Systems Medicine

Kristel Van Steen, PhD[2] (*)

kristel.vansteen@ulg.ac.be

(*) WELBIO, GIGA-R, Medical Genomics, University of Liège, Belgium

Systems Medicine Lab, KU Leuven, Belgium

OUTLINE

- Why looking at “interactions”?
- Case study: pancreatic cancer
- MB-MDR: a decade’s work; Lessons learned
- Implications
  - Risk prediction
  - Molecular reclassification of disease
  - Omics integration
- Take-home message

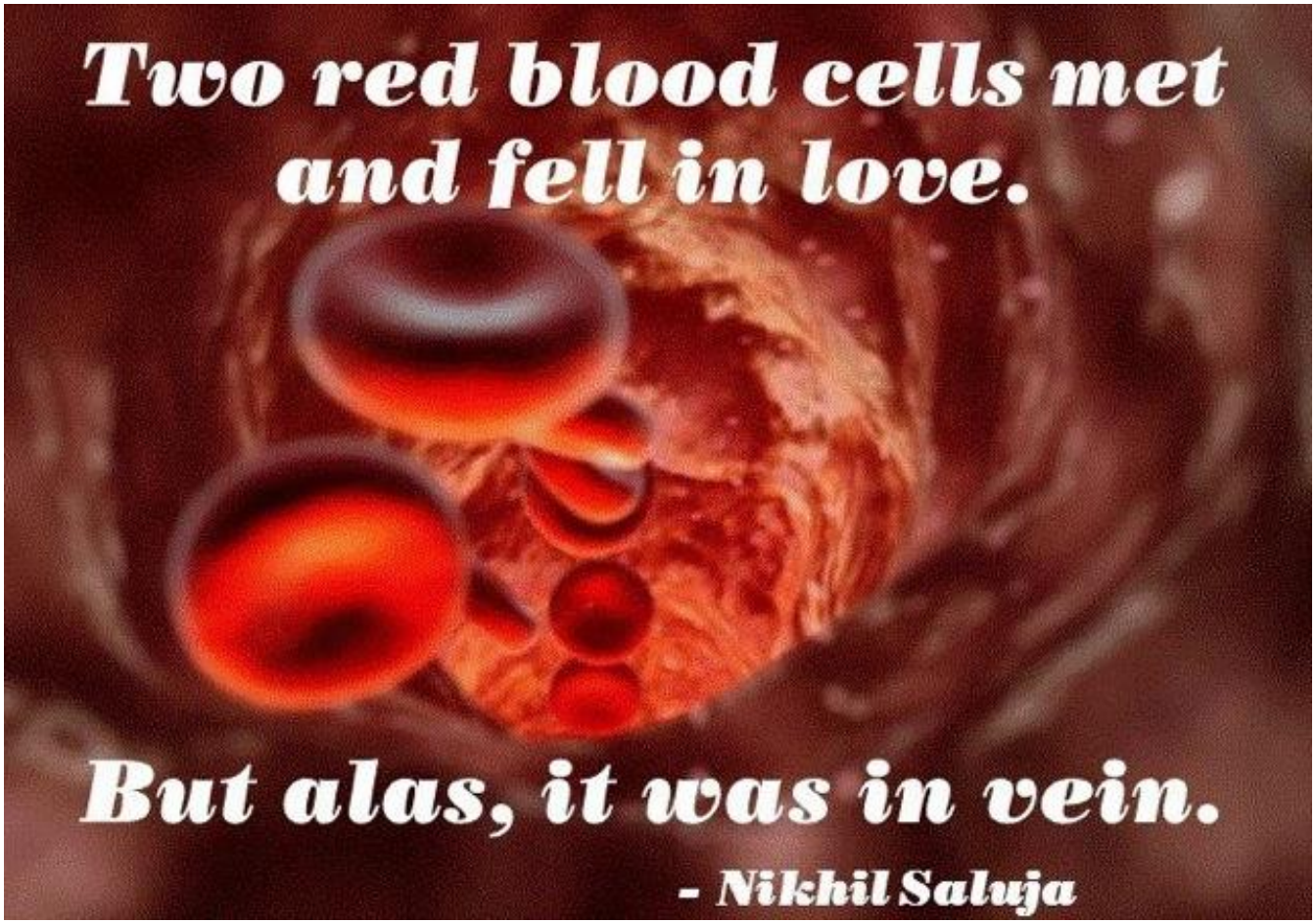
http://bio3.giga.ulg.ac.be/

BIO

Why looking at “interactions”?

including SNP based

A tale of … multiple … stories

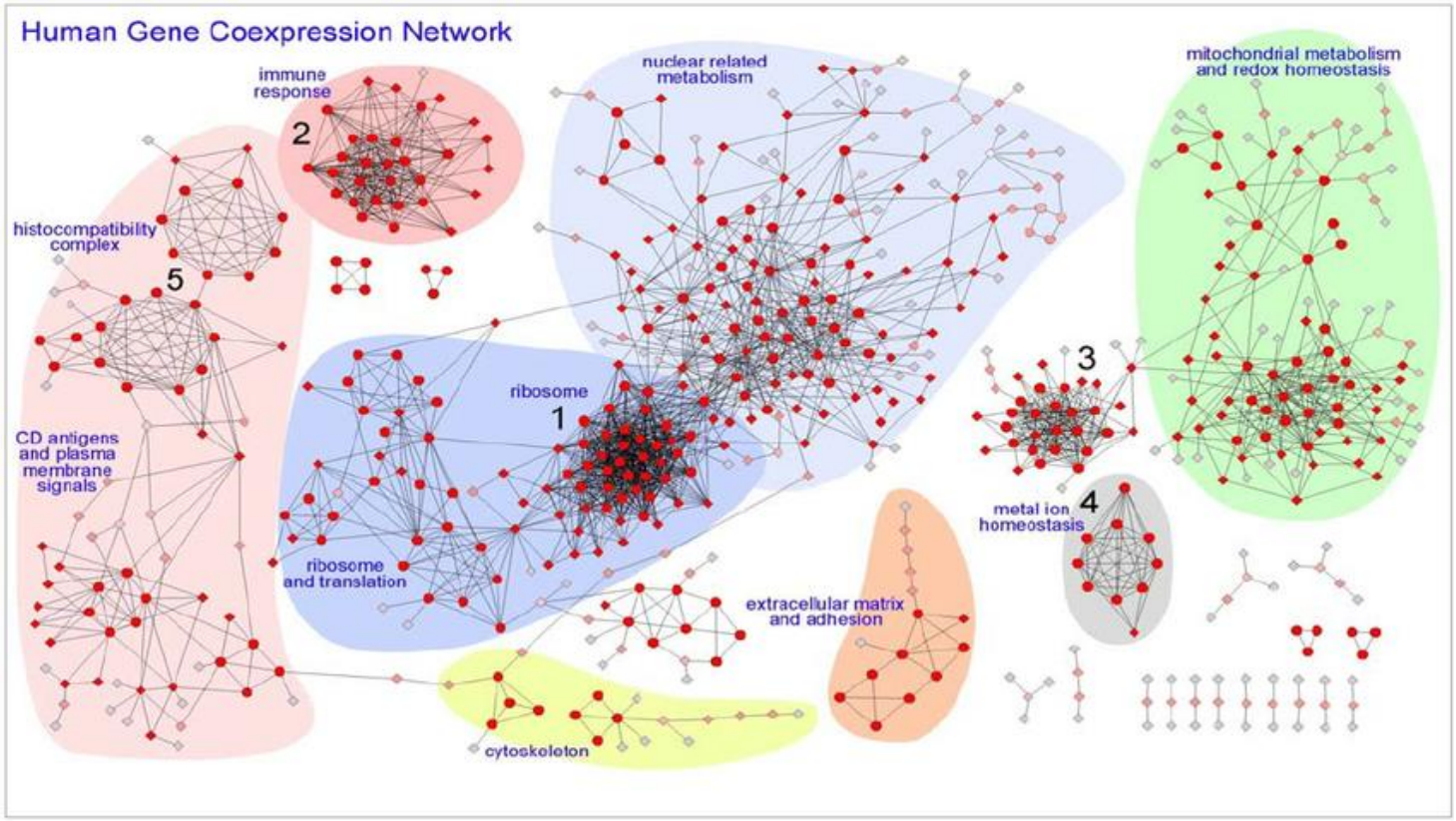


Biological interactions

- Biological interactions are the effects that the organisms in a community have on one another. In the natural world no organism exists in absolute isolation, and thus every organism must interact with the environment and other organisms.
- An organism's interactions with its environment are fundamental to the survival of that organism and the functioning of the ecosystem as a whole.

Gene-gene interactions

- Inference about gene-gene interactions using microarray data

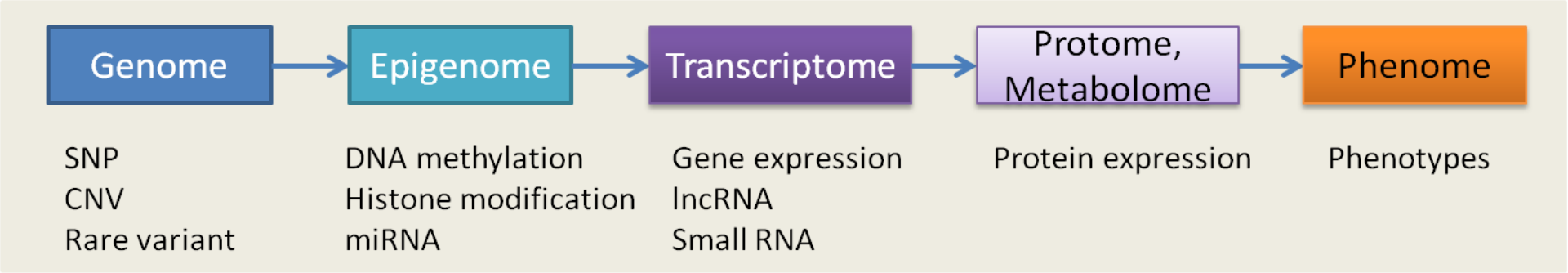


(Prieto et al. 2008)

Omics data as a starting point

- Roughly, omics data is a generic term that describes genome-scale data sets that emerge from high-throughput technologies
- These data describe virtually all biomolecules in a cell (e.g., proteins, metabolites)

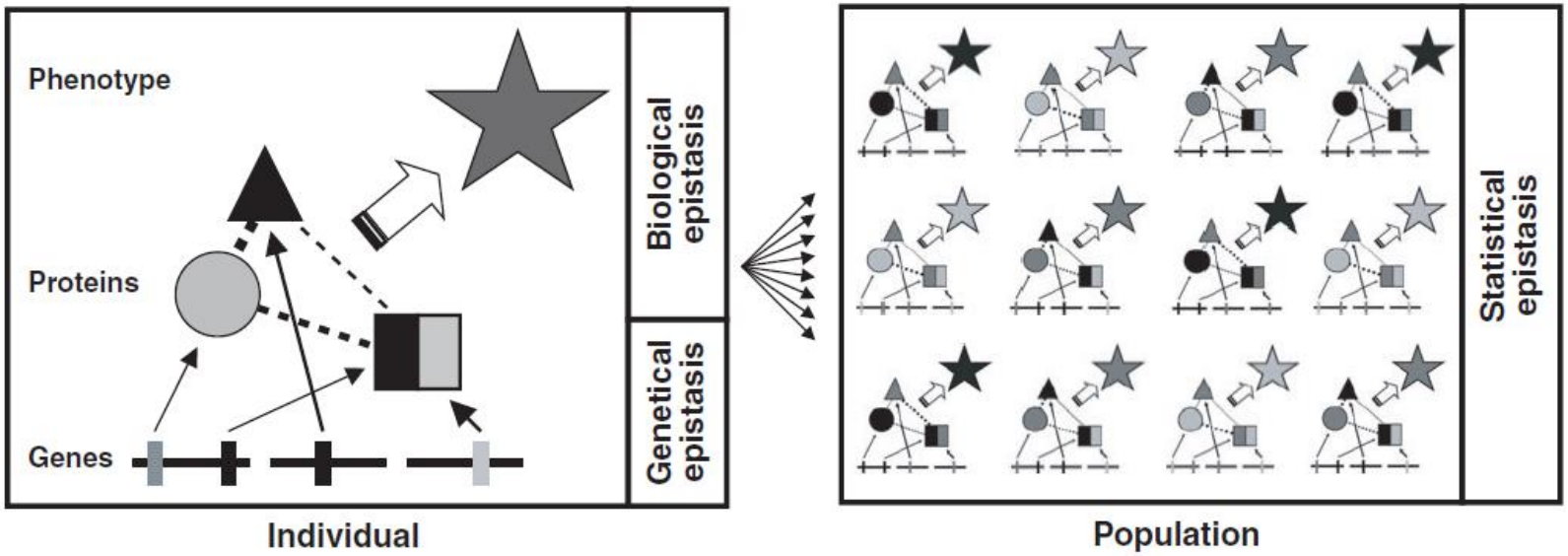
(Joyce and Palsson 2006)



(courtesy figure Maggie Wang)

DNA-DNA interactions

- Two or more DNA variations may “interact” either directly to change transcription or translation levels, or indirectly by way of their protein product (to alter disease risk separate from their independent effects)

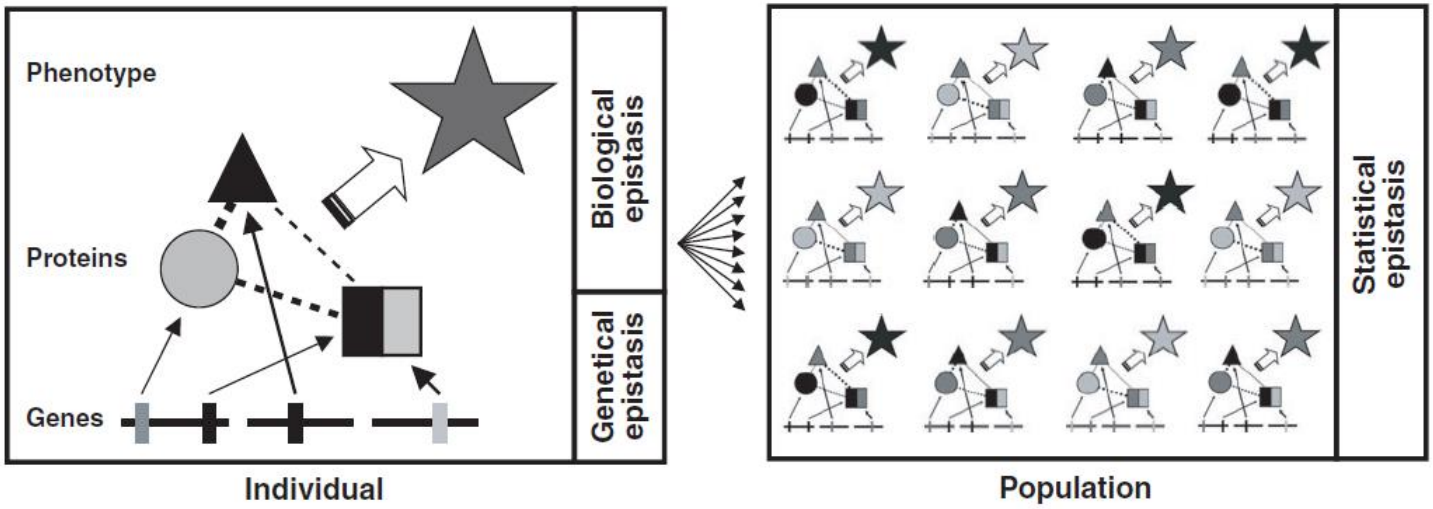


(Moore 2005)

Formal definition of epistasis

(Moore 2005; Moore and Williams 2005)

- The original definition (**driven by biology**) refers to a variant or allele at one locus preventing the variant at another locus from manifesting its effect (William Bateson 1861-1926).
- A later definition of epistasis (**driven by statistics**) is expressed in terms of deviations from a model of additive multiple effects (Ronald Fisher 1890-1962).



Occurrences of "epistasis" – model organisms

- During HSPH post-doc 2003-2005
  - Epistatic QTLs without individual effects have been found in various organisms, such as birds, mammals, insects (Drosophila Melanogaster – fruit fly) and plants.
  - Other similar studies have reported only low levels of epistasis or no epistasis at all, despite being thorough and involving large sample sizes[35–37].
- No single mode of inheritance can be expected to be the rule in all populations and traits; "complex" complex trait regulation; epistatic heterogeneity, which is likely to be contextual

(Carlborg et al. 2004)

Occurrences of epistasis – humans

- Canalization is a form of stabilizing selection to explain the buffering of phenotypes to genetic and environmental perturbations

(Waddington 1942)

Evolution tends to keep our blood pressure and glucose levels within healthy ranges (i.e., evolution of the "system" to a robust level), resistant to most genetic and environmental stimuli

- The consequence is an underlying genetic architecture that is comprised of networks of genes that are redundant and robust (*trans – epistasis*)

(Moore et al. 2009)

Deviations from these healthy ranges are often categorized as "disease", such as hypertension and diabetes

Unexplained heritability (from GWAs)

- The **statistical definition** for heritability defines it as the proportion of phenotypic variance attributable to genetic variance.
- The "sensical" definition defines it as the extent to which genetic individual differences contribute to individual differences in observed behavior (or phenotypic individual differences).
- The proportion of **heritability explained** by a set of variants is the ratio of (i) the heritability due to these variants (numerator), estimated directly from their observed effects, to (ii) the total heritability (denominator), inferred indirectly from population data.

(Maher 2008, Zuk et al. 2012)

Unexplained heritability

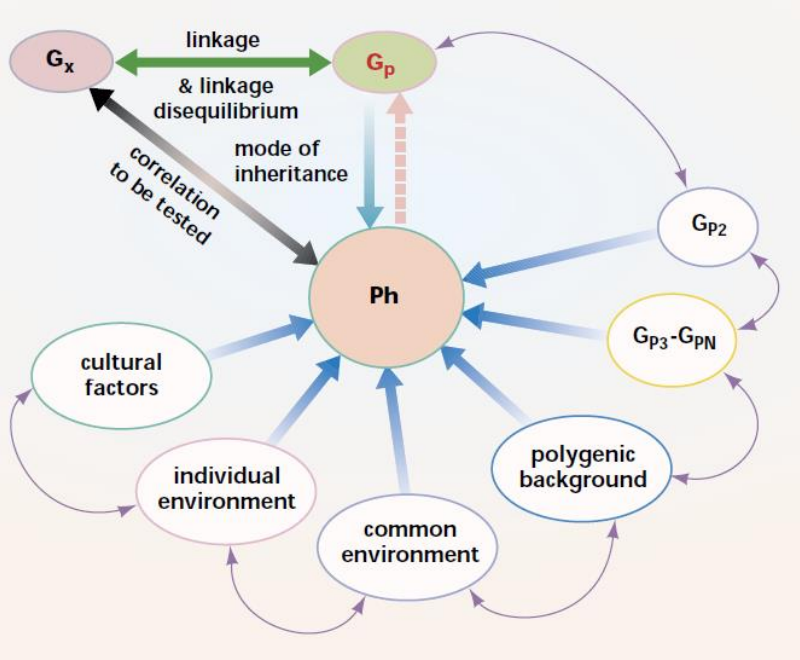
Explanation	Rationale	Comments
Overestimated heritability estimates	These estimates are typically performed in the absence of gene-gene or gene-environment interactions (Young et al. 2014)	Limiting pathway modeling suggests that epistasis could account for missing heritability in complex diseases (Zuk et al. 2012)
Rare genetic variants	Resequencing studies (e.g., WES) could identify rare genetic determinants of large effect size (Zuk et al. 2014)	Limited evidence for rare variants of major effect in complex diseases accounting for large amount of genetic variation – most rare variants analysis methods currently suffer from increased type I errors (Derkach et al. 2014)
Phenotypic and genetic heterogeneity	Most complex diseases are like syndromes with multiple potentially overlapping disease subtypes	Improvements in phenotyping of complex diseases will be required to understand genetic architecture.

Explanation	Rationale	Comments
Interactions	Gene-gene and gene-environment interactions are likely to be important for complex diseases (Moore et al 2005) Roughly 80% of the currently missing heritability for Crohn's disease could be due to genetic interactions, if the disease involves interaction among three pathways (Zuk et al. 2012)	Limited *replicated* evidence for statistical interactions in complex diseases; network-based approaches may be helpful (Hu et al. 2011)

(adapted from Silverman et al. 2012)

(Hayden 2010)
« Life is Complicated »)

The complexity of complex diseases

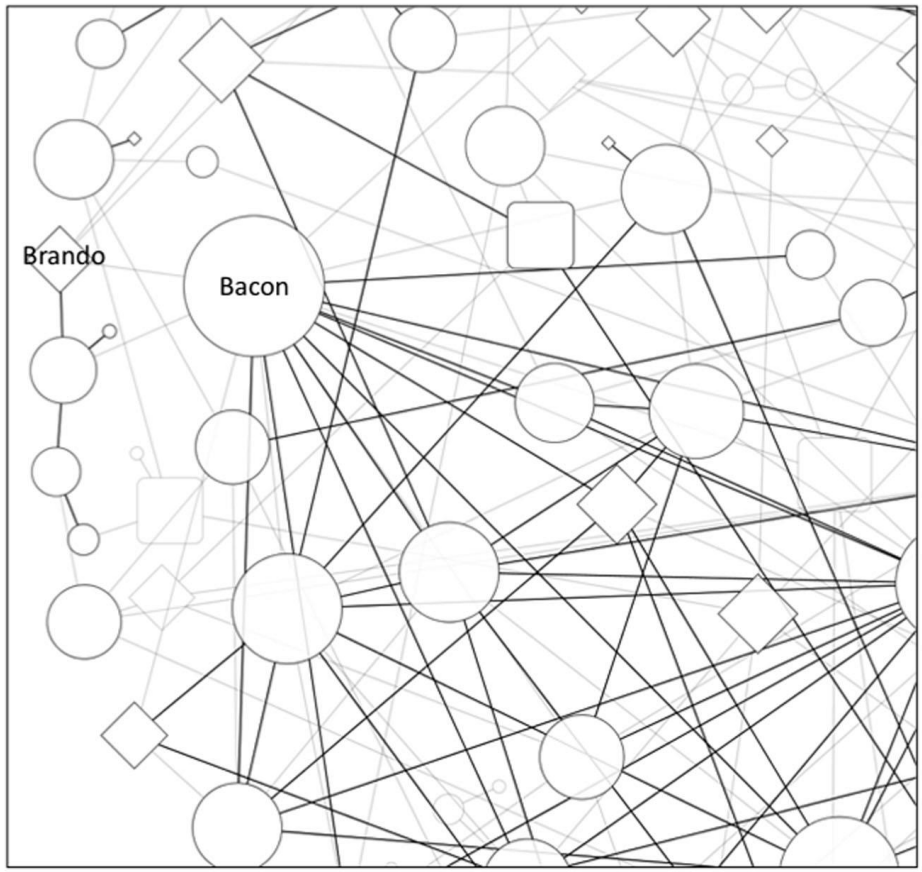


(Weiss and Terwilliger 2000)

There are likely to be *many* susceptibility genes each with combinations of *rare and common* alleles and genotypes that impact disease susceptibility primarily through *non-linear interactions* with *genetic and environmental* factors

(Moore 2008)

Disappointing results for human complex traits [… to date with SNPs]



Edges **represent small gene–gene interactions between SNPs.** Gray nodes and edges **have weaker interactions.** Circle nodes **represent SNPs that do not have a significant main effect.** The diamond nodes **represent significant main effect association.** The size of the node **is proportional to the number of connections.**

(McKinney et al. 2012)

Disappointing results for human complex traits [… to date with SNPs]

- Expectations of the first hour seem to be poorly met:
  - Different schools: heritability composition - additivity (Polderman et al. 2015: Meta-analysis of the heritability of human traits based on fifty years of twin studies vs relevance of SNP-based interactions) vs epistasis oriented starting point (incl. "Moore et al")
    - "LDAK": $E[H^2_{SNP}] = (MAF_{SNP}(1-MAF_{SNP}))^{1+\alpha} \quad w_{SNP} r_{SNP}$
      - "GCTA": $w_{SNP} = 1$; $r_{SNP} = 1$; "LDSC": $\alpha = -1$ + extra parameters
      - "LDSC" will typically have standard errors 25–100% higher than those from "GCTA" (Bulik-Sullivan 2015) (Speed et al. 2017)
    - $H^2_I = H^2 - H^2_{M,SNP1} - H^2_{M,SNP2}$ (Winham et al. 2012)

$$H^2_{M,SNP1} = \frac{1}{P\,(1-P)} \sum_{i=0}^{2} [\sum_{j=0}^{2} Prob(G_{ij})][\sum_{j=0}^{2} Prob(D|G_{ij}) Prob(G_{ij}) - P]^2$$

Disappointing results for human complex traits [… to date with SNPs]

- Expectations of the first hour seem to be poorly met
  - There is an abundance of methodological approaches
  - Interestingly, widely accepted protocol to perform a Genome-Wide Association Interaction Study (GWAIS) is still lacking
  - Possible explanations:
    - many difficulties (technical, statistical, computational) involved in performing large-scale epistasis screening
    - and in inferring biological evidence from statistical findings

Human Genetics (2019) 138:293–305
https://doi.org/10.1007/s00439-019-01987-w

REVIEW

How to increase our belief in discovered statistical interactions via large-scale association studies?

K. Van Steen[1,2] · J. H. Moore[3]

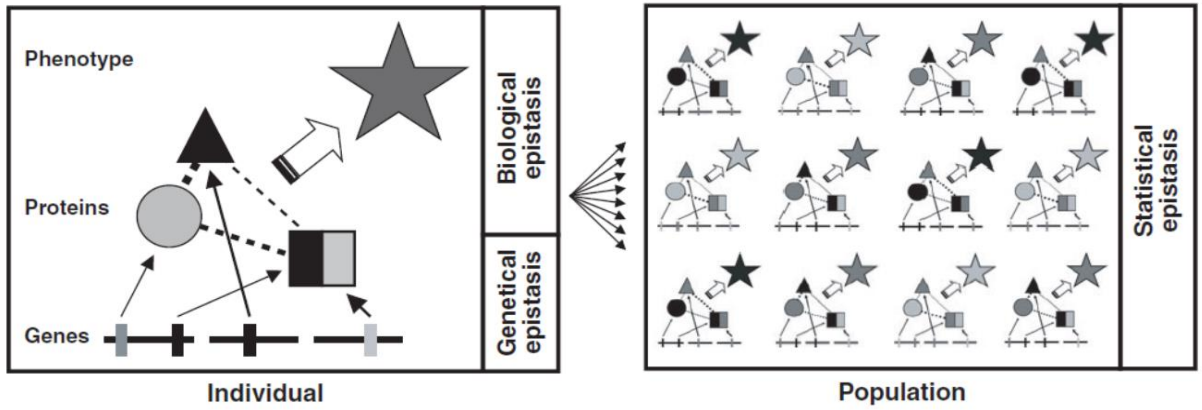
Received: 26 July 2018 / Accepted: 20 February 2019 / Published online: 6 March 2019

Abstract

The understanding that differences in biological epistasis may impact disease risk, diagnosis, or disease management stands in wide contrast to the unavailability of widely accepted large-scale epistasis analysis protocols. Several choices in the analysis workflow will impact false-positive and false-negative rates. One of these choices relates to the exploitation of particular modelling or testing strategies. The strengths and limitations of these need to be well understood, as well as the contexts in which these hold. This will contribute to determining the potentially complementary value of epistasis detection workflows and is expected to increase replication success with biological relevance. In this contribution, we take a recently introduced regression-based epistasis detection tool as a leading example to review the key elements that need to be considered to fully appreciate the value of analytical epistasis detection performance assessments. We point out unresolved hurdles and give our perspectives towards overcoming these.

What's in a name ?

- The original definition (**driven by biology**) refers to a variant or allele at one locus preventing the variant at another locus from manifesting its effect (William Bateson 1861-1926).



(Moore 2005)

- **Grown into a more general theory and applications framework** for the analysis of interactions across and between -omics strata.

Link to “the” interactome

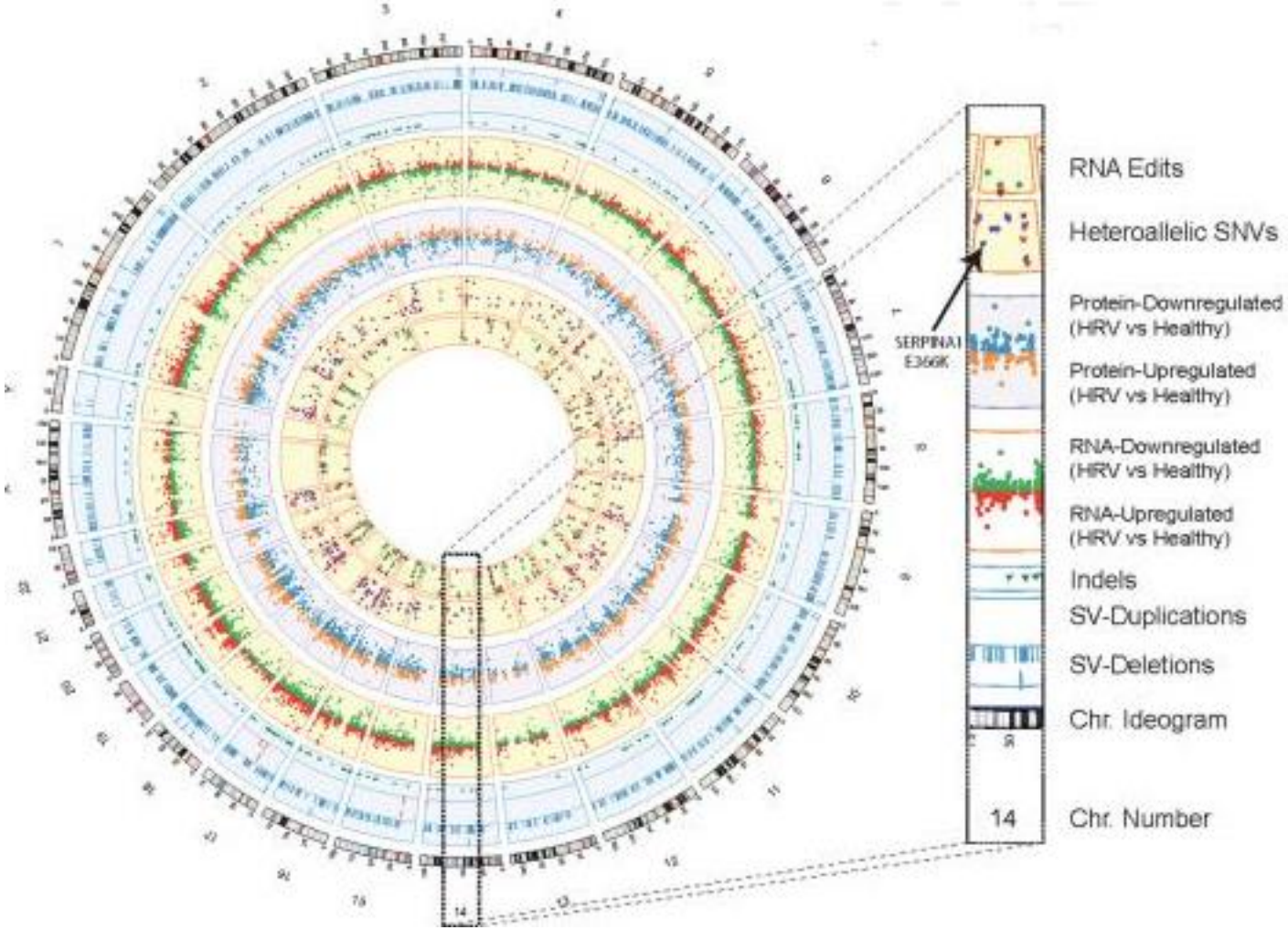
The **interactome** refers to the entire complement of interactions between DNA, RNA, proteins and metabolites within a cell. These interactions are influenced by genetic alterations and environmental stimuli. As a consequence, the interactome should be examined or considered in ***particular contexts***.

Case study: Pancreatic Cancer

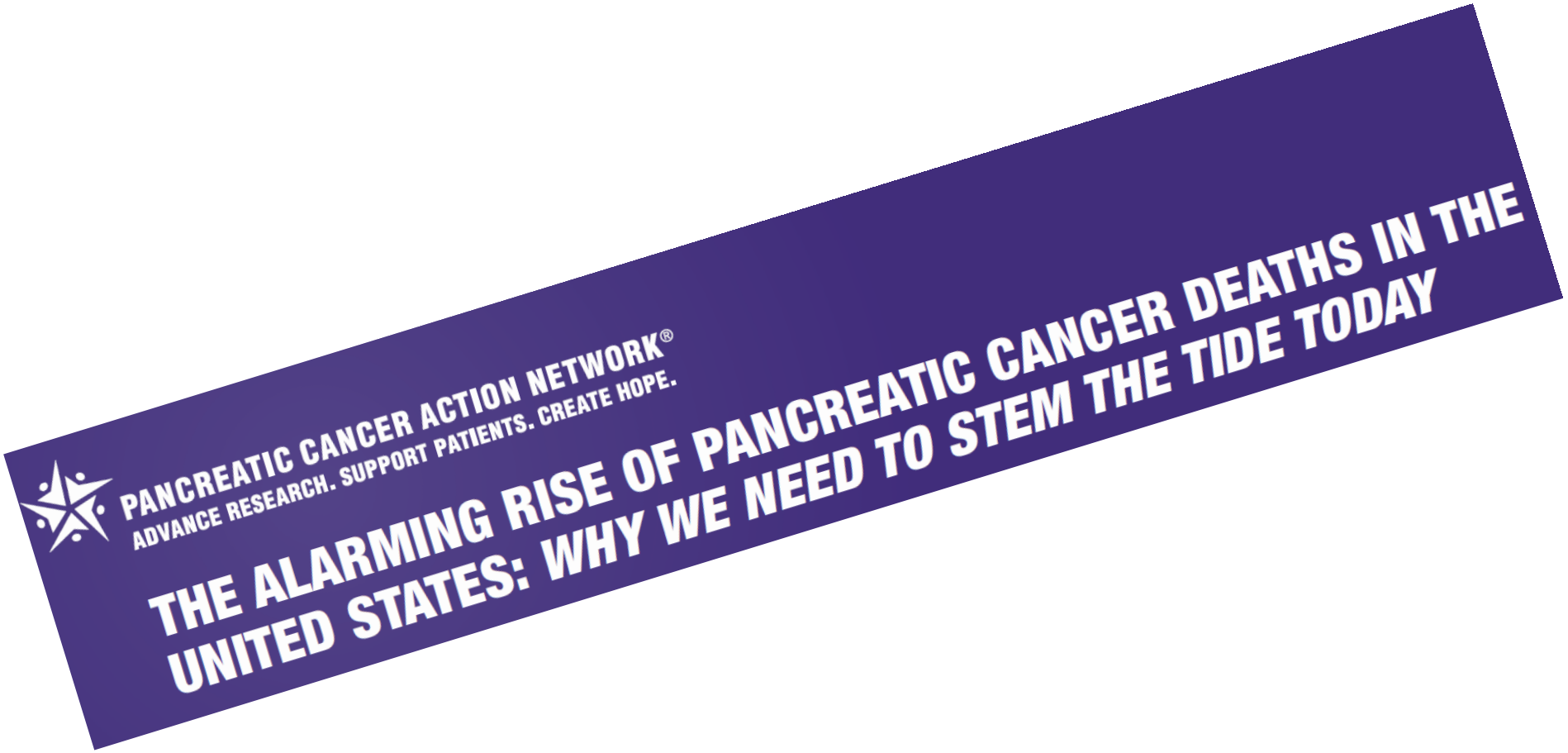
(opportunity)

Data context: Bioinformatics data availability

(Chen et al. 2012)



Disease context: complex “complex diseases”



Addressing complexity in “complex diseases” - pancreatic cancer

“Because effective systemic therapy capable of controlling the aggressive pancreatic cancer biology is currently lacking, the need for a better understanding of detailed mechanisms underlying pancreatic cancer development and progression is **URGENT***”*

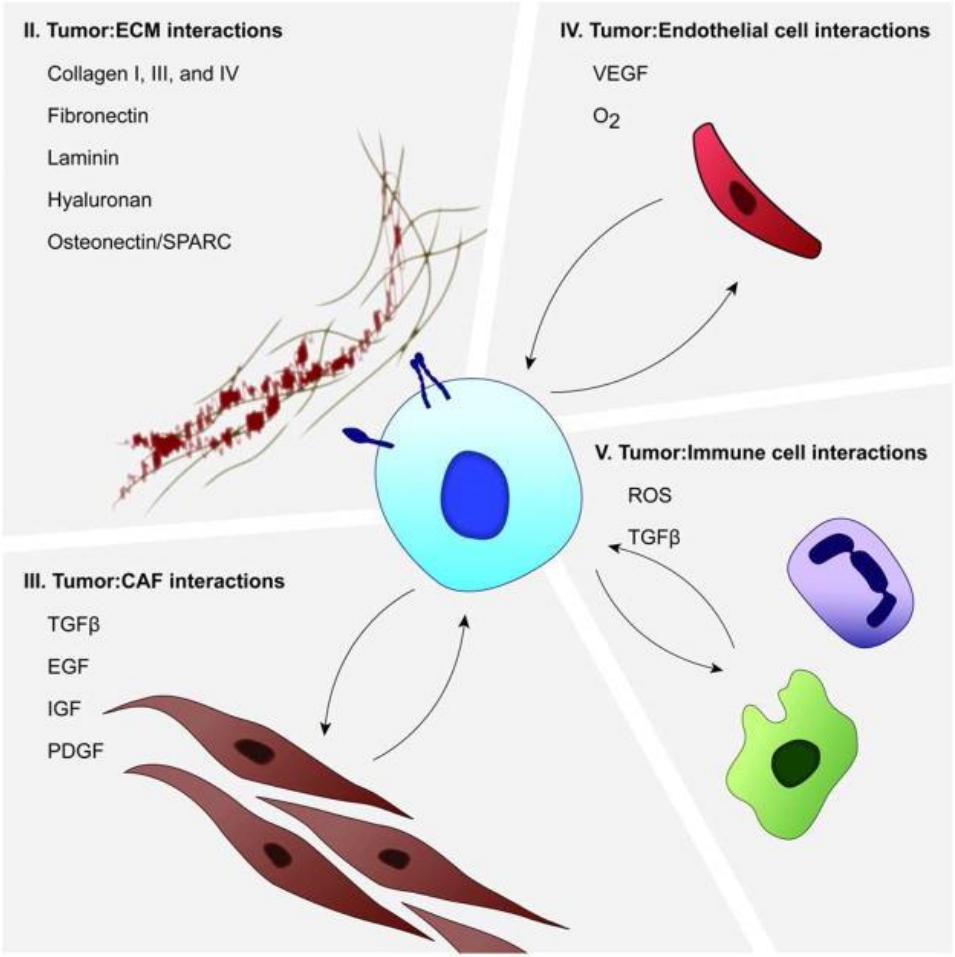
(Xie and Xie 2015)

Examples of interactions in pancreatic cancer

Tumor-stromal interactions

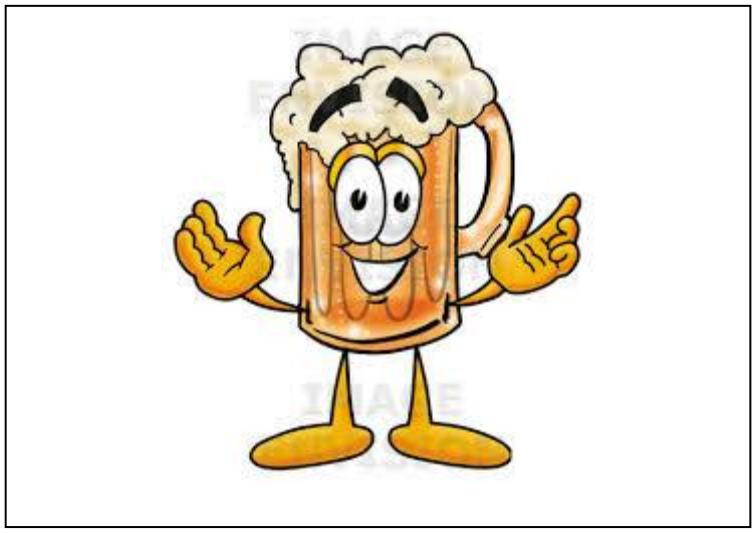
- Treatments focusing on pancreatic cancer cells alone have failed to significantly improve patient outcome over many decades
- Research efforts have now moved to understanding the pathophysiology of the stromal reaction and its role in cancer progression

(Whatcott et al. 2014)



Gene-environment interactions

(Jansen et al. 2015)



Formal definition of gene-environment interactions

- Also gene-environment interactions can be defined in a statistical or a biological way.
- A **biological gene-environment** interaction occurs when one or more genetic and one or more environmental factors participate in the same causal mechanism in the same individual (Yang and Khoury 1997; Rothman et al. 2008)
- As with gene-gene interactions, a **statistical gene-environment** interaction does not imply any inference about a specific biological mode of action. It is based on modeling a sample of individuals.

Formal definition of epistasis

- In practice, when modeling or testing, it may only be possible to detect **effect modification** from real-life data and not **interaction**, or interaction but not effect modification.
- Whereas an interaction effect for "exposures" X_1 and X_2 relies on a symmetric role for both X_1 and X_2 ,
  an effect modification relies on a conditioning argument (for instance on X_2) (VanderWeele 2009a)
- The distinction between both effect types is often concealed in regression analysis ... (Robins et al. 2000; North et al. 2005)

Comparison between gene–gene and gene–environment issues

- Conceptually many similar issues in terms of definition and mathematical modelling.
- In practice, some clear differences emerge.
- For G x E:
  - We generally have to decide which environments to measure / test; these are typically only a few (often < 100)
  - Measurement error (lifestyle) and unknown confounding
  - Risk estimation, important for screening strategies and public health interventions

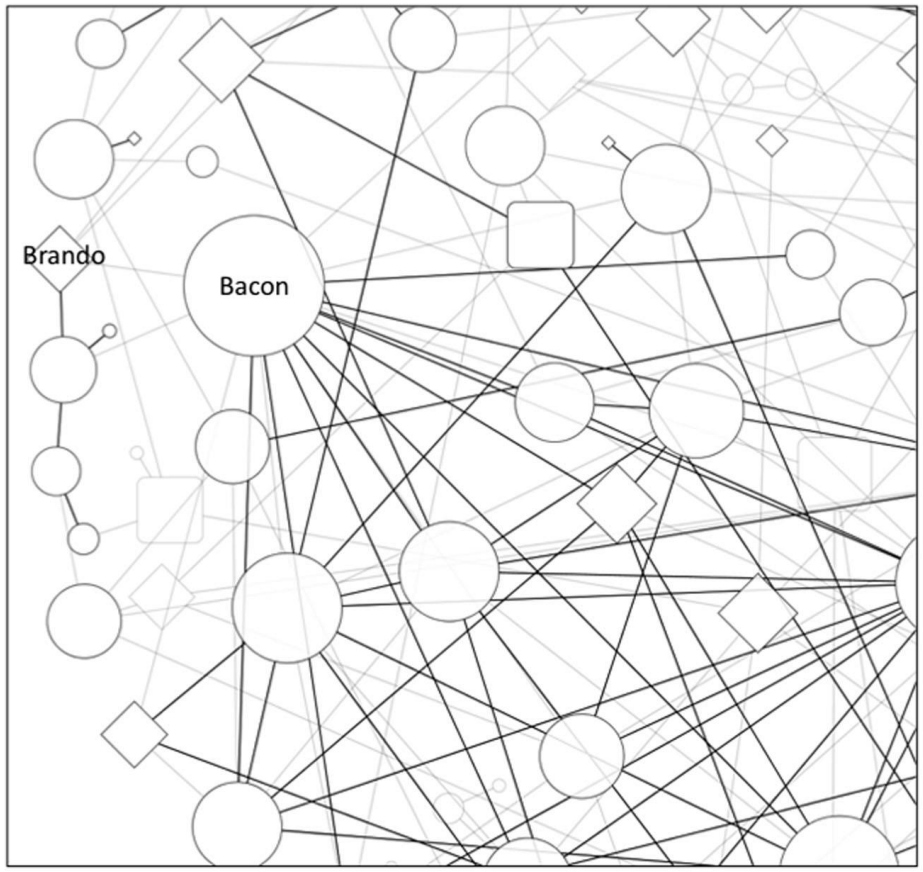
(courtesy slide EUPancreas WG2 Training School, Antwerp, 2016)

Comparison between gene–gene and gene–environment issues

- For G x G
  - Assuming we have GWAS data, we have already measured the genetic factors of interest
  - Adequate error rates (except for newer sequencing technologies)
  - (Hundred) thousands of variants
  - Higher-order interactions may reflect the complex biological wiring of complex diseases (whereas G x E often restricts attention to pairwise interactions)

(courtesy slide EUPancreas WG2 Training School, Antwerp, 2016)

Looking for higher-order interactions



Edges represent small gene–gene interactions between SNPs.

Gray nodes and edges have weaker interactions.
Circle nodes represent SNPs that do not have a significant main effect.
The diamond nodes represent significant main effect association.

The size of the node is proportional to the number of connections.

(McKinney et al 2012)

Some references

Published in final edited form as:
Hum Genet. 2012 October ; 131(10): 1591–1613. doi:10.1007/s00439-012-1192-0.

Challenges and Opportunities in Genome-Wide Environmental Interaction (GWEI) studies

Hugues Aschard[1], **Sharon Lutz**[2,*], **Bärbel Maus**[3,4,*], **Eric J. Duell**[5], **Tasha Fingerlin**[2], **Nilanjan Chatterjee**[6], **Peter Kraft**[1,7], and **Kristel Van Steen**[3,4]

Hum Genet (2014) 133:1343–1358
DOI 10.1007/s00439-014-1480-y

REVIEW PAPER

Practical aspects of genome-wide association interaction analysis

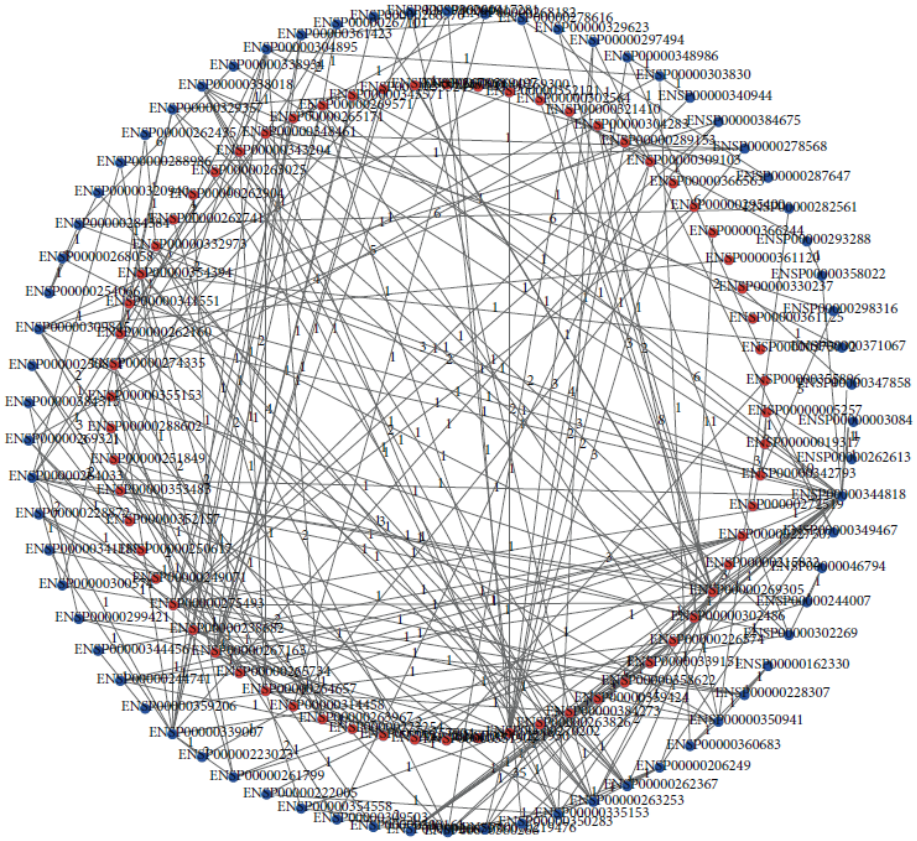
Elena S. Gusareva · Kristel Van Steen

Protein-protein interactions

(Yuan et al. 2015)

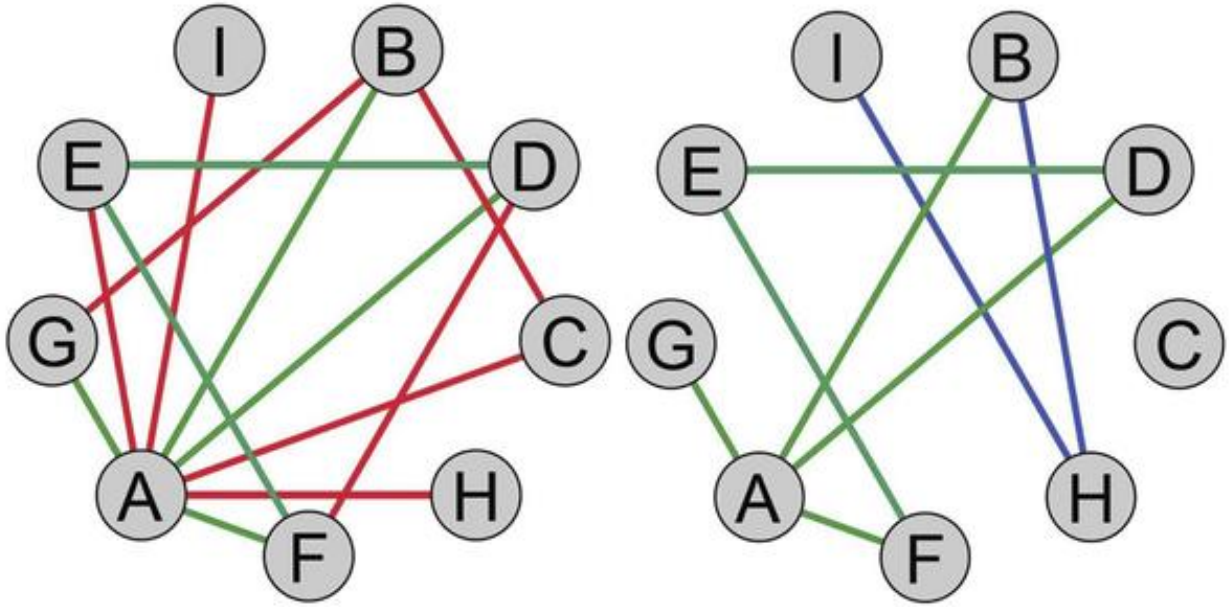
A graph consisting of 2,080 shortest paths:

- The nodes on the inner circle (red nodes) represent 65 PC-related genes.
- The nodes on the outer circle (blue nodes) represent 69 shortest path genes.
- The numbers on the edges represent the weights of the edges.



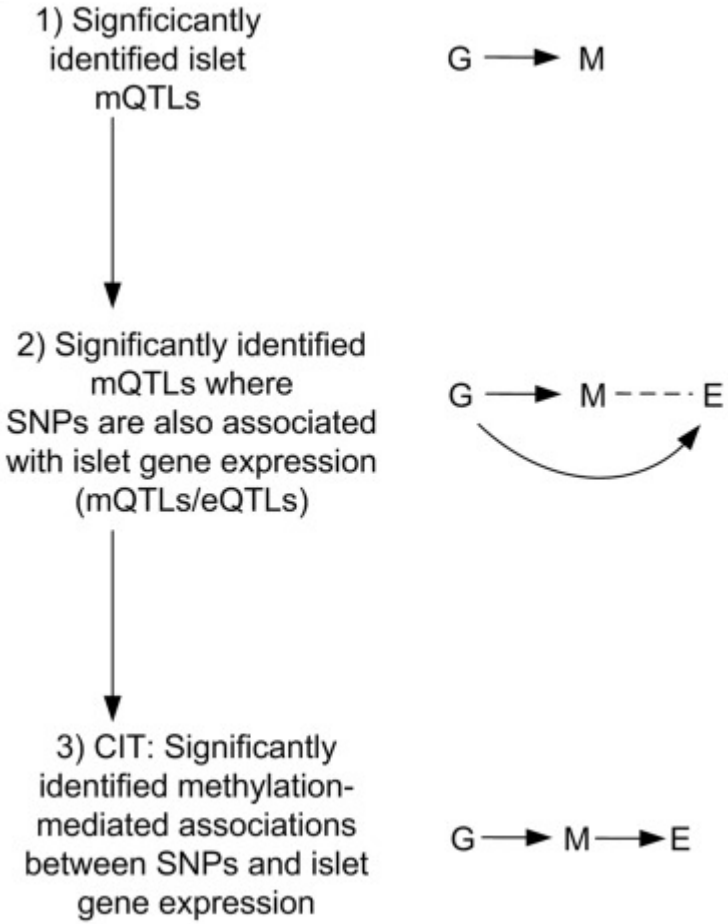
Gene-coexpression networks

(Anglani et al. 2014)



- Healthy condition on the left and disease-affected tissue on the right. Green links remain unchanged in the two phenotypes
- Red connections are loss from healthy to cancer network
- Blue edges are novel connections in the cancer tissue

Genetic-epigenetic mechanistic interactions (pancreatic islets)



cis-mQTLs:	*trans-mQTLs*:
67,438 SNP-CpG pairs (36,783 unique SNPs and 11,735 unique CpGs)	2,562 SNP-CpG pairs (1,465 unique SNPs and 383 unique CpGs)

cis-mQTLs/eQTLs:	*trans-mQTLs/eQTLs*:
2,805 SNP-CpG-mRNA combinations (243 unique SNPs, 276 unique CpGs and 46 unique mRNA transcripts)	37 SNP-CpG-mRNA combinations (22 unique SNPs, 12 unique CpGs and 8 unique mRNA transcripts)

cis-mQTLs/eQTLs:	*trans-mQTLs/eQTLs*:
28 SNP-CpG-mRNA combinations (17 unique SNPs, 14 unique CpGs and 5 unique mRNA transcripts)	4 SNP-CpG-mRNA combinations (2 unique SNPs, 4 unique CpGs and 1 unique mRNA transcript)

Gene-gene interactions using SNPs?

(Olsson et al. 2014)

GWAS Catalogue – "Pancreas Cancer"

Wolpin BM (PMID: 25086665)	2014-08-03	Nat Genet	Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer.	Pancreatic cancer	10

Initial sample description	1,582 European ancestry cases, 5,203 European ancestry controls
Initial ancestry (country of recruitment)	6785 European (U.S., Australia, France, Germany, Netherlands, Denmark, Finland, Norway, Sweden, U.K., Greece, Italy, Spain)
Replication sample description	6,101 European ancestry cases, 9,194 European ancestry controls
Replication ancestry (country of recruitment)	15295 European (Canada, U.S., France, Germany, Netherlands, Denmark, Finland, Norway, Sweden, U.K., Greece, Italy, Spain)
Platform [SNPs passing QC]	Illumina [608202]

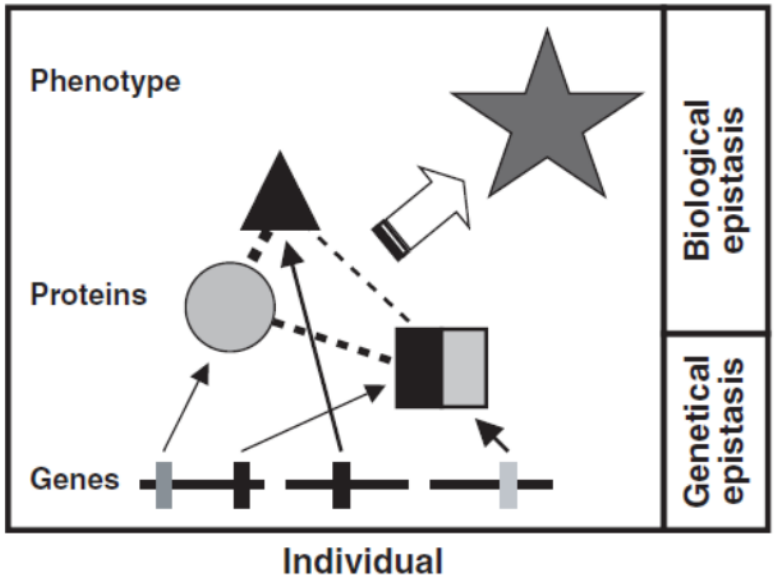
(http://www.ebi.ac.uk/gwas/search?query=pancreas%20cancer#study)

Model-Based Multifactor Dimensionality Reduction

a decade's work

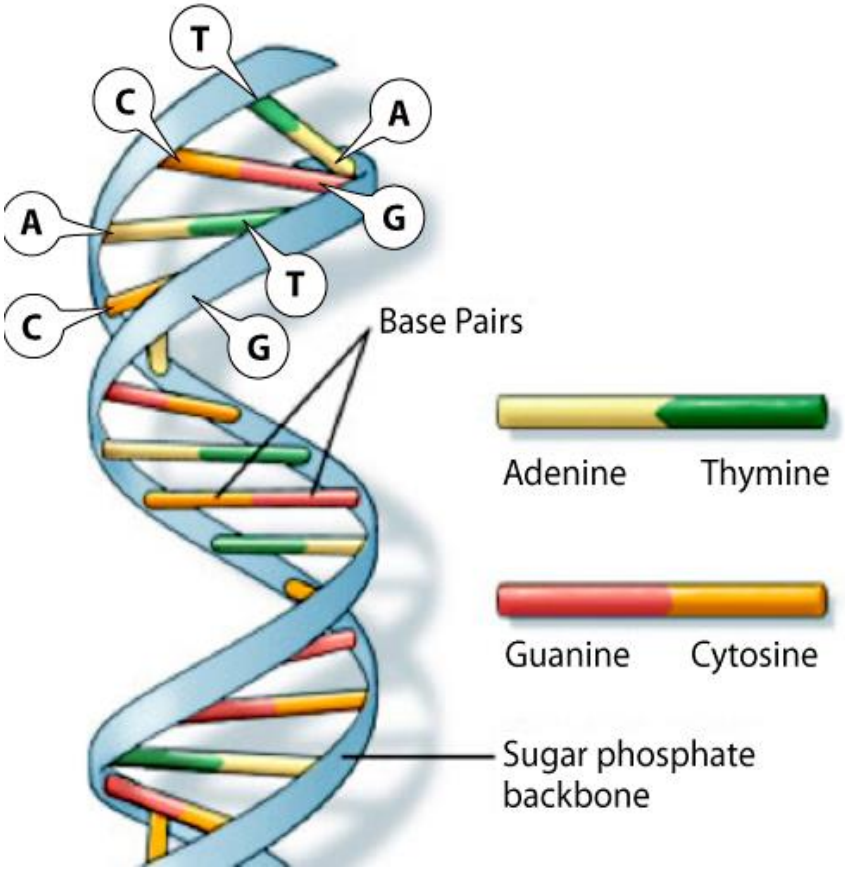
Recall DNA-DNA interactions: biological viewpoint

- Two or more DNA variations may interact either directly to change transcription or translation levels, or indirectly by way of their protein product (to alter disease risk separate from their independent effects)



(Moore 2005)

Recall DNA-DNA interactions: non-biological viewpoint



- Epistasis (**driven by statistics**) is expressed in terms of deviations from a model of additive multiple effects.
- This might be on either a linear or logarithmic scale, which implies different definitions (Ronald Fisher 1890-1962).

(Logistic) Regression

- Alternatively, we can assume additive effects of each allele at each locus, leading to a single interaction term (instead of 4 next!)

		Locus H	
Locus G	2	1	0
2	$\beta_0 + 2\beta_G + 2\beta_H + 4\beta$	$\beta_0 + 2\beta_G + \beta_H + 2\beta$	$\beta_0 + 2\beta_G$
1	$\beta_0 + \beta_G + 2\beta_H + 2\beta$	$\beta_0 + \beta_G + \beta_H + \beta$	$\beta_0 + \beta_G$
0	$\beta_0 + 2\beta_H$	$\beta_0 + \beta_H$	β_0

- This corresponds in statistical analysis packages to the model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_G X_1 + \beta_H X_2 + \beta X_1 X_2$$

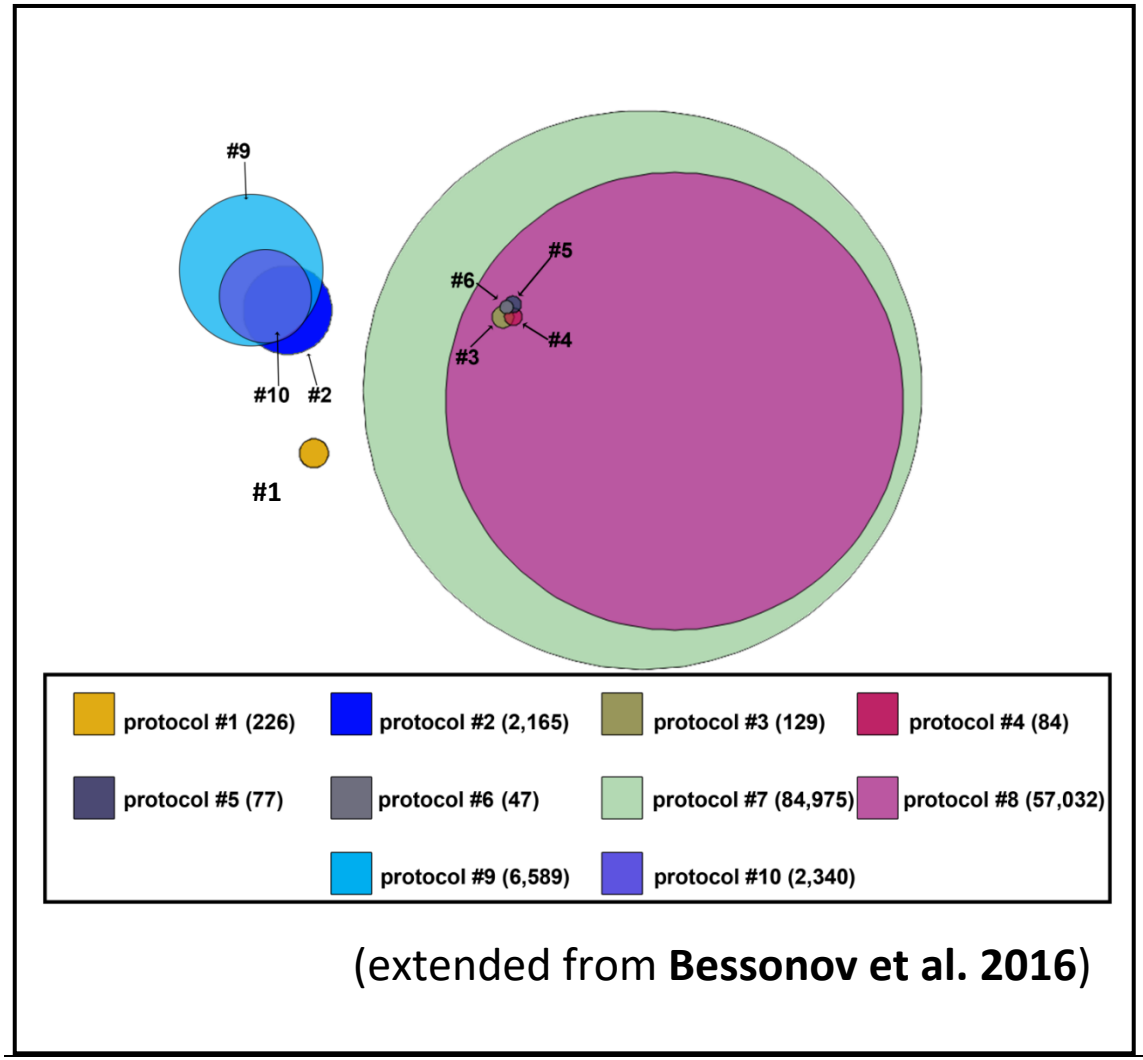
and **dosage encoding for X1 and X2**.

(Logistic) Regression

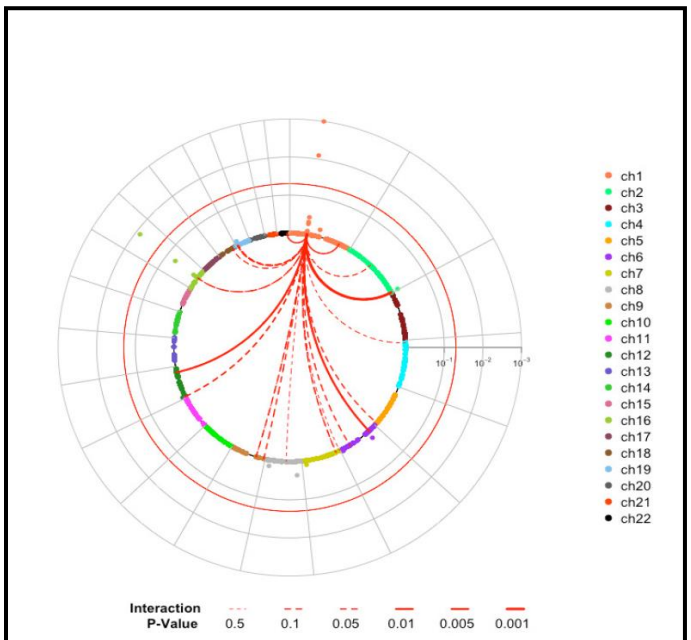
- Most general saturated (9 parameter) genotype model allows all 9 penetrances to take different values
- Log odds is modelled in terms of a baseline effect (β_0), main effects of locus *G* (β_{G1}, β_{G2}), main effects of locus *H* (β_{H1}, β_{H2}), 4 int. terms
- This corresponds in statistical analysis packages to **encoding X1, X2 (0,1,2) as a "factor"**

	Locus H		
Locus G	2	1	0
2	$\beta_0+\beta_{G2}+\beta_{H2}+\beta_{22}$	$\beta_0+\beta_{G2}+\beta_{H1}+\beta_{21}$	$\beta_0+\beta_{G2}$
1	$\beta_0+\beta_{G1}+\beta_{H2}+\beta_{12}$	$\beta_0+\beta_{G1}+\beta_{H1}+\beta_{11}$	$\beta_0+\beta_{G1}$
0	$\beta_0+\beta_{H2}$	$\beta_0+\beta_{H1}$	β_0

Importance of SNP encoding scheme (Ankylosing Spondylitis; WTCCC2 - ~2000 cases + 5000 controls)



(extended from **Bessonov et al. 2016**)

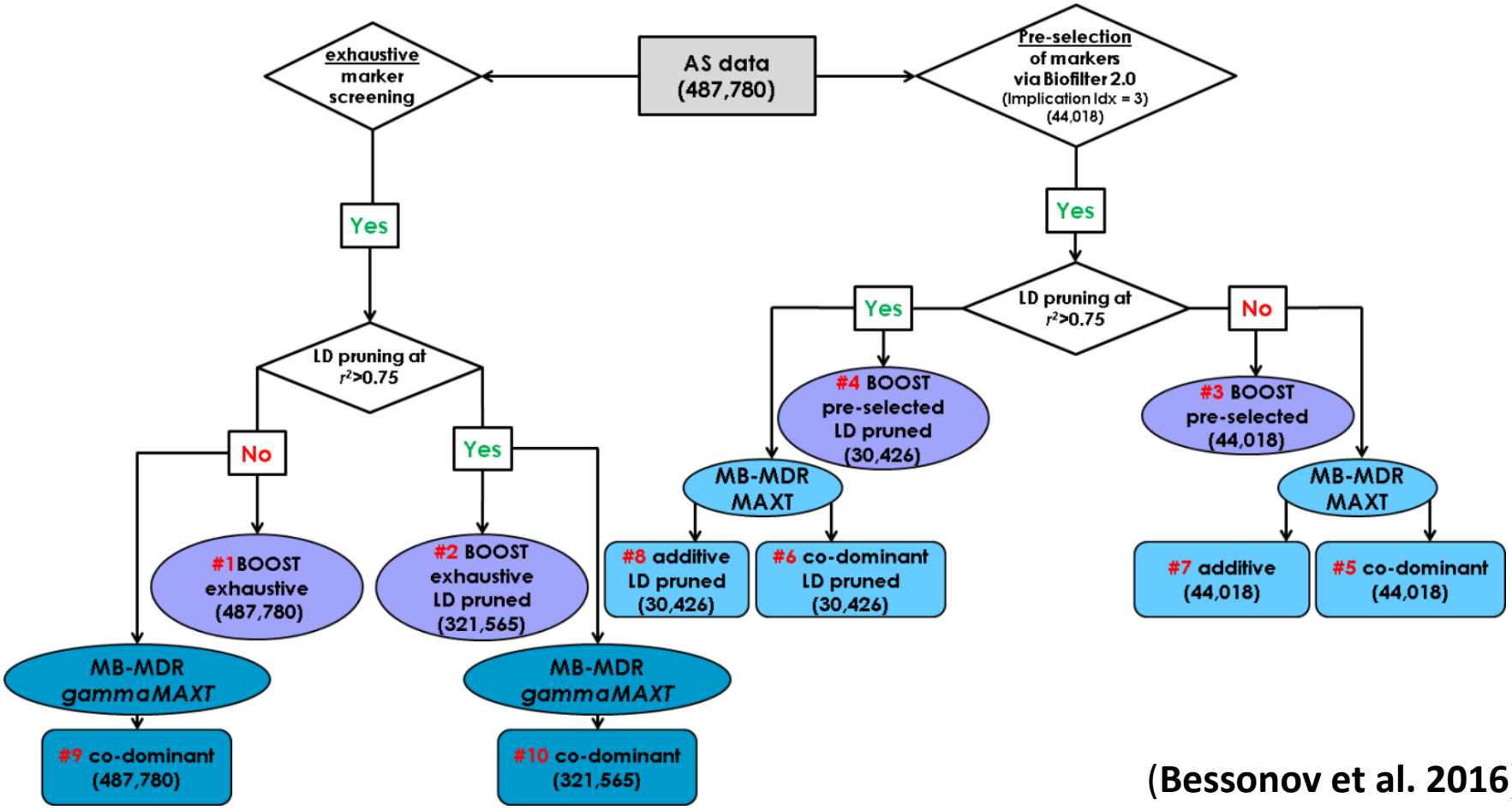


MOGPLOT

Manhattan plot (main effects) over GRAIL plot (interactions)

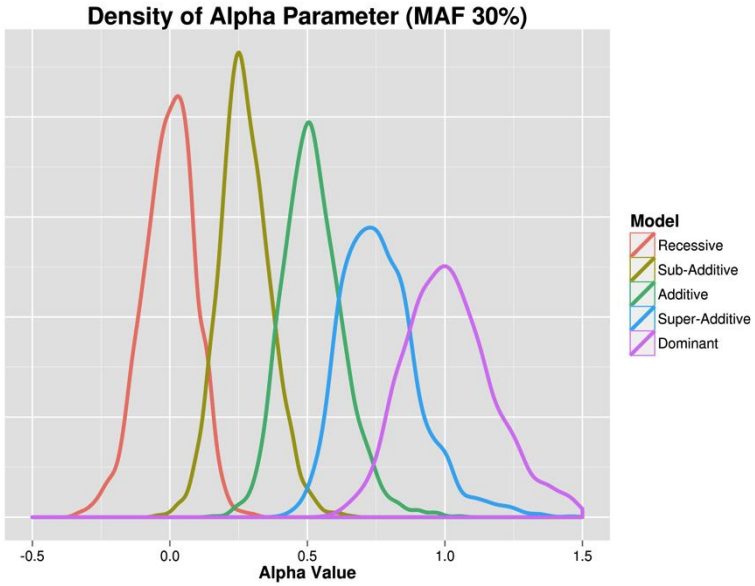
(**Van Lishout et al. 2016**)

Stability of results: analytic REPLICATION



(**Bessonov et al. 2016**)

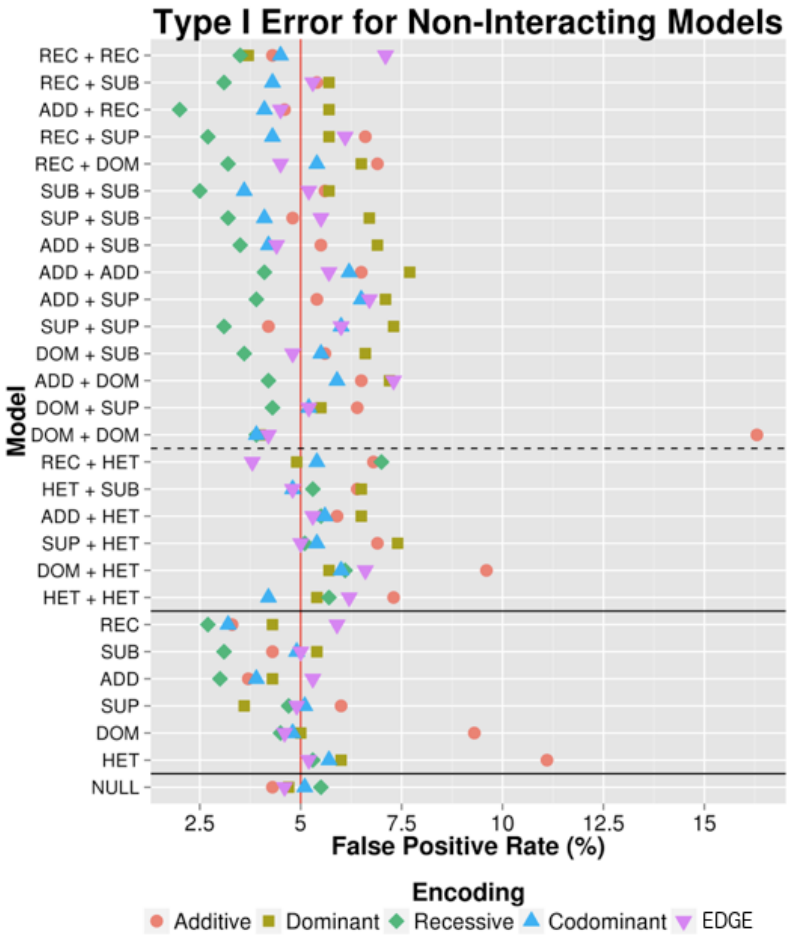
Importance of SNP encoding scheme (**Hall et al. 2019** – submitted)



"Biological Action"	Homozygous Referent	Heterozygous	Homozygous Alternate
EDGE	0	**α**	1

$$Y \sim \beta_{Het}SNP_{Het} + \beta_{HA}SNP_{HA}$$

$$\alpha = \beta_{Het} / \beta_{HA}$$



Encoding (Hall et al. 2020 – under review)

Table 1. Examples of possible proportional genotype risk underlying genetic loci

Biological Action	Homozygous Referent (AA)	Heterozygous (Aa)	Homozygous Alternate (aa)
Recessive (REC)	0%	0%	100%
Sub-Additive (SUB	0%	25%	100%
Additive (ADD)	0%	50%	100%
Super-Additive (SUP)	0%	75%	100%
Dominant (DOM)	0%	100%	100%

Binary genomic markers???

Bioinformatics, 33(12), 2017, 1820–1828
doi: 10.1093/bioinformatics/btx071
Advance Access Publication Date: 14 February 2017
Original Paper

Genetics and population analysis

Genome-wide genetic heterogeneity discovery with categorical covariates

Felipe Llinares-López[1,2,*,†], Laetitia Papaxanthos[1,2,*,†], Dean Bodenham[1,2], Damian Roqueiro[1,2], COPDGene Investigators[3] and Karsten Borgwardt[1,2,*]

[1]Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland, [2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland and [3]COPDGene® Study

Note about analytic comparisons: conceptual differences

- **Regression modelling:** predicting the average response from covariates [maximizing predictive power] *versus* characterizing relationships [control of confounders, account for effect modifiers]

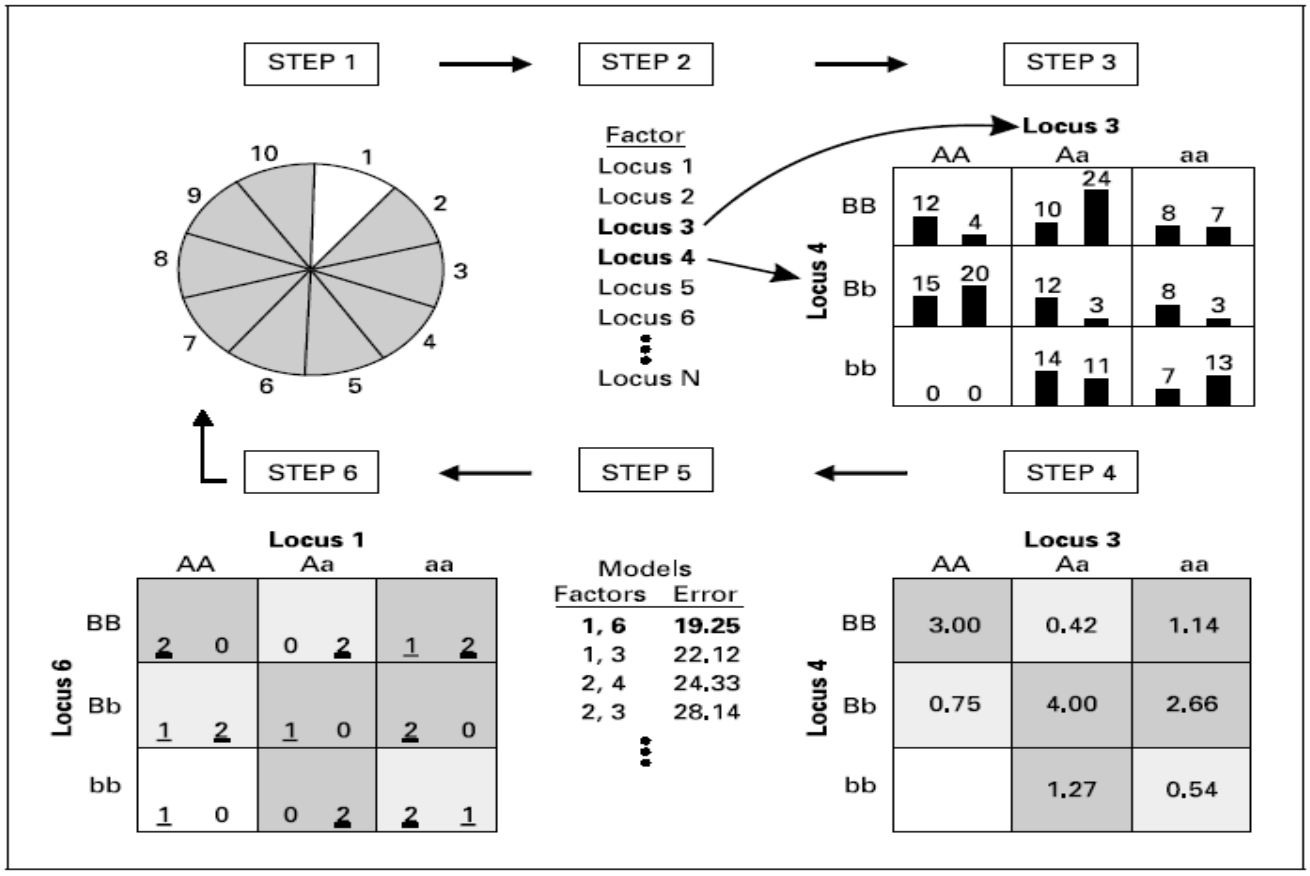
- **Deep neural networks** have been recognized as some of the best performing machine learning methods:

Methods	Accuracy
Deep learning	**68.78**
RF	55.85
LR	67.07
Naïve Bayes	62.68
GBM	65.85

(Uppu et al. 2016)

Historical notes about MB-MDR

- Start: Multifactor Dimensionality Reduction by MD Ritchie et al (2001)



Which dimensions are reduced?

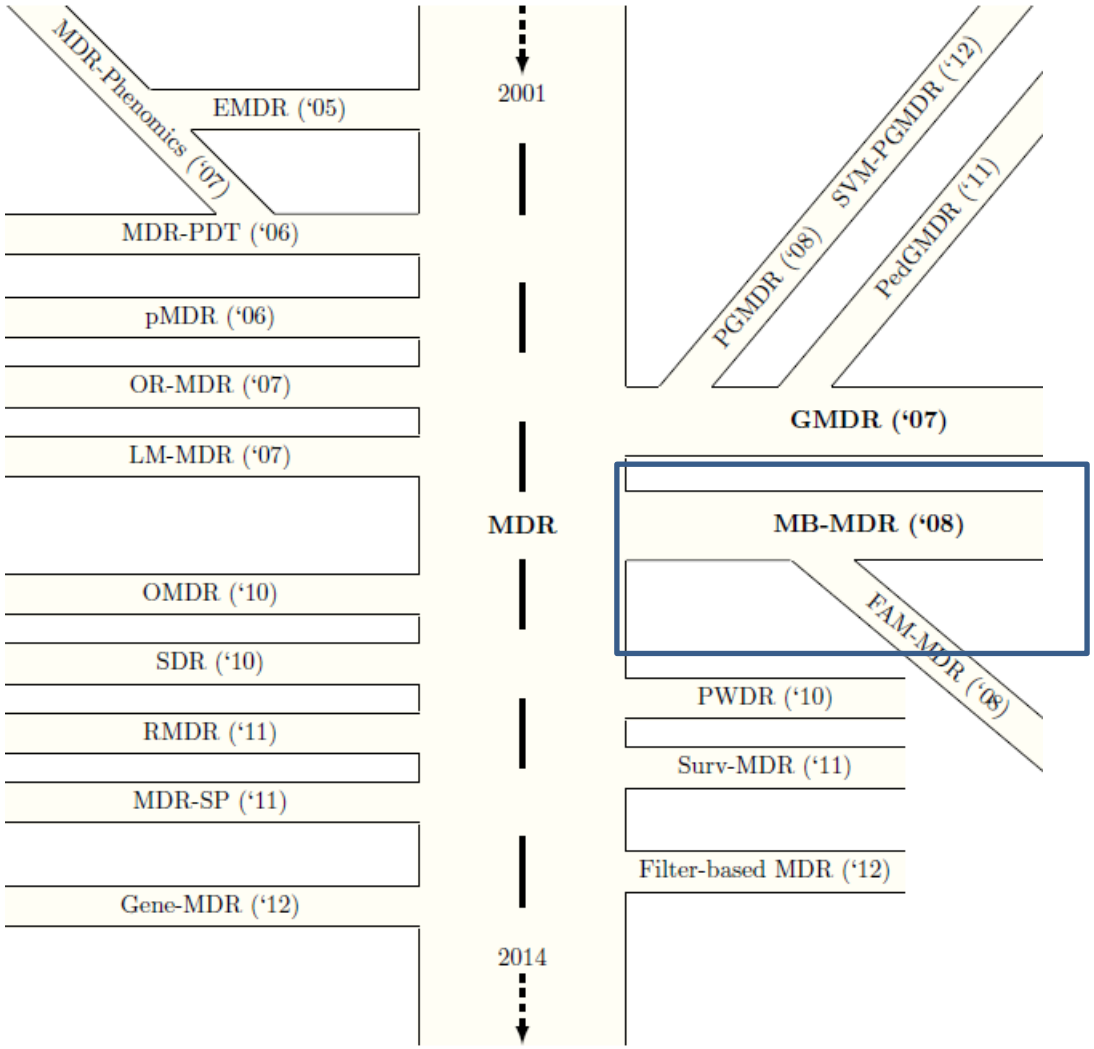
- The estimated degrees of freedom for MDR and LR using K=1, 2 and 3 factors (standard errors in parentheses). LR exact refers to the asymptotic exact degrees of freedom

	Number of Factors K *		
Method	**1**	**2**	**3**
MDR	1.9 (0.13)	5.6 (0.20)	17.4 (0.37)
LR	2.1 (0.4)	8.0 (0.26)	26.8 (0.53)
LR exact	2	8	26

(Park and Hastie 2007)

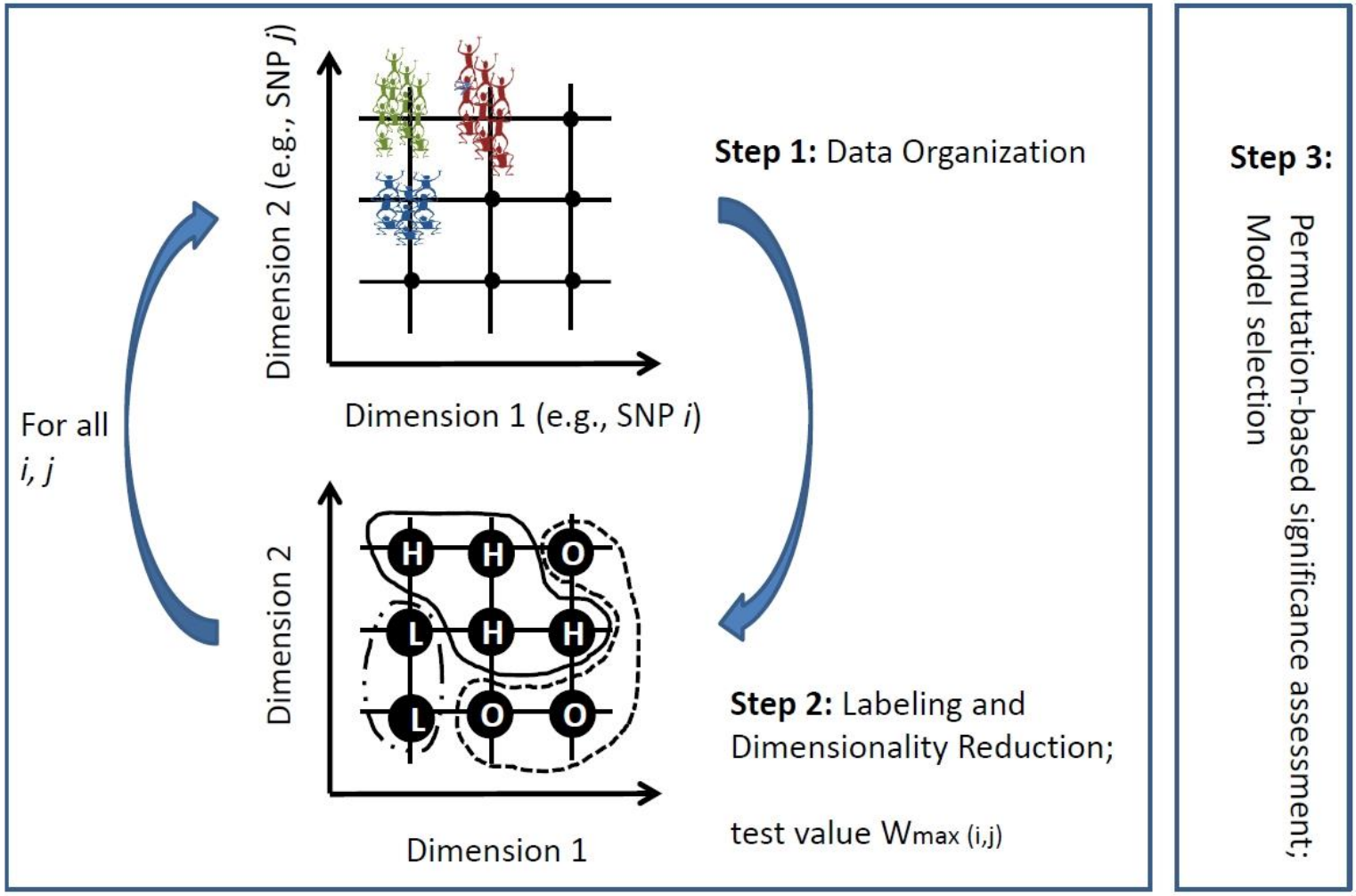
* ”K-way” interaction

Several MDR roads lead to Rome



(Gola et al. 2015)

Model-Based Multifactor Dimensionality Reduction (MB-MDR)



Current versions of MB-MDR

- Computation time is invested in
  - optimal **association tests** to label multi-locus genotype combinations and
  - in statistically valid permutation-based methods to assess **joint statistical significance** of multiple SNP pairs
- Labels are related to substantially improve/worsen trait values (H/L). In case there is **no** such **evidence**, the multi-locus label is not forced to be H or L (but will be O).
- In the **presence of main effects**, MB in MB-MDR ensures false positive control at 5%

Global versus specific modeling

- Model-Based MDR by Calle et al (2008a,b)

Table 3. MB-MDR first step analysis for interaction between SNP 40 and SNP 252 in the bladder cancer study

SNP 40 x SNP 252 genotypes	Cases	Controls	OR	p-value	Category
c1 = (0,0)	88	77	1.01	0.9303	0
c2 = (0,1)	102	114	0.73	0.0562	L
c3 = (0,2)	38	34	0.98	1.0000	0
c4 = (1,0)	50	59	0.76	0.1229	0
c5 = (1,1)	96	37	2.68	0.0000	H
c6 = (1,2)	18	28	0.55	0.0675	L
c7 = (2,0)	12	6	1.99	0.3399	0
c8 = (2,1)	14	18	0.67	0.3668	0
c9 = (2,2)	6	6	0.84	1.0000	0

H: High risk; L: Low risk; 0: No evidence

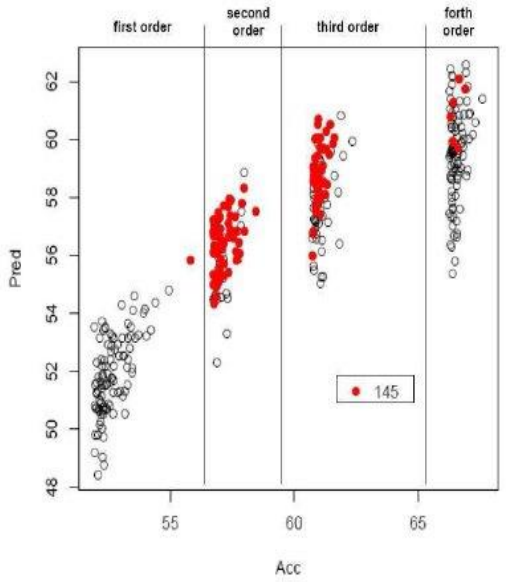
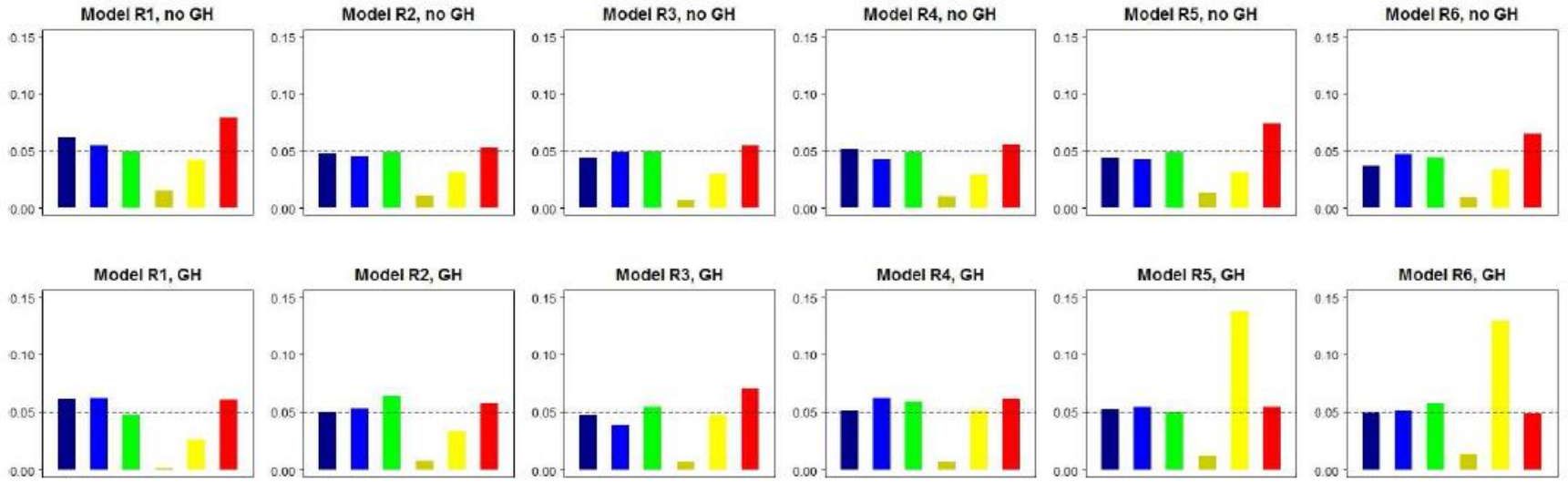


Fig. 1. Average Balanced Training accuracy (Acc) versus Average Balanced Predictive accuracy (Pred) for the 100 models with higher balanced training accuracy for the whole sample. First, second, third and forth order interactions are considered.

Comparative performance of 2-locus MB-MDR

- False positives

  (example: pure epistasis scenario's; unpublished - 2010)



BOOST (dark blue)

EpiCruncher optimal options (light blue)

MB-MDR (green)

PLINK epistasis (dark yellow)

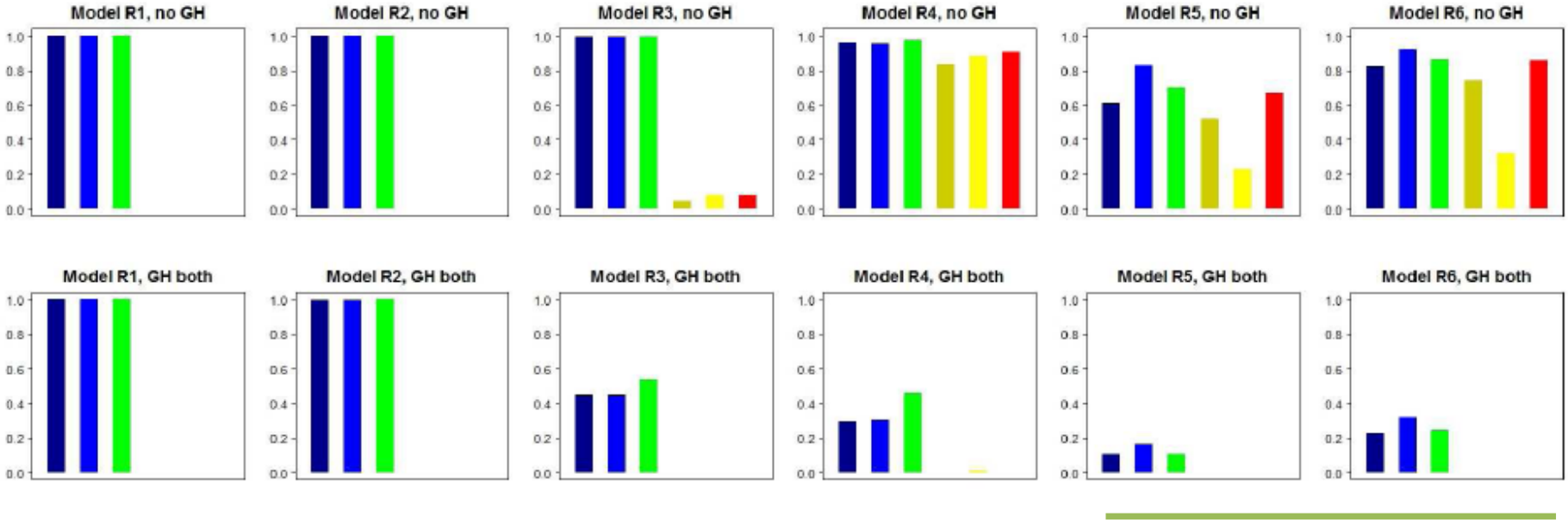
PLINK fast epistasis (light yellow)

EPIBLASTER (red)

Comparative performance of 2-locus MB-MDR

- Power performance

(example: pure epistasis scenario's; unpublished - 2010)



BOOST (dark blue)

EpiCruncher optimal options (light blue)

MB-MDR (green)

PLINK epistasis (dark yellow)

PLINK fast epistasis (light yellow)

EPIBLASTER (red)

Performance summary

Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
False Positives (%)											
MB	**MDR**	**MB**	**MDR**	**MB**	**MDR**	**MB**	**MDR**	**MB**	**MDR**	**MB**	**MDR**
6	9	4	5	6	17	5	13	5	21	5	23
Power (%)											
MB	**MDR**	**MB**	**MDR**	**MB**	**MDR**	**MB**	**MDR**	**MB**	**MDR**	**MB**	**MDR**
100	99	100	100	100	95	100	93	93	62	97	73
MB-MDR (MB): p_c = 0.1, T = *H* vs *L* test; MDR: default options, screening over 1-5 order models											

(**Cattaert et al. 2011**)

Model 1, $p = 0.5$

	BB	Bb	bb
AA	0	0.1	0
Aa	0.1	0	0.1
aa	0	0.1	0

Model 3, $p = 0.25$

	BB	Bb	bb
AA	0.08	0.07	0.05
Aa	0.1	0	0.1
aa	0.03	0.1	0.04

Model 5, $p = 0.1$

	BB	Bb	bb
AA	0.07	0.05	0.02
Aa	0.05	0.09	0.01
aa	0.02	0.01	0.03

Model 2, $p = 0.5$

	BB	Bb	bb
AA	0	0	0.1
Aa	0	0.05	0
aa	0.1	0	0

Model 4, $p = 0.25$

	BB	Bb	bb
AA	0	0.01	0.09
Aa	0.04	0.01	0.08
aa	0.07	0.09	0.03

Model 6, $p = 0.1$

	BB	Bb	Bb
AA	0.09	0.001	0.02
Aa	0.08	0.07	0.005
aa	0.003	0.007	0.02

Performance

Original Paper

Human Heredity

Hum Hered 2015;79:157–167
DOI: 10.1159/000381286

Published online: July 28, 2015

Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis

Ramouna Fouladi Kyrylo Bessonov François Van Lishout Kristel Van Steen

Systems and Modeling Unit, Montefiore Institute, and Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium


LEARNING
LEARNING
LEARNING
EMBRACE COMPLEXITY
(http://thebusyba.com)

Learning from data

- **Calle**, M. L., Urrea, V., Vellalta, G., Malats, N. & Van Steen, K. (2008a) Model-Based Multifactor Dimensionality Reduction for detecting interactions in high-dimensional genomic data. Technical Report No. 24, Department of Systems Biology, Universitat de Vic, http://www.recercat.net/handle/2072/5001 **[technical report, first mentioning MB-MDR]**
- **Calle** M, Urrea V, Malats N, Van Steen K. (2008) Improving strategies for detecting genetic patterns of disease susceptibility in association studies – Statistics in Medicine 27 (30): 6532-6546 **[MB-MDR with Wald tests and MAF dependent empirical test distributions]**
- **Calle** ML, Urrea V, Van Steen K (2010) mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits. Bioinformatics Applications Note 26 (17): 2198-2199 **[first MB-MDR software tool, in R]**

- **Cattaert** T, Urrea V, Naj AC, De Lobel L, De Wit V, Fu M, Mahachie John JM, Shen H, Calle ML, Ritchie MD, Edwards T, Van Steen K. (2010) FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals, PLoS One 5 (4). **[first implementation of MB-MDR in C++, with improved features on multiple testing correction and improved association tests + recommendations on handling family-based designs]**

- **Cattaert** T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, Urrea V, Ritchie MD, Van Steen K (2010) Model-Based Multifactor Dimensionality Reduction for detecting epistasis in case-control data in the presence of noise *(invited paper)*. Ann Hum Genet. 2011 Jan;75(1):78-89 **[detailed study of C++ MB-MDR performance with binary traits]**
- **Mahachie John** JM, Cattaert T, De Lobel L, Van Lishout F, Empain A, Van Steen K (2011) Comparison of genetic association strategies in the presence of rare alleles. BMC Proceedings, 5(Suppl 9):S32 **[first explorations on C++ MB-MDR applied to rare variants]**
- **Mahachie John** JM, Cattaert T, Van Lishout F, Van Steen K (2011) Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data. European Journal of Human Genetics 19, 696-703. **[detailed study of C++ MB-MDR performance with quantitative traits]**
- **Van Steen** K (2011) Travelling the world of gene-gene interactions *(invited paper)*. Brief Bioinform 2012, Jan; 13(1):1-19. **[positioning of MB-MDR in general epistasis context]**
- Aschard et al. … **Van Steen K** (2012) Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Hum Genet. 2012 Oct;131(10):1591-613 **[connection between GxG and GxE interaction detection]**

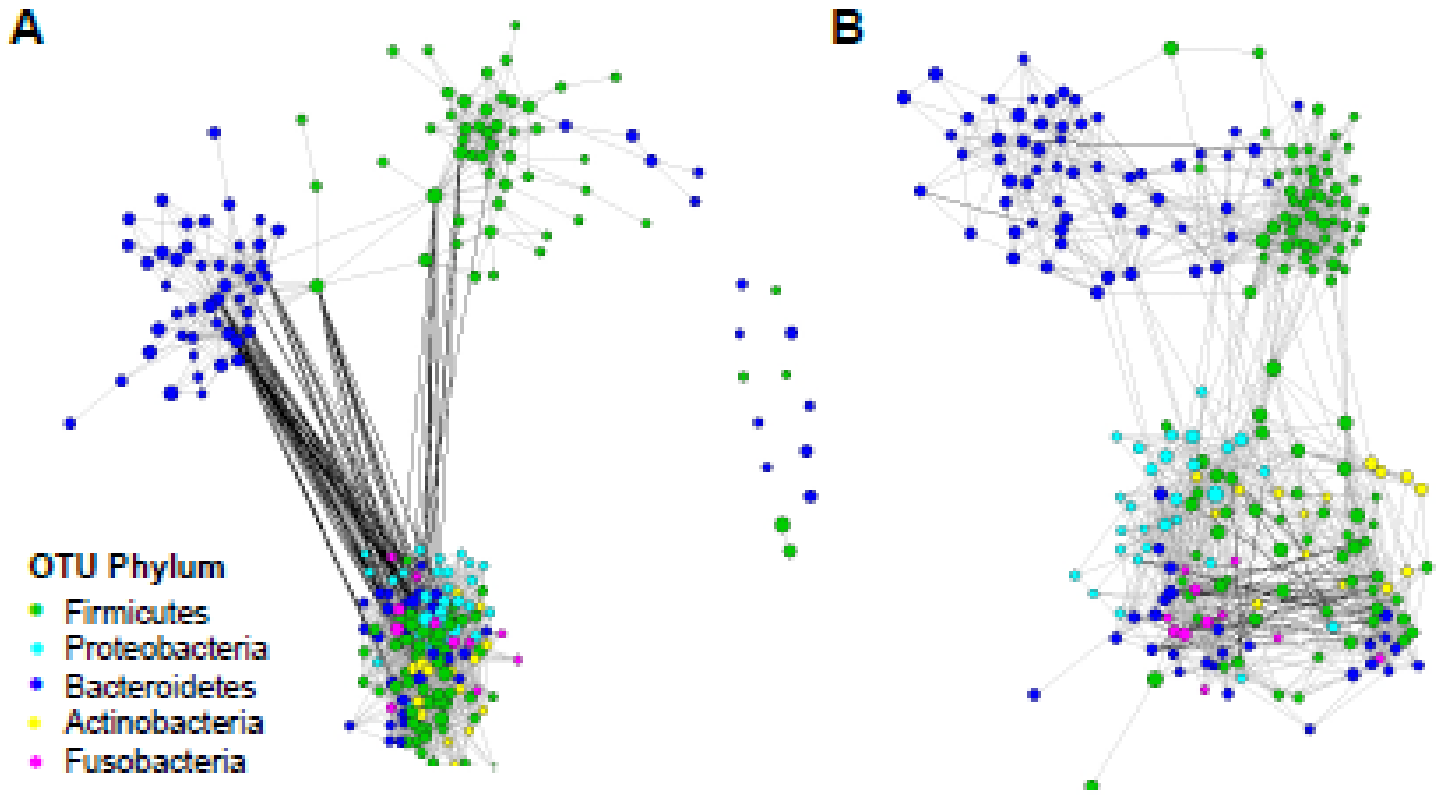
- **Mahachie John** JM , Cattaert T , Van Lishout F , Gusareva ES , Van Steen K (2012) Lower-Order Effects Adjustment in Quantitative Traits Model-Based Multifactor Dimensionality Reduction. PLoS ONE 7(1): e29594. doi:10.1371/journal.pone.0029594 **[recommendations on lower-order effects adjustments]**
- **Mahachie John** JM, Van Lishout F, Gusareva ES, Van Steen K (2013) A Robustness Study of Parametric and Non-parametric Tests in Model-Based Multifactor Dimensionality Reduction for Epistasis Detection. BioData Min. 2013 Apr 25;6(1):9 **[recommendations on QT analysis]**
- **Van Lishout** F, Mahachie John JM, Gusareva ES, Urrea V, Cleynen I, Theâtre E, Charloteaux B, Calle ML, Wehenkel L, Van Steen K (2012) An efficient algorithm to perform multiple testing in epistasis screening. BMC Bioinformatics. 2013 Apr 24;14:138 **[C++ MB-MDR made faster!]**
- **Gusareva ES,** Van Steen K (2014) Practical aspects of genome-wide association interaction analysis. Hum Genet 133(11):1343-58 **[GWAI analysis protocol]**
- **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: an example on ankylosing spondylitis. Hum Genet - accepted **[non-robustness of GWAI analysis protocols]**

- **Van Lishout F,** Gadaleta F, Moore JH, Wehenkel L, Van Steen K (2015) gammaMAXT: a fast multiple-testing correction algorithm – Nov 20;8:36. doi: 10.1186/s13040-015-0069-x. eCollection 2015. **[C++ MB-MDR made SUPER-fast]**
- **Fouladi R**, Bessonov K, Van Lishout F, Van Steen K (2015) Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis. Hum Hered 79(3-4):157-67 **[aggregating based on similarity measures to deal with DNA-seq data or multi-omics]**

- Boulesteix AL, Janitza S, Hapfelmeier A, **Van Steen K**, Strobl C (2015) Letter to the Editor: **On the term 'interaction'** and related phrases in the literature on Random Forests. Briefings in Bioinformatics 16(2): 338-345.
- 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: an example on ankylosing spondylitis. Human Genetics 134(7): 761-773. … … … … …
- Ritchie MD and **Van Steen K** (2018) The search for gene-gene interactions in genome-wide association studies**: challenges** in abundance of methods, practical considerations, and biological interpretation. Ann Transl Med 6:157.
- Gusareva et al. …**Van Steen K** (2018) Male-specific epistasis between WWC1 and TLN2 genes is associated with **Alzheimer's disease**. Neurobiol Aging 72:188.e3-188.e1 (2018).

What have we learned?

Tower of Babel - Microbiome interactions (Gusareva)

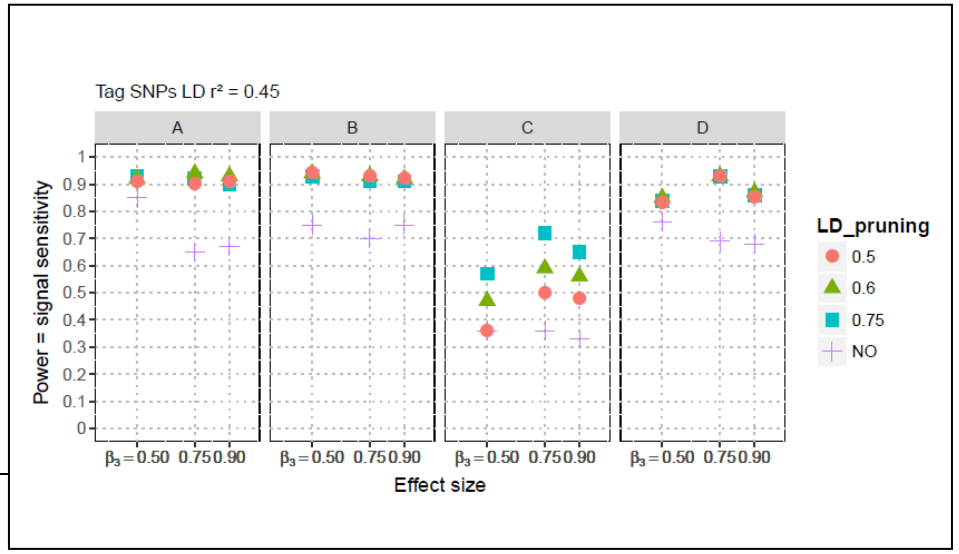


(Cougoul et al. 2019)

Linkage disequilibrium (LD) as a merit AND a nuisance

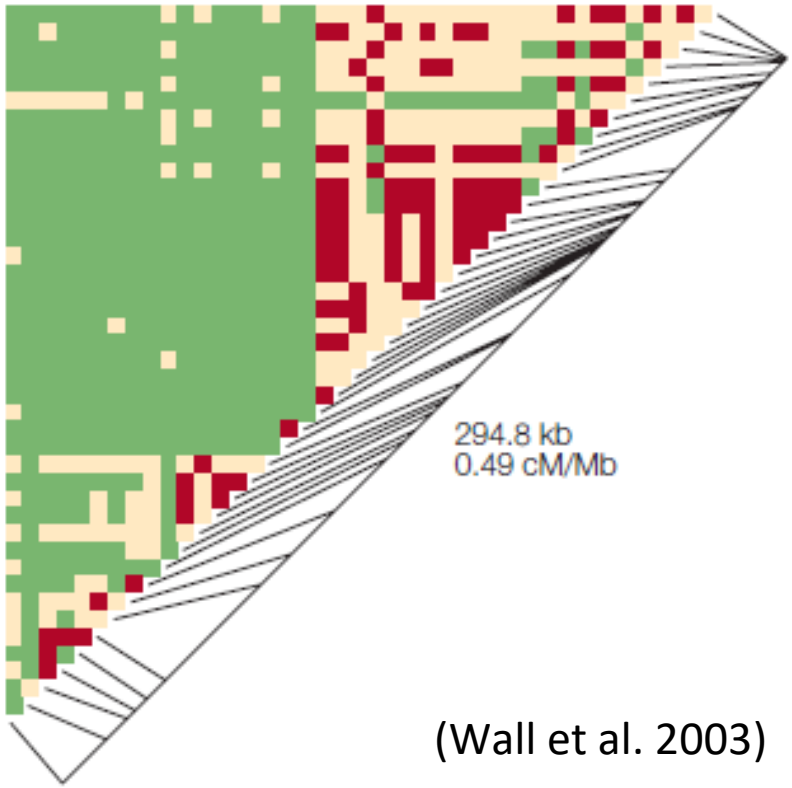
Merit: increased signal sensitivity

- LD pruning to avoid/reduce confounding between LD and epistasis (*redundant epistasis*)
- LD blocks to capture signals (**Joiret et al 2019** – under review):
  - Exact signal sensitivity may be low when actual actors were pruned out
  - No pruning gives the lowest signal sensitivity
  - Sufficient pruning gives acceptable signal sensitity
  - Lowest power when DSLs reside at the boundaries of LD regions (scenario C)



Different ways to determine LD blocks

- LD-block computation methods
  - often do not allow intermediate regions of low LD between strongly associated SNP pairs:
    - small blocks,
    - high between-block correlations (Kim et al 2018)



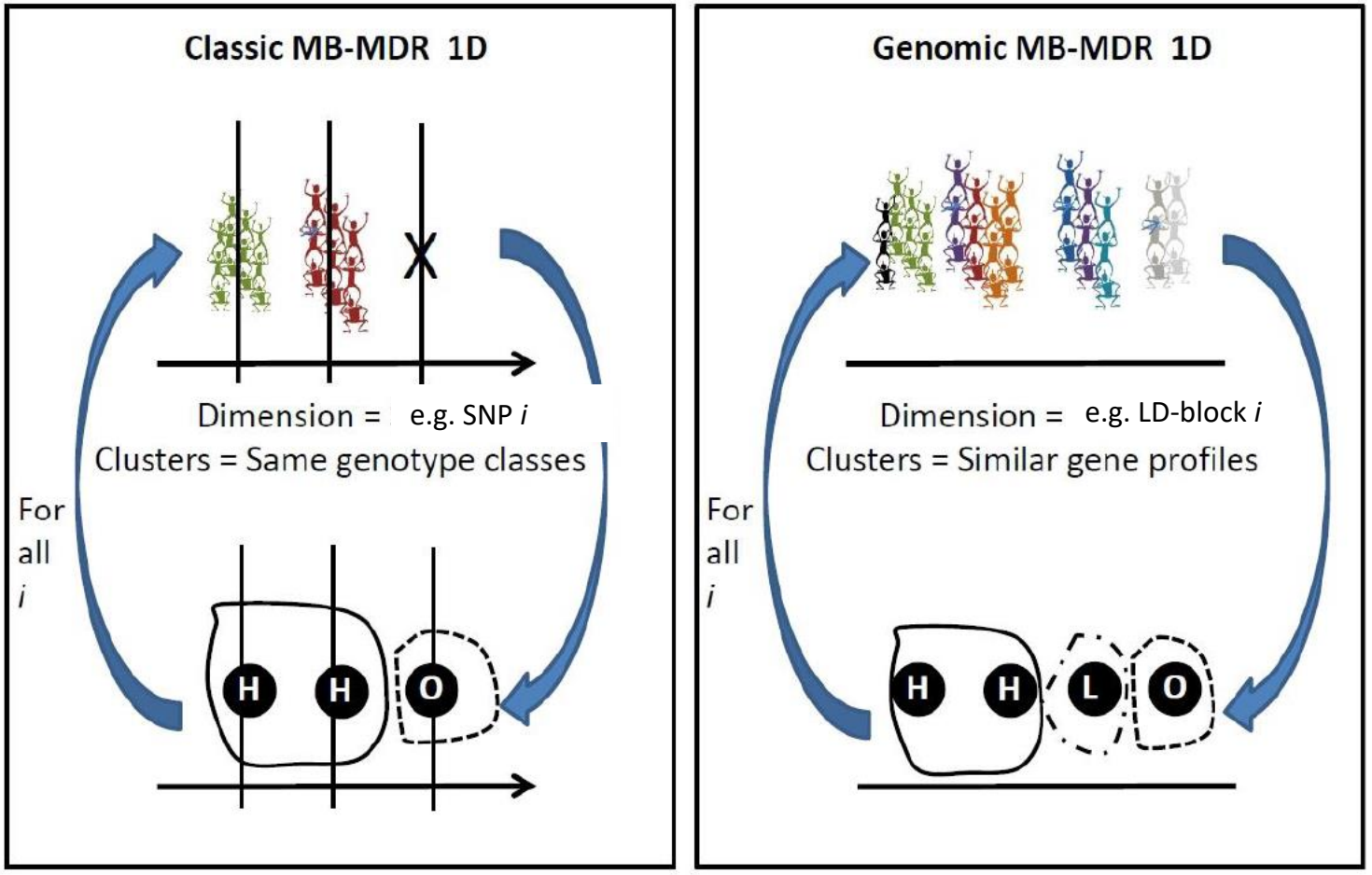
(Wall et al. 2003)

Different ways to determine LD blocks (Junior et al.)

- **Big-LD** produces larger LD blocks compared to existing methods (e.g. MATILDE, Haploview, MIG ++, or S-MIG ++); LD blocks better agree with recombination hotspot locations determined by sperm-typing experiments; per population (Kim et al. 2018)

  - Adaption in progress to facilitate downstream analysis and processing multi-ethnic groups jointly (e.g., multi-population GWAIS):
    - Population-corrected r^2, using genetic origin / admixture proportions of individual genomes (Mangin et al. 2012)

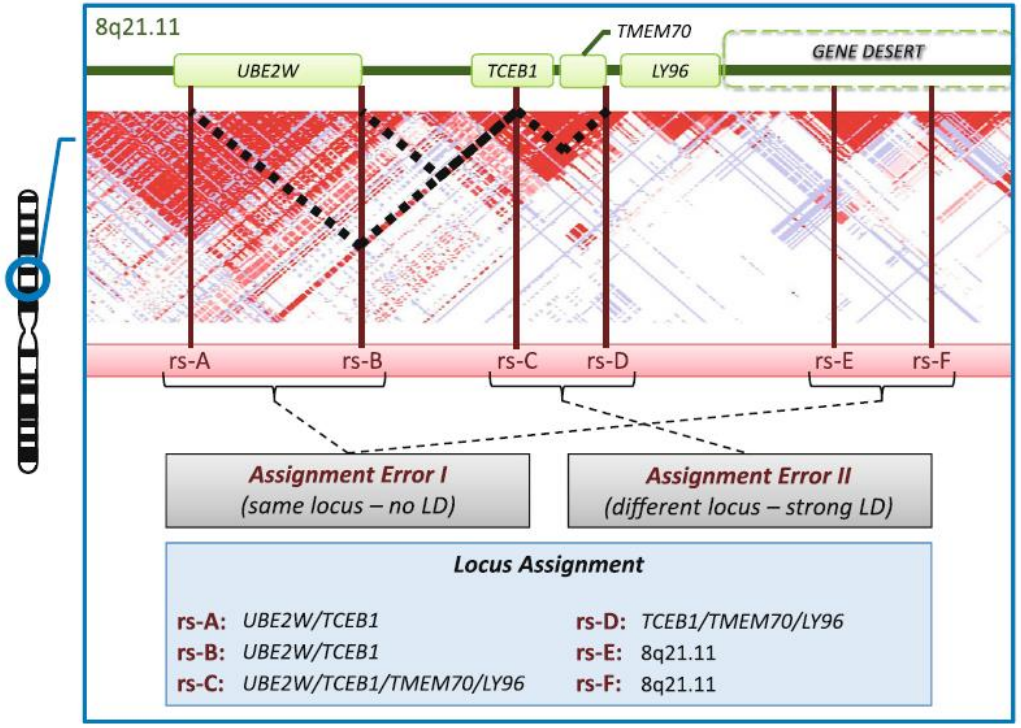
Big-LD and MB-MDR ←→ "phantom epistasis (de los Campo et al. 2019)"



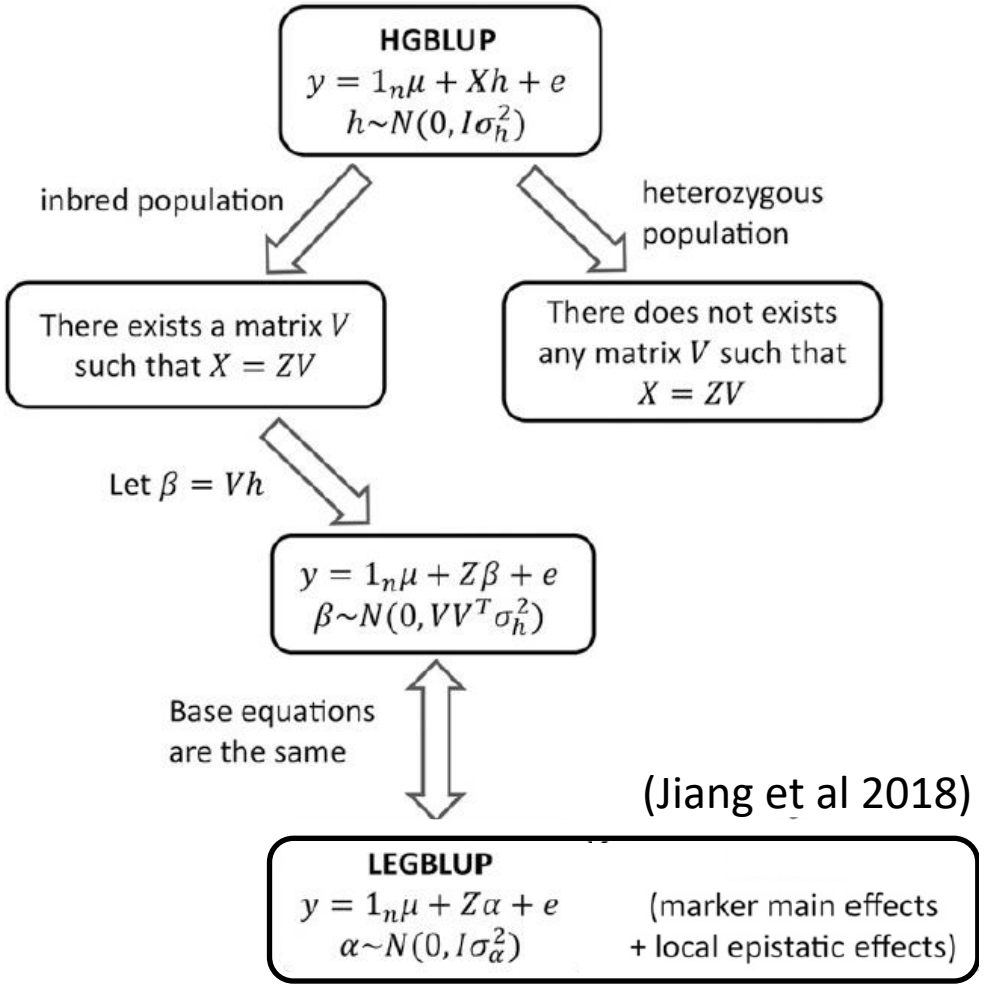
Big-LD and MB-MDR ←→ classical enrichment (Junior et al.)

- **LD based locus assignment and its error sources**
  - whole genomic region captured by SNPs in strong LD (say $r^2 \geq 0.8$) with the marker originally reported in a GWAS
  - loci: genes located within this region

  (Arnold et al 2012)



Big-LD and MB-MDR ←→ within block epistasis



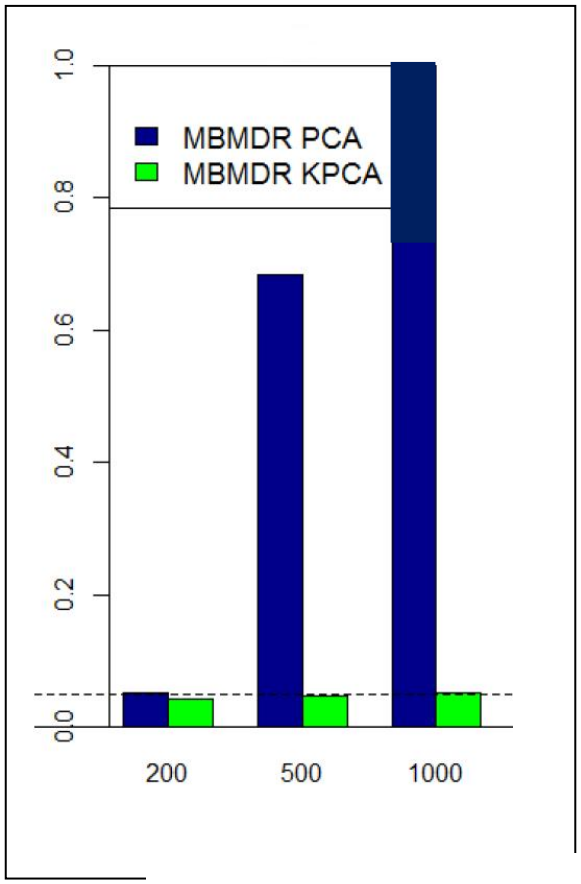
(Jiang et al 2018)

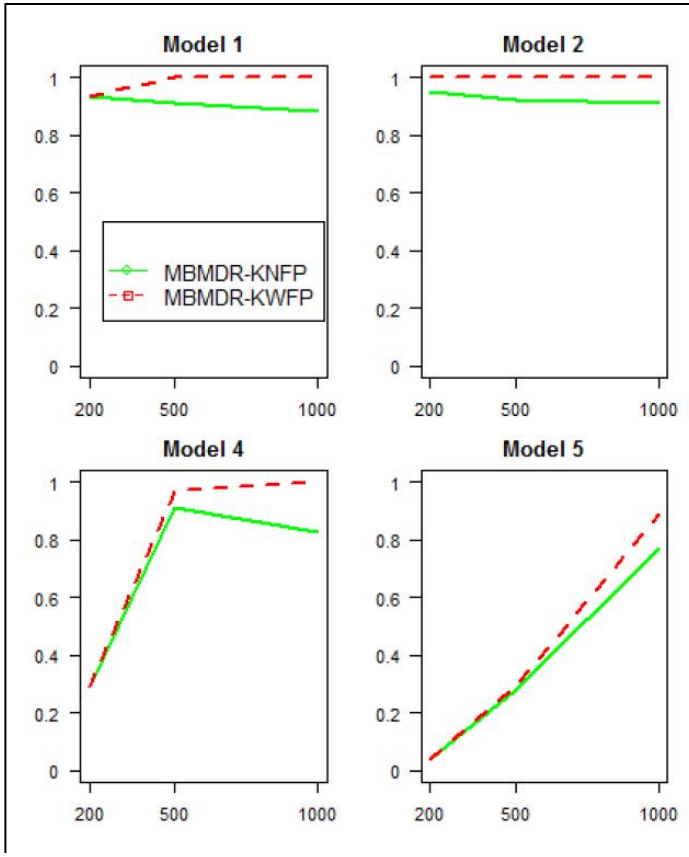
Published in final edited form as:
Nature. 2014 April 10; 508(7495): 249–253. doi:10.1038/nature13005.

Detection and replication of epistasis influencing transcription in humans

Gibran Hemani[1,2,*], Konstantin Shakhbazov[1,2], Harm-Jan Westra[3], Tonu Esko[4,5,6], Anjali K. Henders[7], Allan F. McRae[1,2], Jian Yang[1], Greg Gibson[8], Nicholas G. Martin[7], Andres Metspalu[4], Lude Franke[3], Grant W. Montgomery[7,+], Peter M Visscher[1,2,+], and Joseph E Powell[1,2,+]

Nuisance: **Inadequate capturing of population structure** (no cat cov!)





Left : 60/40 CC ratio, structural epistasis according to corresponding full penetrance Ritchie epistasis model (**Abegaz et al. 2019** - submitted)

Below : 50/50 (200+200) (**Cattaert et al. 2011**)

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Noise	MB	MDR	MB	MDR	MB	MDR	MB	MDR	MB	MDR	MB	MDR
None	100	99	100	100	100	95	100	93	93	62	97	73

Non-continuous axes of genetic variation ???

Bioinformatics, 33(12), 2017, 1820–1828
doi: 10.1093/bioinformatics/btx071
Advance Access Publication Date: 14 February 2017
Original Paper

Genetics and population analysis

Genome-wide genetic heterogeneity discovery with categorical covariates

Felipe Llinares-López[1,2,*,†], Laetitia Papaxanthos[1,2,*,†], Dean Bodenham[1,2], Damian Roqueiro[1,2], COPDGene Investigators[3] and Karsten Borgwardt[1,2,*]

[1]Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland, [2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland and [3]COPDGene® Study

Lack of obvious correspondence between biology and statistics

- From the literature (~ interaction-specific vs two-locus hypotheses):
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387
  - …
  - Moore and Williams (2005) BioEssays 27:637–646
  - Phillips (2008) Nat Rev Genet 9:855-867
  - Clayton DG (2009) PLoS Genet 5(7): e1000540
  - Wang, Elston and Zhu (2010) Hum Hered 70:269-277
  - …

  - **Van Steen et al. (2012) Brief Bioinform. 13(1):1-19**
  - **Aschard et al. (2012) Hum Genet 131(10):1591-1613**
  - **Gusareva and Van Steen (2014) Hum Genet 133(11):1343-58**

- In either case: statistical interactions DO imply JOINT involvement

Understanding molecular mechanisms of epistasis [→ Vidal lab]

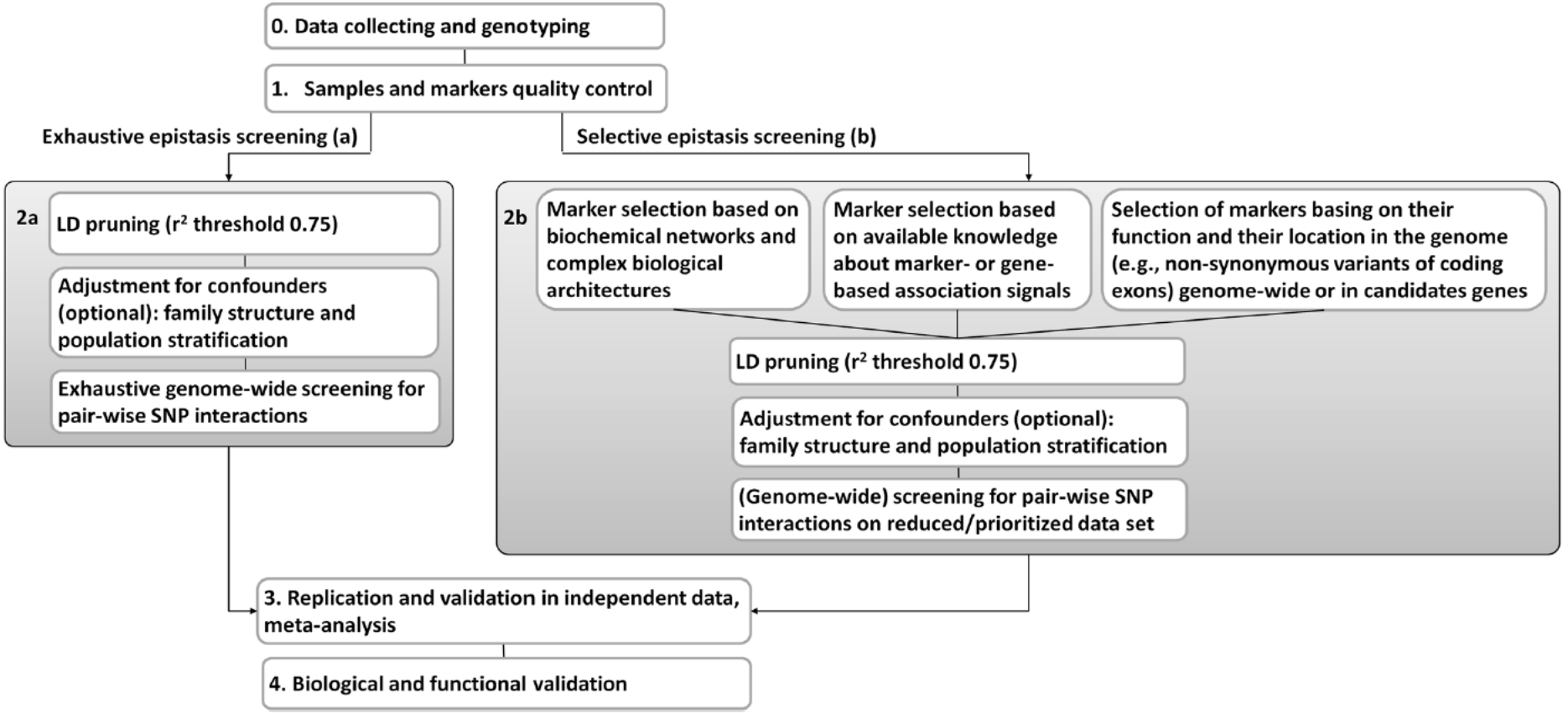
- Best evidence for pervasiveness of epistasis (wrt strong LOF mutations) are derived from large-scale reverse genetics screens

  Pairs of mutations (or RNA interference treatments) are systematically combined and effects on viability or growth are determined (yeast: Costanzo et al. 2010)

- Conclusion:
  - For majority of LOFs the effect can be influenced by perturbing the activity of many additional genes
  - Mutations with weaker effects on proteins are understudied. It needs to be determined what their potential epistatic role is.
  - Multiple molecular mechanisms underly similar epistatic interactions

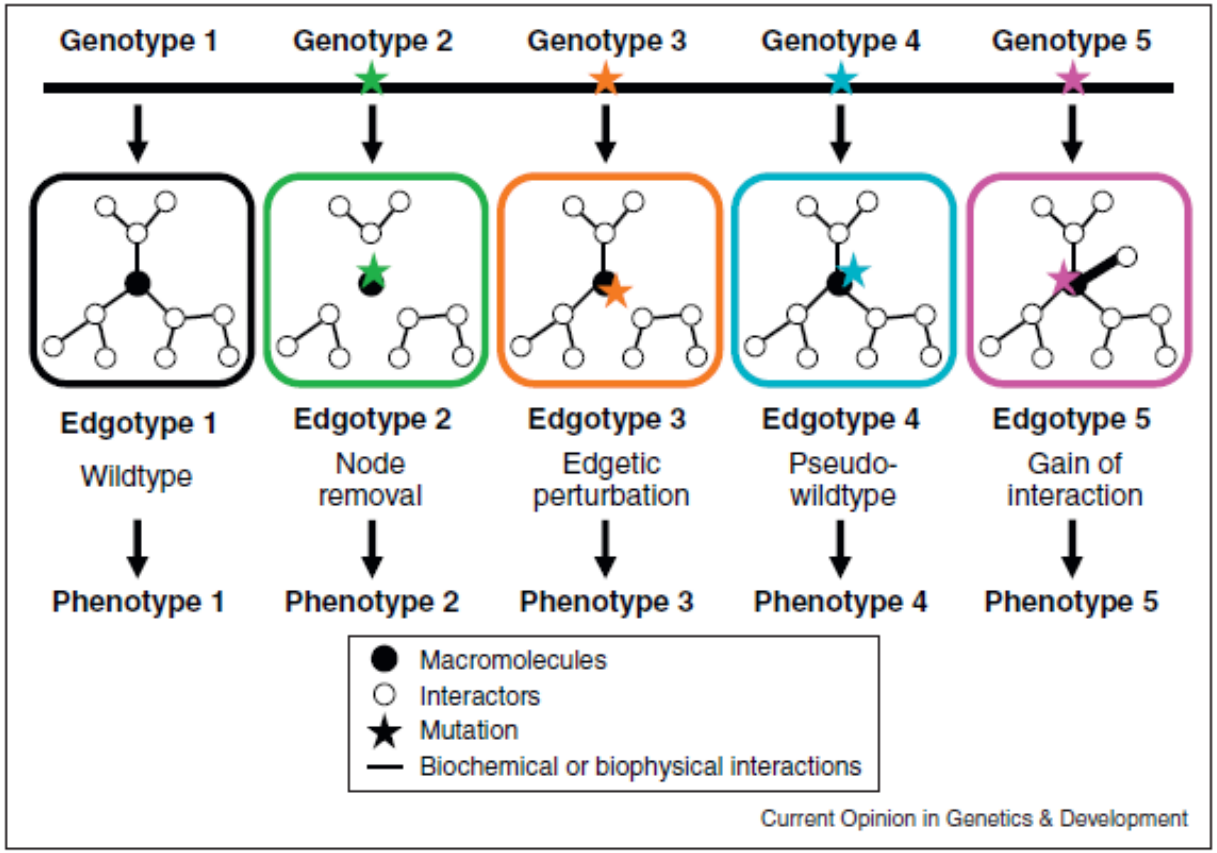
(Lehner 2011)

A priori use of biological (functional) knowledge



(Gusareva et al. 2014)

Knowledge boosting: PPI perturbations



(Sahni et al. 2013)

The space of PPI perturbations

Multi-scale perturbations of protein interactomes reveal their mechanisms of regulation, robustness and insights into genotype–phenotype maps

Marie Filteau,* Hélène Vignaud,* Samuel Rochette,* Guillaume Diss,* Andrée-Ève Chrétien,* Caroline M. Berger and Christian R. Landry

Corresponding author: Christian Landry, Département de Biologie, Institut de Biologie Intégrative et des Systèmes, Université Laval, Room 3106, Pavillon Charles-Eugène-Marchand 1030, Avenue de la Médecine, Québec (Québec) G1V 0A6, Canada. Tel.: 418-656-3954; Fax: 418-656-7176; E-mail: Christian.landry@bio.ulaval.ca
*These authors contributed equally to this work.

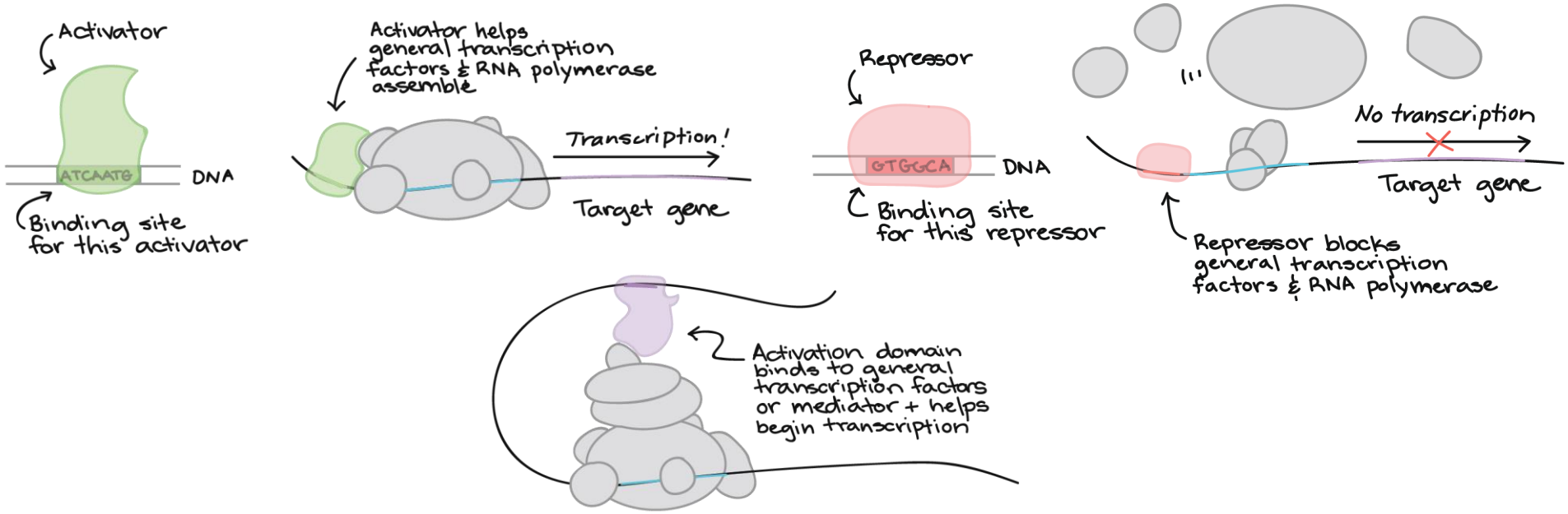
(Filteau et al. 2015)

The space of PPI perturbations

- Zhong et al. 2009: Interactomes of 29 mutant alleles of genes implicated in five human Mendelian disorders. Each allele was cloned and its interactions tested with ~ 8100 other proteins (edgetic perturbations)
  - **Different mutant alleles of the same protein can cause different perturbations on their PPI profiles**
- Sahni et al. 2015: ~ 2500 mutant alleles of proteins implicated in human diseases and their ~1000 corresponding wild-type proteins; Y2H to test the interactions of these proteins with a set of ~ 7200 human ORFs
  - **~ One third of the mutations, the cause of the disease may result from perturbations of PPIs**

“Message passing” between layers of information

- Example: Transcription Factors (TFs)



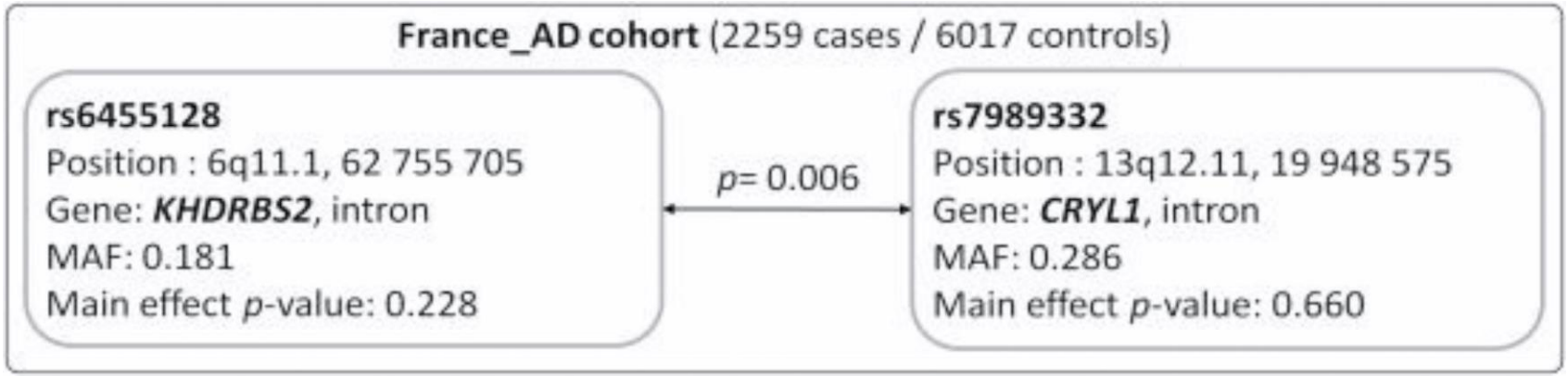
TF interaction network (TranSignalTM TF ProteinArray) ⟷ Gene co-expression network

TRRUSTv2 (Han et al. 2018)

Replication and validation

Genome-wide association interaction analysis for Alzheimer's disease

Elena S. Gusareva[1,2], Minerva M. Carrasquillo[3], Céline Bellenguez[4,5,6], Elise Cuyvers[7,8], Samuel Colon[3], Neill R. Graff-Radford[9], Ronald C. Petersen[10], Dennis W. Dickson[3], Jestinah M. Mahachie Johna[1,2], Kyrylo Bessonov[1,2], Christine Van Broeckhoven[7,8], The GERAD1 Consortium, Denise Harold[11], Julie Williams[11], Philippe Amouyel[4,5,6], Kristel Sleegers[7,8], Nilüfer Ertekin-Taner[9], Jean-Charles Lambert[4,5,6], and **Kristel Van Steen[1,2]**



Which extent of replication is required?

4. Replication analysis with alternative methods for epistasis detection: follow up the selected set of markers
(MB-MDR$_{2D}$ analysis, SD plot,
logistic regression-based methods)

5. Replication of epistasis in the independent data and biological validation

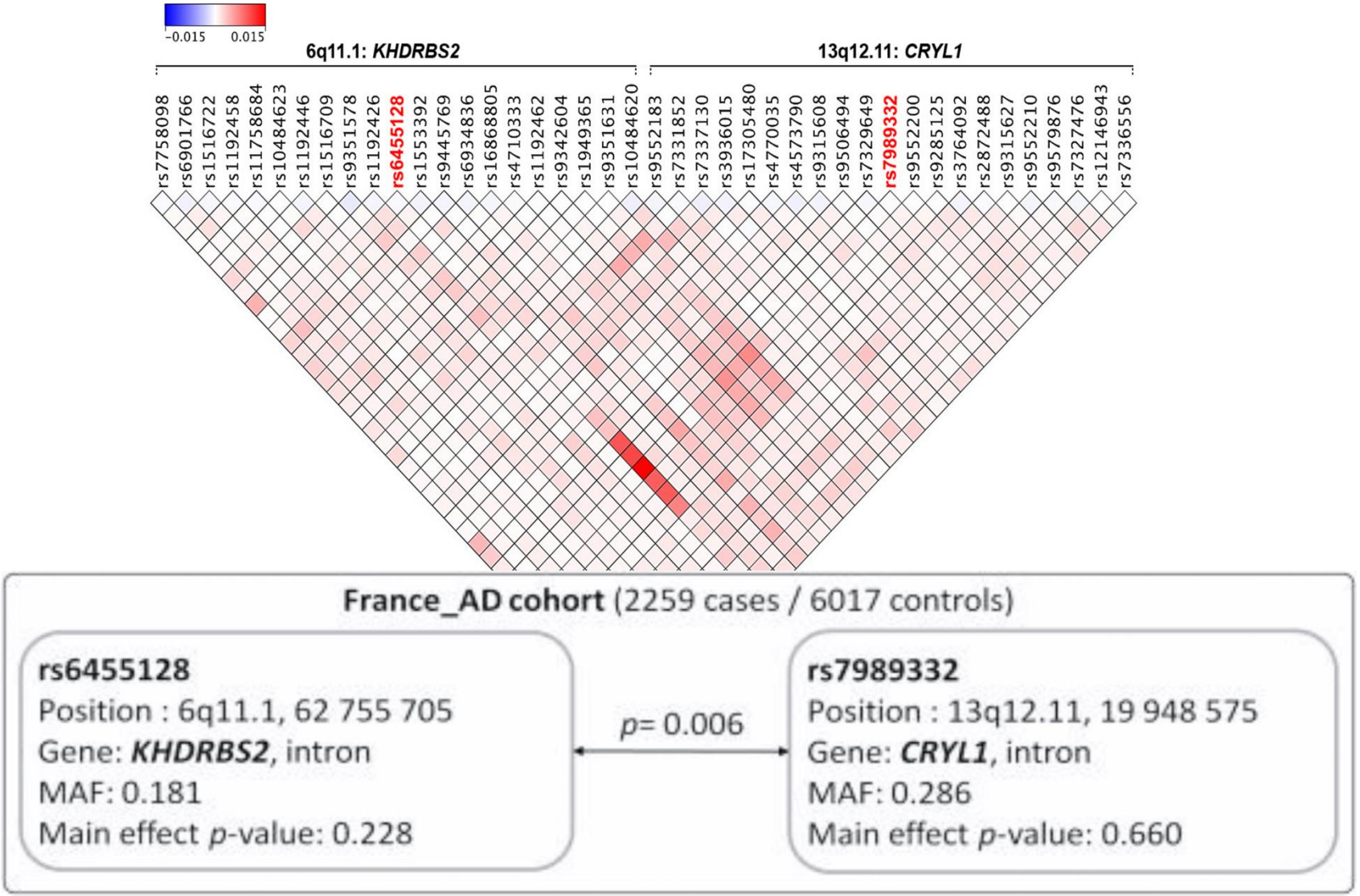
"This study in particular demonstrates an alternative approach to elucidate the functional repercussions of epistasis."

Review Article

Bridging the Gap between Statistical and Biological Epistasis in Alzheimer's Disease

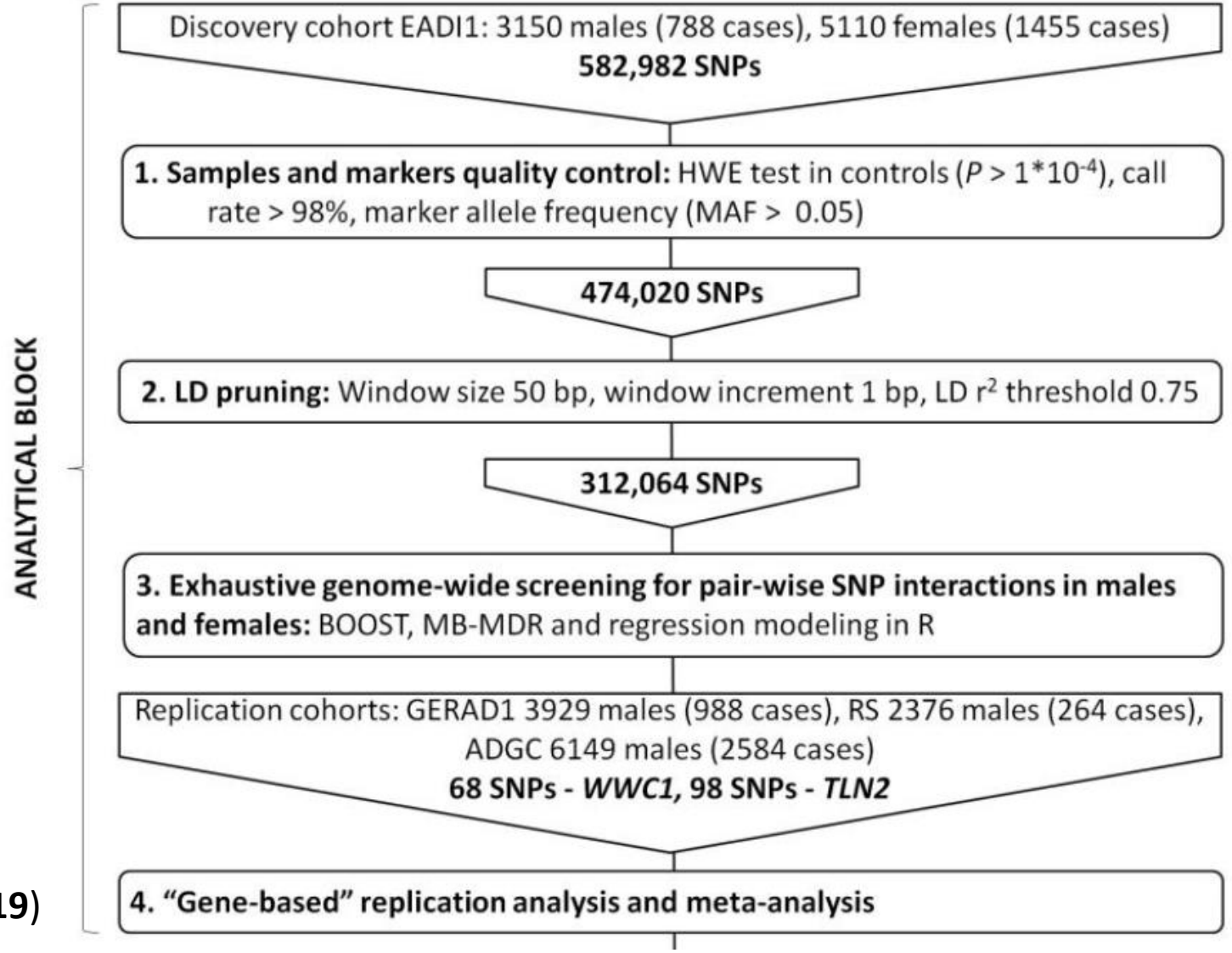
Mark T. W. Ebbert, Perry G. Ridge, and John S. K. Kauwe

Which extent of replication is required?

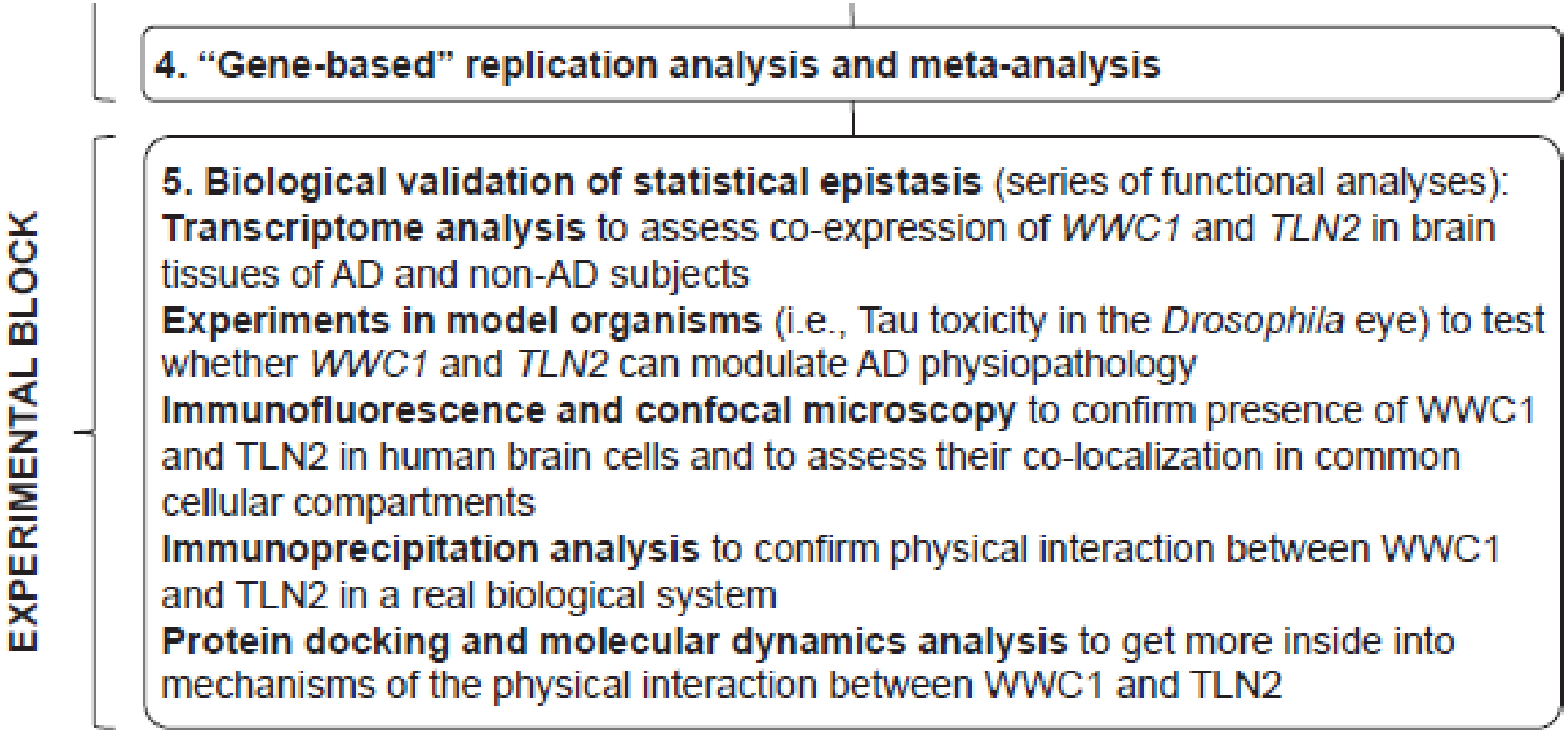


Sex-specific interactions for AD

(**Gusareva et al 2019**)



Which extent of (experimental) validation is required?



(**Gusareva et al 2019**)

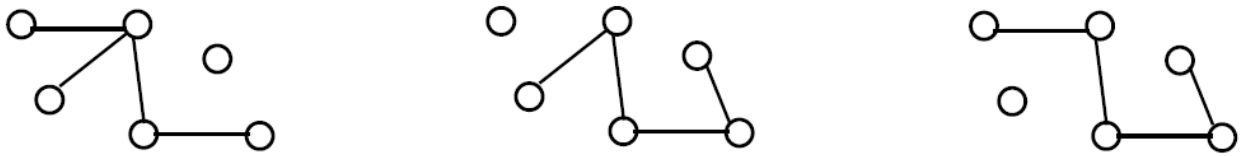
Functional in-silico analyses - pairs or unique SNP set from pairs

- Hemani et al. 2014:
  - Are the SNPs in transcriptionally active regions?
    - Cell-type specific enrichment of active chromatin
  - Colocalization of corresponding chromosomal regions in a cell?
  - Which genomic features are covered by the index SNPs (LD r^2>0.8; 0.5Mb)
    - Predicted promoter region including TSS, Predicted promoter flanking region, Predicted enhancer, Predicted weak enhancer or open chromatin cis regulatory element, CTCF enriched element, Predicted transcribed region, or Predicted Repressed or Low Activity region positions.
  - Enrichment of transcription factor binding sites around index SNPs?
- **Final protocol paper**: pre (Moore et al.) / post in silico (Hemani et al.); BIOGRID with stringent validation criteria; and minimal lab work (**Gusareva et al.**)

Epistasis meta-analysis is hampered by analytic heterogeneity

Statistical Epistasis Networks
(nodes: e.g., SNPs → genes; edges: statistical evidence)

Network 1 | Network 2 | Network 3



Heterogeneity

- Same data, different analytics
- Different data, same analytics (incl. meta-analysis)

We have all found "our" interactions, now what?
(drawing – ensemble – conclusions)

(Duroux et al)

- Consensus clustering + meta clustering = Multiple Consensus Clustering (MCC: Zhang and Li 2011)
- Similarity Network Fusion (Wang et al 2014)
- Regularized unsupervised multiple kernel learning (Speicher et al 2015)

- Statistical comparison of networks (Fraiman et al 2018)
- Differential graphlet community analysis (Wong et al 2015)

Computational efficiency – pragmatic consideration

- Graphics processing units (GPUs),
  as alternative powerful and cost-effective parallel processing units (Putz et al. 2013)
- Cloud computing infrastructures,
  although these do not offer unlimited possibilities (Wang et al. 2011)
- Hardware oriented solutions,
  such as those based on field-programmable gate array (FPGA) architecture (Gundlach et al. 2016)

Computational feasibility

Multiple testing correction via “MAXT” in MBMDR-3.0.3:

SNPs	Sequential version Binary trait	Sequential version Continuous trait	Parallel workflow Binary trait	Parallel workflow Continuous trait
10^2	45 sec	1 min 35 sec	<1sec	<1sec
10^3	1 hour 16 min	2 hours 39 min	38 sec	1 min 17 sec
10^4	5 days 13 hours	11 days 19 hours	1 hour 3 min	2 hours 14 min
10^5	≈ 1.5 year	≈ 3 years	4 days 9 hours	≈ 9 days

The parallel workflow was tested on a cluster composed of 10 blades, containing each four Quad-Core AMD Opteron(tm) Processor 2352 2.1 GHz. The sequential executions were performed on a single core of this cluster. The results prefixed by the symbol “≈” are extrapolated.

(Van Lishout et al. 2013)

Computational feasibility: approximating vs exact

Multiple testing correction via “gammaMAXT” in MBMDR-4.2.2:

SNPs	Sequential version Binary trait	Parallel workflow Binary trait	Sequential version Continuous trait	Parallel workflow Continuous trait
10^3	13 min 33 sec	20 sec	13 min 18 sec	18 sec
10^4	52 min 15 sec	1 min 05 sec	56 min 14 sec	53 sec
10^5	64 hours 35 min	22 min 15 sec	70 hours 03 min	20 min 28 sec
10^6	$\approx$ 270 days	25 hours 12 min	$\approx$ 290 days	24 hours 06 min

The parallel workflow was tested on a 256-core computer cluster (Intel L5420 2.5 GHz 1333 MHz FSB). The sequential executions were performed on a single core of this cluster. The results prefixed by the symbol “$\approx$” are extrapolated.

(Van Lishout et al. 2015)

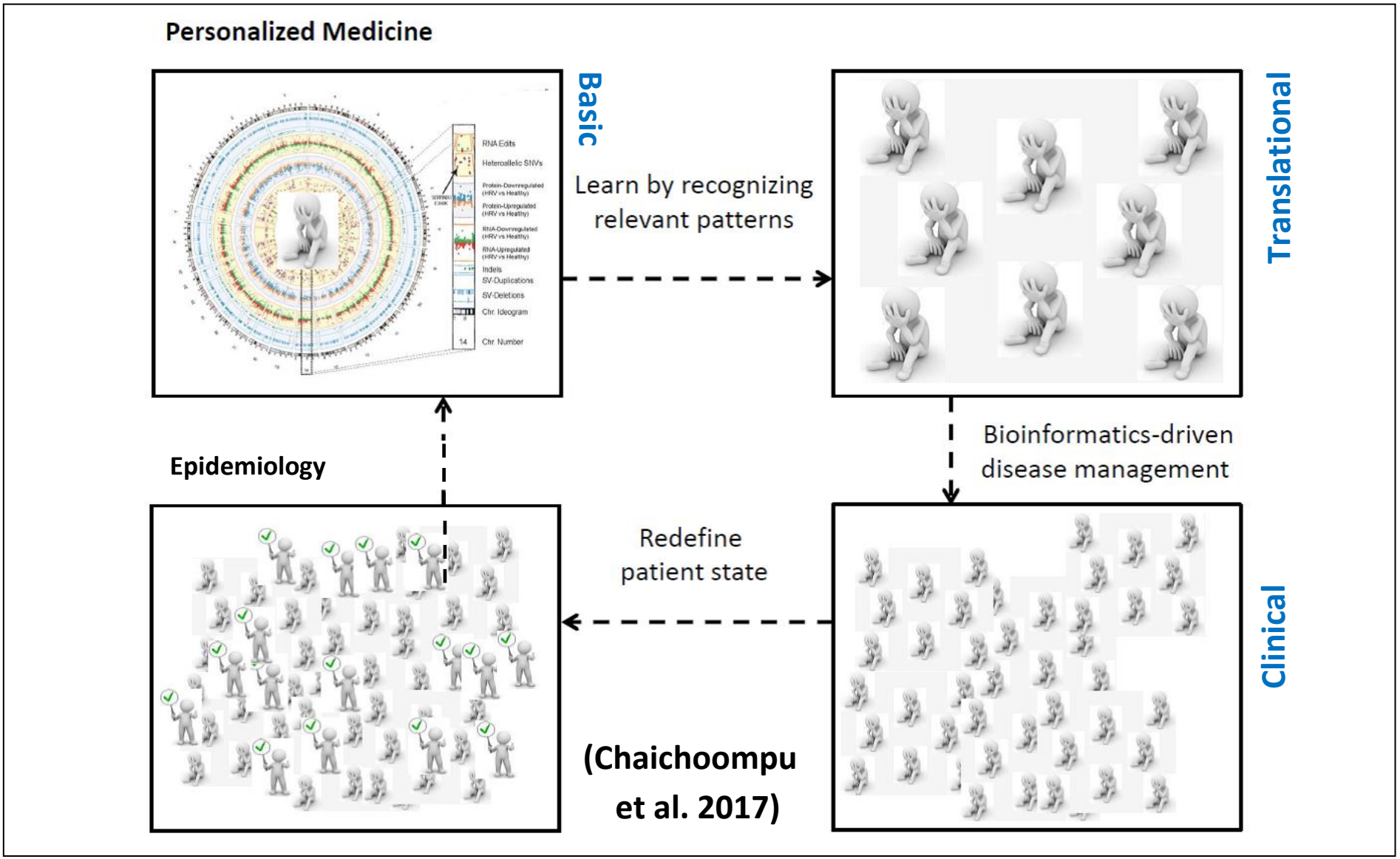
What are the implications?

Individual risk (trait) prediction – **deep MBMDR** (~McGill,CA)

- MB-MDR was never tested for its predictive ability
  - Note: MDR uses cross-validation and prediction accuracy as measure to select the most optimal interaction model
- Collaborators extended **MB-MDR** to generate **prediction rules**
- The new algorithm (available in R) can use information hidden in interactions more efficiently than two other state-of-the-art algorithms; it clearly **outperforms Random Forest and Elastic Net** if interactions are present.
- The performance of these algorithms is comparable if no interactions are present

(Gola et al. 2019)

Molecular reclassification of disease [→Sharma lab, Michael Cho]



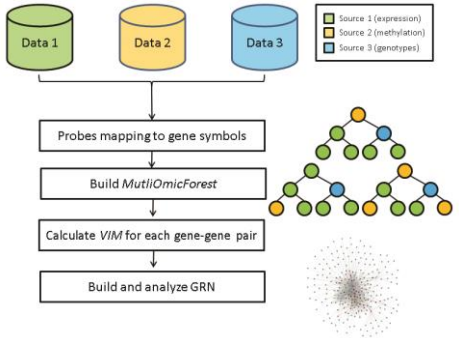
Integration - GRNs by integrating multi-omics data

Integrative networks approaches

- Adapt Lasso? Nine "Lasso's" compared incl new **LABNet** for gene expression networks (paper: Gadaleta)
  - Multicollinearity + high-dimensionality are concern for all
  - GO LABNet (MCC: correl. coefficient between observed and predicted binary classifications)
- New computational approach to analyse *gene expression and methylation* profiles via regression analysis and network-based techniques - **Regression2Net** (paper: Gadaleta; Bessonov)
  - SNF between EE-net and EM-net

Conditional Inference Forests (CIFs)

- RF VIMs: bias towards correlated predictor variables. Created **new VIMs for CIFs**
  - Optimal performance of CIF_{cond} followed by CIF_{mean} (paper: Bessonov)
  - **Bring EXTRA trees idea in CIF**
- **MultiOmicForest** (thesis : Bessonov)
  - Extends CIF_{mean} to 3-omics (diff scales)
  - Non-linearity via GAMs: $g\big(E(Y)\big) = \beta_0 + f_1(x_1) + \cdots + f_m(x_m)$

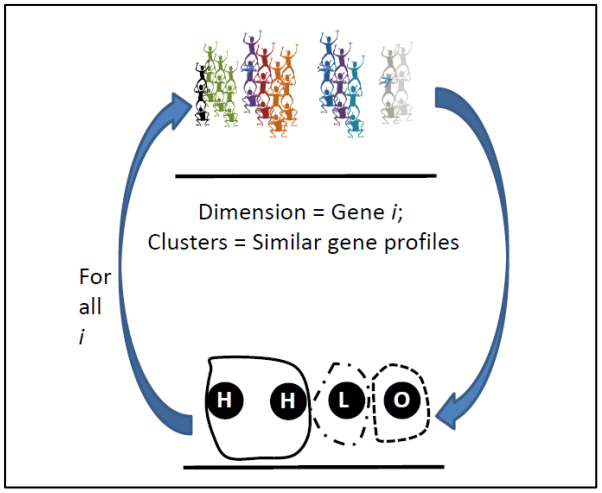


BIO3's approach

- **Data integration** (heterogeneous data types) – WELL PROGRESSING

Ex: MB-MDR + defining a smaller "system"

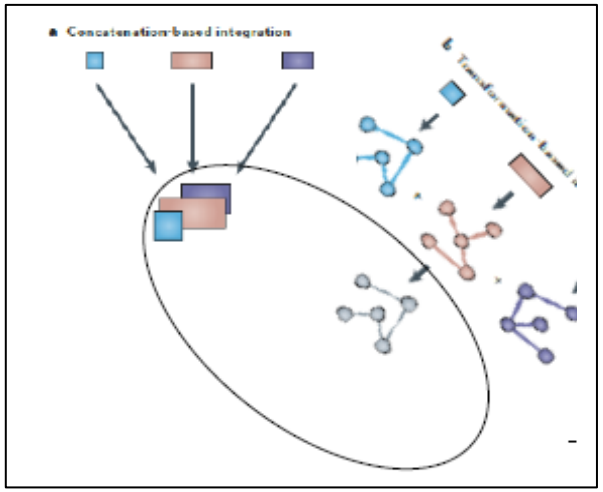
to create omics integrative units of analysis



(DEST*in*CT : MB-MDR)

- Component-based
- Kernel-based *(big)*
- Network-based

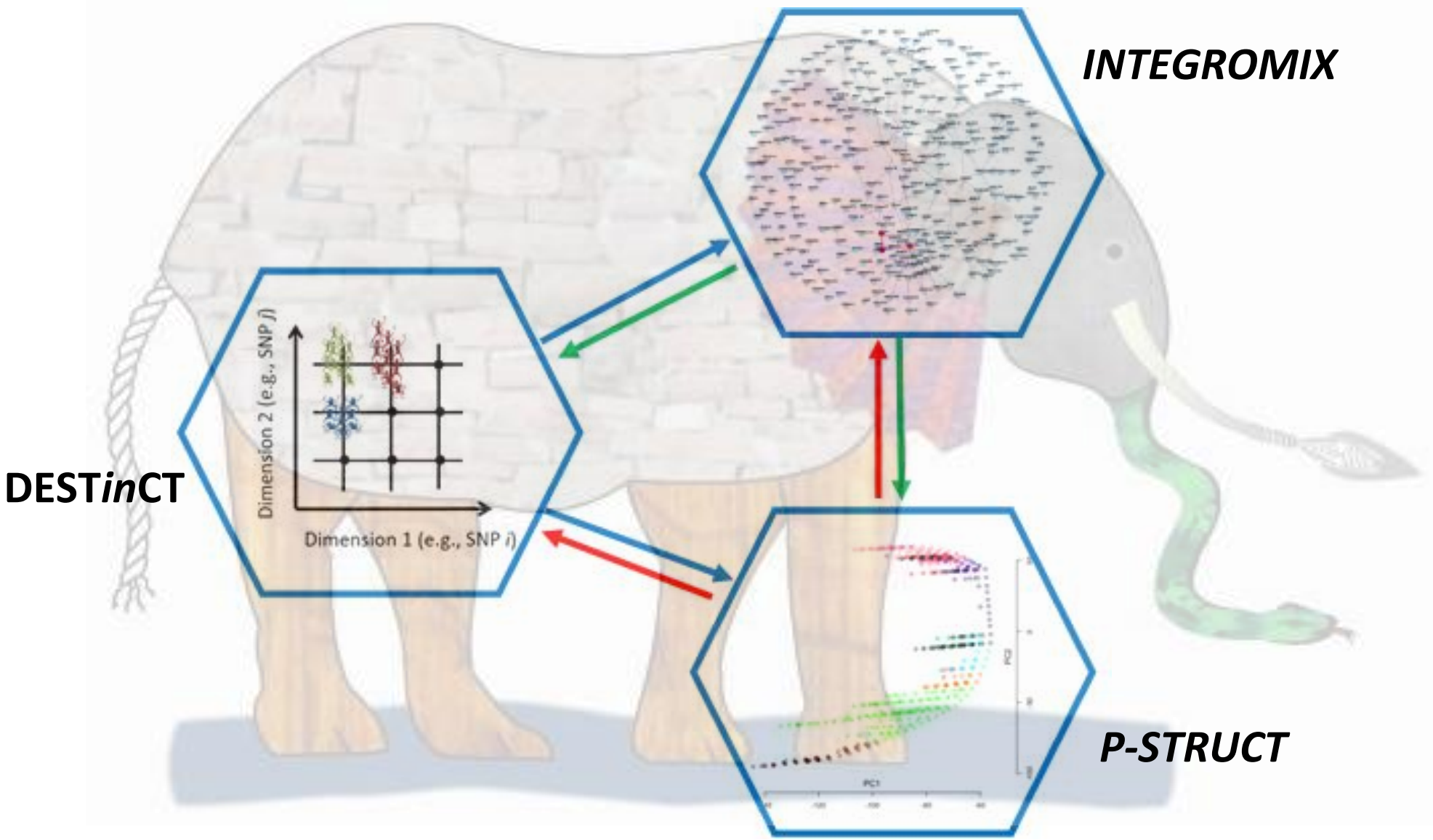
(PhD thesis **Fouladi 2018**)



(Ritchie et al. 2015)

- **Analytic integration** (modelling paradigms) – INFANCY

Take-home messages


INTEGROMIX
DESTinCT
Dimension 2 (e.g., SNP j)
Dimension 1 (e.g., SNP i)
P-STRUCT

Through the looking-glass

Alice doesn't play by the conventional rules of a little girl during the 1800s; she's up for whatever comes her way and is willing to take a chance on the unexpected with brilliant results.

(Lewis Carroll)

Questions?

Main supporting doc to this class (complementing course slides)

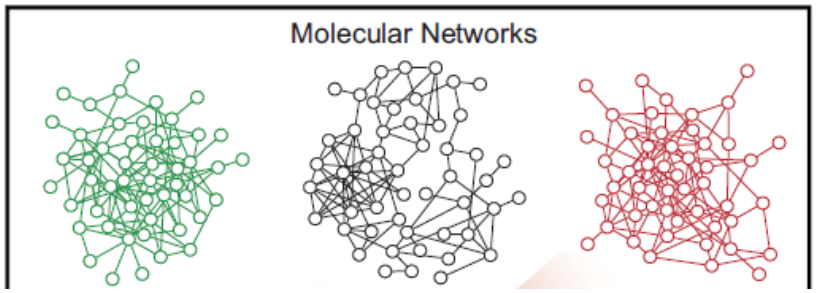
√

Article

Cell Systems

Systematic Evaluation of Molecular Networks for Discovery of Disease Genes

Graphical Abstract



Authors

Justin K. Huang, Daniel E. Carlin, Michael Ku Yu, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo, Trey Ideker

Correspondence

jkh013@ucsd.edu

√

Human Genetics (2019) 138:293–305
https://doi.org/10.1007/s00439-019-01987-w

REVIEW

How to increase our belief in discovered statistical interactions via large-scale association studies?

K. Van Steen[1,2] · J. H. Moore[3]

Homework assignment II

+

Check out the document

“Critical evaluation of a paper/report”