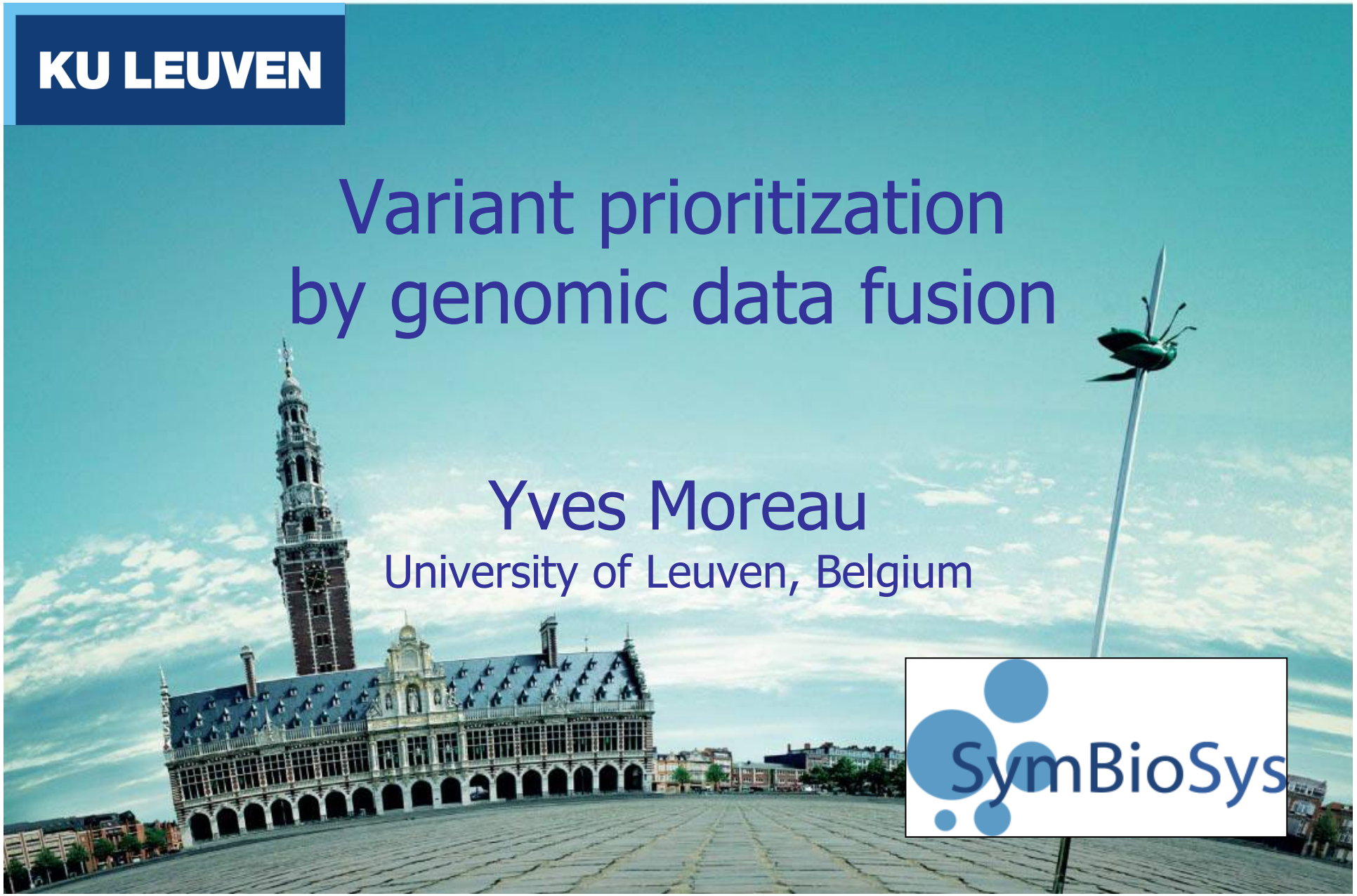
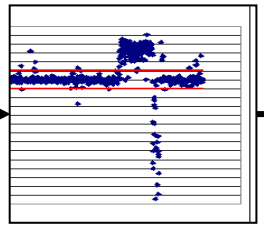

KU LEUVEN
Variant prioritization
by genomic data fusion
Yves Moreau
University of Leuven, Belgium
SymBioSys

Disease gene discovery in rare congenital disorders

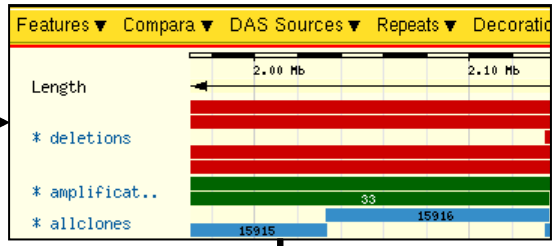
Patients with congenital & acquired disorders

CGH microarrays NGS sequencing

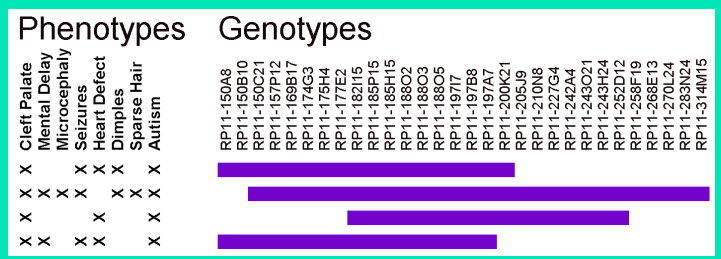
Statistical analysis



Location of chromosomal imbalances



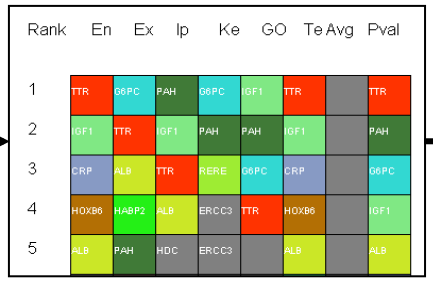
Databasing



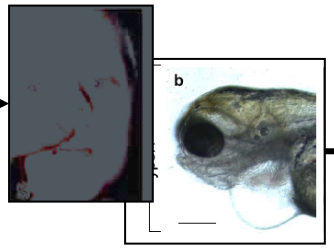
- Map chromosomal abnormalities
- Improved diagnosis

Discover new disease causing genes and explain their function

Prioritized candidate genes



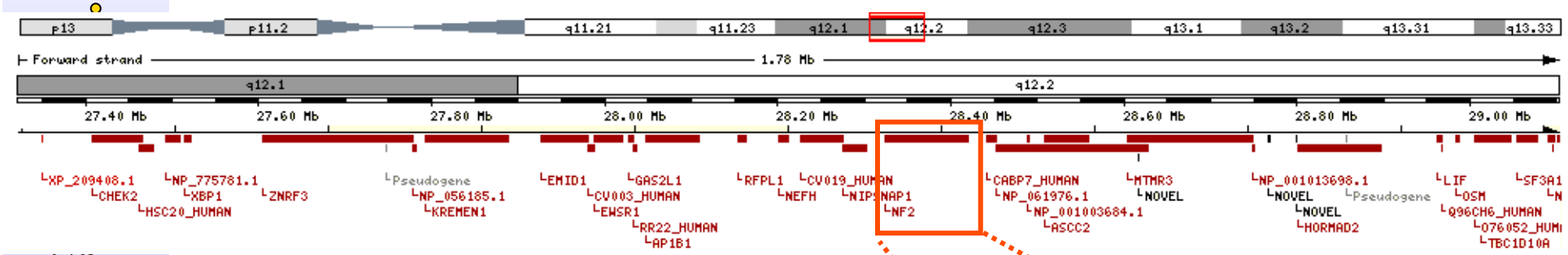
Validation



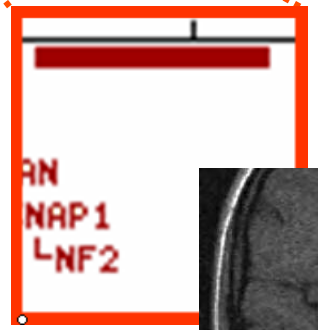
Genetic diagnosis

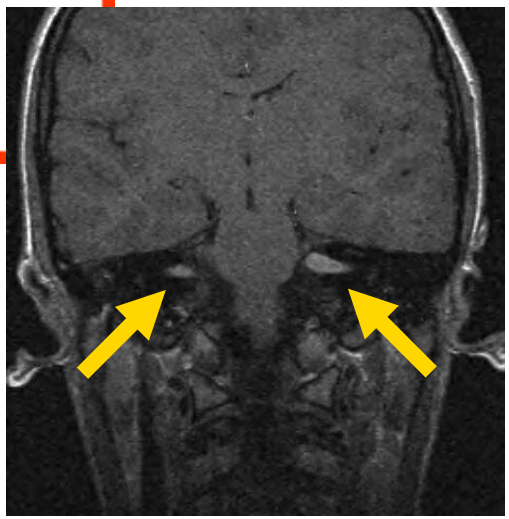
- Main medical goals
  - End diagnostic odyssey
  - Estimate risk for next pregnancy
  - Predict disease progression, life expectancy, etc.
- Patient - deletion del(22)(q12.2)
  - Pulmonary valve stenosis
  - Cleft uvula
  - Mild dysmorphism
  - Mild learning difficulties
  - High myopia

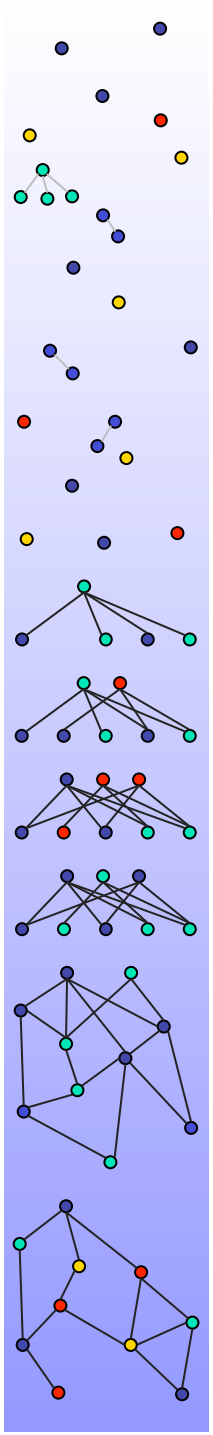
Deletion del(22)(q12.2)



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



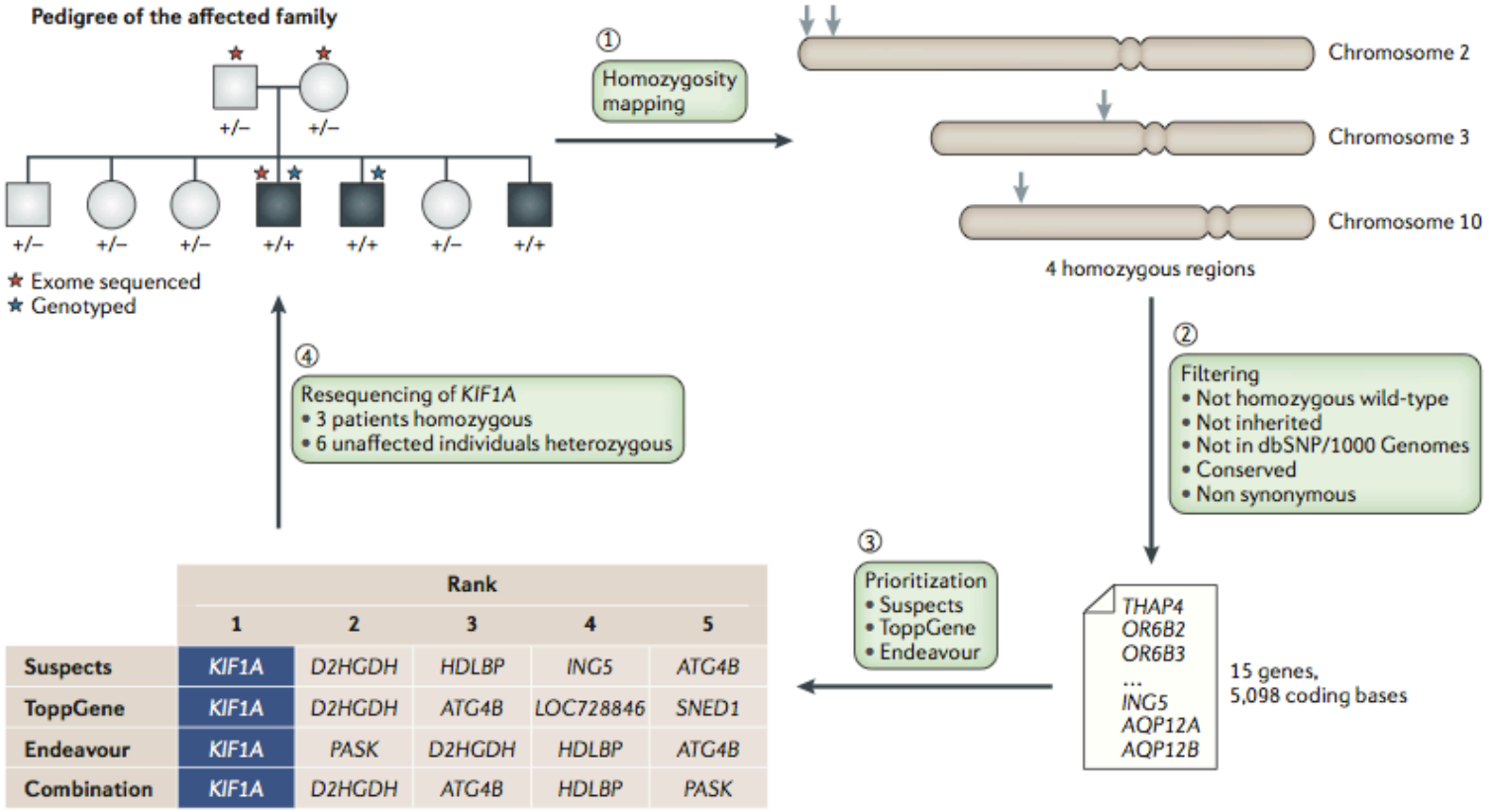




Exome sequencing

- Clinical sequencing of whole genomes is around the corner
  - But data will be hard to interpret
- Exome sequencing
  - Routine clinical use has started
  - More conserved, fewer mutations, easier to interpret
- Some mutations are easy to interpret, but in most cases it will still be hard to identify which mutation causes disease
  - Can variants be prioritized?
  - Existing tools for variant deleteriousness prediction (SIFT, Polyphen, MutationTaster etc.) fall short

Exome sequencing and gene prioritization



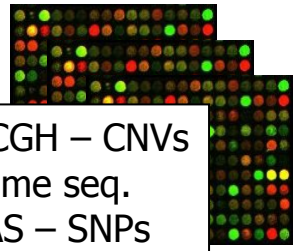
	Rank				
	1	2	3	4	5
Suspects	*KIF1A*	*D2HGDH*	*HDLBP*	*ING5*	*ATG4B*
ToppGene	*KIF1A*	*D2HGDH*	*ATG4B*	*LOC728846*	*SNED1*
Endeavour	*KIF1A*	*PASK*	*D2HGDH*	*HDLBP*	*ATG4B*
Combination	*KIF1A*	*D2HGDH*	*ATG4B*	*HDLBP*	*PASK*

Erlich, Y. et al. Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. Genome Res. 21, 658–664 (2011).

Nature Reviews | Genetics

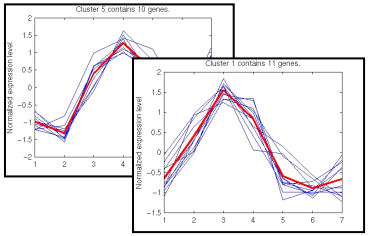
Candidate gene prioritization

High-throughput genomics



Array CGH – CNVs
Exome seq.
GWAS – SNPs
Expression

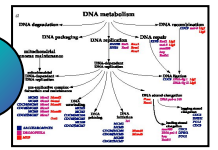
Data analysis

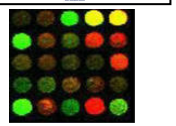


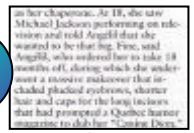
Candidate genes

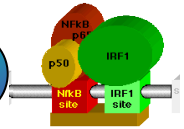
Name	Ensembl
TTR	ENSG00000118271
PAH	ENSG00000171759
G6PC	ENSG00000131482
IGF1	ENSG00000017427
ALB	ENSG00000163631
CRP	ENSG00000132693
HABP2	ENSG00000148702
IF	ENSG00000138799
FST	ENSG00000134363
ARAF1	ENSG00000078061
HMGA2	ENSG00000149948
C9	ENSG00000113600
PCBP2	ENSG00000111406
HOXB6	ENSG00000108511
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
	ENSG00000184508
	ENSG00000113460
TGFB3	ENSG00000
C10R1	ENS
N	ENS
PDG	ENSG
PD	
NR	
NFYA	
	ENSG 06537
MMP3|MMP1	ENSG00000149968

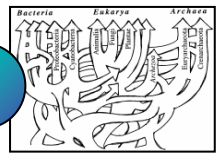
Information sources







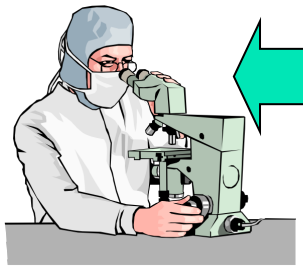




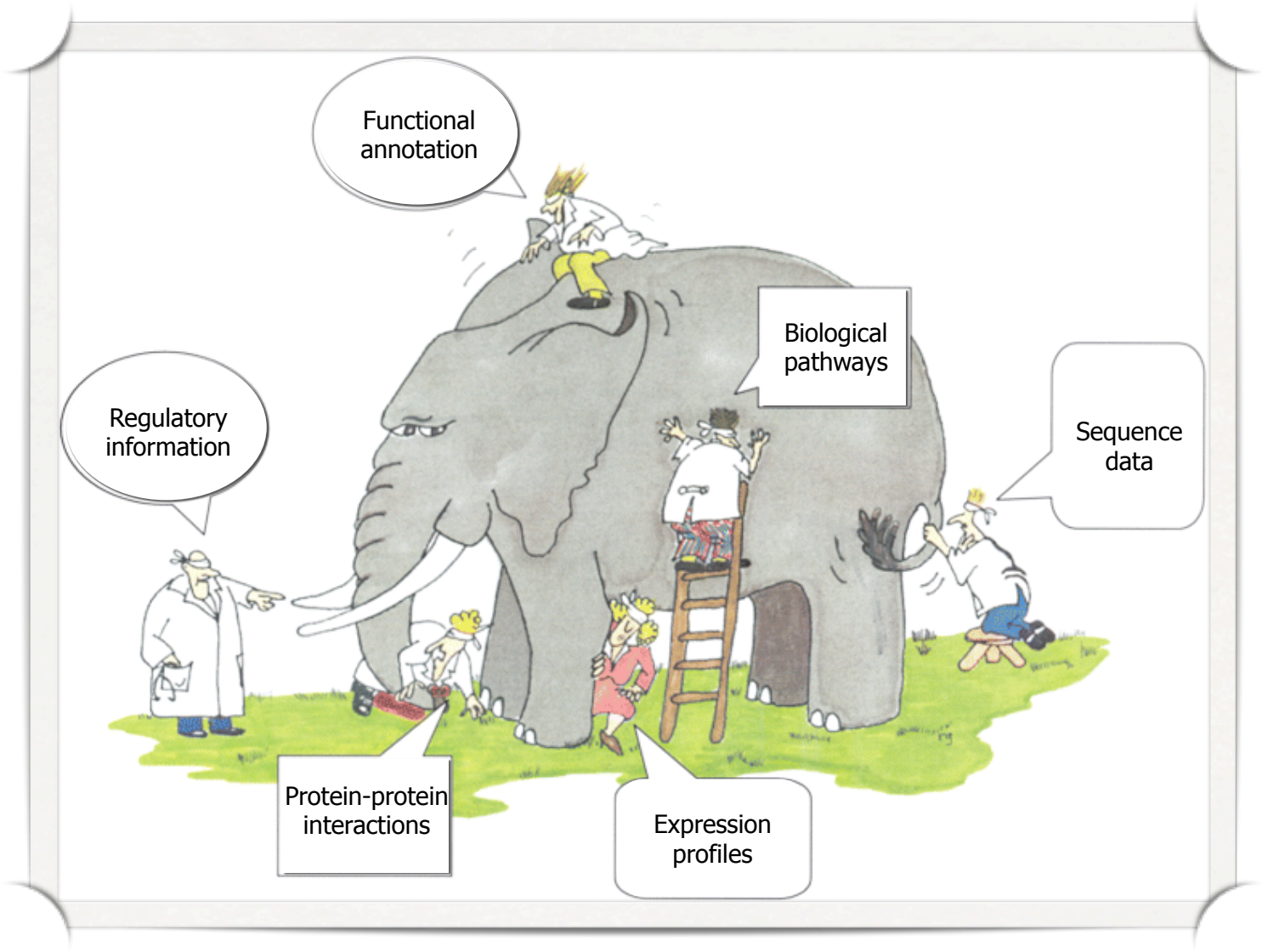
Candidate prioritization

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		
7	PAH		C1QR1	ARAF1	HDC	NR4A2		HABP
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

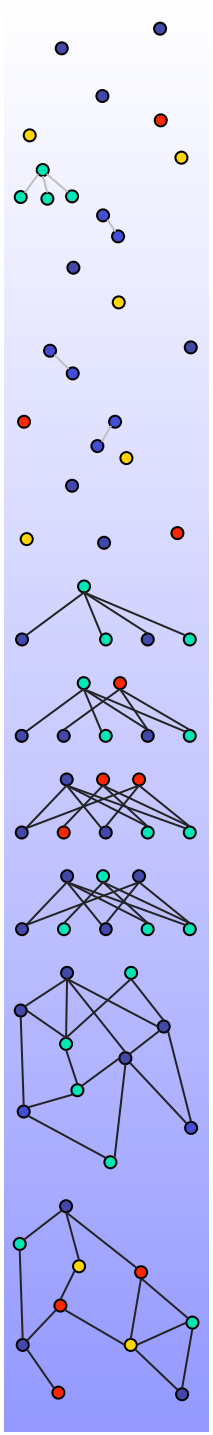
Validation



Data fusion



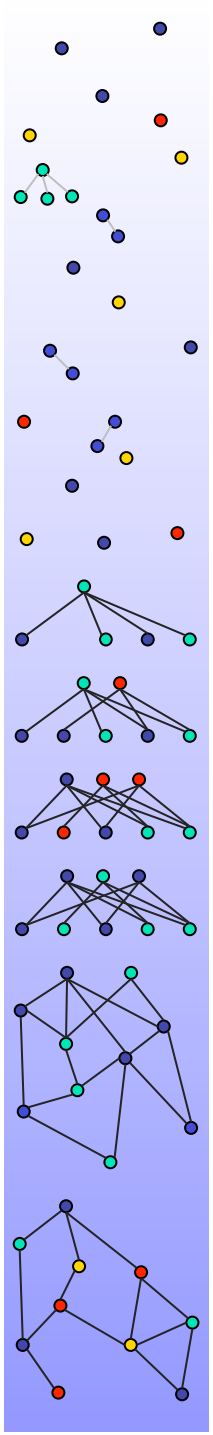
KATHOLIEKE UNIVERSITEIT LEUVEN



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



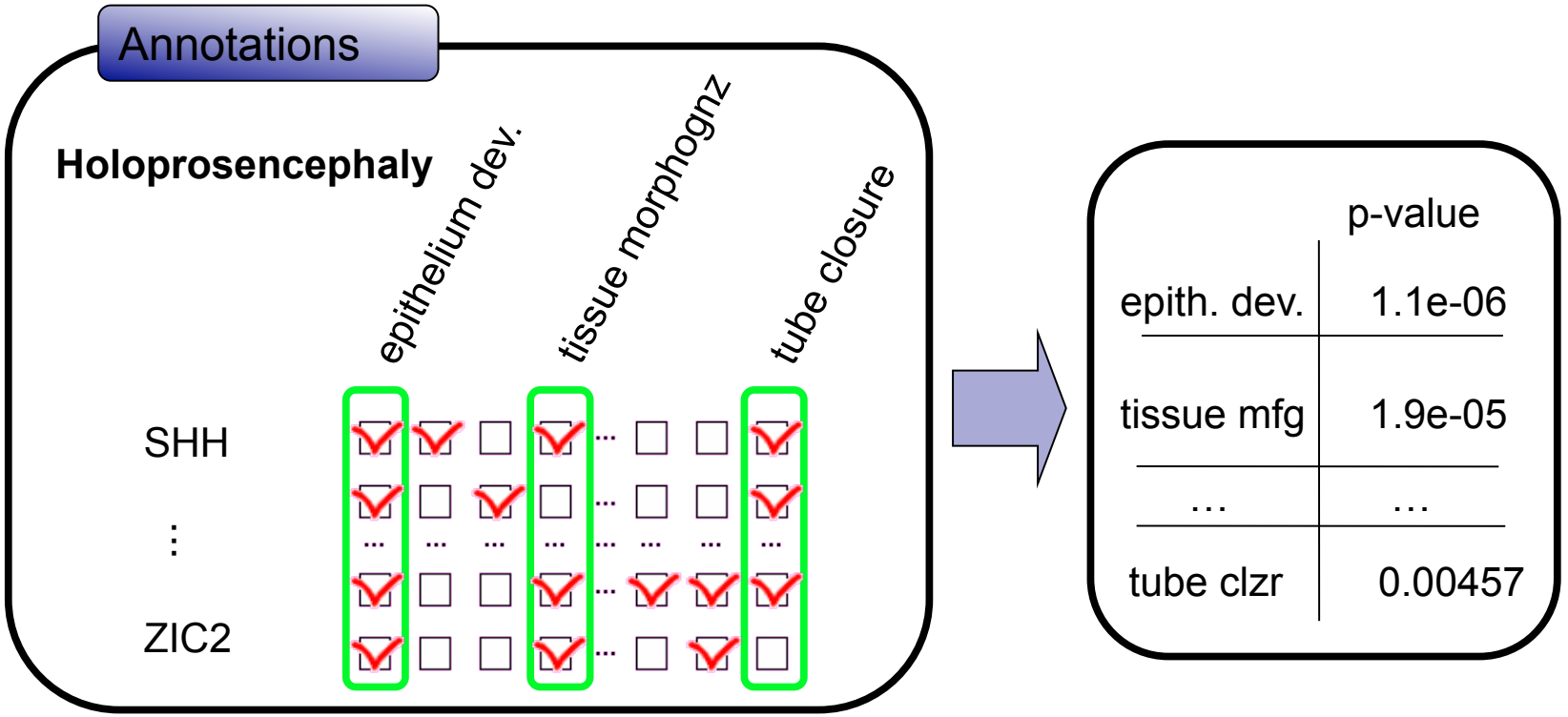
Responsible for MODY, an uncommon monogenetic form of early onset T2D.

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)

Profiling known genes (Gene Ontology)

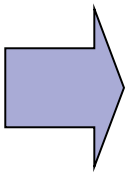


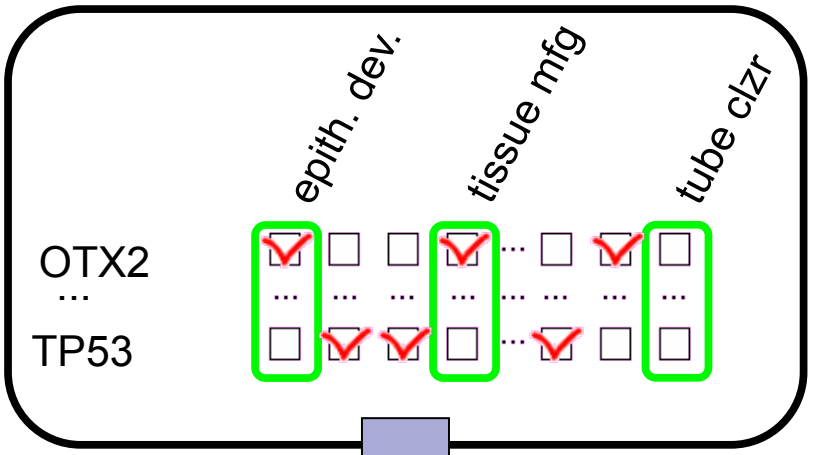
- A term is over-represented if its frequency inside the training set is significantly larger than its frequency over the genome
  - E.g., Gene Ontology, Interpro, KEGG

Scoring candidates

Annotations

	p-value
epith. dev.	1.1e-06
tissue mfg	1.9e-05
…	…
tube clzr	0.00457



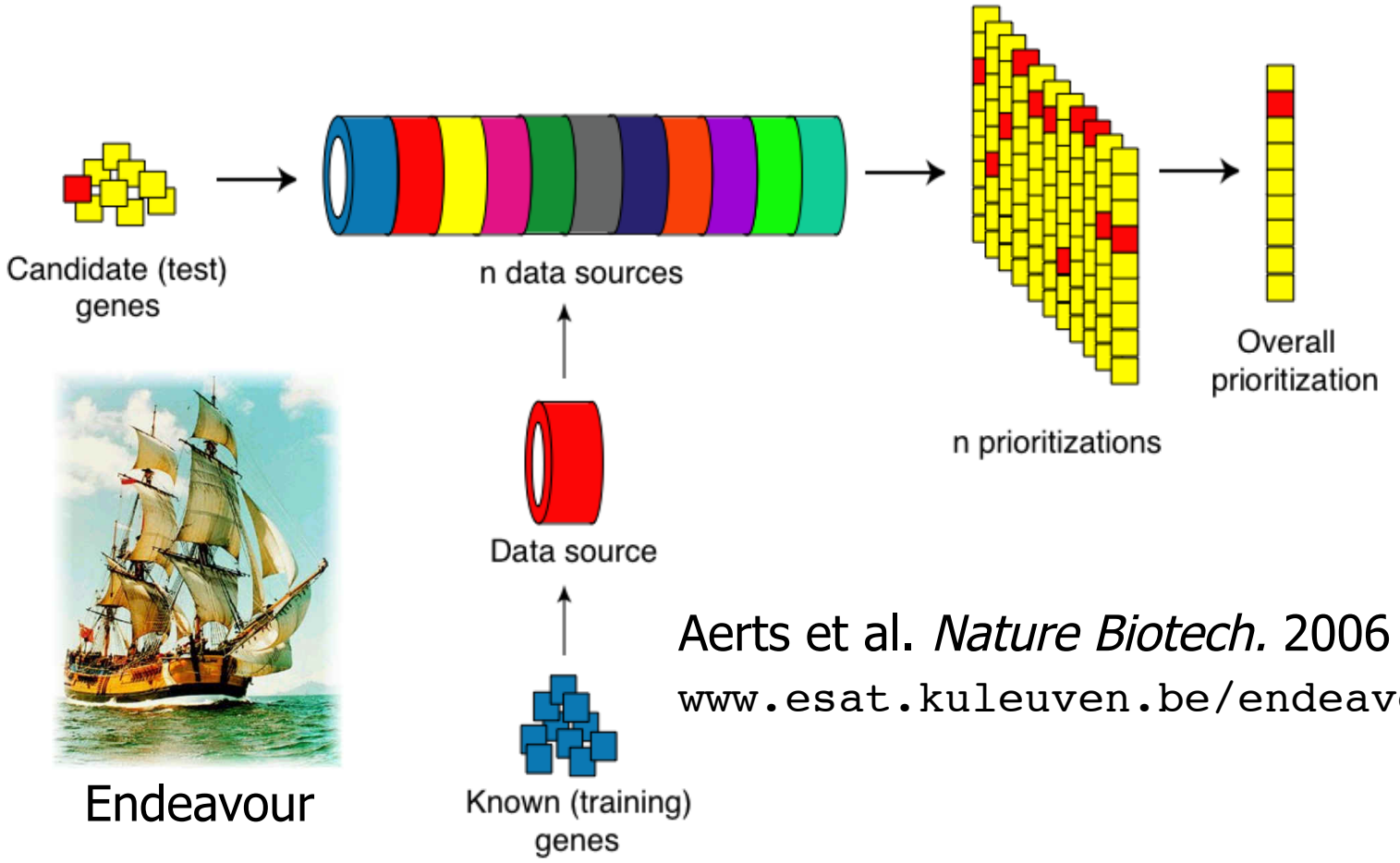


Scoring derived from Fisher's omnibus statistic

- $S = -2 \Sigma_i \log p_i$

	p-value
OTX2	0.0005
…	…
TP53	1.0

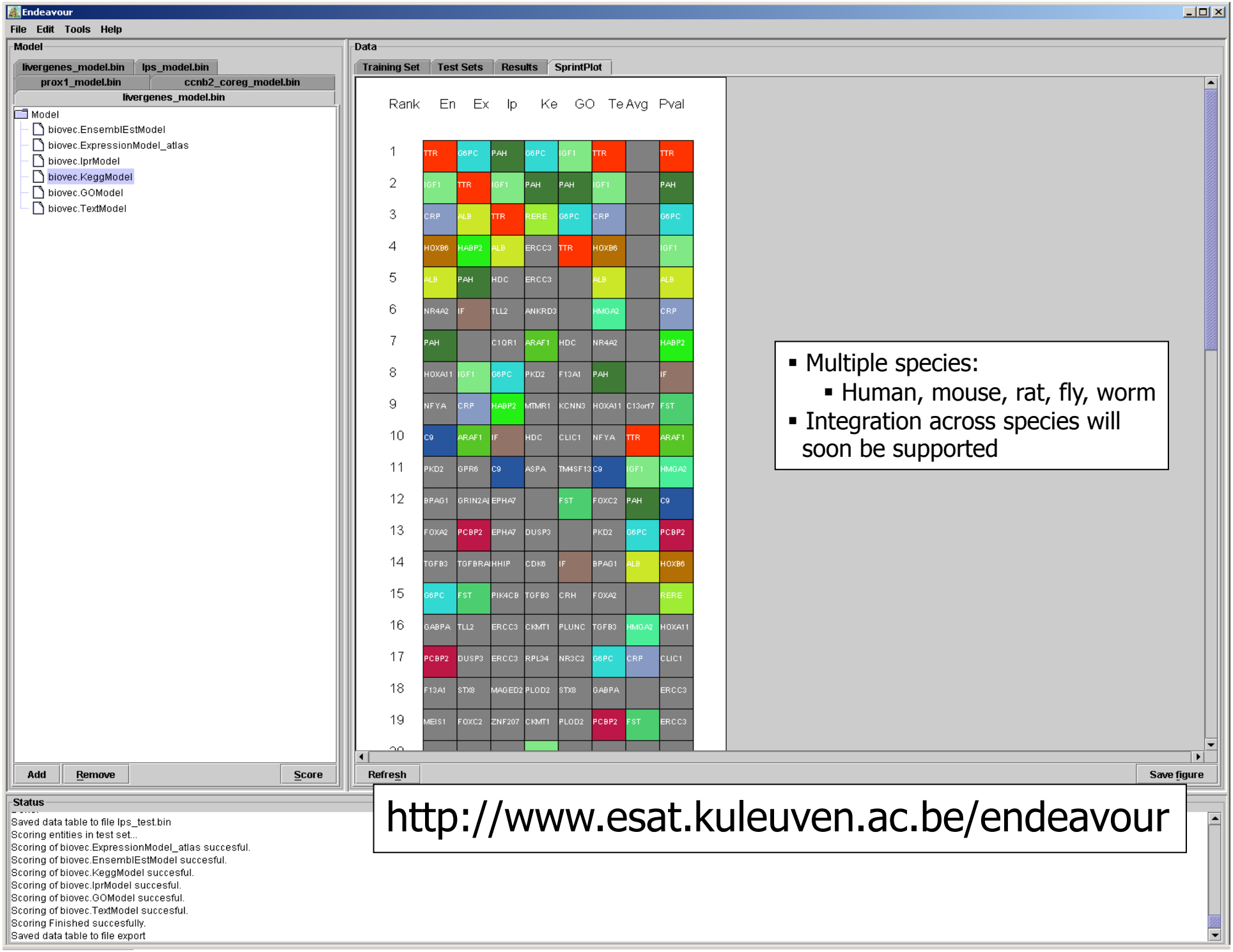
Data fusion with order statistics



Endeavour

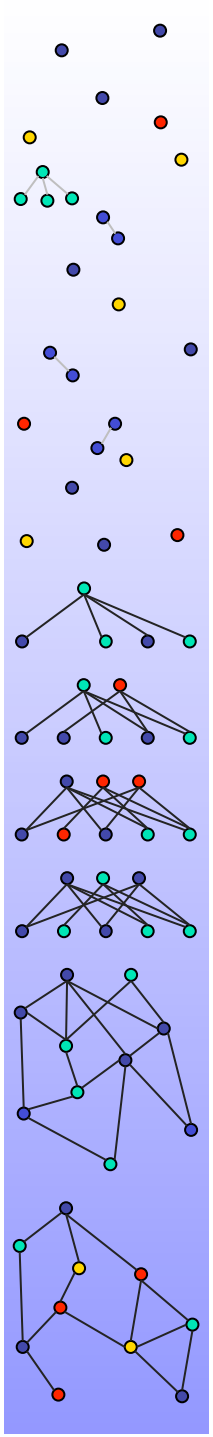
Aerts et al. *Nature Biotech.* 2006

`www.esat.kuleuven.be/endeavour`



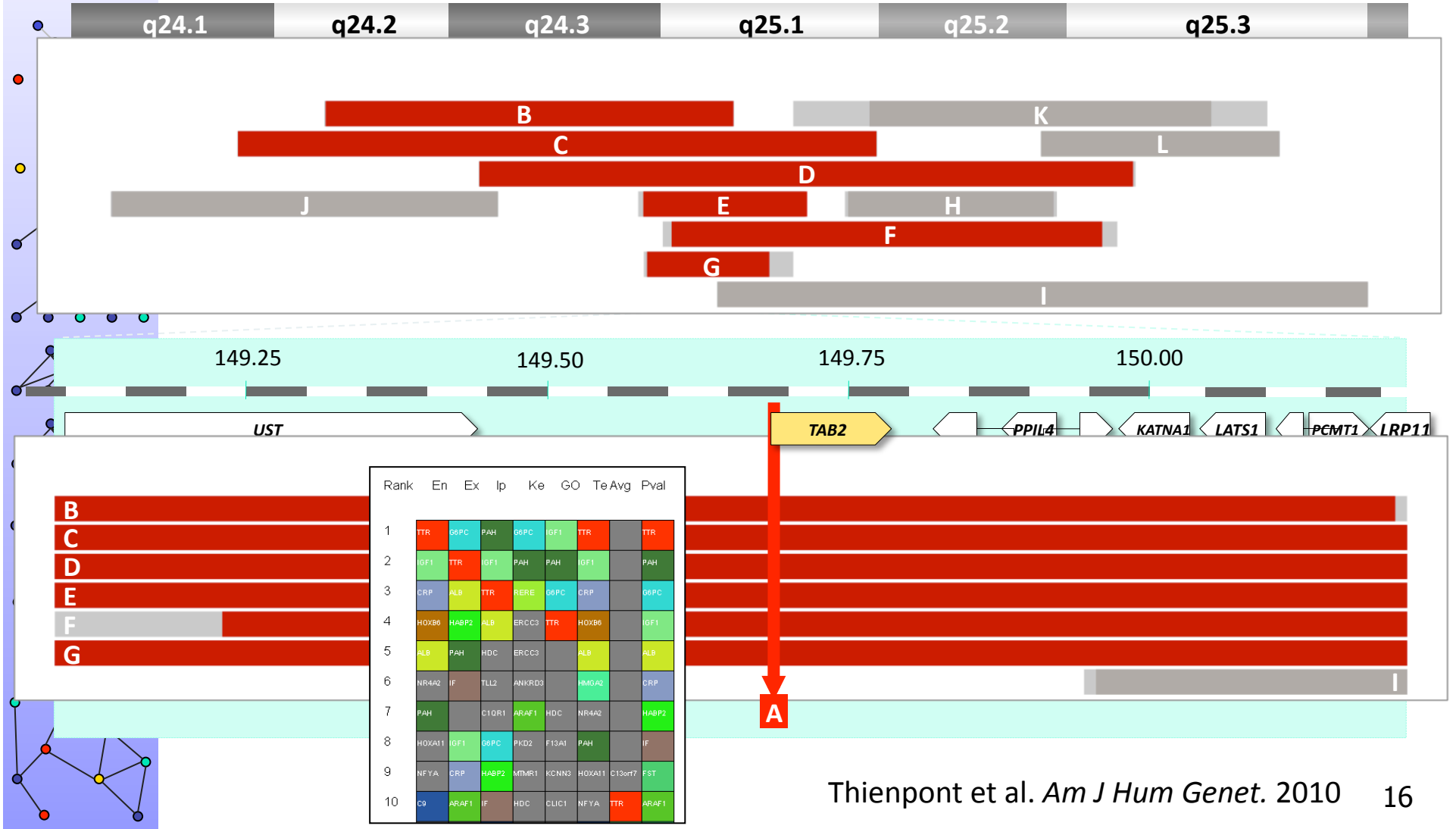
- Multiple species:
  - Human, mouse, rat, fly, worm
- Integration across species will soon be supported

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



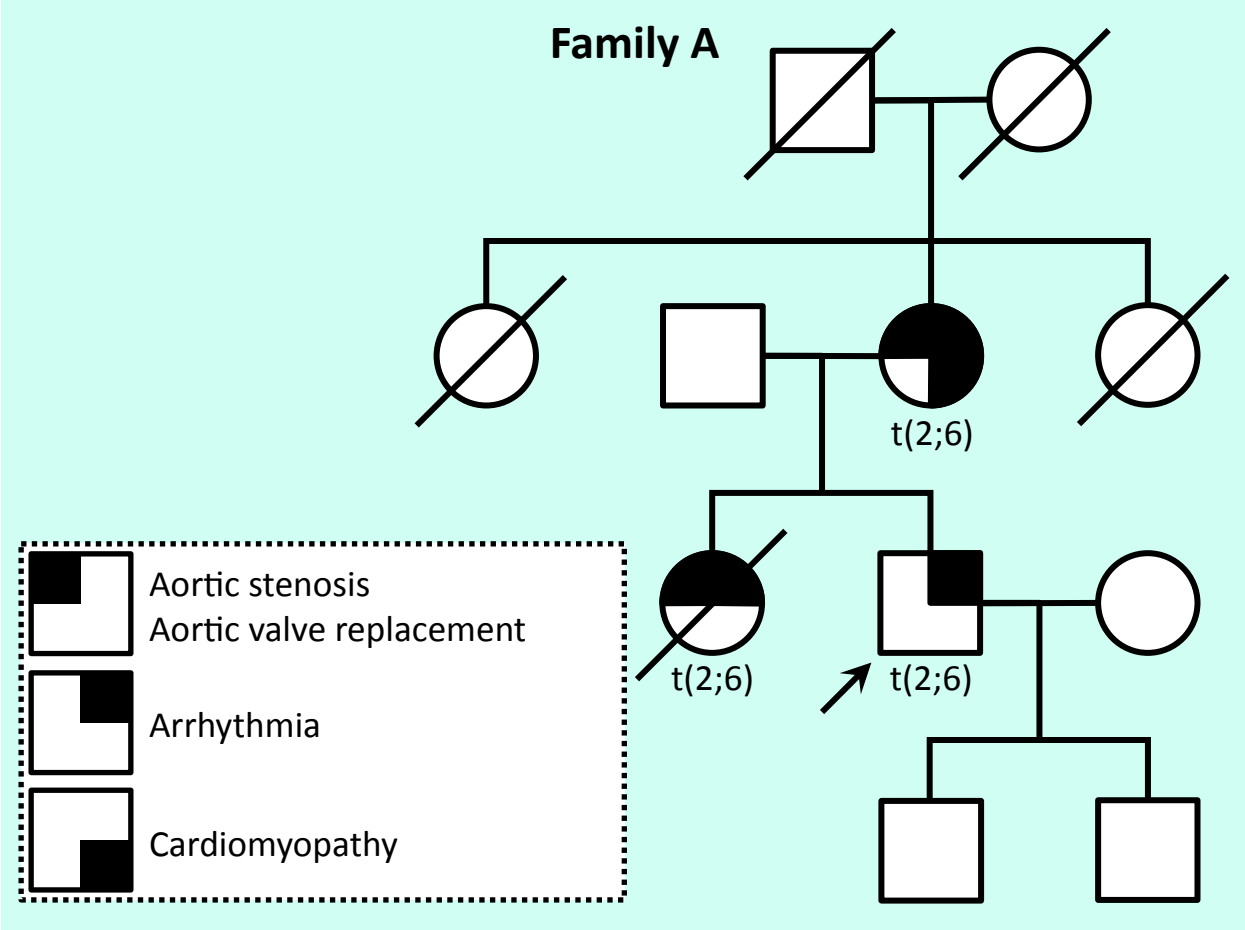
Prioritization for a monogenic disorder

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



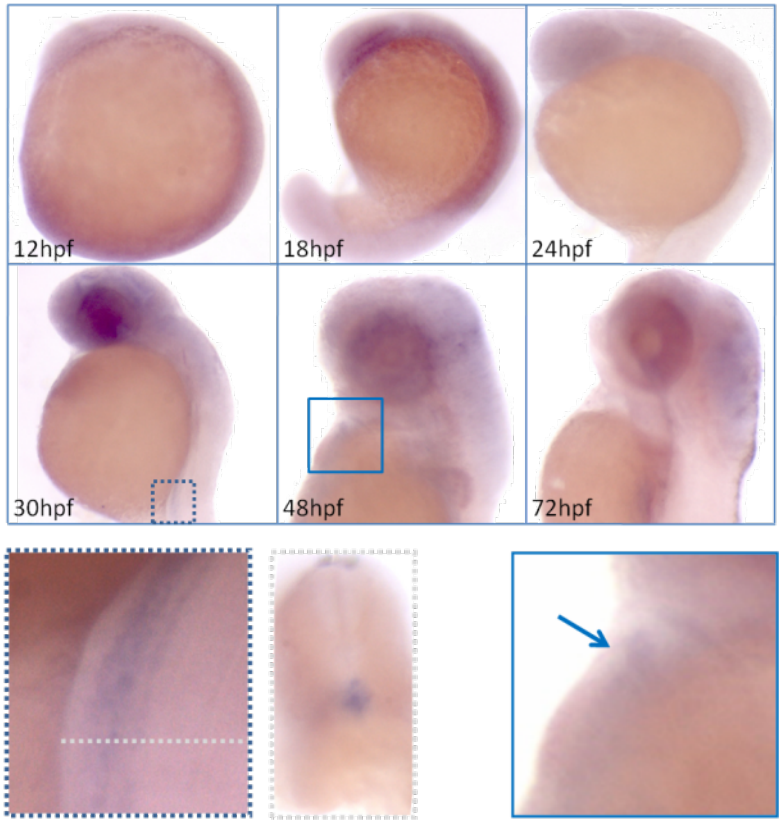
Thienpont et al. *Am J Hum Genet.* 2010

Translocation t(2;6)(q21;q25)

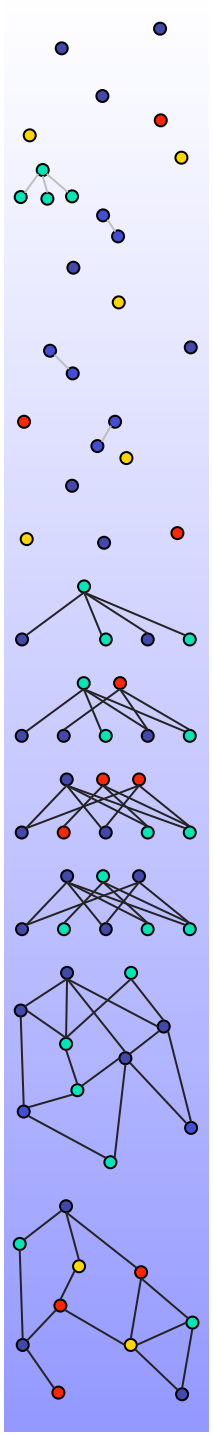


Thienpont et al. *Am J Hum Genet.* 2010

Zebrafish morpholino knock-down

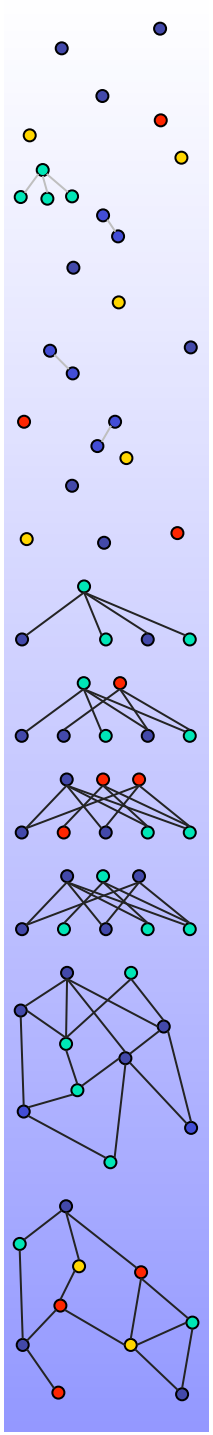


Thienpont et al. *Am J Hum Genet.* 2010



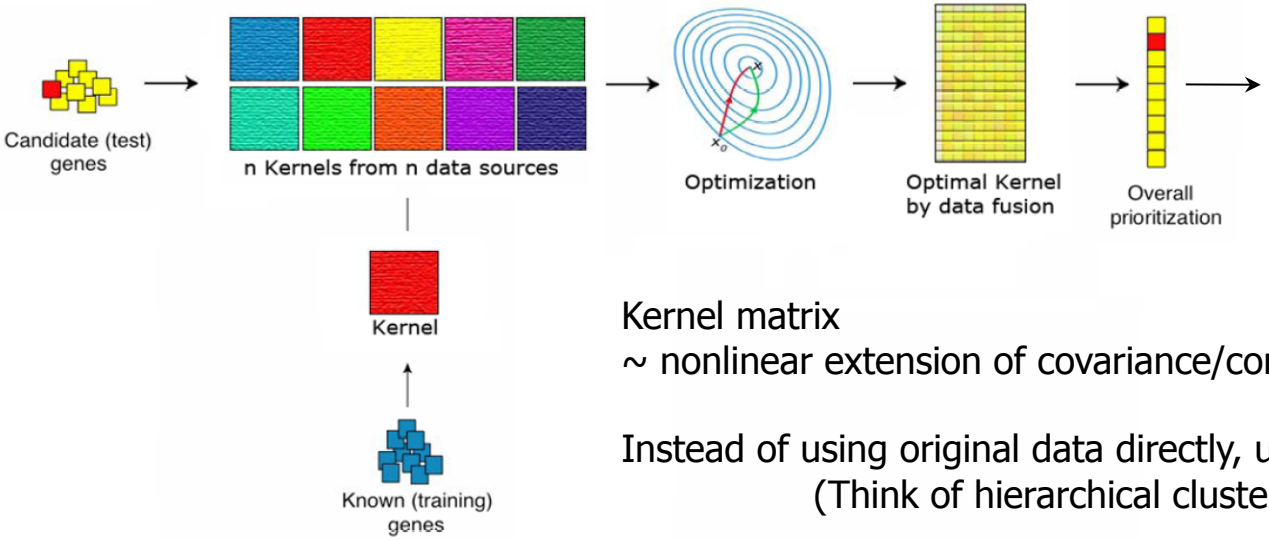
Mutation sequencing

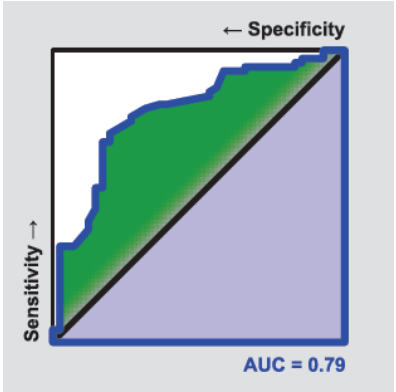
- Sequencing of TAB2 in 270 CHD patients revealed 2 missense mutations



Kernel methods for genomic data fusion

Kernel-based genomic data fusion





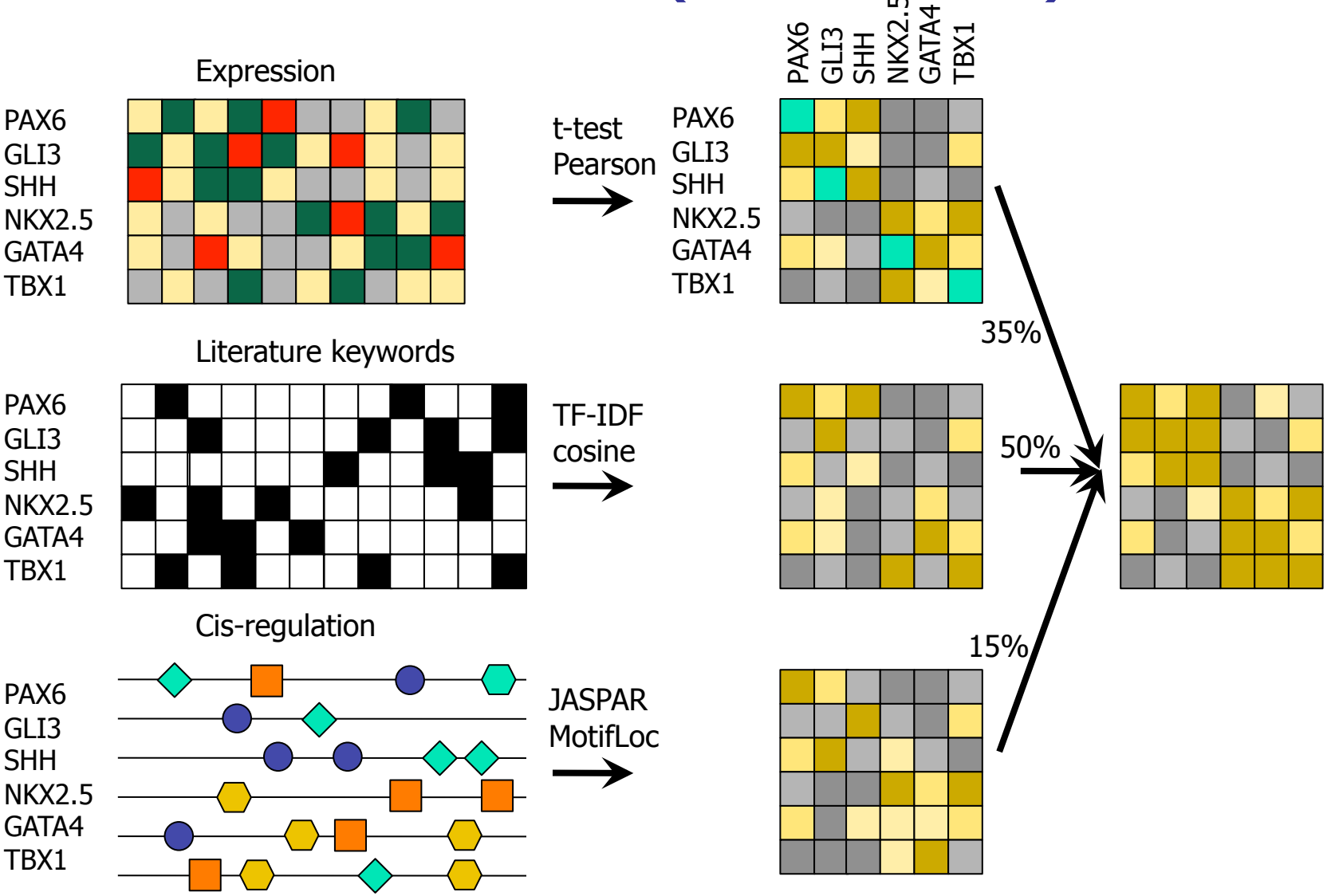
Kernel matrix
~ nonlinear extension of covariance/correlation matrix

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

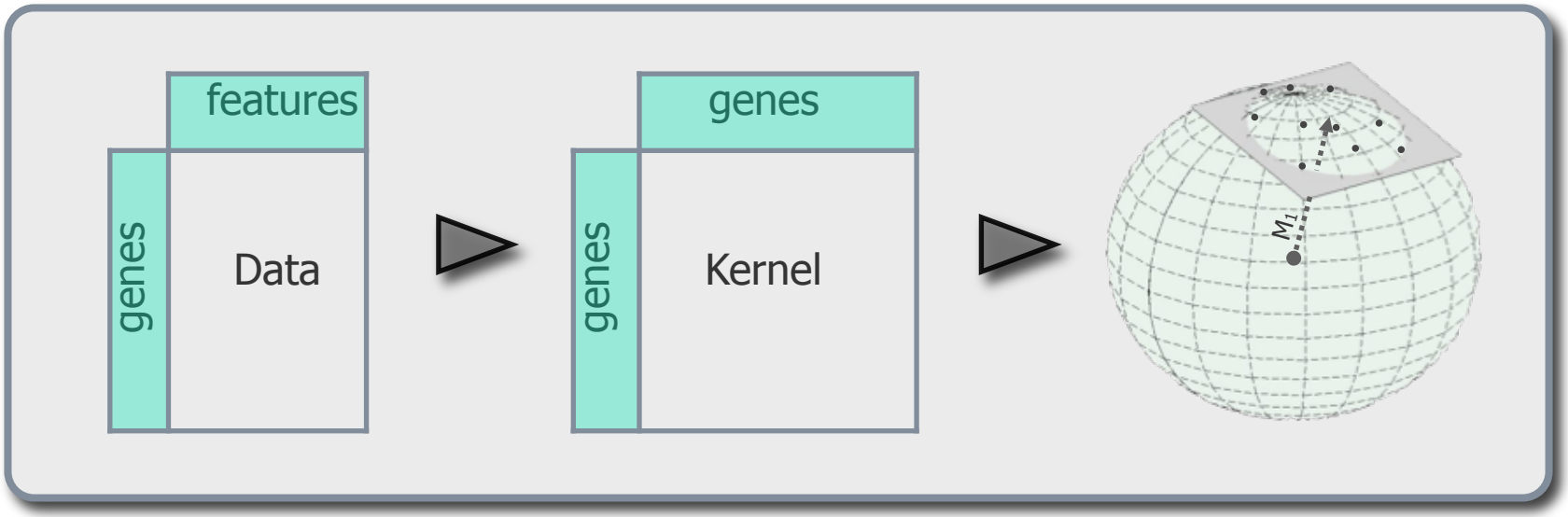
Advantage 1: kernel matrices form a single type of object, regardless of the heterogeneity of the original data types

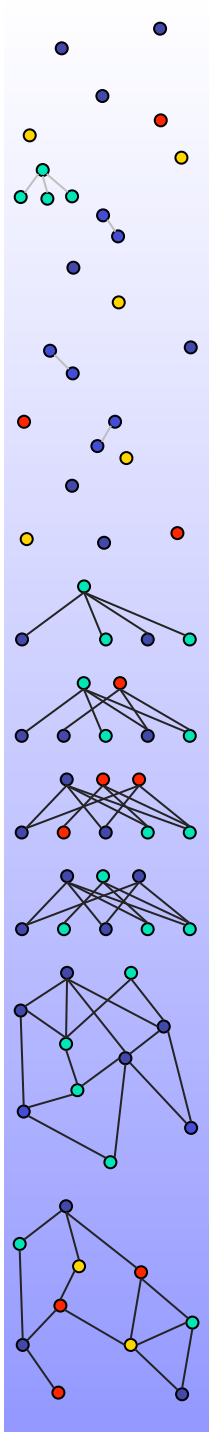
Advantage 2: all machine learning methods can be applied to kernels (classification, clustering, prioritization, ranking, etc.)

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



Prioritization by novelty detection





One-class support vector machine

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

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

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

$\vec{w}$: 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

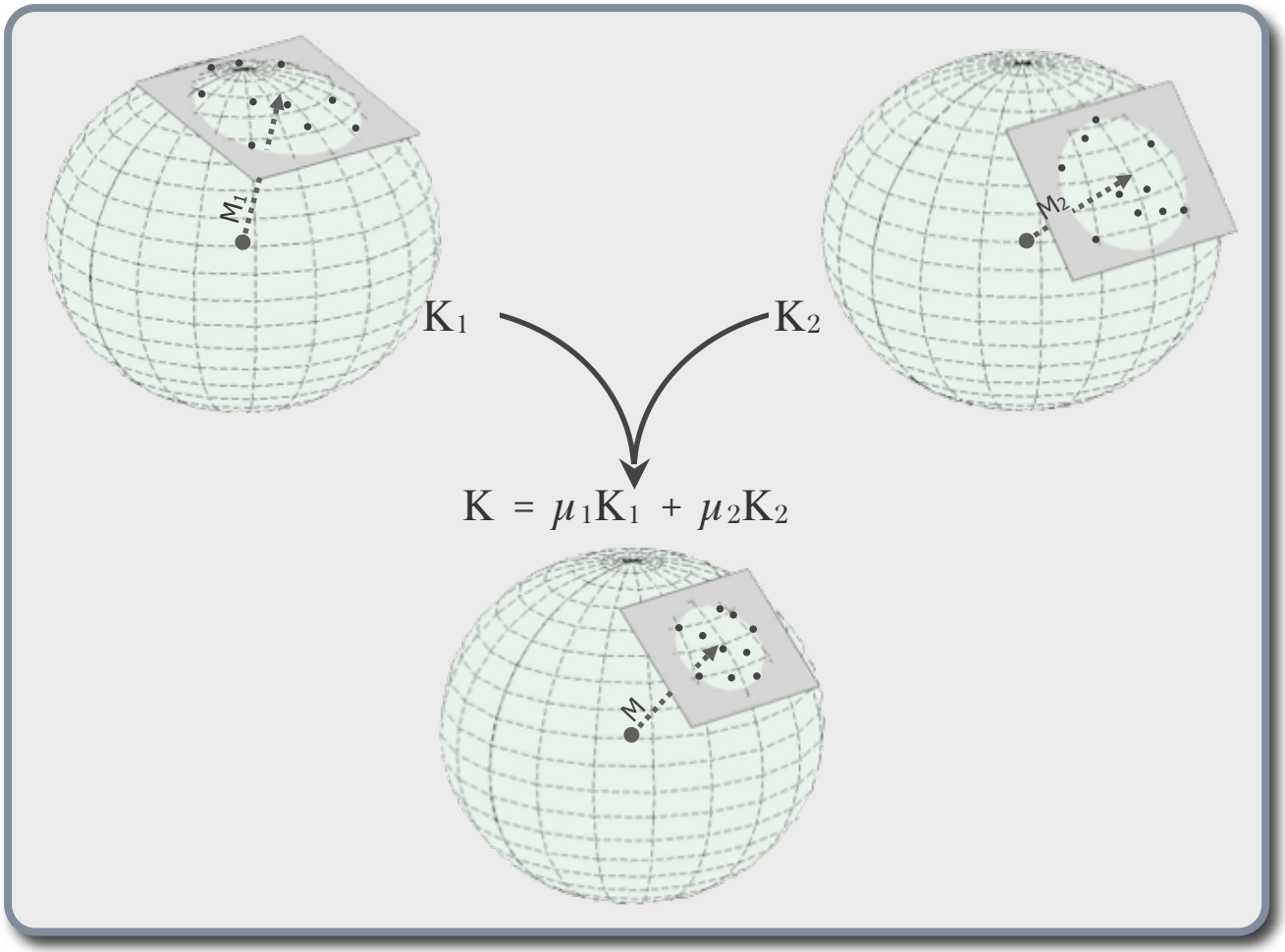
$$\boxed{\text{D:}} \min_{\vec{\alpha}} \vec{\alpha}^T K \vec{\alpha}$$

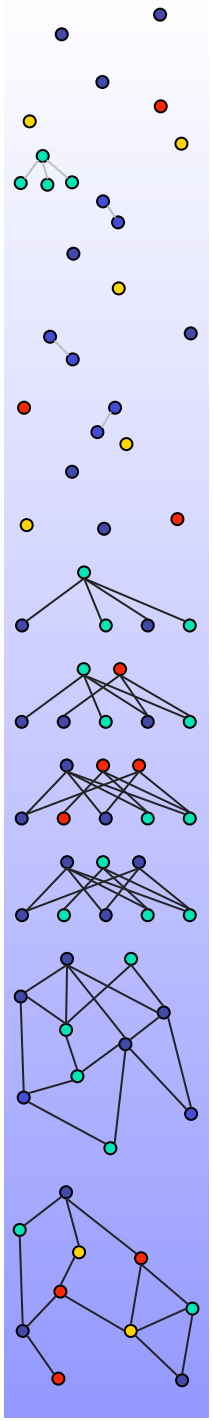
$$\text{s.t.} \quad 0 \leq \alpha_k \leq \frac{1}{\nu N}, \quad k = 1, ..., N$$

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

α_k: the dual variables
K: the kernel matrix

Kernel fusion for novelty detection





Kernel fusion in one-class SVM

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

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

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

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

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

s_j: dummy variables

L_2 vs. L_∞ kernel fusion

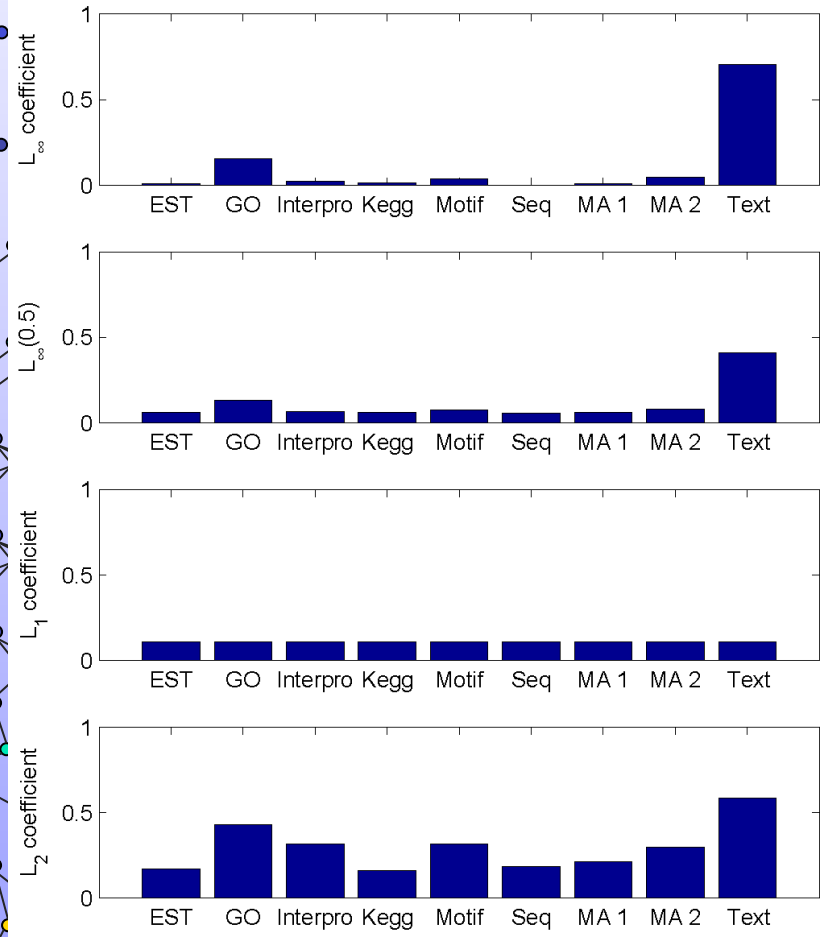
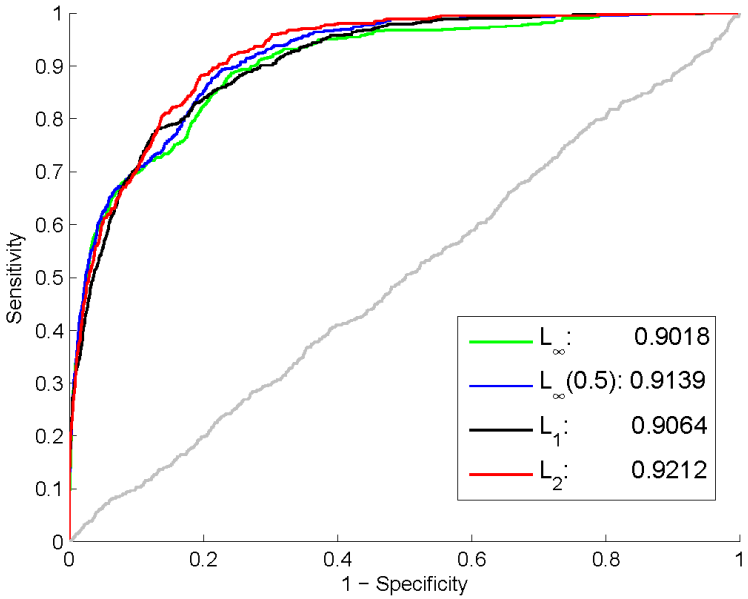
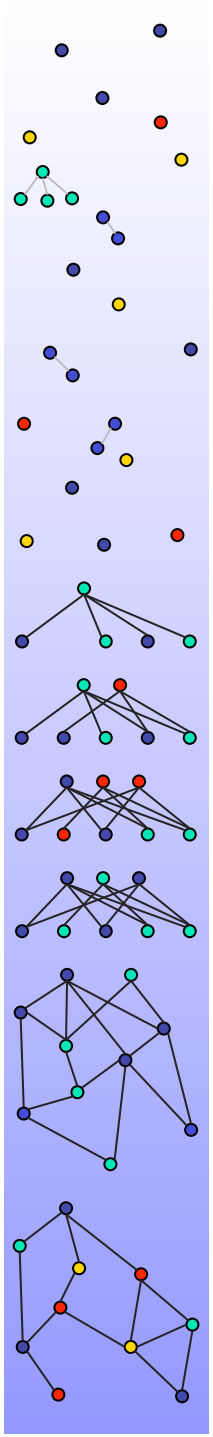


Table 1: AUC values of LOO performance evaluated from 20 random repetitions. The paired Spearman correlation scores indicate the similarities of rankings obtained by different approaches compared with the target rankings (denoted as -).

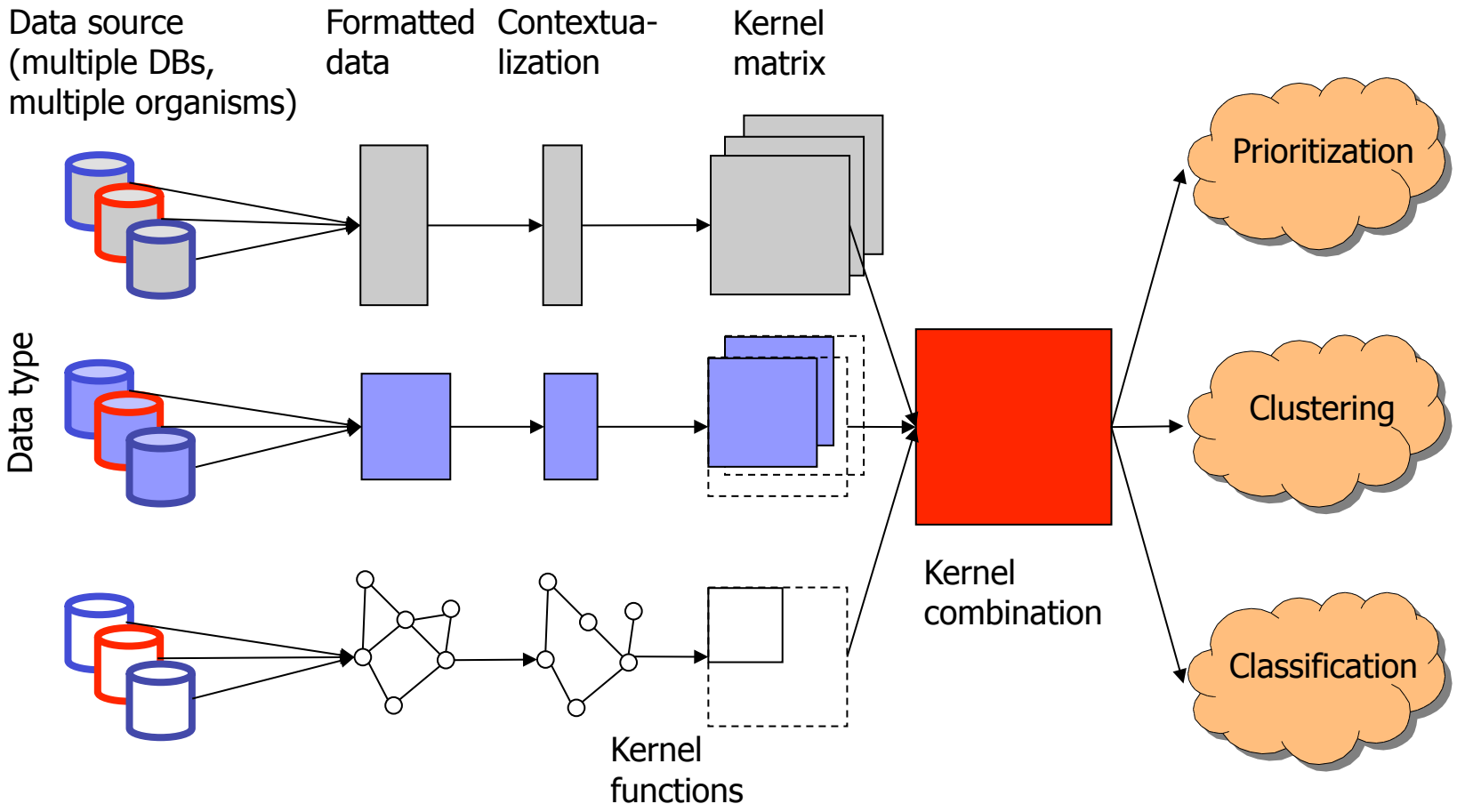
	AUC	corr	corr	corr	corr
L_∞	0.9045(0.0043)	-	0.94	0.66	0.82
$L_\infty(0.5)$	0.9176(0.0040)	0.94	-	0.82	0.92
L_1	0.9103(0.0035)	0.66	0.82	-	0.90
L_2	**0.9219(0.0034)**	0.82	0.92	0.90	-

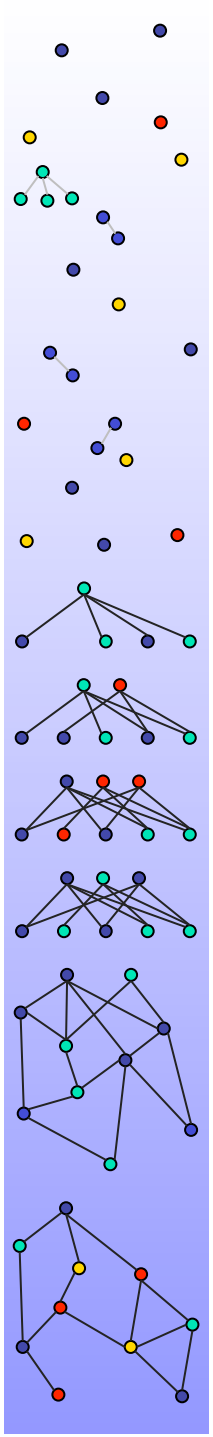




A framework for kernel data fusion

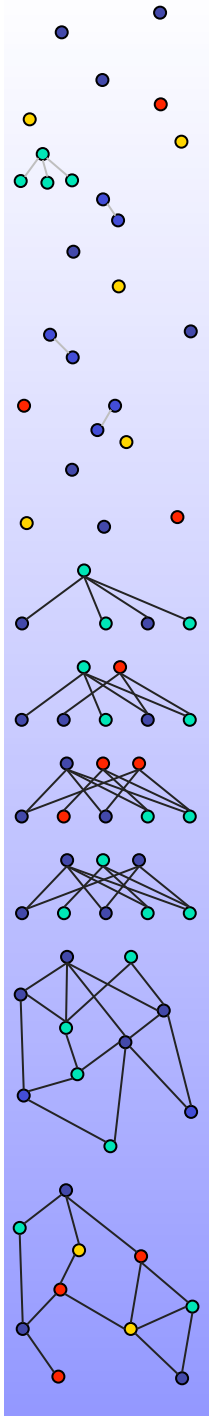
Kernel data fusion





ETkL: Extract, Transform, Kernelize, Learn

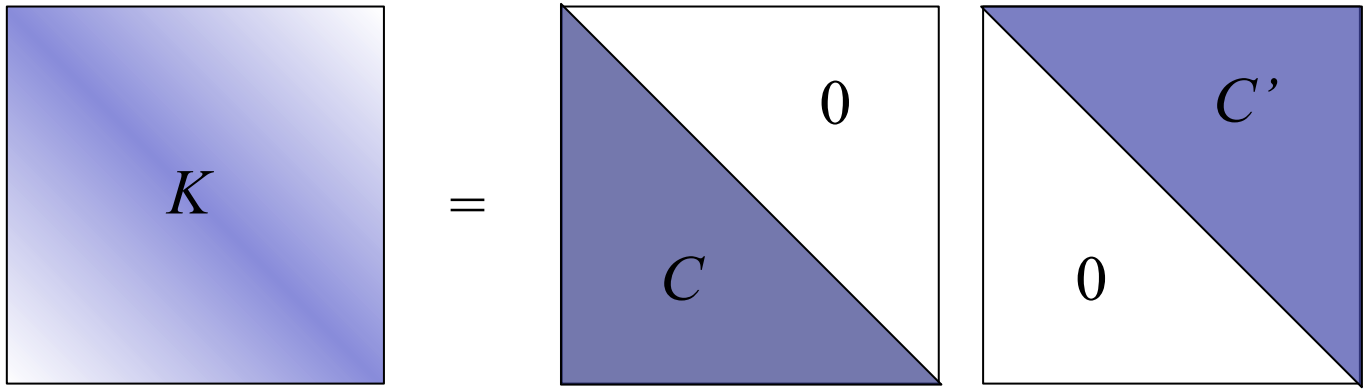
- Systematic multi-tier framework for data integration
  - Resembles multi-tier architecture of complex IT systems and Extract-Transform-Load methodology of datawarehousing
    1. Database / web service sources
    2. Data reconciliation, cleaning, and warehousing, etc.
    3. Scaling, normalization, 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

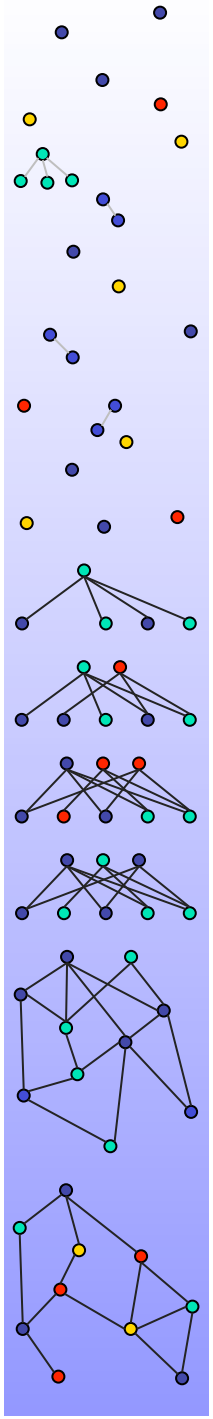


Handling large kernel matrices

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

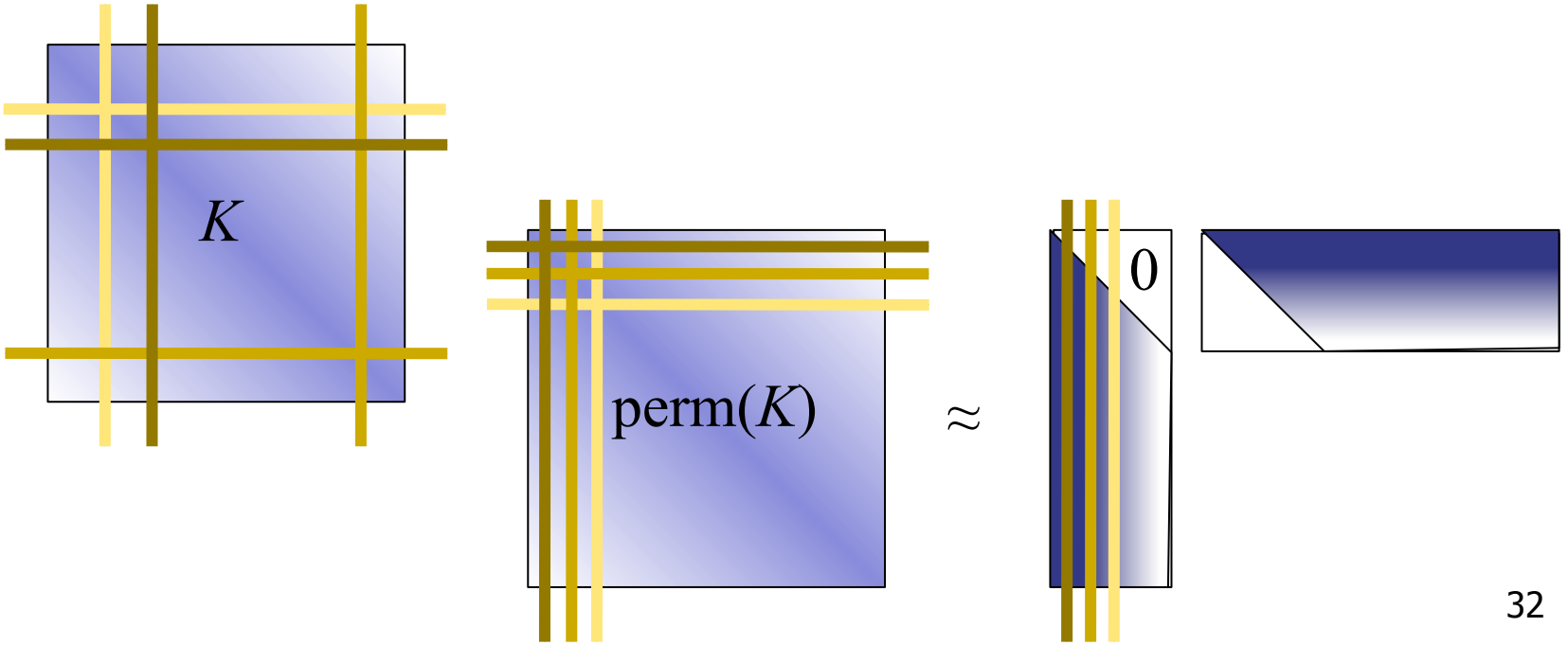
$$\exists C\,(\text{lower triangular} \;\&\; \text{unique}) : K = CC'$$

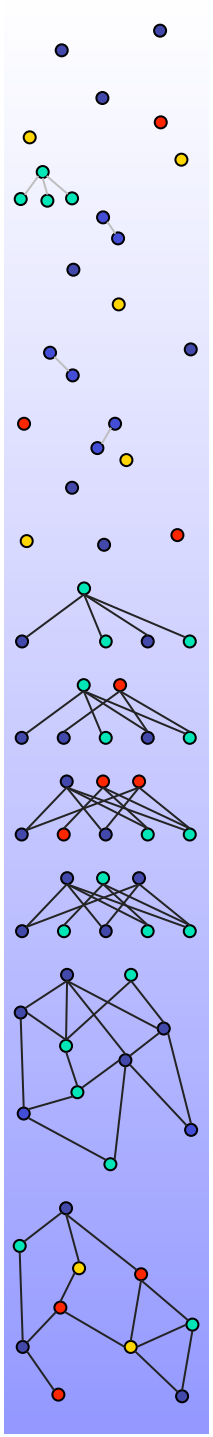




Incomplete Cholesky decomposition

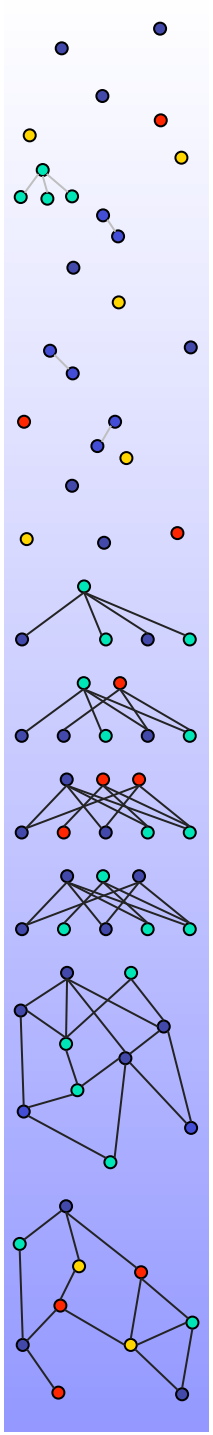
- Incomplete Cholesky
  - K symmetric positive semidefinite
  - Limit to rank $r \leq \mathrm{rank}(K)$
  - Add pivoting to capture more informative rows/columns first
  - Limit information loss to e.g. 5%



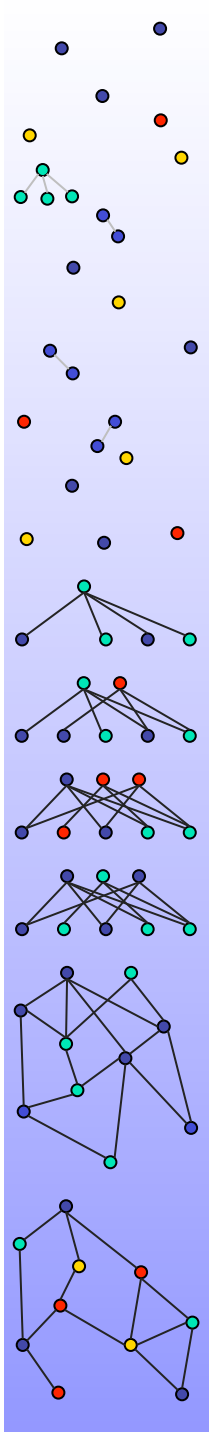


The No-Voodoo principle

- 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




extasy
Variant Prioritization



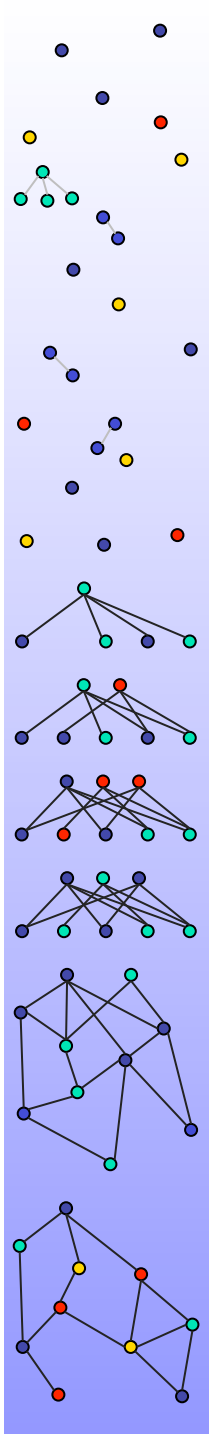
Challenges

- About 2,000 rare coding variants per patient
- About 5 *de novo* coding variants per patient

- Tractable by filtering
  - Loss-of-function (truncating, splice site) mutations
  - Two patients with de novo variants in same gene
  - Recessive mutations in inbred families
  - Multiple patients with rare variants in the same gene (association)

- Challenging
  - What about locus heterogeneity?
  - What about compound heterozygotes?
  - What about oligogenic disorders?

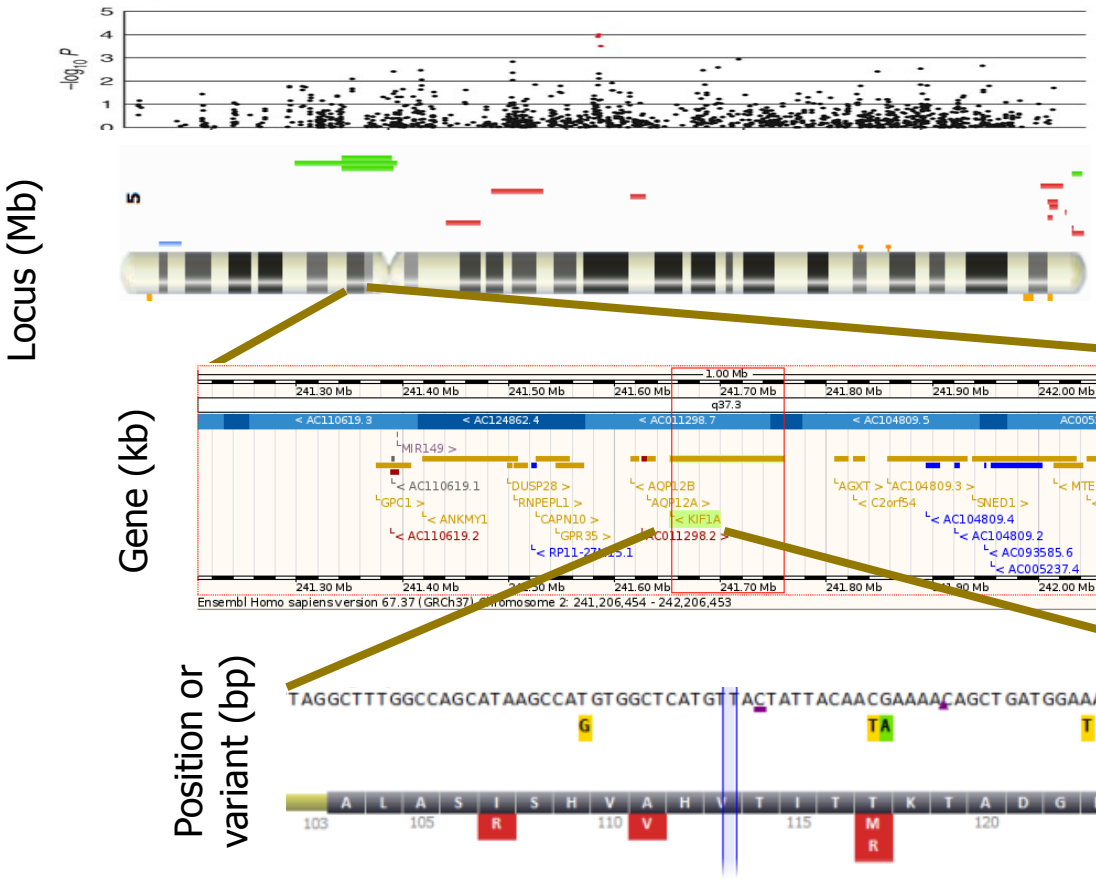
  ➔ Need to prioritize variants



Variant prioritization

- Variant and basepair level
  - *Structural change:* change from one nucleotide to the another will change the amino-acid encoded at that position, which will change the structure of the protein and thus its function
  - *Association:* variant is present more often in patients than controls
  - *Conservation:* position at which the variant is found is highly conversed across species and evolution is apparently reluctant to see this position changed
- Gene level
  - *Haploinsufficiency:* gene in which the variant is found is putatively haploinsufficient
  - *Gene prioritization:* gene in which the variant is found is known to be involved or is putatively involved in the phenotype of interest
- Locus level
  - *Locus mapping:* region of the genome in which variant is found is associated (CNV, association, linkage) with phenotype of interest

Variant prioritization



Genome-wide association studies

Copy number variation

Haploinsufficiency

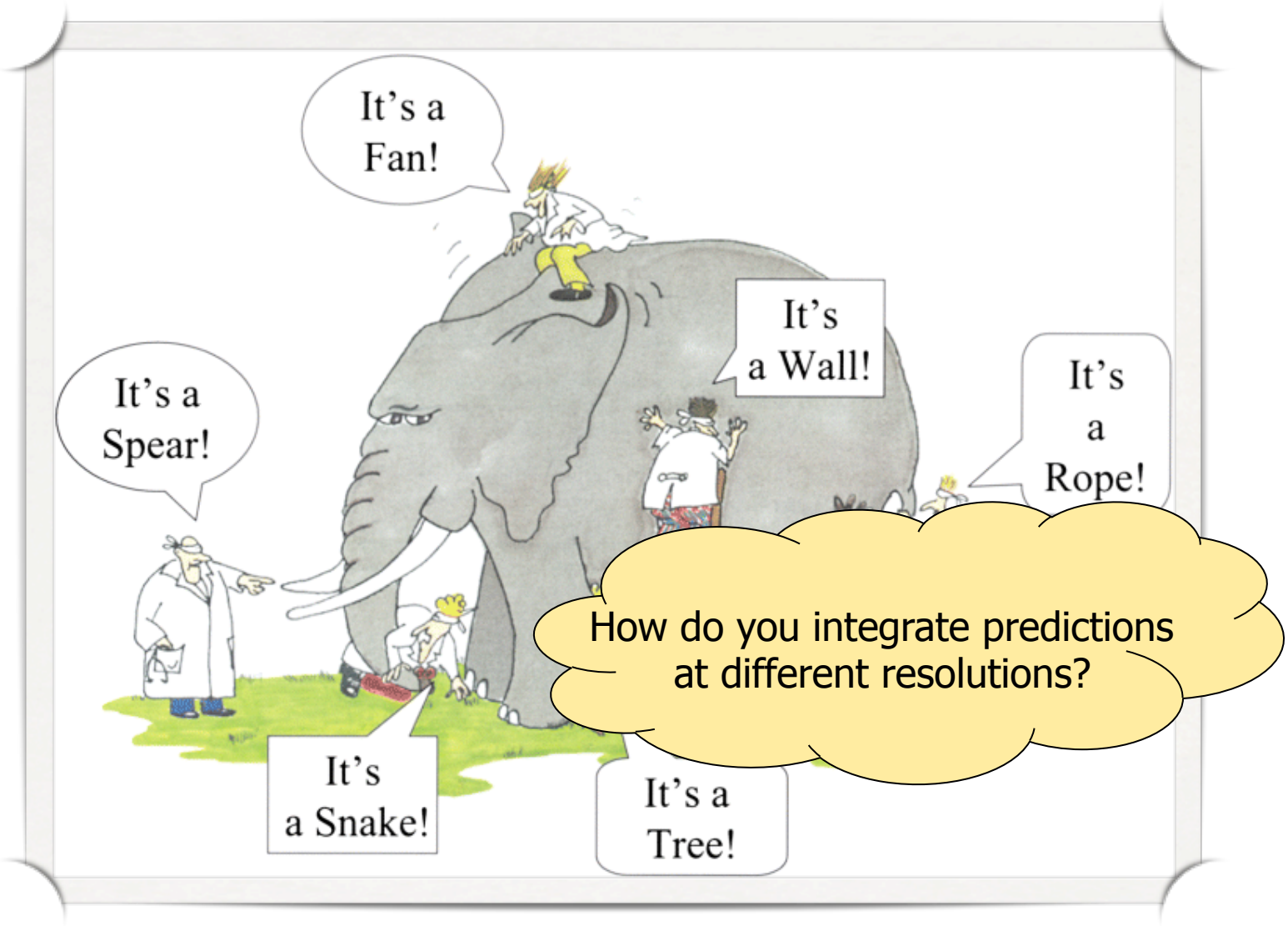
Gene prioritization

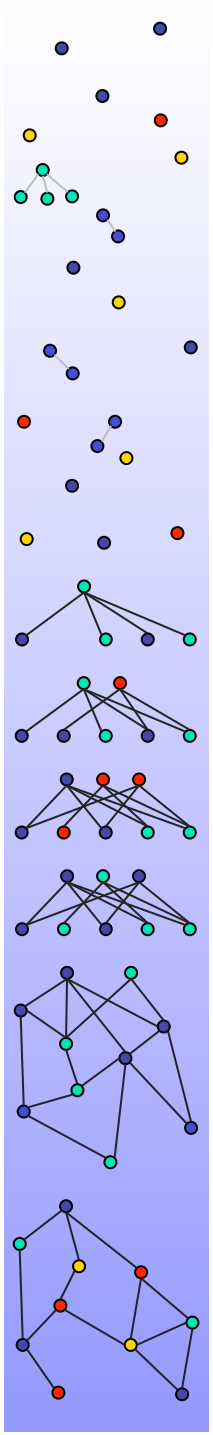
Variant association

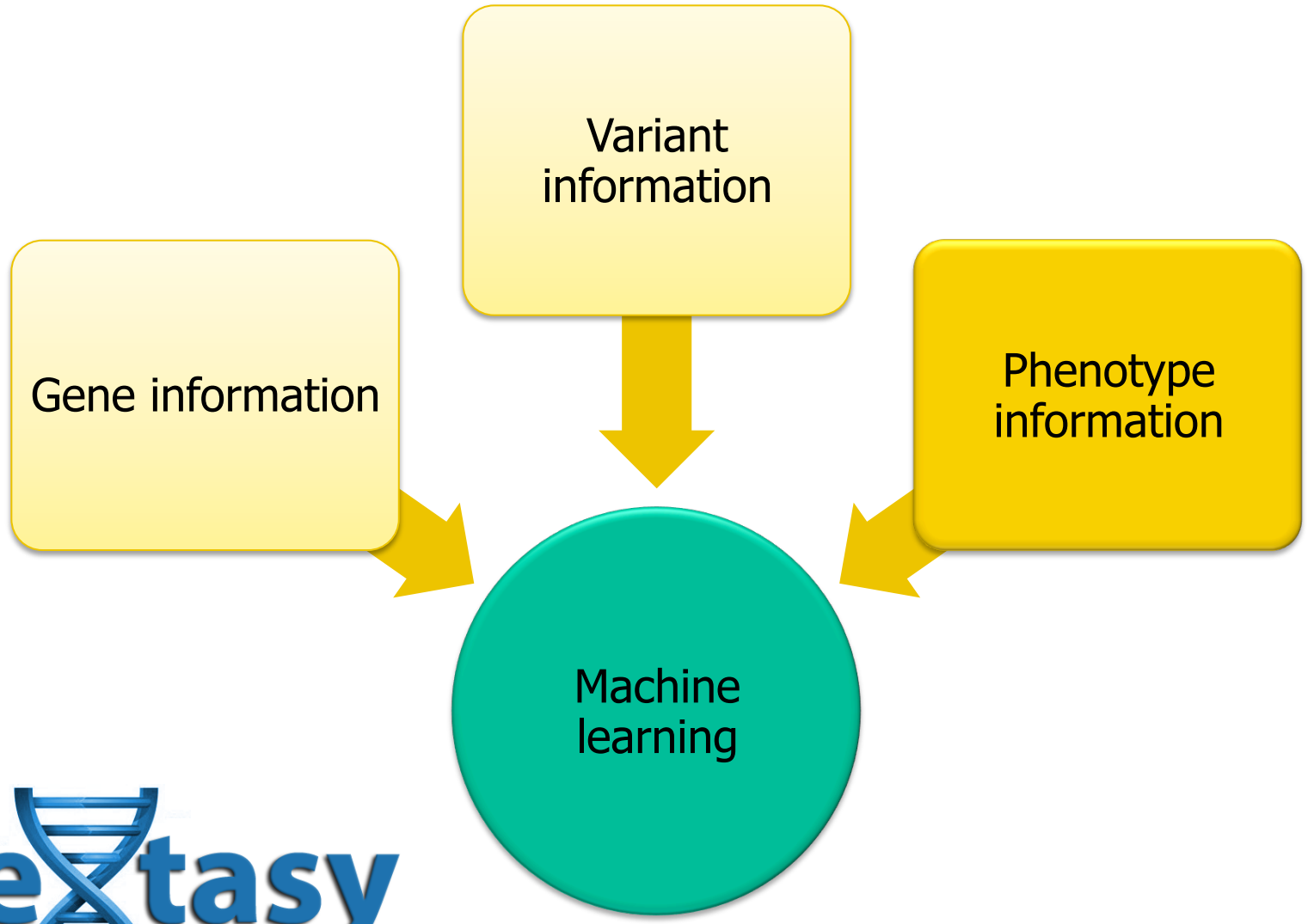
Conservation

Structural effects

Variant prioritization

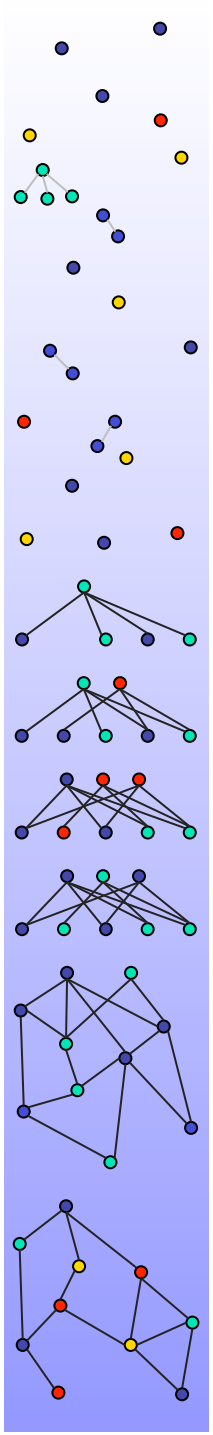






eXtasy
Variant Prioritization

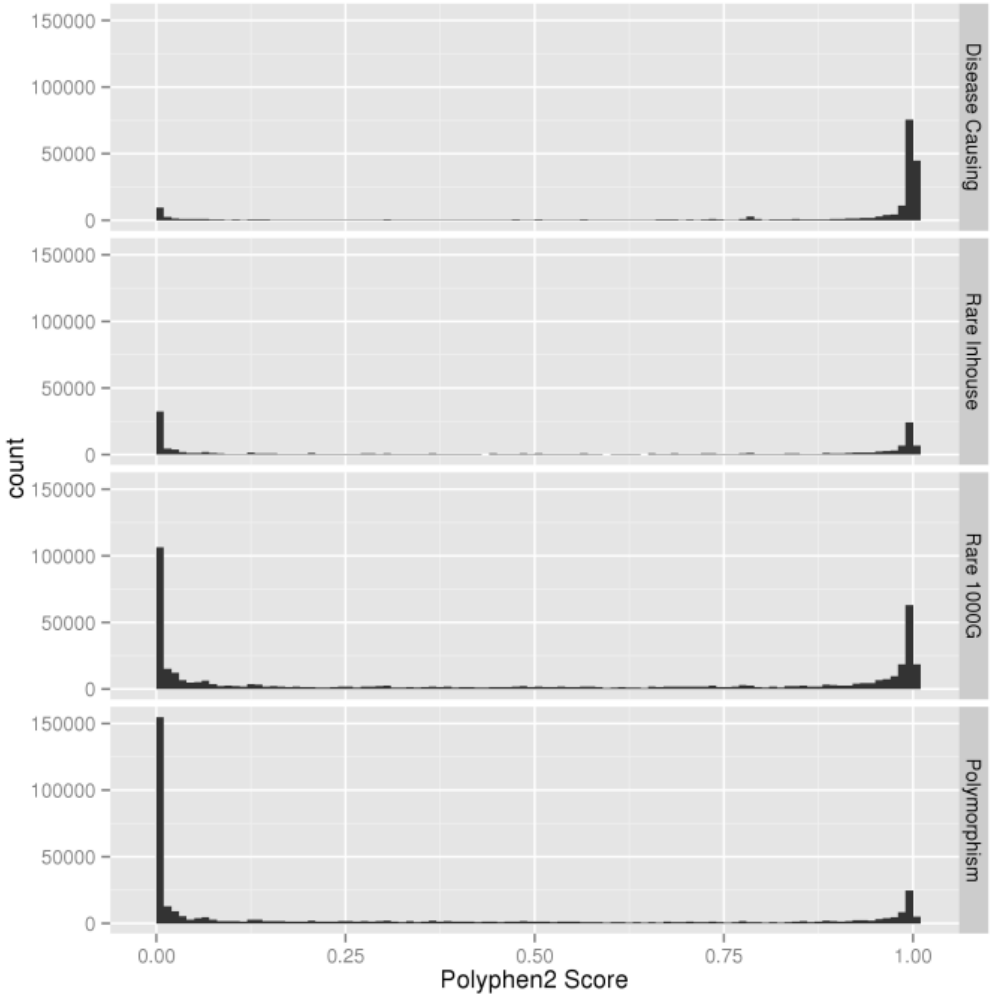
homes.esat.kuleuven.be/~bioiuser/eXtasy/



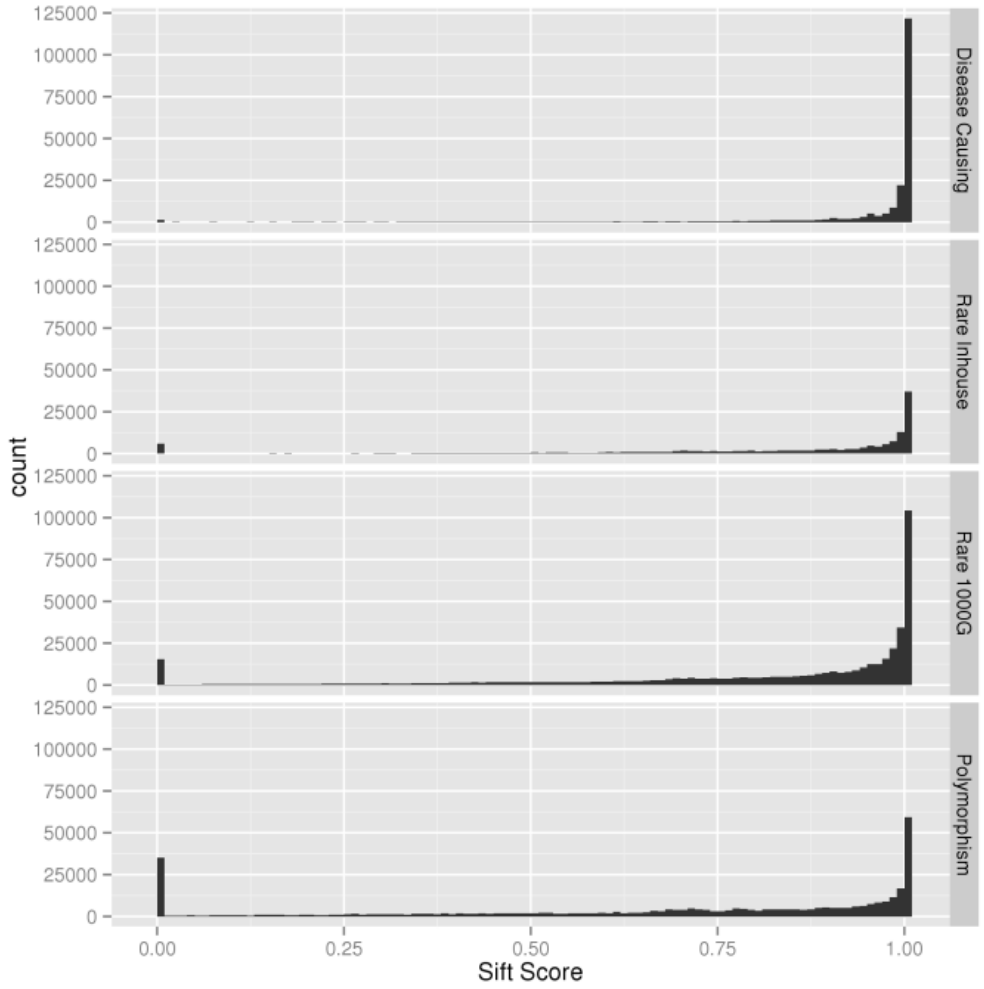
Data sets

- HGMD: 24,454 variants in 1,142 HPO terms
  - HGMD terms mapped to HPO
  - At least three genes for training of Endeavour
- Control sets (sampled 500/phenotype):
  - Polymorphisms: MAF > 1%, 1000G, 43,724 variants
  - Rare
    - MAF < 1%, 1000G, 43,724 variants
    - In-house, > 20X coverage, 257, 556 variants
- Scores from different sources mapped directly from highest to lowest level
- Existing method perform poorly on rare *a priori* benign variants vs. polymorphisms

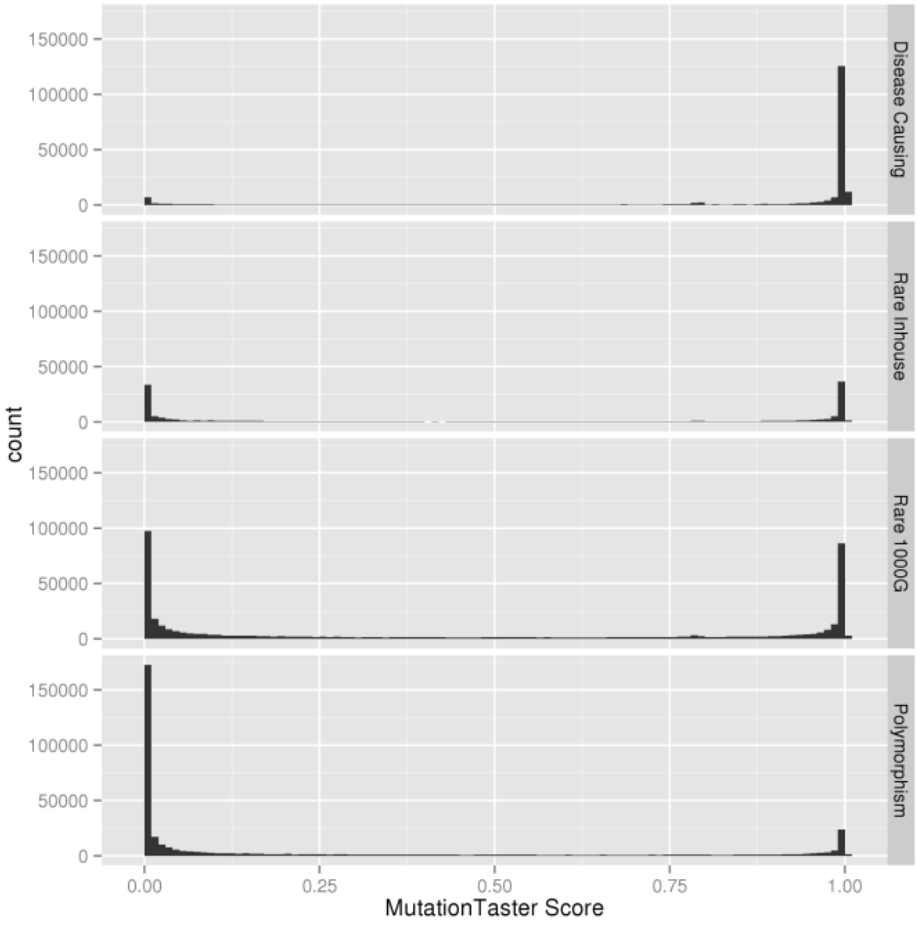
Polyphen2

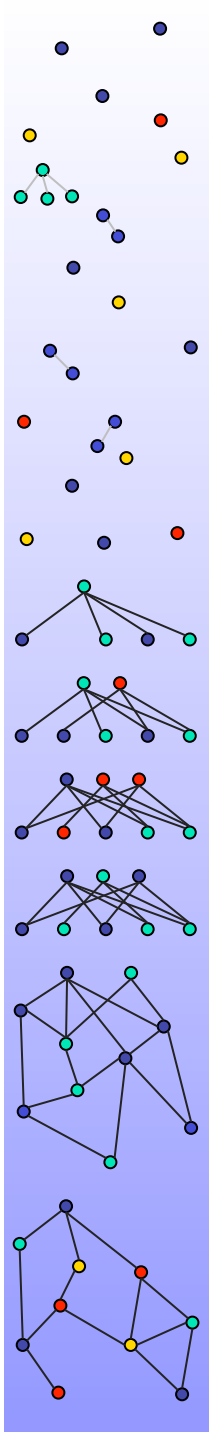

150000
100000
50000
0
Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
0.00
0.25
0.50
0.75
1.00
Polyphen2 Score

SIFT


125000
100000
75000
50000
25000
0
count
Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
0.00
0.25
0.50
0.75
1.00
Sift Score

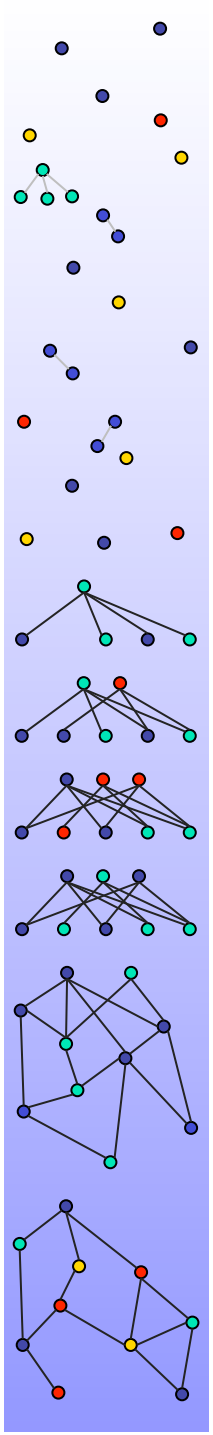
MutationTaster

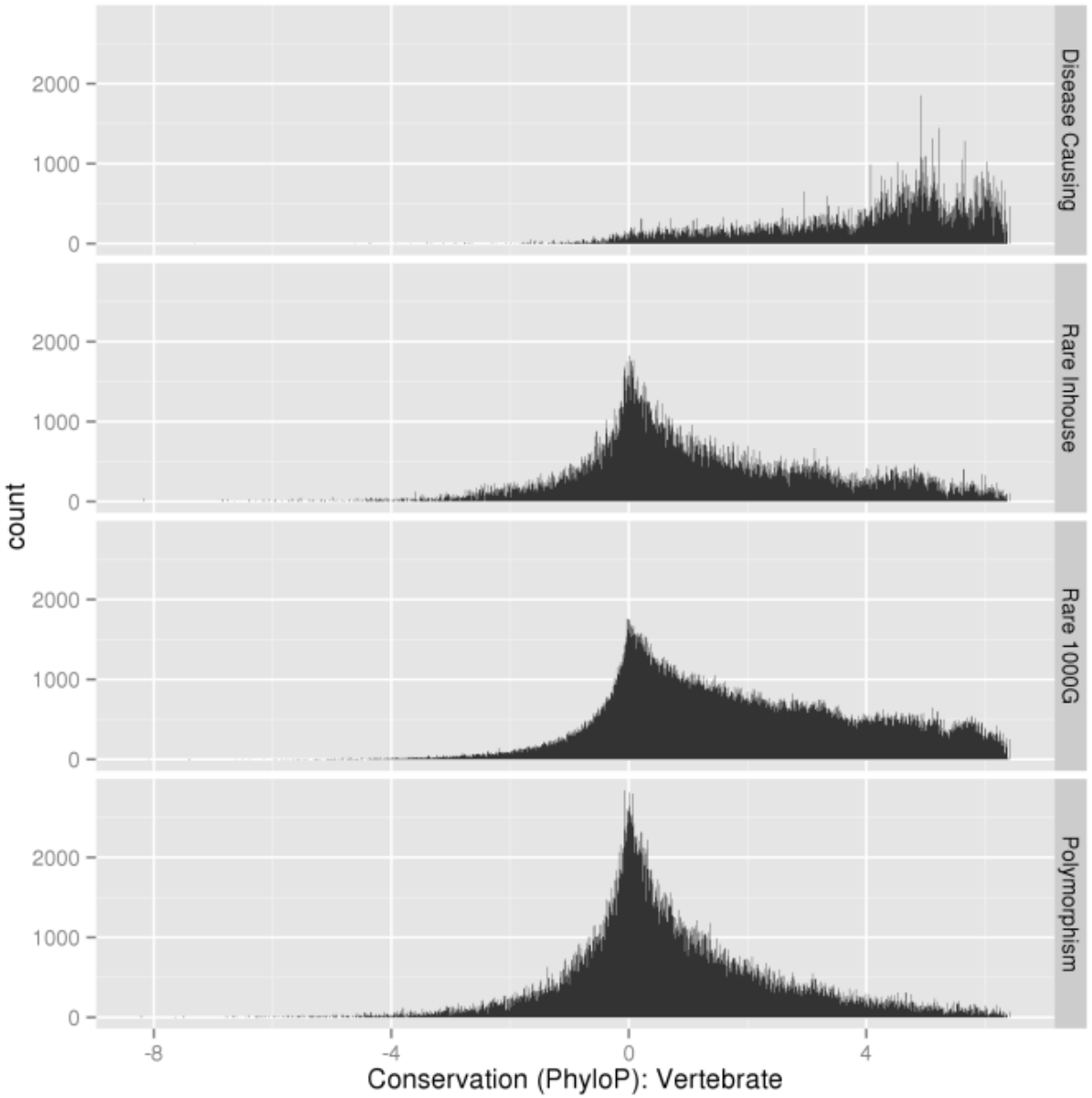


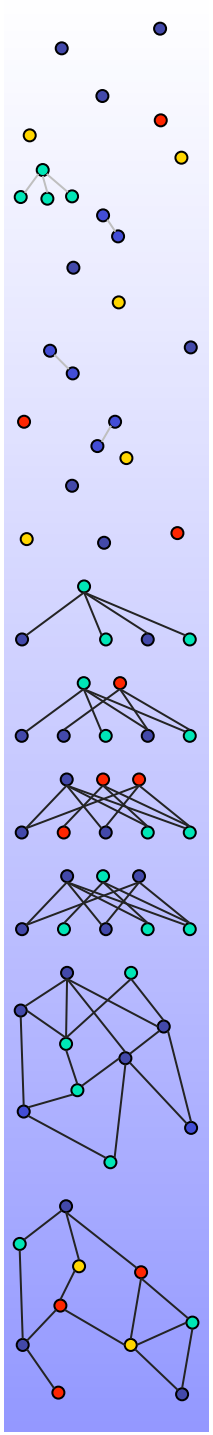


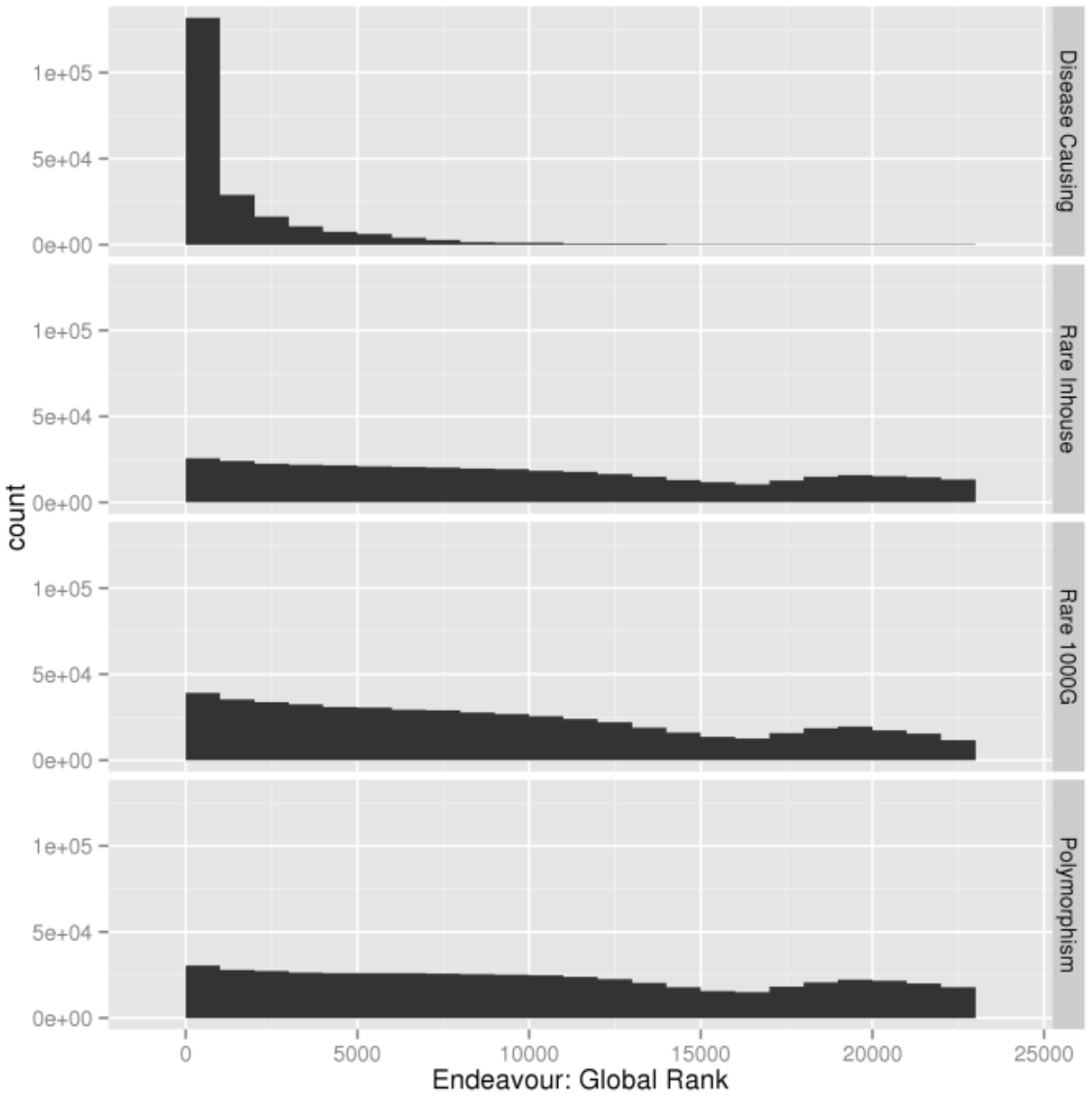
Where is the problem?

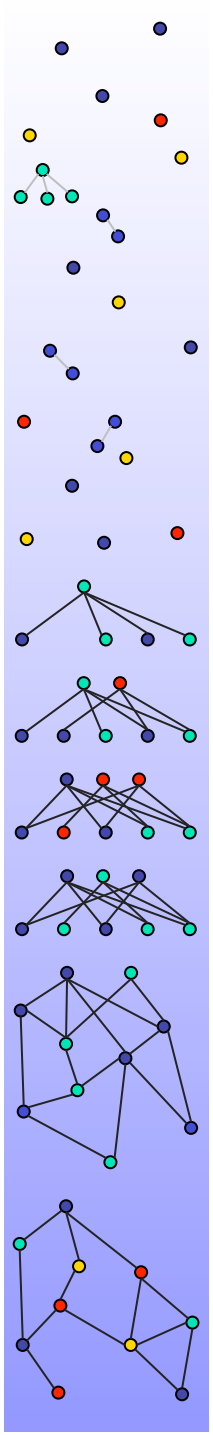
- Previous methods trained to distinguish disease-causing variants from common SNPs, not rare variants
- "Deleterious" variant = variant that affects gene function
  - Deleterious variants may not be disease causing
  - "Mildly deleterious" – Kryukov et al. (2007)
  - "Accelerated population growth and weak purifying selection" – Tennessen et al. (2012)
- Bad training sets?
- What if they are deleterious but not specific for our desired phenotype?

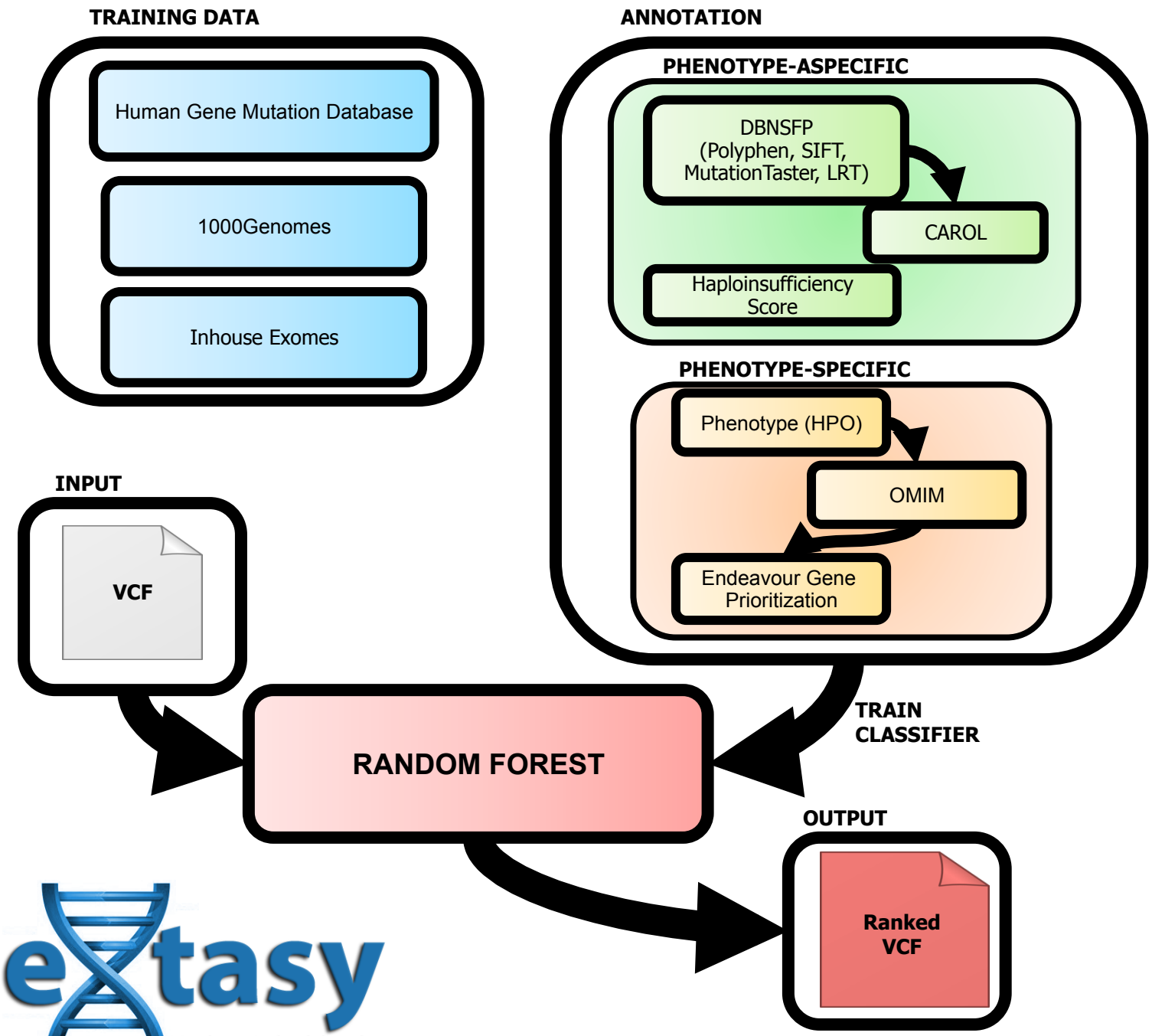


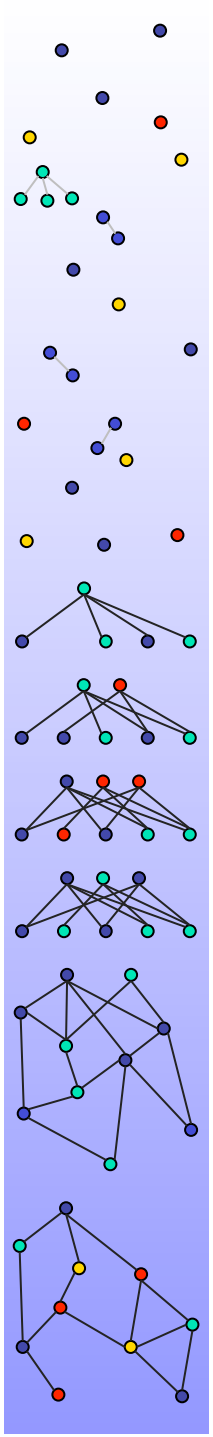

Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
2000
1000
0
-8
-4
0
4
Conservation (PhyloP): Vertebrate



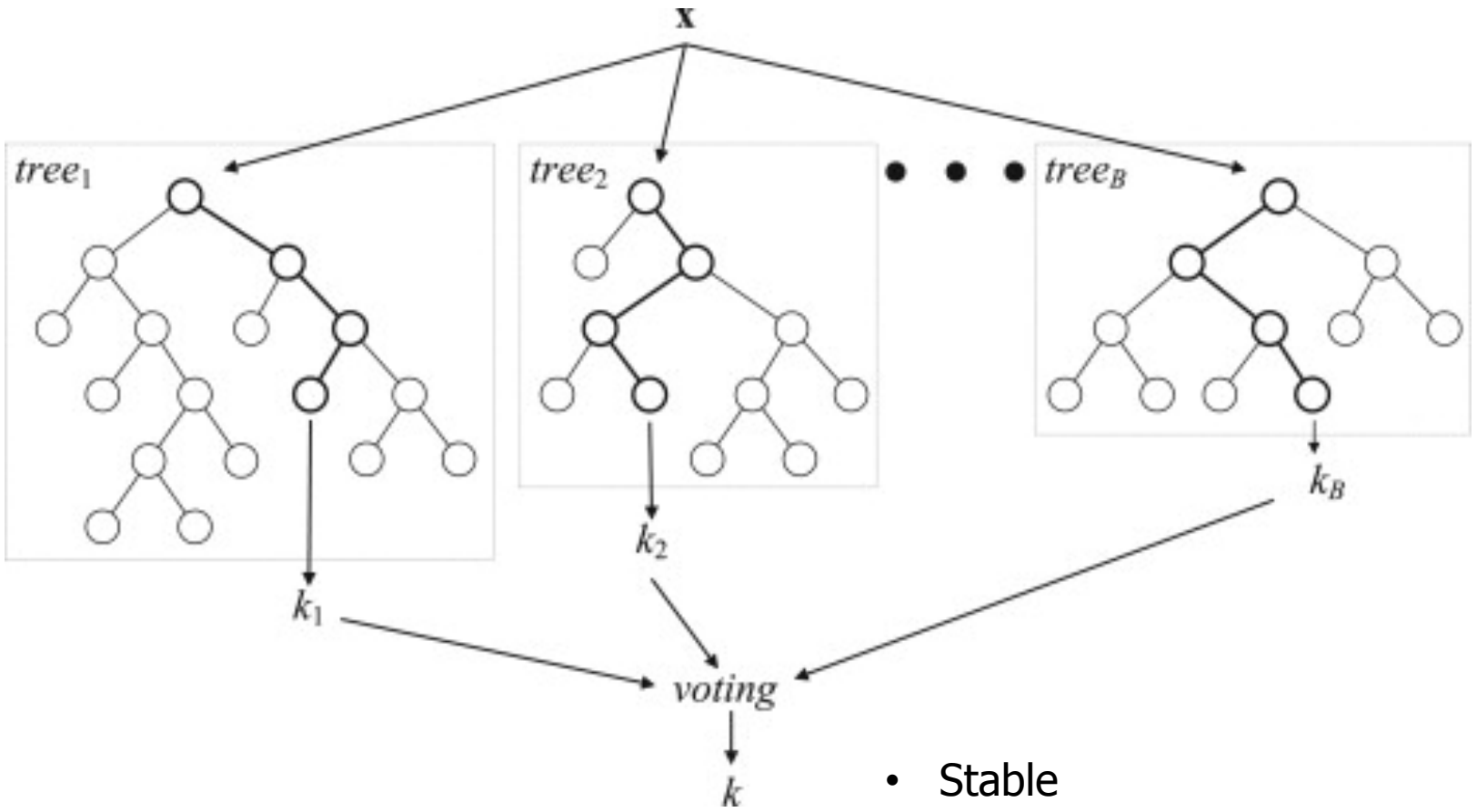

Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
1e+05
5e+04
0e+00
0
5000
10000
15000
20000
25000
Endeavour: Global Rank



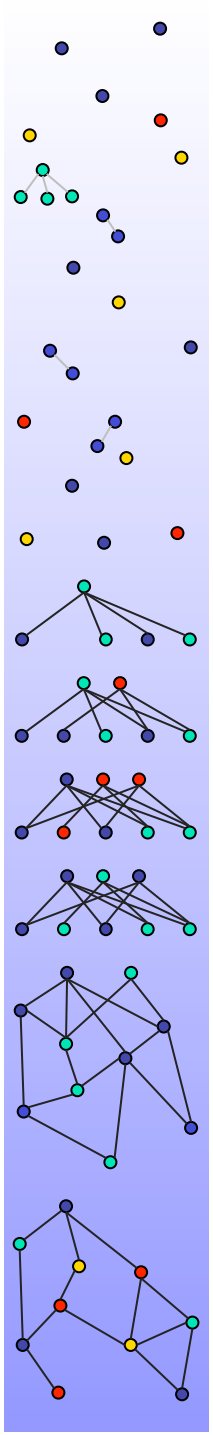

TRAINING DATA
Human Gene Mutation Database
1000Genomes
Inhouse Exomes
ANNOTATION
PHENOTYPE-ASPECIFIC
DBNSFP
(Polyphen, SIFT,
MutationTaster, LRT)
CAROL
Haploinsufficiency
Score
PHENOTYPE-SPECIFIC
Phenotype (HPO)
OMIM
Endeavour Gene
Prioritization
INPUT
VCF
RANDOM FOREST
TRAIN
CLASSIFIER
OUTPUT
Ranked
VCF
extasy
Variant Prioritization



Random forests

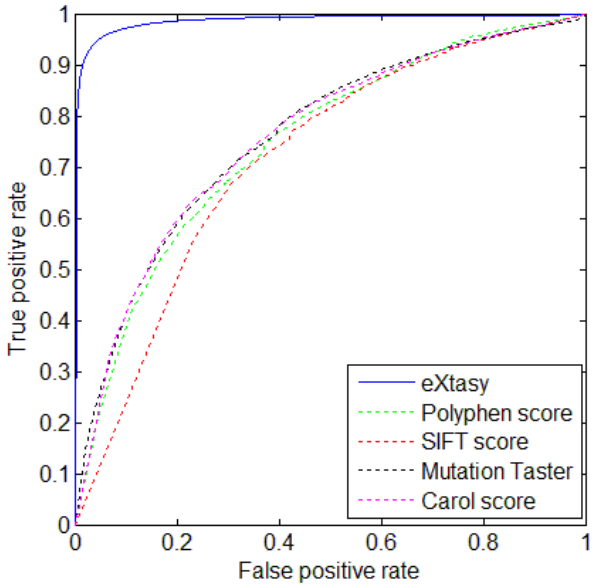


- Stable
- Fast
- Semi-interpretable

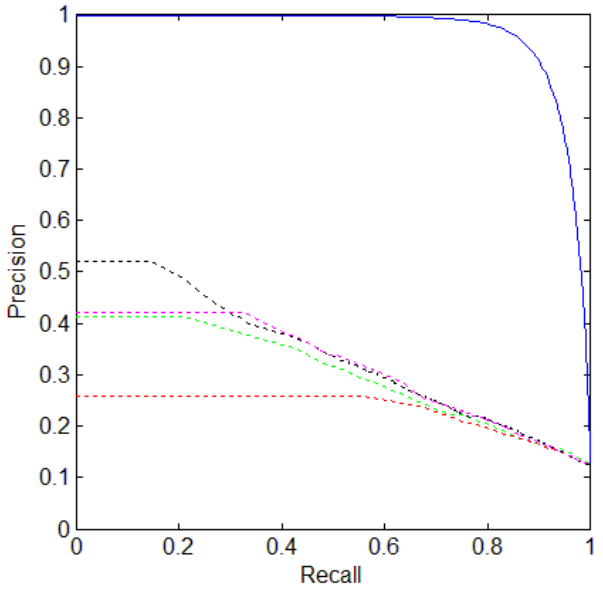


Rare variants

A

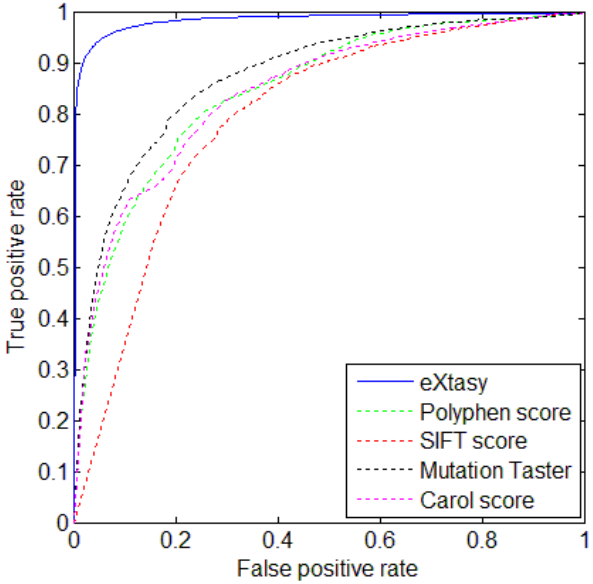


B

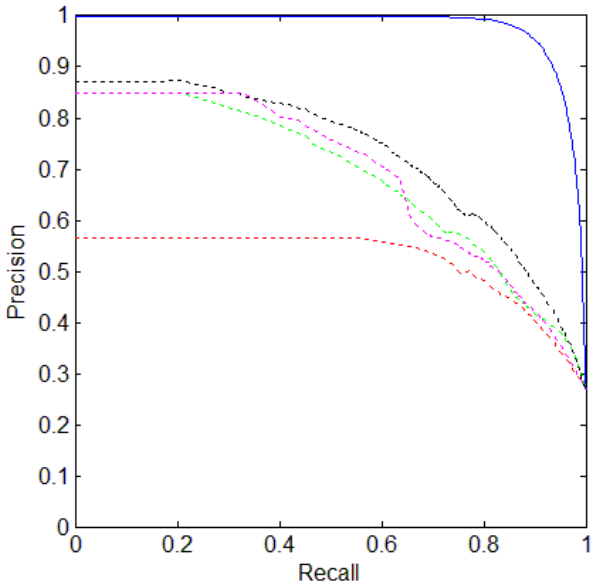


Polymorphisms

C

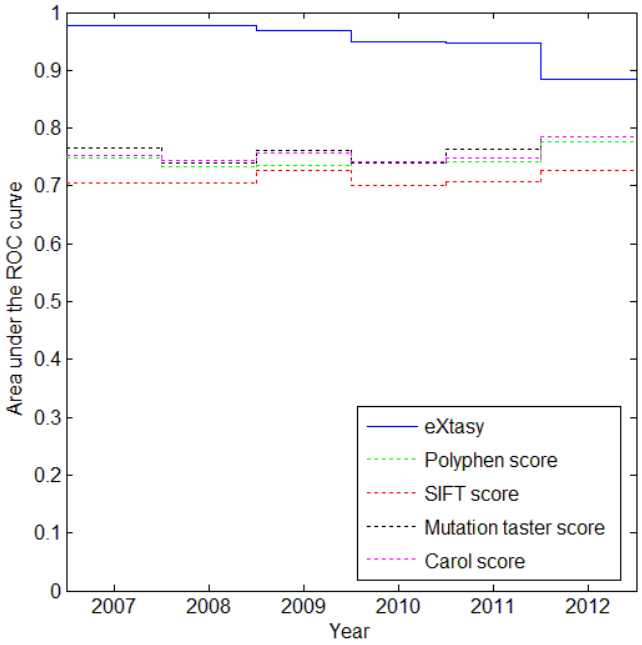


D

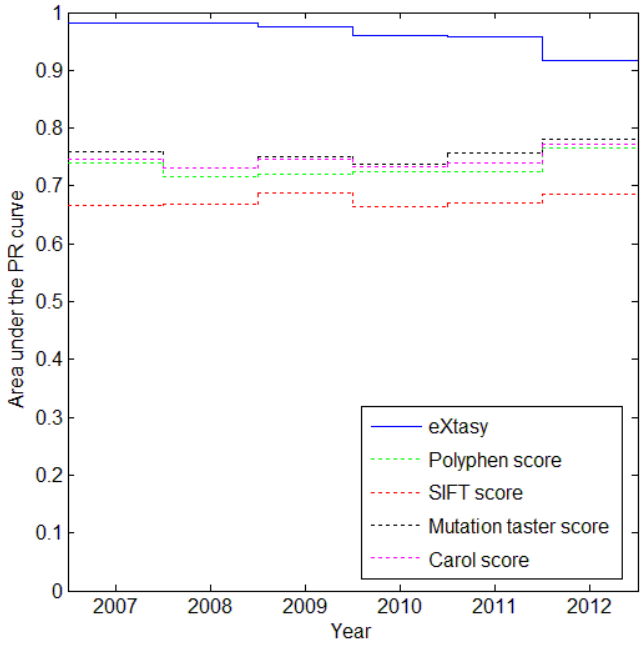


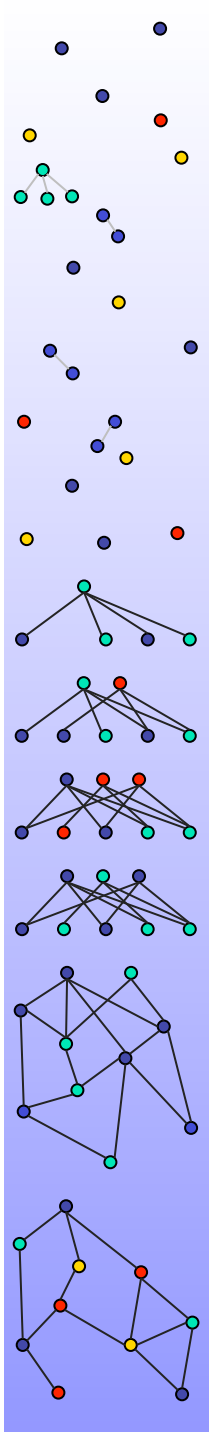
Temporal stratification

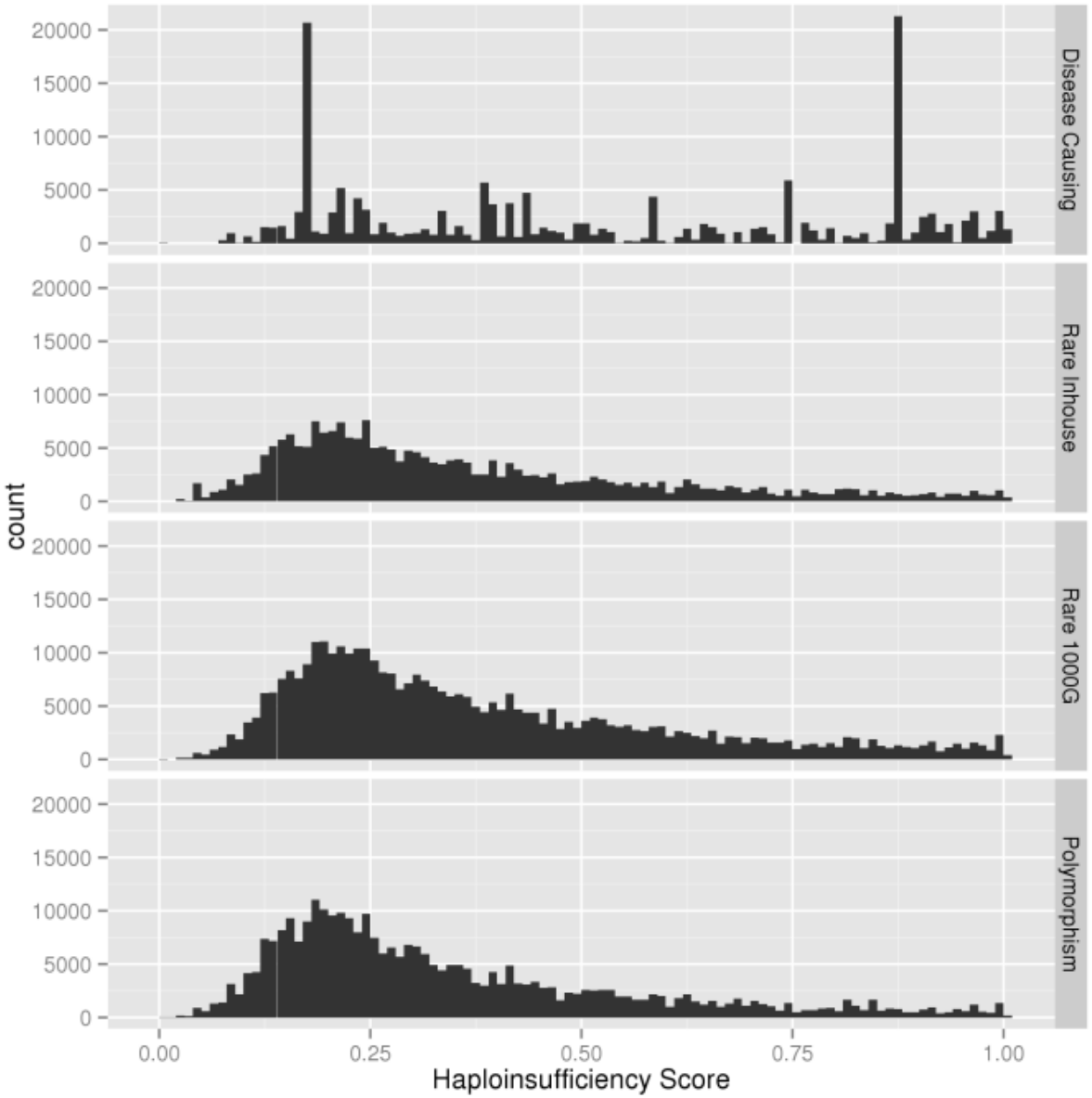
A

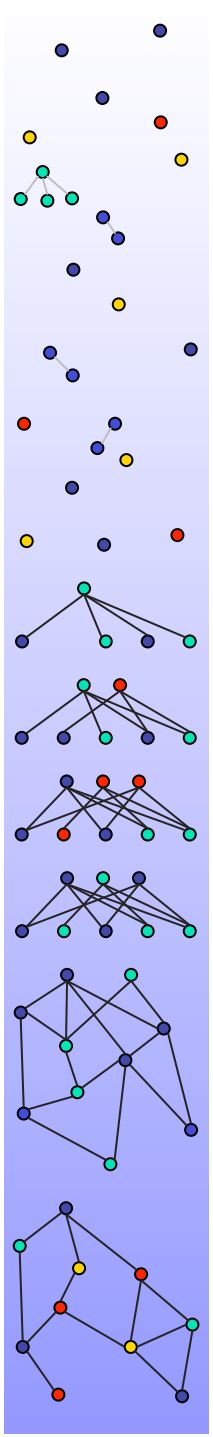


B



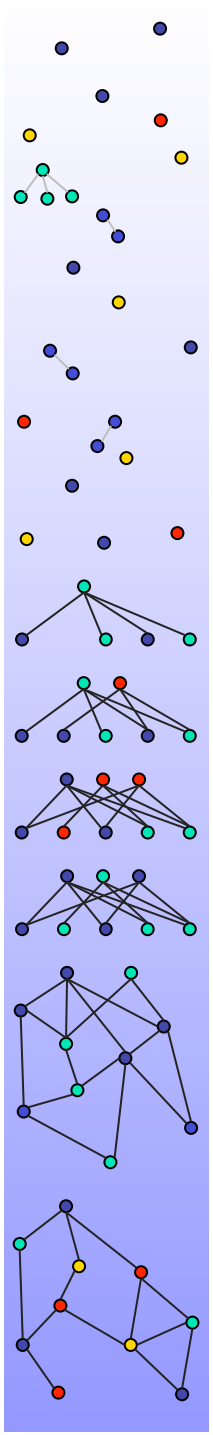



Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
20000
15000
10000
5000
0
0.00
0.25
0.50
0.75
1.00
Haploinsufficiency Score



What's the catch?

- Data sets are biased
  - Benchmark on known mutations
  - Retrospective benchmarks are overoptimistic!
- High proportion of negative variants
  - Despite good discrimination, still lots of false positives



Run eXtasy online:

eXtasy Variant Prioritization

Email: moreau@esat.kuleuven.be

HPO term: HP:0000347 Micrognathia

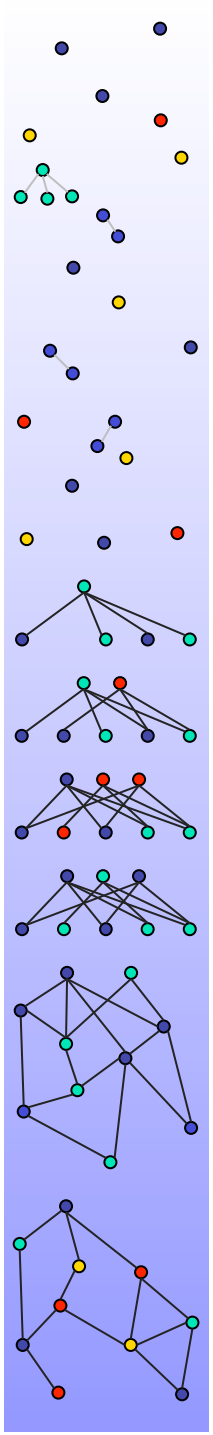
VCF file: Choose File miller.vcf

Submit

Example Data: miller.vcf,schinzel_giedion.vcf

We provide two example vcf files which were generated by adding published disease causing variants for Miller syndrome (causative gene: DHODH, Ng et al., 2010, Nature Genetics) or Schinzel-Giedion syndrome (causative gene: SETBP1, Hoischen et al., 2010, Nature Genetics) to a publicly available VCF file of the exome of a healthy individual (obtained from here). These files can be prioritized against any of the phenotype terms which characterize the syndromes. For Schinzel-Giedion this could for example be *HP:0009924 (Hypoplasia/aplasia involving the nose)* or for Miller syndrome this could be *HP:0000347 (Micrognathia)*.

homes.esat.kuleuven.be/~bioiuser/eXtasy/



Conclusions and perspectives

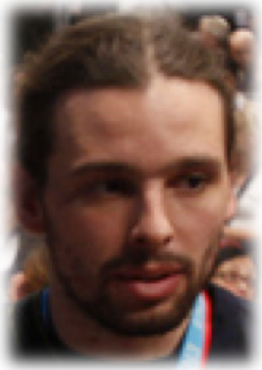
- Genomic data fusion for disease gene prioritization
- Kernel methods for genomic data fusion
- Extract, Transform, kernelize & Learn

- Phenotype information improves variant prioritization
- Importance of reference data
  - Common SNPs
  - Rare *a priori* benign variants
  - *Common and rare variants from local population*
- Scoring for multiple phenotypes
- Further integration with locus info (GWAS, CNV)
- Further integration with variant association scoring
- Scoring other mutations (synonymous, indels, noncoding)

`homes.esat.kuleuven.be/~bioiuser/eXtasy/`

KU LEUVEN

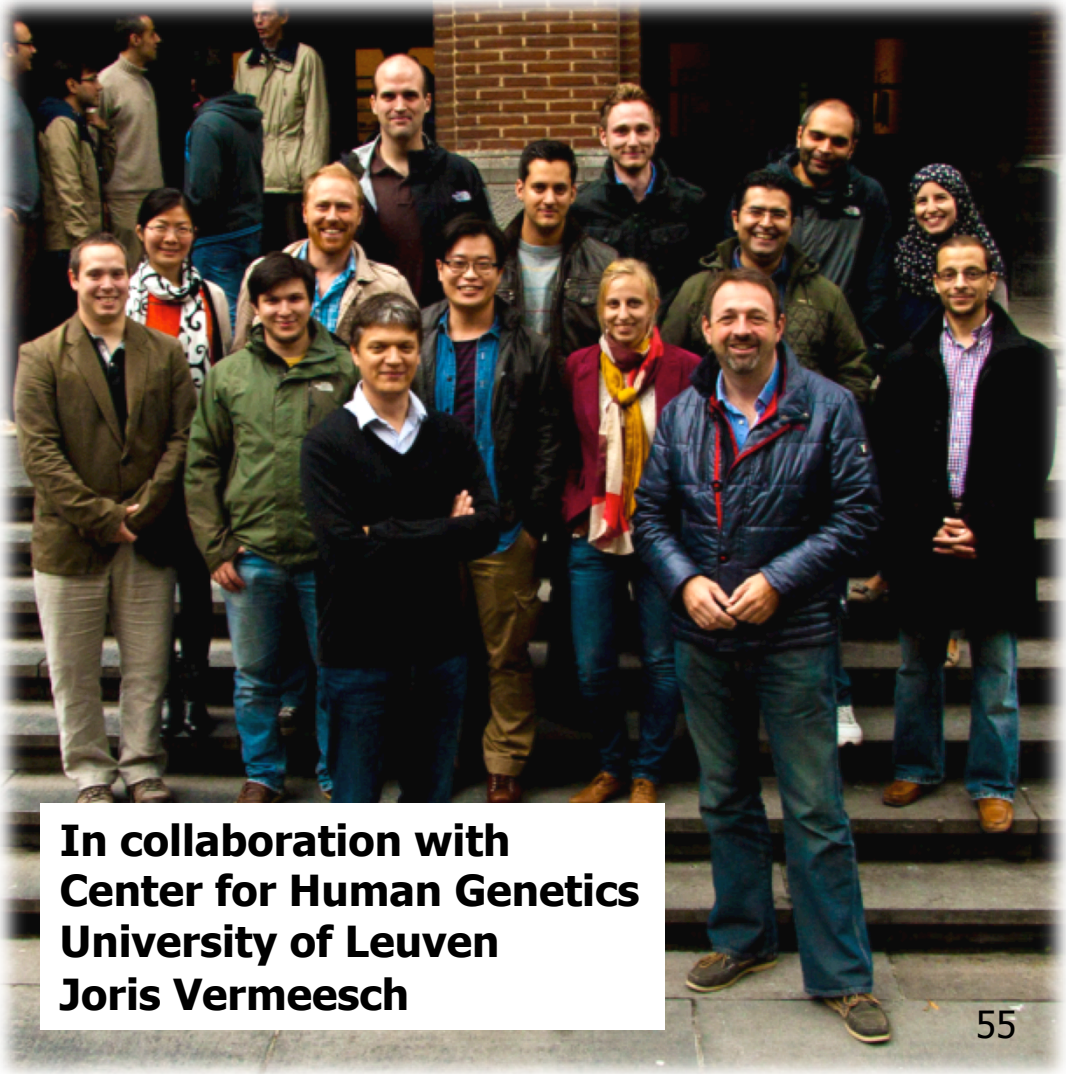
SymBioSys



Leo Tranchevent

Alejandro Sifrim

Dusan Popovic



In collaboration with
Center for Human Genetics
University of Leuven
Joris Vermeesch