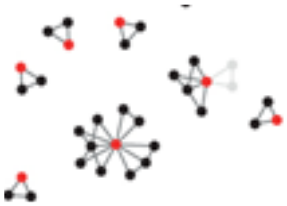
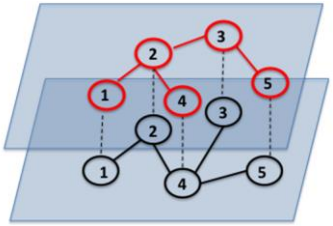
From Statistical to Biological Interactions via Omics Integration



University of Liège

Kyrylo Bessonov

4 July 2016

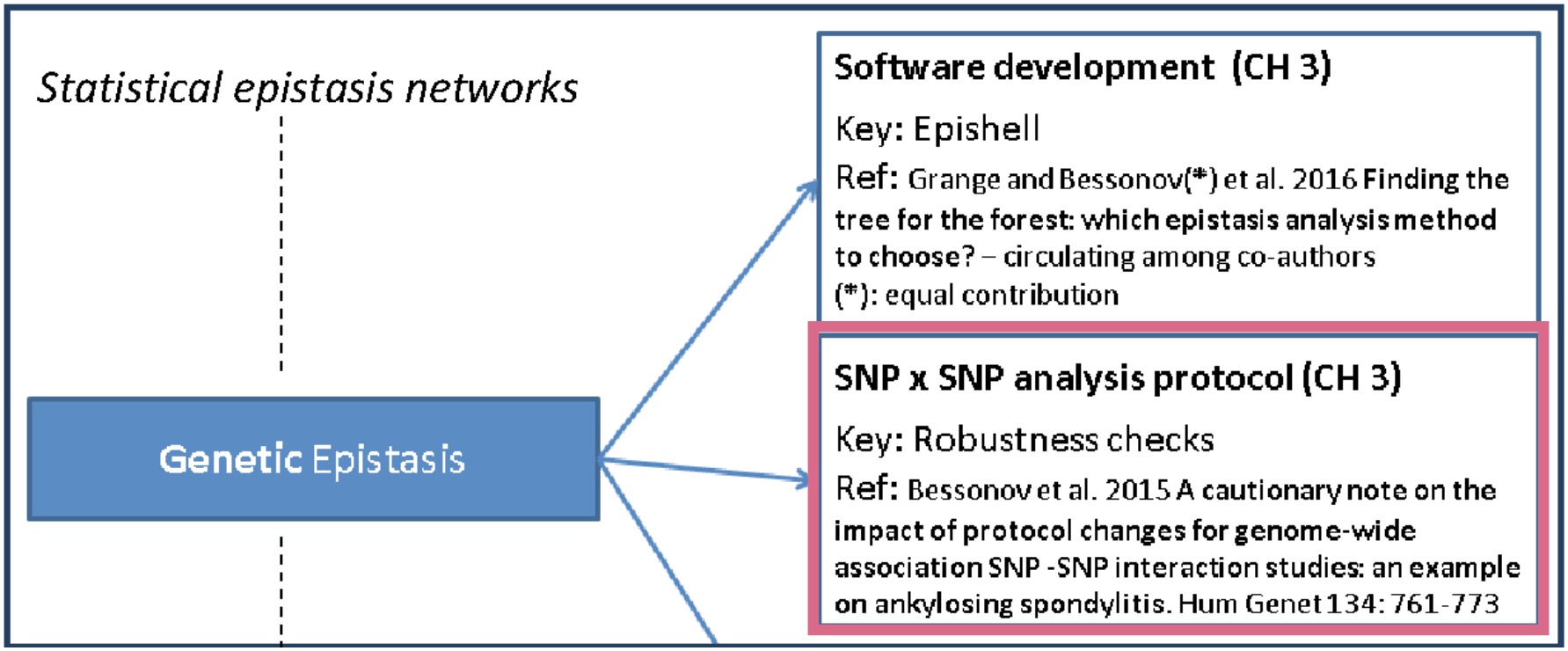


Outline

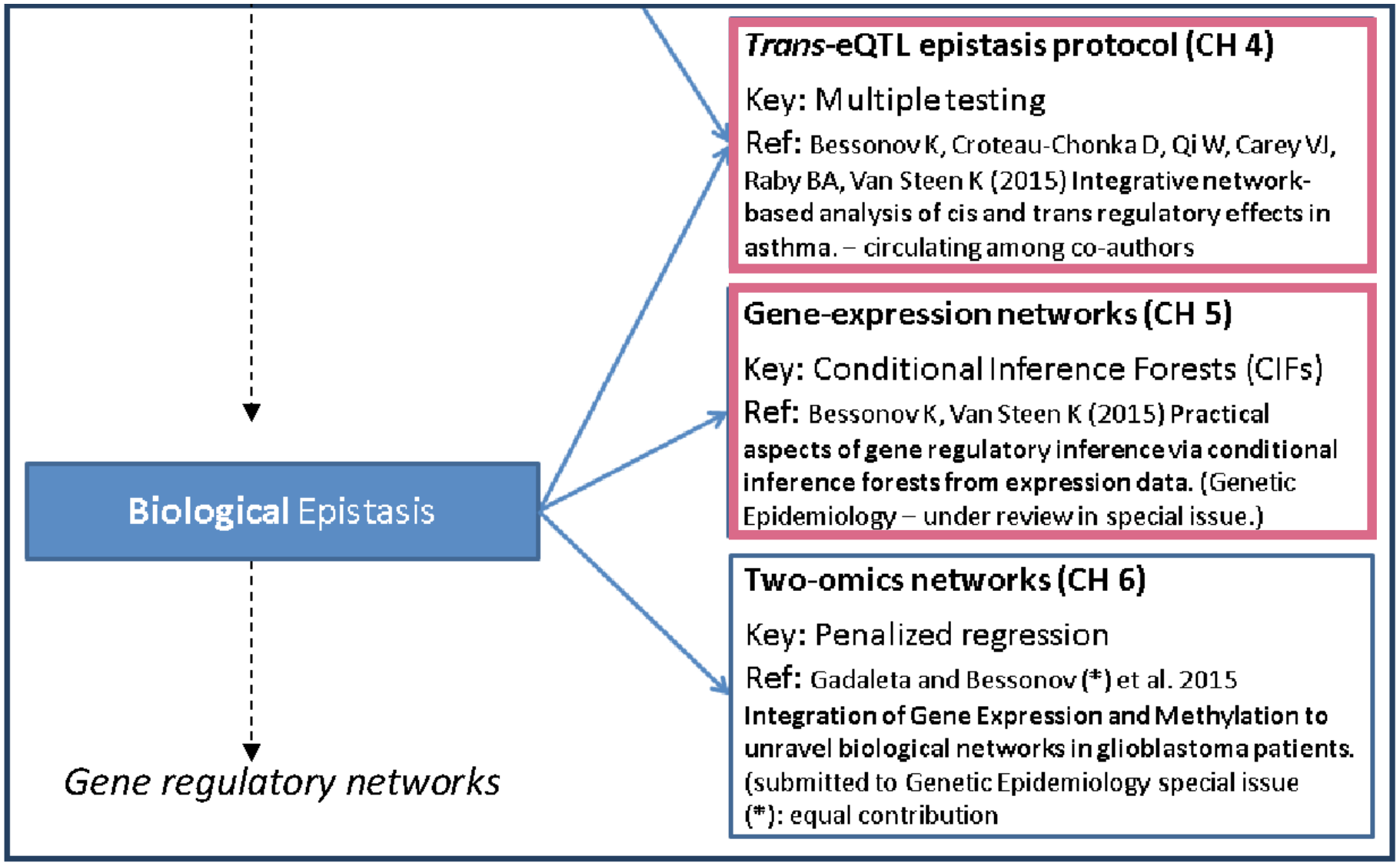
1. Thesis overview
2. Concepts
3. Genome-genome interactions
4. *Trans*-eQTL epistasis protocol
5. Gene expression networks
6. General conclusions
7. Future directions

Thesis Overview

Contributions 1 of 2

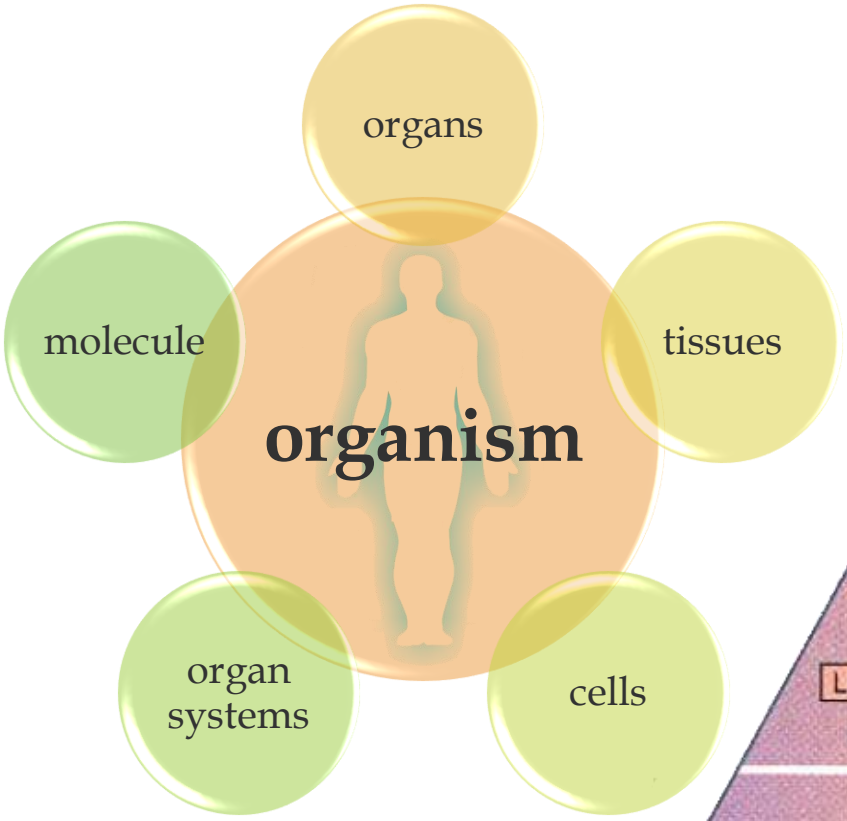


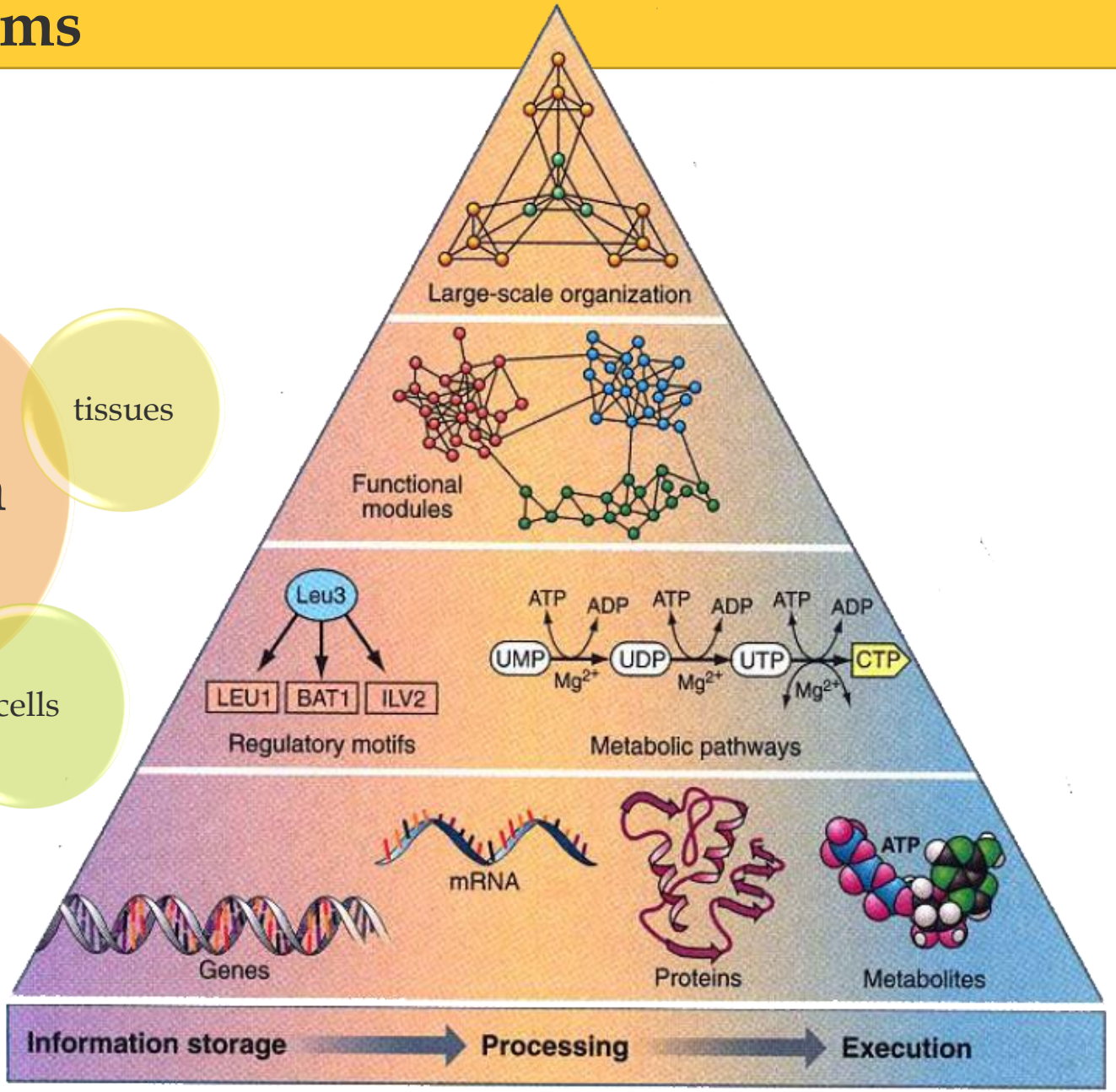
Contributions 2 of 2



Concepts

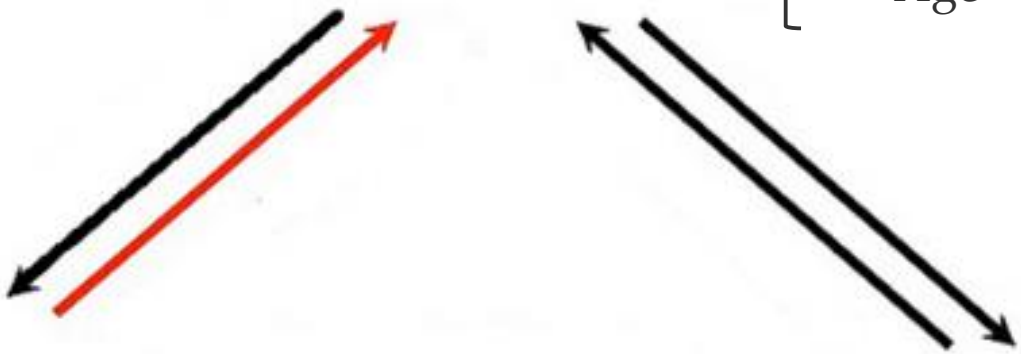
Biological systems





*adapted from Zoltán N., and Barabási. "Life's complexity pyramid."*Science* 298.5594 (2002): 763-764.

Omics data



Environmental (*E*)

- Smoking
- Age

Omics (*O*)

Gen**omics**

Epigen**omics**

Transcript**omics**

Prote**omics**

Metabol**omics**

Phenotypic (*P*)

Phen**omics**

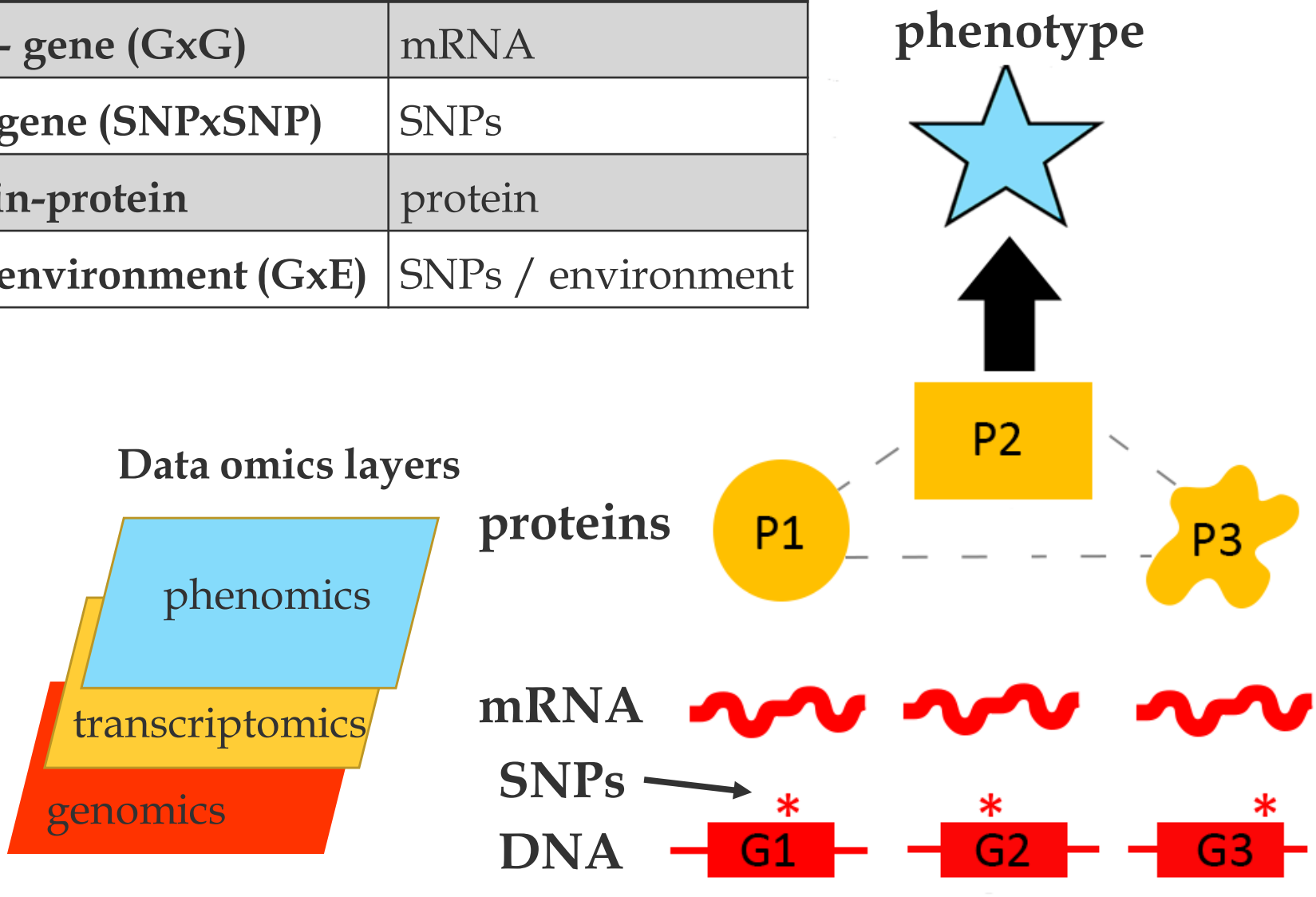
- Cancer
- Diabetes

Genomics: Single Nucleotide Polymorphisms

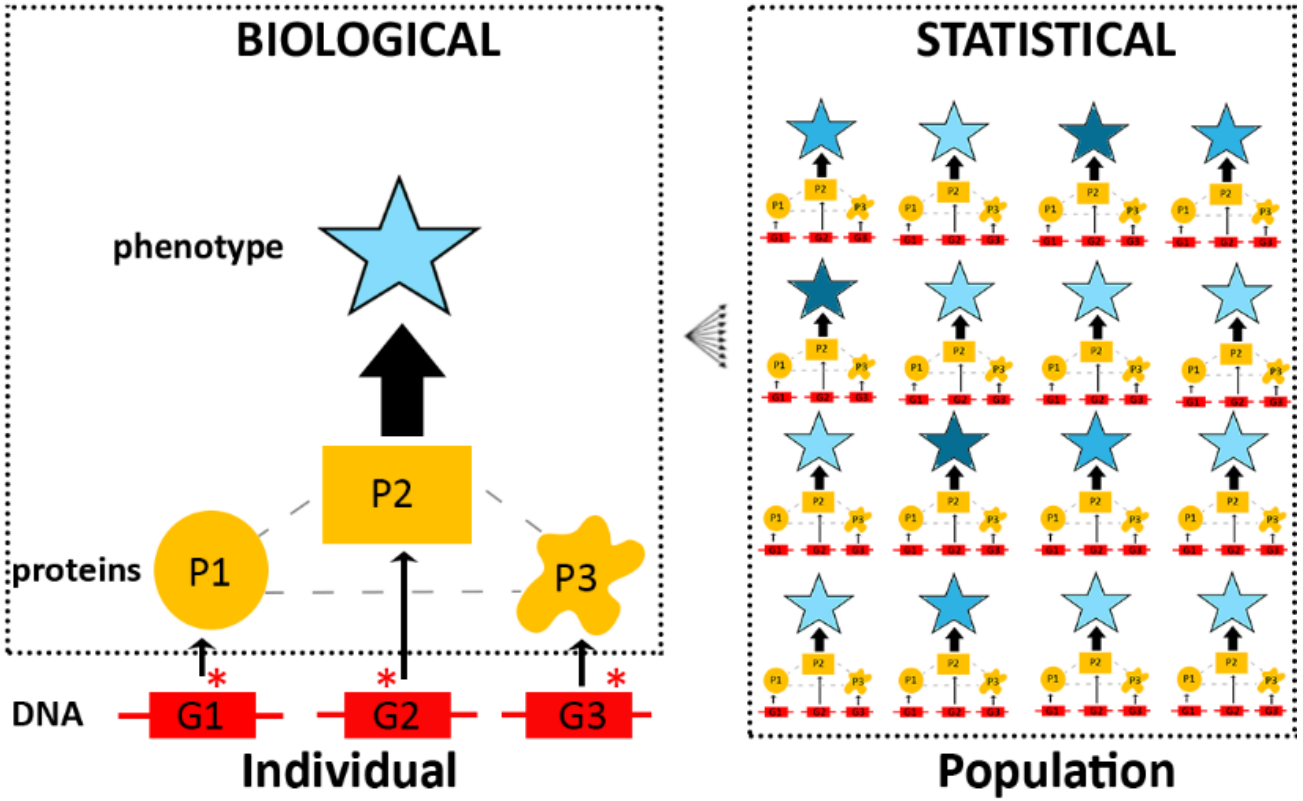
- Locus – physical location in the genome
- SNP - genetic marker
- Phenotype - observable trait

Interactions

Interaction	Source
gene – gene (GxG)	mRNA
gene-gene (SNPxSNP)	SNPs
protein-protein	protein
gene-environment (GxE)	SNPs / environment

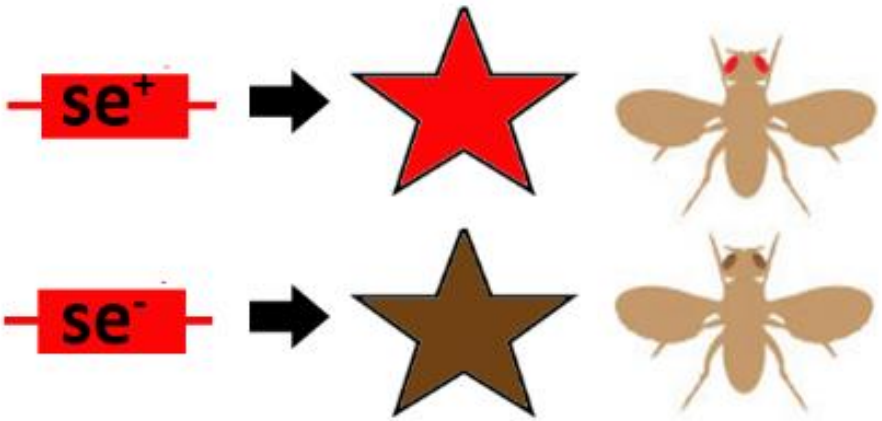


Epistasis

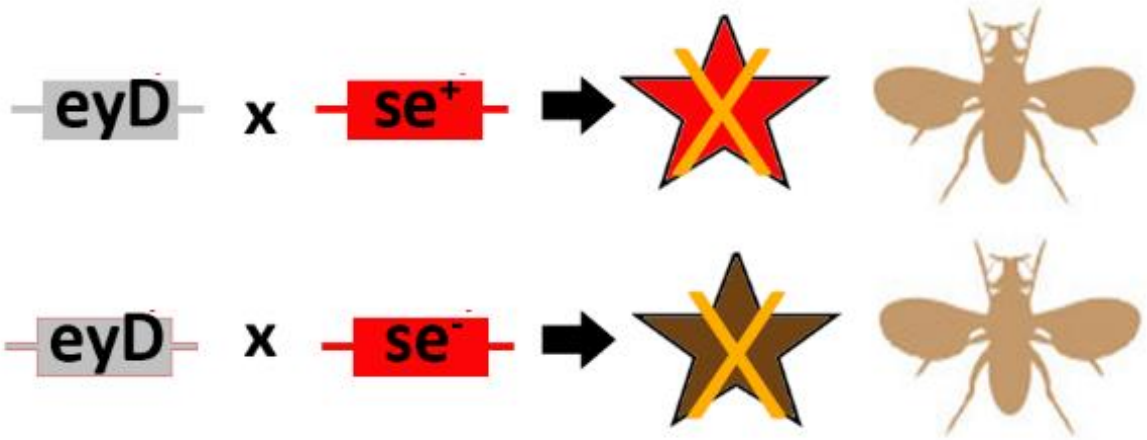


Biological	Statistical
One gene or allele **masking** the phenotypic expression of the other genes or alleles in the interaction.	Departure from a specific **linear model** describing the relationship between predictive factors (here assumed to be alleles at different genetic loci)
~ not necessarily symmetric	~ symmetric in regression framework

Biological epistasis

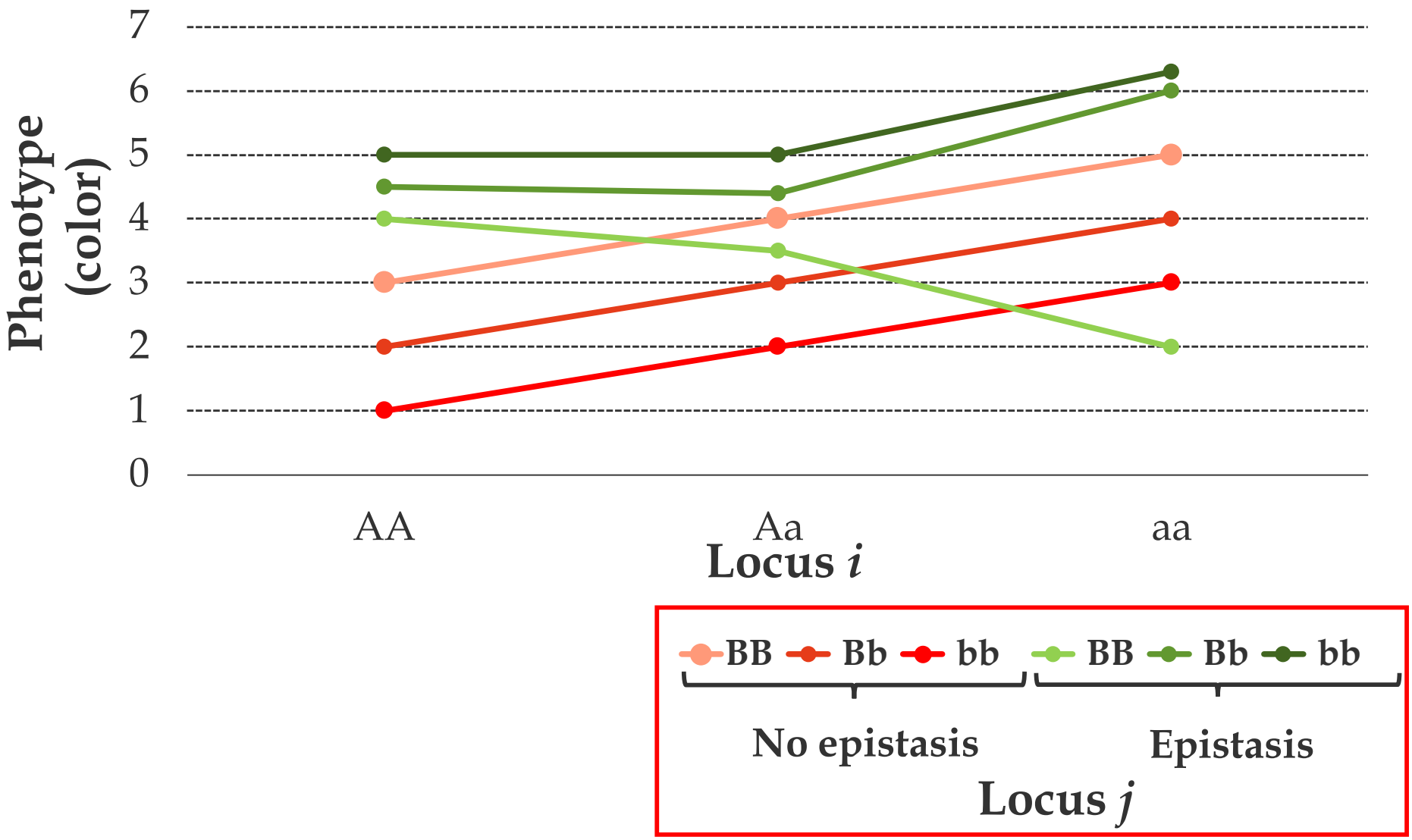


epistasis

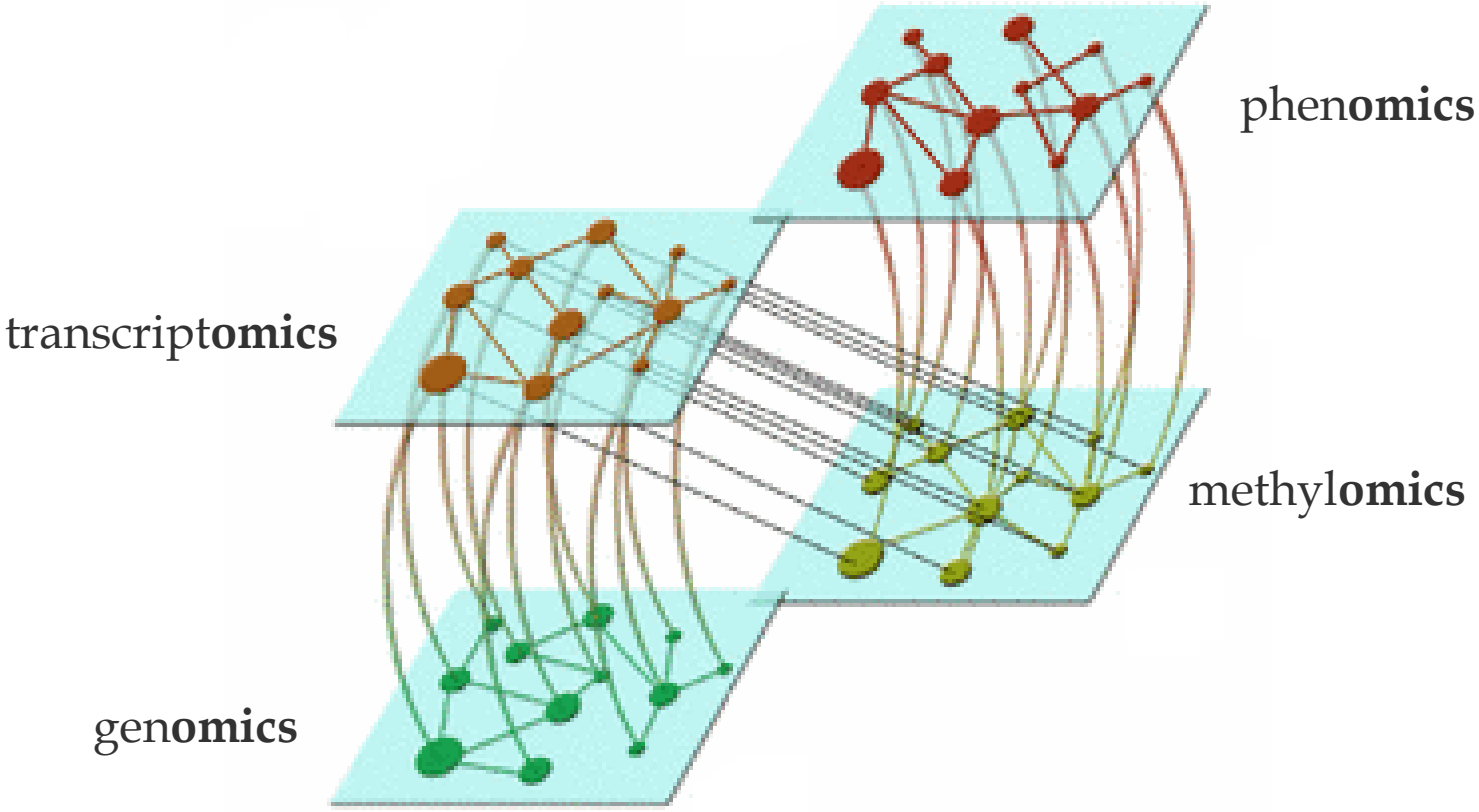


- **se^{+}** red eyes (dominant)
- **se^{-}** brown eyes (recessive)
- **eyD** no eyes (dominant)

Statistical epistasis



Omics integration

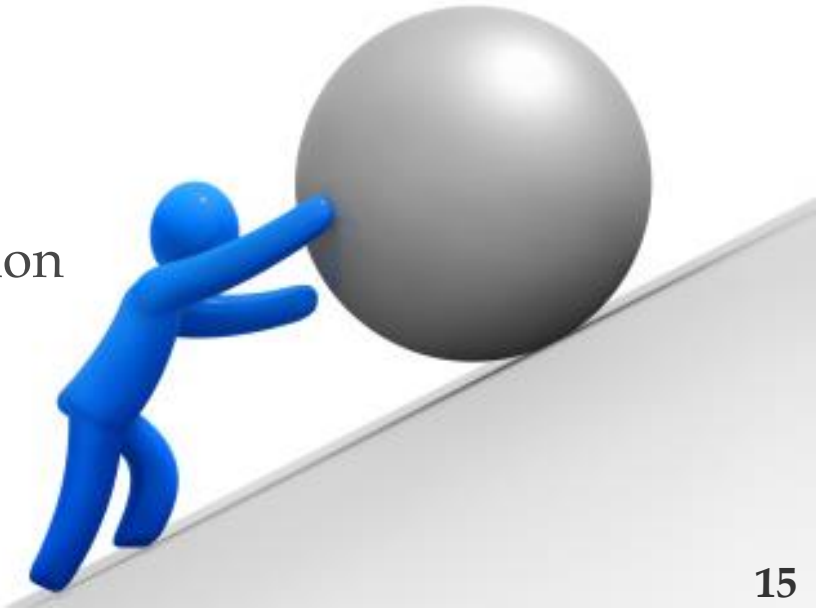


- Single-omics (transcriptomics / transcriptomics)
- Multi-omics (transcriptomics / metabolomics)

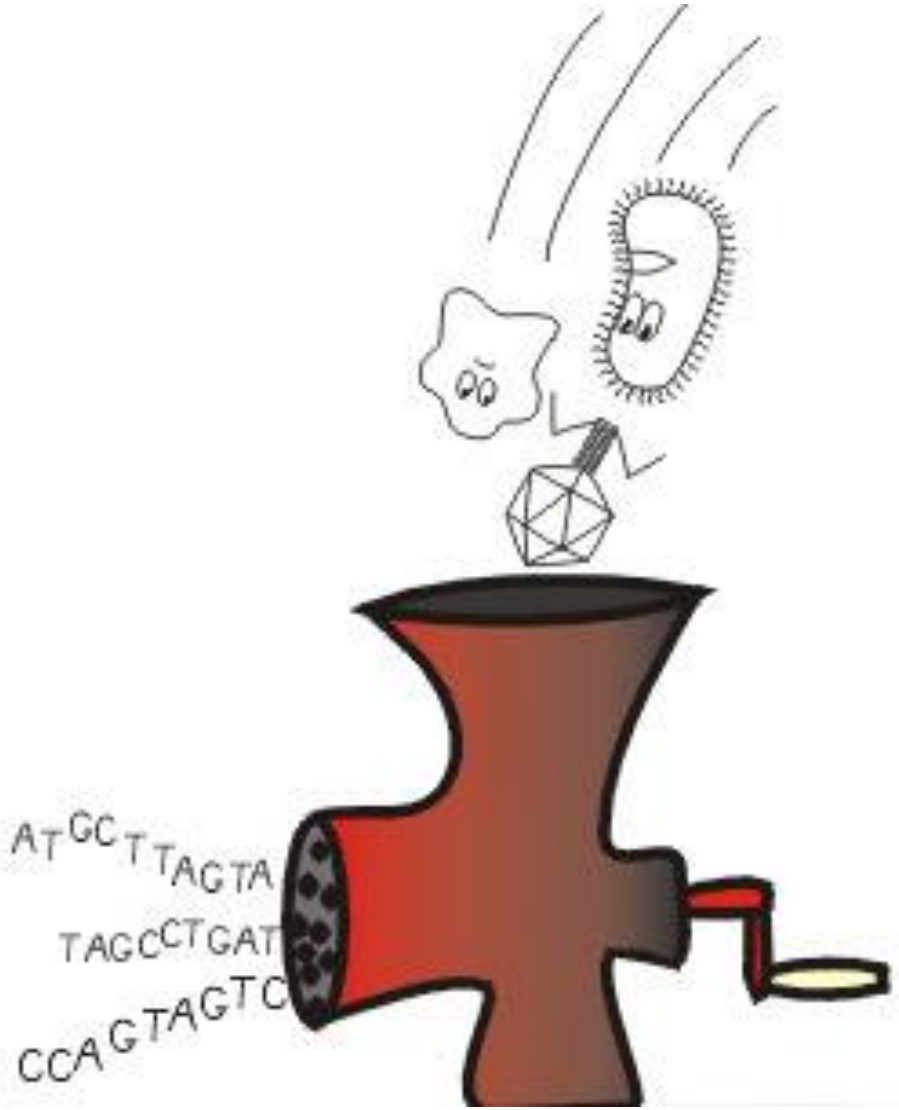
Challenges

- Data
  - Storage
  - Accessibility
  - Standardization
- Analysis
  - "Curse of dimensionality

    Large *p*, small *n* problem
  - Systems view in omics integration

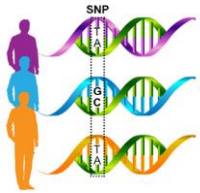


Nowadays ...

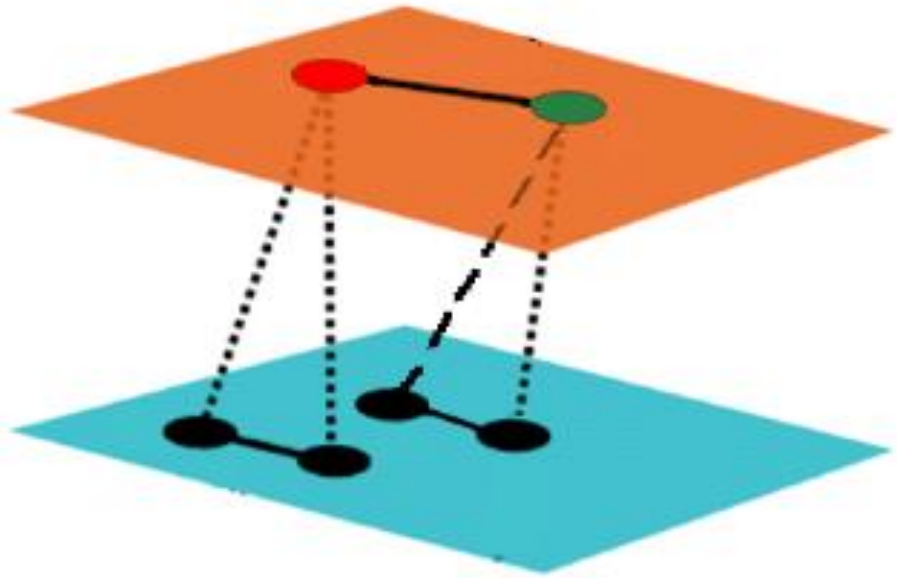


Genome-genome interactions

The impact of protocol changes for genome-wide association SNP x SNP interaction



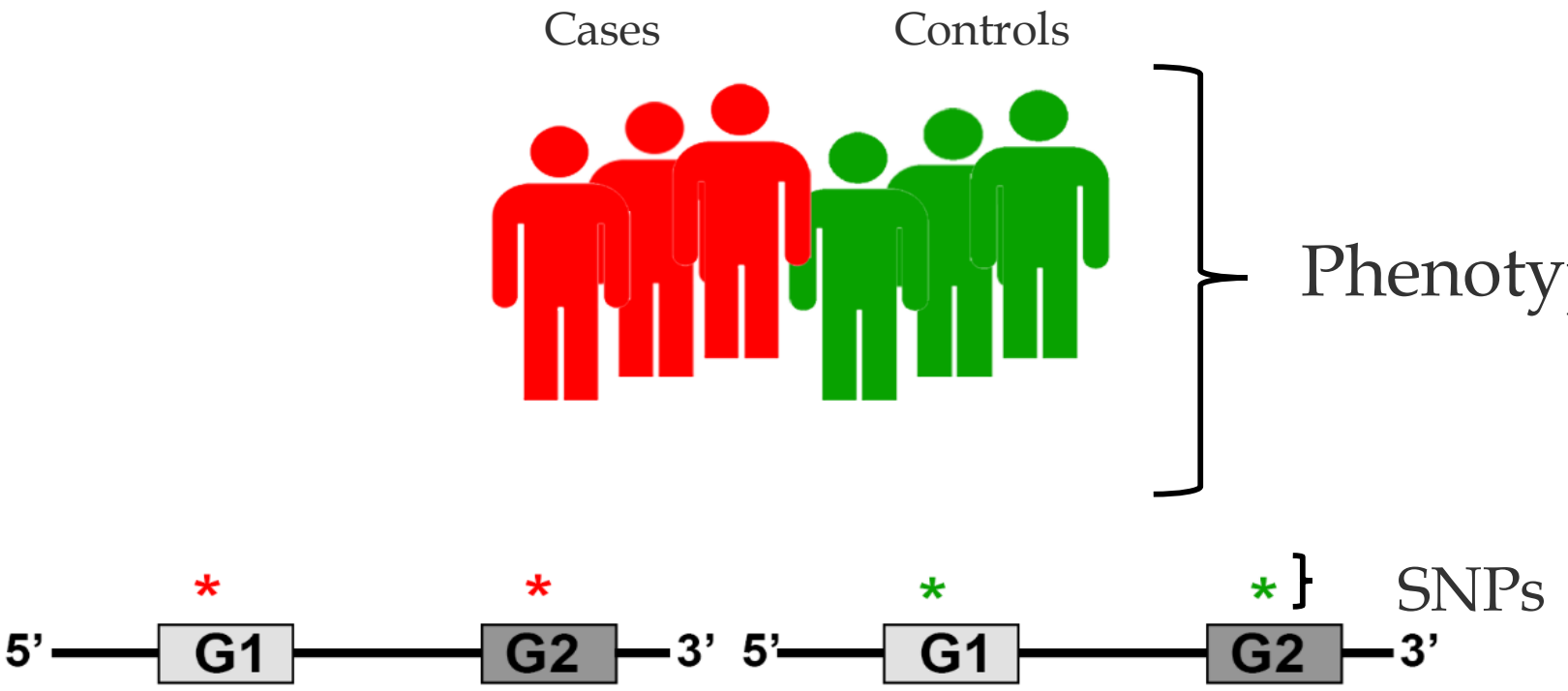
Context: genome - phenome



Phenotypic data layer
(case / control)

Genotypic data layer
(SNP)

Context: genome – phenome interactions

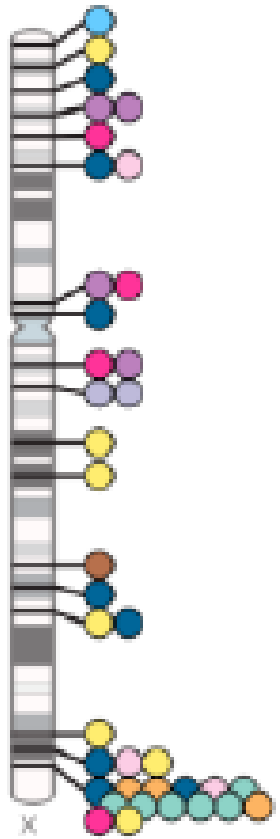


- Which pair of markers affects phenotype?
  - Predictors - SNPs
  - Trait – phenotype
- Linkage disequilibrium (LD)
  - Association between alleles

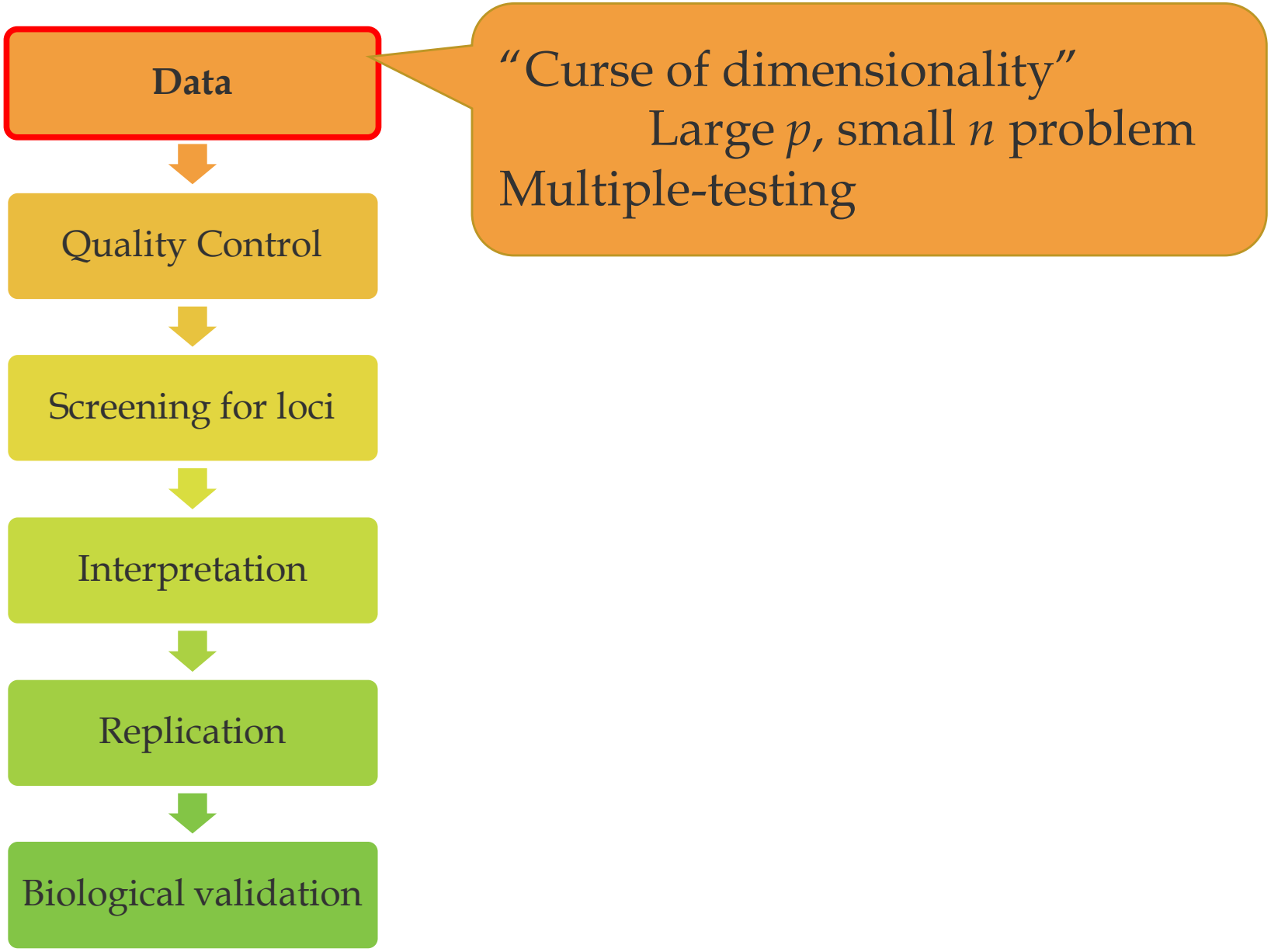
Context: GWAI

- Genome-wide association interaction studies (GWAI)
- Goal
  - Gene – gene interactions
  - Assumes large number of individuals
- Linear regression model

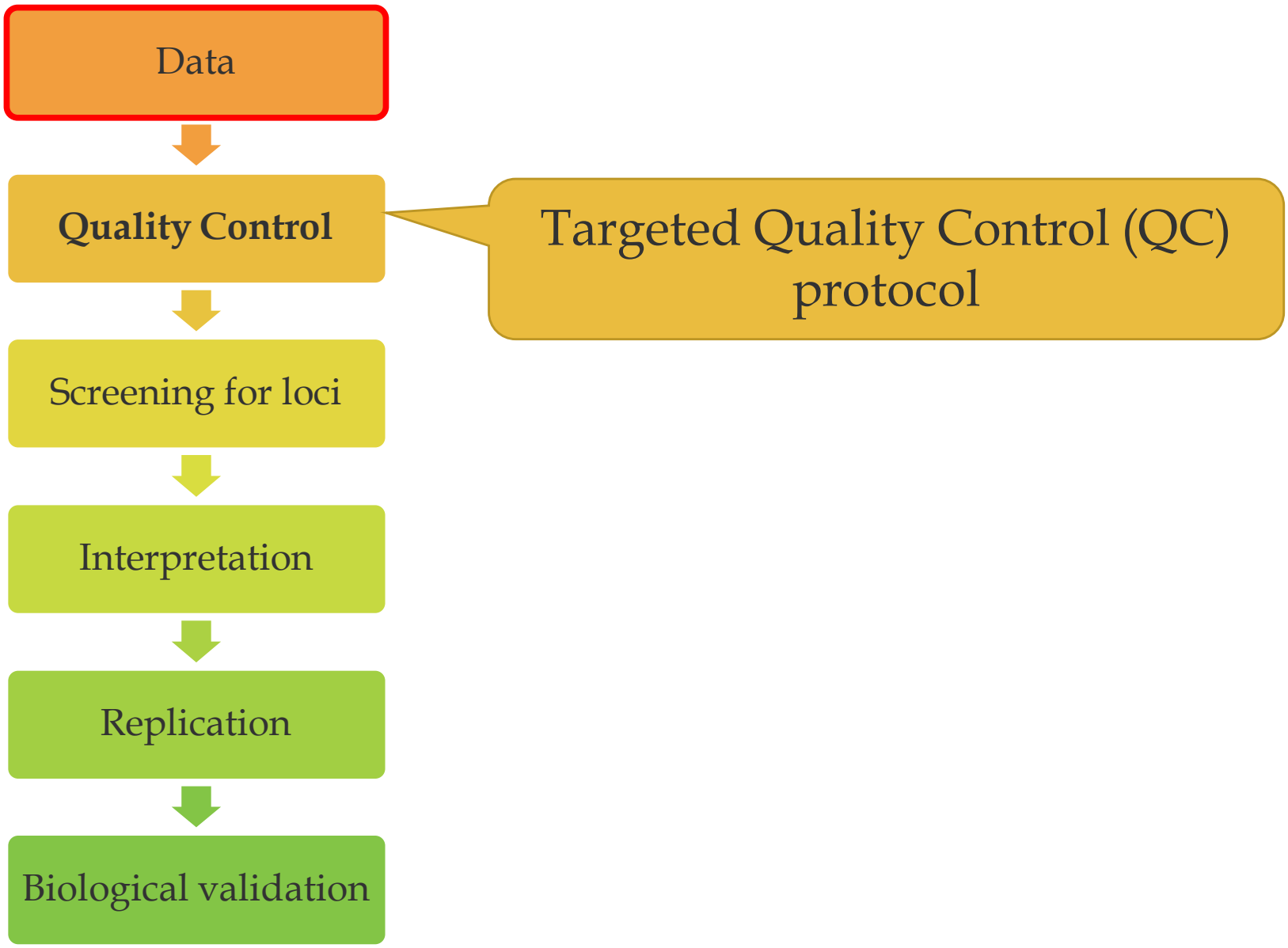
$$Y_{trait} = \beta_o + \beta_1 X_{locus\ i} + \beta_2 X_{locus\ j} + \boldsymbol{\beta_3 X_{locus\ i} * X_{locus\ j}} + \varepsilon_i$$



Strategy: GWAI protocol



Strategy: GWAI protocol



Strategy: GWAI protocol

Data

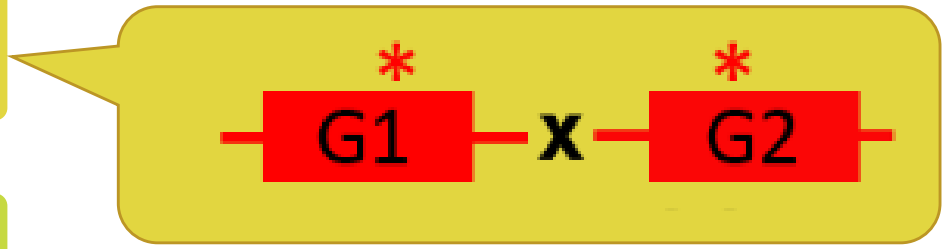
Quality Control

Screening for loci

Interpretation

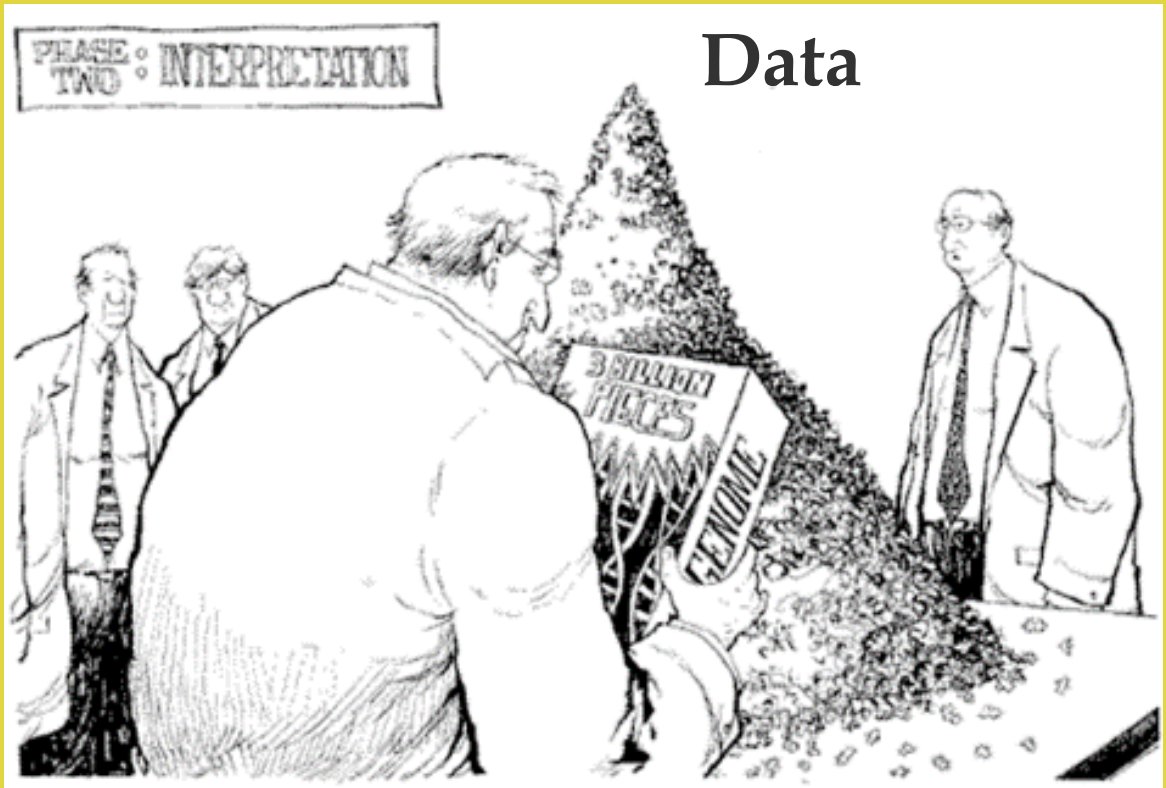
Replication

Biological validation

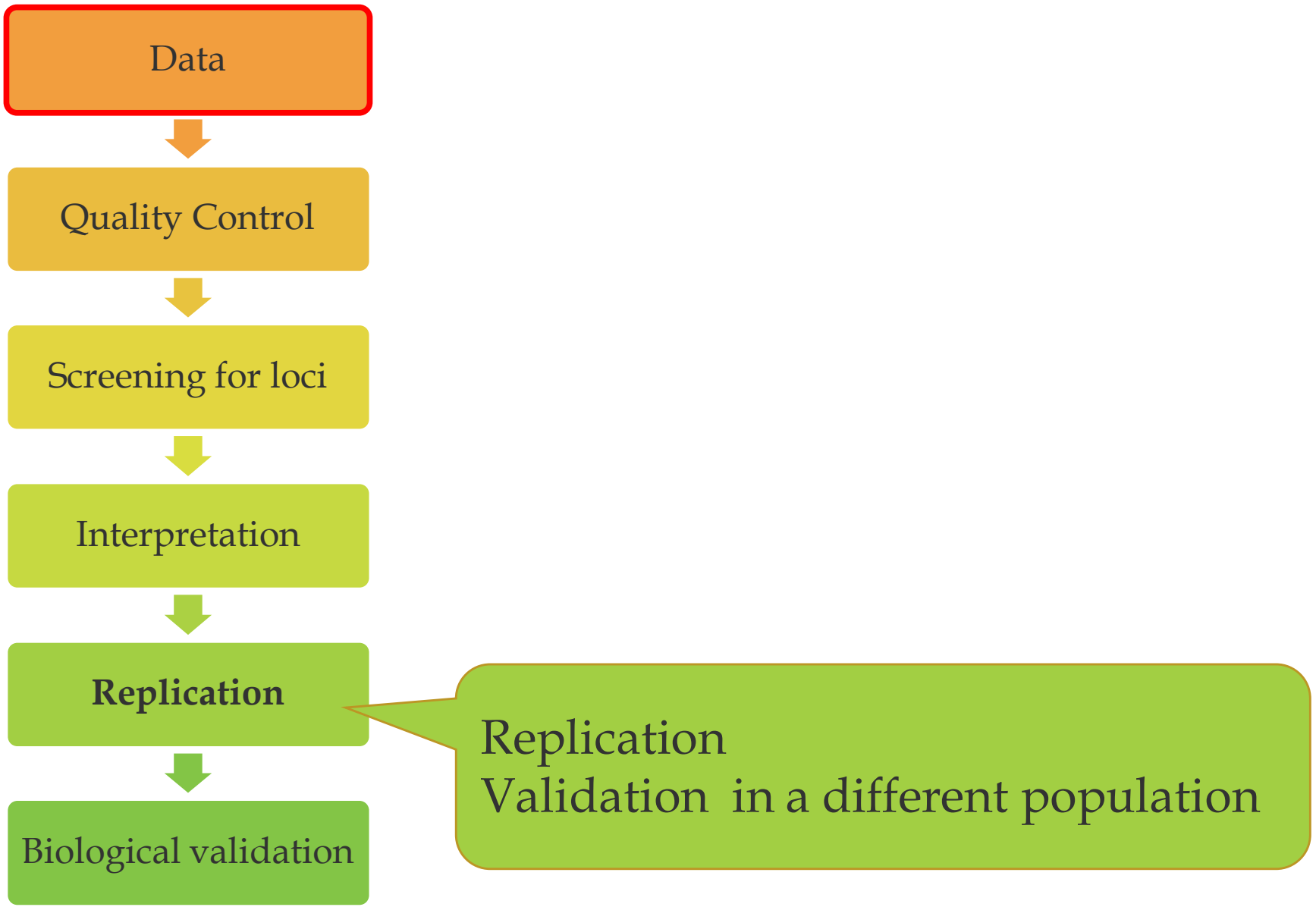


Strategy: GWAI protocol

- Data
- Quality Control
- Screening for loci
- **Interpretation**
- Replication
- Biological validation



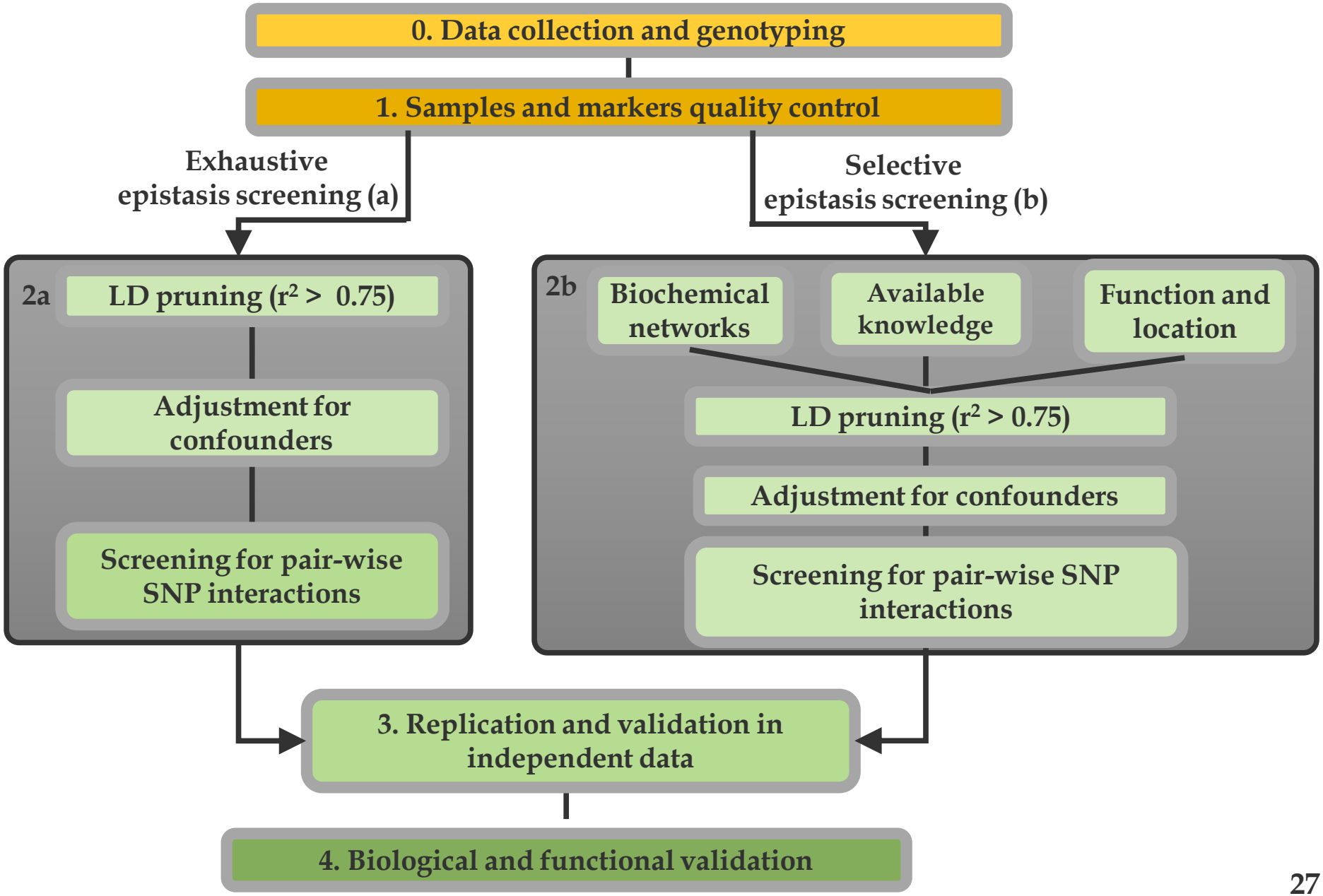
Strategy: GWAI protocol



Strategy: GWAI protocol

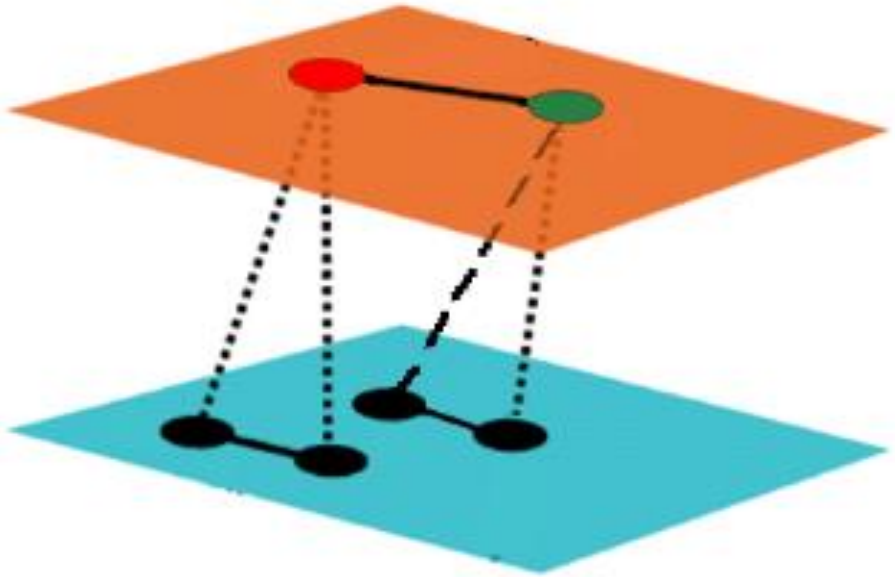
- Data
- Quality Control
- Screening for loci
- Interpretation
- Replication
- **Biological validation**

Strategy: GWAI protocol[4]



Problem

- No standard GWAI protocol exists
  - Choice of parameters
    - Dataset
    - Encoding
      - Additive
      - Co-dominant
    - LD pruning



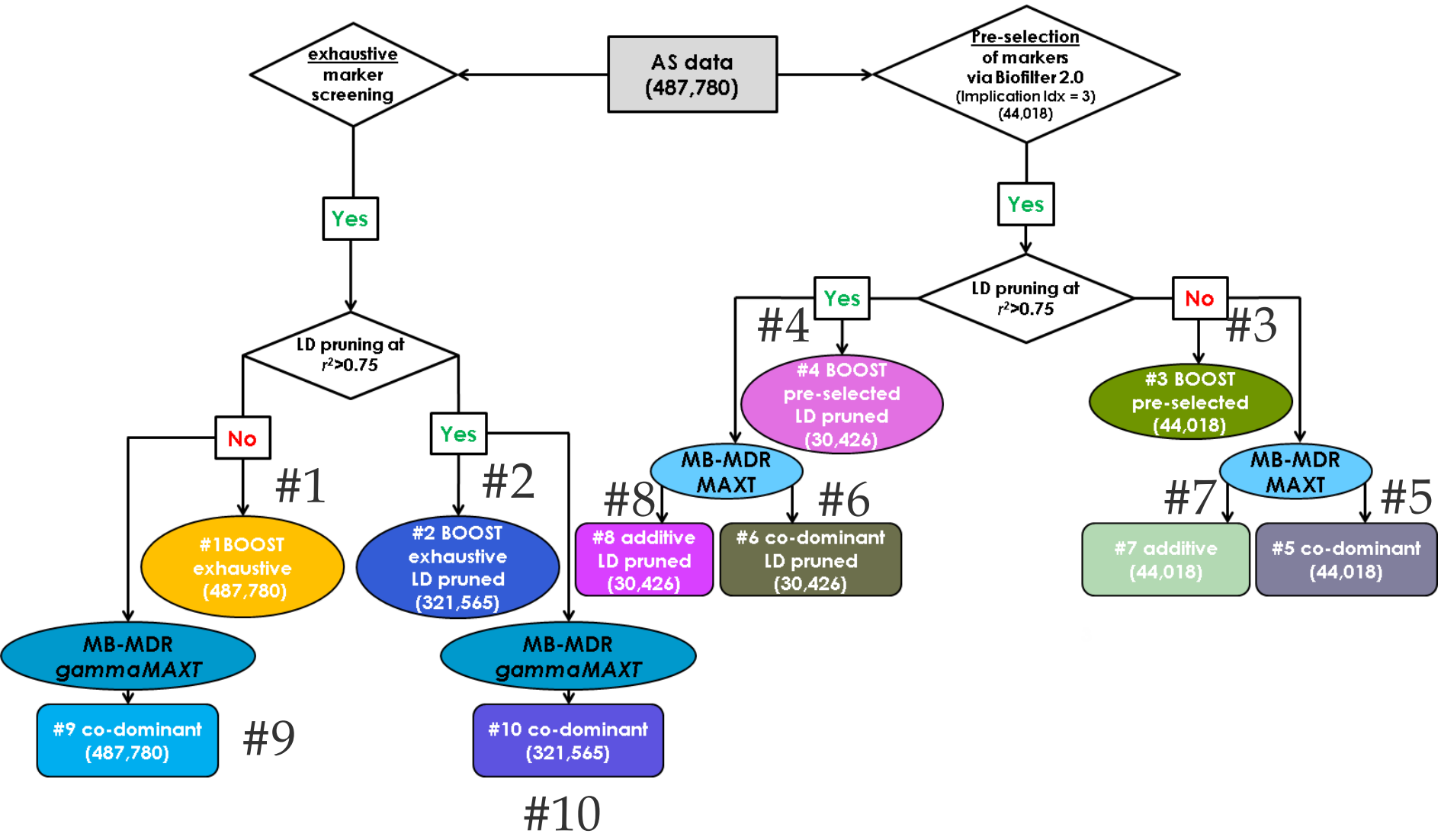
- Impact on the final epistasis findings

Application: Ankylosing spondylitis data

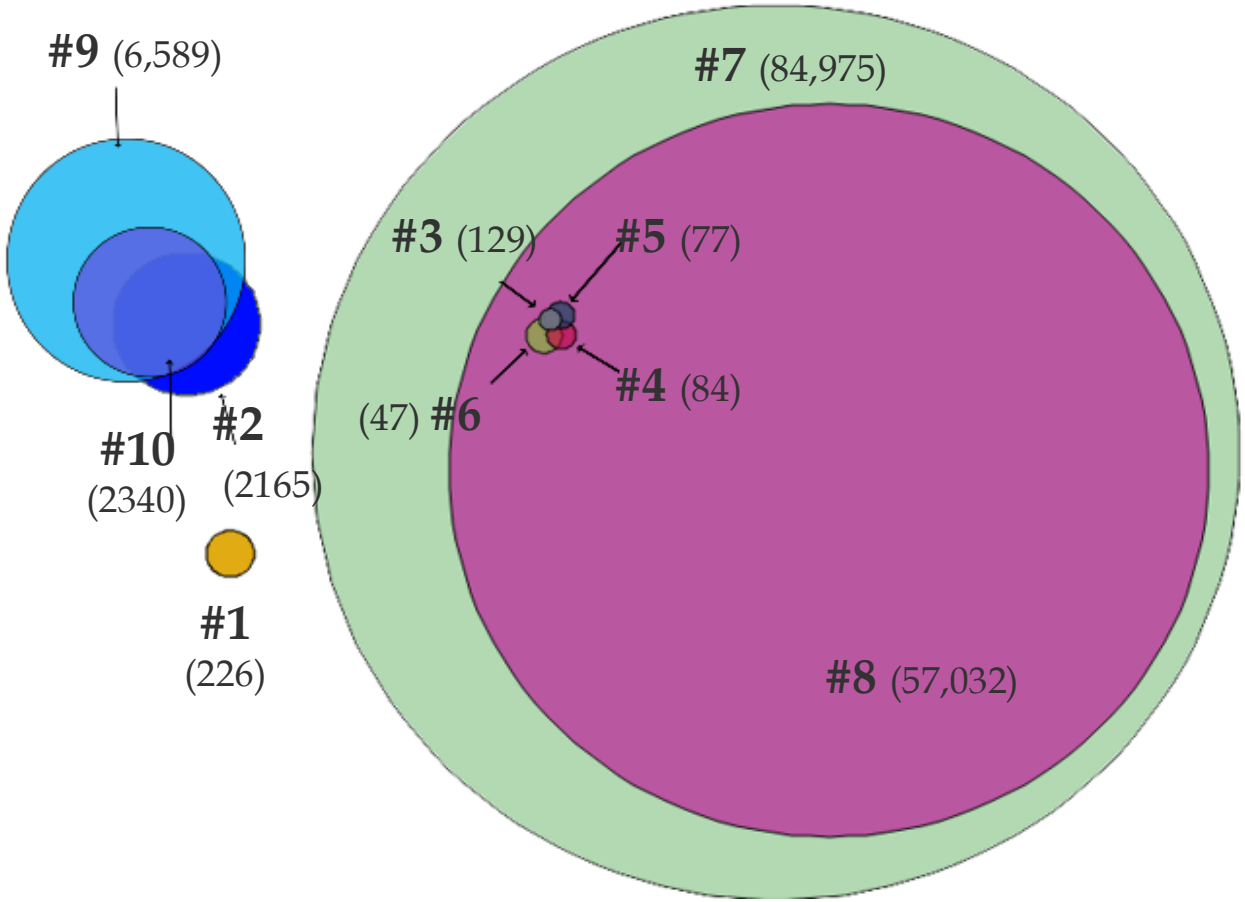
- Cases*
  - 2005 ankylosing spondylitis (AS)
- Controls*
  - 3000 British 1958 Birth Cohort (BC)
  - 3000 National Blood Donors (NBS)
- Source
  - Wellcome Trust Case Control Consortium (WTCCC2)

* European ancestry

Application: 8+2 GWAI protocols



Application: results overlap



(significant SNP pairs)

Application: percent overlap

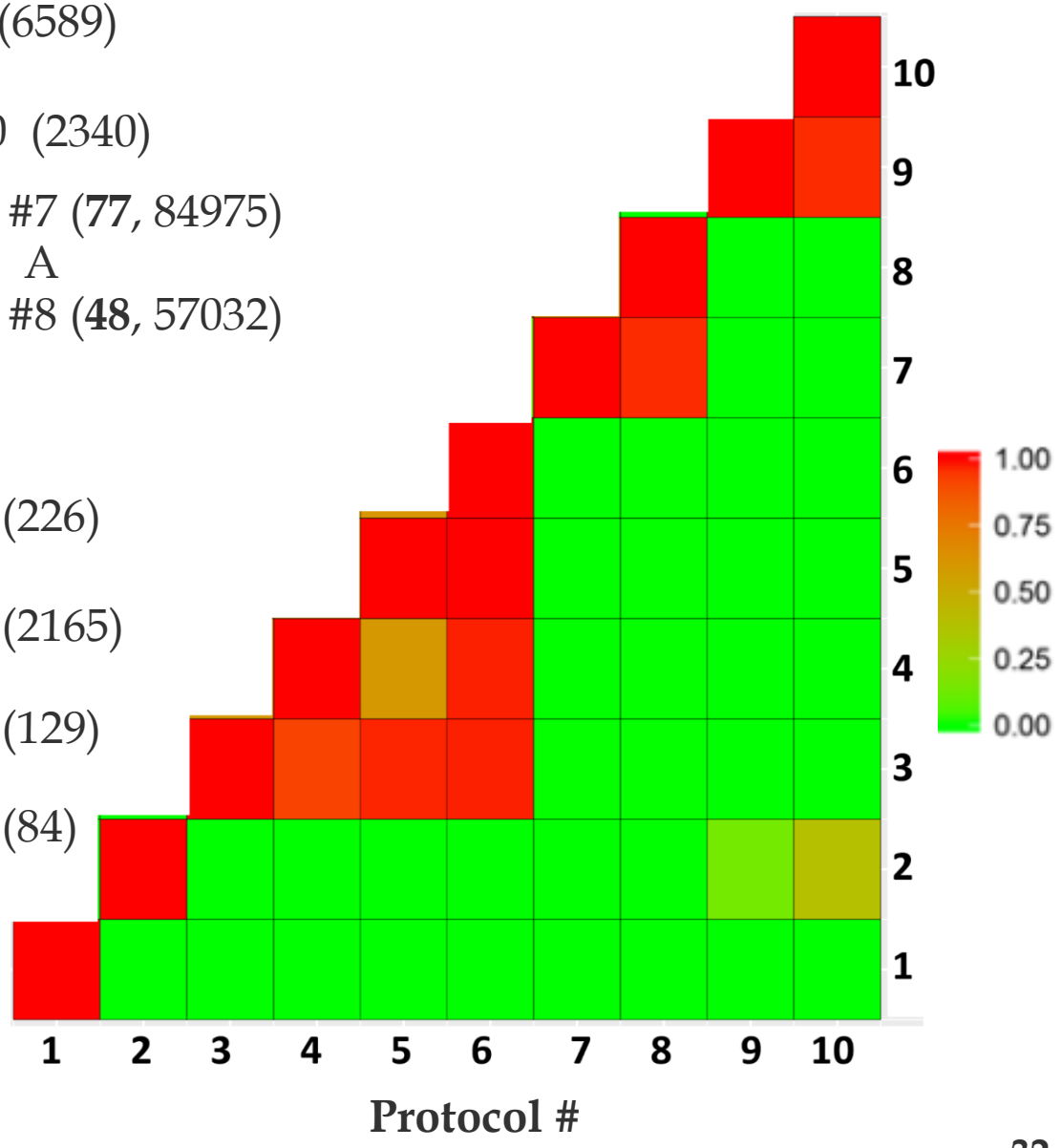
MB-MDR

- Exhaustive
  - LD up → #9 (6589)
  - LD down → #10 (2340)
- Biofilter
  - LD up → **#5**, #7 (**77**, 84975)
  - LD down → **#6**, #8 (**48**, 57032)

C A

BOOST

- Exhaustive
  - LD up → #1 (226)
  - LD down → #2 (2165)
- Biofilter
  - LD up → #3 (129)
  - LD down → #4 (84)

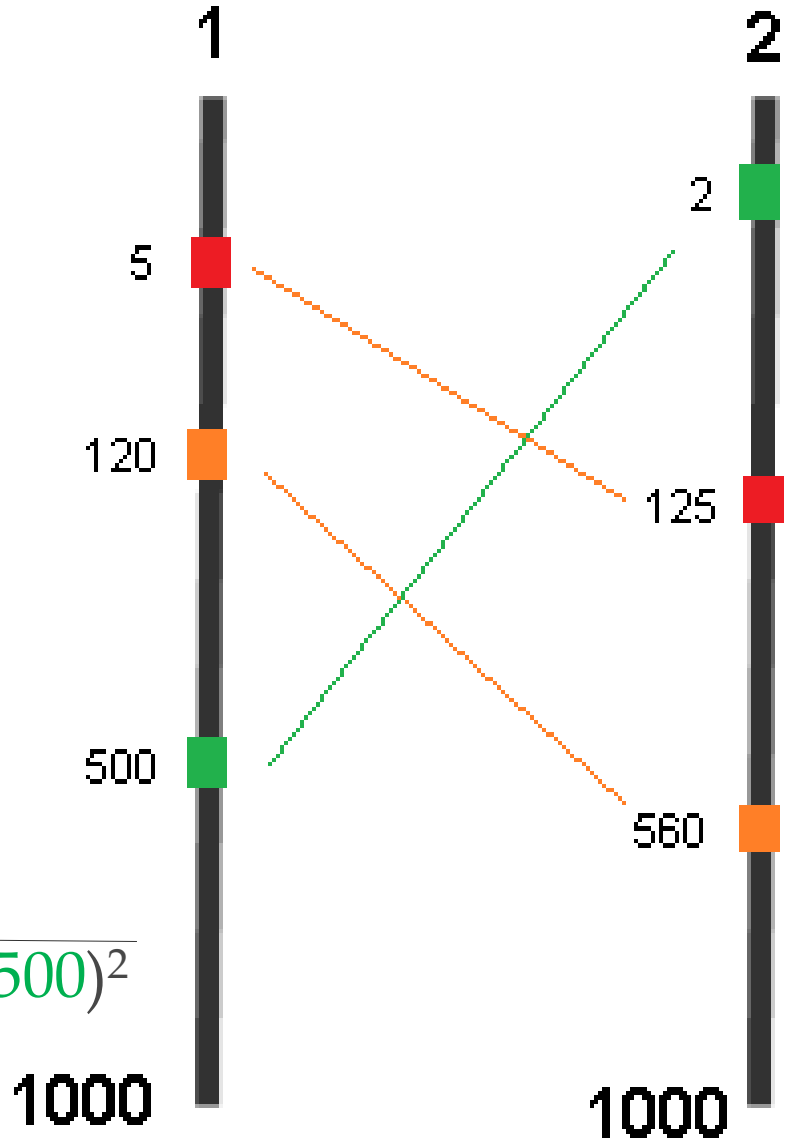


Application: distance

- Sort results
  - Highest to lowest significance
- 207 common SNPs
  - All protocols
  - Significant and Non-significant
- Get rank values`
- Calculate Euclidian distance

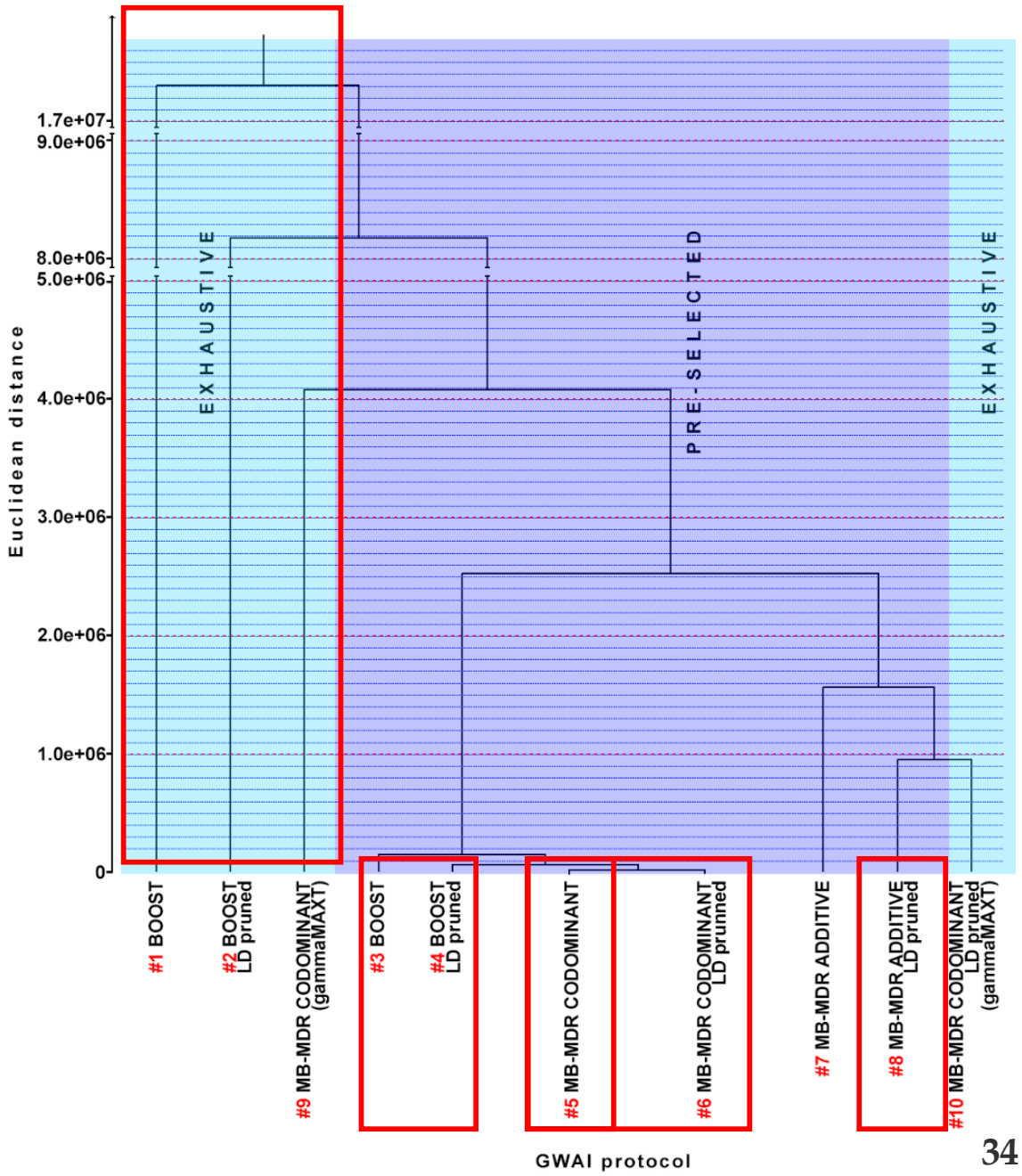
$$D(1,2)=\sqrt{(5-125)^2+(5-500)^2+(120-500)^2}$$

$$D(1,2)= 675.28$$



Application: clustering of protocols

- 207 ranks
  - Common SNPs pairs
  - Not all significant
- Marker selection
  - Protocols
    - #1-#2
    - #3-#8
- Encoding
  - Protocols #5-#8
- LD pruning
  - Protocols
    - #3 and #4
    - #5 and #6

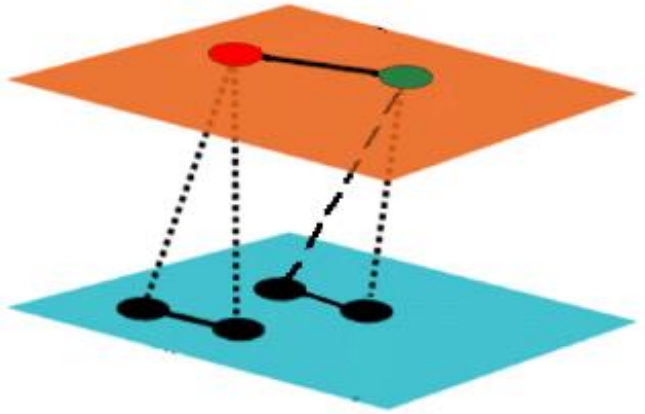


Application: biological relevance

GO ID	GO Term Description	p-value*
GO:0007411	**axon guidance**	**7.9E-77**
GO:0030168	platelet activation	3.9E-58
GO:0055085	transmembrane transport	3.0E-50
GO:0007268	**synaptic transmission**	**2.0E-36**

* Fisher's method (combined topGO *p*-values from 10 protocols)

Conclusions



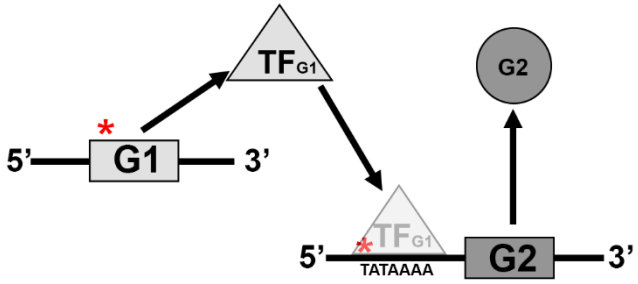
- 10 GWAI protocols
  - Dramatic changes
  - Key factors
    - Input markers
    - Tool selection
    - Encoding of lower order effects (additive / co-dominant)
    - LD pruning
- Impact strength

Higher → Lower

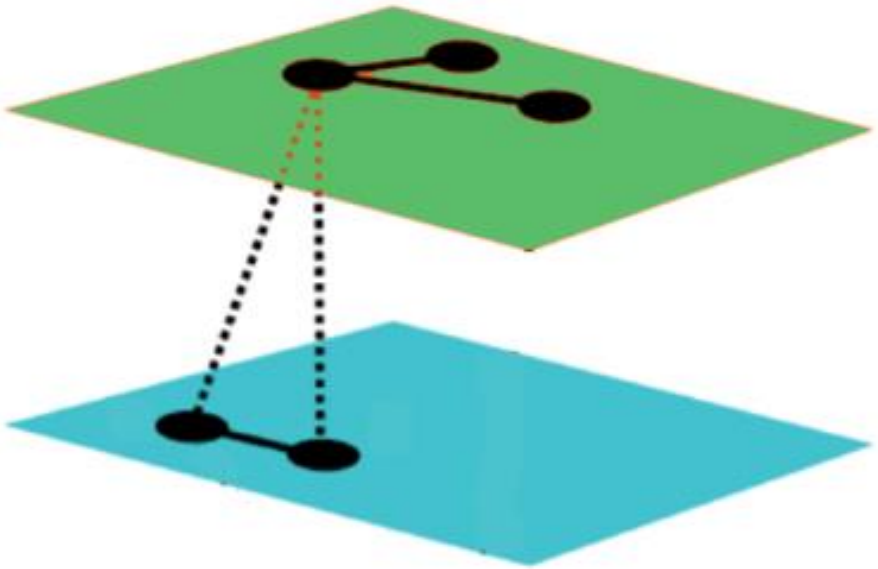
Dataset > Encoding > LD pruning

Trans-eQTL epistasis protocol

Integrative network-based analysis of cis and trans regulatory effects in asthma



Context: genome - transcriptome



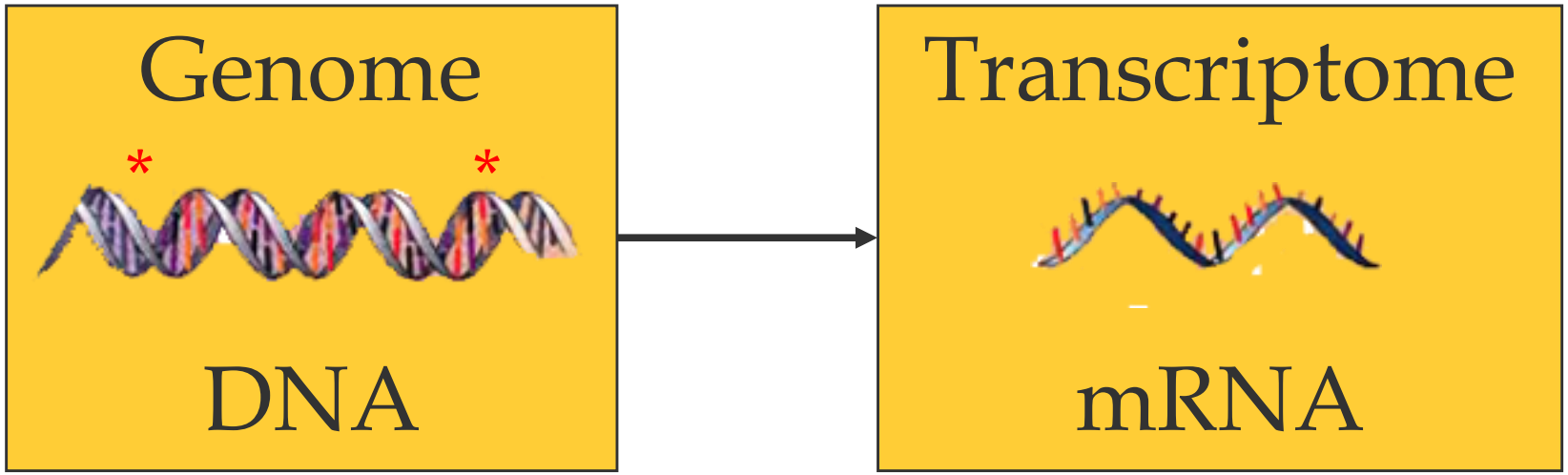
Transcriptomic data layer

Genomic data layer

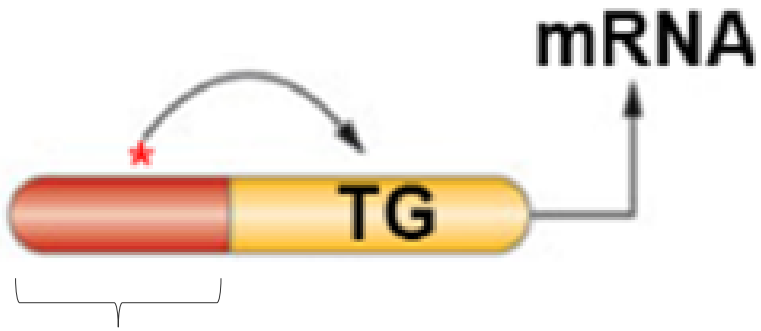
- Trait – expression (microarrays / RNA*seq*)
- Predictors – genotypic data (SNP arrays)

Context: problem

- Identification of genome – transcriptome interactions
- Avoid statistical artifacts
- Build epistatic statistical model (network)



Context: *Cis* eQTL



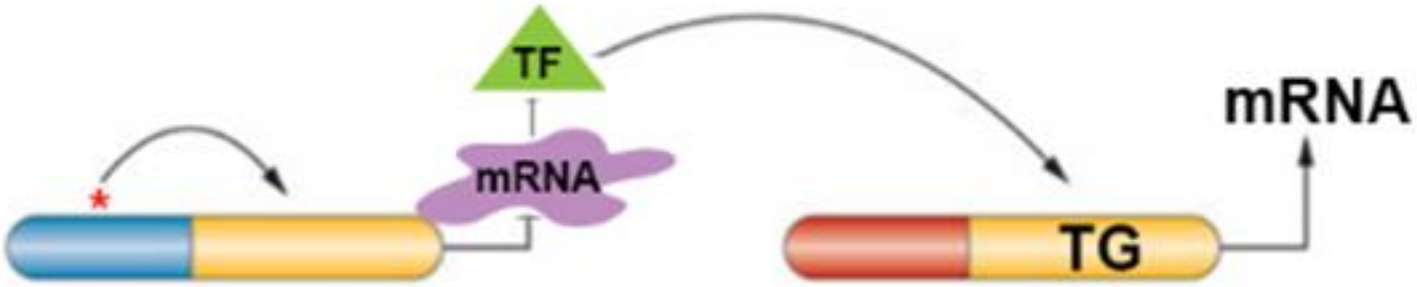
'neighborhood'

* - SNP / locus

- **TG** – target gene
- **[mRNA]** amount of expressed mRNA

- Expression quantitative trait loci (eQTL)
- Marker in the TG 'neighborhood'

Context: *Trans* eQTL

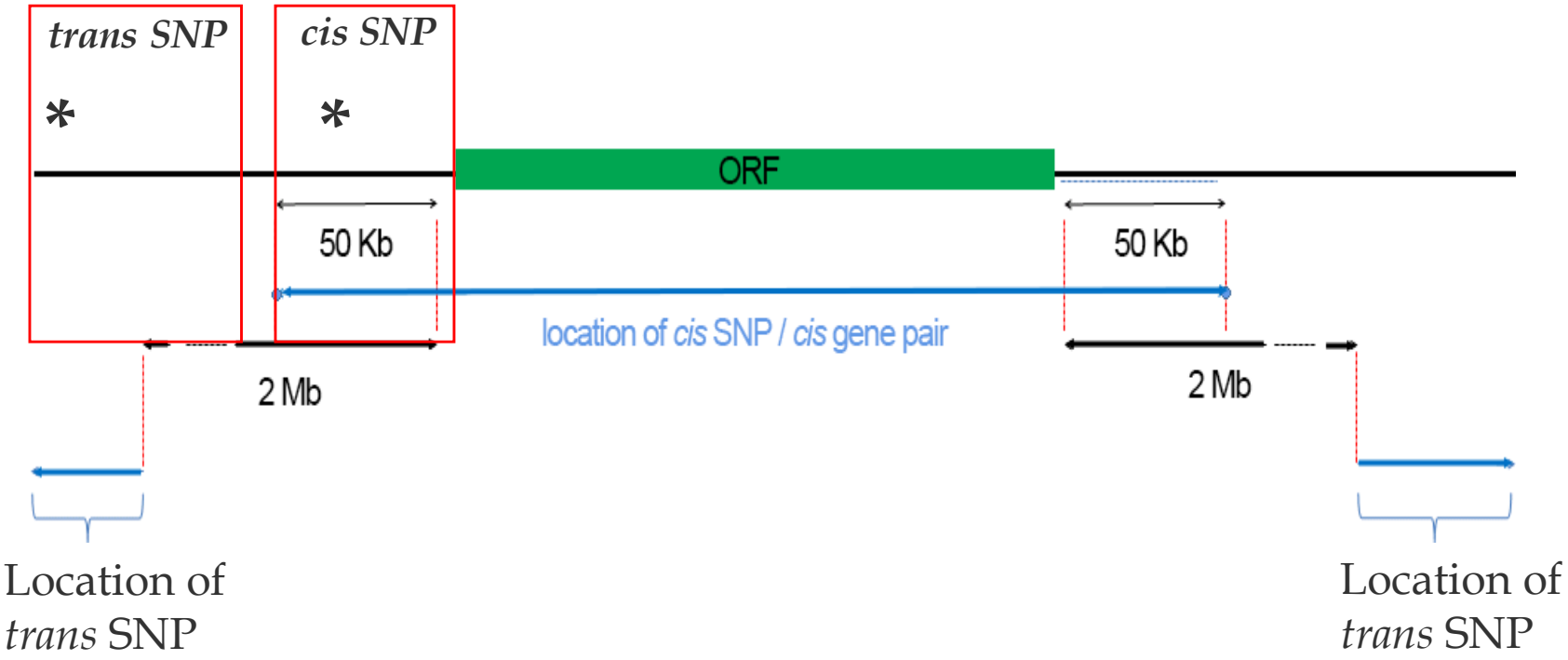


- Distant marker affecting a TG
- TF – transcription factor

Context: epistatic *trans/cis* eQTL

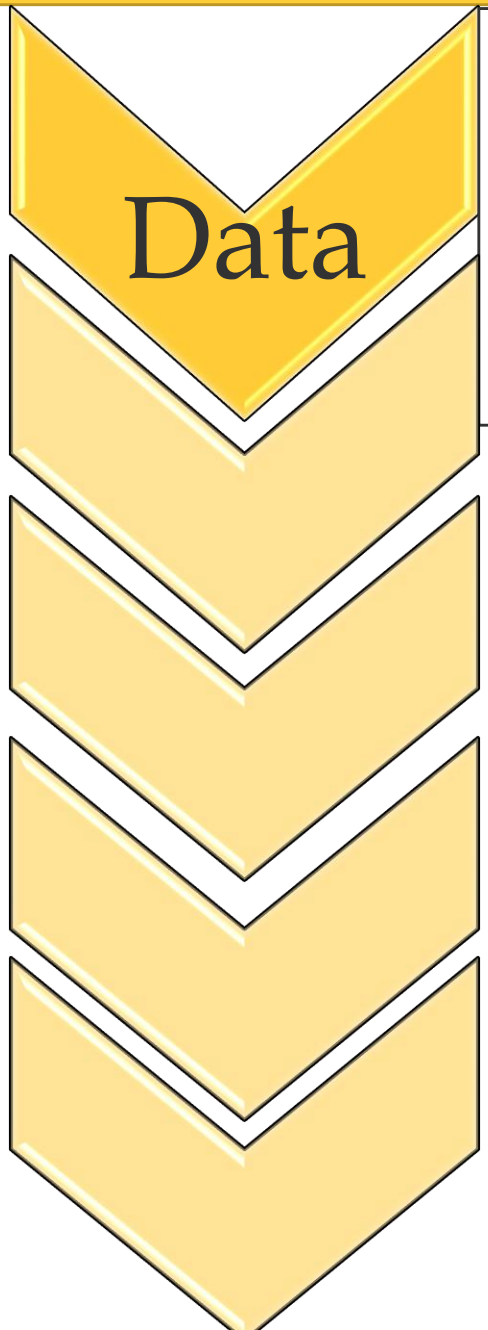
- Interaction
  - Between *trans* and *cis* loci
  - *Trans* locus modifies effect of *cis* locus on the TG
    - SNP_{trans} x SNP_{cis} → [TG]

Context: physical *trans/cis* loci mapping



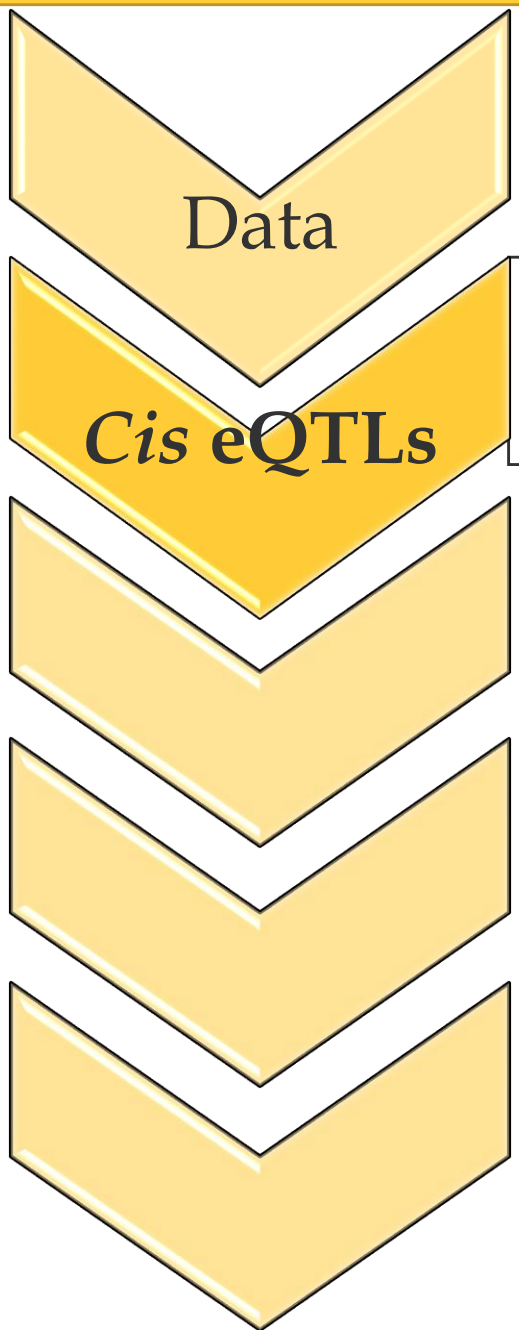
- ORF – open reading frame
  - Codes a gene product (introns + exons)

Strategy: *trans-eQTL epistasis protocol*



- Childhood Asthma management program (CAMP)
- 177 asthmatics (smokers / non-smokers)
- Expression - microarrays
- Genotypic - SNP arrays

Strategy: *trans-eQTL epistasis protocol*



- Generalized Least Squares (GLS)
- 1763 *cis* eQTLs

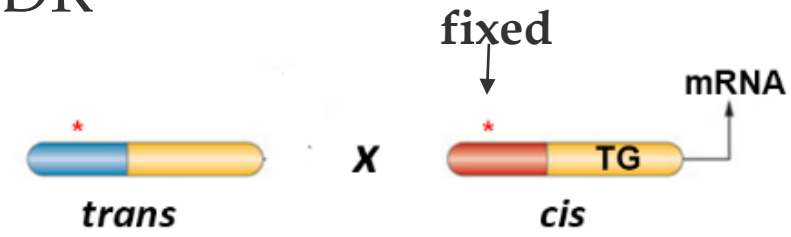
Strategy: *trans-eQTL epistasis protocol*

Data

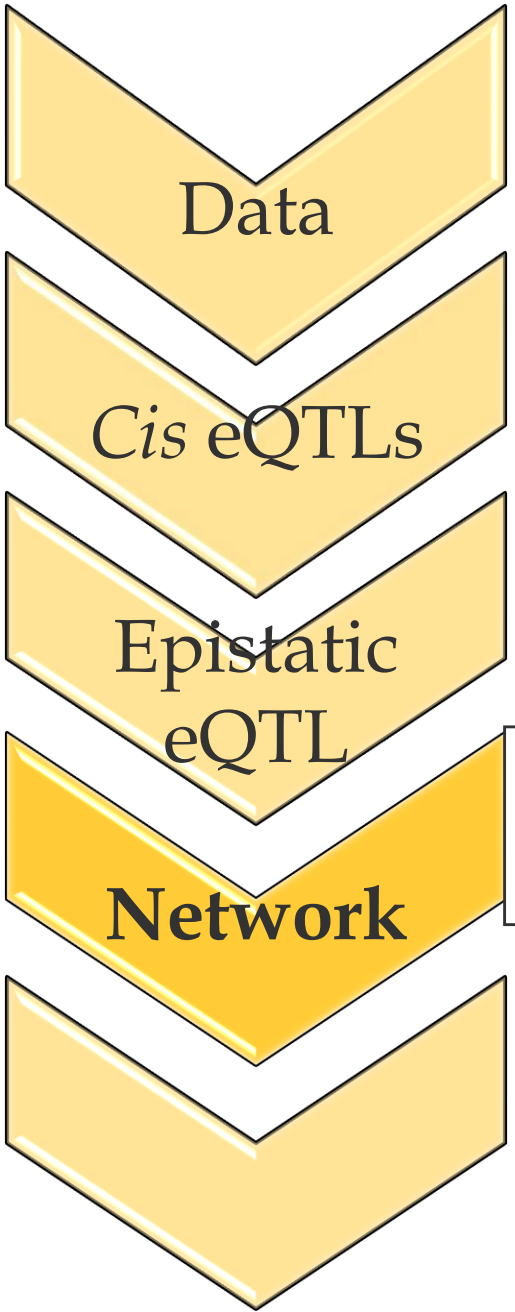
Cis eQTLs

Epistatic eQTLs

- For 1763 *cis* eQTLs
- Find epistatic signals (*trans/cis* eQTLs)
  - MB-MDR
  - Step-wise permutation
    - Step 1 - 10^3 permutations
    - Step 2 - 10^7 permutations
- MB-MDR

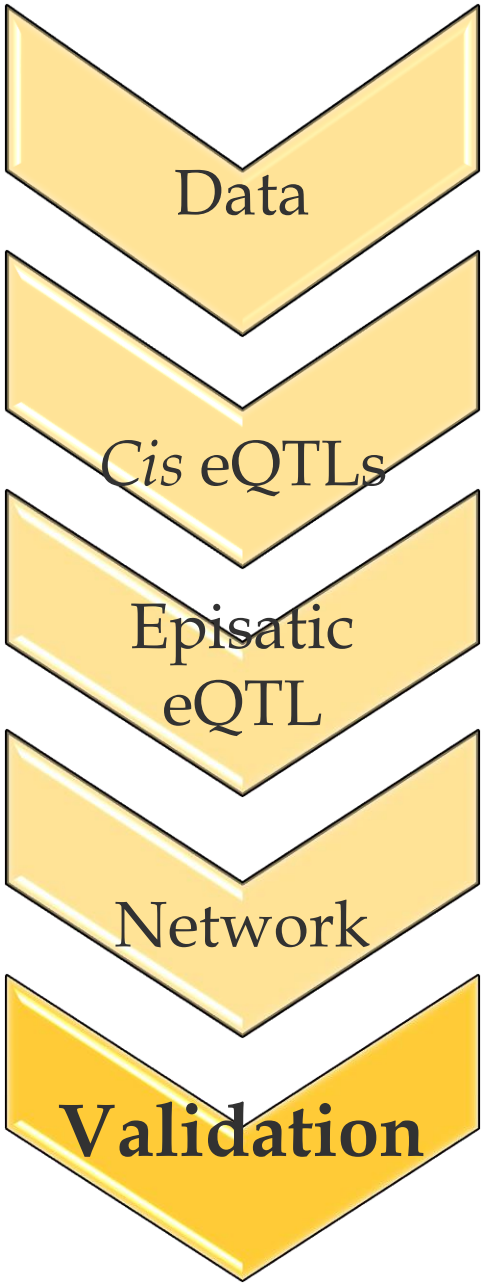


Strategy: *trans-eQTL epistasis protocol*



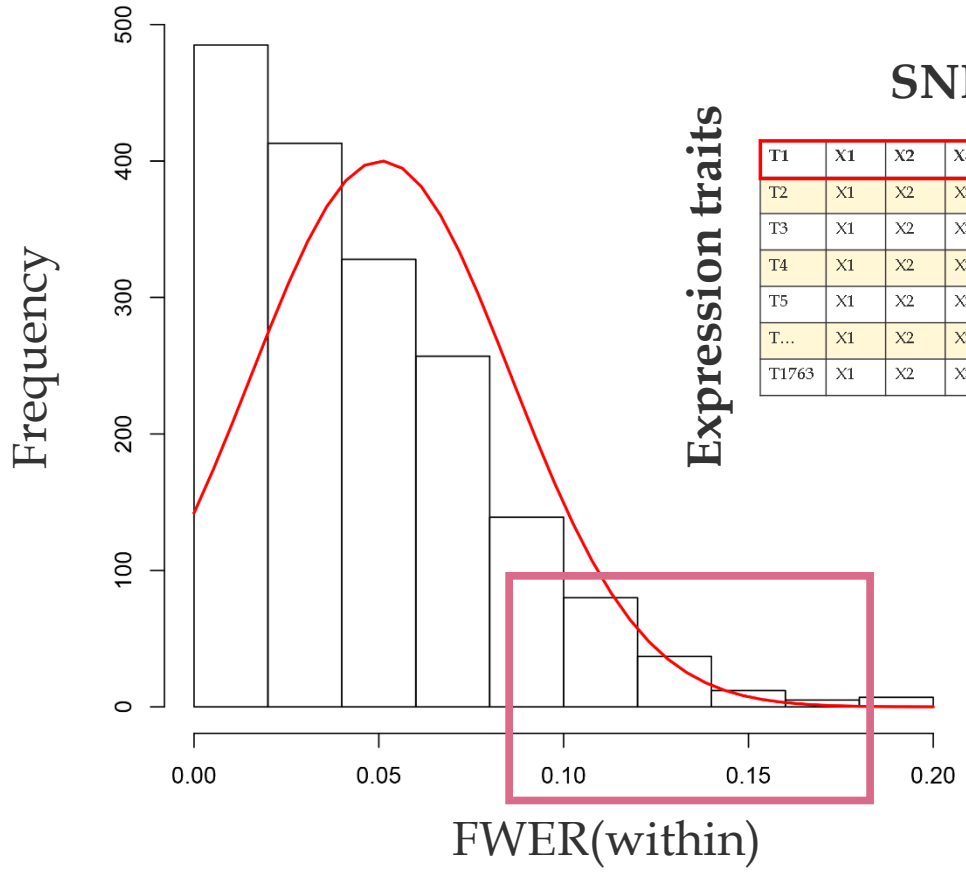
- Build statistical epistatic network
- *Trans* x *cis*, *cis* x *cis*, *trans* x *trans* interactions

Strategy: *trans-eQTL epistasis protocol*



- Disease etiology
- Previous knowledge

Simulations: null data



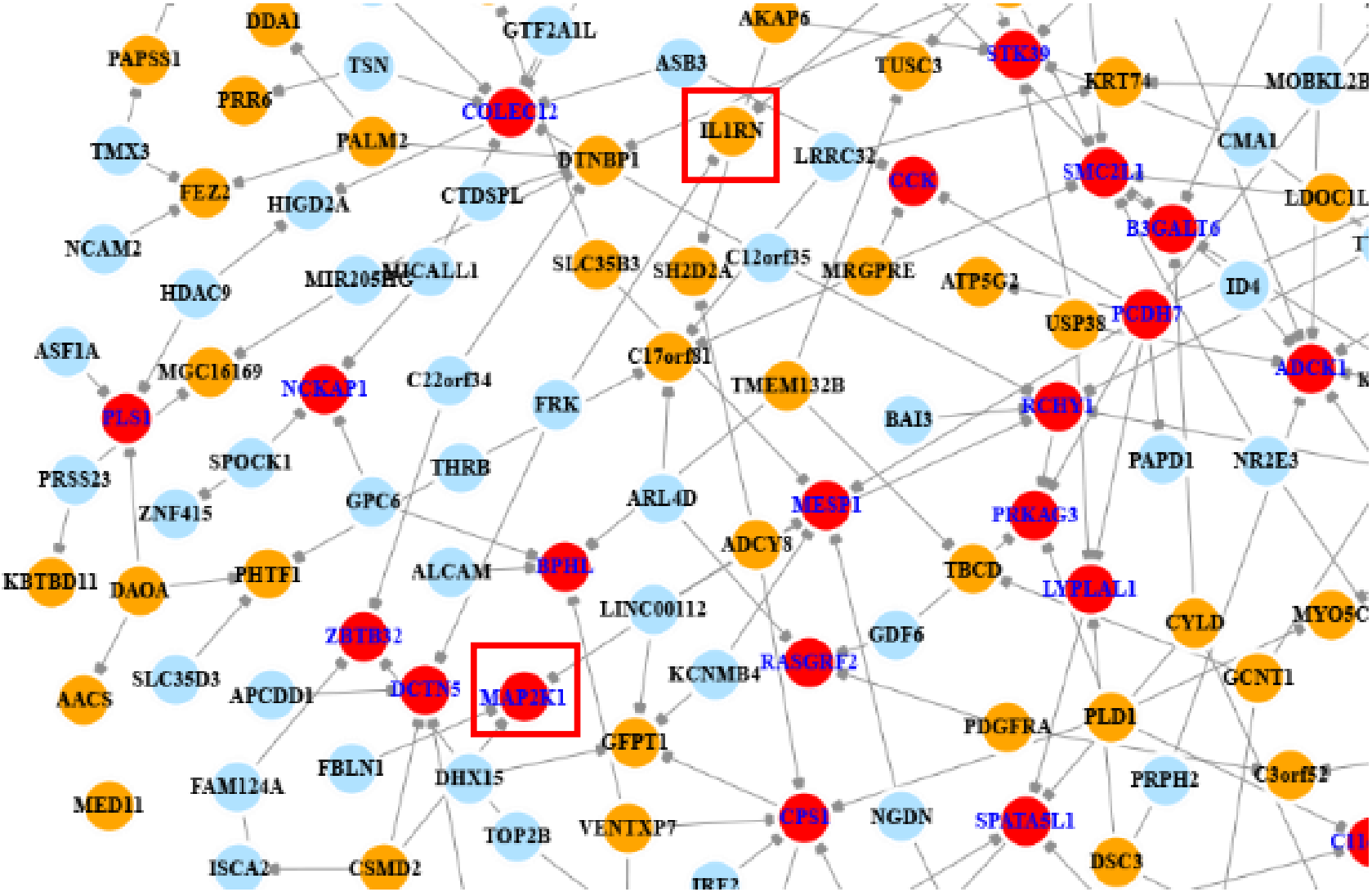
SNP x SNP pairs

Expression traits

T1	X1	X2	X3	X4	X5	X6	X7	X8	X9
T2	X1	X2	X3	X4	X5	X6	X7	X8	X9
T3	X1	X2	X3	X4	X5	X6	X7	X8	X9
T4	X1	X2	X3	X4	X5	X6	X7	X8	X9
T5	X1	X2	X3	X4	X5	X6	X7	X8	X9
T…	X1	X2	X3	X4	X5	X6	X7	X8	X9
T1763	X1	X2	X3	X4	X5	X6	X7	X8	X9

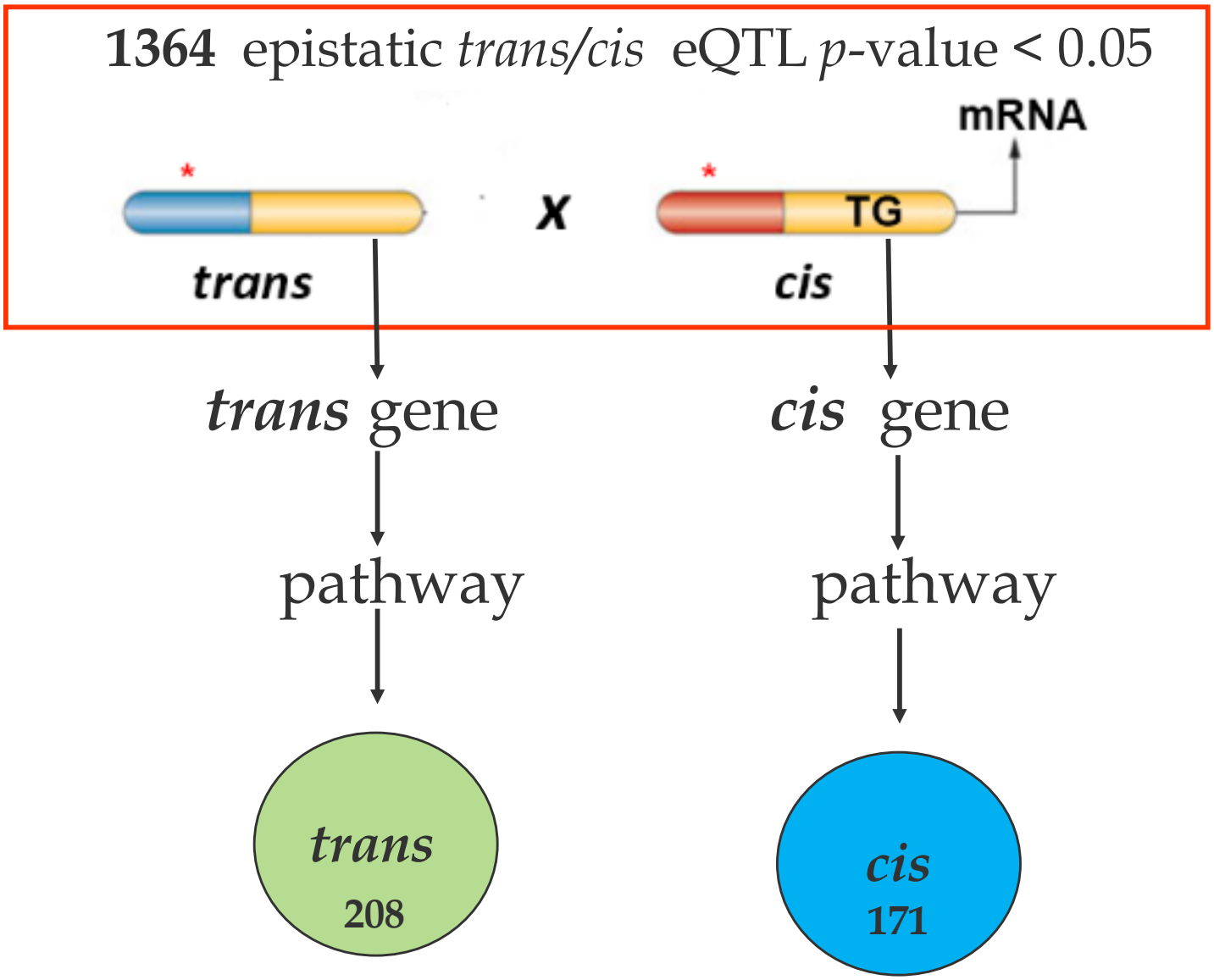
- FWER
  - within each *trans/cis* eQTL run
  - Mean 0.056
  - Median 0.04

Applications: statistical epistatic network

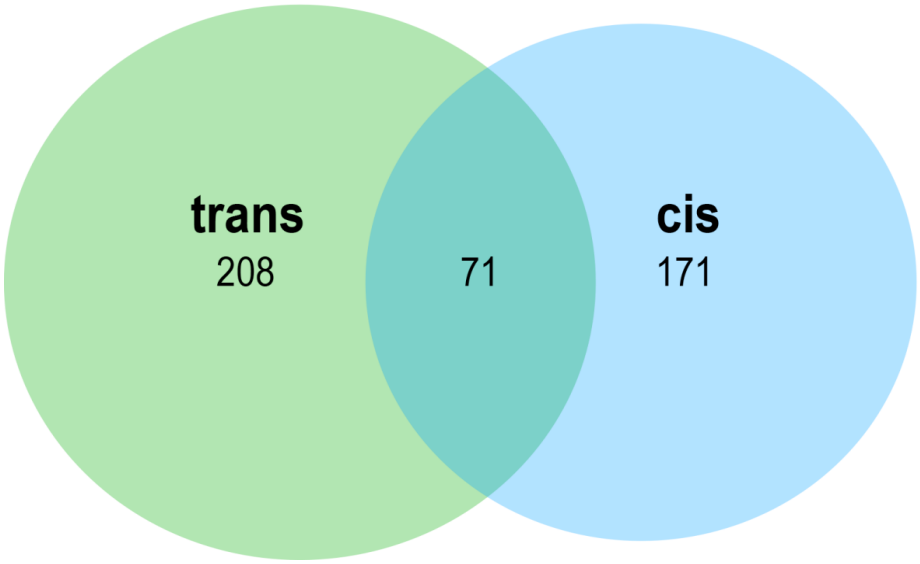


- 1459 nodes
- red: high
- orange: average

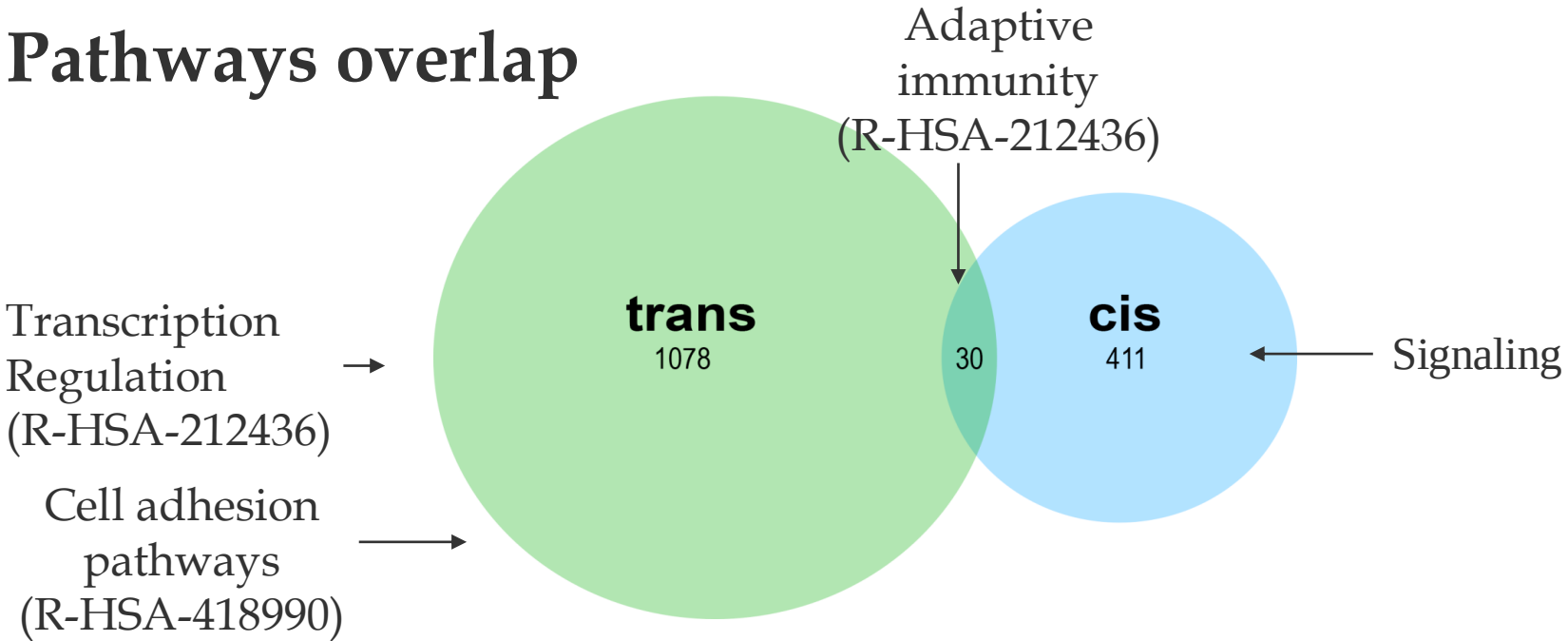
Application: mapping to pathways



Applications: significant genes overlap



Pathways overlap



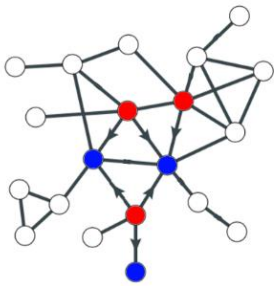
Conclusions

- Impact of genetic component on expression
  - Higher order interactions
    - *trans/cis* epistatic effects
- Global interaction map
    - Epistatic network
- Disease-relevant results

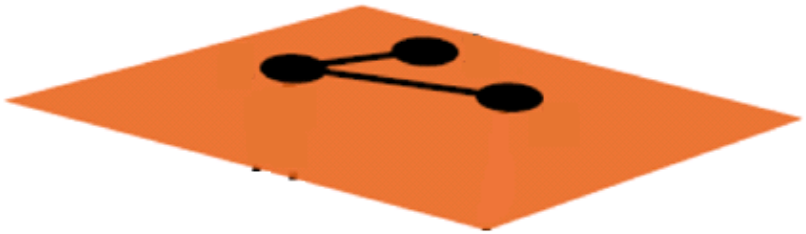
Gene expression networks

Practical aspects of gene regulatory network inference

(CIFs)

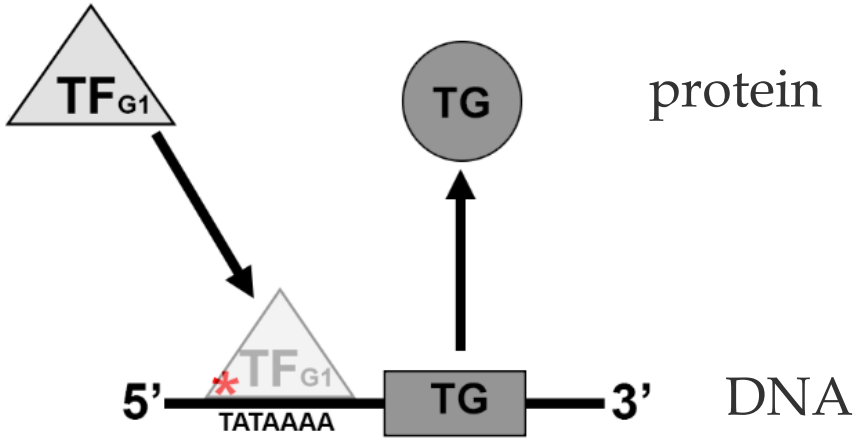


Context: transcriptome - transcriptome

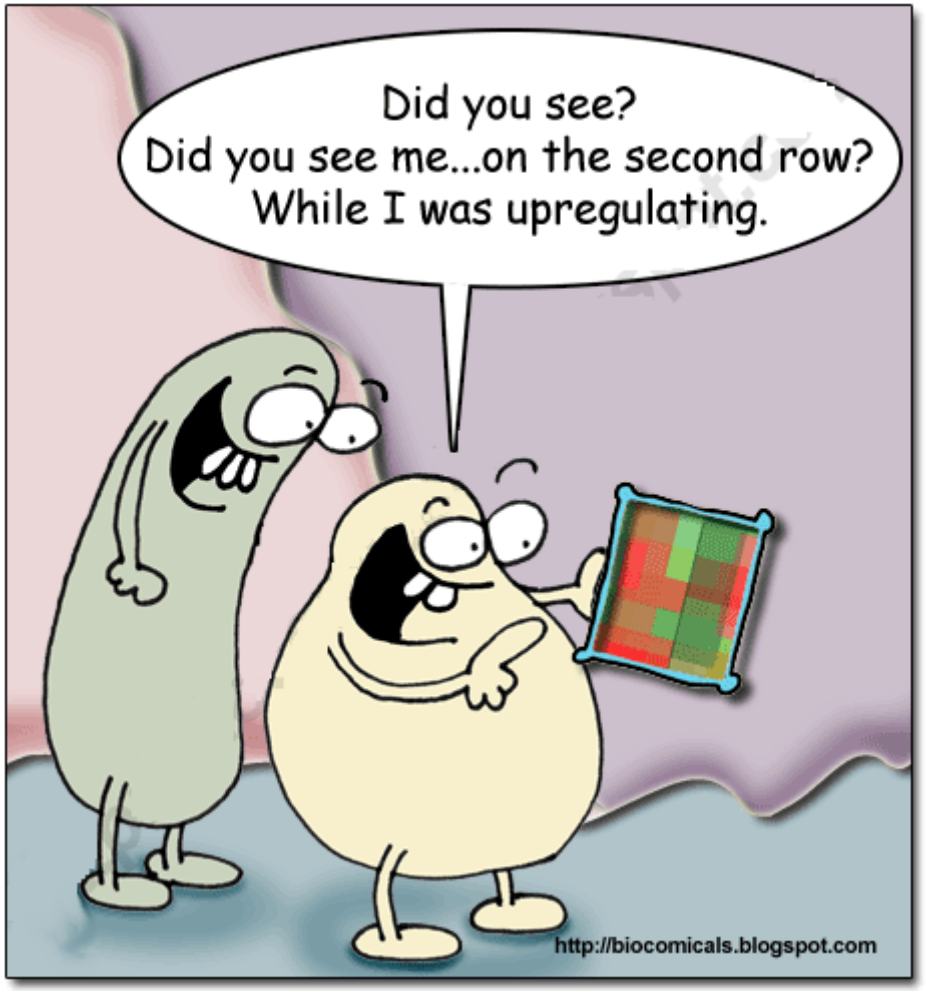


Transcriptomic data layer

- Trait – target gene (TG)
- Predictors –transcription factor (TF)

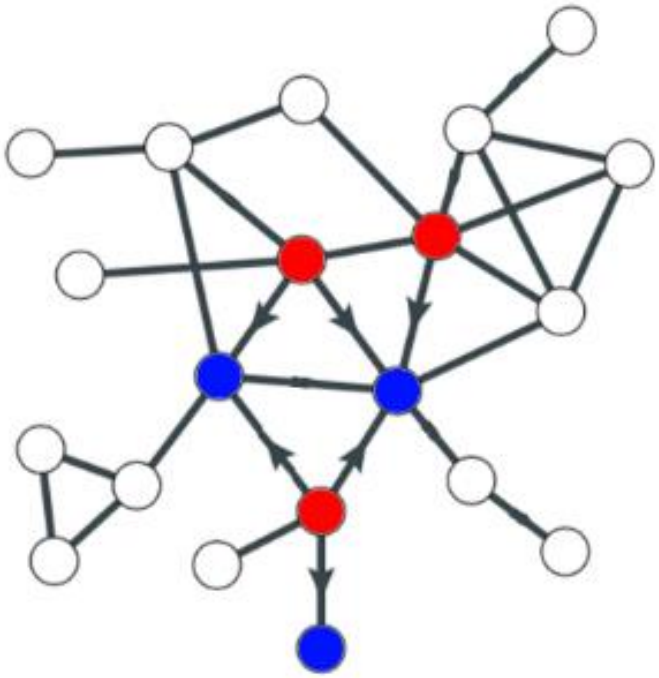


Context: transcriptome - transcriptome



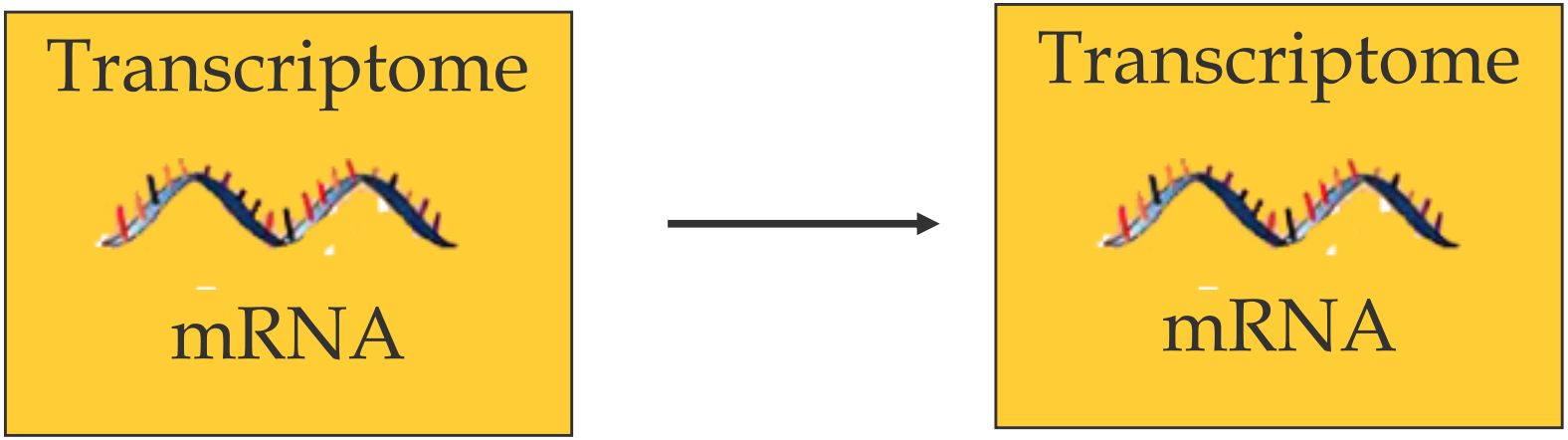
Context: transcriptional networks

- Regulators
  - Transcription factors
- Targets
  - Target genes
- Expression data
- Directed edges



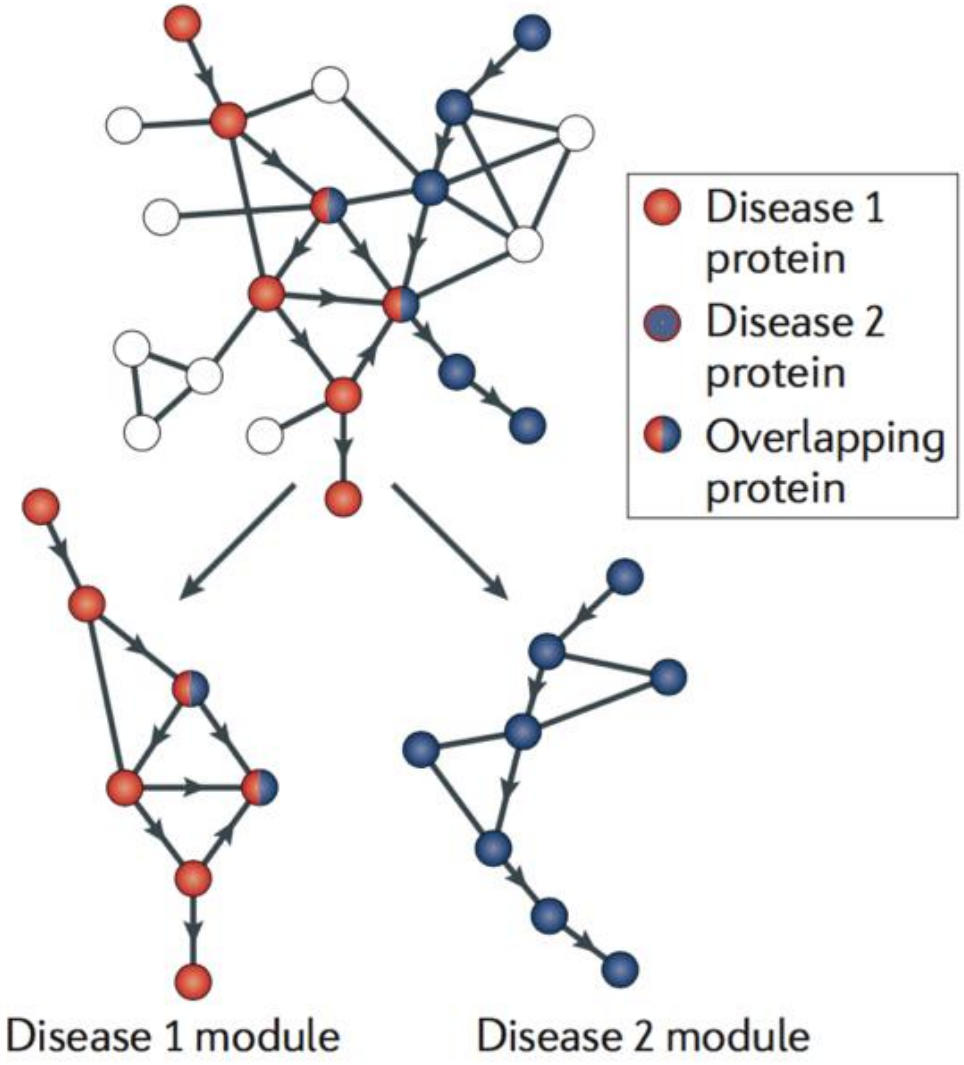
Context: problem

- Infer a transcriptional network
  - Correlation structure of genes
  - Scaling (>1000 genes)
  - *mtry* parameter
    - performance impact

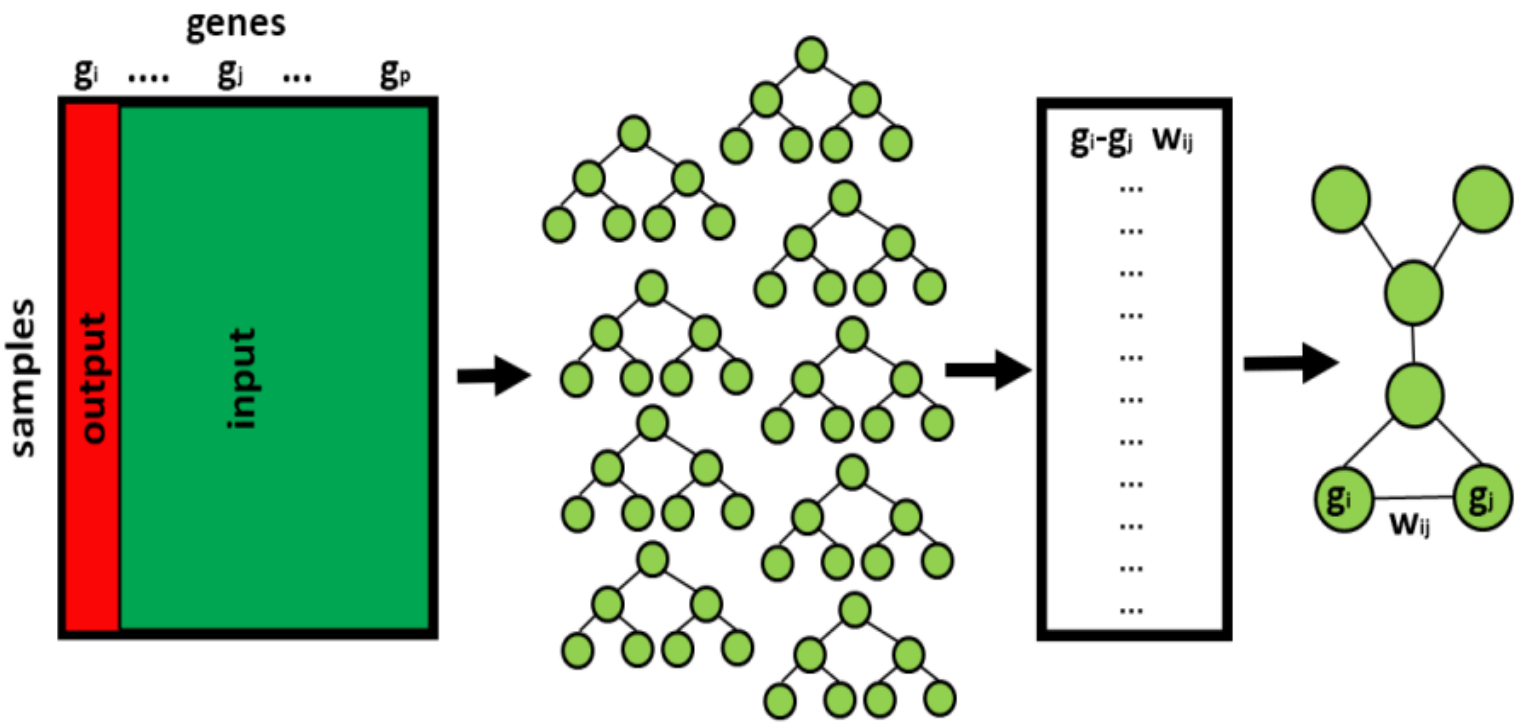


Context: network medicine

- Diseases
  - Share genes
  - Classify
  - Etiology



Strategy: network inference via trees

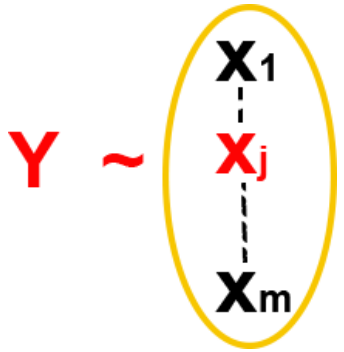


G1	G2	G3	G4	G5	G6	G7	G8

Strategy: Conditional Inference Forest

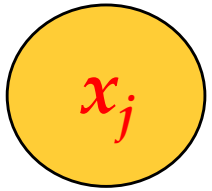
- **Select** randomly *m* variables (*mtry*) $X=\{x_1, \ldots, x_m\}$
- **For each** x_i in X test "global" null hypothesis H_0

$$H_0 = \cap_{j=1}^{m} H_0^j \text{ and } H_0^j : D(\mathbf{Y}|X_j) = D(\mathbf{Y})$$



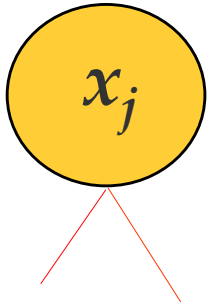
- **Select one** covariate x_j with largest c_{max}

$$c_{max}(\mathbf{t}, \mu, \Sigma) = \max_{k=1,\ldots,pq} \left| \frac{(t-\mu)_k}{\sqrt{(\Sigma)_{kk}}} \right| = \left| \frac{t-\mu}{\sqrt{\Sigma}} \right|$$



- **Assign** x_j to a node
- **Split** x_j
  - Maximize split test statistic c_{split}

$$c_{max}(\mathbf{t}^A_{j*}, \mu^A_{j*}, \Sigma^A_{j*}) = max_k \left| \frac{(\mathbf{t}^A - \mu)_k}{\sqrt{(\Sigma)_{kk}}} \right|$$



Strategy: Why CIFs?

Advantages

- Threshold available
  - ➢ useful in the absence of a gold standard
- Global test of independence*
  - ➢ Avoids bias in variable selection [5]
    - 2-stages: 1) node variable selection and 2) splitting
  - ➢ Accommodates different measurement scales
- Handles correlated variables
  - ➢ Conditional permutation scheme (CIF_{cond})

* Strasser H, Weber C (1999) **On the asymptotic theory of permutation statistics.**

Strategy: Why not CIFs?

Disadvantages

- Computation time
  - In the presence of multicollinearity (f.i. gene co-expression) the conditional variable importance measure is advocated

- Selects the features with the best "linear" association to the outcome
  - Tends to miss non-linear associations
    - Proposed solution
      - Generalized additive models (GAM)

Strategy: CIF variants

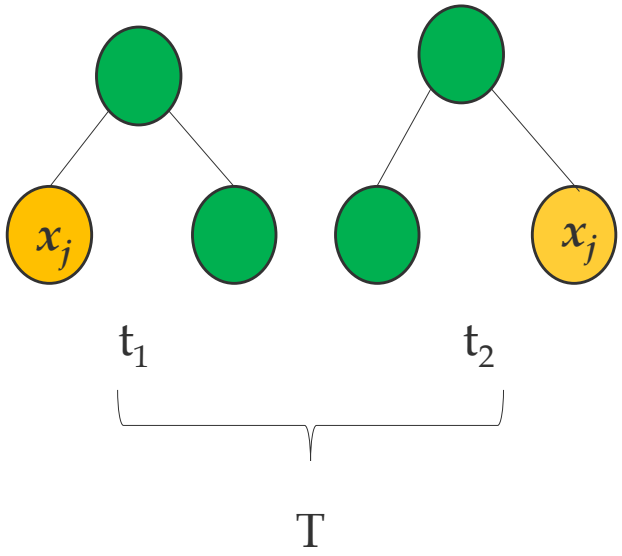
- **CIT**
  - Single conditional inference tree
- **CIF**
  - original CIF
  - classical permutation scheme
- **CIF$_{cond}$**
  - original CIF
  - conditional permutation scheme
- **CIF$_{mean}$**
  - CIF without permutation
  - Averaging of node *p*-values or test-statistics
- **RF** – Random Forest

Strategy: CIF_{mean}

- No permutations are required
- Multiple-test control at each node
  - Bonferroni (samples)
  - Monte-Carlo
- Variable importance for each x_j
  - Average over n trees where x_j is present

$$\frac{\sum_t^T p_{Xjt}}{n(X_j{}^t)}$$

Number of trees containing x_j



Results: DREAM data

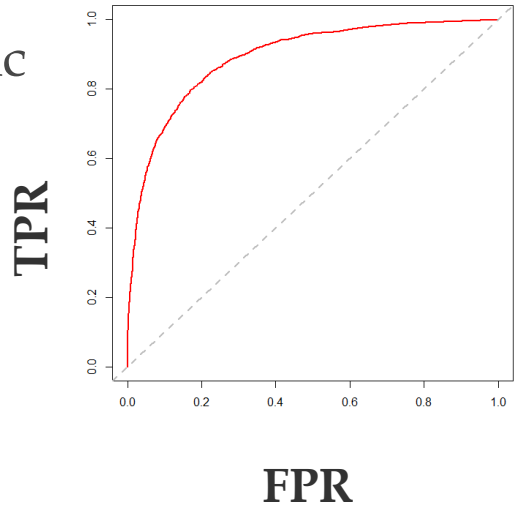
- **D**ialogue for **R**everse **E**ngineering **A**ssessments and **M**ethods
- Gold Standard available
- Predict
  - Gene regulatory network
- Data
  - Expression

Results: DREAM Data

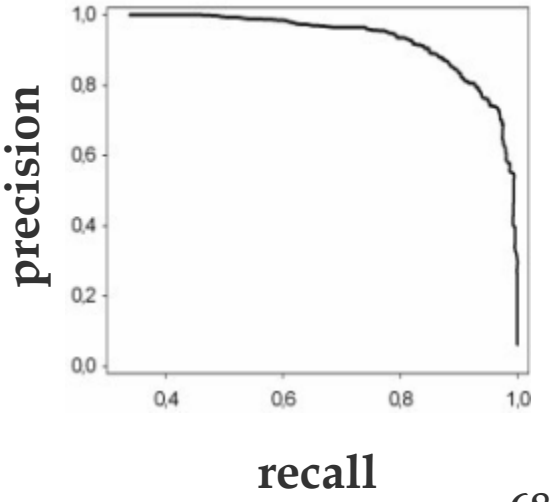
Dataset	GS available	Real-life	Nr of genes	Nr of TFs	Nr of Samples
DREAM2 (E.coli)	Y	Y	3456	320	300
DREAM4 network 1	Y	N	100	100**	100
DREAM4 network 2	Y	N	100	100**	100
DREAM4 network 3	Y	N	100	100**	100
DREAM4 network 4	Y	N	100	100**	100
DREAM4 network 5	Y	N	100	100**	100
DREAM5 network 1	Y	N	1643	195	805
DREAM5 network 2 (E.coli)	Y	Y	4511	334	805
DREAM5 network 3 (S.cerevisiae)	Y	Y	5950	333	536

Results: measures of evaluation

- AUROC
  - Area Under Receiver Operating Characteristic
  - TPR / FPR

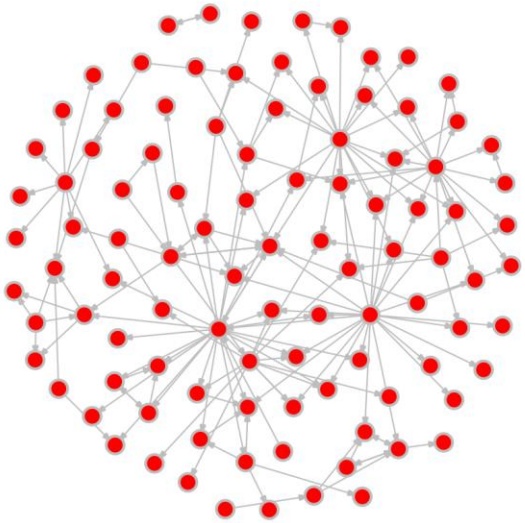


- AUPR
  - Area Under Precision Recall
  - Precision / Recall

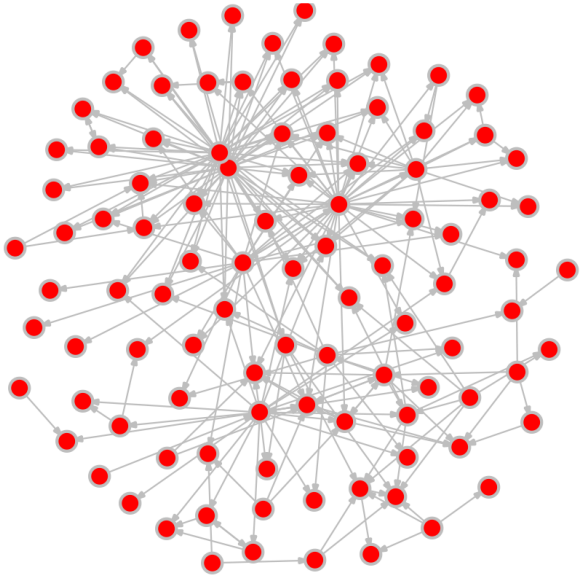


- DREAM 4/5 score
  - 1/2 * (ROC score + PR score)
  - 25,000 of random networks (re-sampling)

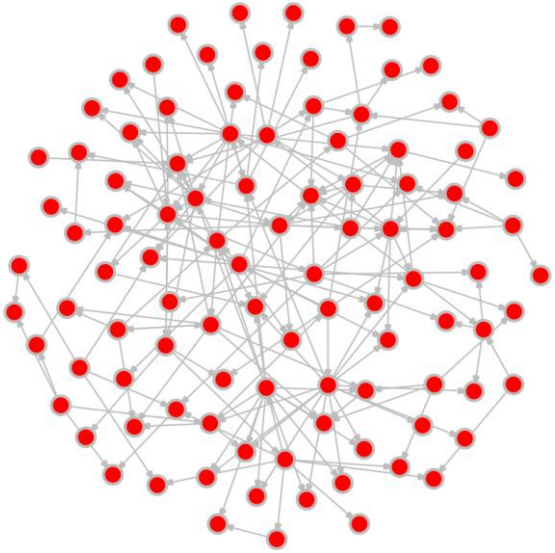
Results: DREAM 4 gold standards



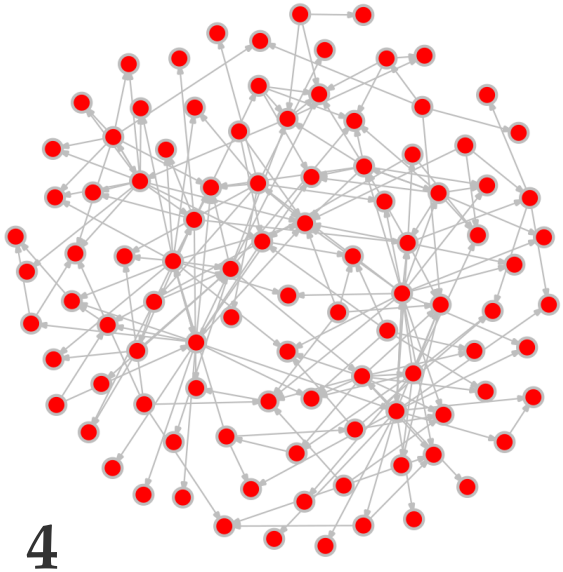
1



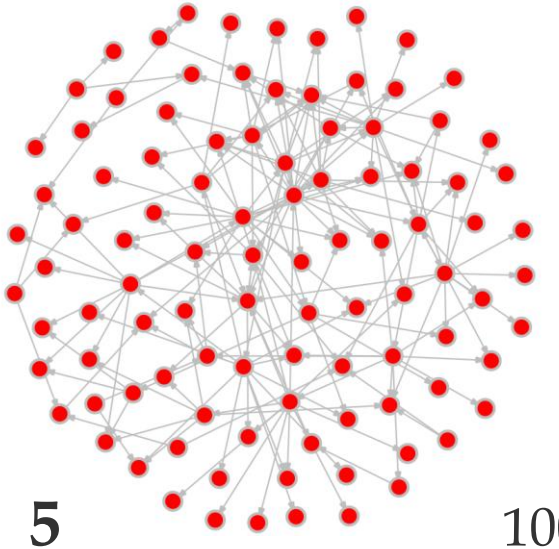
2



3



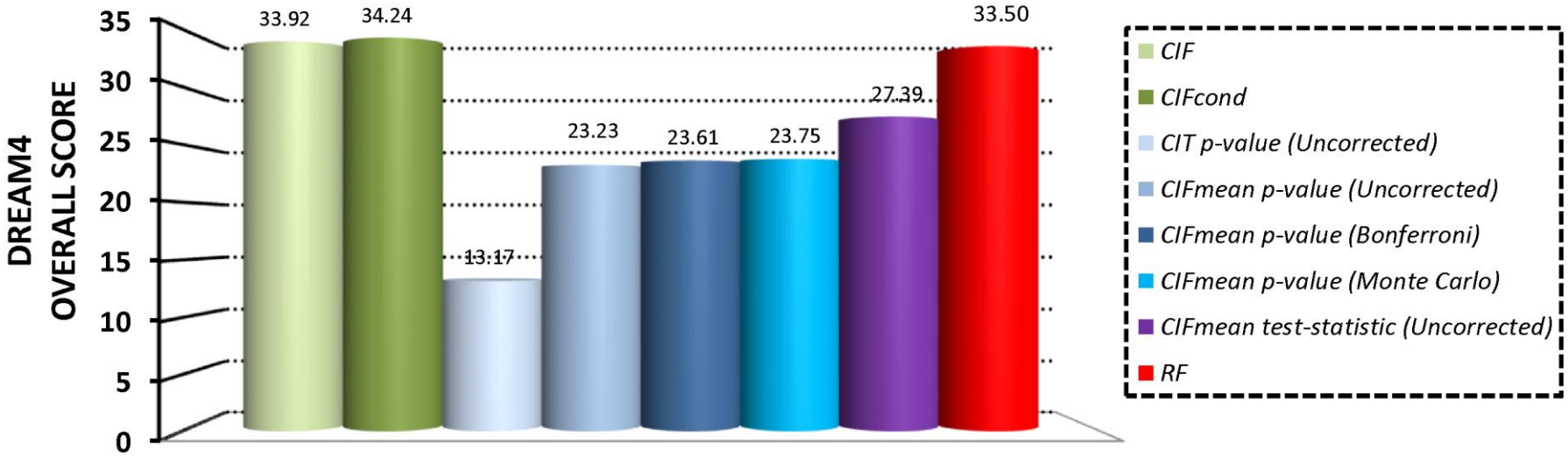
4



5

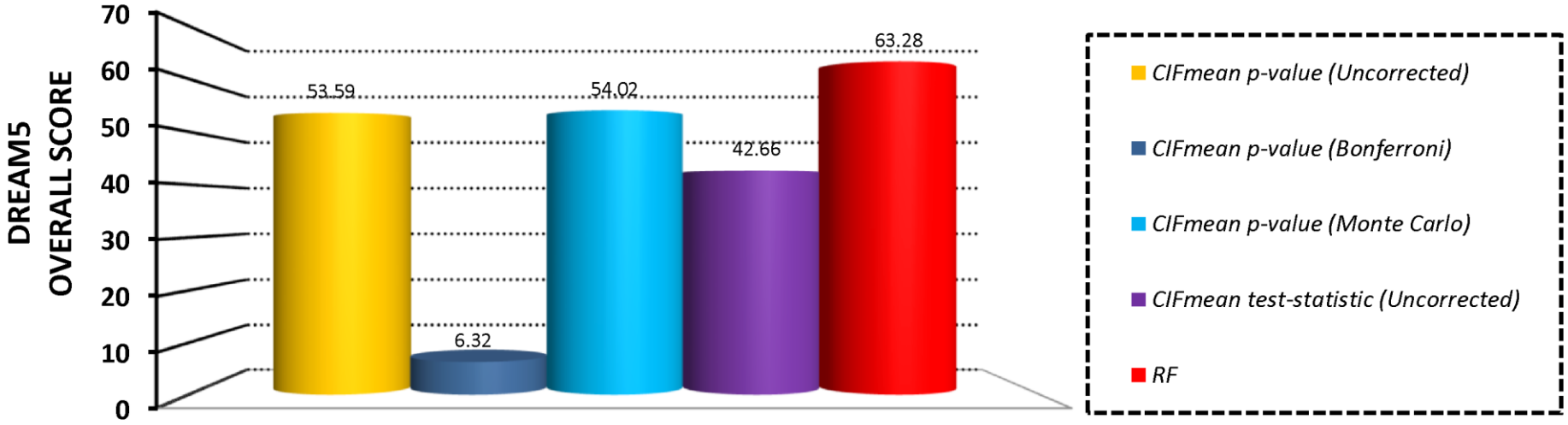
100 nodes each

Results: DREAM4



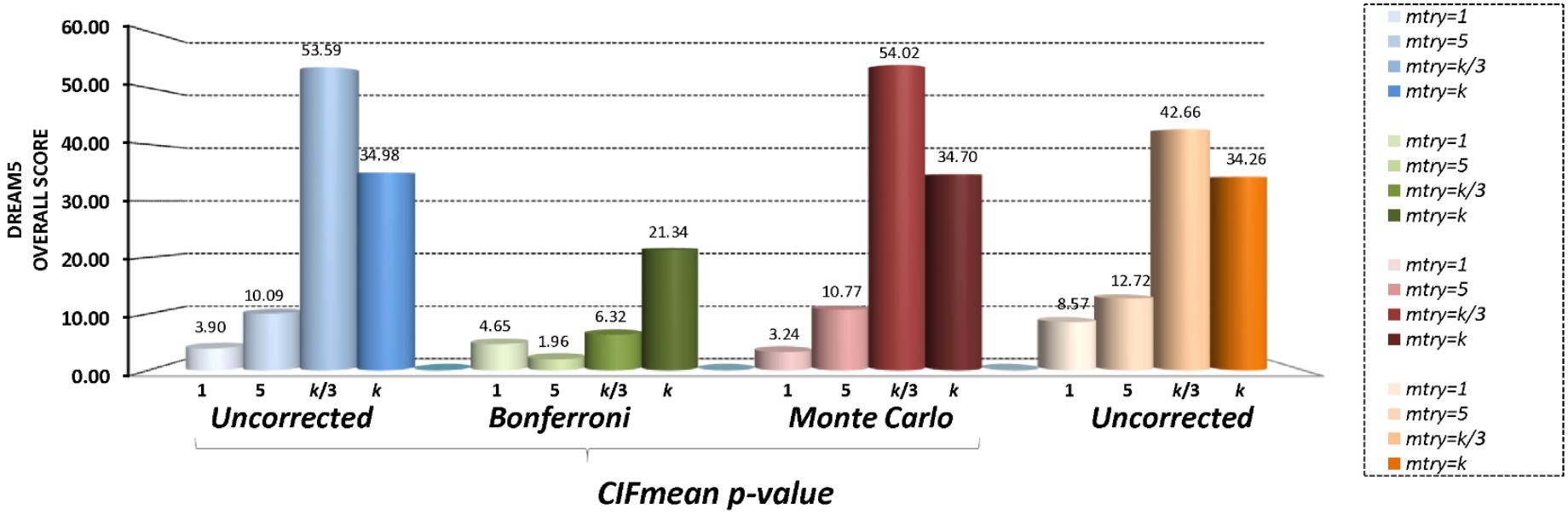
- Single tree (CIT)
  - Poor performance
- CIF_{cond} performance slightly better than RF

Results: DREAM 5



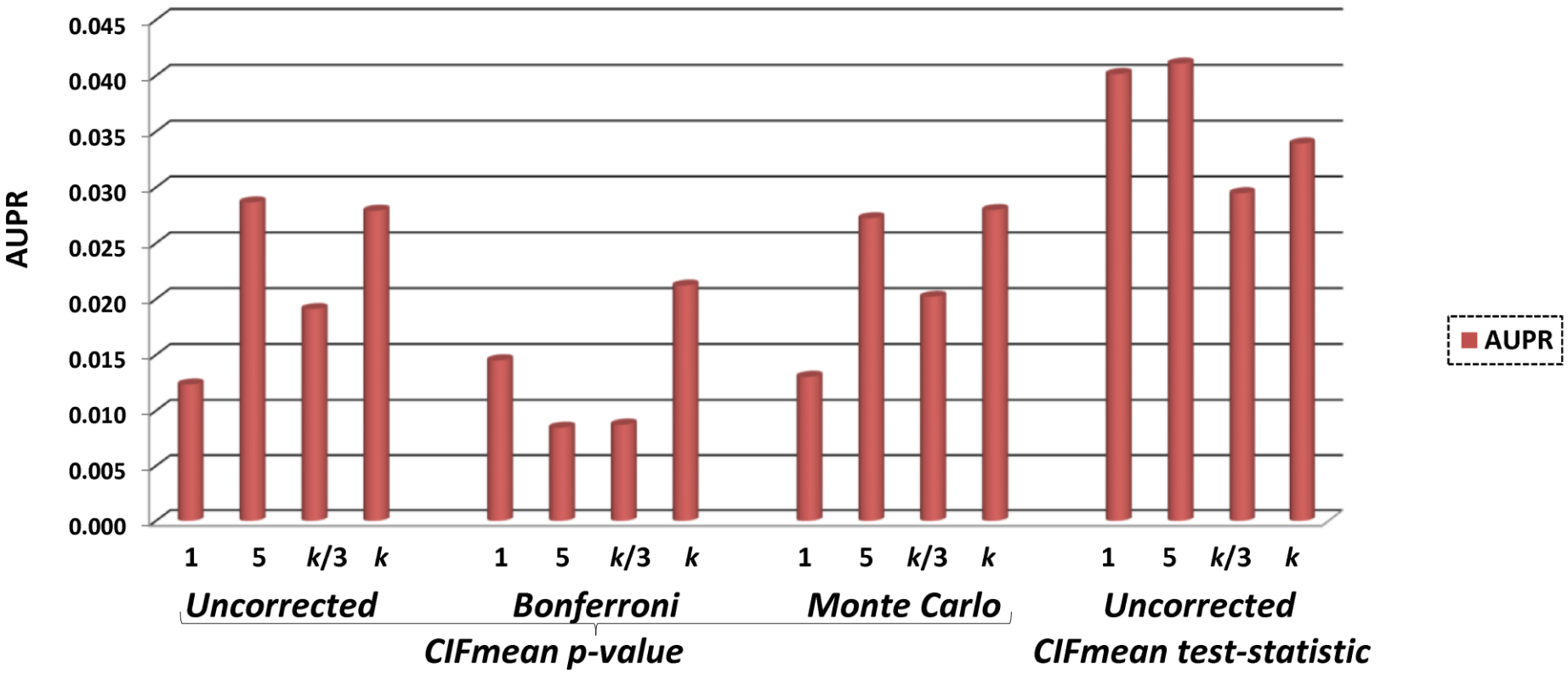
- CIF_{mean} comparable to RF and GENIE3 performance
- Little gain from Monte-Carlo MT
- Bonferroni is not to be recommended

Results: DREAM5 - *mtry*



- *mtry* parameter
  - Significant performance impact
  - Here *k/3* is the top performer

DREAM2 - *mtry*



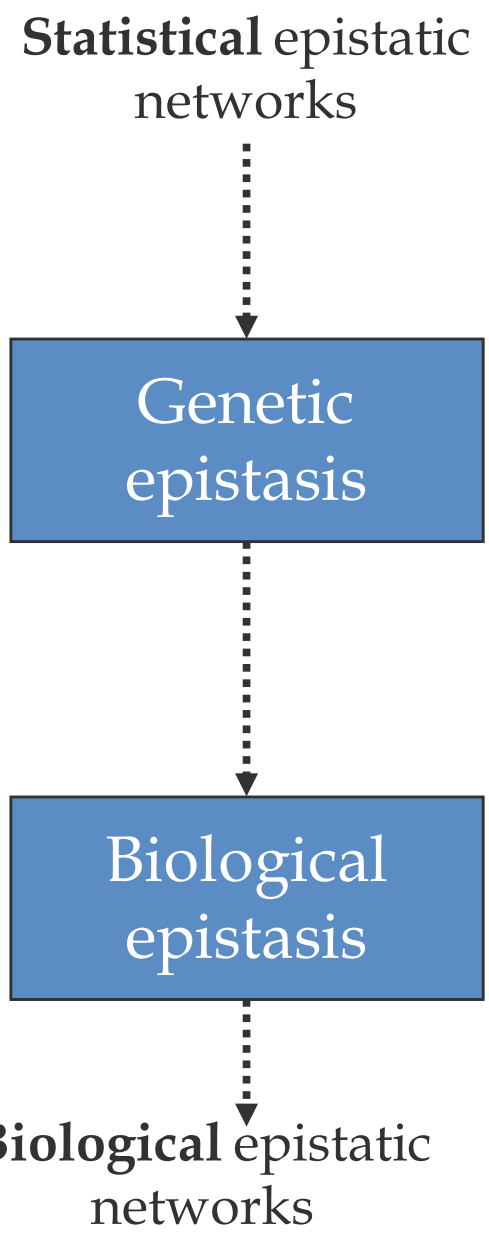
- *mtry* parameter
  - Significant performance impact
  - here *k*=5 is the top performer

Conclusions

- CIFs provide comparable performance to RF
- CIFs are scalable
  - Multi-thread runs
  - CIF_{mean} (12 min/100 genes/100 samples /1CPU)
- CIFs imply statistically sound variable selection
  - Significance-based threshold selection
  - No gold standard needed (CIF_{mean})

General conclusions

Conclusions



- Small protocol changes in epistasis screening can have a major impact on replication and validation **follow-up** studies
- Using prior information helps in obtaining more robust results, yet limits the detection of novel (not previously reported) gene-gene interactions
- Sometimes pragmatic approaches to feature selection need to taken in very small *n* datasets
- Even in the absence of multicollinearity or highly correlated features, CIF_{cond} showed comparable results with RF (DREAM4 score).

Future directions

Future perspectives

Genome-Genome Interactions

1. Optimal LD pruning threshold definition
   - Determine the lower bounds for LD pruning (now $r^2 > 0.75$)
2. Epistatic hits aggregation over protocols
   - Optimally combine complementary epistatic evidences from different epistasis detection routes

Perspectives

Trans-eQTL epistasis protocol

1. Apply the protocol to sufficiently large datasets (large n)
   - Carry out a thorough evaluation of false positives (FWER)
   - Assess the impact of the step-wise procedure (MB-MDR) on false positives

Gene expression networks

1. Increase computational efficiency (speed) in the conditional variable importance computations

Papers (14 published)

Statistical genetics / genetic epidemiology related to my PhD thesis

1. **Bessonov K**, Gusareva ES, Van Steen K (2015) A cautionary note on the impact of protocol changes for genome-wide association SNP x SNP interaction studies.Hum Genet 134:761-773
2. Pineda S, Gomez-Rubio P, Picornell A, **Bessonov K**, Márquez M, Kogevinas M, Real FX, Van Steen K, Malats N. Framework for the integration … . Hum Hered. 2015;79(3-4):124-36
3. Bollen L, Vande Casteele N, Peeters M, **Bessonov K**, Van Steen K, Rutgeerts P, *et al.* . Short-term Effect of Infliximab … . Inflamm Bowel Dis. 2015 Mar;21(3):570-8
4. Gusareva ES, … , Dickson DW, Mahachie John JM, **Bessonov K**, Van Steen K, *et al.* Genome-Wide Association … . Neurobiol Aging. 2014 Nov;35(11):2436-43
5. Fouladi R, **Bessonov K**, Van Lishout F, Van Steen K. Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis. Hum Hered. 2015;79(3-4):157-67
6. **Bessonov K**, Van Steen K (2015) Practical aspects of gene regulatory inference via conditional inference forests from expression data. (Genetic Epidemiology submitted July 2015)
7. Francesco G¶, **Bessonov K**¶, Van Steen K (2015) Integration of Gene Expression and Methylation to unravel biological networks … (submitted to Genetic Epidemiology)

In preparation / submitted

1. **Bessonov K**, Van Steen K (2015) Practical aspects of gene regulatory inference via conditional inference forests from expression data. (Genetic Epidemiology submitted July 2015)
2. **Bessonov K**, Croteau-Chonka D, Qi W, Carey VJ, Raby BA, Van Steen K (2015) Integrative network-based analysis of cis and trans regulatory effects in asthma.
3. Schleich F., **Bessonov K**, Van Steen K (2015). Exhaled volatile organic compounds are able to discriminate between neutrophilic and eosinophilic asthma. (submitted) - Patent #203-17

Data mining/molecular dynamics related

1. **Bessonov, K.,** Harauz, G.(2010) "In silico study of the myelin basic protein C-terminal α-helical peptide in DMPC and mixed DMPC/DMPE lipid bilayers." Studies by Undergraduate Researchers at Guelph 4:1 [http://www.criticalimprov.com/index.php/surg/article/view/1102]
2. Luiza Antonie and **Kyrylo Bessonov** (2012), "Biologically Relevant Association Rules for Classification of Microarray Data", Applied Computing Review (ACR) 2012; 12(1)
3. **K. Bessonov**, K.A. Vassall, G. Harauz, "Parameterization of the proline analogue Aze for molecular dynamics simulations …", *Journal of Molecular Graphics and Modelling* (2012)
4. **Bessonov K**, "Functional Analyses of NSF1 in wine yeast using Interconnected Correlation Clustering … " 2012, submitted to *PLoS ONE* (Manuscript #: PONE-D-12-35841)

Biology related

1. **Bessonov K,** Bamm VV, Harauz G. "Misincorporation of the proline homologue Aze (azetidine-2-carboxylic acid) into recombinant …." Phytochemistry 2010; 71(5-6): 502-507
2. Berg L, Koch T, Heerkens T, **Bessonov K**, Thomsen P, Betts D. "Chondrogenic potential of mesenchymal stromal cells derived ..." Vet Comp Orthop Traumatol. 2009; 22(5): 363-70
3. **Kyrylo Bessonov** and Dr. George Harauz. "In-silico study of the myelin basic protein C-terminal α-helical peptide in DMPC and mixed DMPC/DMPE lipid bilayers." Studies by Undergraduate Researchers at Guelph 2010; 4(1).
4. Lopamudra Homchaudhuri, Miguel De Avila, Stina B. Nilsson, **Kyrylo Bessonov,** Graham S.T. Smith, Vladimir V. Bamm, Abdiwahab A. Musse, George Harauz,and Joan M. Boggs. "Secondary Structure and Solvent Accessibility of a Calmodulin-Binding C-Terminal Segment of Membrane-Associated Myelin Basic Protein." Biochemistry 2010; 49(41):8955-66
5. Mumdooh A.M Ahmed, Miguel De Avila, Eugenia Polverini, **Kyrylo Bessonov,** Vladimir V. Bamm, George Harauz. "Solution NMR structure and molecular dynamics simulations of murine 18.5-kDa myelin basic protein segment (S72-S107) in association with dodecylphosphocholine micelles". Biochemistry

Acknowledgements

- Supervisor
  - Prof. Dr. Dr. Kristel Van Steen
- BIO3 lab
- Committee members
  - Pierre Geurts ▪ Vincent Bours ▪ Benno Schwikowski
  - Patrick Meyer ▪ Monika Stoll
- Funding
  - FNRS, COST BM1204, ITN MLPM
  - University of Liege

References

1. Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, et al. (2011) **Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility.** *Nat Genet* 43: 761-767.
2. Van Lishout F, Mahachie John JM, Gusareva ES, Urrea V, Cleynen I, et al. (2013) **An efficient algorithm to perform multiple testing in epistasis screening.** *BMC Bioinformatics* 14: 138.
3. Borish, L. A. R. R. Y., et al. "**Detection of alveolar macrophage-derived IL-1 beta in asthma. Inhibition with corticosteroids.**" The Journal of Immunology149.9 (1992): 3078-3082.
4. Gusareva, Elena S., and Kristel Van Steen. "Practical aspects of genome-wide association interaction analysis." Human genetics 133.11 (2014): 1343-1358.
5. Strobl, Carolin, Torsten Hothorn, and Achim Zeileis. "Party on!." (2009).