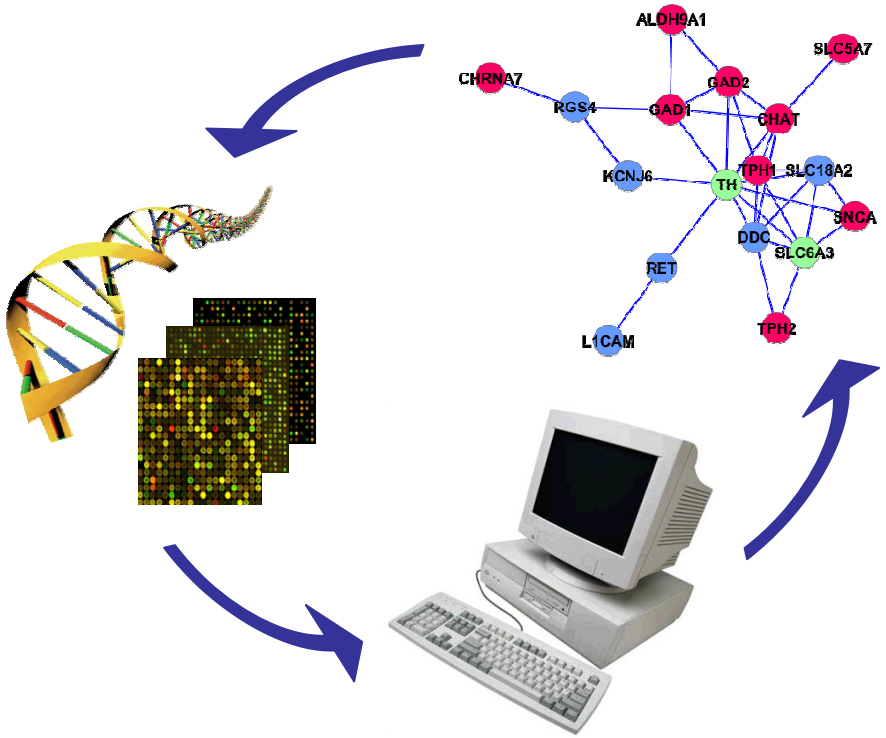


Integrating prior biological knowledge into omics data analysis

Enrico Glaab, Luxembourg Centre for Systems Biomedicine

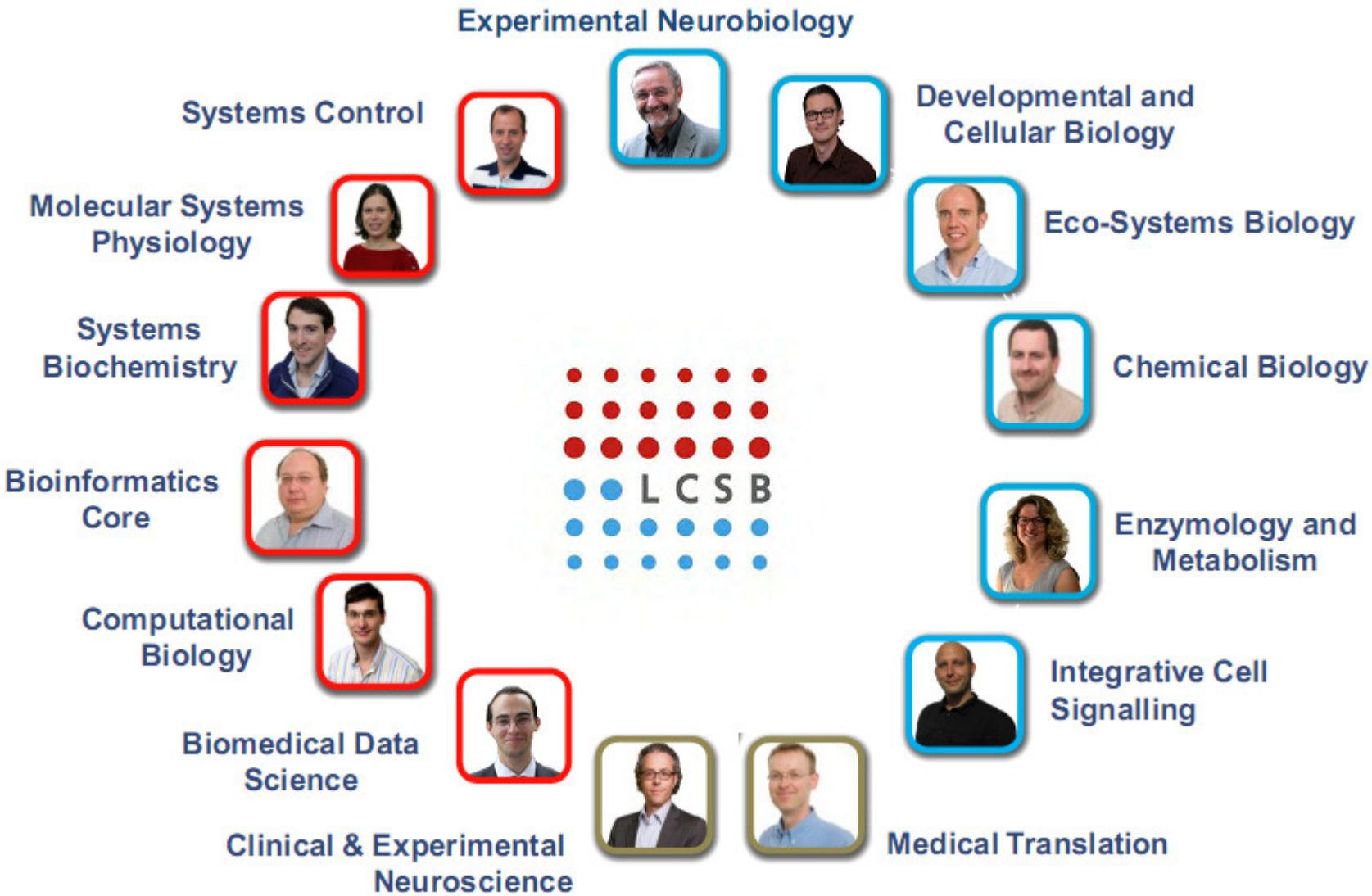
Outline

- Background and focus of research group
- Addressing common statistical challenges in omics data analysis
- Exploiting information from cellular pathways and molecular networks for machine learning analyses of omics data
- Summary & Discussion

Luxembourg Centre for Systems Biomedicine (LCSB)

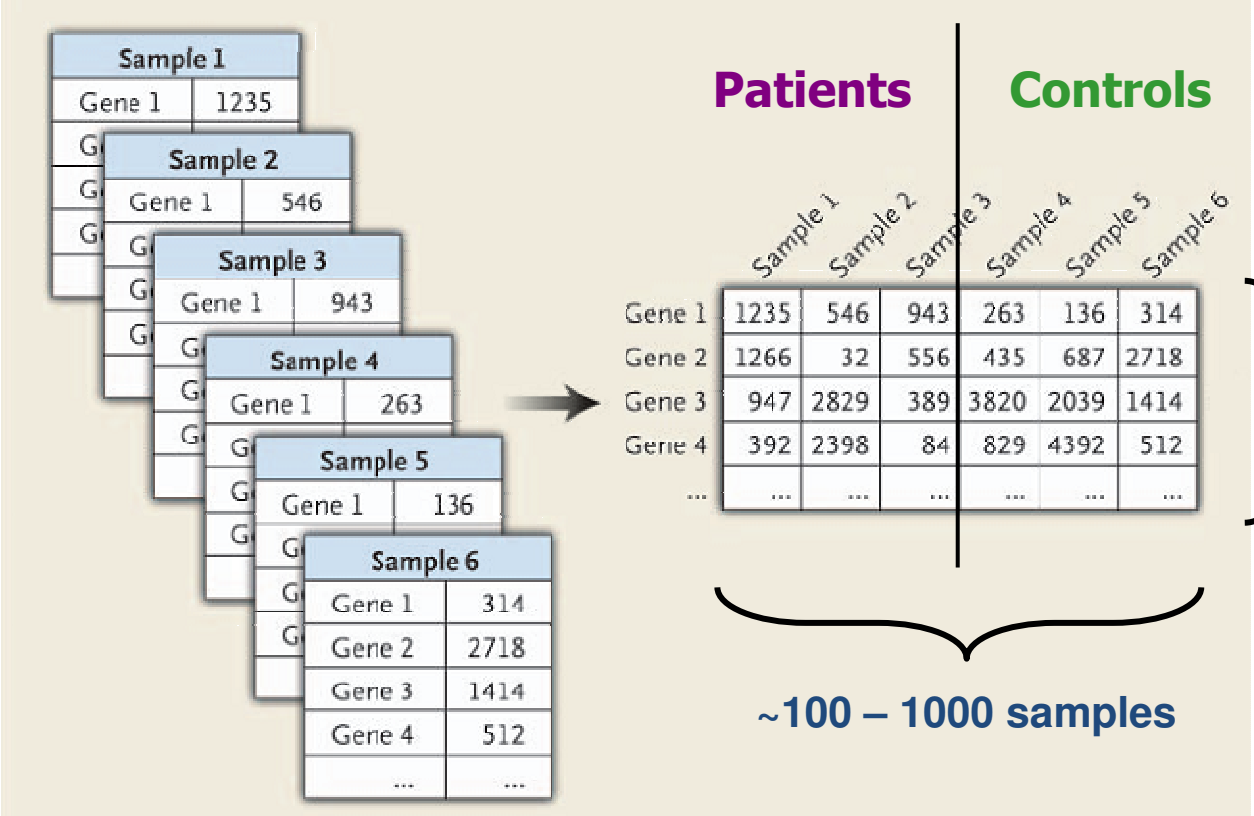
Luxembourg Centre
for Systems Biomedicine

LCSB – Research Groups & Interdisciplinarity



Research Focus & Main Goals

<u>Research focus</u>: Analysis of omics data from case/control studies



"Small N, Large P" problem

→ **<u>GOALS</u>: Interpret biological differences between patients and controls, identify candidate disease genes & biomarkers for validation**

Overview of machine learning analysis types for omics data

Unsupervised Analyses:
(no sample labels used)

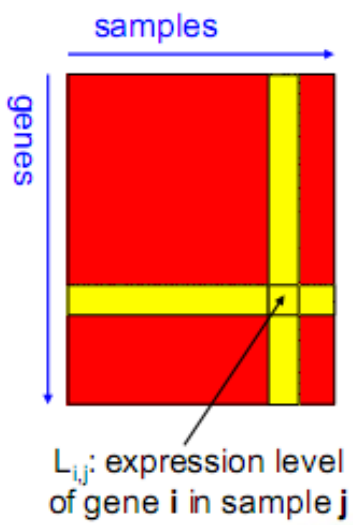
Complexity:

- Clustering samples (columns) *"hard"*
- Clustering biomolecules (rows) *"hard"*
- Bi-Clustering (rows & columns) *"hard"*

Supervised Analyses
(using labelled training data):

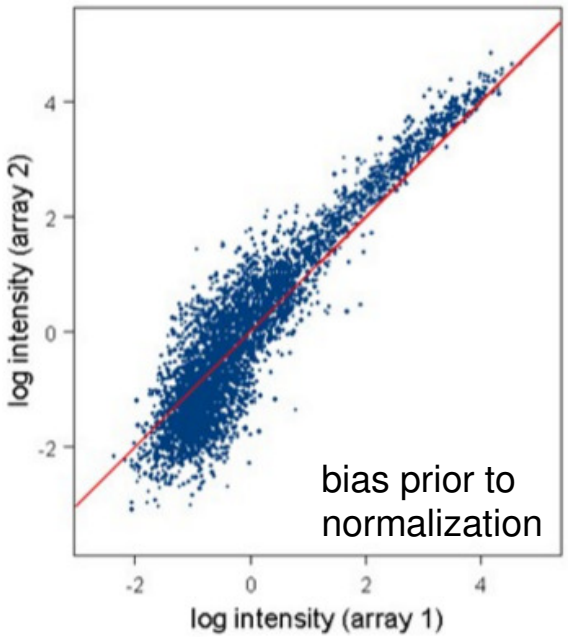
- Differential expression analysis *"easy"*
- Pathway enrichment analysis *"easy"*
- Network/causal reasoning canalysis *"hard"*
- Sample classification/regression *"hard"*
- Gene/protein function classification *"hard"*

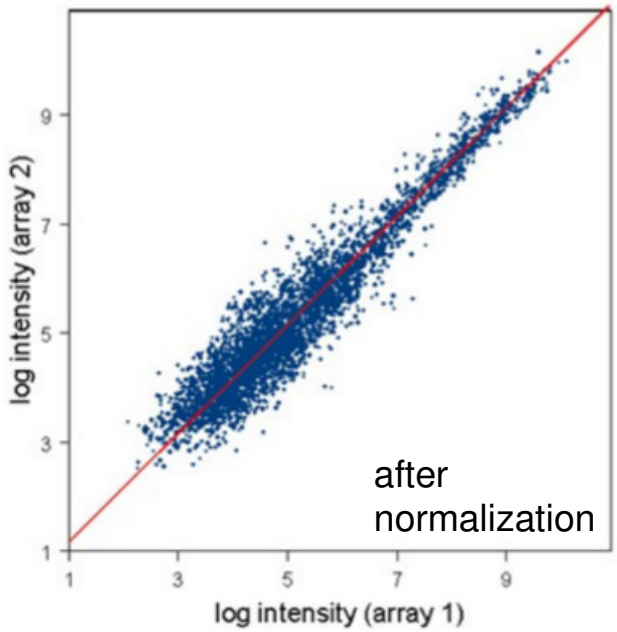
Example: High-throughput gene expression data



Common challenges for functional genomics data analysis (1)

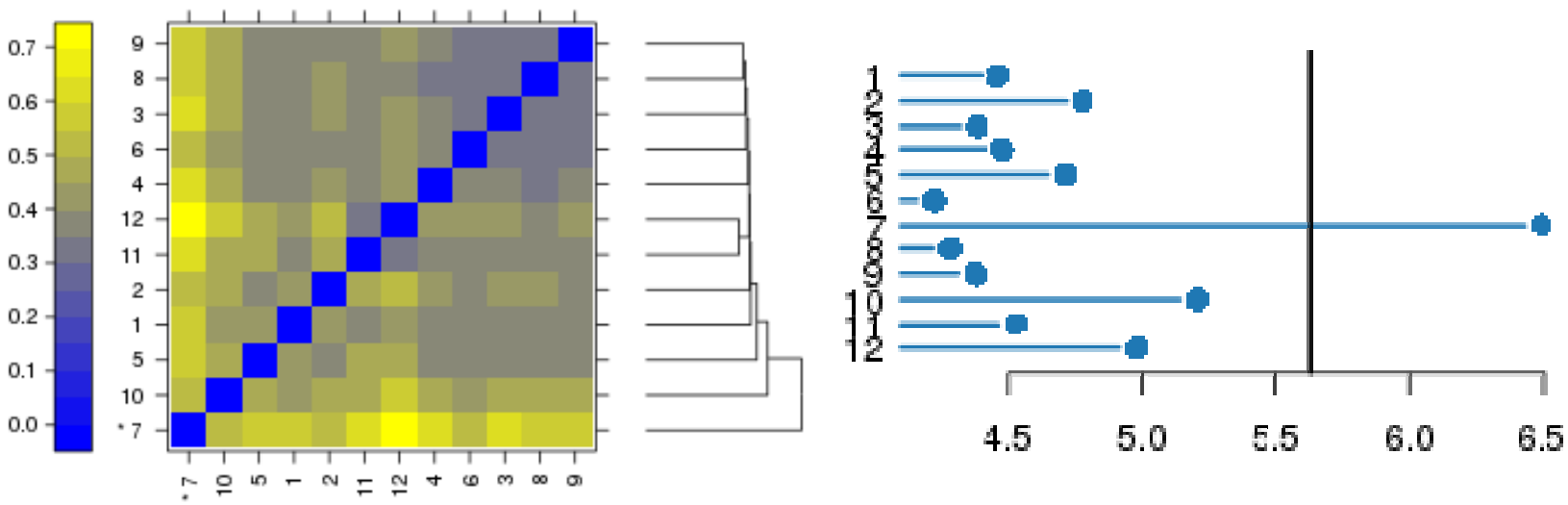
- Small number of samples in relation to large number of biomolecules ("Small N, Large P" problem) → "curse of dimensionality"
- Large number of uninformative and/or functionally redundant biomolecules (shared function & expression/activation pattern)
- Real signal shifted and scaled by additive and multiplicative noise





Common challenges for functional genomics data analysis (2)

- Outliers (among biomolecules or samples) and transcriptional amplification in sample subset
- Imbalances in no. of samples per condition (e.g. lack of control samples)
- Confounding factors and inadequate matching of patients & controls



False colour heat map (left) and bar chart (right) of distances between microarrays

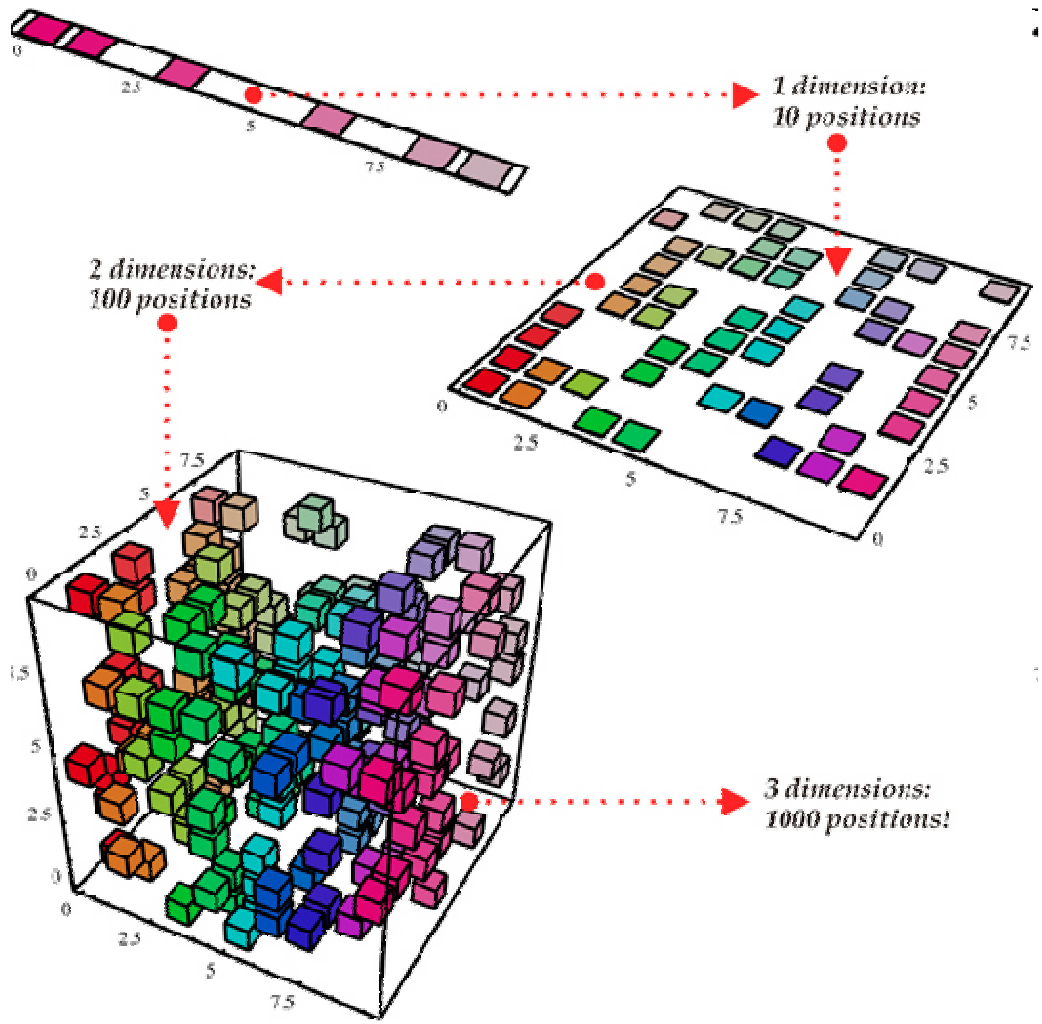
The “curse of dimensionality“

For increasing numbers of biomolecules/features:

- the space spanned by these features grows exponentially

  → the available data becomes sparse

  → more data points needed to train reliable diagnostic models



Strategies to address common issues in omics analysis

Statistical approaches:

- Use dimension reduction techniques, dedicated methods to exploit on-chip replicates and spike-in controls, model averaging methods for machine learning (e.g. ensemble classification, consensus clustering)

Data integration methods / exploiting prior biological knowledge:

- Apply meta-analyses across multiple studies, combining information across complementary omics & clinical data in supervised machine learning models
- Analyse and integrate data on the level of cellular pathways & molecular networks

Computer-assisted study design / power calculation:

- Design the study with a sufficient number of replicates per class/condition and balanced classes; reduce impact of confounding factors via algorithmic sample selection/matching

Using prior knowledge in omics data analysis - Overview

Data integration at the level of biomolecules:

- Exploit functional relationships between biomolecules:
  → cellular pathway membership
  → protein complex membership
  → interaction in gene regulatory or protein interaction networks

- Exploit biomolecular relationships across different omics:
  → genes encoding proteins
  → enzymes converting metabolites
  …

	Patient samples			Control samples		
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Gene 1	1235	546	943	263	136	314
Gene 2	1266	32	556	435	687	2718
Gene 3	947	2829	389	3820	2039	1414
Gene 4	392	2398	84	829	4392	512
…	…	…	…	…	…	…

Data integration at the level of samples:

- Exploit meta information for each sample (clinical data, sample quality, storage duration)
- Exploit correlation patterns across different omics data collected for the same samples

…

Pathway-based sample classification (PathVar software)

Motivation: Gene/protein expression alterations in diseases tend to be co-ordinated at the level of cellular pathways

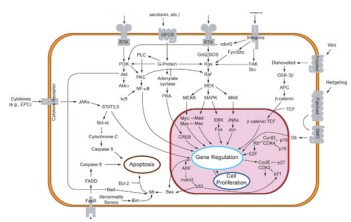
→ Use "**pathway fingerprints**" (weighted sums of gene expression levels from all pathway members) as candidate biomarkers with increased robustness

Pathway databases

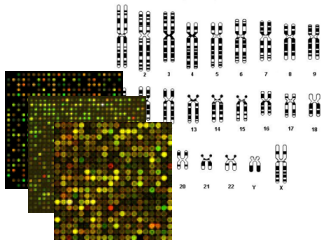
Cellular pathways

Omics data

Min. size: 10 genes

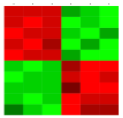


Map expression levels



Compute weighted sum of expression levels for each pathway (e.g. using PCA)

Pathway-level activity measures (fingerprints)



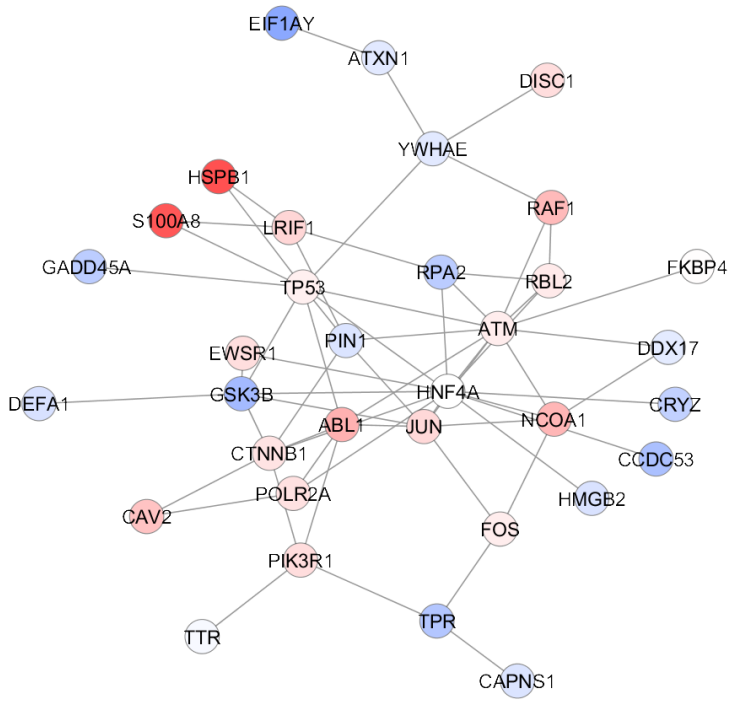
Pathway-level sample classification results

- Map omics data onto Gene Ontology (GO) biological processes (example: Parkinson's disease case/control post-mortem brain transcriptomics data)
- Use "pathway fingerprints" and a Support Vector Machine for classification (10-fold cross-validation; feature selection: empirical Bayes moderated t-statistic)

	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

Molecular networks as prior knowledge (GenePEN software)

Motivation: Disease-associated perturbations are often localized in biological networks. Finding these network clusters may help us to develop more robust biomarker models.



Example sub-network (Meta-analysis of 8 post-mortem microarray datasets for *substantia nigra* tissue) :

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

Question: How can we find clustered gene/protein groups efficiently, accounting for their predictivity and connectedness in the network?

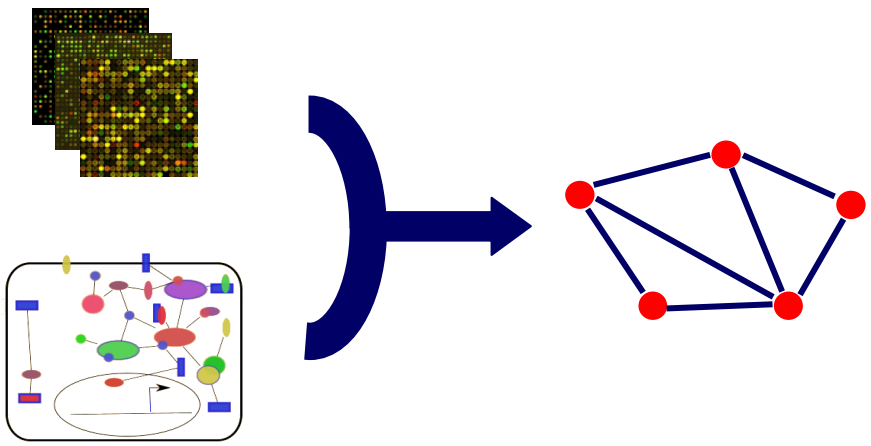
GenePEN - Workflow

Input:

- Omics dataset **X** (p rows = biomolecules, n columns = samples)
- Class labels **y** (e.g. “patient” or “control”)
- Table **A** of interactions/similarities between rows in X (e.g. protein-protein interactions)

Output:

- A subset of discriminative biomolecules (rows in X) representing a **connected** component in A (→ an altered sub-network) that provides a signature to classify new samples



GenePEN - Approach

Idea: Cast the feature selection as an optimization problem, maximizing two quantities:

- the estimated diagnostic prediction accuracy of the classifier
- the connectedness of selected features/biomolecules in the network

→ formulate an objective function (details not shown):

$$\min_{w} \underline{loss(w)} + \underline{\lambda} \cdot \underline{penalty(w)}$$

loss-function (minimize error) trade-off parameter penalty-function (network grouping)

→ Output after optimization procedure: A selection of features (***w***) that minimizes the objective function (features which minimize the prediction error and are well-grouped in the network)

GenePEN – Application to Parkinson‘s disease data

- **Parkinson’s disease test dataset**: Microarray gene expression data from *post mortem* brain samples (*substantia nigra*) of 43 PD patients and 50 controls (Zhang et al., 2005)

- **Network data**: Human genome-scale protein-protein interaction network constructed from 80,543 public, direct physical interactions between 10,042 proteins.

- **Comparison to other approaches**: GenePEN was compared against related methods with other penalty functions:

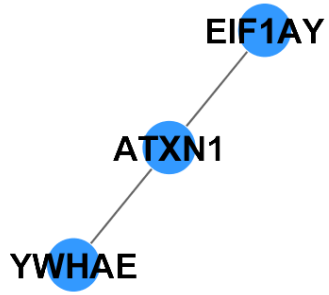
→ Lasso (Tibshirani, 1996) → sparse feature selection, but no feature grouping

→ Elastic Net (Zou & Hastie, 2005) → cannot account for external network data

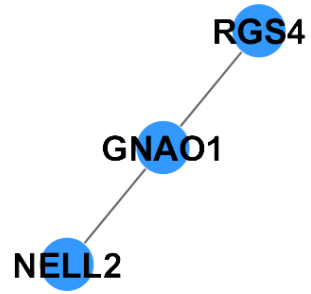
→ Pairwise Elastic Net (Lorbert, 2010 & 2013) → can take external network data into account to achieve a partial grouping of features

Comparison: Largest clusters found for 50 selected genes

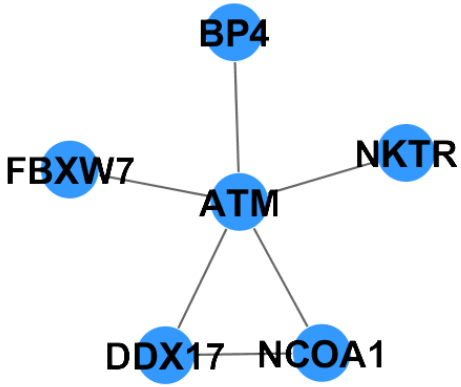
Lasso



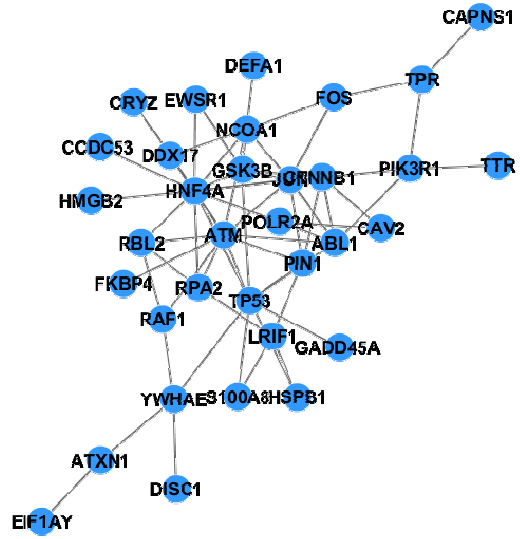
Elastic Net



PEN (2010)



PEN (2013)



GenePEN → cluster of 34 genes
(accuracy comparable to best alternative)

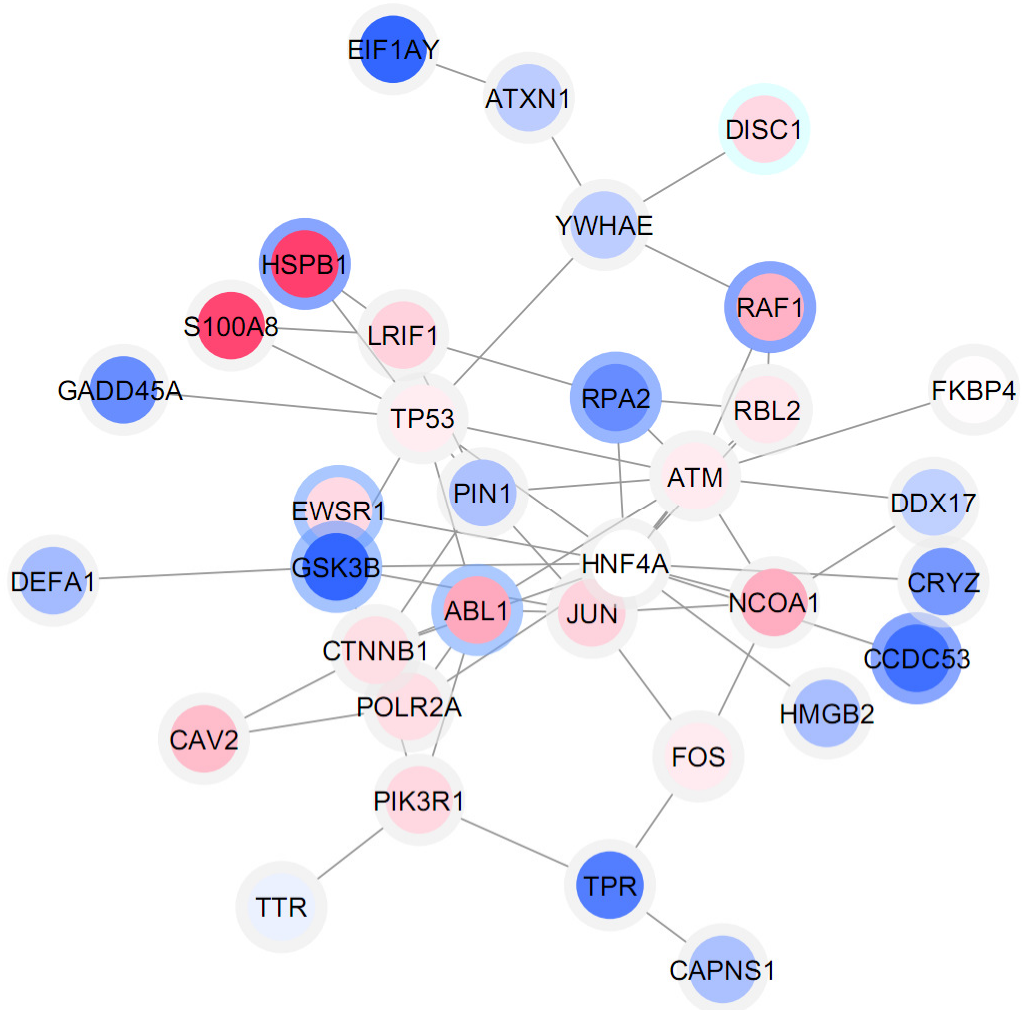
GenePEN: Biological analysis of predictive sub-networks

Largest connected graph component identified for Parkinson's disease:

- red = over-expressed in PD
  blue = under-expressed in PD
  node borders = individual statistical significance (from gray to blue with increasing significance)

- individually significant genes are significantly over-represented in the sub-network (p = 0.01)

- GSK3B contains polymorphisms associated with Parkinson's disease



Summary

- Many tools are available to address statistical challenges in omics data analysis

  → computer-guided **study design**, dedicated **normalization** methods,

  exploiting **prior knowledge** from molecular networks and pathways

- **PathVar** uses "pathway activity fingerprints" derived from omics data and known pathway definitions to build robust diagnostic machine learning models

- **GenePEN** discovers discriminative sub-networks for diagnostic sample classification and enables an interpretation of disease-associated molecular alterations at the network level

References

1. E. Glaab, Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification, Briefings in Bioinformatics (2015), 17(3), pp. 440
2. 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
3. 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), pp. 221
4. 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 (2016), in press (doi: 10.1007/s12035-016-0299-z)
5. E. Glaab, R. Schneider, *RepExplore: Addressing technical replicate variance in proteomics and metabolomics data analysis*, Bioinformatics (2015), 31(13), pp. 2235
6. E. Glaab, *Building a virtual ligand screening pipeline using free software: a survey*, Briefings in Bioinformatics (2015), 17(2), pp. 352
7. E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, A. Valencia. *EnrichNet: network-based gene set enrichment analysis*, Bioinformatics, 28(18):i451-i457, 2012
8. E. Glaab, R. Schneider, *PathVar: analysis of gene and protein expression variance in cellular pathways using microarray data*, Bioinformatics, 28(3):446-447, 2012
9. 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
10. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. *TopoGSA: network topological gene set analysis*, Bioinformatics, 26(9):1271-1272, 2010
11. 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
12. H. O. Habashy, D. G. Powe, E. Glaab, N. Krasnogor, J. M. Garibaldi, E. A. Rakha, G. Ball, A. R Green, C. Caldas, I. O. Ellis, *RERG (Ras-related and oestrogen-regulated growth-inhibitor) expression in breast cancer: A marker of ER-positive luminal-like subtype*, Breast Cancer Research and Treatment, 128(2):315-326, 2011
13. 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
14. 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
15. 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