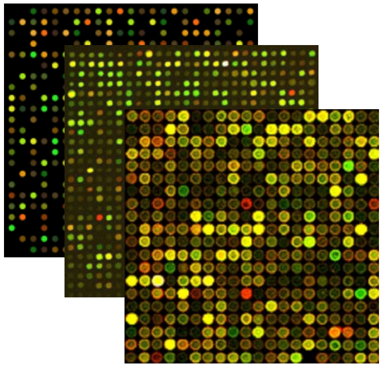
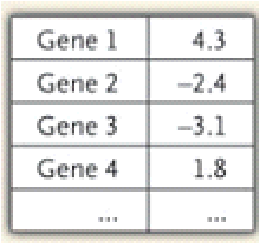
Algorithmic improvement of public cellular pathway and process definitions

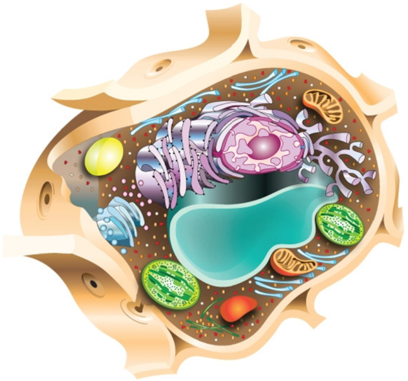
Enrico Glaab
Luxembourg Centre for Systems Biomedicine

Motivation for pathway analyses





Gene 1	4.3
Gene 2	–2.4
Gene 3	–3.1
Gene 4	1.8
...	...



- How do the changes in omics data relate to known cellular functions?
- Are there specific cellular pathways / molecular networks which display an over-representation of changes in my data?

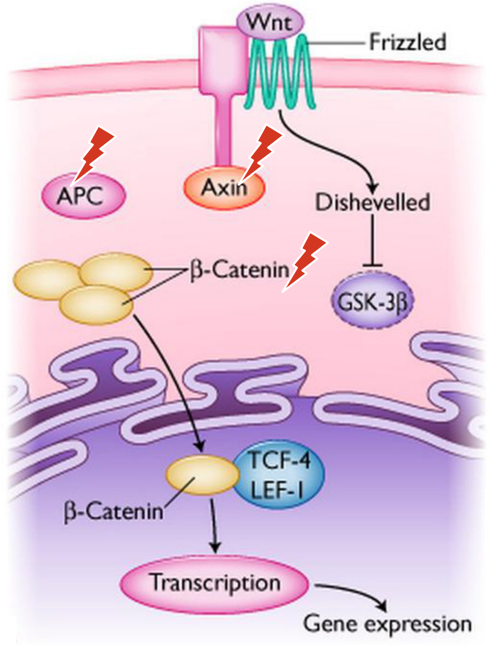
Motivation (2): Complex diseases as pathway perturbations

Alterations in different biomolecules of a cellular pathway or network can cause similar disruptions downstream

Example: Colorectal carcinoma

- Mutation deactivating APC has the same overall effect as mutations preventing degradation of β-catenin (Segditsas et al., 2006)

→ **Strategy**: Analyze alterations at the level of molecular networks and pathways to complement single gene/protein level analyses



Wnt/β-catenin signaling pathway
(⚡ = affected by disease-related mutations)

Motivation (3): The “curse of dimensionality“

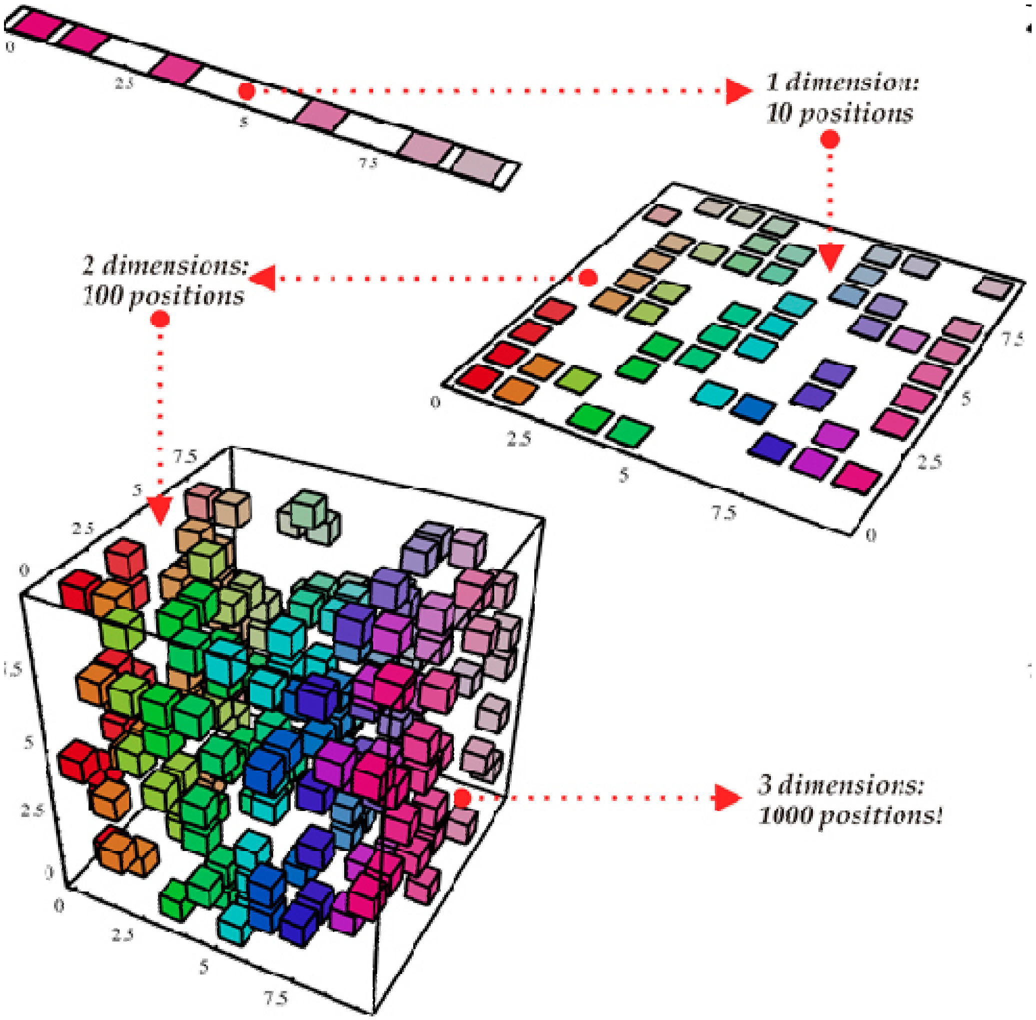
When analyzing increasing numbers of genes (features):

- the space spanned by these features grows exponentially (no. of features = no. of dimensions)

  → the available data tends to become sparse

  → discrimination between different sample groups (e.g. patients vs. controls) becomes more difficult

→ **Strategy**: Use pathway activity representations of the data to reduce the number of dimensions



Pathway / gene set resources

- Many public databases on functional gene sets and pathways available
- Both generic, multi-organism pathway collections and specialized collections (e.g. disease pathways such as the PD map)
- Format standardization efforts underway (BioPax, SBGN/SBML)

BIOCARTA

KEGG Kyoto Encyclopedia of Genes and Genomes

WIKIPATHWAYS

GO the Gene Ontology

Reactome

Representations of pathways / functional gene groups

GENE SETS

Mitochondrion

Gene.1
Gene.2
Gene.3

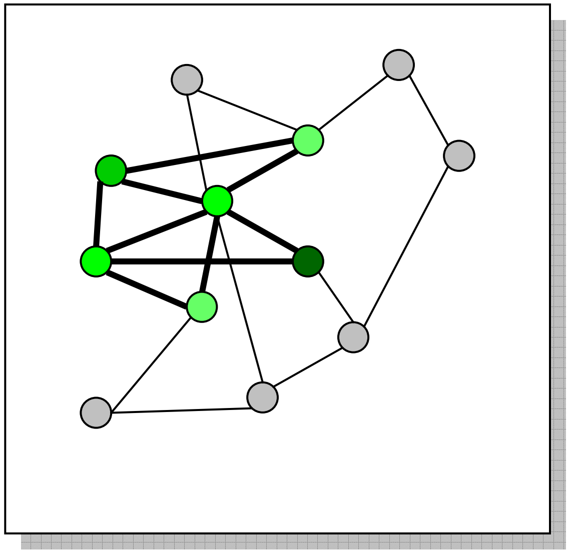
P53 signaling

Gene.2
Gene.4
Gene.5

→ Find gene sets whose members are enriched among the differentially expressed genes

→ pure statistical scoring

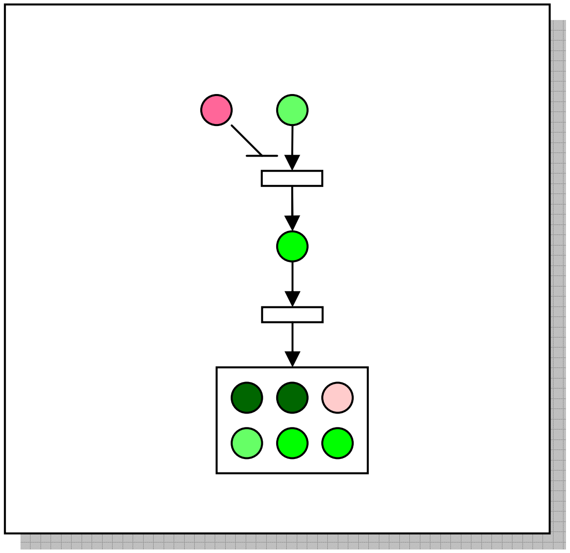
NETWORKS



→ Identify network regions enriched in expression alterations

→ scoring topological + expression criteria

DIAGRAMS

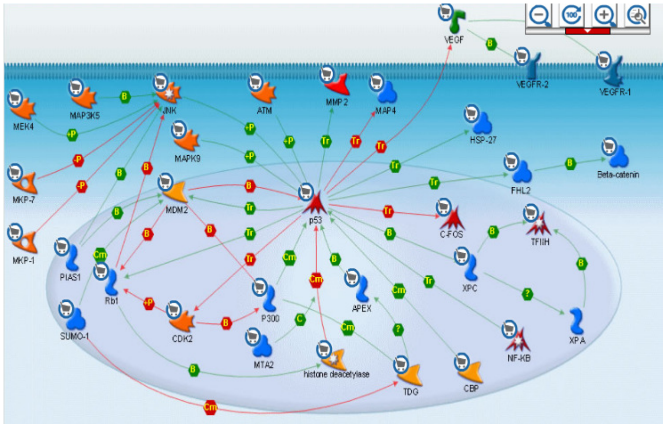


→ Score pathways with regulatory consistent expression alterations

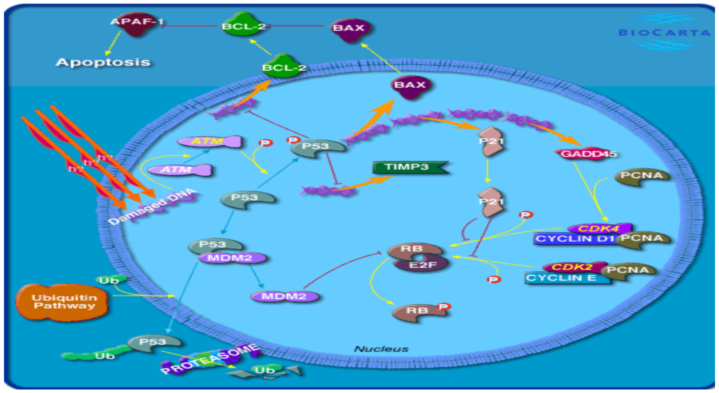
→ scoring topology + expression changes + consistency criteria

Inconsistencies between pathway definitions

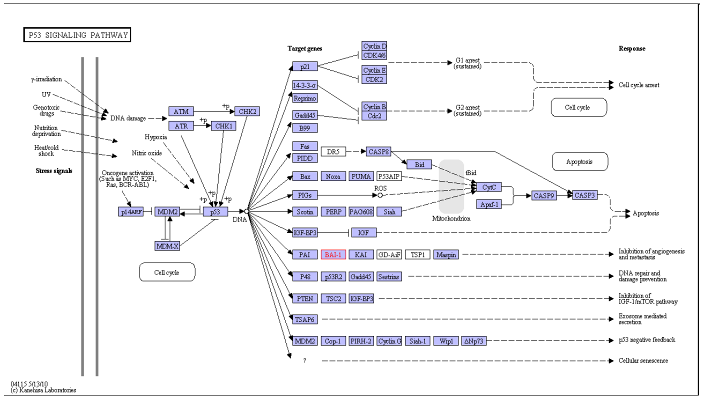
- Pathways are usually manually curated → subjective decisions on members & boundaries
- A pathway defined for the same cellular process may look entirely different in two separate databases, e.g. “p53 signaling“:



Invitrogen iPath (p53 signaling)



BioCarta (p53 signaling)



KEGG (p53 signaling)

Improving pathway definitions using networks

- **Questions**: Can we make pathway definitions more objective? Can we improve existing pathways according to quantitative criteria (compactness, connectivity, density)?

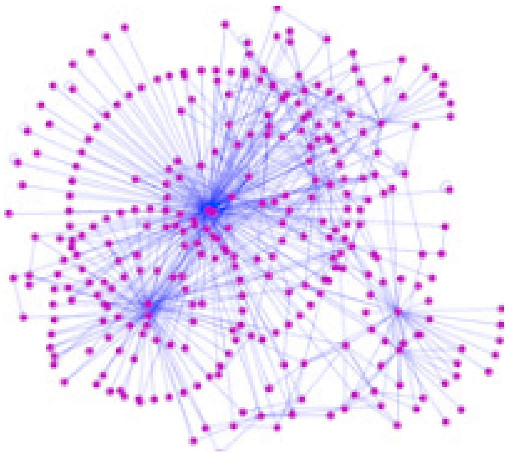
- **Strategy**: Use genome-scale networks to redefine pathways:

  - protein-protein interactions

  - genetic interactions

  - gene co-expression relations

  → large-scale, higher coverage, less biased
  → can also reveal communication between pathways (“cross-talk“)

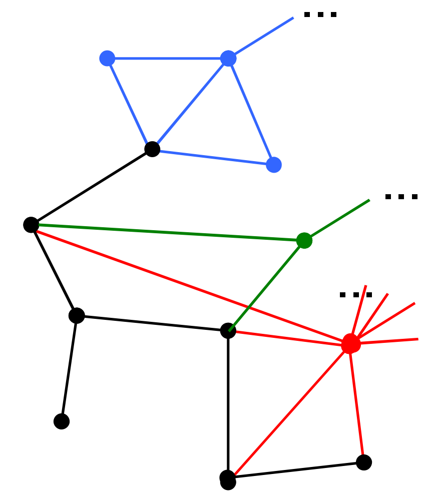


PathExpand: Network-based pathway extension

- **Idea**: Extend pathways by adding genes that are ”strongly connected“ to the pathway-nodes and increase the pathway-“compactness“ in a network.

Pathway extension criteria: Add a node v to set P if:

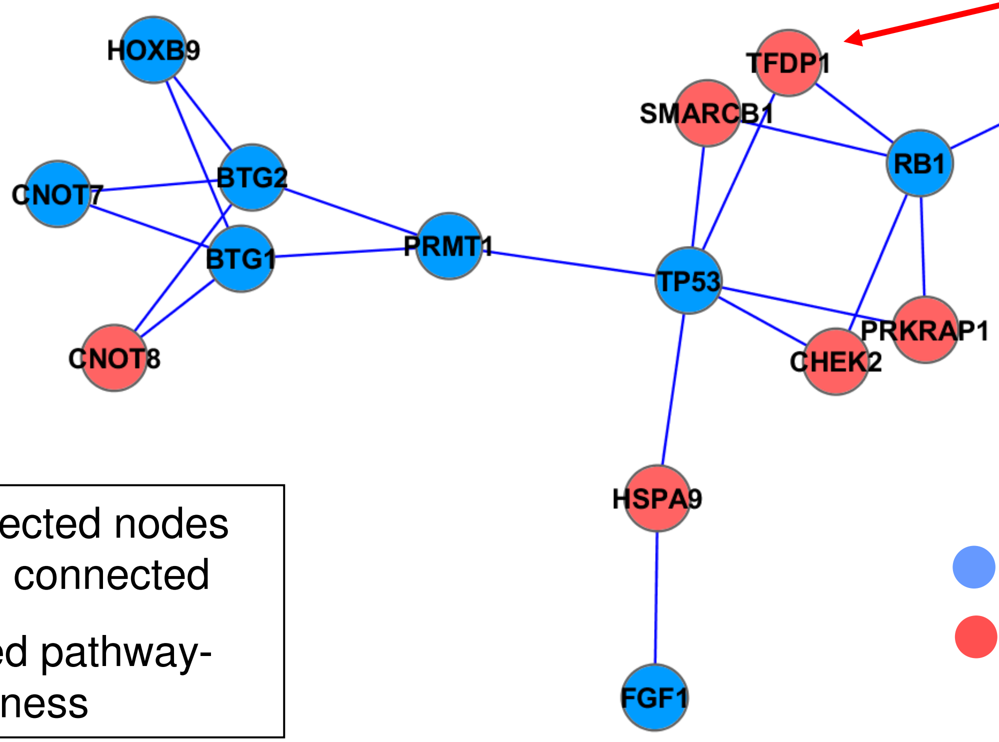
- v has a pathway-neighbour and degree(v) > 1; and
- #pathway-links(v,p) / #outside-links(v,p) > T_1; or
- #triangle-links(v,p) / #possible_triangles(v,p) > T_2; or
- #pathway-links(v,p) / #pathway-nodes(p) > T_3; and
- avg. shortest path distance in {P,v} smaller than in P



black = pathway-nodes
red **blue** **green** = nodes added based on different criteria

PathExpand: Example

Known cancer pathway: "*BTG family proteins and cell cycle regulation*" (BioCarta)



Added known cancer gene

● original pathway

● added nodes

→ Disconnected nodes become connected

→ increased pathway-compactness

PathExpand: Cross-validation

Question: Can randomly deleted genes in the original pathways be recovered by the expansion?

→ 3-step cross-validation procedure:

1. Randomly remove 10% of the pathway members (among proteins with at least one partner in the pathway)

2. Apply the proposed extension procedure as well as 100 random extensions (random sampling among candidates)

3. Estimate p-value-like significance scores:

$$\sum_{i \in P} \left(\frac{\sum_{i=1}^{100} I(recovery\_random_i >= recovery\_proposed)}{100} \right) / |P|$$

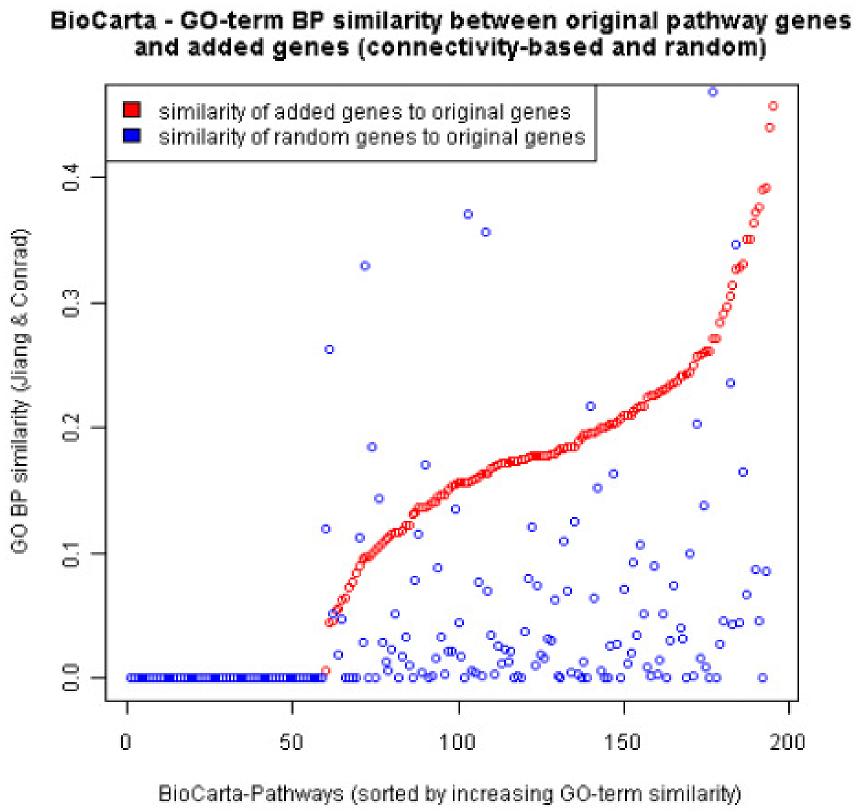
PathExpand: Semantic similarity analysis

- **Goal**: Quantify pairwise similarities between protein annotations

  Method: Jiang & Conrath's semantic GO term similarity measure

- Compute avg. GO-term similarity between pathway-proteins and added proteins

  → compare to random extension model



Biological applications (1): Alzheimer‘s disease

- More than 20 proteins

  annotated in our

  PPI network

- 5 proteins added by the

  extension process (circled)

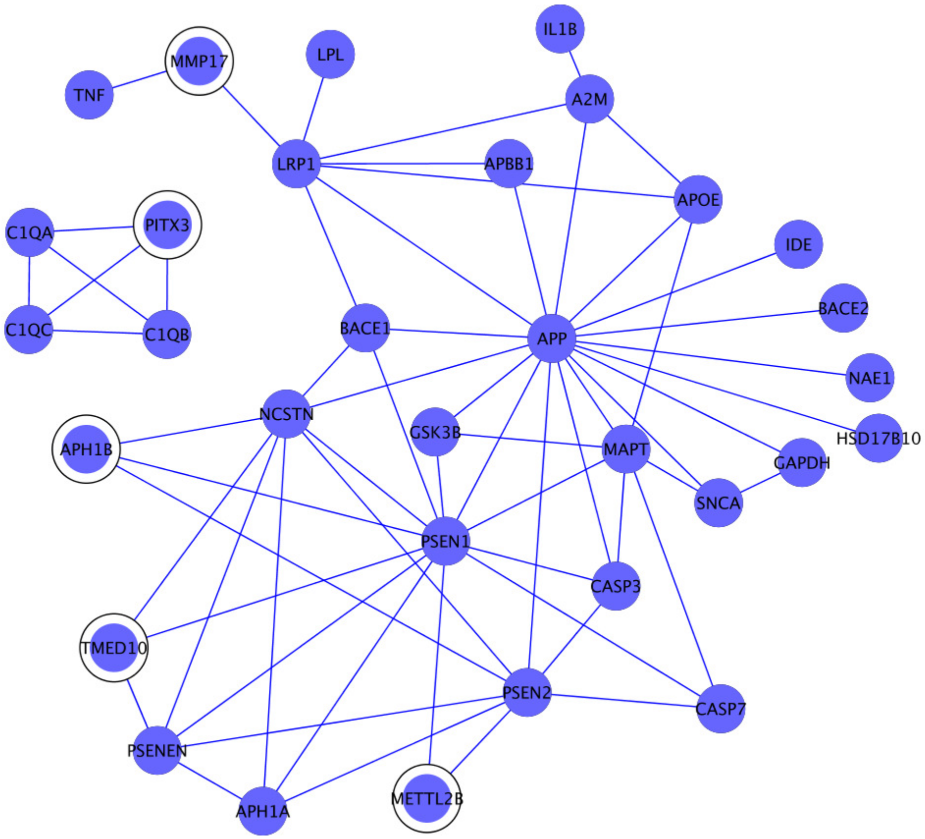
- 3 known to be associated

  with the disease

- 2 novel candidates:

  METTL2B, TMED10*

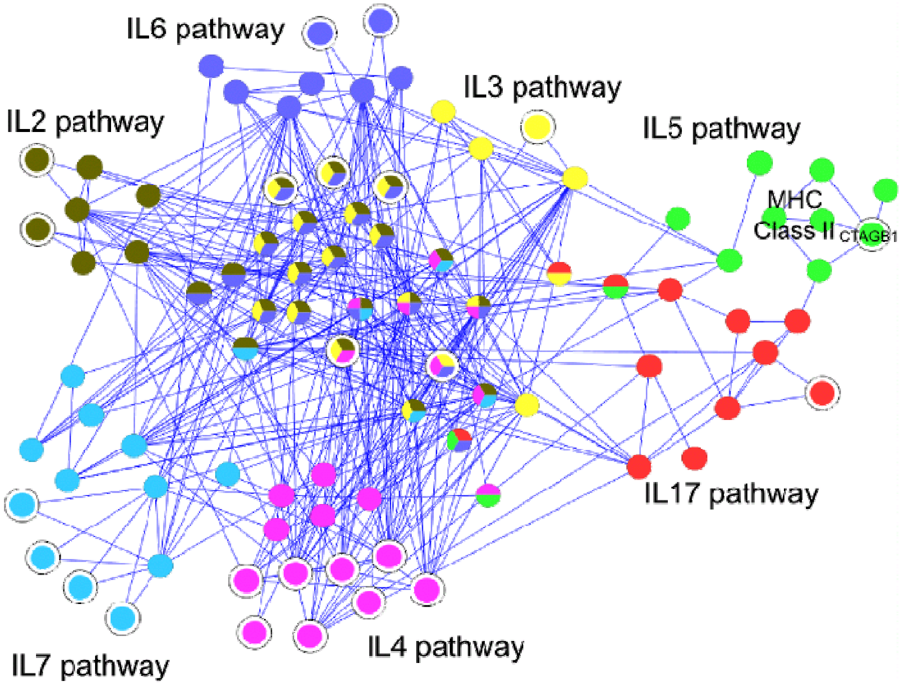
  (*putative early-onset AD mutations reported)



KEGG Alzheimer disease pathway mapped on human PPI-network

Biological applications (2): Interleukin signaling

- Complex system of intracellular signaling cascades
- New putative pathway regulators identified
- New "cross-talk proteins" identified (associated with multiple pathways)



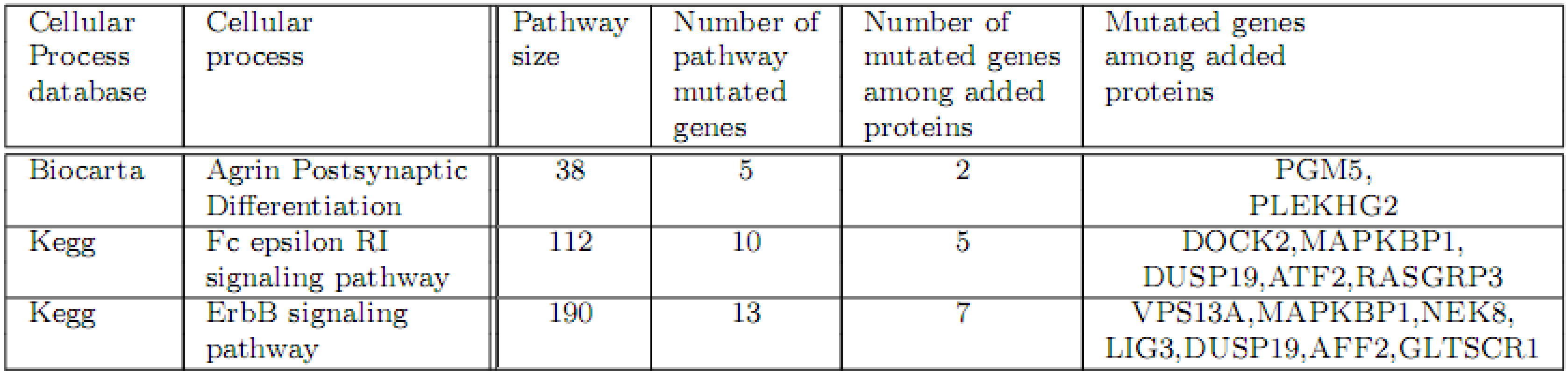
Two functions: pathway-regulation & pathway-communication?

Biological applications (3): Enrichment analysis

Classical approach: Test enrichment of experimentally derived gene sets in cellular pathway members (one-sided Fisher exact test)

→ **Idea**: replace original pathways by extended versions

Example: Enrichment analysis for pancreatic cancer mutated genes:

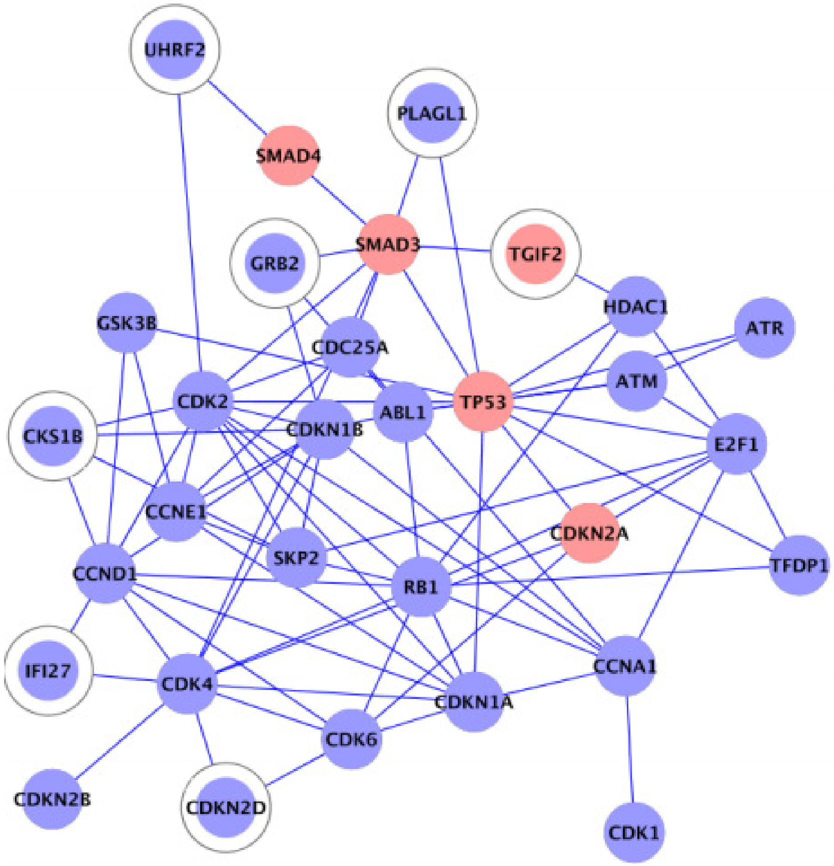


Cellular Process database	Cellular process	Pathway size	Number of pathway mutated genes	Number of mutated genes among added proteins	Mutated genes among added proteins
Biocarta	Agrin Postsynaptic Differentiation	38	5	2	PGM5, PLEKHG2
Kegg	Fc epsilon RI signaling pathway	112	10	5	DOCK2,MAPKBP1, DUSP19,ATF2,RASGRP3
Kegg	ErbB signaling pathway	190	13	7	VPS13A,MAPKBP1,NEK8, LIG3,DUSP19,AFF2,GLTSCR1

Biological applications (4): Pancreatic cancer

- “Cell cycle G1/S check point process” - extension procedure adds 7 proteins

- 6 of the added proteins are involved in cell cycle regulation
- the 7th (TGIF2) is known to be mutated in pancreatic cancer
- points to functional role of added proteins



PathExpand: Conclusion & Summary

- The method integrates two sources of information, extending **canonical pathways** using large-scale **protein interaction data**

- Three **evaluated methods**: cross-validation, GO-term semantic similarity and enrichment analysis

- Extended pathways are more compact and provide insights on on **pathway regulators**, the **cross-talk** between pathways and gene set **functional enrichment**

References

1. E. Glaab, J.P. Trezzi, A. Greuel, C. Jäger, Z. Hodak, A. Drzezga, L. Timmermann, M. Tittgemeyer, N. J. Diederich, C. Eggers, Integrative analysis of blood metabolomics and PET brain neuroimaging data for Parkinson's disease, Neurobiology of Disease (2019), Vol. 124, No. 1, pp. 555
2. E. Glaab, *Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification*, Briefings in Bioinformatics (2015), 17(3), pp. 440
3. E. Glaab, R. Schneider, *Comparative pathway and network analysis of brain transcriptome changes during adult aging and in Parkinson's disease*, Neurobiology of Disease (2015), 74, 1-13
4. N. Vlassis, E. Glaab, *GenePEN: analysis of network activity alterations in complex diseases via the pairwise elastic net*, Statistical Applications in Genetics and Molecular Biology (2015), 14(2), 221
5. Z. Zhang, P. P. Jung, V. Grouès, P. May, C. Linster, E. Glaab, *Web-based QTL linkage analysis and bulk segregant analysis of yeast sequencing data*, GigaScience (2019), 8(6), 1-18
6. S. Köglsberger, M. L. Cordero-Maldonado, P. Antony, J. I. Forster, P. Garcia, M. Buttini, A. Crawford, E. Glaab, *Gender-specific expression of ubiquitin-specific peptidase 9 modulates tau expression and phosphorylation: possible implications for tauopathies*, Molecular Neurobiology (2017), 54(10), pp. 7979
7. L. Grandbarbe, S. Gabel, E. Koncina, G. Dorban, T. Heurtaux, C. Birck, E. Glaab, A. Michelucci, P. Heuschling, *Inflammation promotes a conversion of astrocytes into neural progenitor cells via NF-kB activation*, Molecular Neurobiology (2016), Vol. 53, No. 8, 5041-5055
8. S. Kleiderman, J. Sá, A. Teixeira, C. Brito, S. Gutbier, L. Evje, M. Hadera, E. Glaab, M. Henry, S. Agapios, P. Alves, U. Sonnewald, M. Leist, *Functional and phenotypic differences of pure populations of stem cell-derived astrocytes and neuronal precursor cells*, Glia (2016), Vol. 64, No. 5, 695-715
9. E. Glaab, R. Schneider, *RepExplore: Addressing technical replicate variance in proteomics and metabolomics data analysis*, Bioinformatics (2015), 31(13), pp. 2235
10. E. Glaab, *Building a virtual ligand screening pipeline using free software: a survey*, Briefings in Bioinformatics (2015), 17(2), pp. 352
11. E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, A. Valencia. *EnrichNet: network-based gene set enrichment analysis*, Bioinformatics, 28(18):i451-i457, 2012
12. E. Glaab, R. Schneider, *PathVar: analysis of gene and protein expression variance in cellular pathways using microarray data*, Bioinformatics, 28(3):446-447, 2012
13. E. Glaab, J. Bacardit, J. M. Garibaldi, N. Krasnogor, *Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data*, PLoS ONE, 7(7):e39932, 2012
14. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. *TopoGSA: network topological gene set analysis*, Bioinformatics, 26(9):1271-1272, 2010
15. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. *Extending pathways and processes using molecular interaction networks to analyse cancer genome data*, BMC Bioinformatics, 11(1):597, 2010
16. E. Glaab, J. M. Garibaldi and N. Krasnogor. *ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization*, BMC Bioinformatics,10:358, 2009
17. E. Glaab, J. M. Garibaldi, N. Krasnogor. *Learning pathway-based decision rules to classify microarray cancer samples*, German Conference on Bioinformatics 2010, Lecture Notes in Informatics (LNI), 173, 123-134
18. E. Glaab, J. M. Garibaldi and N. Krasnogor. *VRMLGen: An R-package for 3D Data Visualization on the Web*, Journal of Statistical Software, 36(8),1-18,2010