database for Alzheimer’s disease. *Alzheimer’s and Dementia*, 11(10):1212–1221.
- [Noble et al., 2013] Noble, W., Hanger, D. P., Miller, C. C., and Lovestone, S. (2013). The importance of tau phosphorylation for neurodegenerative diseases.
- [Novère et al., 2009] Novère, N. L., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S. E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T. C., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S., Wu, G., Goryanin, I., Kell, D. B., Sander, C., Sauro, H., Snoep, J. L., Kohn, K., and Kitano, H. (2009). The Systems Biology Graphical Notation.
- [Oda and Kitano, 2006] Oda, K. and Kitano, H. (2006). A comprehensive map of the toll-like receptor signaling network.
- [O’Donoghue et al., 2010] O’Donoghue, S. I., Horn, H., Pafilis, E., Haag, S., Kuhn, M., Satagopam, V. P., Schneider, R., and Jensen, L. J. (2010). Reflect: A practical approach to web semantics. *Journal of Web Semantics*, 8(2-3):182–189.
- [Oemig and Snelick, 2016] Oemig, F. and Snelick, R. (2016). Healthcare Standards Landscape. In *Healthcare Interoperability Standards Compliance Handbook*, pages 75–103.

- [Olney et al., 2017] Olney, N. T., Alquezar, C., Ramos, E. M., Nana, A. L., Fong, J. C., Karydas, A. M., Taylor, J. B., Stephens, M. L., Argouarch, A. R., Van Berlo, V. A., Dokuru, D. R., Sherr, E. H., Jicha, G. A., Dillon, W. P., Desikan, R. S., De May, M., Seeley, W. W., Coppola, G., Miller, B. L., and Kao, A. W. (2017). Linking tuberous sclerosis complex, excessive mTOR signaling, and age-related neurodegeneration: a new association between TSC1 mutation and frontotemporal dementia.
- [Osowski and Urano, 2011] Osowski, C. M. and Urano, F. (2011). *Measuring ER Stress and the Unfolded Protein Response Using Mammalian Tissue Culture System*, volume 490.
- [Ostaszewski et al., 2018] Ostaszewski, M., Gebel, S., Kuperstein, I., Mazein, A., Zinovyev, A., Dogrusoz, U., Hasenauer, J., Fleming, R. M. T., Le Novère, N., Gawron, P., Ligon, T., Niarakis, A., Nickerson, D., Weindl, D., Balling, R., Barillot, E., Auffray, C., and Schneider, R. (2018). Community-driven roadmap for integrated disease maps. *Briefings in Bioinformatics*.
- [Panahiazar et al., 2014] Panahiazar, M., Taslimitehrani, V., Jadhav, A., and Pathak, J. (2014). Empowering Personalized Medicine with Big Data and Semantic Web Technology: Promises, Challenges, and Use Cases. *Proceedings : ... IEEE International Conference on Big Data. IEEE International Conference on Big Data*, 2014:790–795.
- [Perez-Riverol et al., 2018] Perez-Riverol, Y., Zorin, A., Dass, G., Glon, M., Vizcaino, J. A., Jarnuczak, A., Petryszak, R., Ping, P., and Hermjakob, H. (2018). Quantifying the impact of public omics data. *bioRxiv*, page 282517.
- [Prinz et al., 2011] Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets?
- [Raghupathi and Raghupathi, 2014] Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1):3.

- [Rance et al., 2016] Rance, B., Canuel, V., Countouris, H., Laurent-Puig, P., and Burgun, A. (2016). Integrating Heterogeneous Biomedical Data for Cancer Research: the CARPEM infrastructure. *Applied Clinical Informatics*, 7(2):260–274.
- [Reddy, 2013] Reddy, P. H. (2013). Amyloid beta-induced glycogen synthase kinase 3 β phosphorylated VDAC1 in Alzheimer’s disease: Implications for synaptic dysfunction and neuronal damage.
- [Redman, 2016] Redman, T. C. (2016). Bad Data Costs the U.S. \$3 Trillion Per Year. *Harvard Business Review*, pages 1–6.
- [Rigden and Fernández, 2018] Rigden, D. J. and Fernández, X. M. (2018). The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Research*.
- [Rodriguez et al., 2016] Rodriguez, N., Pettit, J. B., Dalle Pezze, P., Li, L., Henry, A., van Iersel, M. P., Jalowicki, G., Kutmon, M., Natarajan, K. N., Tolnay, D., Stefan, M. I., Evelo, C. T., and Le Novère, N. (2016). The systems biology format converter. *BMC Bioinformatics*.
- [Rorie et al., 2017] Rorie, D. A., Flynn, R. W., Grieve, K., Doney, A., Mackenzie, I., MacDonald, T. M., and Rogers, A. (2017). Electronic case report forms and electronic data capture within clinical trials and pharmacoepidemiology.
- [Sano and Reed, 2013] Sano, R. and Reed, J. C. (2013). ER stress-induced cell death mechanisms.
- [Sansone et al., 2018] Sansone, S.-A., Mcquilton, P., Rocca-serra, P., Gonzalez-beltran, A., Izzo, M., Lister, A., and Thurston, M. (2018). FAIRsharing: working with and for the community to describe and link data standards , repositories and policies. *bioRxiv*, (1):0–9.
- [Satagopam et al., 2016] Satagopam, V., Gu, W., Eifes, S., Gawron, P., Ostaszewski, M., Gebel, S., Barbosa-Silva, A., Balling, R., and Schneider, R. (2016). Integration and Visualization of Translational Medicine Data for Better Understanding of Human Diseases. *Big data*, 4(2):97–108.

- [Scharm et al., 2015] Scharm, M., Wolkenhauer, O., and Waltemath, D. (2015). An algorithm to detect and communicate the differences in computational models describing biological systems. *Bioinformatics*, 32(4):563–570.
- [Schultze and Rosenstiel, 2018] Schultze, J. L. and Rosenstiel, P. (2018). Systems Medicine in Chronic Inflammatory Diseases.
- [Schumacher et al., 2014] Schumacher, A., Rujan, T., and Hoefkens, J. (2014). A collaborative approach to develop a multi-omics data analytics platform for translational research. *Applied & translational genomics*, 3(4):105–8.
- [Scott-Brown and Papachristodoulou, 2017] Scott-Brown, J. and Papachristodoulou, A. (2017). sbml-diff : A Tool for Visually Comparing SBML Models in Synthetic Biology. *ACS Synthetic Biology*, 6(7):1225–1229.
- [Shah et al., 2017] Shah, S. Z. A., Zhao, D., Hussain, T., and Yang, L. (2017). The role of unfolded protein response and mitogen-activated protein kinase signaling in neurodegenerative diseases with special focus on prion diseases.
- [Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*.
- [Shvaiko, 2005] Shvaiko, P. (2005). A Survey of Schema-based Matching Approaches. *Journal on Data Semantics*, 3730:146–171.
- [Smith and Brochhausen, 2010] Smith, B. and Brochhausen, M. (2010). Putting biomedical ontologies to work. *Methods of Information in Medicine*, 49(2):135–140.
- [Stanford et al., 2015] Stanford, N. J., Wolstencroft, K., Golebiewski, M., Kania, R., Juty, N., Tomlinson, C., Owen, S., Butcher, S., Hermjakob, H., Le Novere, N., Mueller, W., Snoep, J., and Goble, C. (2015). The evolution of standards and data management practices in systems biology. *Molecular Systems Biology*, 11(12):851–851.

- [Strömbäck et al., 2007] Strömbäck, L., Hall, D., and Lambrix, P. (2007). A review of standards for data exchange within systems biology.
- [Strömbäck et al., 2006] Strömbäck, L., Jakoniene, V., Tan, H., and Lambrix, P. (2006). Representing, storing and accessing molecular interaction data: A review of models and tools.
- [Sudduth et al., 2013] Sudduth, T. L., Schmitt, F. A., Nelson, P. T., and Wilcock, D. M. (2013). Neuroinflammatory phenotype in early Alzheimer’s disease. *Neurobiology of Aging*, 34(4):1051–1059.
- [Szalma et al., 2010] Szalma, S., Koka, V., Khasanova, T., and Perakslis, E. D. (2010). Effective knowledge management in translational medicine. *Journal of Translational Medicine*, 8.
- [Tanzi and Bertram, 2005] Tanzi, R. E. and Bertram, L. (2005). Twenty years of the Alzheimer’s disease amyloid hypothesis: A genetic perspective.
- [Tibbetts, 2011] Tibbetts, H. (2011). \$3 Trillion Problem: Three Best Practices for Today’s Dirty Data Pandemic — Hollis Tibbetts. 2011.
- [Ubeda et al., 2004] Ubeda, M., Kemp, D. M., and Habener, J. F. (2004). Glucose-induced expression of the cyclin-dependent protein kinase 5 activator p35 involved in Alzheimer’s disease regulates insulin gene transcription in pancreatic β -cells. *Endocrinology*, 145(6):3023–3031.
- [Urrea et al., 2013] Urrea, H., Dufey, E., Lisbona, F., Rojas-Rivera, D., and Hetz, C. (2013). When ER stress reaches a dead end.
- [Vassar et al., 2009] Vassar, R., Kovacs, D. M., Yan, R., and Wong, P. C. (2009). The β -Secretase Enzyme BACE in Health and Alzheimer’s Disease: Regulation, Cell Biology, Function, and Therapeutic Potential. *Journal of Neuroscience*, 29(41):12787–12794.
- [Viswanathan et al., 2008] Viswanathan, G. A., Seto, J., Patil, S., Nudelman, G., and Sealfon, S. C. (2008). Getting Started in Biological Pathway Construction and Analysis. *PLoS Computational Biology*, 4(2):e16.

- [Vivanco and Sawyers, 2002] Vivanco, I. and Sawyers, C. L. (2002). The phosphatidylinositol 3-Kinase AKT pathway in human cancer. *Nature reviews Cancer*, 2(7):489–501.
- [Waltemath et al., 2011] Waltemath, D., Adams, R., Bergmann, F. T., Hucka, M., Kolpakov, F., Miller, A. K., Moraru, I. I., Nickerson, D., Sahle, S., Snoep, J. L., and Le Novère, N. (2011). Reproducible computational biology experiments with SED-ML - The Simulation Experiment Description Markup Language. *BMC Systems Biology*, 5.
- [Waltemath et al., 2015] Waltemath, D., Bergmann, F. T., Chaouiya, C., Czauderna, T., Gleeson, P., Goble, C., Golebiewski, M., Hucka, M., Juty, N., Krebs, O., Le Novère, N., Mi, H., Moraru, I. I., Myers, C. J., Nickerson, D., Olivier, B. G., Rodriguez, N., Schreiber, F., Smith, L., Zhang, F., and Bonnet, E. (2015). Meeting report from the fourth meeting of the Computational Modeling in Biology Network (COMBINE). *Standards in Genomic Sciences*, 9(3):1285–1301.
- [Waltemath and Wolkenhauer, 2016a] Waltemath, D. and Wolkenhauer, O. (2016a). How Modeling Standards, Software, and Initiatives Support Reproducibility in Systems Biology and Systems Medicine. *IEEE Transactions on Biomedical Engineering*.
- [Waltemath and Wolkenhauer, 2016b] Waltemath, D. and Wolkenhauer, O. (2016b). How Modeling Standards, Software, and Initiatives Support Reproducibility in Systems Biology and Systems Medicine. *IEEE Transactions on Biomedical Engineering*, 63(10):1999–2006.
- [Walther et al., 2011] Walther, B., Hossin, S., Townend, J., Abernethy, N., Parker, D., and Jeffries, D. (2011). Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. *PLoS ONE*.
- [Wang et al., 2015] Wang, Q., Liu, Y., and Zhou, J. (2015). Neuroinflammation in Parkinson’s disease and its potential as therapeutic target.

- [Wang et al., 2018] Wang, Y., Kung, L. A., and Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126:3–13.
- [Wilkinson et al., 2016] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., t Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.
- [Williams et al., 2012] Williams, A. J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E. L., Evelo, C. T., Blomberg, N., Ecker, G., Goble, C., and Mons, B. (2012). Open PHACTS: Semantic interoperability for drug discovery.
- [Wolstencroft et al., 2017] Wolstencroft, K., Krebs, O., Snoep, J. L., Stanford, N. J., Bacall, F., Golebiewski, M., Kuzyakiv, R., Nguyen, Q., Owen, S., Soiland-Reyes, S., Straszewski, J., Van Niekerk, D. D., Williams, A. R., Malmström, L., Rinn, B., Müller, W., and Goble, C. (2017). FAIRDOMHub: A repository and collaboration environment for sharing systems biology research. *Nucleic Acids Research*, 45(D1):D404–D407.
- [Wolstencroft et al., 2015] Wolstencroft, K., Owen, S., Krebs, O., Nguyen, Q., Stanford, N. J., Golebiewski, M., Weidemann, A., Bittkowski, M., An, L., Shockley, D., Snoep, J. L., Mueller, W., and Goble, C. (2015). SEEK: A systems biology data and model management platform. *BMC Systems Biology*, 9(1).

- [Xie et al., 2017] Xie, L., Draizen, E. J., and Bourne, P. E. (2017). Harnessing Big Data for Systems Pharmacology. *Annual Review of Pharmacology and Toxicology*, 57(1):245–262.
- [Yarza et al., 2016] Yarza, R., Vela, S., Solas, M., and Ramirez, M. J. (2016). c-Jun N-terminal kinase (JNK) signaling as a therapeutic target for Alzheimer’s disease.
- [Yates et al., 2016] Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*.
- [Zhang et al., 2011] Zhang, X., Tang, N., Hadden, T. J., and Rishi, A. K. (2011). Akt, FoxO and regulation of apoptosis.
- [Zhong and Sternberg, 2007] Zhong, W. and Sternberg, P. W. (2007). Automated data integration for developmental biological research. *Development (Cambridge, England)*, 134(18):3227–38.