

# Living in a World of Interactions

**Kristel Van Steen, PhD<sup>2</sup> (\*)**

(\*) WELBIO, GIGA-R Medical Genomics (BIO3), University of Liège, Belgium

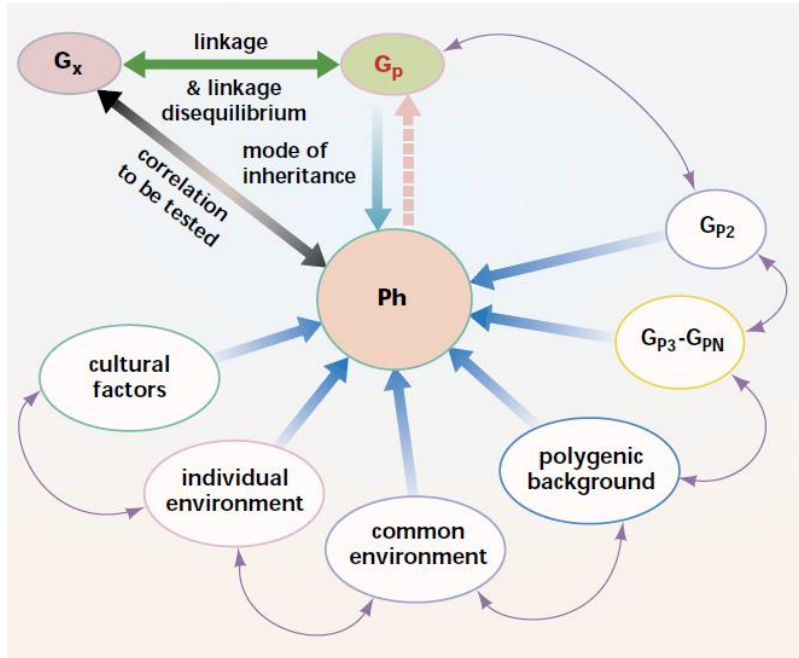
---

# Outline

- Motive
  - Opportunity
    - Data context
    - Disease context
  - Means
    - MB-MDR
    - GWAI protocol
  - Take-home messages
-

# MOTIVE

# 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)

## “Interactions” in humans come natural

- From an evolutionary biology perspective, for a phenotype to be buffered against the effects of mutations, it must have an underlying genetic architecture that is comprised of networks of genes that are redundant and robust.
- The existence of these networks creates dependencies among the genes in the network and is realized as gene-gene interactions or (*trans-*) epistasis.
- This suggests that epistasis is not only important in determining variation in natural and human populations, but should also be more widespread than initially thought (rather than being a limited phenomenon).

(Moore et al. 2005)

---

## Unexplained heritability

- 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

<b>Explanation</b>	<b>Rationale</b>	<b>Comments</b>
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 <i>replicated</i> evidence for statistical interactions in complex diseases; network-based approaches may be helpful (Hu et al. 2011)

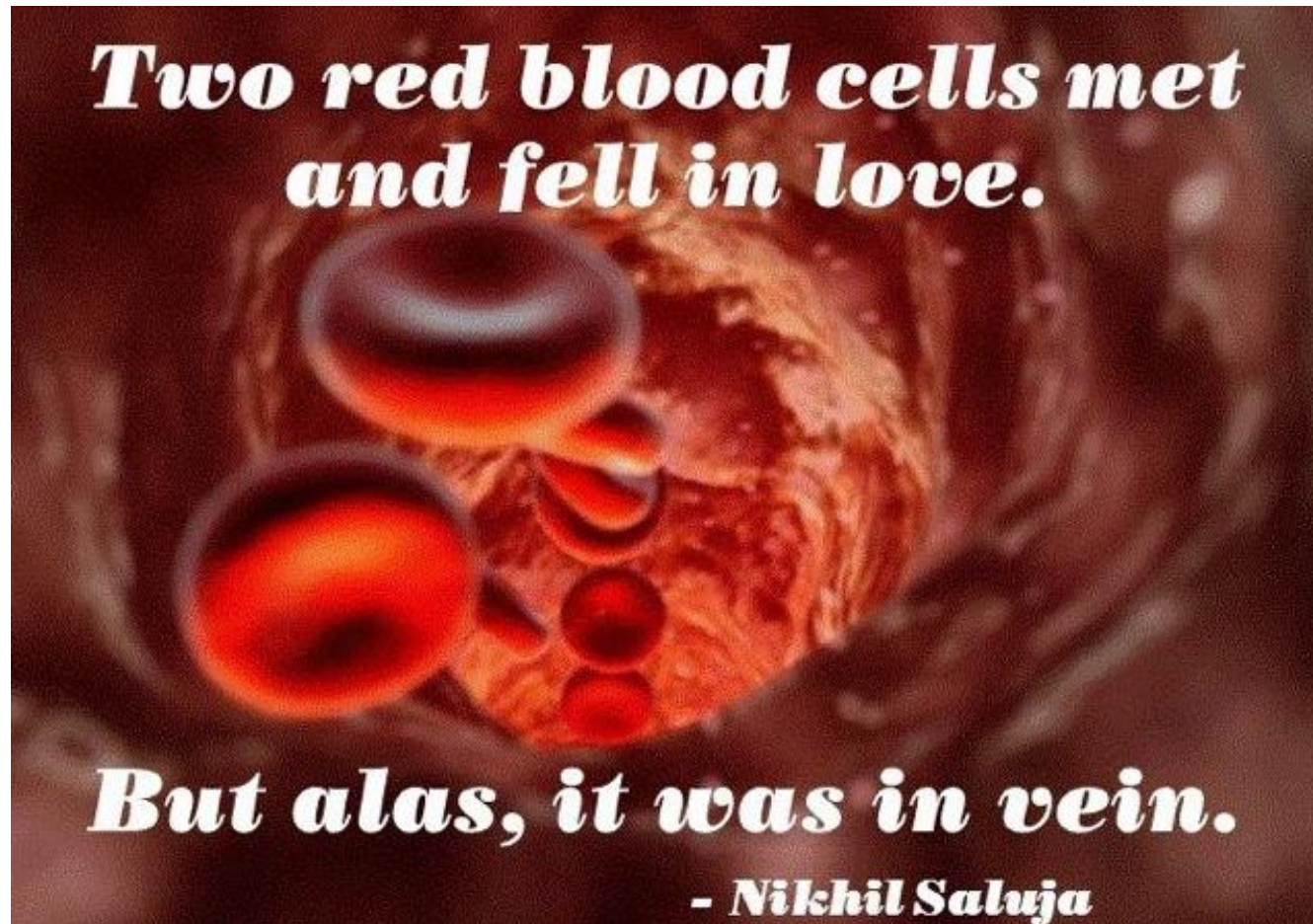
(adapted from Silverman et al. 2012)



(Hayden 2010)  
« Life is Complicated »)



## 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

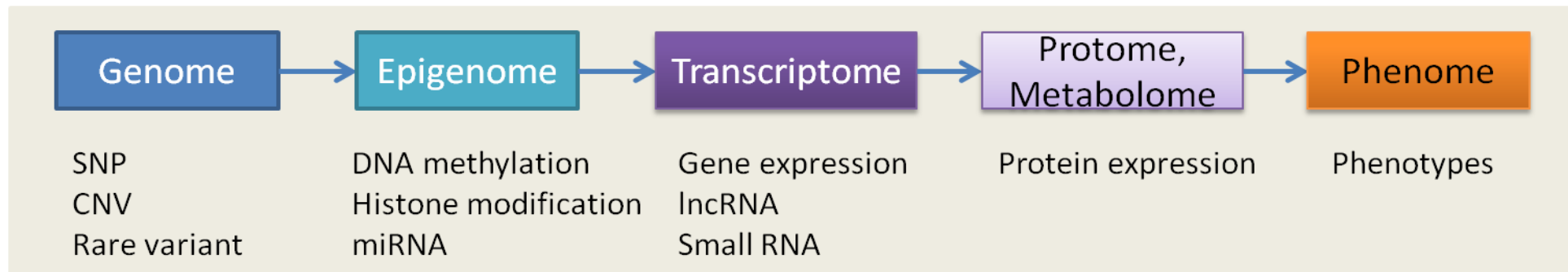


(Elton 1968; Wikipedia)

## Omic 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)

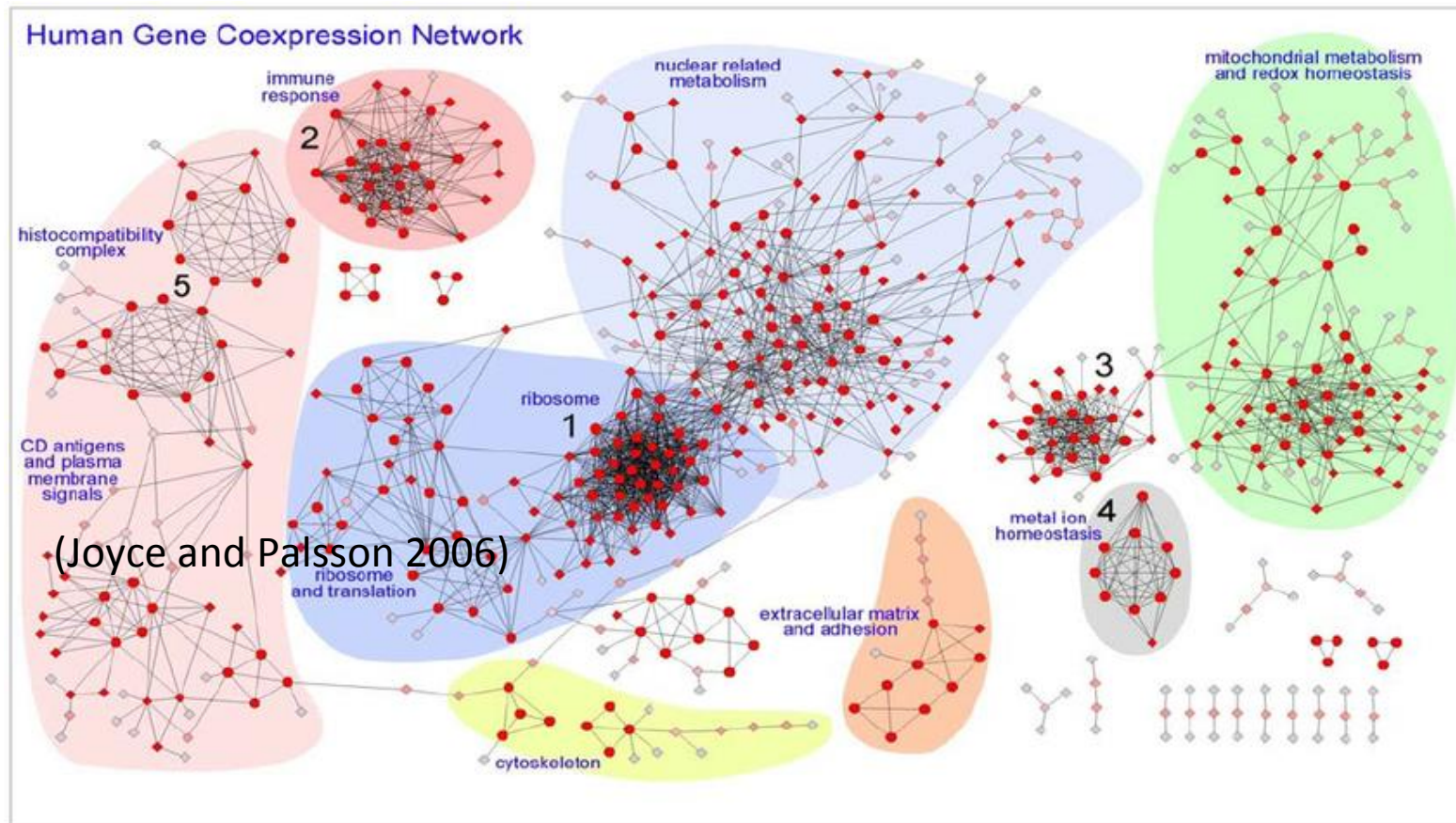


(courtesy figure Maggie Wang)



## Gene-gene interactions (epistasis)

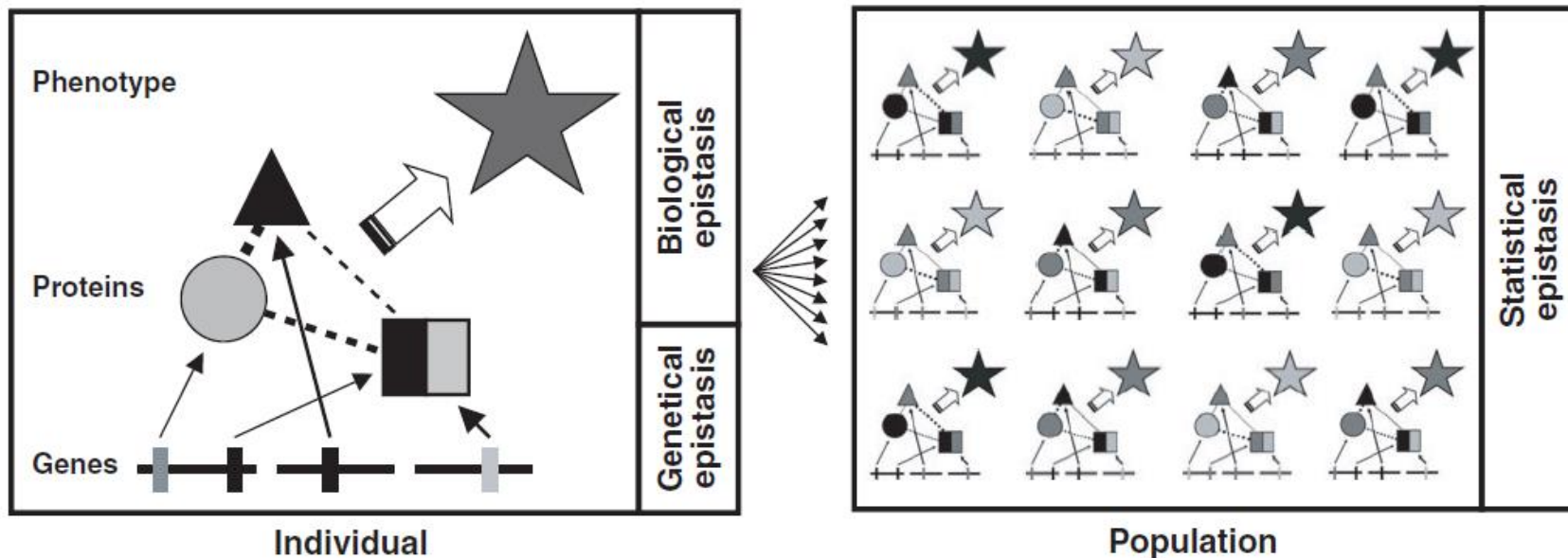
- Inference about gene-gene interactions using microarray data



(Prieto et al. 2008)

## DNA-DNA interactions

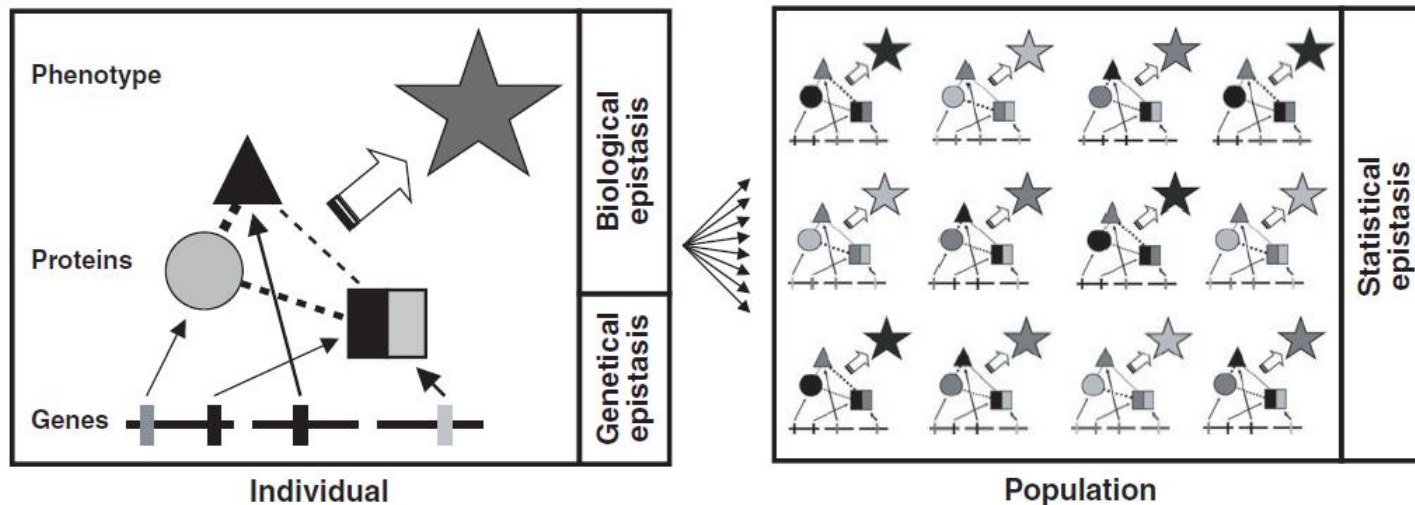
- Two or more DNA variations may “interact” either directly to change transcription or translation levels, or indirectly by way of their protein product (to alter disease risk separate from their independent effects)



(Moore 2005)

## Formal definition of epistasis

- 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).



(Moore 2005; Moore and Williams 2005)

## Correspondence between statistical and biological interactions

- Much discussion in the literature (primarily gene-gene):
  - Siemiatycki and Thomas (1981) Int J Epidemiol 10:383-387
  - Thompson (1991) J Clin Epidemiol 44:221-232
  - Phillips (1998) Genetics 149:1167-1171
  - Cordell (2002) Hum Molec Genet 11:2463-2468
  - McClay and van den Oord (2006) J Theor Biol 240:149-159
  - 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
- Conclusions: 1) little direct correspondence (physical interactions) ;  
2) statistical interaction DOES imply joint involvement

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

---

## 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  $G \times E$ :
  - 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)

---

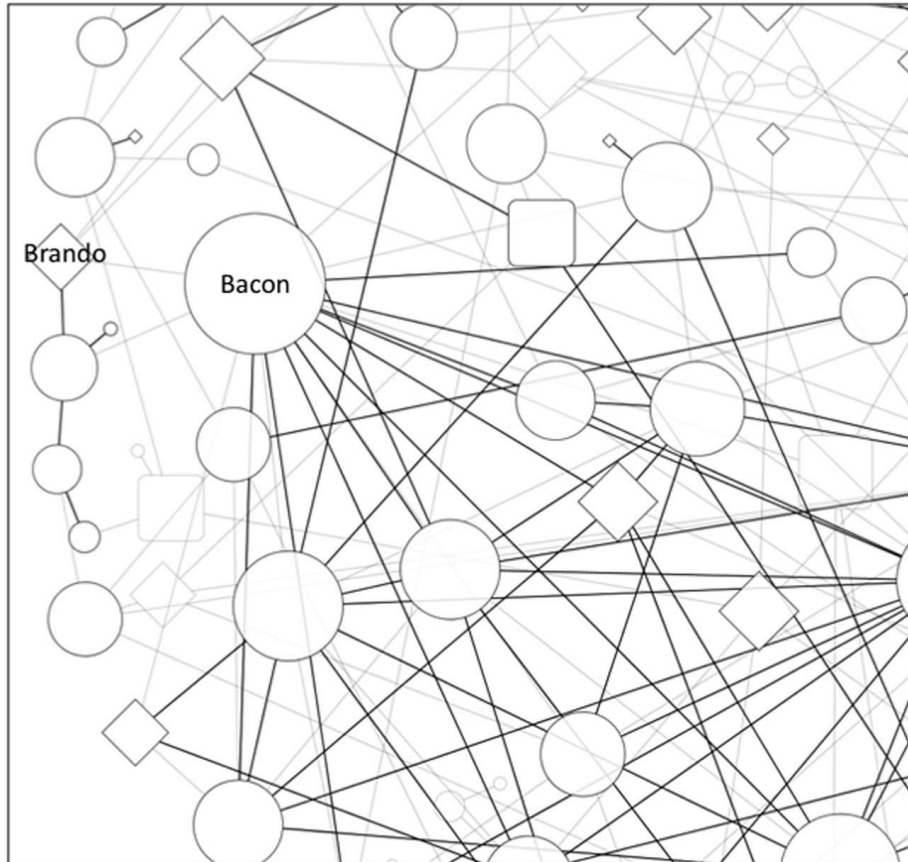
## Comparison between gene–gene and gene–environment issues

- 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<sup>1</sup>, Sharon Lutz<sup>2,\*</sup>, Bärbel Maus<sup>3,4,\*</sup>, Eric J. Duell<sup>5</sup>, Tasha Fingerlin<sup>2</sup>, Nilanjan Chatterjee<sup>6</sup>, Peter Kraft<sup>1,7</sup>, and Kristel Van Steen<sup>3,4</sup>

*Hum Genet* (2014) 133:1343–1358

DOI 10.1007/s00439-014-1480-y

REVIEW PAPER

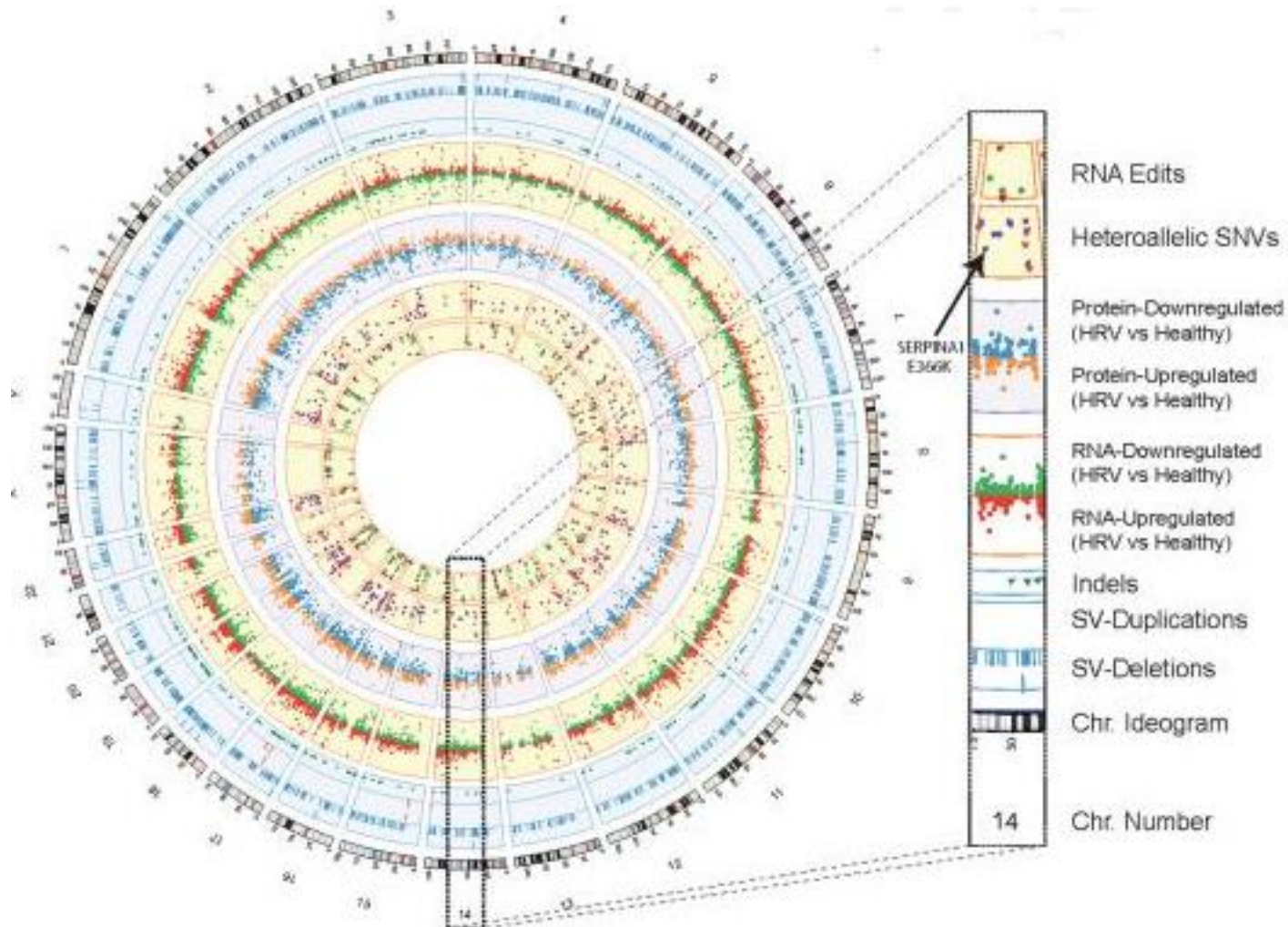
### **Practical aspects of genome-wide association interaction analysis**

Elena S. Gusareva · Kristel Van Steen

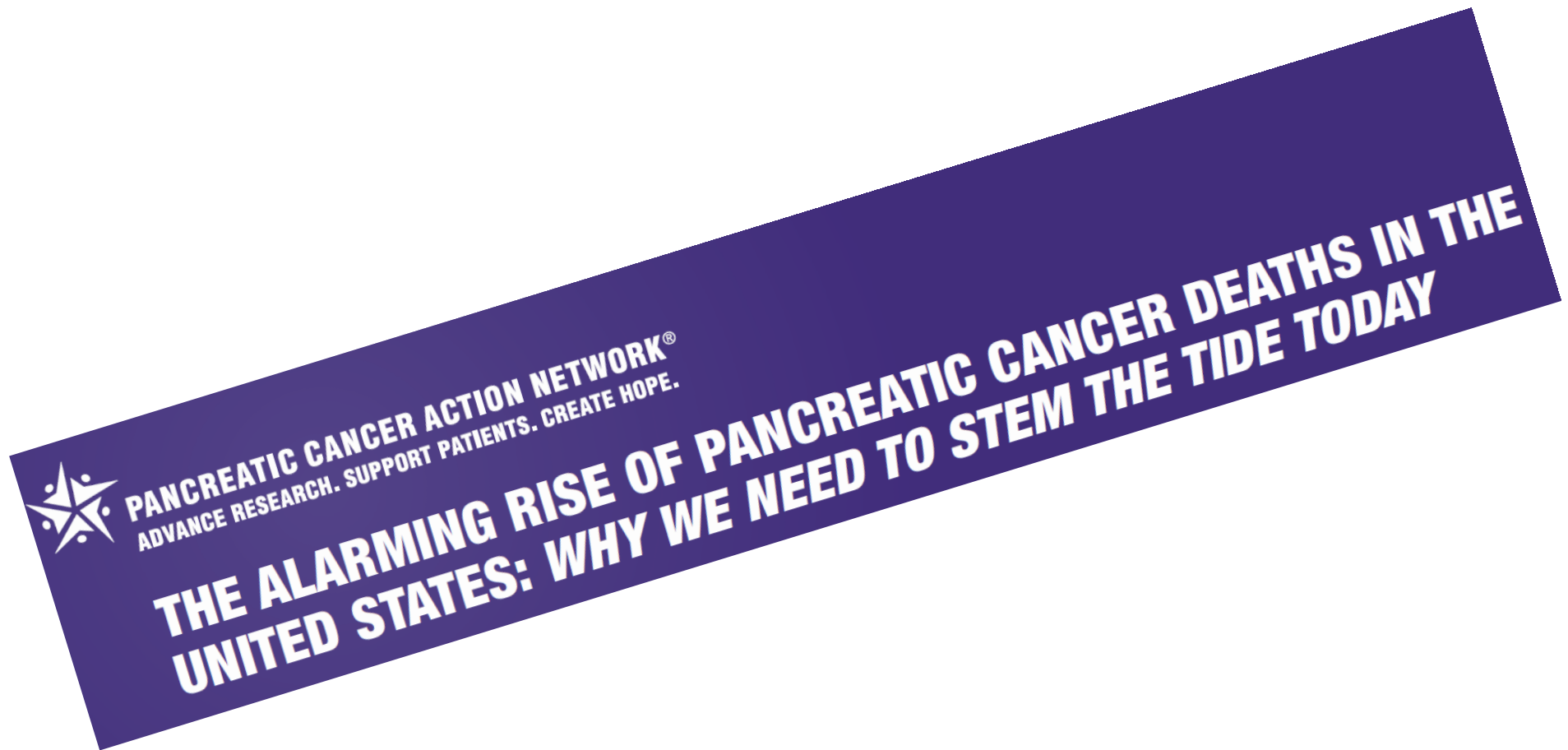
# OPPORTUNITY

# 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)

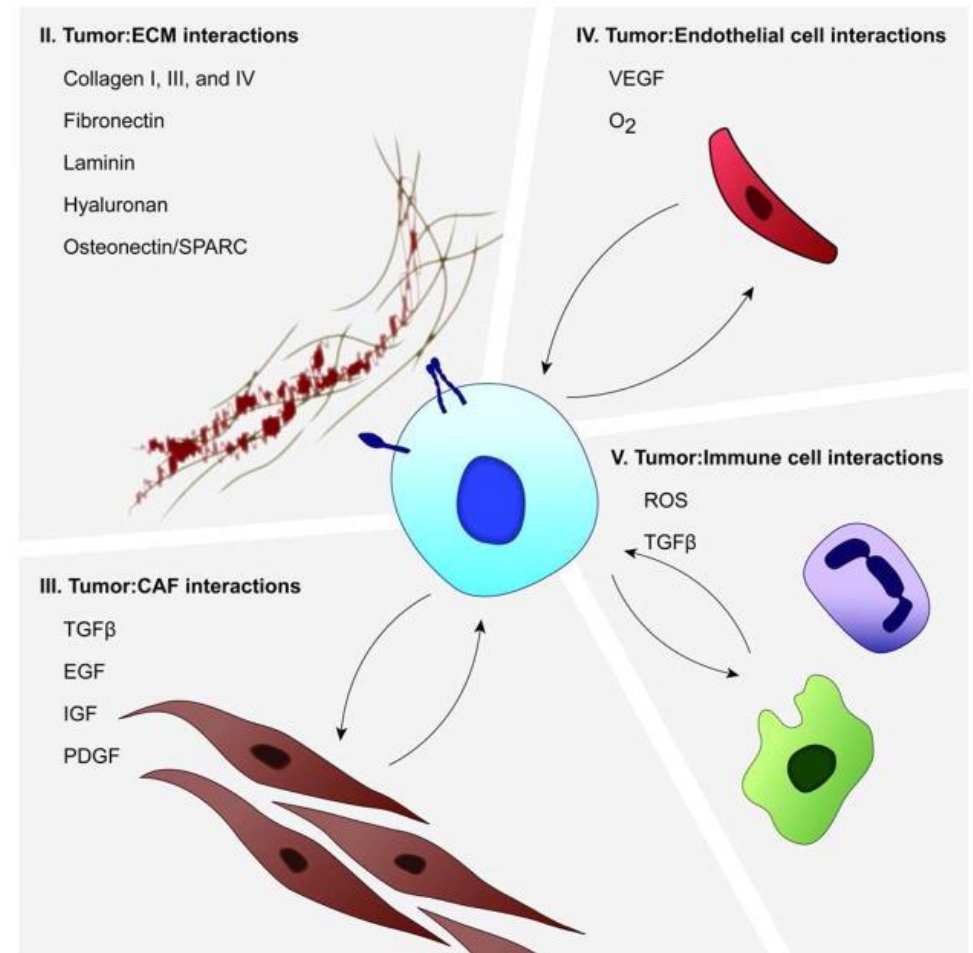
---

# 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

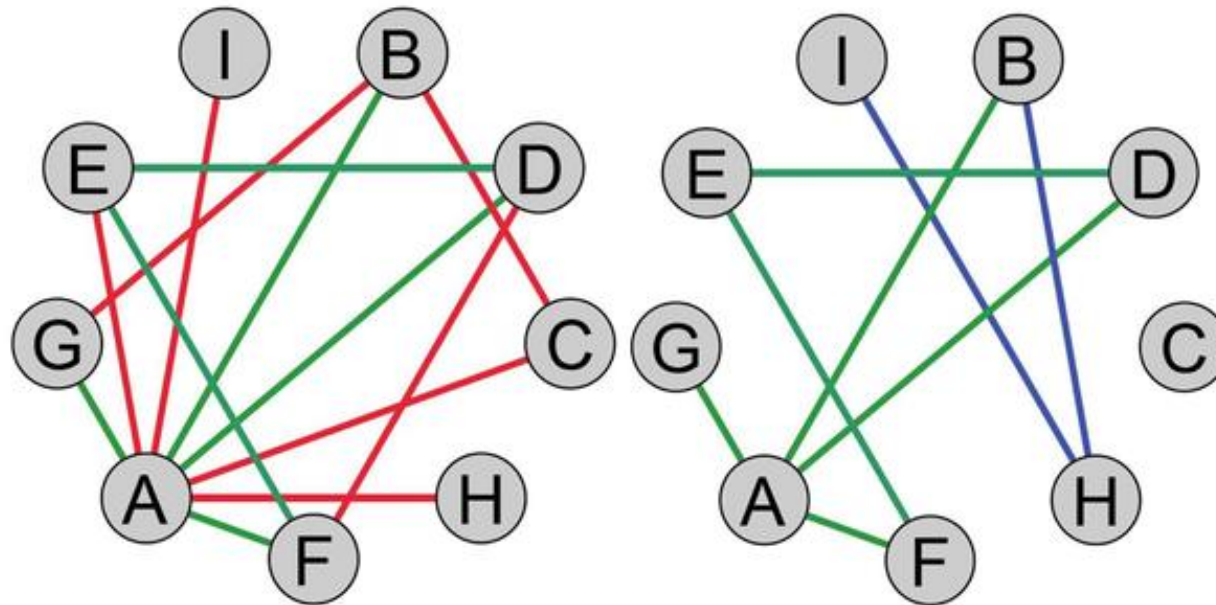
(Jansen et al. 2015)





## Gene-coexpression networks

(Anglani et al. 2014)

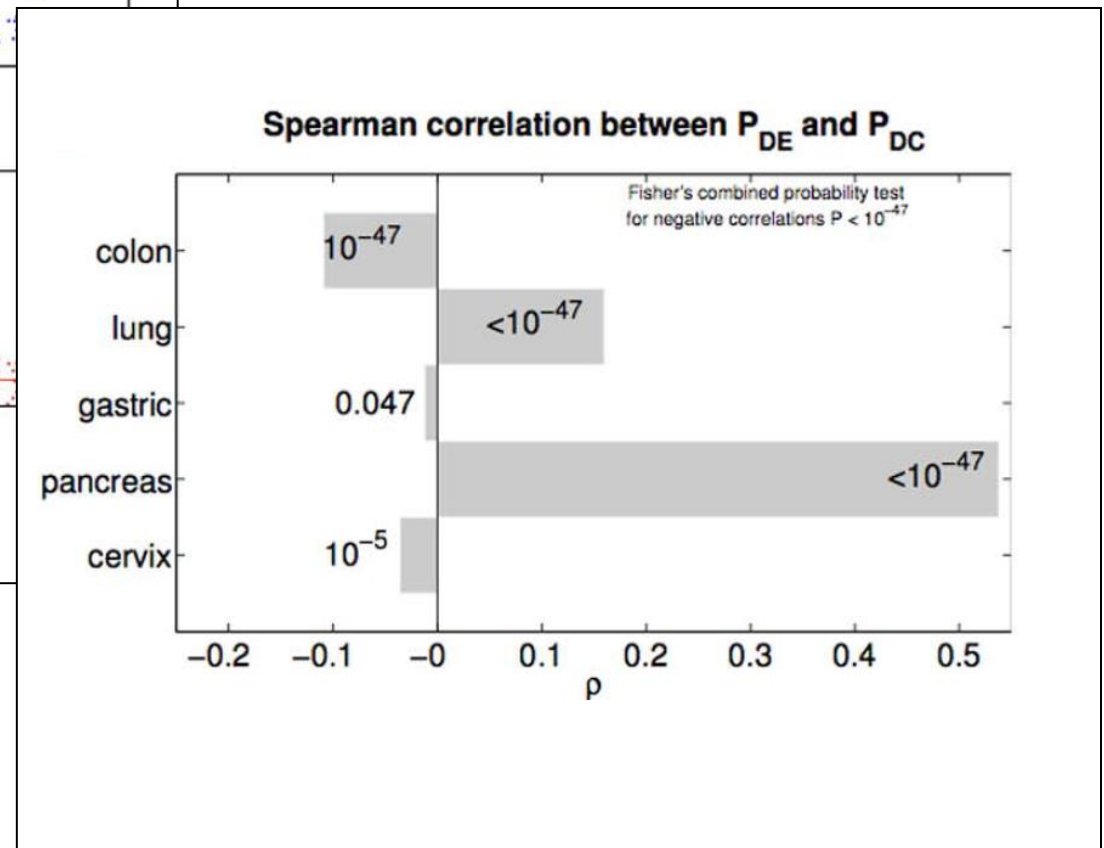
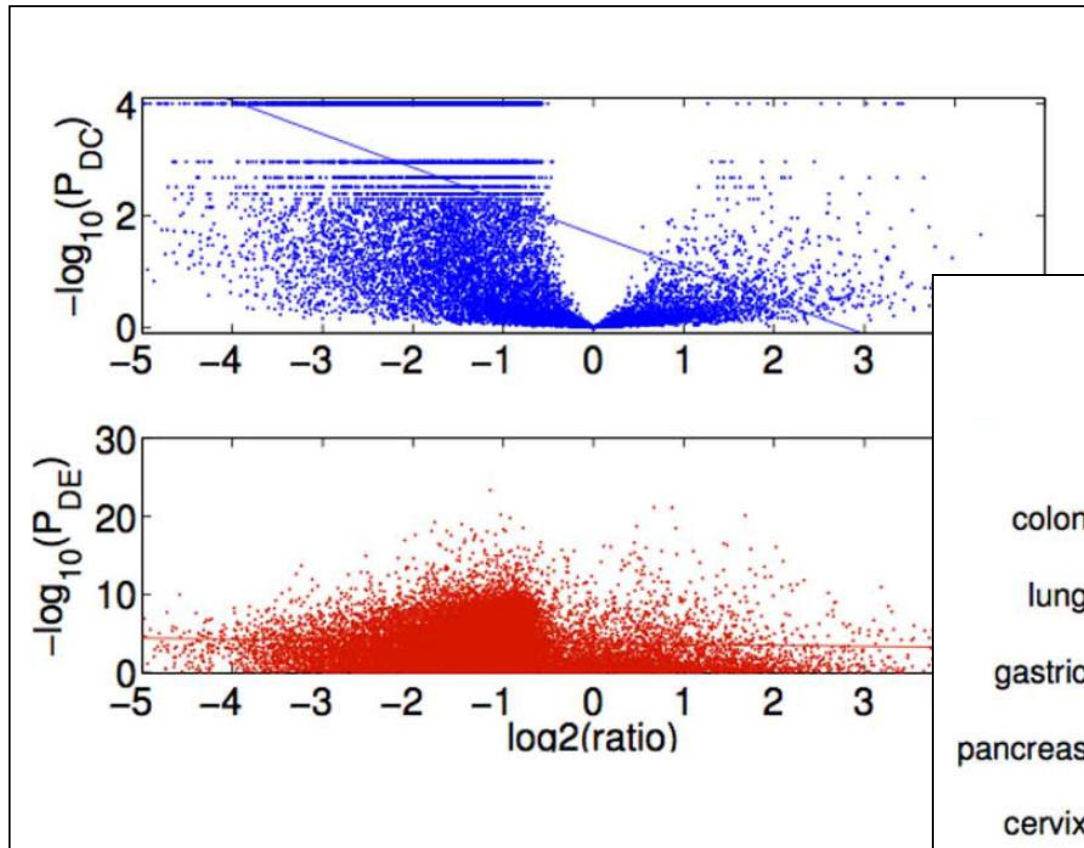


- Healthy condition on the left and disease-affected tissue on the right. Green links remain unchanged in the two phenotypes
  - Red connections are loss from healthy to cancer network
  - Blue edges are novel connections in the cancer tissue
-

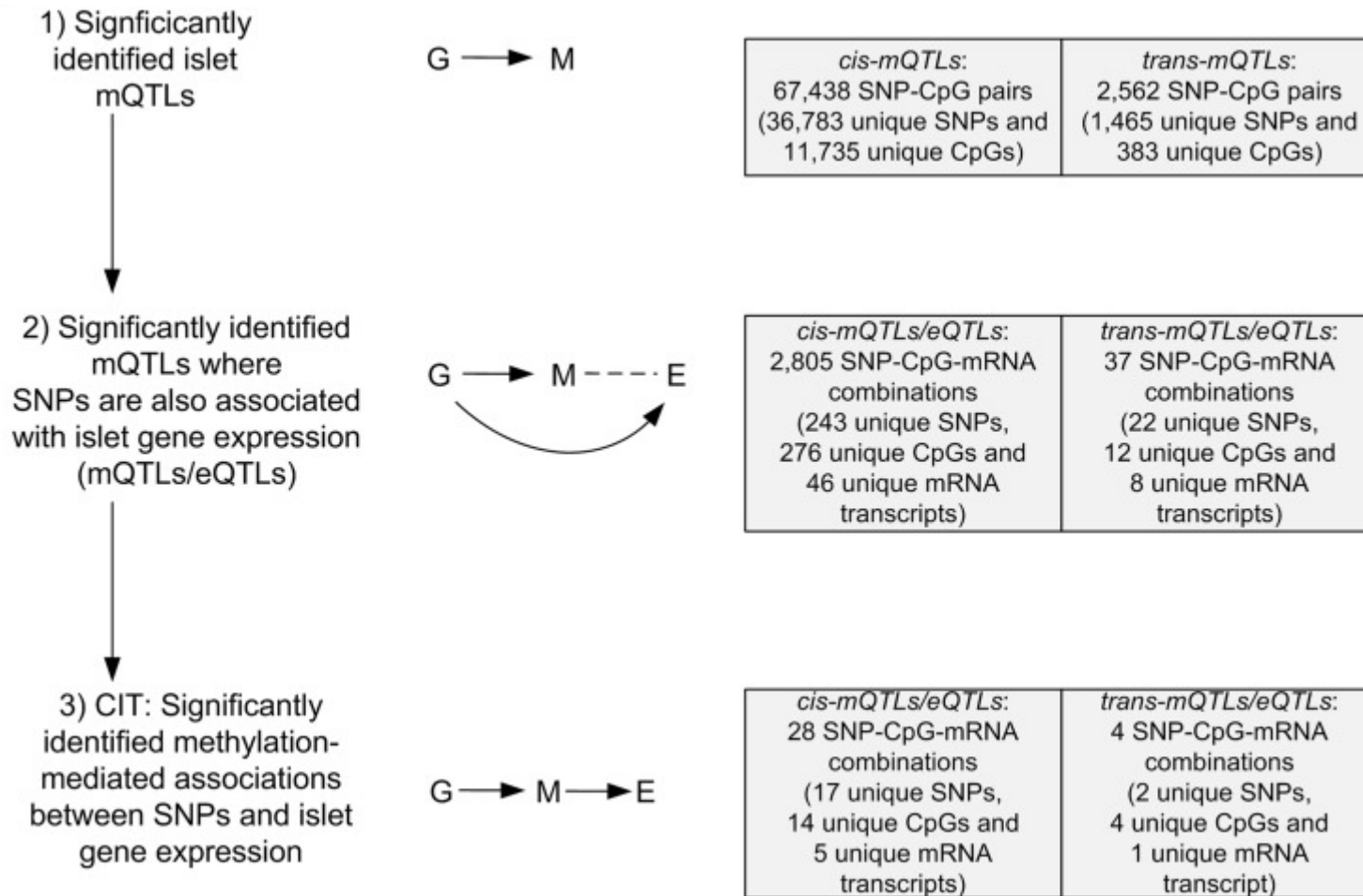


# Gene co-expression networks

(Anglani et al. 2014)





## Genetic-epigenetic mechanistic interactions (pancreatic islets)



(Olsson et al. 2014)

## Gene-gene interactions using SNPs?

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

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



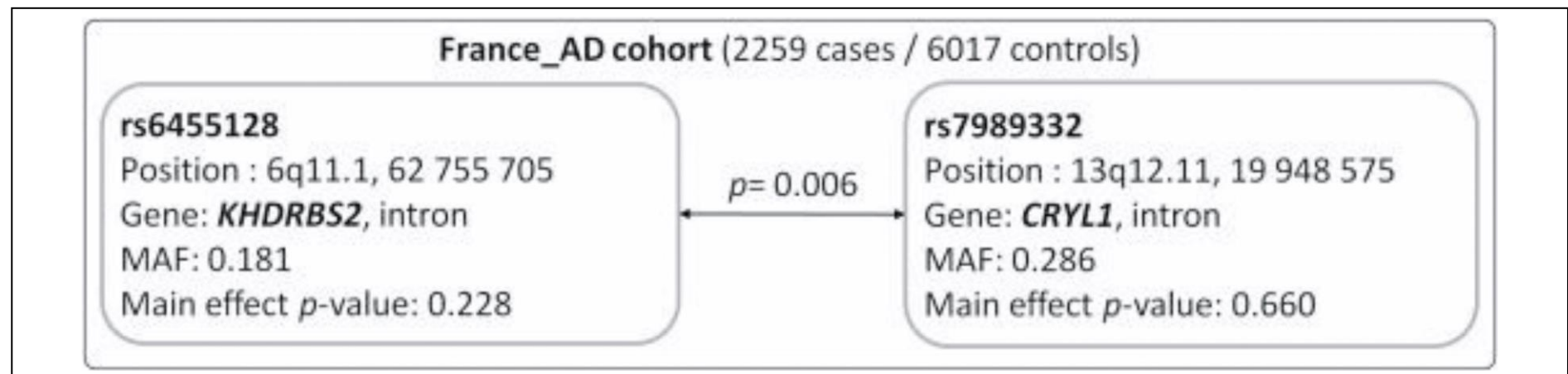
## Epistasis appearance versus detection

- Examples of DNA-DNA interactions from model organisms (Carlborg and Haley 2004):
    - Epistatic QTLs without individual effects have been found in various organisms, such as birds, mammals, *Drosophila melanogaster* 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.
  - Indicates the complexity with which multifactorial traits are regulated; no single mode of inheritance can be expected to be the rule in all populations and traits...
-

## The only source of knowledge is experience (Albert Einstein)

### Genome-wide association interaction analysis for Alzheimer's disease

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



(Gusareva et al. 2014)

## The only source of knowledge is experience (Albert Einstein)

### **Genome-wide association interaction analysis for Alzheimer's disease**

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

(Gusareva et al. 2014)

*“This particular study in particular demonstrates an effective approach to elucidate the functional repercussions of epistasis”*

(Ebbert et al. 2015)

# MEANS

*Although there is growing appreciation that attempting to map genetic interactions in humans may be a fruitful endeavor, there is no consensus as to the best strategy for their detection, particularly in the case of genome-wide association where the number of potential comparisons is enormous*

(Evans et al. 2006)

---

## One popular method singled out: (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 ( $\beta_0$ ), main effects of locus  $G$  ( $\beta_{G1}, \beta_{G2}$ ), main effects of locus  $H$  ( $\beta_{H1}, \beta_{H2}$ ), 4 interaction terms
- This corresponds in statistical analysis packages to coding  $X1, X2$  (0,1,2) as a “factor”

Locus G	Locus H		
	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	$\beta_0 + \beta_{H2}$	$\beta_0 + \beta_{H1}$	$\beta_0$

## One popular method singled out: (logistic) regression

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

		Locus H		
		2	1	0
Locus G	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	$\beta_0 + 2\beta_H$	$\beta_0 + \beta_H$	$\beta_0$

corresponding to the model (**X1 and X2 coded as (0,1,2)**):

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


---

## Concerns with (logistic) regression models

- There is no direct correspondence between the interaction effects and the underlying penetrance based models displaying some kind of epistasis effect (North et al. 2005) → interpretation and biological translation
  - Standard regression techniques are hampered by inflated false positives, and diminished power caused by the presence of sparse data and multiple testing problems, even in small simulated data sets only including 10 SNPS (Vermeulen et al. 2007) → assessing significance
  - Unknown confounders or wrong model assumptions: model misspecification → robust modeling
-



## Investigating multi-locus correlations instead

- Test whether correlation is different in cases and controls via testing a “log OR” for association between two loci
- Examples :
  - Fast epistasis (PLINK)

Locus G	Locus H			Locus G	Locus H	
	2	1	0		$H_1$	$H_2$
2	$a$	$b$	$c$	$G_1$	$A = 4a + 2b + 2d + e$	$B = 4c + 2b + 2f + e$
1	$d$	$e$	$f$	$G_2$	$C = 4g + 2h + 2d + e$	$D = 4i + 2h + 2f + e$
0	$g$	$h$	$i$			

See Ueki and Cordell (2012) for a correct variance estimation for the log(OR) (CASSI / PLINK 9.1)

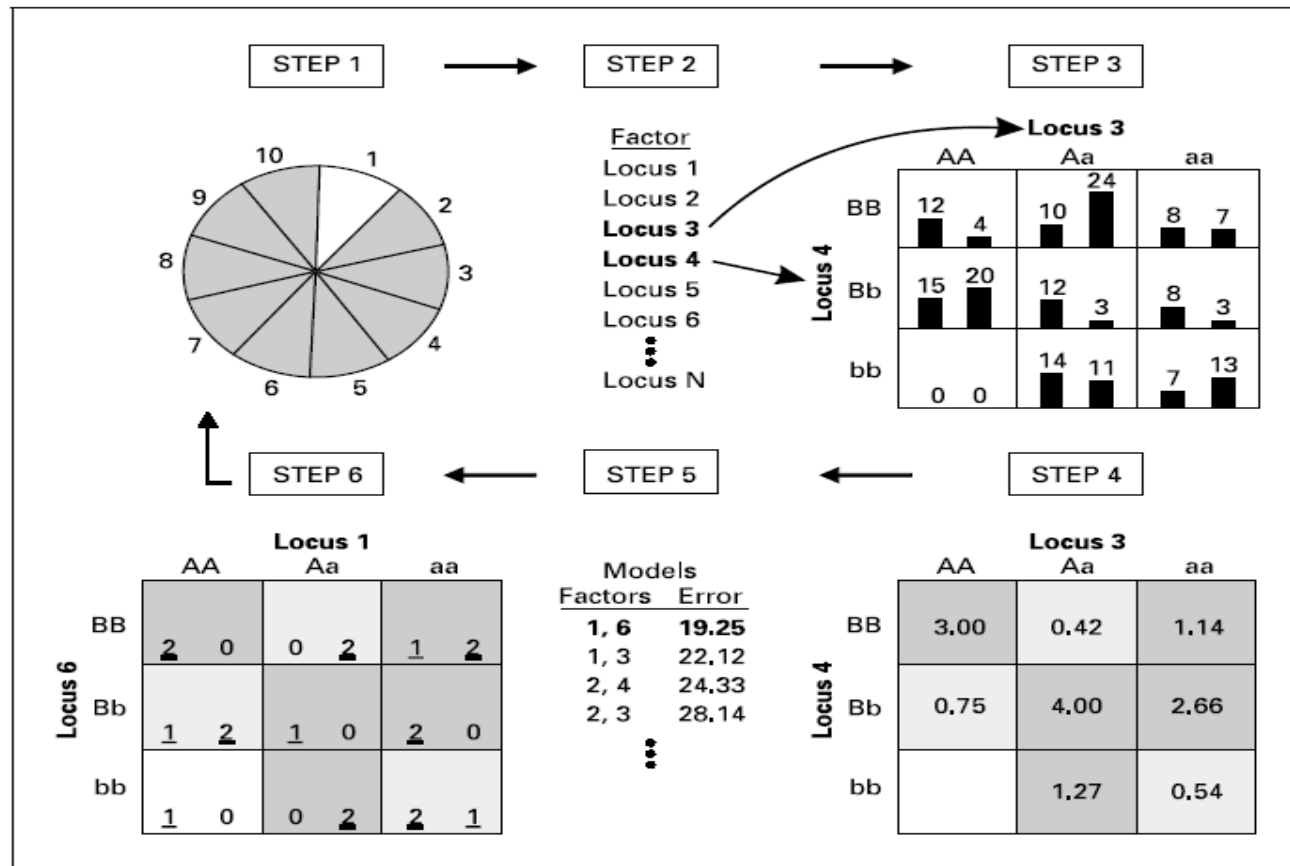
- EPIBLASTER (Kam-Thong et al. 2011)

# **Model-Based Multifactor Dimensionality Reduction (MB-MDR)**

---

# Historical notes about MB-MDR

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



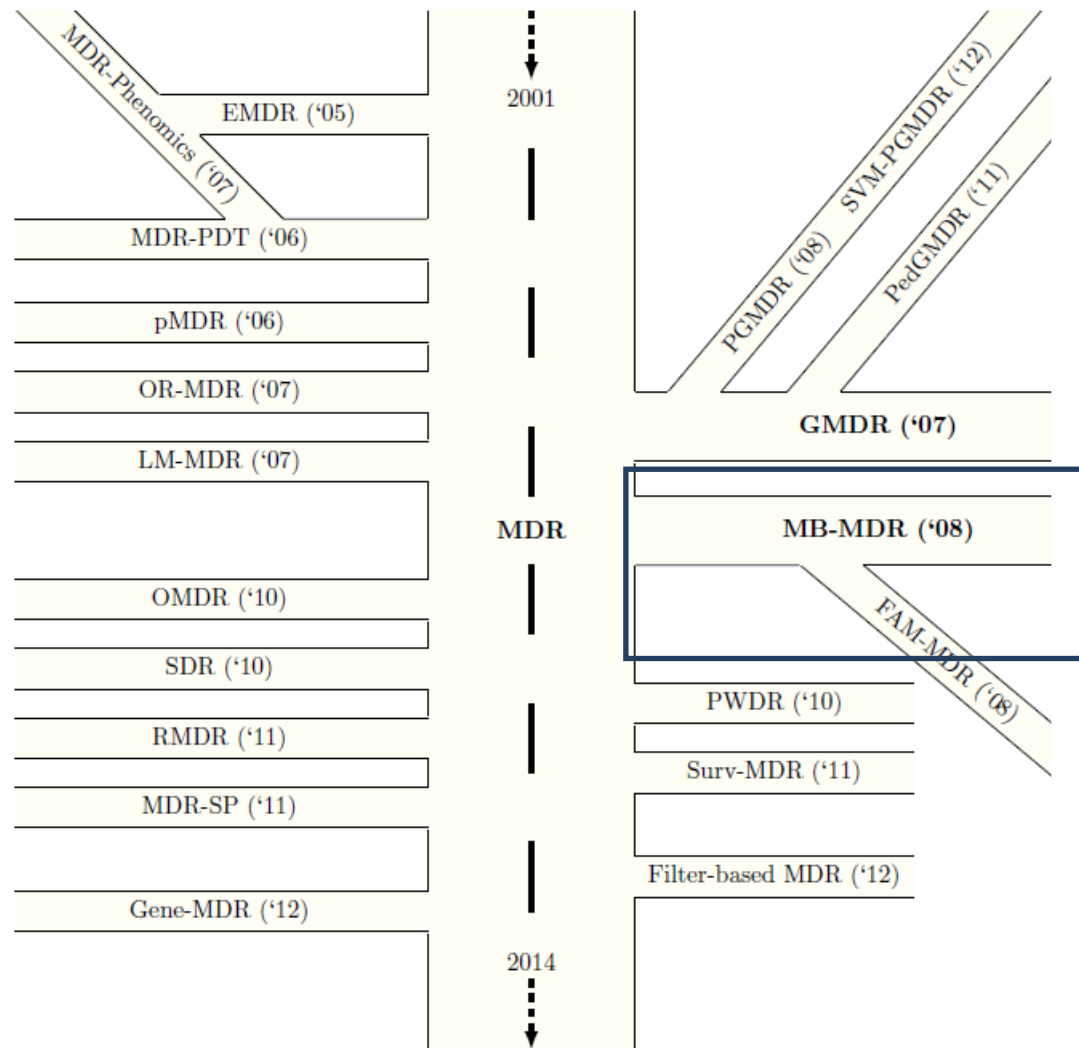
## Which dimensions are reduced?

- 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

Method	Number of Factors $K$		
	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)

# Several MDR roads lead to Rome



(Gola et al. 2015)

## Shift from prediction to association

- Model-Based MDR by Calle et al (2008a)
    - Rather, computation time is invested in optimal **association tests** to prioritize multilocus genotype combinations and in statistically valid permutation-based methods to assess **joint statistical significance**
    - Results of association tests are used to “label” multilocus genotype cells (for instance: increased / **no evidence**/ reduced risk, based on sign of “effect”) and to “quantify” the multilocus signal wrt the trait of interest, “**above and beyond** lower order signals”
-

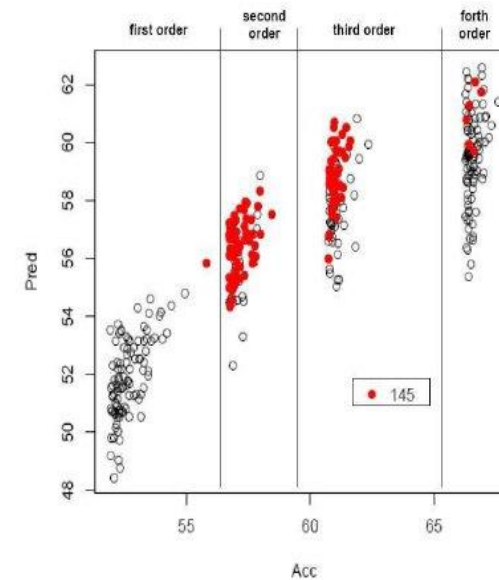
## Global versus specific modeling

- 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

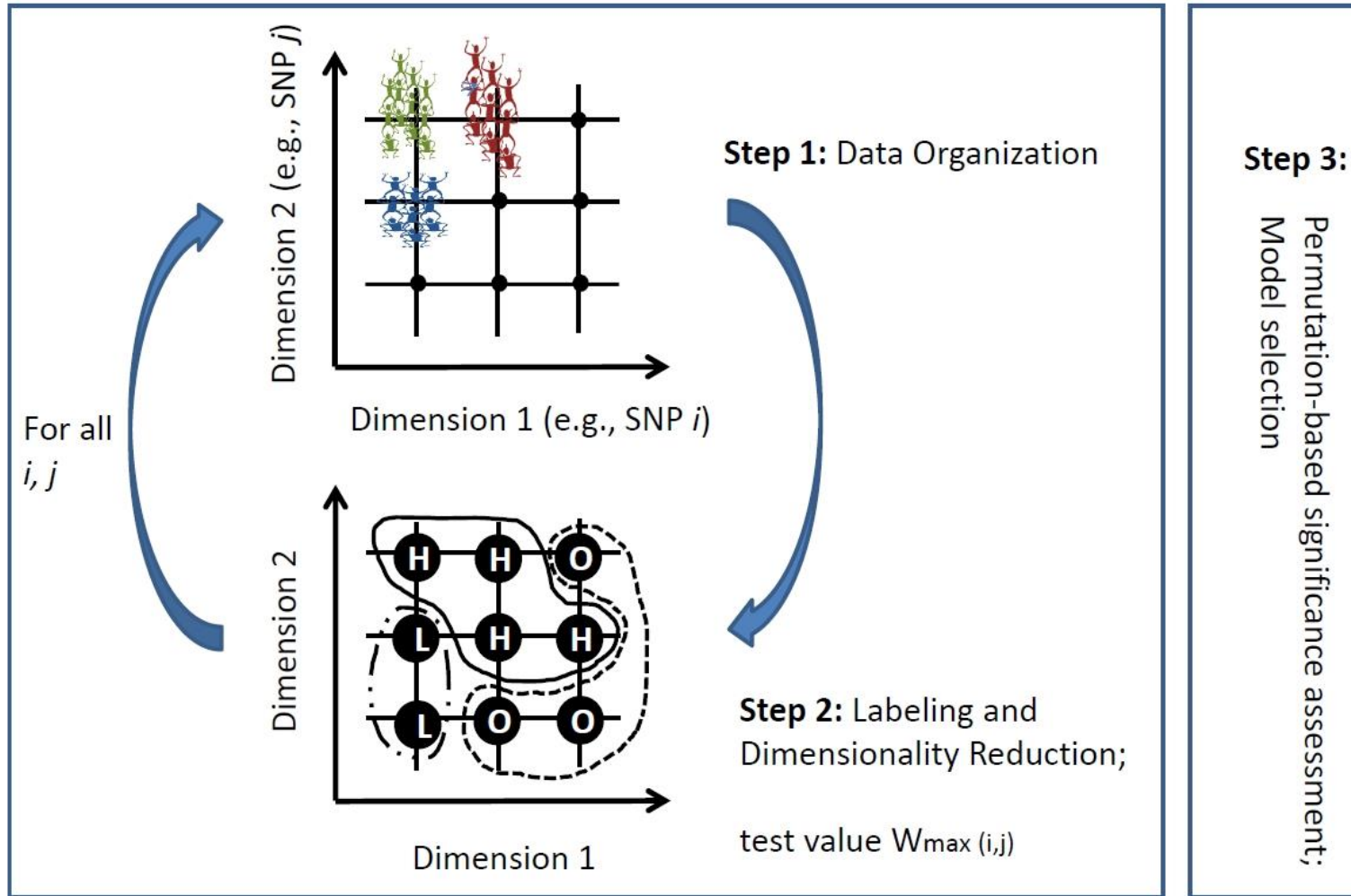
SNP 40 x SNP 252 genotypes	Cases	Controls	OR	p-value	Category
c1 = (0,0)	88	77	1.01	0.9303	0
c2 = (0,1)	102	114	0.73	0.0562	L
c3 = (0,2)	38	34	0.98	1.0000	0
c4 = (1,0)	50	59	0.76	0.1229	0
c5 = (1,1)	96	37	2.68	0.0000	H
c6 = (1,2)	18	28	0.55	0.0675	L
c7 = (2,0)	12	6	1.99	0.3399	0
c8 = (2,1)	14	18	0.67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

# Model-Based Multifactor Dimensionality Reduction (MB-MDR)





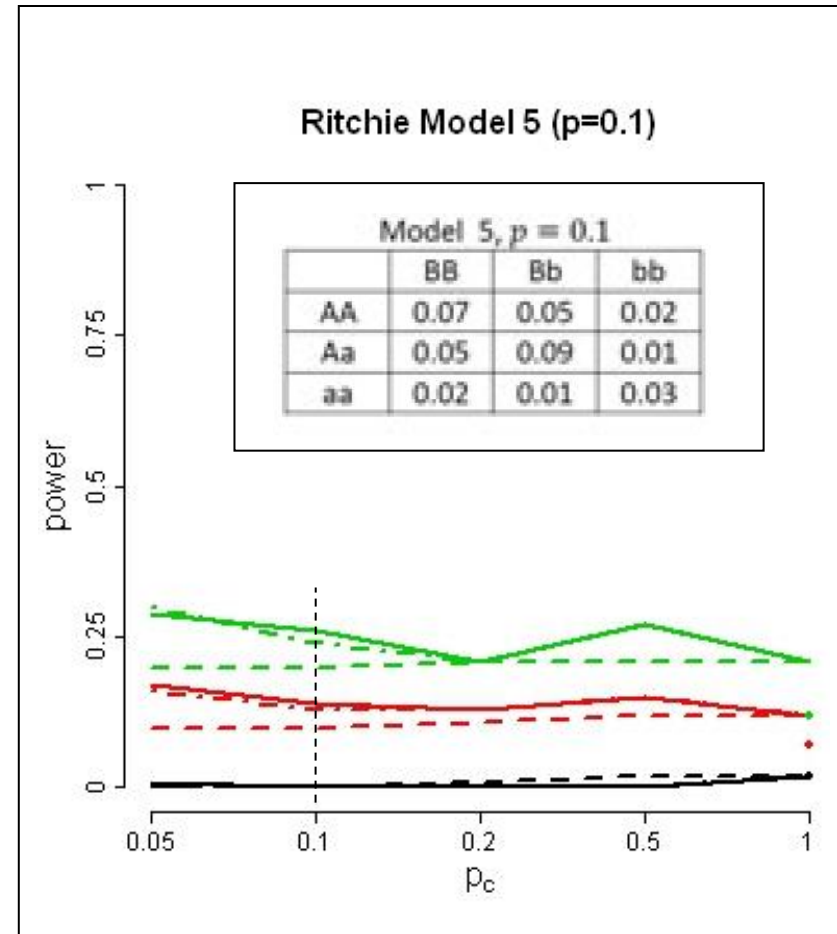
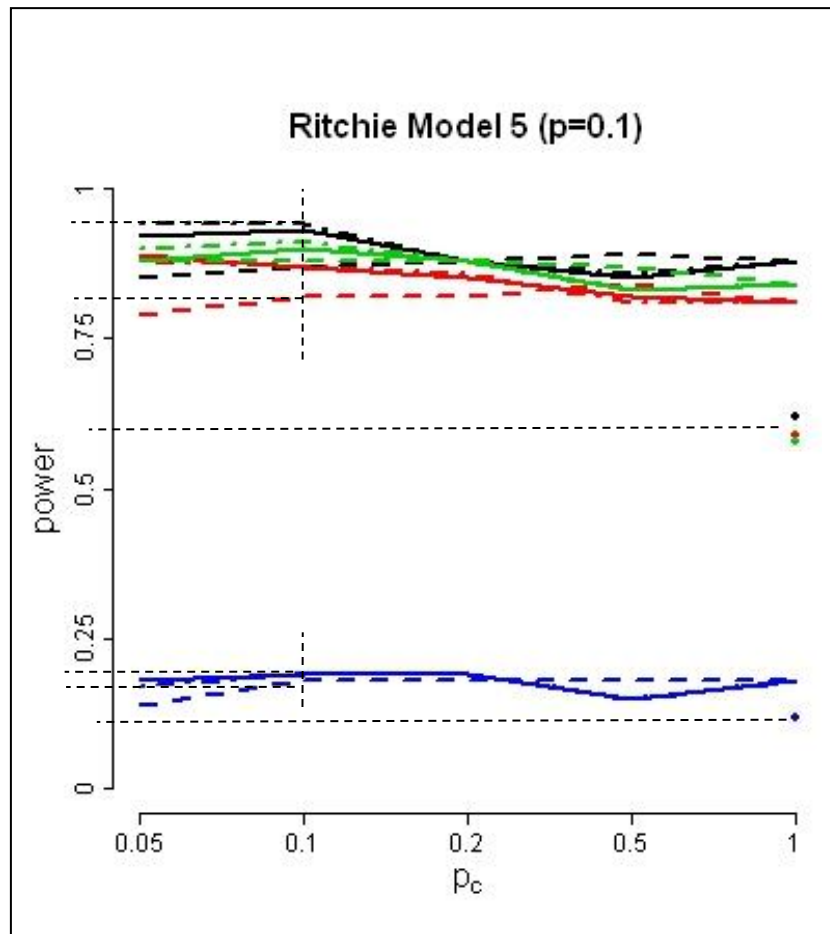
## MB-MDR

Summary of the steps involved in MB-MDR analysis:

- For every  $k$  variates (e.g., SNPs), Step 1 is a data organization step in which individuals are (naturally) allocated to  $k$ -dimensional (genotype) profiles.
  - Step 2 labels individuals according to their profiles in multi-dimensional space and liberal association tests. Individuals with the same label are merged into a single group.
  - Extreme groups are contrasted to each other via an association test, leading to a test value  $W_{max}$  for the selected  $k$ -tuple.
  - The final  $k$ -models are selected in Step 3, using permutation-based significance assessments and adequate multiple testing control.
-

# Power performance (absence/presence of genetic heterogeneity)

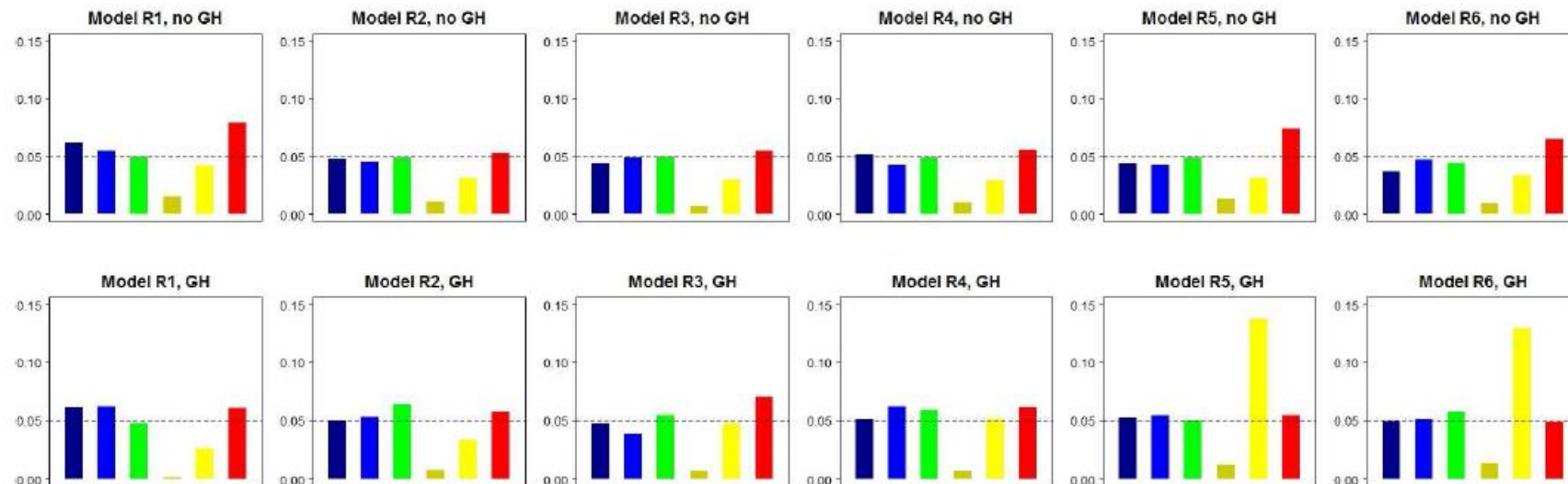
- Model-Based MDR by Cattaert et al. 2011



## 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)

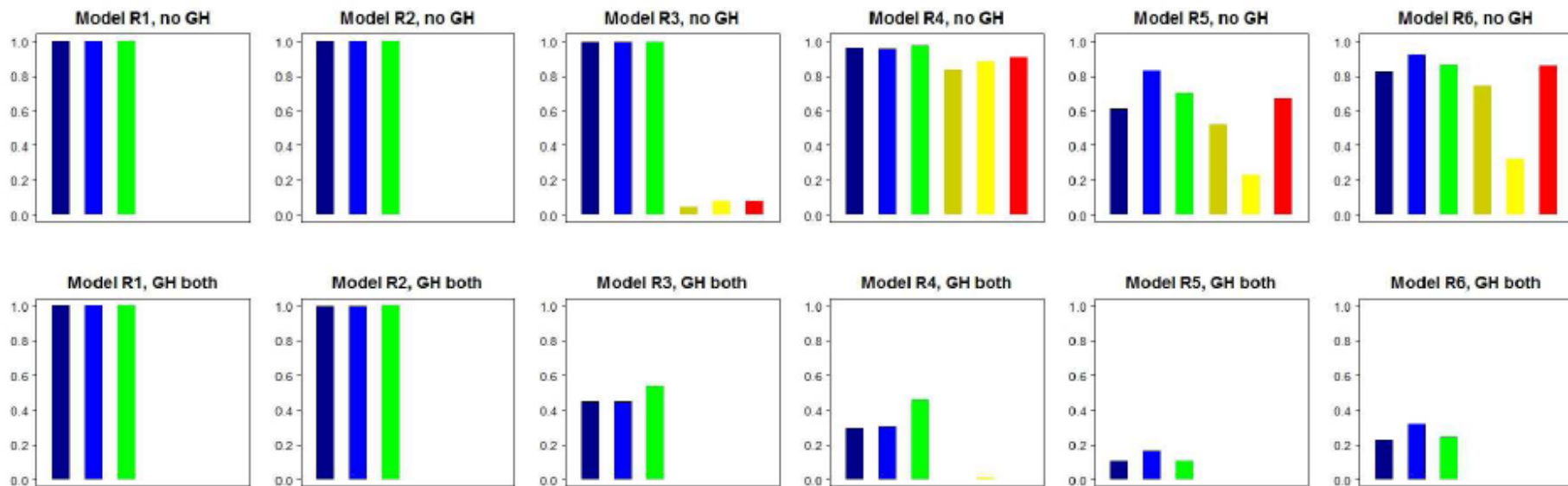
PLINK fast epistasis (light yellow)

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)

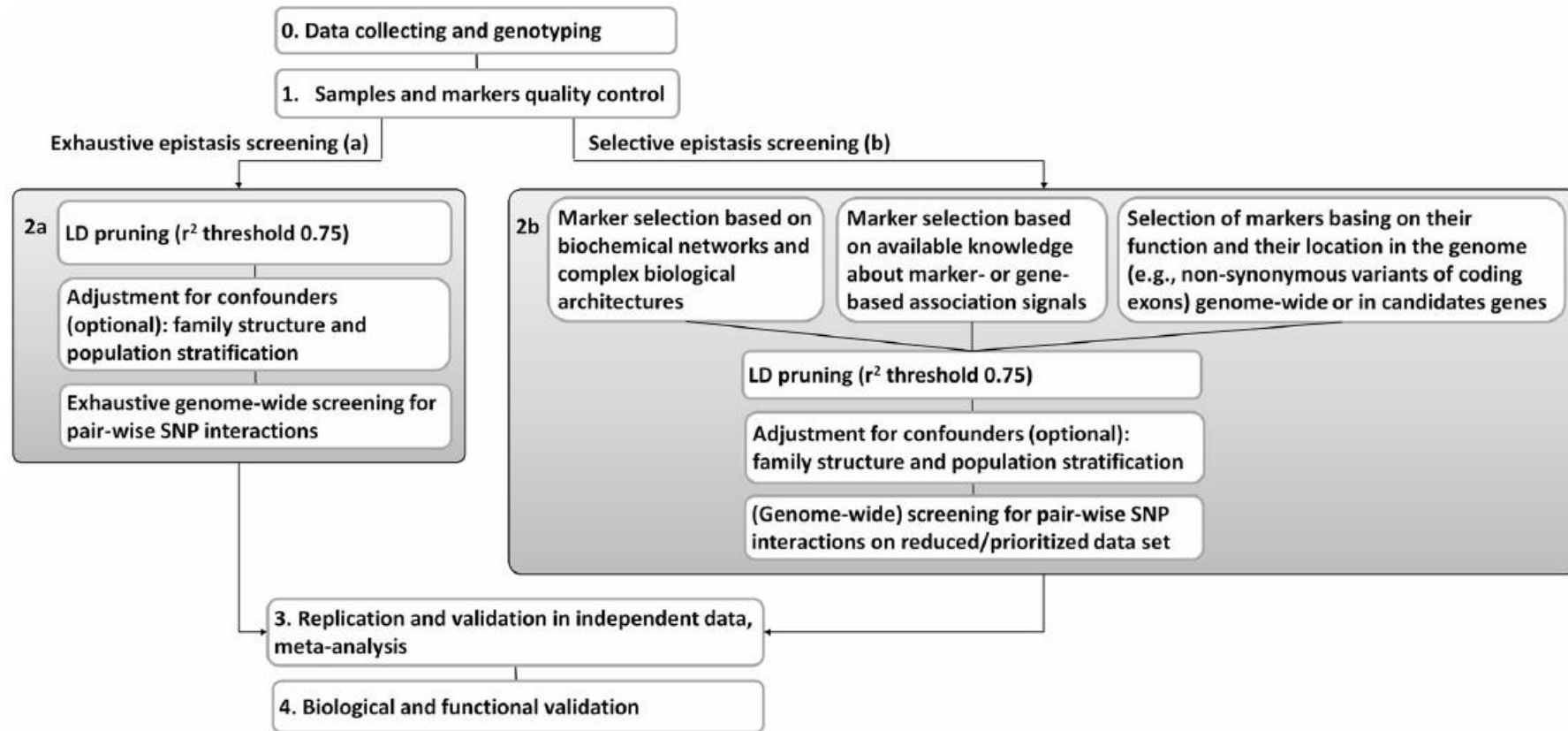
PLINK epistasis (dark yellow)

PLINK fast epistasis (light yellow)

EPIBLASTER (red)

# Towards a GWAI Protocol

# The BIO3 GWAI protocol



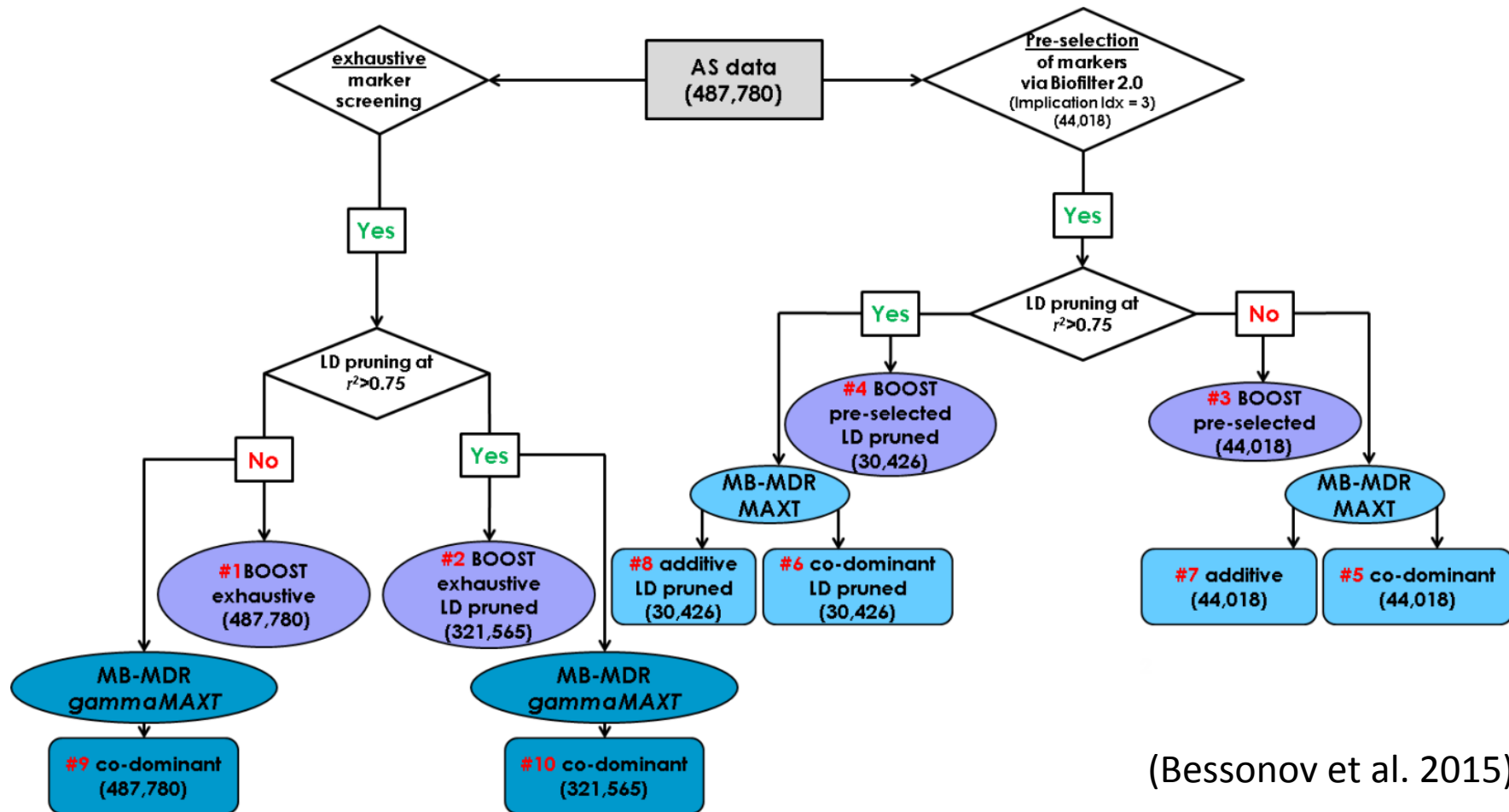
(Gusareva et al. 2014)

These critical steps are paramount to the *success* of GWAI studies

## Prior biological information

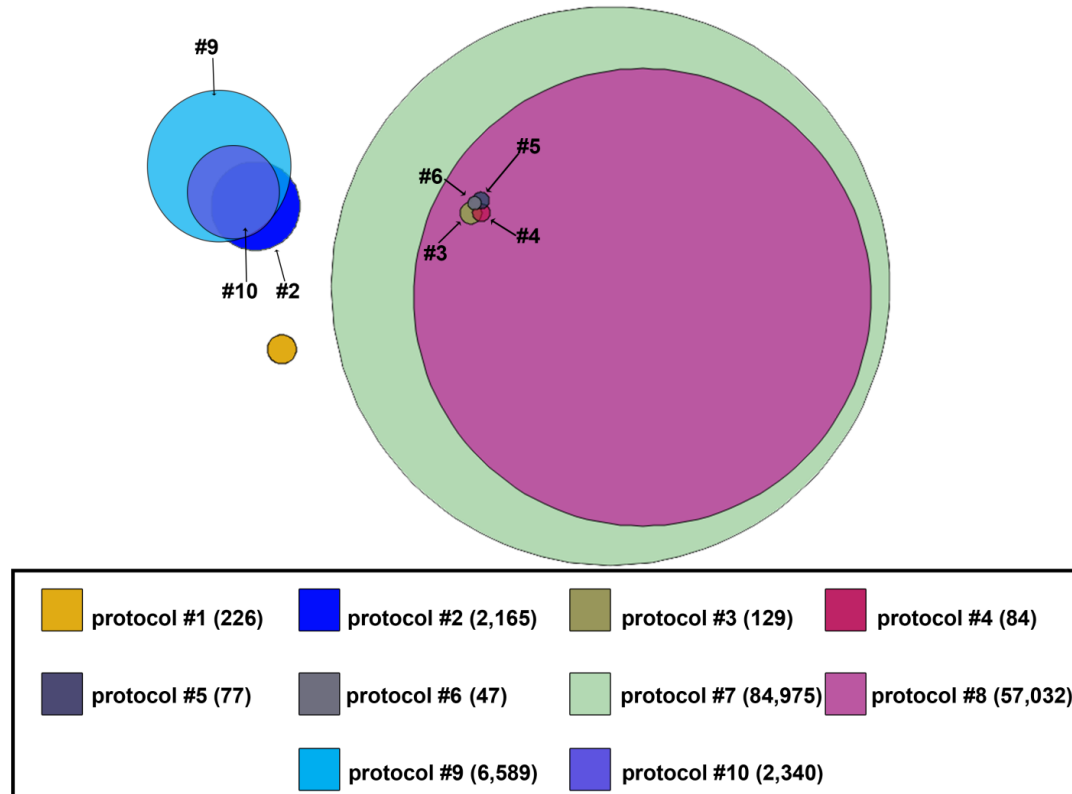
- Some researchers incorporate prior biological “knowledge”:
    - Allow for uncertainty involved in the data source entries
    - Acknowledge the complementary characteristics of each of the available data sources
    - Think about the “significance” of evidence scores
  - The advantage is reduced data dimension and potentially saving costs
  - The draw-back is to be restricted by the biological assumption: hypothesis-driven versus hypothesis generating analysis
-

# Slight protocol changes may lead to huge differences in results





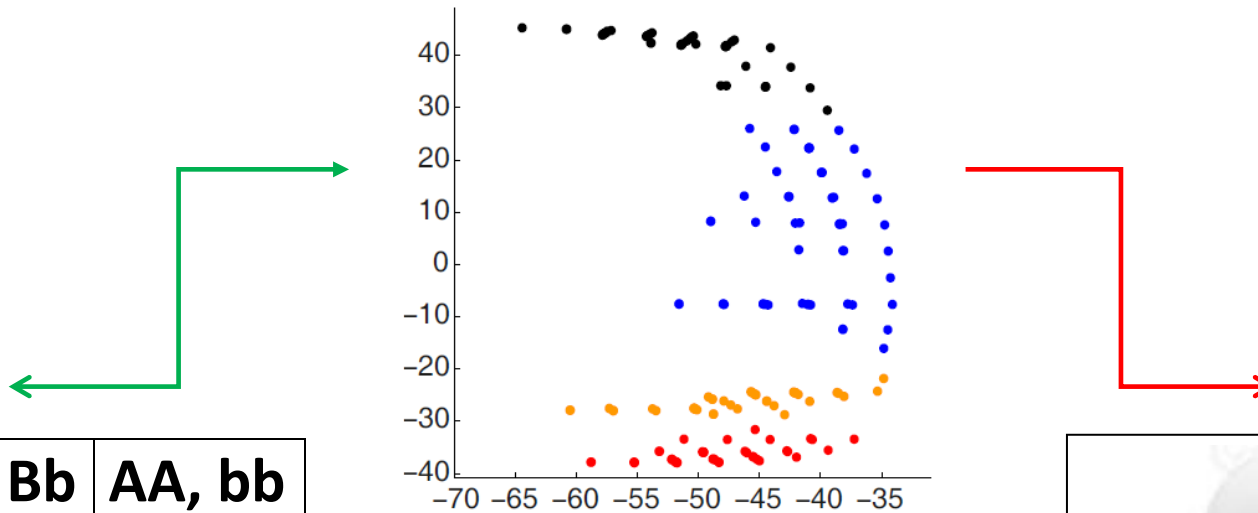
## Slight protocol changes may lead to huge differences in results



- **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]**

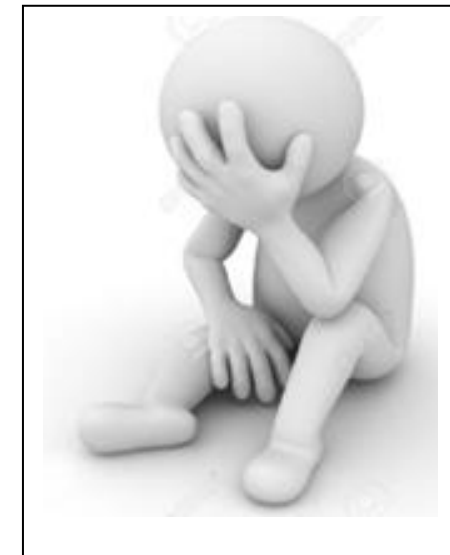
# Confounding by shared genetic ancestry

(MB-MDR and pair-specific genomic control – Van Lishout)



<b>AA, BB</b>	<b>AA, Bb</b>	<b>AA, bb</b>
<b>Aa, BB</b>	<b>Aa, Bb</b>	<b>Aa, bb</b>
<b>aa, BB</b>	<b>aa, Bb</b>	<b>Aa, bb</b>

Multilocus Exposure

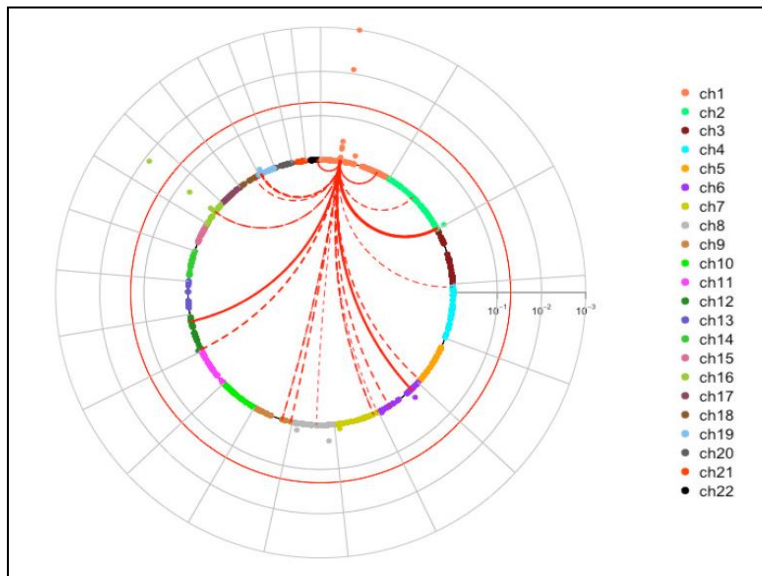


Disease Outcome



## Replication

*“Leaving aside for the moment **what replication means** or should mean in the context of GWAS, even for the currently so-called replicated genetic interactions it is unclear to what extent **a false positive has been replicated** due to the adopted methodological strategy itself or whether the replication of epistasis is not solely attributed to main effects (such as HLA effects) not properly accounted for.”*



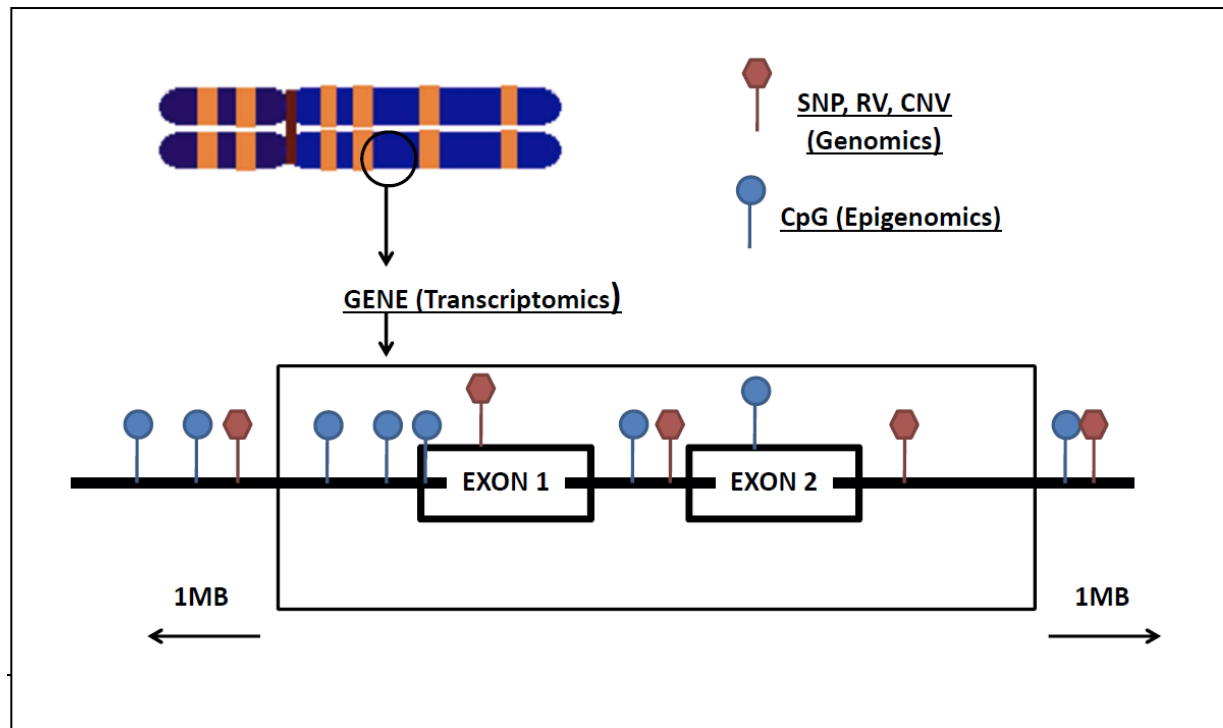
(Ritchie and Van Steen 2016)

**MogPlot** (Van Lishout)

## Replication

*“Genome-wide SNP genotyping platforms consist predominantly of **tagSNPs** from across the genome. Most of these SNPs are not causal and have no functional consequences. **When two or more tagSNPs are combined in a genetic interaction model, is it reasonable to assume that the same combination of tagSNPs interacts in an independent dataset?**”*

(Ritchie and Van Steen 2016)



(Pineda, CNIO/Ulg 2015)

# TAKE HOME MESSAGES



## Learning from data (synthetic + real-life)

- **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**]
  - **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 (2012) 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, Théâtre E, Charlotiaux B, Calle ML, Wehenkel L, Van Steen K (2013) An efficient algorithm to perform multiple testing in epistasis screening. *BMC Bioinformatics* 14:138 [**C++ MB-MDR made faster!**]
- **Van Lishout F**, Gadaleta F, Moore JH, Wehenkel L, Van Steen K (2015) gammaMAXT: a fast multiple-testing correction algorithm *BioData Mining* 8:36 [**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. *Human Heredity* – accepted [**aggregating based on similarity measures to deal with DNA-seq data**]

**Contact:** [f.vanlishout@ulg.ac.be](mailto:f.vanlishout@ulg.ac.be) (C++ MB-MDR software engineer)

---

## Learning by data summary

- Backpack items on the road less travelled by, include:

Item	Our label
Speed controller	Gamma MaxT (Van Lishout et al.2015 – submitted)
Population/patient substructure or (cryptic) relatedness chart	MB-MDR for structured populations (Van Lishout et al. 2013 – poster ASHG, manuscript in preparation)
Correlated input features map	Component-based Path Modeling (PLS-PM; Esposito Vinzi @ ERCIM2014 short course)
Replication / Meta-analysis toolkit	Easier to do when units of analysis are at a higher level (such as genes instead of {SNPs, epigenetic markers, miRNAs, ...}) (Gusareva et al. 2014 – GWAI protocol)

# Thank You



“If we put our minds and resources together we will be able to improve things for all those suffering with this disease [pancreas cancer]. The EU has a prominent role to play”

(Françoise Grossetête, MEP)

## KRISTEL VAN STEEN



GIGA-R Medical Genomics - BIO3 Unit  
**University of Liège**  
Belgium



<http://www.statgen.ulg.ac.be/>  
<http://www.montefiore.ulg.ac.be/~kvansteen/>  
Google Scholar, Research Gate

---