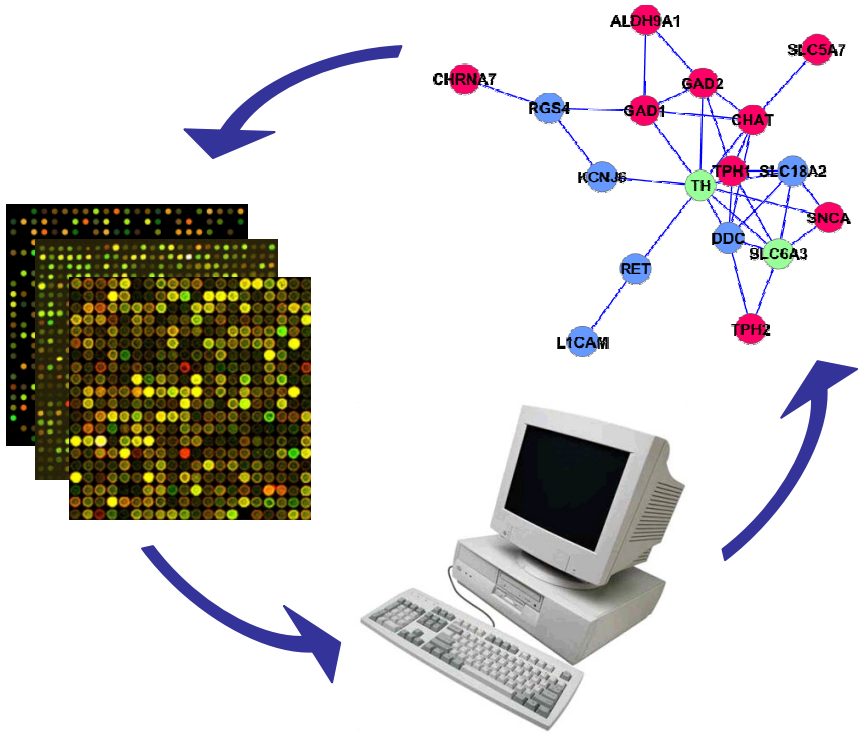
Conference of the German Society for Genetics



Integrated bioinformatics analysis of functional omics and GWAS data for neurodegenerative disorders

Speaker: Enrico Glaab, Luxembourg Centre for Systems Biomedicine

Outline:

- Motivation: Complex diseases as pathway or network perturbations
- Finding disease-associated pathways using omics network analysis (software EnrichNet, joint work with Alfonso Valencia CNIO group)
- Using pathway information to improve omics sample classification
- Application to public omics & GWAS data from Parkinson‘s disease studies

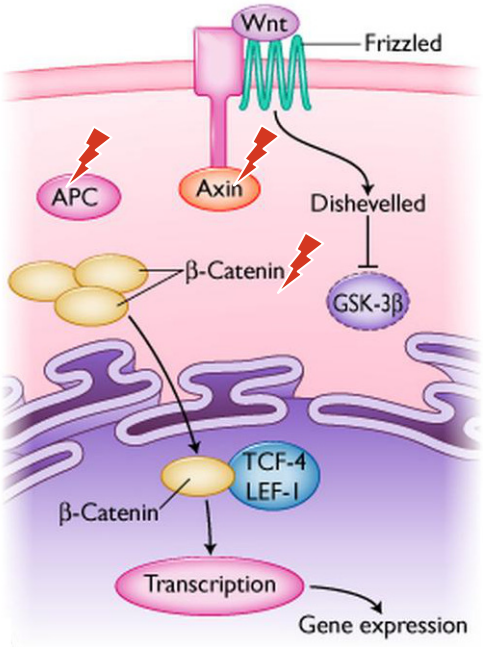
Motivation: Complex diseases as network perturbations

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)

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



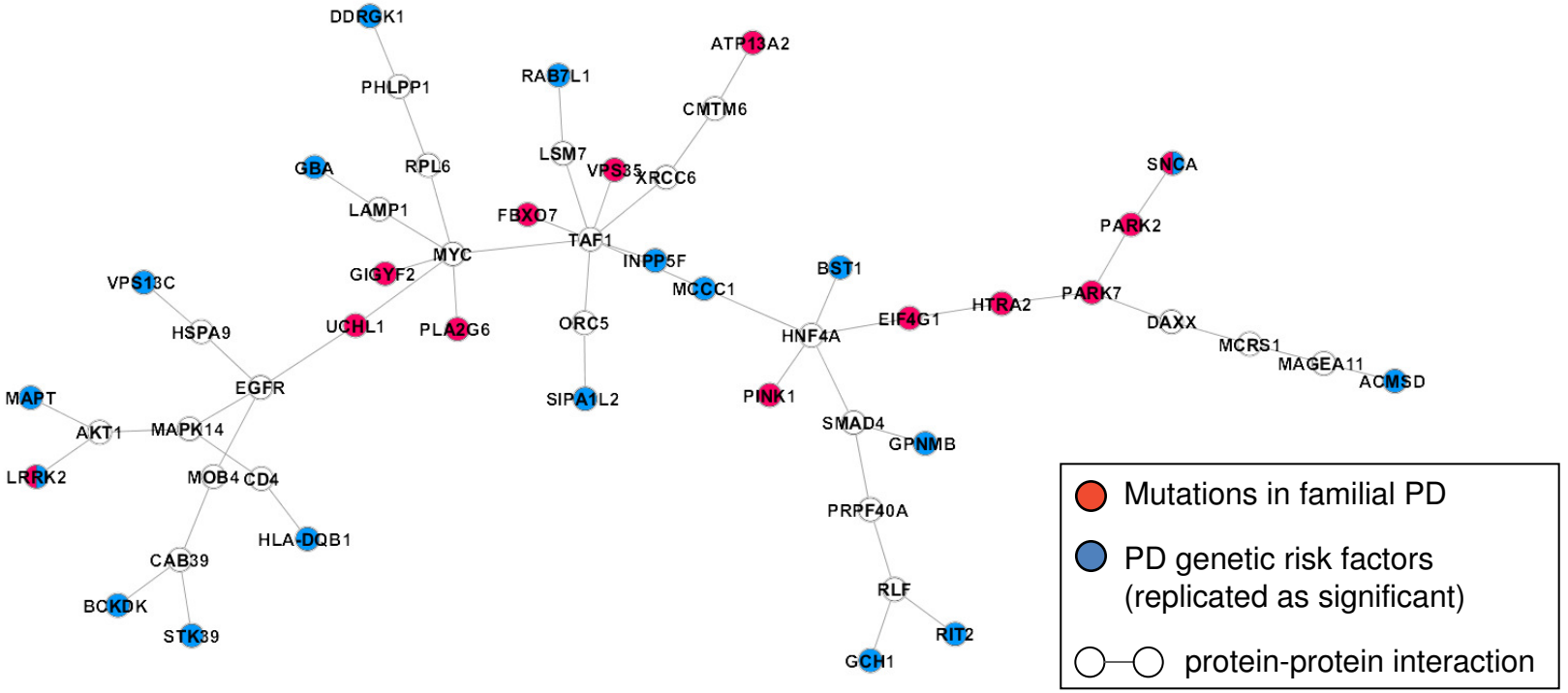
Wnt/β-catenin signaling pathway
(= affected by disease-related mutations)

Network-based approach to Parkinson's disease

Parkinson's disease (PD) as a network perturbation

- **Familial PD**: Multiple disease-causing mutations across different genes
- **Idiopathic PD**: Complex interplay between genetic and environmental risk factors

→ interrelate affected genes in molecular networks



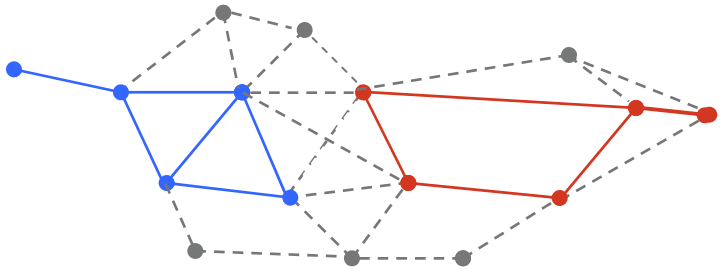
EnrichNet: Network scoring of disease-related pathways

EnrichNet pathway association scoring idea:

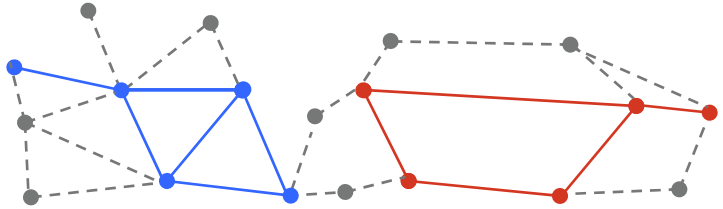
→ Quantify distances in a molecular network and multiplicity and density of interactions between the genes/proteins of interest (considering all possible interconnecting random walks)

Illustration:

Case 1:
many, dense and short inter-connecting routes



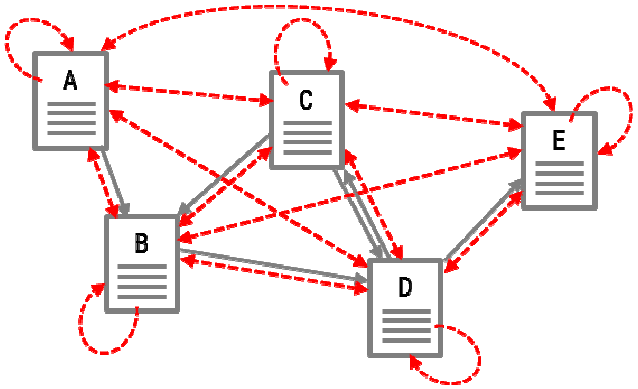
Case 2:
few, sparse and long inter-connecting routes



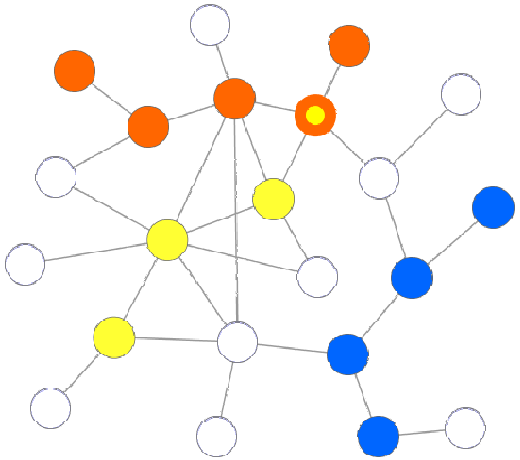
- **target set nodes** (e.g. differentially expressed genes in PD vs. controls)
- **pathway nodes**
- **other nodes**

EnrichNet: Network node relevance scoring

Google "Personalized Page Rank":



Transfer approach to molecular networks



Compute steady-state node visiting probabilities for random walks:

$$p^{t+1} = (1\text{-}r)\ A\ p^t + p^0$$

- A:= network adjacency matrix
- r:= restart probability (here: r = 0.9)
- p_i^t:= probability walker is at node i at time t

target set | target/pathway overlap | pathway 1 | pathway 2

→ **Output**: Relevance scores for each pathway

→ Use distance-weighted aggregation (Xd-score, Olmea et al. 2009)

EnrichNet: Comparative analysis

Comparative analysis on benchmark microarray data:

- compare EnrichNet against classical over-representation analysis using benchmark datasets from the Broad Institute of MIT and Harvard (5 gene expression datasets and 2 pathway databases)

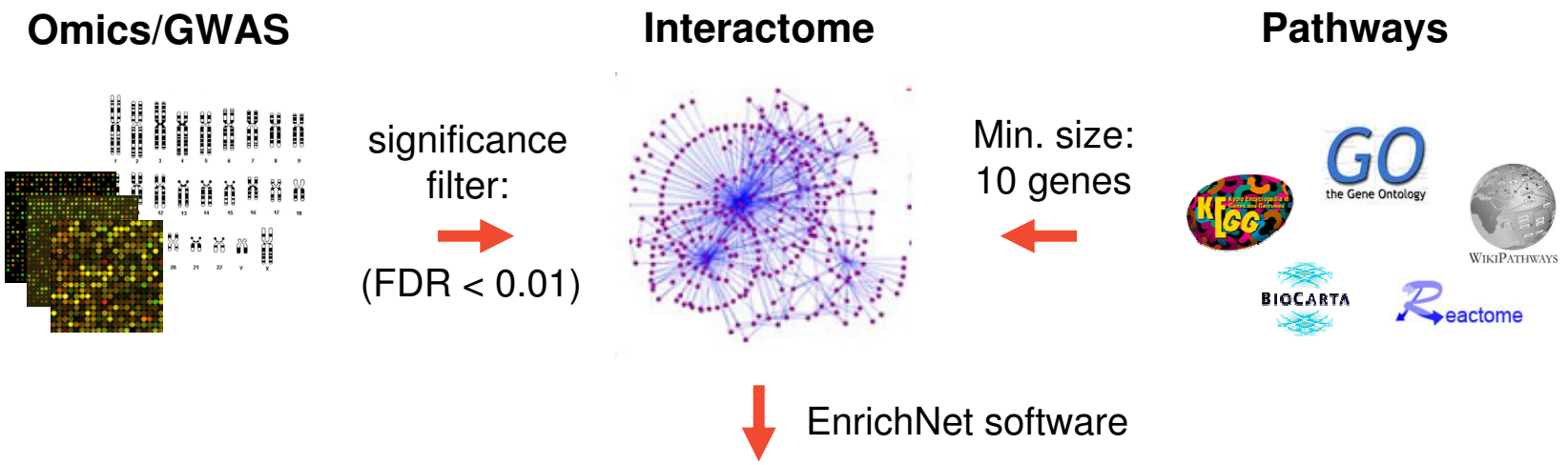
→ EnrichNet provides a consistently higher agreement with benchmark rankings

Dataset	Gene set collection	Fisher exact test Similarity score (p-value)	EnrichNet Similarity score (p-value)
p53	**C1**	13.5 ($p = 0.225$)	36.9 ($p < 0.001$)
	C2	45.6 ($p < 0.001$)	65.2 ($p < 0.001$)
Lung (Boston)	**C1**	2.6 ($p = 0.936$)	40.0 ($p < 0.001$)
	C2	15.0 ($p = 0.302$)	43.7 ($p < 0.001$)
Lung (Michigan)	**C1**	21.2 ($p = 0.028$)	40.8 ($p < 0.001$)
	C2	9.1 ($p = 0.634$)	40.5 ($p = 0.001$)
Colon	**C1**	6.85 ($p = 0.673$)	70.1 ($p < 0.001$)
	C2	22.8 ($p = 0.075$)	94.9 ($p < 0.001$)
Lymphoma	**C1**	8.0 ($p = 0.569$)	65.2 ($p < 0.001$)
	C2	0.94 ($p = 0.985$)	69.8 ($p < 0.001$)

EnrichNet: Parkinson's disease datasets and workflow

Datasets:

- 8 public *post mortem* brain (*substantia nigra*) gene expression datasets from PD case-control studies
- GWAS data from 5 public PD case-control studies
- a genome-scale human protein interaction network (STRING high-confidence network)
- cellular pathway definitions from 5 annotation databases



EnrichNet: Results on Parkinson's disease data

Top-ranked cellular processes for **transcriptomics** data (Gene Ontology):

Gene Ontology process	XD-score	Fisher Q-value	Pathway size	Overlap size
dopamine metabolic process (GO:0042417)	1.851	0.034	14	3
neuron maturation (GO:0042551)	1.722	0.353	10	2
response to herbicide (GO:0009635)	1.722	0.353	10	2
response to amphetamine (GO:0001975)	1.163	0.024	29	4
neurotransmitter transport (GO:0006836)	0.951	0.034	35	4
nervous system development (GO:0007399)	0.203	0.024	320	10

Top-ranked cellular processes for **GWAS** data (Gene Ontology):

Gene Ontology process	XD-score	Fisher Q-value	Pathway size	Overlap size
synaptic vesicle endocytosis (GO:0048488)	1.457	0.537	12	2
regulation of autophagy (GO:0010506)	1.157	0.537	15	2
lipid biosynthetic process (GO:0008610)	1.082	0.537	16	2
receptor clustering (GO:0043113)	1.082	0.537	16	2
long-term synaptic potentiation (GO:0060291)	1.082	0.537	16	2

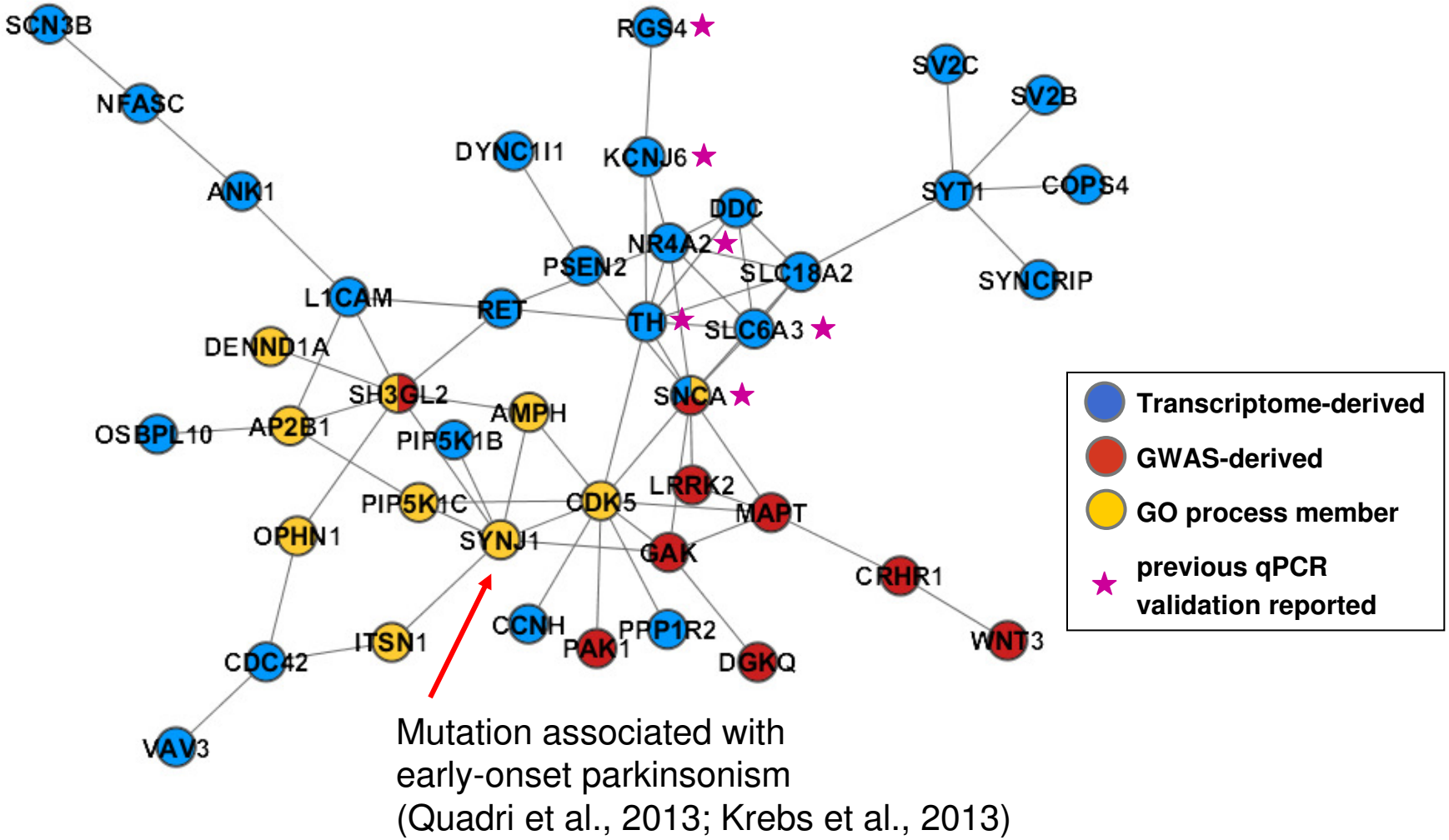
EnrichNet: Integrate transcriptomics and GWAS data

Combine rankings for **transcriptomics** and **GWAS** data (using the sum of standardized scores)
→ Joint top-ranked cellular processes (Gene Ontology):

Gene Ontology process	Trans-criptome XD-score	Trans-criptome overlap	GWAS XD-score	GWAS overlap	Pathway size
synaptic vesicle endocytosis (GO:0048488)	0.672	1	1.457	2	12
response to herbicide (GO:0009635)	1.722	2	0.857	1	10
dopamine metabolic process (GO:0042417)	1.851	3	0.599	1	14
long-term synaptic potentiation (GO:0060291)	0.484	1	1.082	2	16
axon extension (GO:0048675)	0.822	1	0.857	1	10
synaptic transmission, dopaminergic (GO:0001963)	1.307	2	0.649	1	13
regulation of exocytosis (GO:0017157)	1.047	2	0.519	1	16
cellular response to stress (GO:0033554)	0.614	1	0.649	1	13
positive regulation of endocytosis (GO:0045807)	0.565	1	0.599	1	14
positive regulation of release of sequestered calcium ion into cytosol (GO:0051281)	0.565	1	0.599	1	14
positive regulation of protein serine/threonine kinase activity (GO:0071902)	0.565	1	0.599	1	14
regulation of neurotransmitter secretion (GO:0046928)	0.484	1	0.519	1	16
positive regulation of synaptic transmission (GO:0050806)	0.484	1	0.519	1	16
branched chain family amino acid catabolic process (GO:0009083)	0.422	1	0.457	1	18
cellular response to oxidative stress (GO:0034599)	0.422	1	0.457	1	18

EnrichNet: Network visualization for top-ranked GO term

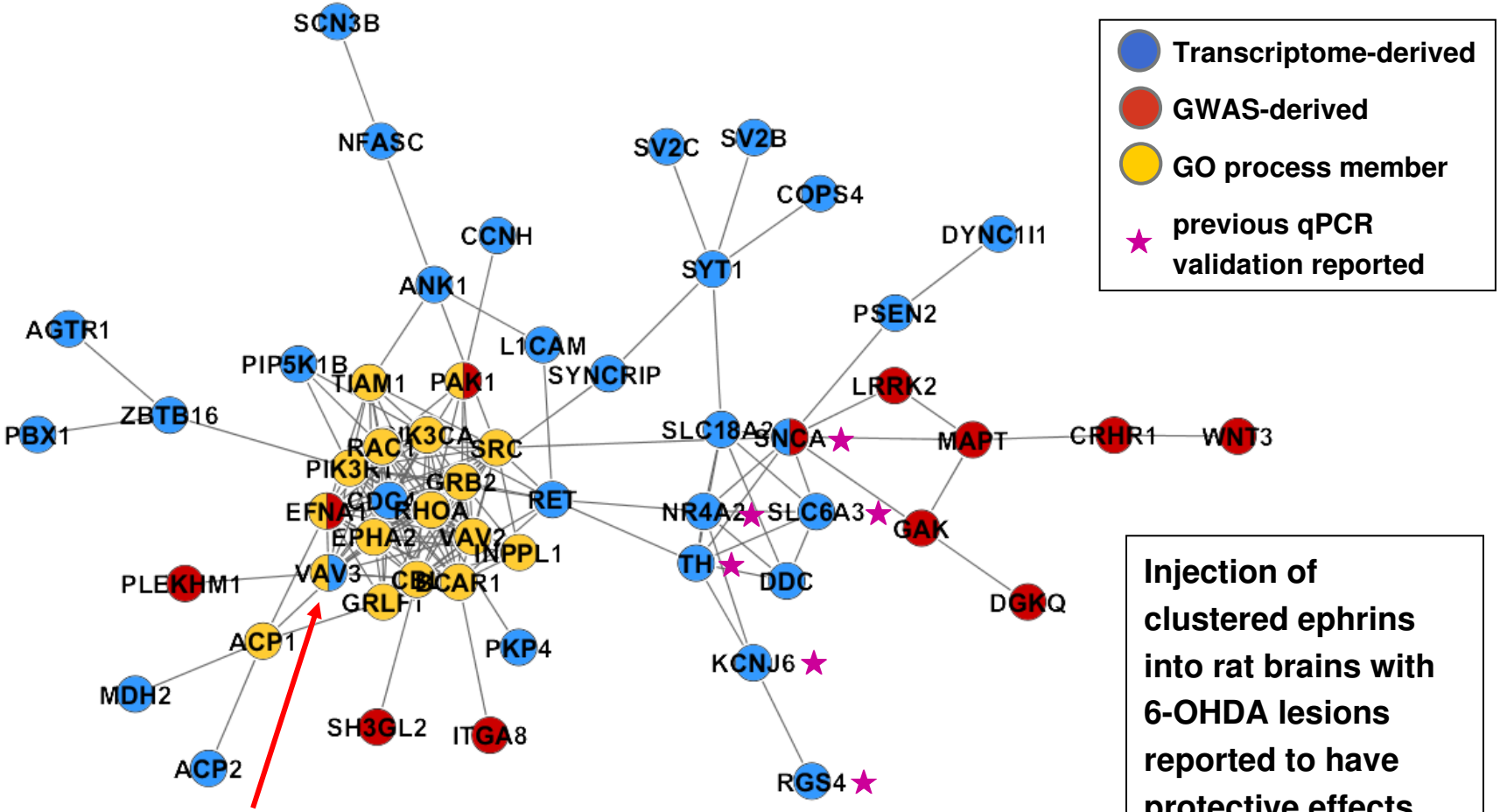
Largest connected network component for “Synaptic vesicle endocytosis” GO process



Mutation associated with early-onset parkinsonism (Quadri et al., 2013; Krebs et al., 2013)

EnrichNet: Network visualization example

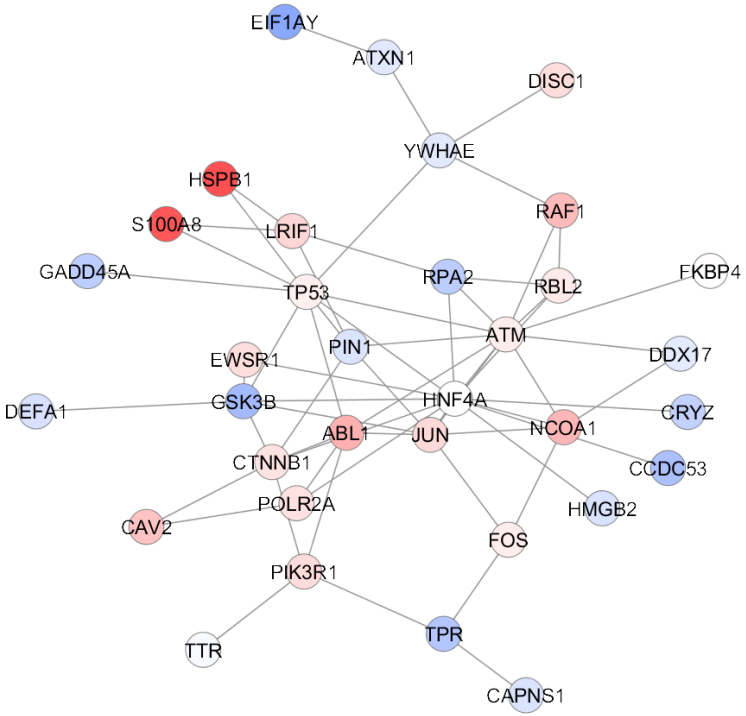
Largest connected network component for “EPHA2 forward signaling” pathway (PID)



Using pathway information for sample classification

Question: Disease-associated expression alterations are often localized in pathways / network clusters – can we use this information to build more robust classification models for biomarker development?

Idea: Collect a dictionary of pathway-level predictive features ("fingerprints") to build machine learning classifiers for omics data (e.g. using mean or variance of expression in pathway, PCA / MDS components of pathway expression)



Example of localized transcriptome changes in PD (data by Zhang et al., 2005):

- Over-expressed in PD
- Under-expressed in PD

Pathway-level sample classification: Cross-validation

Sample classification results on PD brain transcriptomic data

→ map dataset by Moran et al. (2006) onto Gene Ontology (GO) processes
→ use empirical Bayes moderated T-test to select attributes ("pathway fingerprints") and a linear Support Vector Machine for classification (10-fold cross-validation)

	Accuracy and stddev. for different numbers of selected attributes			
Attribute type	**10**	**30**	**50**	**100**
Gene-level model	89.2 ± 14.2	89.2 ± 14.2	92.5 ± 12.1	92.5 ± 12.1
GO - Mean	84.2 ± 13.9	90 ± 12.9	92.5 ± 12.1	89.2 ± 14.2
GO - Median	84.2 ± 13.9	91.7 ± 13.6	**95 ± 10.5**	91.7 ± 13.6
GO - Stddev.	76.7 ± 18.8	81.7 ± 17.5	79.2 ± 20.1	86.7 ± 14.3
GO - Min.	71.7 ± 21.9	68.3 ± 25.1	79.2 ± 23.3	71.7 ± 24.9
GO - Max.	81.7 ± 17.5	84.2 ± 13.9	90 ± 12.9	84.2 ± 18.2
GO - PCA	89.2 ± 14.2	**95 ± 10.5**	92.5 ± 12.1	**95 ± 10.5**
GO - MDS	**91.7 ± 13.6**	86.7 ± 18.5	84.2 ± 18.2	87.5 ± 17.7

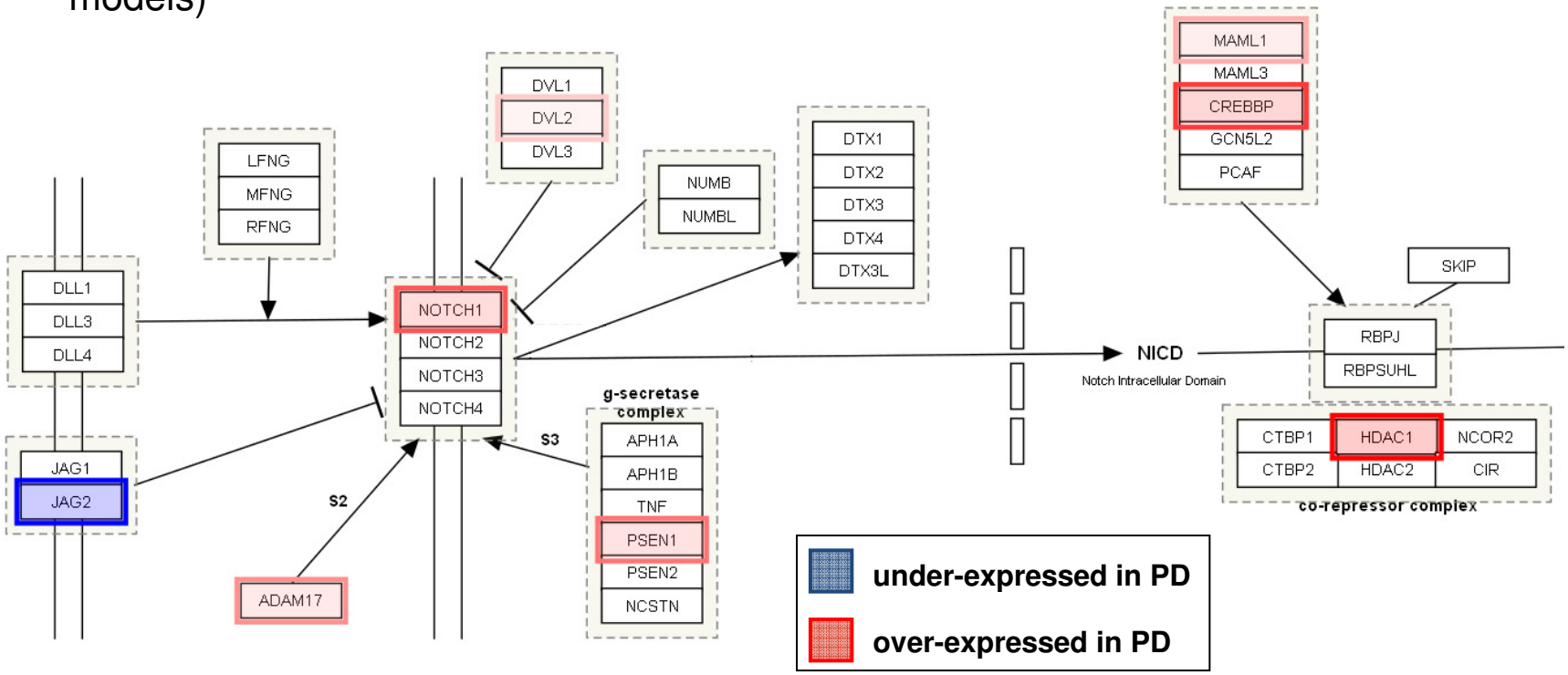
Model limitations: 1) post-mortem samples; 2) tissue-specific (brain, *substantia nigra*) 3) no disease controls (PD vs. unaffected)

Visualizing the pathways with most predictive changes

Example: Notch signaling pathway (Wikipathways)

→ identify pathways which are both enriched in gene activity alterations and contain predictive information for sample classification

→ select representative genes for qPCR validation (iterative refinement of prediction models)



Summary

- We present a software, EnrichNet, for the integrated pathway and network analysis of disease-related omics data
- On public transcriptomics and GWAS datasets for Parkinson‘s disease, the approach identifies new statistically significant pathway associations (not detected with classical overlap-based approach)
- The pathway information can be used to improve machine learning models for omics sample classification

Software available at: www.enrichnet.org
pathvar.embl.de

References

1. Lerner, T. N., & Kreitzer, A. C. (2012). RGS4 is required for dopaminergic control of striatal LTD and susceptibility to parkinsonian motor deficits. *Neuron*, *73*(2), 347-359.

2. Jing, X., Miwa, H., Sawada, T., Nakanishi, I., Kondo, T., Miyajima, M., & Sakaguchi, K. (2012). Ephrin-A1-mediated dopaminergic neurogenesis and angiogenesis in a rat model of Parkinson's disease. *PloS one*, *7*(2), e32019

3. E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, A. Valencia. *EnrichNet: network-based gene set enrichment analysis*, Bioinformatics, 28(18):i451-i457, 2012

4. E. Glaab, R. Schneider, *PathVar: analysis of gene and protein expression variance in cellular pathways using microarray data*, Bioinformatics, 28(3):446-447, 2012

5. 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

6. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. *TopoGSA: network topological gene set analysis*, Bioinformatics, 26(9):1271-1272, 2010

7. 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

8. 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

9. 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

10. E. Glaab, R. Schneider, Comparative pathway and network analysis of brain transcriptome changes during adult aging and in Parkinson's disease, Neurobiology of Disease (2014), 74, 1-13

11. E. Glaab, Analysing Functional Genomic Data Using Novel Ensemble, Consensus and Data Fusion Techniques, PhD Thesis, Nottingham University, 2011

12. 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, in press