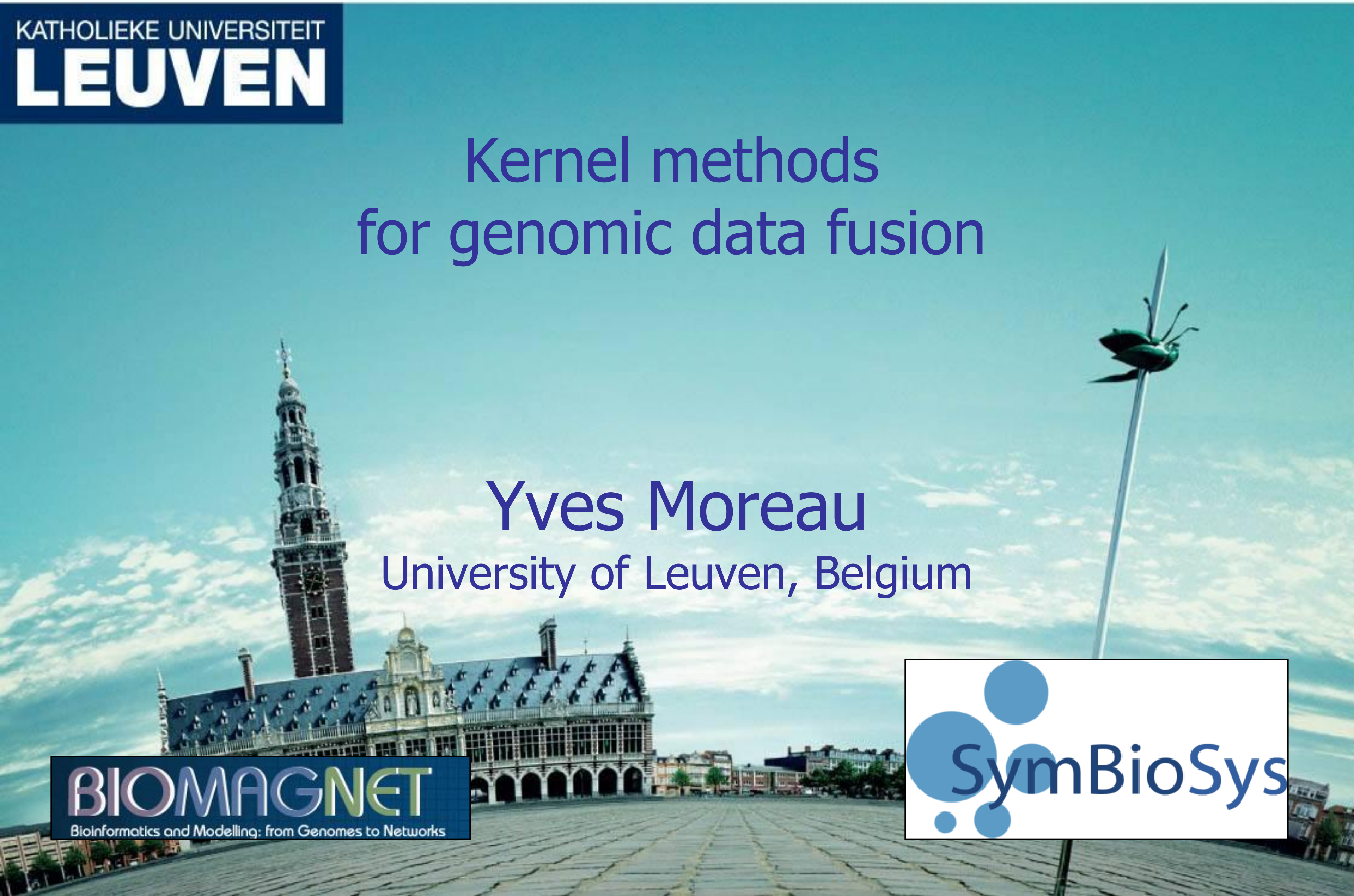
KATHOLIEKE UNIVERSITEIT
LEUVEN



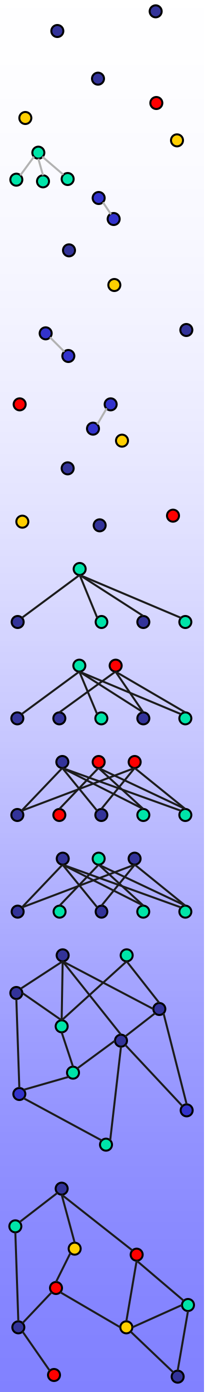
Kernel methods for genomic data fusion

Yves Moreau
University of Leuven, Belgium

BIOMAGNET
Bioinformatics and Modelling: from Genomes to Networks

SymBioSys

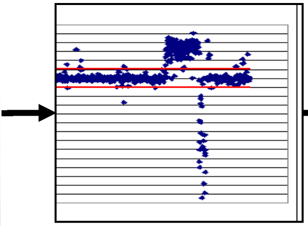
Disease gene discovery in rare congenital disorders



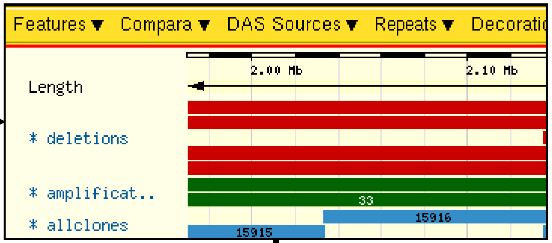
Patients with congenital & acquired disorders

CGH microarrays NGS sequencing

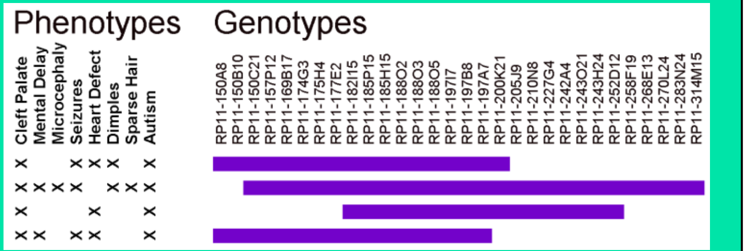
Statistical analysis



Location of chromosomal imbalances



Databasing



Phenotypes

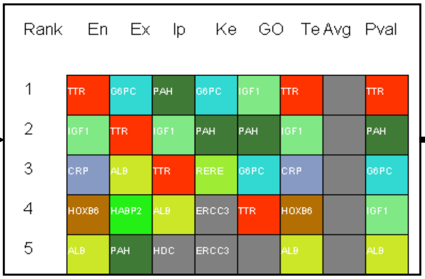
Cleft Palate	Mental Delay	Microcephaly	Seizures	Heart Defect	Dimples	Sparse Hair	Autism
X			X	X	X		X
X	X	X	X		X	X	X
X				X			X
X	X		X	X			X

Genotypes

- Map chromosomal abnormalities
- Improved diagnosis

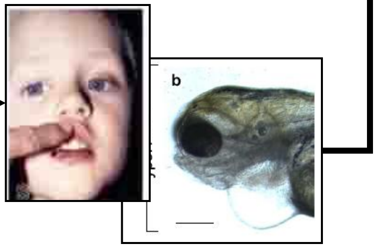
Discover new disease causing genes and explain their function

Prioritized candidate genes



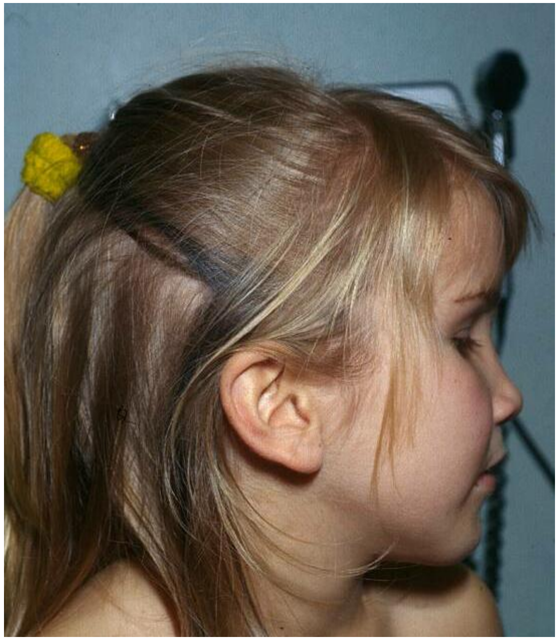
Rank	En	Ex	Ip	Ke	GO	Te	Avg	Pval
1	TTR	G6PC	PAH	G6PC	IGF1	TTR		TTR
2	IGF1	TTR	IGF1	PAH	PAH	IGF1		PAH
3	CRP	ALB	TTR	RERE	G6PC	CRP		G6PC
4	HOXB6	HABP2	ALB	ERCC3	TTR	HOXB6		IGF1
5	ALB	PAH	HDC	ERCC3		ALB		ALB

Validation

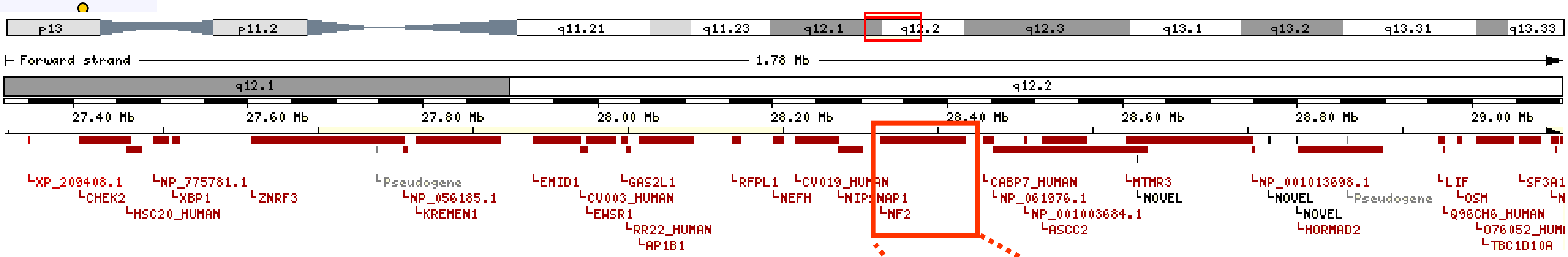


Genetic diagnosis

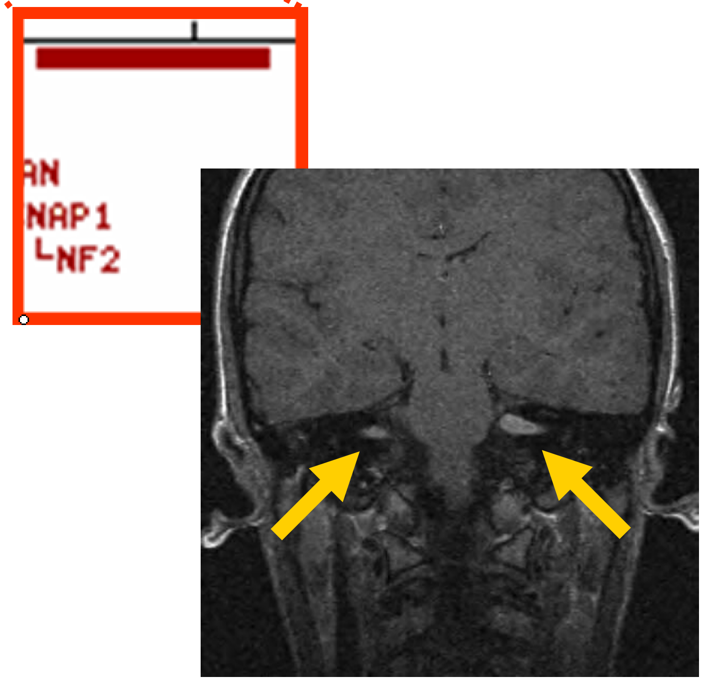
- Main medical goals
  - End diagnostic odyssey
  - Estimate risk for next pregnancy
  - Predict disease progression, life expectancy, etc.
- Patient - deletion del(22)(q12.2)
  - Pulmonary valve stenosis
  - Cleft uvula
  - Mild dysmorphism
  - Mild learning difficulties
  - High myopia



Deletion del(22)(q12.2)

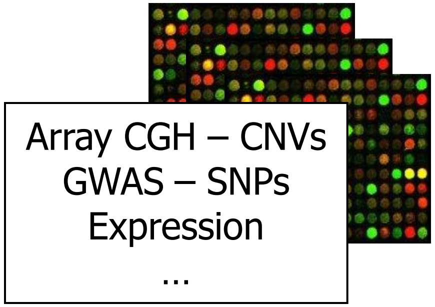


- Deletion on Chromosome 22
  - ~0.8Mb
- Deletion contains NF2
  - NF2 ↔ acoustic neurinomas
  - Benign tumor, BUT
    - Hard to diagnose
    - Severe complications



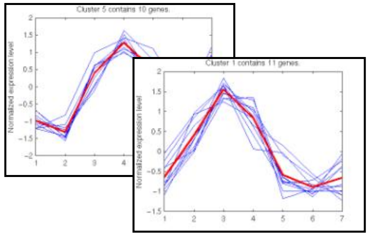
Candidate gene prioritization

High-throughput genomics

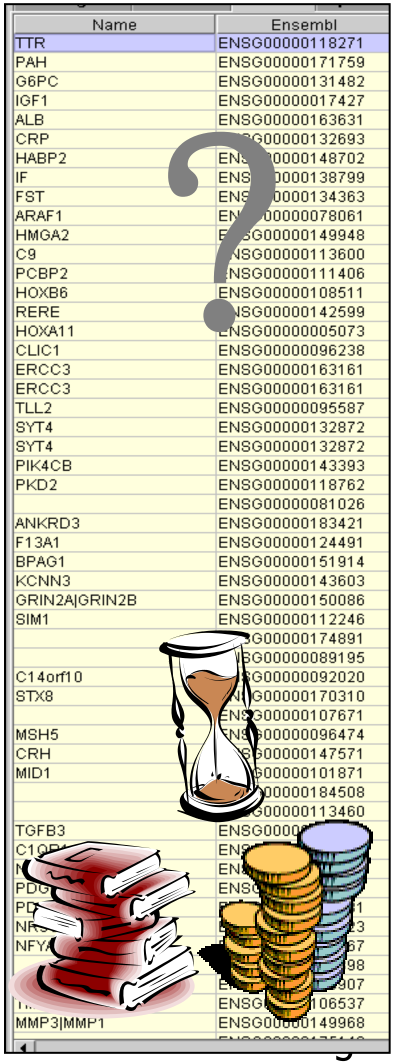


Array CGH – CNVs
GWAS – SNPs
Expression
…

Data analysis

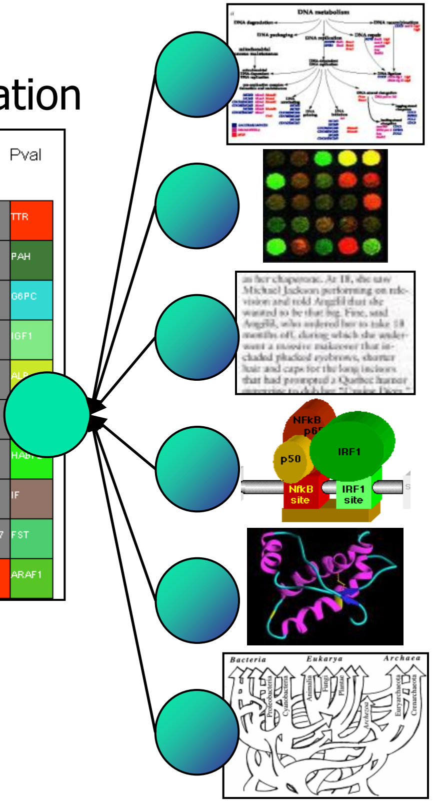


Candidate genes

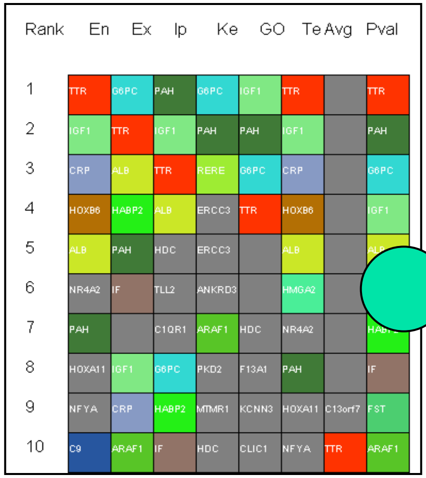


Name	Ensembl
TTR	ENSG00000118271
PAH	ENSG00000171759
G6PC	ENSG00000131482
IGF1	ENSG00000017427
ALB	ENSG00000163631
CRP	ENSG00000132693
HABP2	ENSG00000148702
IF	ENSG00000138799
FST	ENSG00000134363
ARAF1	ENSG00000078061
HMGA2	ENSG00000149948
C9	ENSG00000113600
PCBP2	ENSG00000111406
HOXB6	ENSG00000108511
RERE	ENSG00000142599
HOXA11	ENSG00000005073
CLIC1	ENSG00000096238
ERCC3	ENSG00000163161
ERCC3	ENSG00000163161
TLL2	ENSG00000095587
SYT4	ENSG00000132872
SYT4	ENSG00000132872
PIK4CB	ENSG00000143393
PKD2	ENSG00000118762
	ENSG00000081026
ANKRD3	ENSG00000183421
F13A1	ENSG00000124491
BPAG1	ENSG00000151914
KCNN3	ENSG00000143603
GRIN2A|GRIN2B	ENSG00000150086
SIM1	ENSG00000112246
	ENSG00000174891
	ENSG00000089195
C14orf10	ENSG00000092020
STX8	ENSG00000170310
	ENSG00000107671
MSH5	ENSG00000096474
CRH	ENSG00000147571
MID1	ENSG00000101871
	00000184508
	G00000113460
TGFB3	ENSG0000
C1QR	ENS
N	ENS
PDG	ENSG
PD	1
NR	23
NFYA	67
	98
	E 907
TI	ENSG 106537
MMP3|MMP1	ENSG00000149968

Information sources

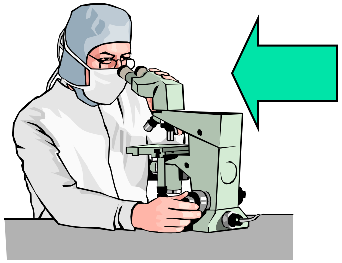


Candidate prioritization

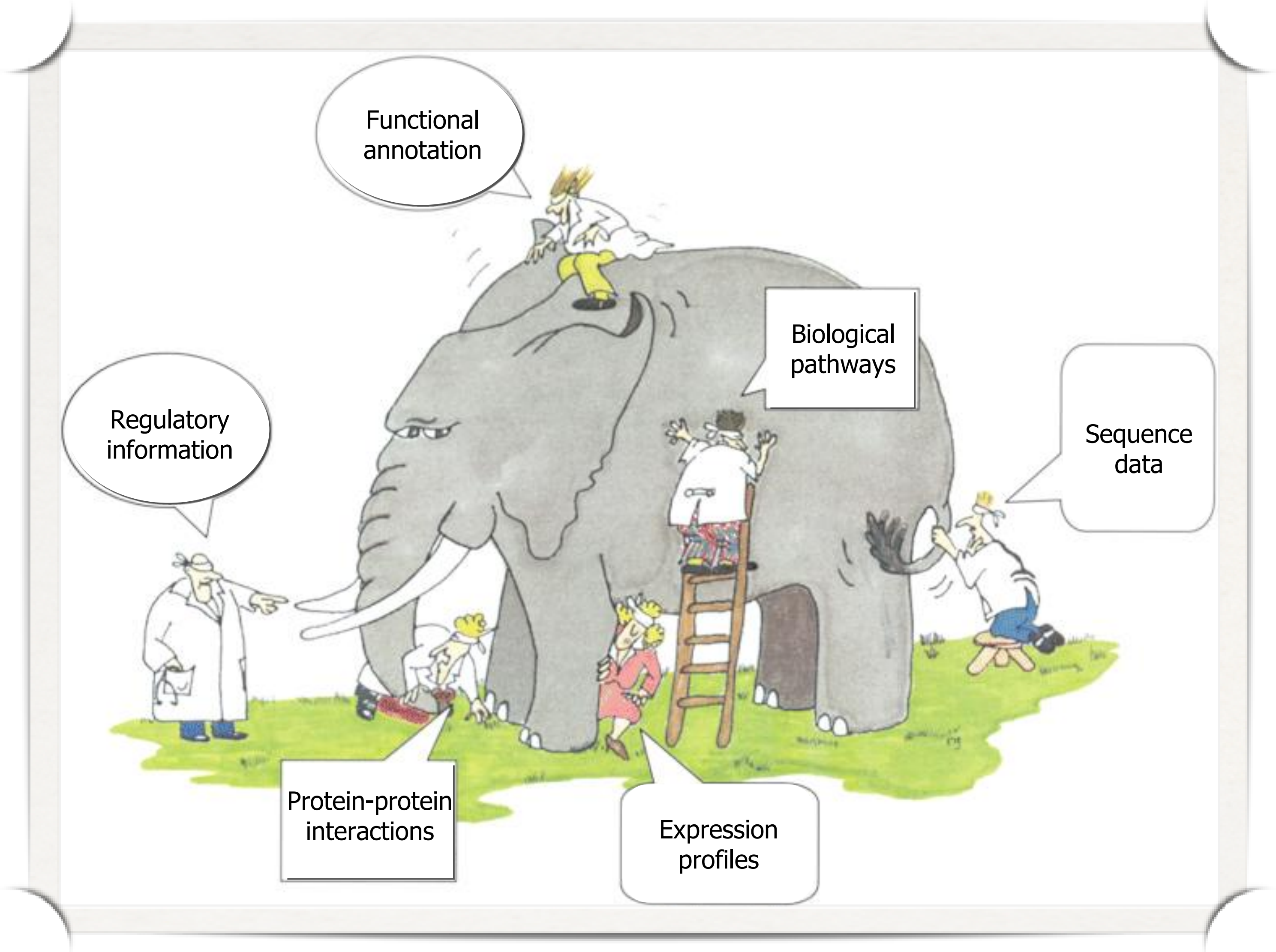


Rank	En	Ex	Ip	Ke	GO	Te	Avg	Pval
1	TTR	G6PC	PAH	G6PC	IGF1	TTR		TTR
2	IGF1	TTR	IGF1	PAH	PAH	IGF1		PAH
3	CRP	ALB	TTR	RERE	G6PC	CRP		G6PC
4	HOXB6	HABP2	ALB	ERCC3	TTR	HOXB6		IGF1
5	ALB	PAH	HDC	ERCC3		ALB		ALB
6	NR4A2	IF	TLL2	ANKRD3		HMGA2		[illegible]
7	PAH		C1QR1	ARAF1	HDC	NR4A2		HAB
8	HOXA11	IGF1	G6PC	PKD2	F13A1	PAH		IF
9	NFYA	CRP	HABP2	MTMR1	KCNN3	HOXA11	C13orf7	FST
10	C9	ARAF1	IF	HDC	CLIC1	NFYA	TTR	ARAF1

Validation



Data fusion



Functional annotation

Biological pathways

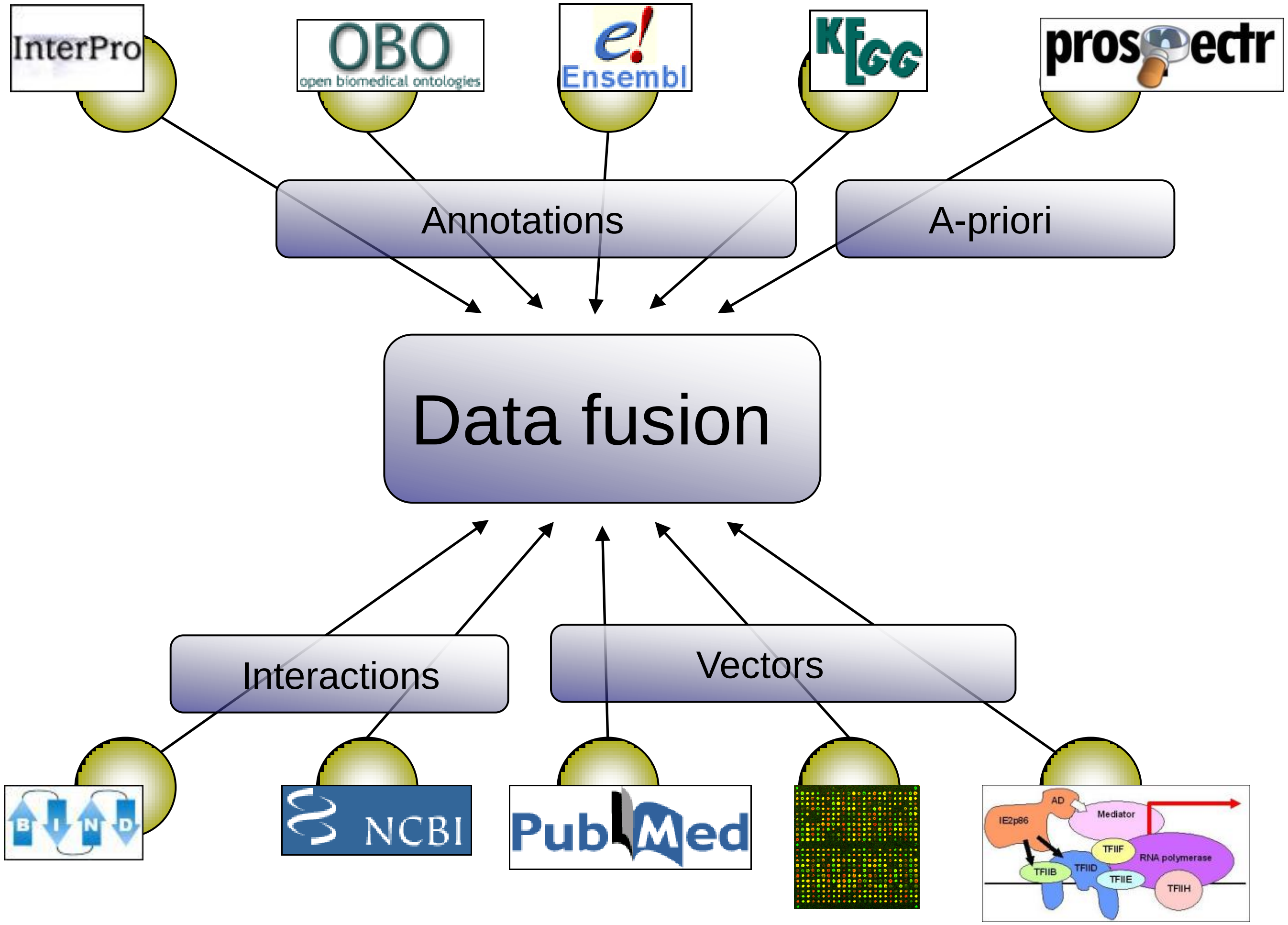
Regulatory information

Sequence data

Protein-protein interactions

Expression profiles

Challenge of heterogeneous data



Annotations

A-priori

Data fusion

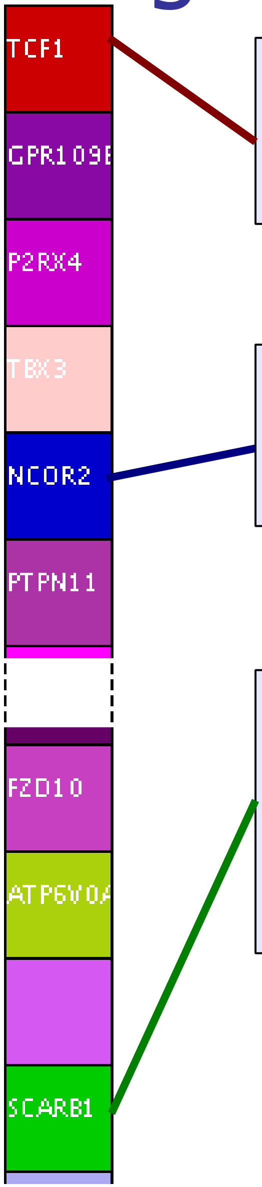
Interactions

Vectors

Prioritization by example

- Known/training genes
  - Type 2 diabetes: 21 known genes in OMIM, 118 known genes in GAD
  - Manually curated gene set from Elbers et al., 2007
    - ACDC, ADRA2A, ADRA2B, ADRB1, ADRB2, ADRB3, LEP, LEPR, NR3C1, UCP1, UCP2, UCP3, PPARG, KCNJ11, TCF7L2
- Candidate/test genes
  - Prioritizations of a known region (from Elbers et al., 2007)
    - 12q24: 327 candidates

Region 12q24: 327 candidates



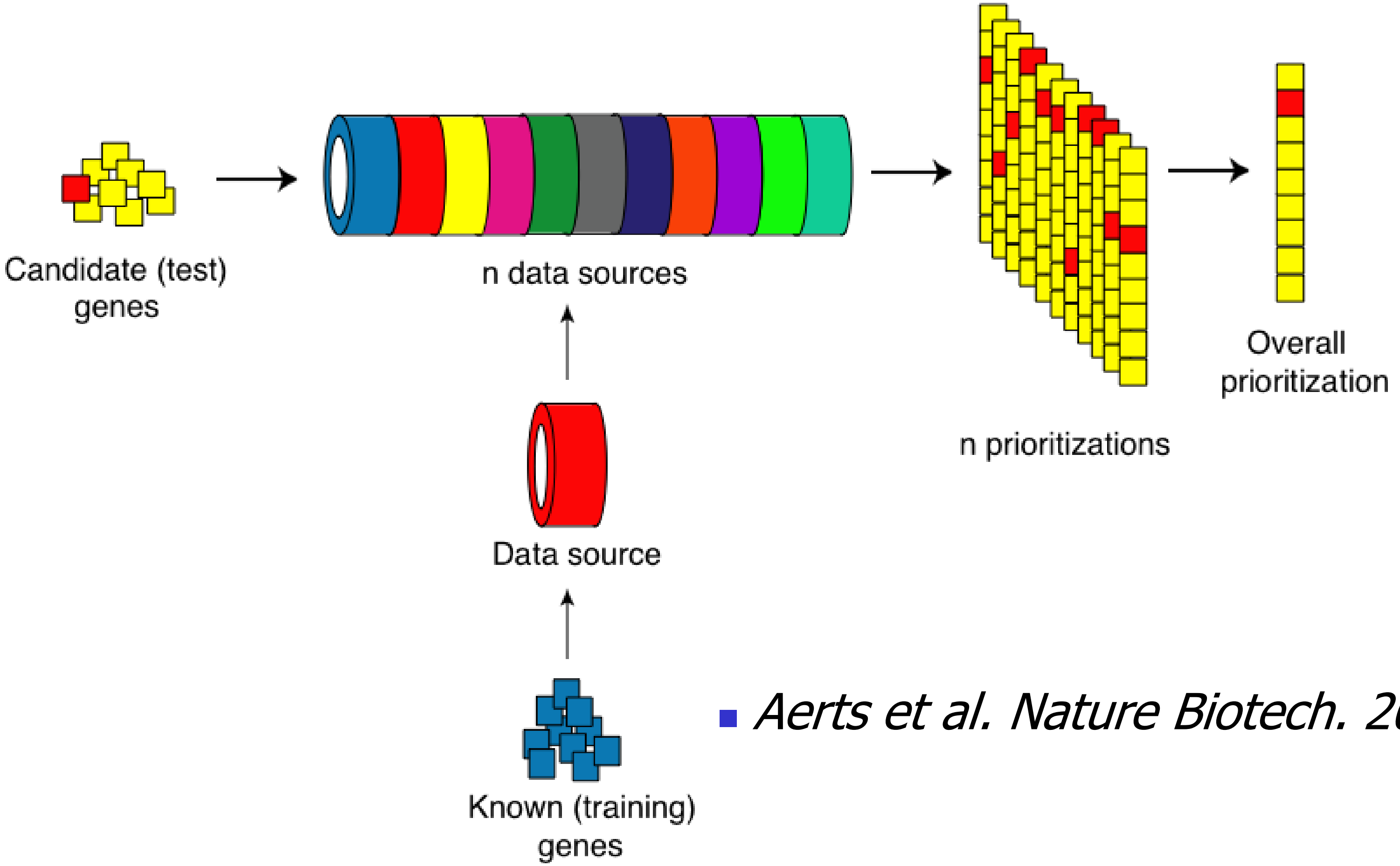
Responsible for MODY, an uncommon monogenetic form of early onset T2D.

NCOR2 has an important role in the adipocyte by inhibiting adipocyte differentiation via repression of PPAR-g activity.

Key component in the reverse cholesterol transport pathway. Genetically associated with differences in insulin sensitivity in healthy subjects

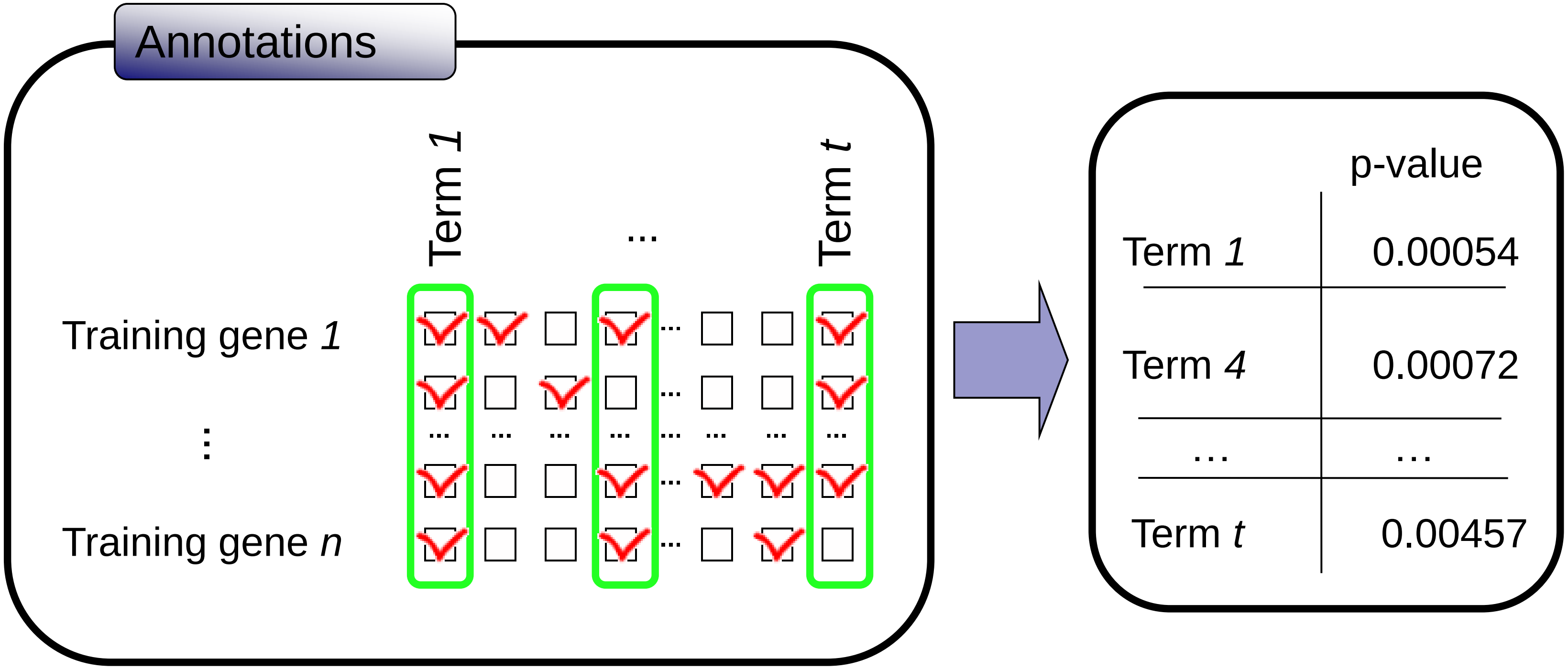
McCarthy *et al.* (2006), Cohen *et al.* (2006), Perez-Martinez *et al.* (2005)

Data fusion with order statistics



- *Aerts et al. Nature Biotech. 2006*

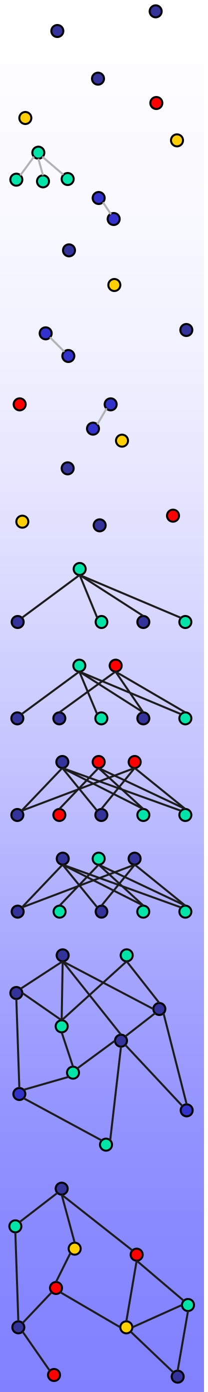
Training of an attribute submodel



	Term *1*				…			Term *t*
Training gene *1*	☑	☑	□	☑	…	□	□	☑
	☑	□	☑	□	…	□	□	☑
⋮	…	…	…	…	…	…	…	…
	☑	□	□	☑	…	☑	☑	☑
Training gene *n*	☑	□	□	☑	…	□	☑	□

	p-value
Term *1*	0.00054
Term *4*	0.00072
…	…
Term *t*	0.00457

- A term is over-represented if its frequency inside the training set is significantly larger than its frequency over the genome
  - Gene Ontology, Interpro, KEGG & EST submodels



Scoring of an attribute submodel

Annotations

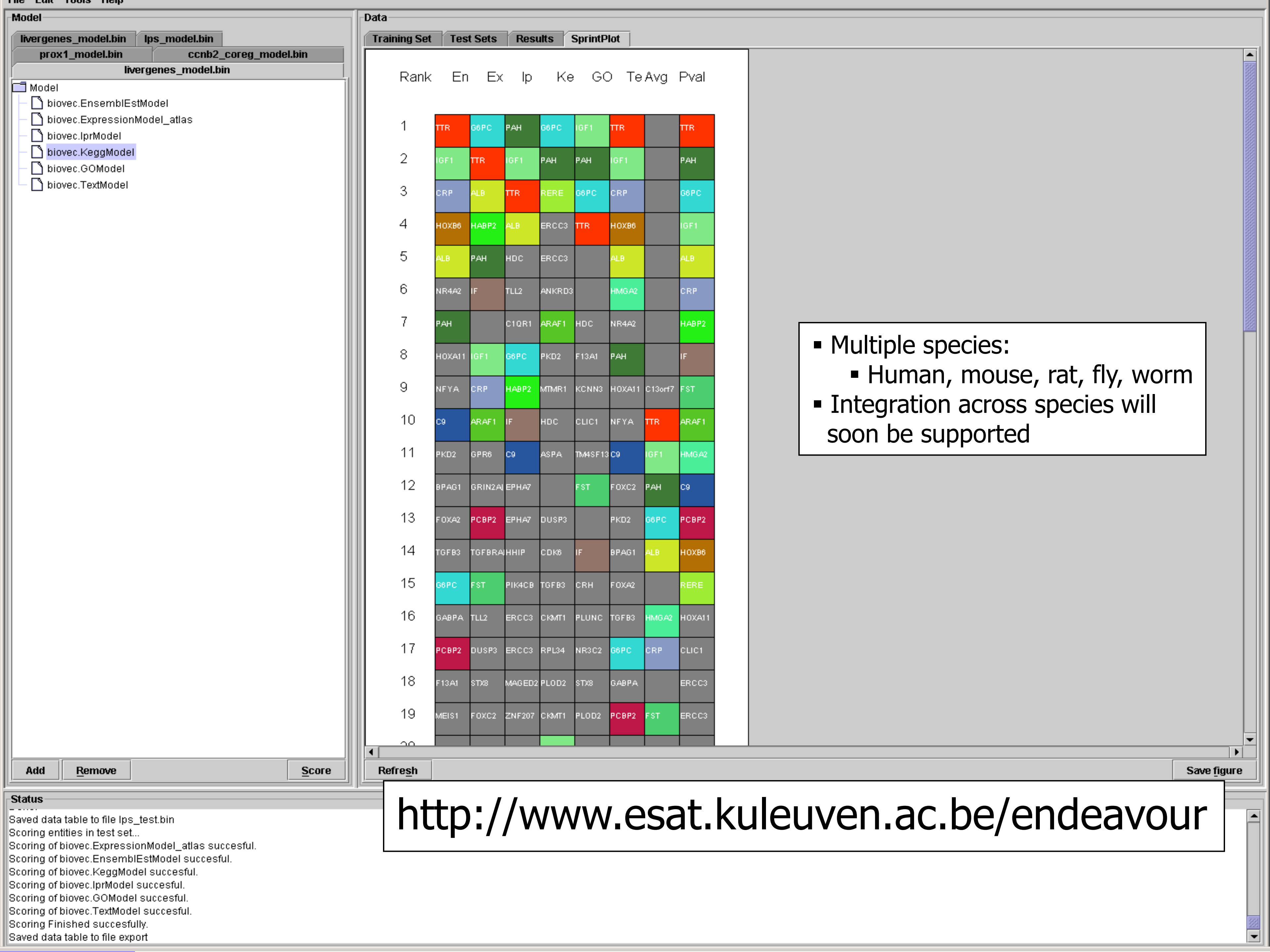
	p-value
Term *1*	0.00054
Term *4*	0.00072
…	…
Term *t*	0.00457

	Term *1*				…			Term *t*
Candidate *1*	✓	□	□	✓	…	□	✓	□
…	…	…	…	…	…	…	…	…
Candidate *m*	□	✓	✓	□	…	✓	□	□

	p-value
Candidate *1*	0.0005
…	…
Candidate *m*	1.0

Scoring derived from Fisher's omnibus statistic

- S = -2 ©$_i$ log p_i



Rank	En	Ex	Ip	Ke	GO	Te	Avg	Pval
1	TTR	G6PC	PAH	G6PC	IGF1	TTR		TTR
2	IGF1	TTR	IGF1	PAH	PAH	IGF1		PAH
3	CRP	ALB	TTR	RERE	G6PC	CRP		G6PC
4	HOXB6	HABP2	ALB	ERCC3	TTR	HOXB6		IGF1
5	ALB	PAH	HDC	ERCC3		ALB		ALB
6	NR4A2	IF	TLL2	ANKRD3		HMGA2		CRP
7	PAH		C1QR1	ARAF1	HDC	NR4A2		HABP2
8	HOXA11	IGF1	G6PC	PKD2	F13A1	PAH		IF
9	NFYA	CRP	HABP2	MTMR1	KCNN3	HOXA11	C13orf7	FST
10	C9	ARAF1	IF	HDC	CLIC1	NFYA	TTR	ARAF1
11	PKD2	GPR6	C9	ASPA	TM4SF13	C9	IGF1	HMGA2
12	BPAG1	GRIN2A	EPHA7		FST	FOXC2	PAH	C9
13	FOXA2	PCBP2	EPHA7	DUSP3		PKD2	G6PC	PCBP2
14	TGFB3	TGFBRA	HHIP	CDK6	IF	BPAG1	ALB	HOXB6
15	G6PC	FST	PIK4CB	TGFB3	CRH	FOXA2		RERE
16	GABPA	TLL2	ERCC3	CKMT1	PLUNC	TGFB3	HMGA2	HOXA11
17	PCBP2	DUSP3	ERCC3	RPL34	NR3C2	G6PC	CRP	CLIC1
18	F13A1	STX8	MAGED2	PLOD2	STX8	GABPA		ERCC3
19	MEIS1	FOXC2	ZNF207	CKMT1	PLOD2	PCBP2	FST	ERCC3

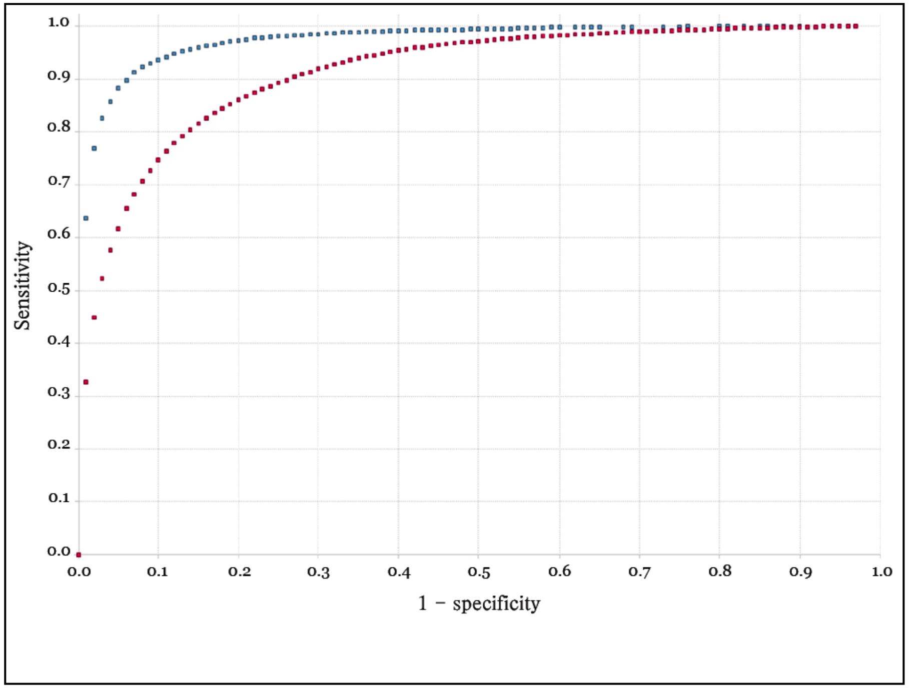
```
Saved data table to file lps_test.bin
Scoring entities in test set...
Scoring of biovec.ExpressionModel_atlas succesful.
Scoring of biovec.EnsemblEstModel succesful.
Scoring of biovec.KeggModel succesful.
Scoring of biovec.IprModel succesful.
Scoring of biovec.GOModel succesful.
Scoring of biovec.TextModel succesful.
Scoring Finished succesfully.
Saved data table to file export
```

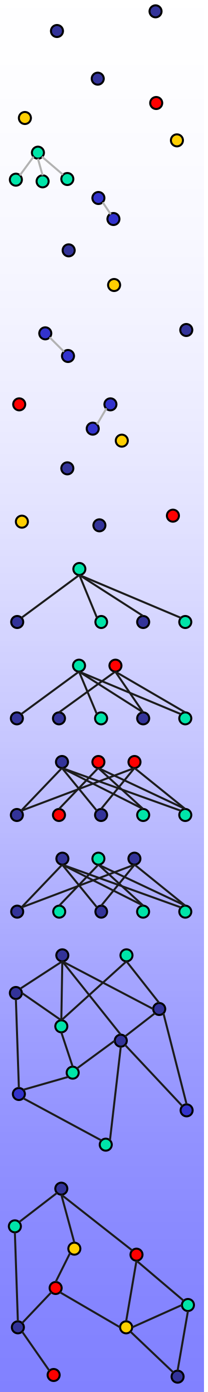
- Multiple species:
  - Human, mouse, rat, fly, worm
- Integration across species will soon be supported

http://www.esat.kuleuven.ac.be/endeavour

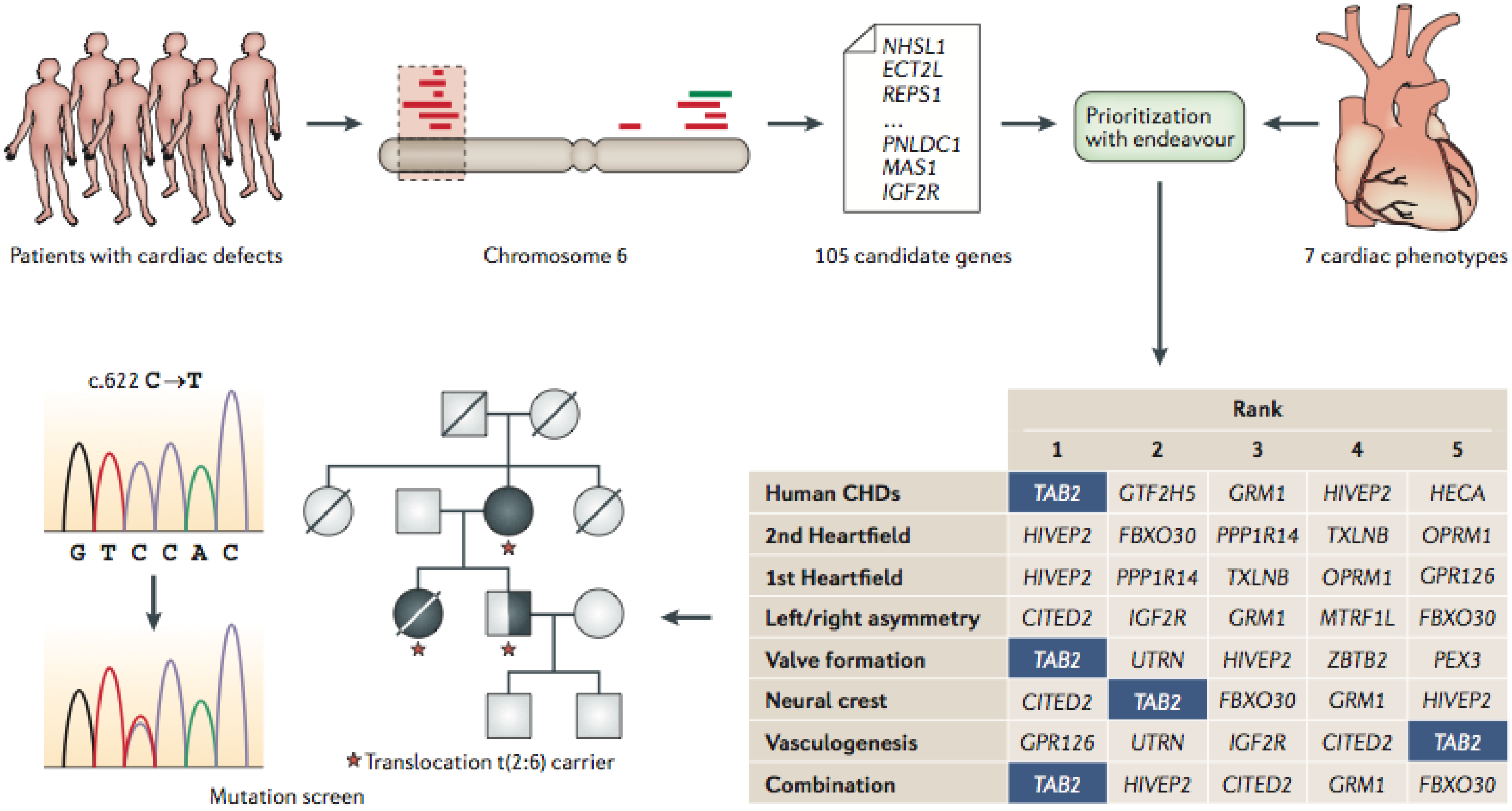
Large-scale statistical validation

- Evaluation by an independent third party (pharma)
- MetaCore pathway and disease maps
  - 454 pathway maps with 10,053 pathway genes
  - 833 disease maps with 12,699 disease genes
- ROC curve for ranks





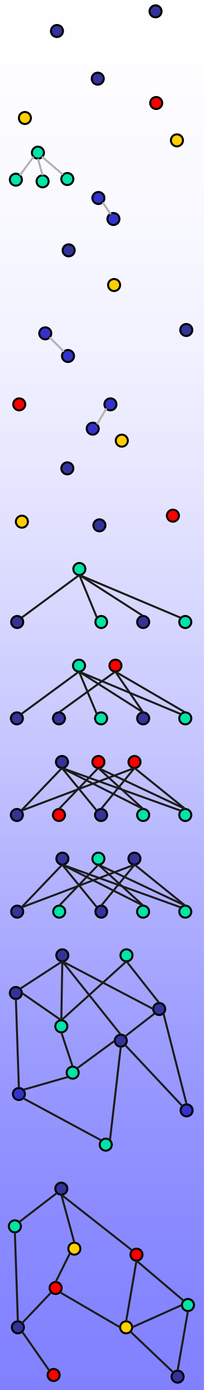
A novel locus for congenital heart defect on chromosome 6q24-25



	Rank				
	1	2	3	4	5
Human CHDs	TAB2	GTF2H5	GRM1	HIVEP2	HECA
2nd Heartfield	HIVEP2	FBXO30	PPP1R14	TXLNB	OPRM1
1st Heartfield	HIVEP2	PPP1R14	TXLNB	OPRM1	GPR126
Left/right asymmetry	CITED2	IGF2R	GRM1	MTRF1L	FBXO30
Valve formation	TAB2	UTRN	HIVEP2	ZBTB2	PEX3
Neural crest	CITED2	TAB2	FBXO30	GRM1	HIVEP2
Vasculogenesis	GPR126	UTRN	IGF2R	CITED2	TAB2
Combination	TAB2	HIVEP2	CITED2	GRM1	FBXO30

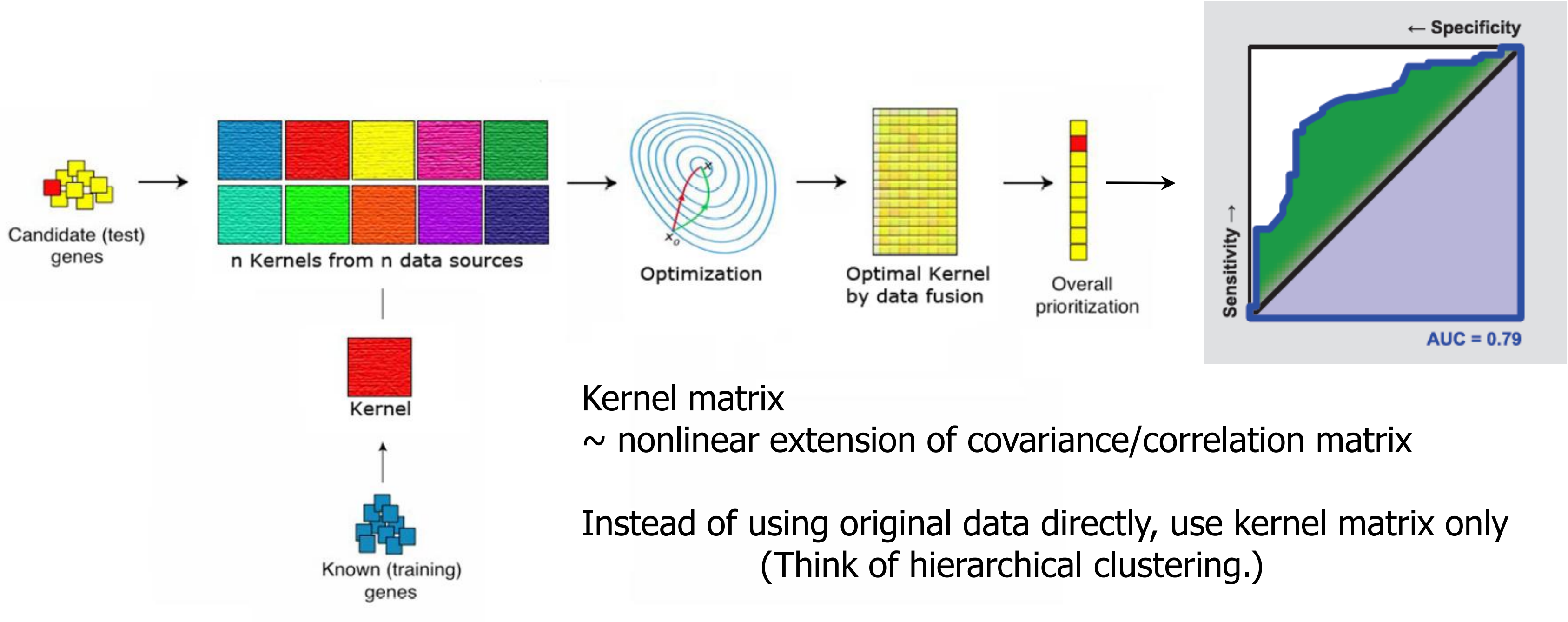
Nature Reviews | Genetics

Thienpont et al. *Am J Hum Genet.* 2010



Kernel methods for genomic data fusion

Kernel-based genomic data fusion



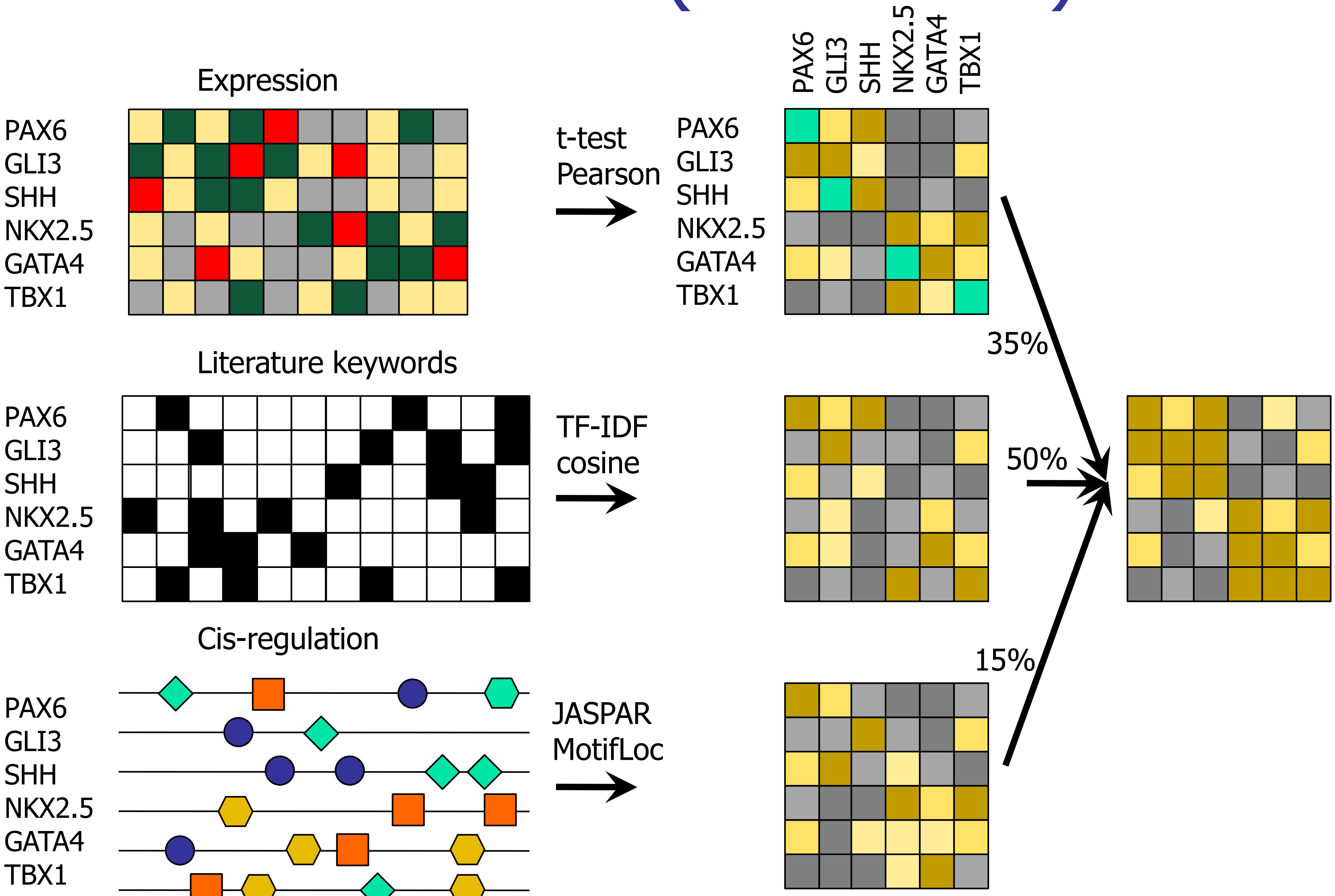
Kernel matrix
~ nonlinear extension of covariance/correlation matrix

Instead of using original data directly, use kernel matrix only
(Think of hierarchical clustering.)

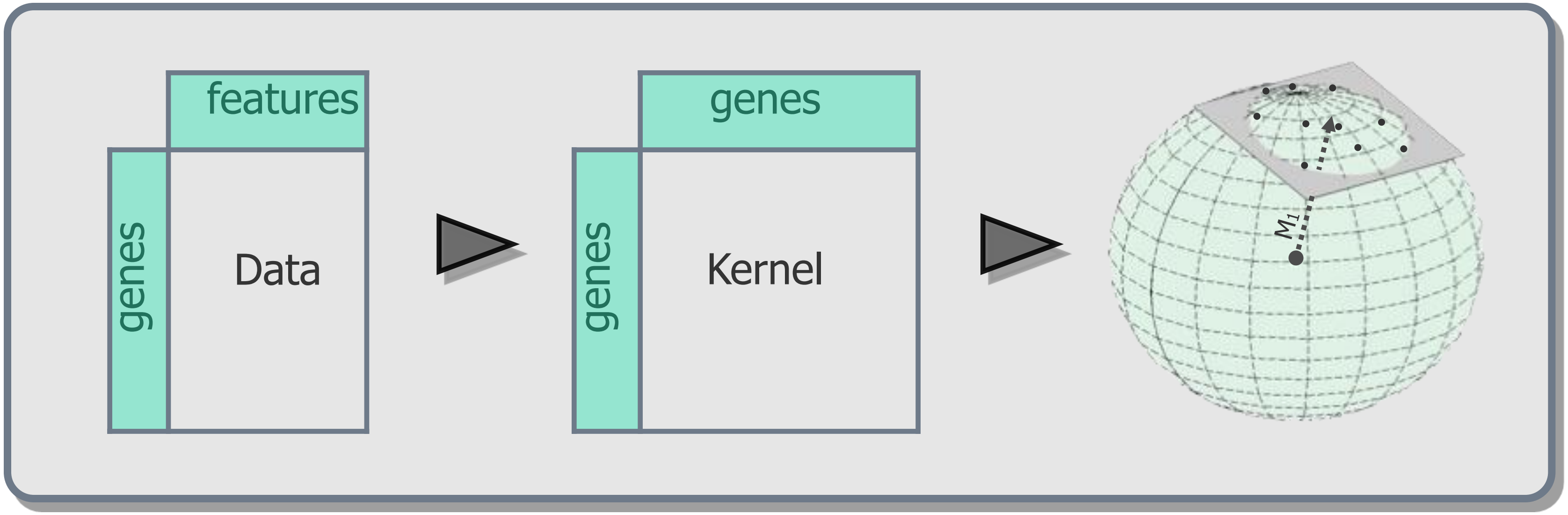
Advantage 1: kernel matrices form a single type of object, regardless of the heterogeneity of the original data types

Advantage 2: all machine learning methods can be applied to kernels (classification, clustering, prioritization, ranking, etc.)

Kernel data fusion (a.k.a. MKL)



Prioritization by novelty detection



features

genes

Data

genes

genes

Kernel

One-class support vector machine

$$\boxed{\text{P:}}\ \min_{\vec{w},\xi,\rho}\ \frac{1}{2}\vec{w}^T\vec{w} - \frac{1}{\nu l}\sum_{k=1}^{l}\xi_k - \rho$$

$$\text{s.t.}\ \ \vec{w}^T\phi(\vec{x}_k) \geq \rho - \xi_k,\ \ k = 1, \ldots, N$$

$$\xi_k \geq 0,\ \ k = 1, ..., N.$$

$\vec{w}$: the norm vector of the separating hyperplane

$\vec{x}_k$: the training samples

ν: a regularization term penalizing the outliers in the training samples

$\phi(\cdot)$: the feature map

ρ: the bias term

ξ_k: the slack variables

N: the number of training samples

$$\boxed{\text{D:}}\ \min_{\vec{\alpha}}\ \vec{\alpha}^T K \vec{\alpha}$$

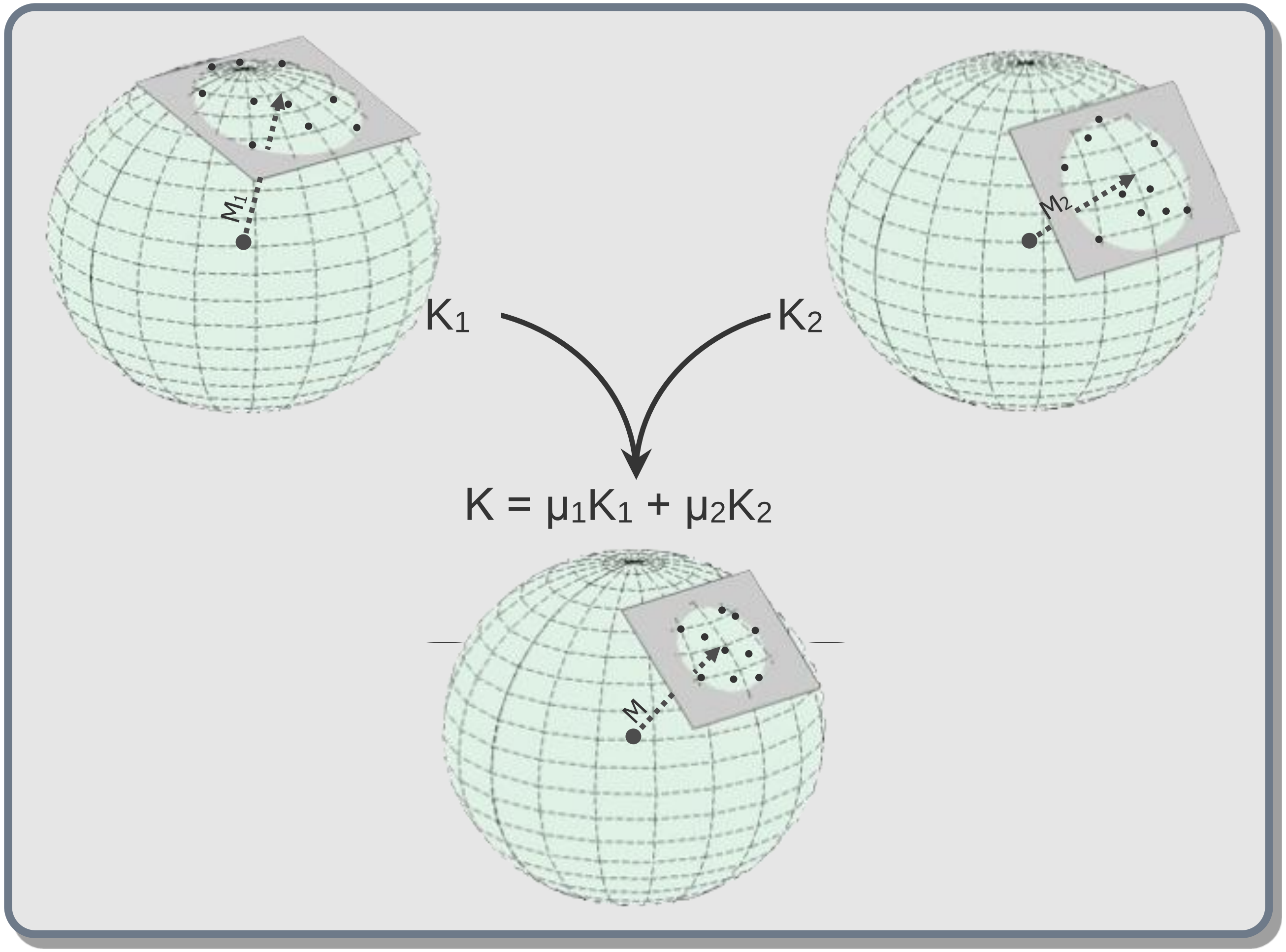
$$\text{s.t.}\ \ 0 \leq \alpha_k \leq \frac{1}{\nu N},\ \ k = 1, ..., N$$

$$\sum_{k=1}^{N}\alpha_k = 1,$$

α_k: the dual variables

K: the kernel matrix

Kernel fusion for novelty detection



$K = \mu_1 K_1 + \mu_2 K_2$

Kernel fusion in one-class SVM

- L_∞-norm kernel fusion (De Bie et al., 2007)

$$\begin{aligned} \min_{\vec{\alpha}} \quad & t \\ \text{s.t.} \quad & t \geq \vec{\alpha}^T K_j \vec{\alpha}, \ \ j = 1, ..., p \\ & 0 \leq \alpha_k \leq \frac{1}{\nu N}, \quad k = 1, ..., N \\ & \sum_{k=1}^{N} \alpha_k = 1, \end{aligned}$$

p: the number of kernel matrices
K_j: the j-th kernel matrix

- L_2-norm kernel fusion (Yu et al., 2009)

$$\begin{aligned} \min_{\vec{\alpha}} \quad & t \\ \text{s.t.} \quad & t \geq ||s_j||_2, \ \ j = 1, ..., p \\ & s_j \geq \vec{\alpha}^T K_j \vec{\alpha}, \ \ j = 1, ..., p \\ & 0 \leq \alpha_k \leq \frac{1}{\nu N}, \ \ k = 1, ..., N \\ & \sum_{k=1}^{N} \alpha_k = 1. \end{aligned}$$

s_j: dummy variables

L_2 vs. L_∞ kernel fusion

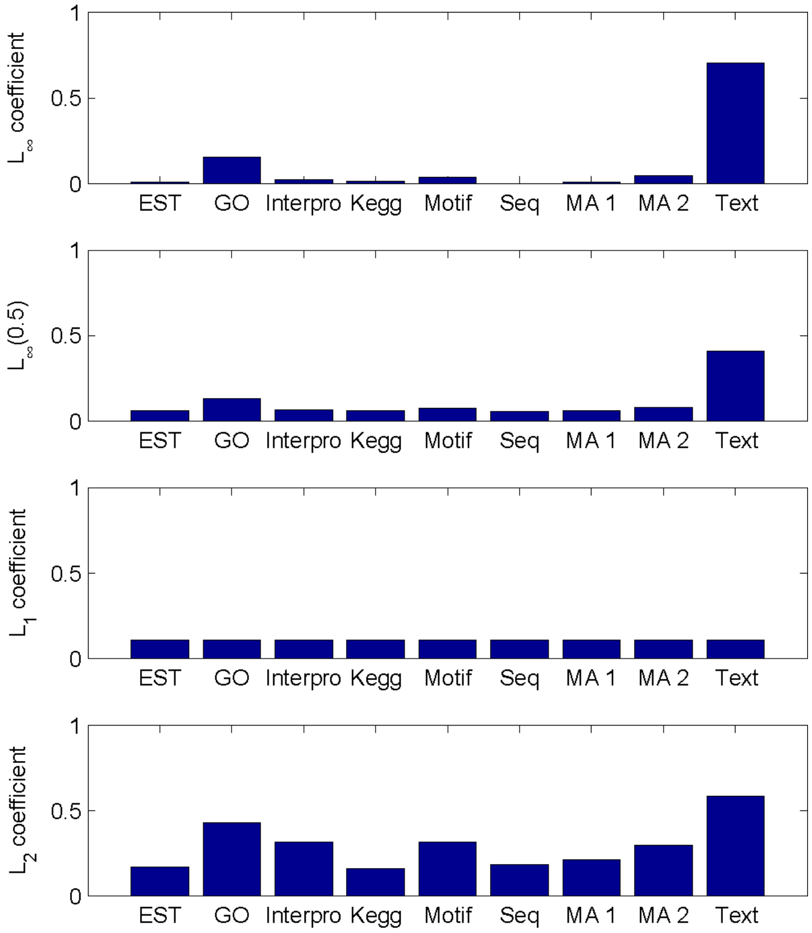
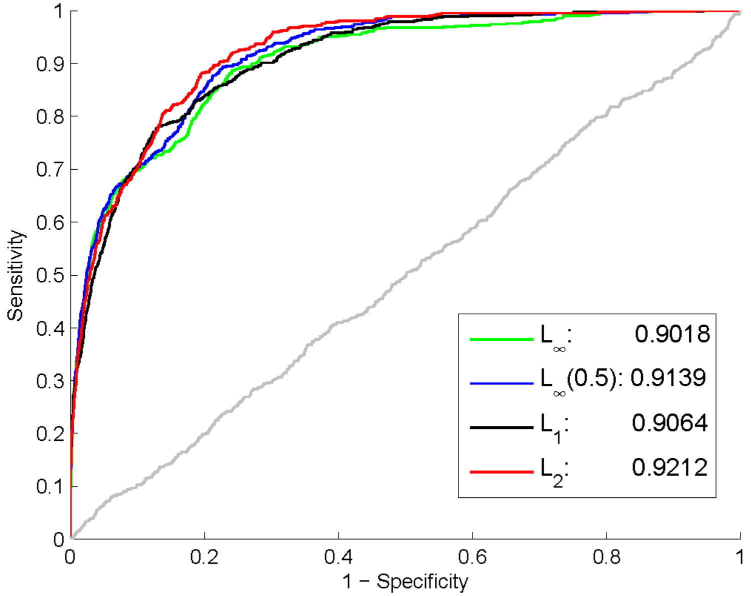


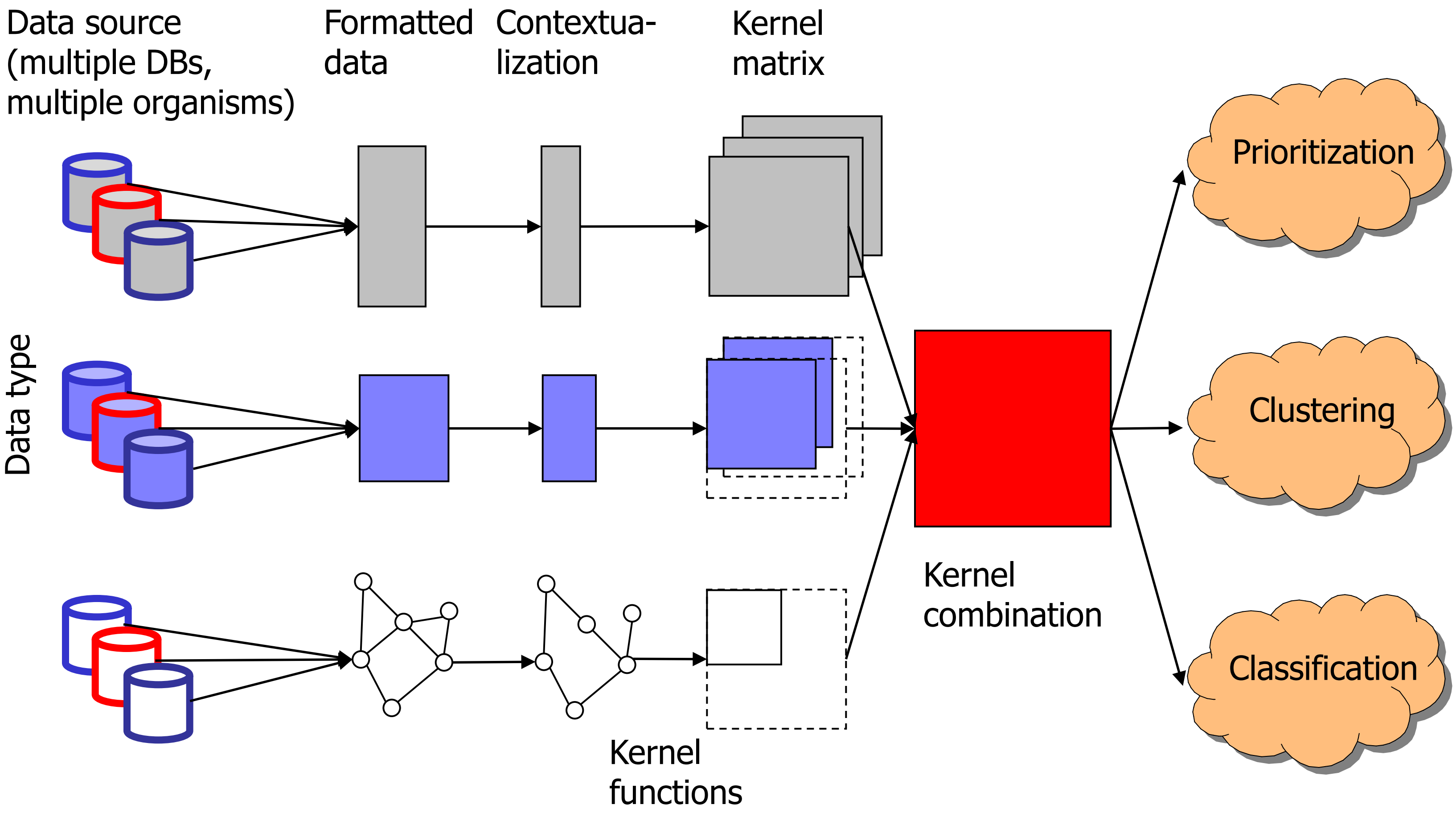
Table 1: AUC values of LOO performance evaluated from 20 random repetitions. The paired Spearman correlation scores indicate the similarities of rankings obtained by different approaches compared with the target rankings (denoted as -).

	AUC	corr	corr	corr	corr
L_∞	0.9045(0.0043)	-	0.94	0.66	0.82
$L_\infty(0.5)$	0.9176(0.0040)	0.94	-	0.82	0.92
L_1	0.9103(0.0035)	0.66	0.82	-	0.90
L_2	**0.9219(0.0034)**	0.82	0.92	0.90	-



A framework for kernel data fusion

Kernel data fusion



ETkL: Extract, Transform, Kernelize, Learn

- Systematic multi-tier framework for data integration
  - Resembles multi-tier architecture of complex IT systems and Extract-Transform-Load methodology of datawarehousing
    1. Database / web service sources
    2. Data reconciliation, cleaning, and warehousing, etc.
    3. Scaling, normalization, feature selection, etc.
    4. Computation and storage of kernels
    5. Learning
  - May require feedback loops (e.g., feature selection)
- Scale up to large, heterogeneous databases
- 20,000 x 20,000 kernel matrices are ugly animals

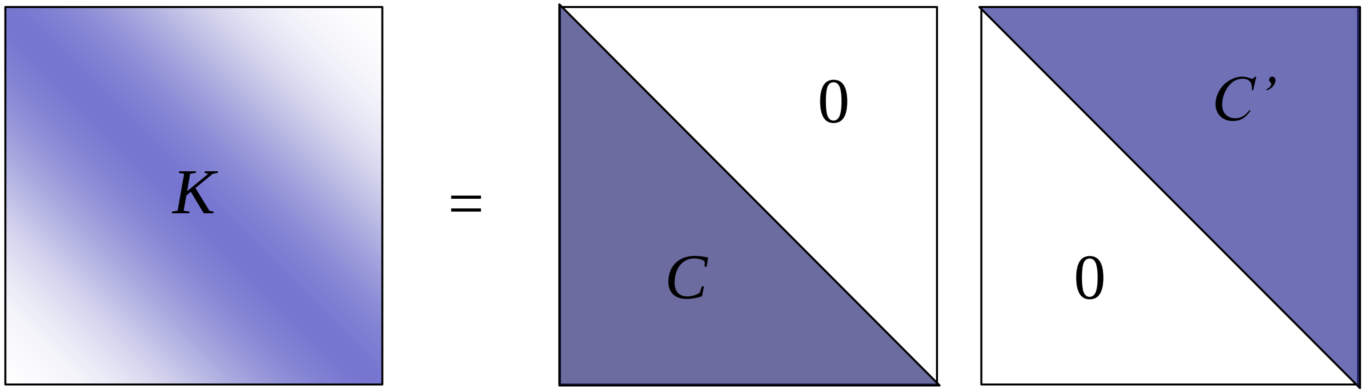
The No-Voodoo principle

- Given a data matrix D for a learning problem, the no voodoo principle states that, in the absence of prior knowledge or arbitrary assumptions, no information can be extracted about the problem except the information provided by the data matrix
  - In particular, no information can be created that wasn’t initially present in the data
    - No amount of bagging, random projection, nonlinear high-dimensional feature map, etc. can extract information that was not present in the data (except through the implicit or explicit injection of constraints into the problem)
  - If two frameworks represent data in ways that are related in a one-to-one fashion, there is nothing that prevents the development of methods with identical accuracy (e.g., random projections vs. spectral methods)
  - If one method outperforms another on a given problem (remember the no free lunch theorem), it is because the methods are more or less efficient (in particular, in terms of generalization performance vs. retrospective accuracy) at capturing the available information or because the methods incorporate explicit or implicit constraints that are more or less relevant to the given learning task

Handling large kernel matrices

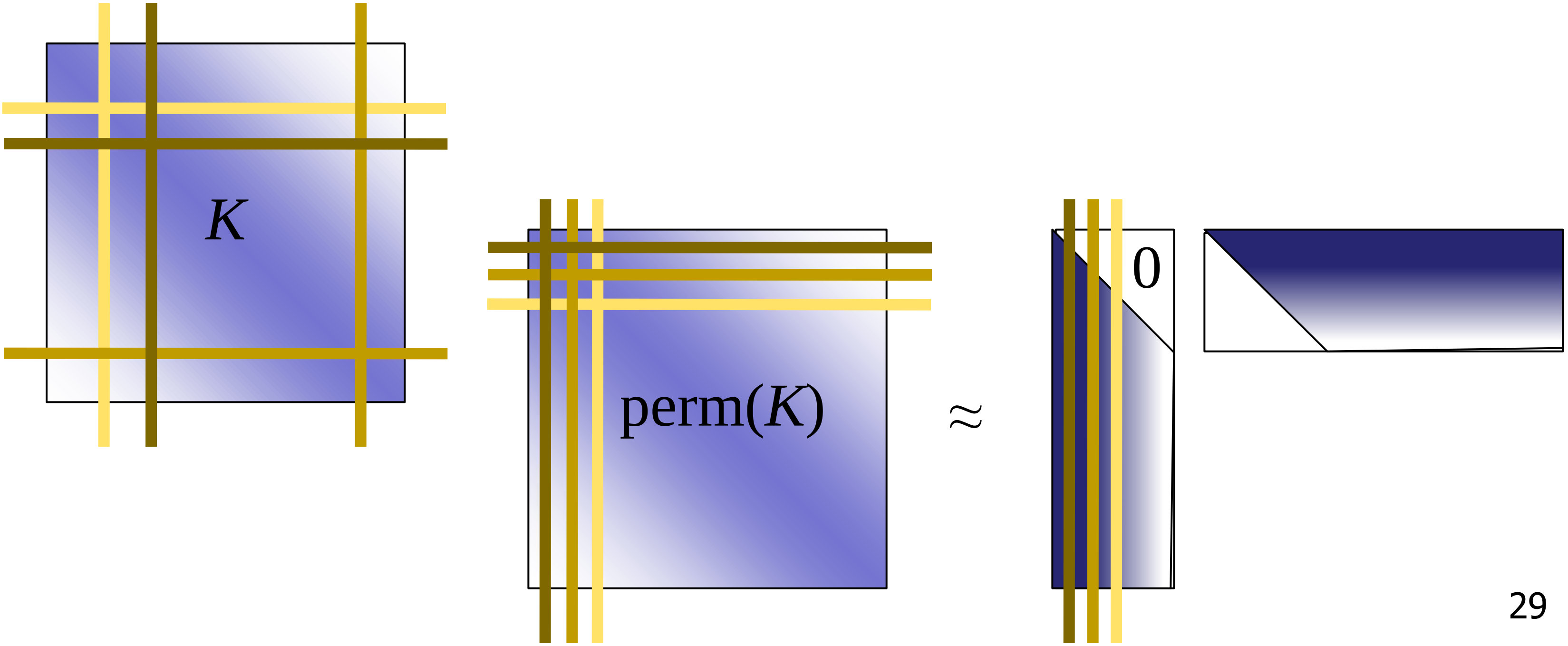
- One way to handle large kernel matrices is via low-rank approximations
  - Store $r \times n$ instead of $n \times n$
- Cholesky decomposition
  - K symmetric positive definite

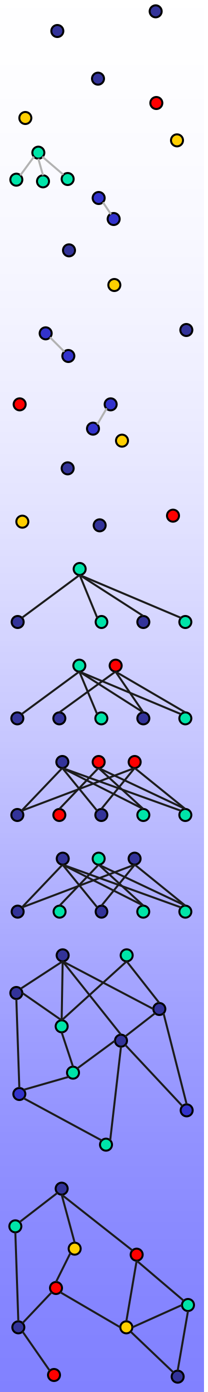
$$\exists C(\text{lower triangular} \;\; \& \;\; \text{unique}) : K = CC'$$



Incomplete Cholesky decomposition

- Incomplete Cholesky
  - K symmetric positive semidefinite
  - Limit to rank $r \leq \mathrm{rank}(K)$
  - Add pivoting to capture more informative rows/columns first
  - Limit information loss to e.g. 5%



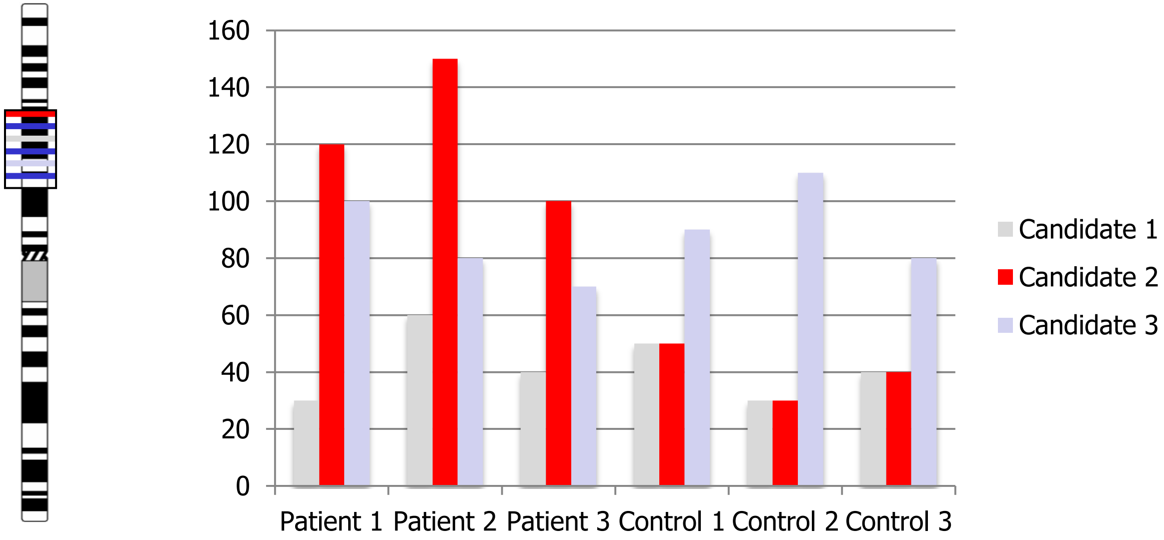


What if no or few genes known for a disease?

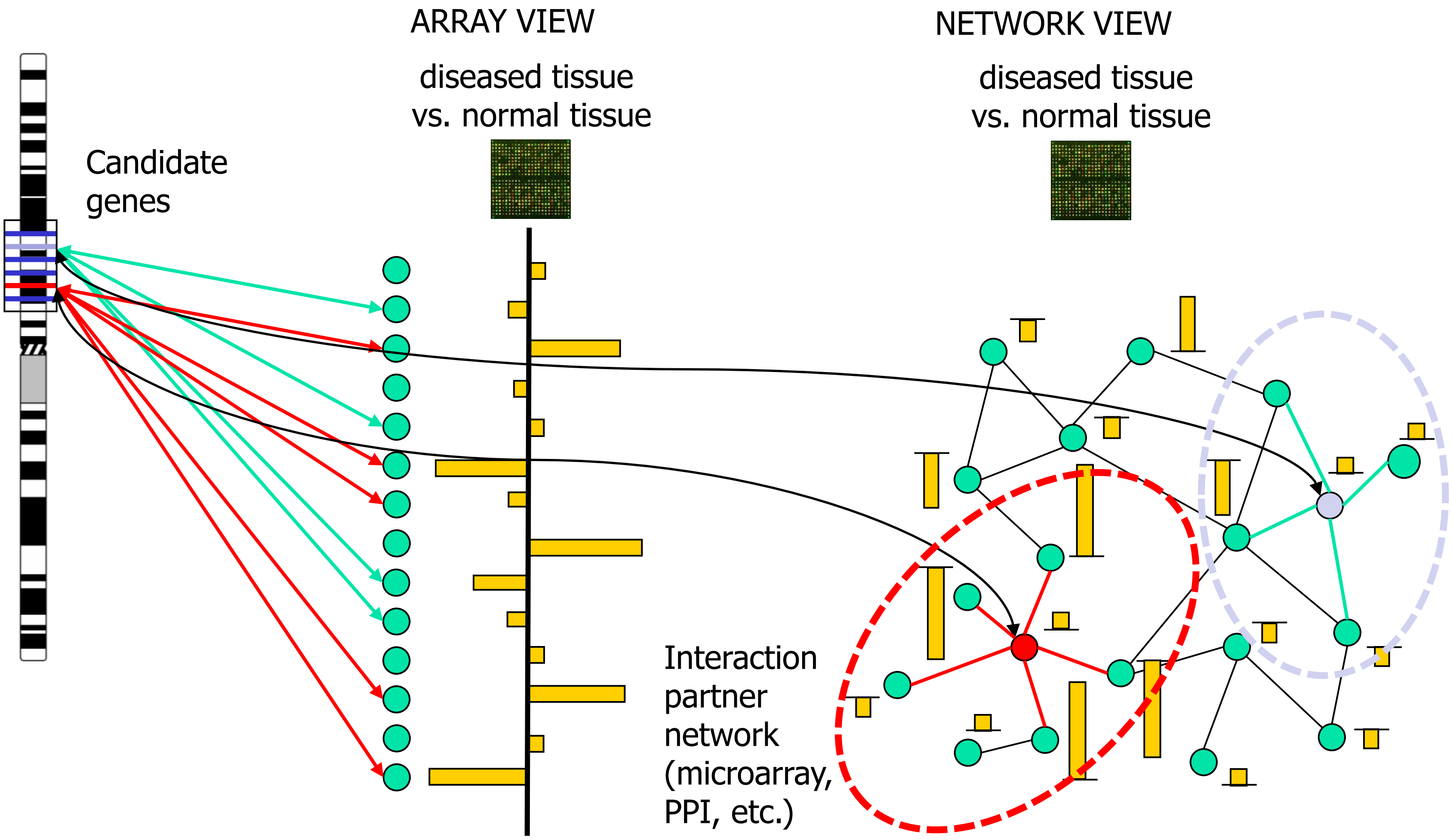
Expression of candidate genes

- For positional cloning, checking expression of candidate genes is standard but has a low yield
  - No guarantee that disease gene itself is perturbed

- Existing prioritization methods (e.g., Endeavour) rely heavily on prior knowledge and hard to achieve “breakthroughs”



Systems biology: network analysis



ARRAY VIEW

diseased tissue
vs. normal tissue

NETWORK VIEW

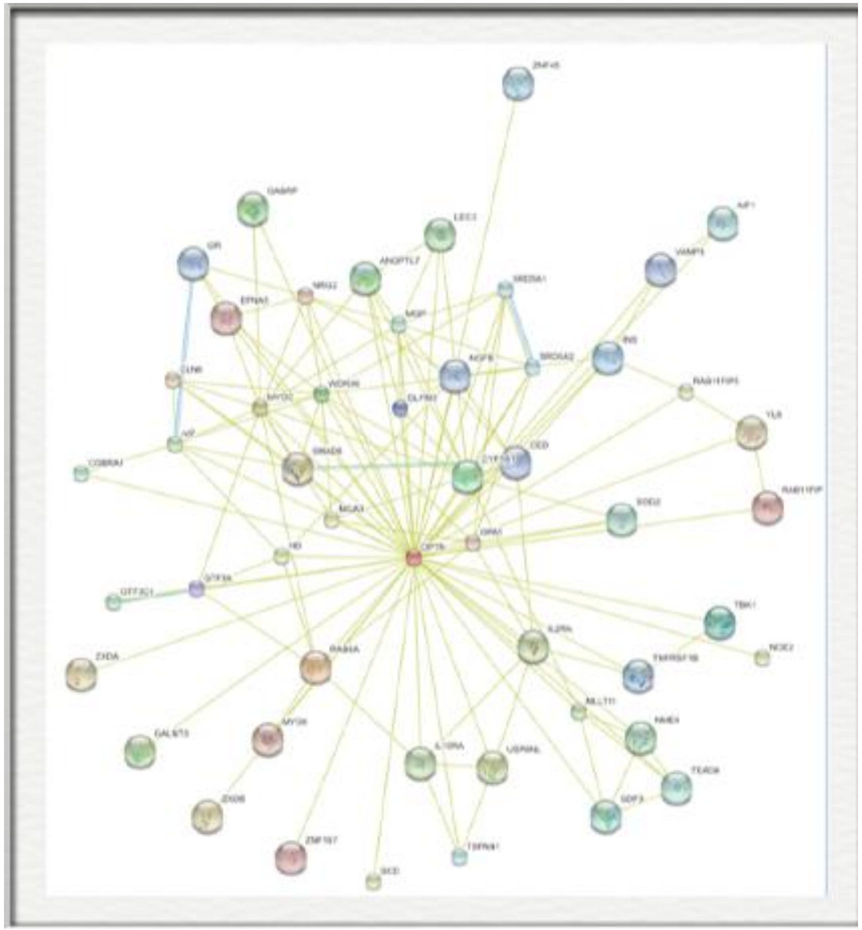
diseased tissue
vs. normal tissue

Candidate
genes

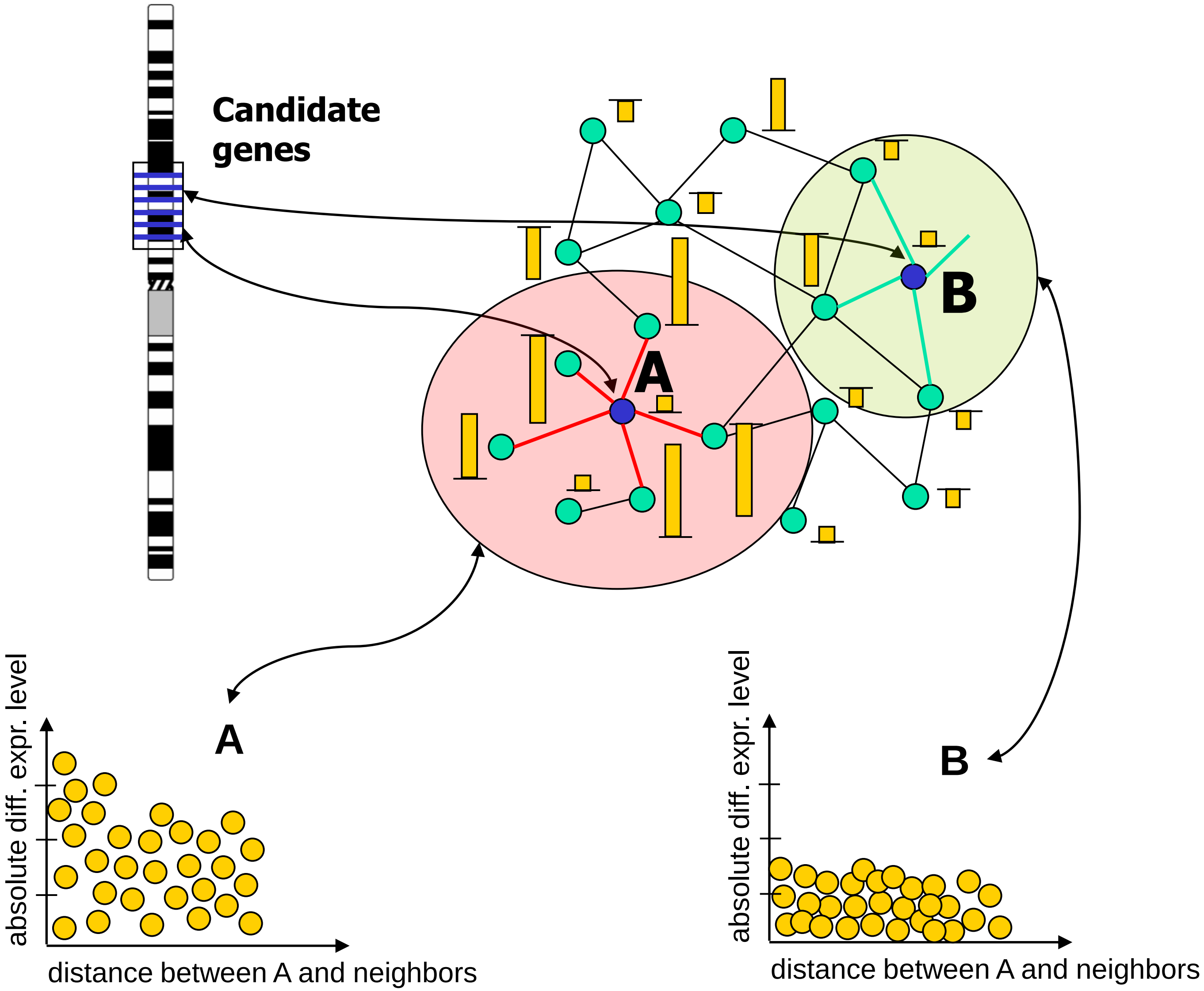
Interaction
partner
network
(microarray,
PPI, etc.)

Integrative protein network

- e.g., STRING



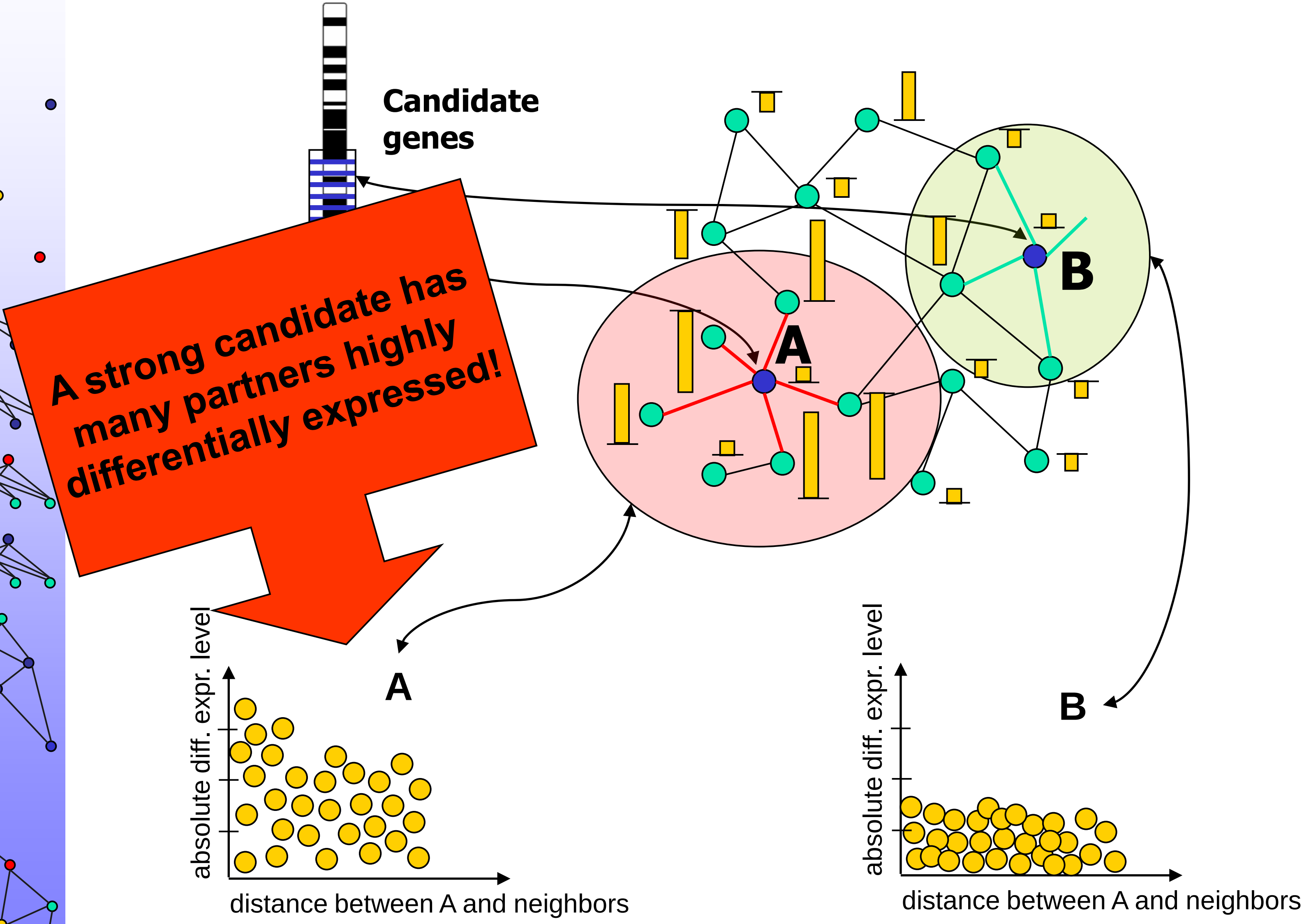
Concept



Candidate genes

A
B
A
absolute diff. expr. level
distance between A and neighbors
B
absolute diff. expr. level
distance between A and neighbors

Concept



Candidate genes

A strong candidate has many partners highly differentially expressed!

A
B
A
absolute diff. expr. level
distance between A and neighbors
B
absolute diff. expr. level
distance between A and neighbors

Methods

- Machine-learning strategies
  - Naive ranking
    - Use only the differential expression of the candidate
  - Direct neighborhood ranking
    - Combine differential expression level of candidate with the average of the differential expression levels of the direct neighbors
  - Kernel ridge regression
    - Smooth a candidate's differential expression level by kernel ridge regression
  - Approximate heat kernel diffusion
    - Discrete low-accuracy approximation to the exponential diffusion kernel $\exp(\alpha L) = \exp(\alpha(D\text{-}A))$ takes direct and indirect association into account
  - Arnoldi diffusion
    - Memory-light high-accuracy approximation to exponential diffusion kernel using Krylov subspaces (Arnoldi algorithm)

Methods & benchmark

- Benchmark: 40 KO experiments in mouse
  - Publicly available data sets from GEO - Affymetrix platform
  - Simple KO versus control
  - How well can we rank the KO gene?
  - Which algorithm and what combination of steps performs best?
- Preprocessing
  - RMA
  - GCRMA
  - MAS5
- Differential expression
  - Log2 ratio
  - Regularized t-statistic (CyberT)
  - Significant log2 ratio
- Different networks
  - STRING7, STRING8
  - PPI Network from BioGRID
  - PPI Network from I2D

Database (mouse)	**Number of genes**	**Number of interactions**	**Average node degree**
STRING v7.1	16,566	820,177	49.5
STRING v8.2	24,442	1,405,375	57.5
BioGRID v2.0.61	1,417	2,026	2.5
I2D v1.72	10,867	79,088	10.6

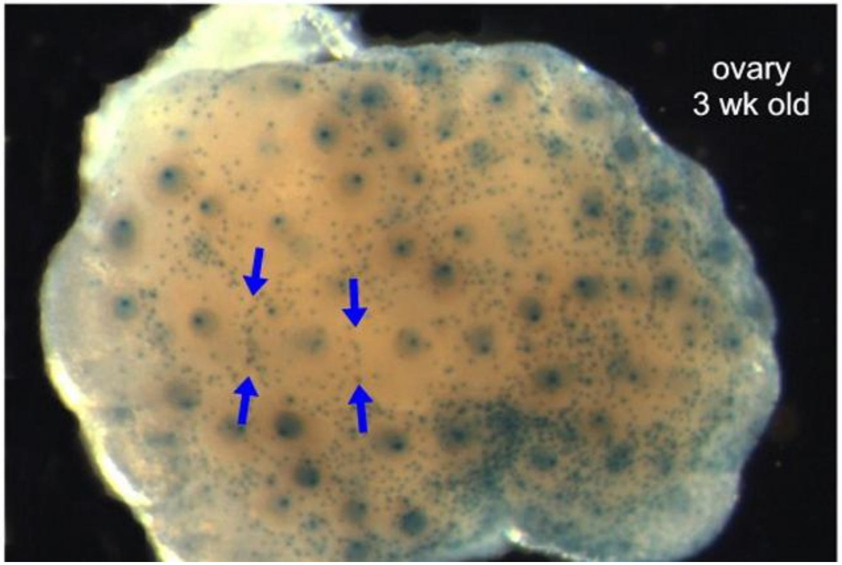
Results

Strategy	AUC	Error reduction relative to baseline
Simple expression ranking	83.0%	baseline
Direct neighborhood ranking	88.0%	26.4%
Kernel ridge regression	86.8%	19.0%
Heat kernel	92.3%	52.8%
Arnoldi diffusion	87.4%	22.7%

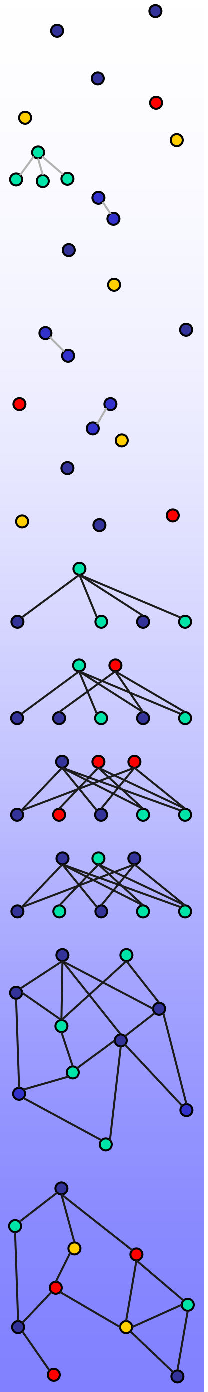
A candidate gene for PCOS

- PolyCystic Ovary Syndrome (PCOS)
  - Major cause of infertility (chronic anovulation)
  - Hormonal dysfunction (hyperandrogenism)
  - Obesity (depending on diagnostic criteria)
  - Oligogenic disorder (no Mendelian inheritance)
- Two confirmed susceptibility loci
  - 19p13.2 -> FBN3 (Fibrillin 3)
  - 5q11.2
    - FST (follistatin) proposed, but infirmed in subsequent validation
- Expression data (GEO GDS2084)
  - Omental (belly) fat from patients vs. control
  - Affymetrix HG-U133A

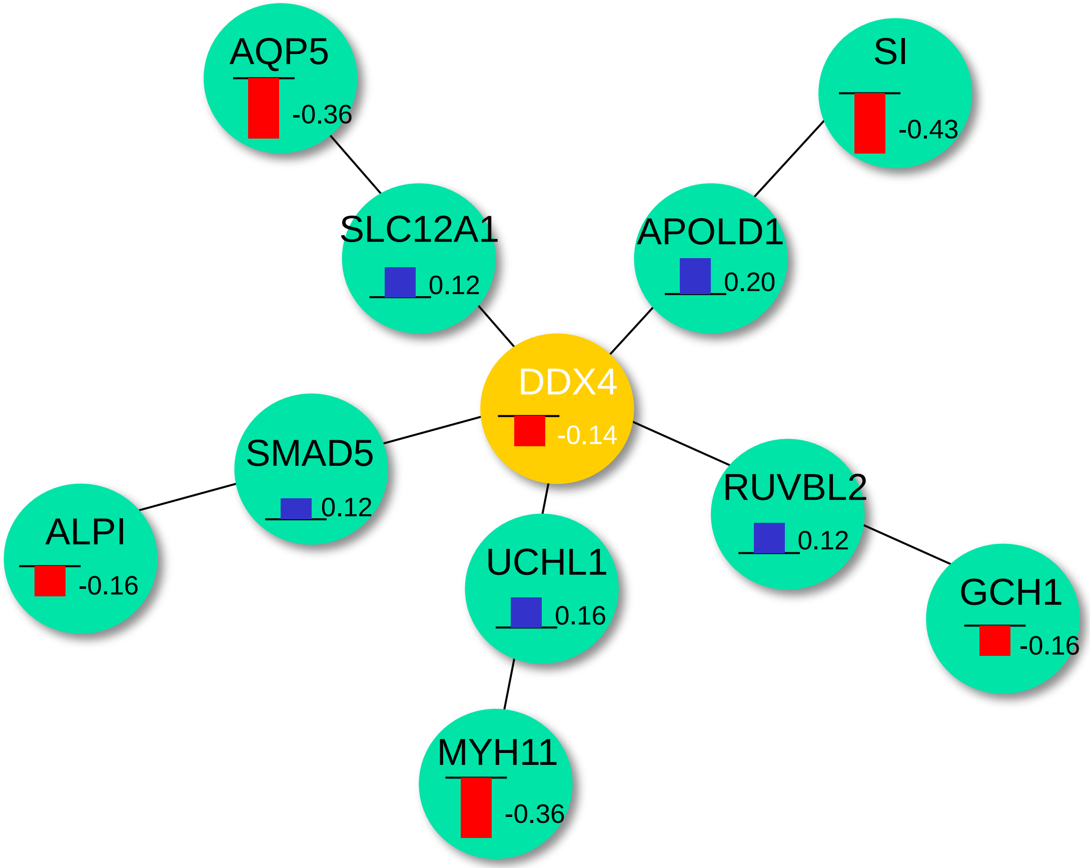
DDX4 as a PCOS candidate



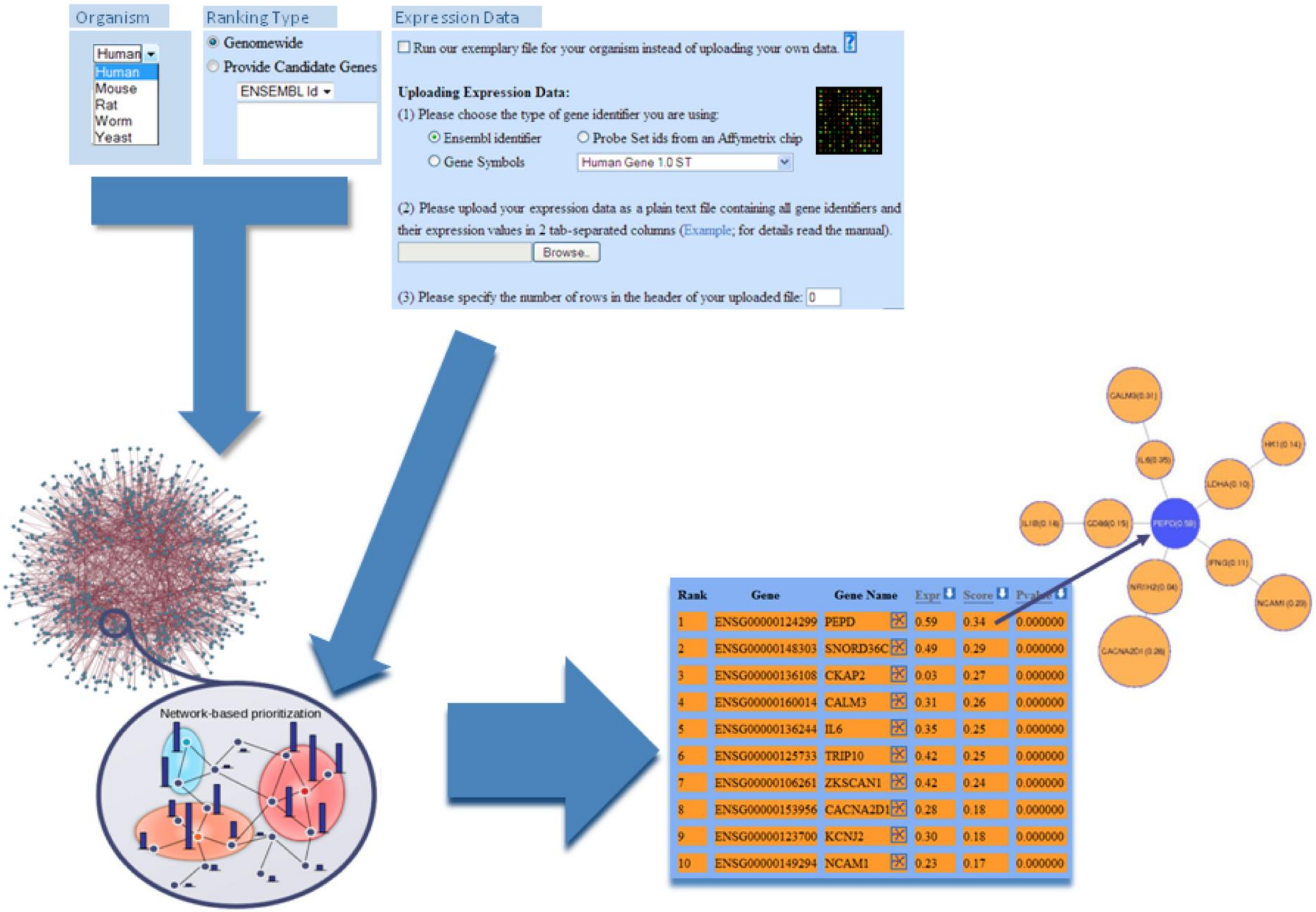
- Prioritization of 5q11.2
  - FST ranks 2nd
  - DDX4 ranks 1st
    - Expression in ovary follicles (image = mouse Ddx4/Vasa)
    - A germline development gene (sperm and ovary only)
      - Plausible mechanism for infertility and hyperandrogenism
      - Mechanism not previously suggested for PCOS
      - Not a perfect candidate (male phenotype in mouse, not female)



DDX4 expression neighborhood



PINTA web tool



Ranking Type: ● Genomewide ○ Provide Candidate Genes

□ Run our exemplary file for your organism instead of uploading your own data.

Uploading Expression Data:

(1) Please choose the type of gene identifier you are using:

● Ensembl identifier ○ Probe Set ids from an Affymetrix chip

○ Gene Symbols

(2) Please upload your expression data as a plain text file containing all gene identifiers and their expression values in 2 tab-separated columns (Example; for details read the manual).

(3) Please specify the number of rows in the header of your uploaded file: 0

Rank	Gene	Gene Name	Expr	Score	Pvalue
1	ENSG00000124299	PEPD	0.59	0.34	0.000000
2	ENSG00000148303	SNORD36C	0.49	0.29	0.000000
3	ENSG00000136108	CKAP2	0.03	0.27	0.000000
4	ENSG00000160014	CALM3	0.31	0.26	0.000000
5	ENSG00000136244	IL6	0.35	0.25	0.000000
6	ENSG00000125733	TRIP10	0.42	0.25	0.000000
7	ENSG00000106261	ZKSCAN1	0.42	0.24	0.000000
8	ENSG00000153956	CACNA2D1	0.28	0.18	0.000000
9	ENSG00000123700	KCNJ2	0.30	0.18	0.000000
10	ENSG00000149294	NCAM1	0.23	0.17	0.000000

Krylov subspace methods

- Function of a square matrix $f(A)$
  - Matrix inverse
  - Matrix exponential (e.g., exponential of graph Laplacian)
  - Computationally challenging for large matrices
- In applications, often no need for matrix function $f(A)$ directly, but only its evaluation $f(A).v$ at a point v
- Krylov methods – simplified argument
  - Cayley-Hamilton theorem for characteristic polynomial of square n x n matrix A

$$p_n(\lambda) = \det(\lambda I_n - A)$$

$$p_n(A) = 0$$

  - Any power of A higher or equal to n can be expressed in function of $A^{n-1}, A^{n-2}, \ldots, A^2, A, I$

- Any matrix function $f(A)$ can be expressed as a polynomial of degree $n\text{-}1$
- $f(A).v$ can be expressed as a linear combination of

$$\left\{v, Av, A^2v, \ldots, A^{n-2}v, A^{n-1}v\right\}$$

- Krylov methods consists in projecting $f(A).v$ onto the subspace

$$S_m = \mathrm{span}\left\{v, Av, A^2v, \ldots, A^{m-2}v, A^{m-1}v\right\}$$

- Only requires matrix-vector operations!
- The set of spanning vectors is kept orthogonal via QR orthogonalization
- In practice, often fast convergence $m \ll n\text{-}1$

Networks vs. kernels

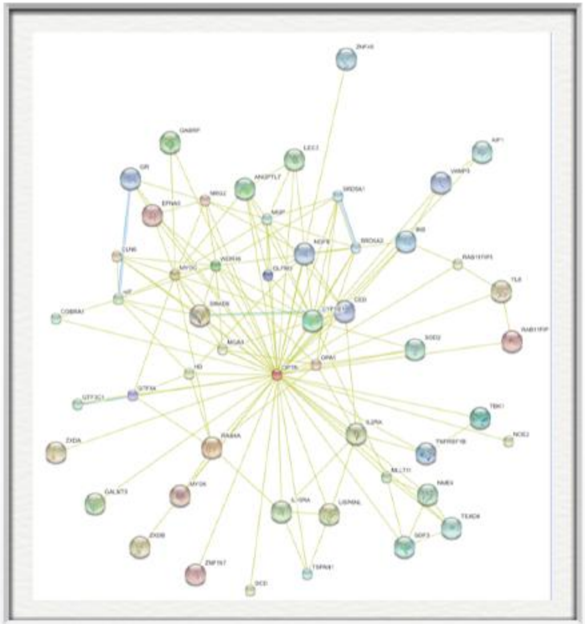
<Rant> A network is a matrix is a network

- And a symmetric similarity matrix is an undirected graph
- Implicitly, we mean more by a network
  - Sparse matrix
  - Edges have some underlying biological reality
  - e.g., KEGG metabolic network from one organism or regulonDB yeast transcriptional network
- Most predicted protein networks do not have such properties
  - Usually calculated much like similarity matrices
  - Why handle them as "biological" networks?
  - Networks useful in visualization, but should not be misleading
  - Network representation usually involves heavy thresholding and creates an information bottleneck </Rant>

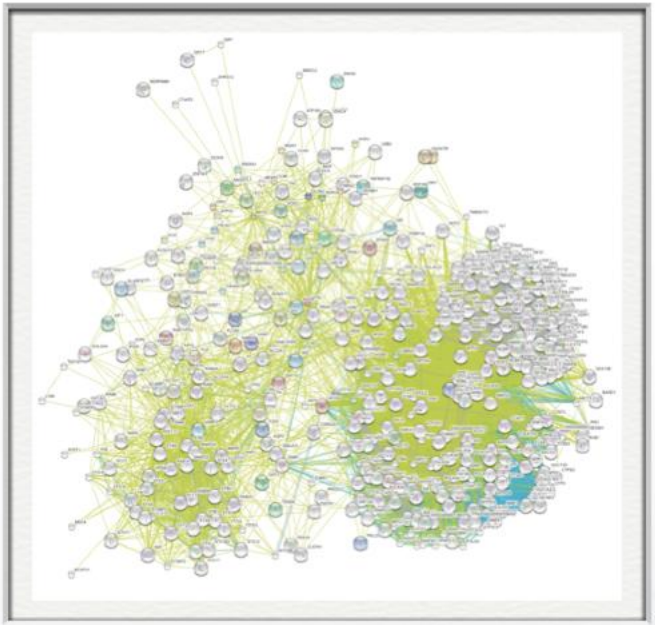
- Gillis and Pavlidis suggest that networks and similarity matrices are almost equivalent under the no-voodoo principle

What’s wrong with network propagation?

- Kernel diffusion and network module biomarkers all seem to perform less strongly than expected.
- What is wrong?
  - Nothing? (Our expectation is unrealistic)



Best 50



Best 500

Curse of the small world?

1. Data is not specific enough to highlight the right network neighborhoods
   -> improve experimental design (e.g., factorial design)
2. Our data is improperly scaled or normalized
   (propagating apples and oranges)
3. Our networks are bad
   (STRING > BioGrid, coverage more important than specificity)
4. Network is in fact thresholded propagation matrix already (e.g., STRING)
   (further propagation does not help much)
5. Our notion of neighborhood and diffusion is unsuitable (curse of the small world)
   (rough approximation to heat kernel works better than accurate one)
6. Our randomization procedures are unsuitable
   (propagation results are still apples and oranges)
7. Uncertainty propagation is the bottleneck
   (more complex model propagates more noise and thus destroys the advantage of added knowledge)

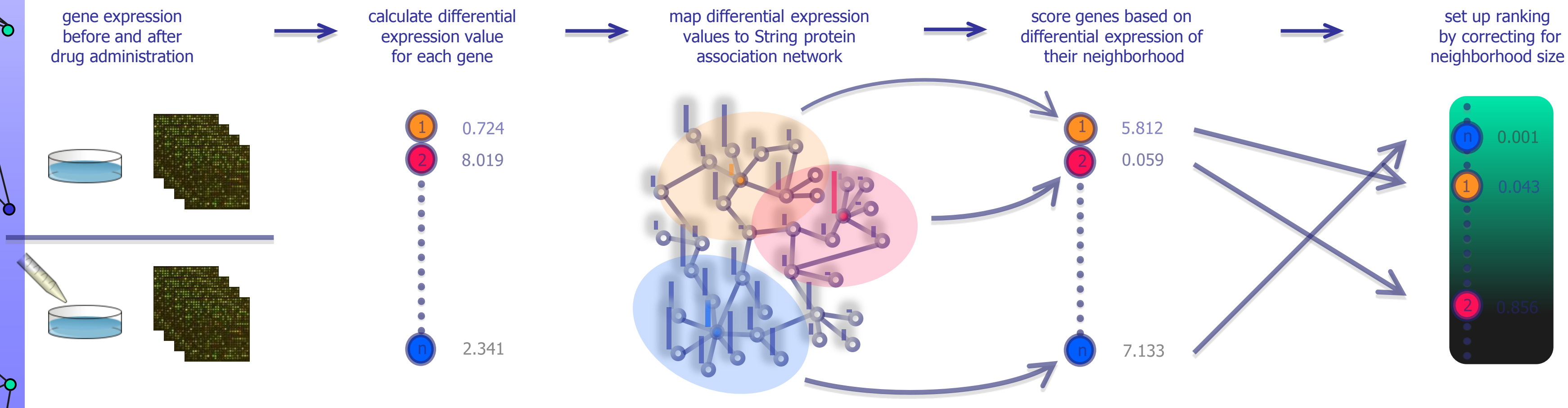
Drug target prioritization

Target prioritization

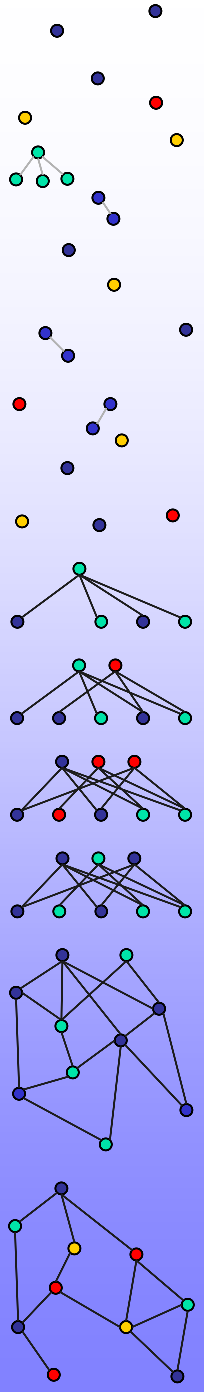
- One drug, many targets
  - Many targets are unknown
    - Side effects
    - Synergistic effects
  - Candidates identified by phenotypic screen
  - Knowledge of a drug candidate’s mode of action can help drug development
- Predict targets based on gene expression following treatment
  - Distinguish between genes targeted by the drug and indirectly regulated genes

Network analysis of drug response

- Gene expression integrated with protein associations
- Neighborhood analysis
- Gene prioritization based on differential expression of functionally related network neighborhood



gene expression before and after drug administration → calculate differential expression value for each gene → map differential expression values to String protein association network → score genes based on differential expression of their neighborhood → set up ranking by correcting for neighborhood size



Method

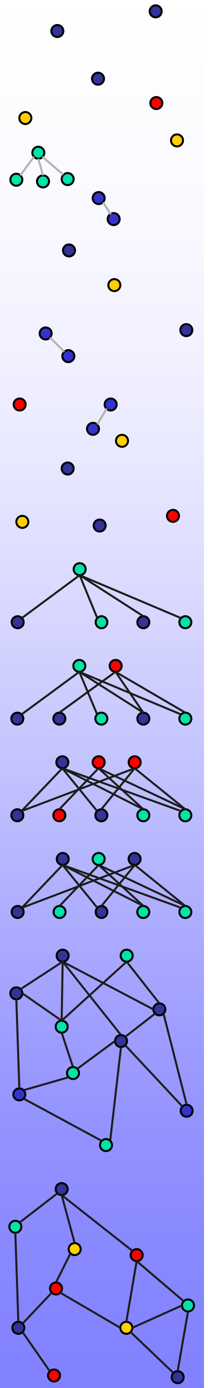
- Filtering
  - No filtering
- Expression measure
  - log ratio
- Network
  - STRING 8.2
- Parameters
  - N = 1, 2, 3,ɑ = 0.9

    $$p_\alpha = p_0\left(I - \frac{\alpha}{N}L\right)^N$$

  - ɑ = 0.9, β=0.1

    $$p_\alpha = p_0\left(I - \alpha L + \frac{\alpha^2}{4}L^2\right) \Rightarrow p_\alpha = p_0\left(I - \alpha L + \beta L^2\right)$$

Preliminary results: monoclonal antibodies



- Monoclonal antibodies specifically bind to one target

- 7 datasets from Gene Expression Omnibus:
  - tocilizumab: IL6
  - bevacizumab: VEGFA
  - rituximab: MS4A1
  - infliximab: TNF
  - h10H5: IGF1R
  - anti-CD25: IL2RA
  - LY2439821: IL17A

- Can we identify the target from expression response?

Preliminary results: monoclonal antibodies

		tocilizumab	bevacizumab	rituximab	infliximab	h10H5	anti-CD25	LY2439821	# in top 5%	# in top 10%
differential expression		1,057	6,896	176	4,281	12,279	522	1,992	2	3
α=0.9	N=1	982	99	342	142	9,109	48	227	5	6
	N=2	597	279	268	2,254	5,055	102	766	5	5
	N=3	720	446	151	2,628	4,645	148	840	4	5
α=0.9, β=0.1		763	93	186	758	7,780	38	454	6	6

- Why does the h10H5 target IGF1R rank this low?
  - Bad experiment?
  - Bad method?
  - No downstream transcriptional effect?
- Test other IGF1R inhibitor: BMS754807
  - ➔ For α=0.9 and β=0.1 IGF1R ranked at position 381

Preliminary results: chemical drugs

- Chemical drugs can bind multiple targets

- 7 datasets from Gene Expression Omnibus
  - letrozole: CYP19A1
  - bicalutamide: AR
  - calcitriol: VDR
  - methylprednisolone: NR3C1
  - gefitinib: EGFR
  - methotrexate: DHFR
  - progesterone: PGR

- Can we identify the target from expression response?

Preliminary results: chemical drugs

		letrozole	bicalutamide	calcitriol	methyl-prednisolone	gefitinib	methotrexate	progesterone	# in top 5%	# in top 10%
differential expression		14,055	700	4,262	4,316	8,612	871	79	2	3
α=0.9	**N=1**	2,460	23	887	5,109	1,848	650	919	2	4
	N=2	1,658	7	387	6,402	915	683	118	4	5
	N=3	2,030	11	595	5,283	990	500	112	4	5
α=0.9, β=0.1		1,349	1	55	8,246	1,076	819	325	4	6

- Why does the methylprednisolone target NR3C1 rank this low?
  - Bad experiment?
  - Bad method?
  - No downstream transcriptional effect?
- Test other NR3C1 agonist: fluticasone
  - ➔ For α=0.9 and β=0.1 NR3C1 ranked at position 40

KATHOLIEKE UNIVERSITEIT LEUVEN

SymBioSys

BIOMAGNET
Bioinformatics and Modelling: from Genomes to Networks

K.U.L. ESAT-SCD: L. Tranchevent, Y. Shi, D. Nitsch, , R. Barriot, S. Leach, B. Coessens, S. Van Vooren

K.U.L. CME-UZ: J. Vermeesch, K. Devriendt, B. Thienpont, F. Hannes, J. Breckpot

K.U.L. VIB: D. Lambrechts, S. Maity, P. Carmeliet, S. Aerts, B. Hassan, P. Van Loo, P. Marynen

U. Bristol: T. De Bie

INESC-ID, Lisbon: J. Gonçalves, S.Madeira

Novartis: S. Schuierer, U. Dengler