Beyond GWAS

Opportunities and Challenges of Large-Scale Epistasis Screening

Kristel Van Steen, PhD[2] (*)

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

Department of Human Genetics (Systems Medicine), KULeuven, Leuven, Belgium

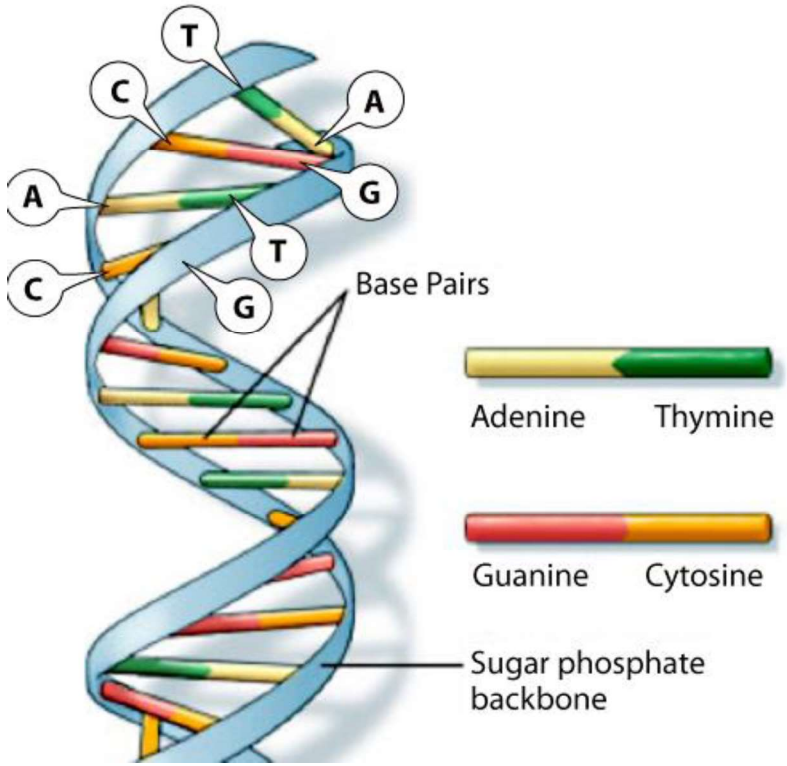
Outline

- Why unveiling interactions?
- How to identify interactions?
- Challenges?
  - Toy analytic example: MB-MDR
- Take-home messages

Why unveiling interactions?

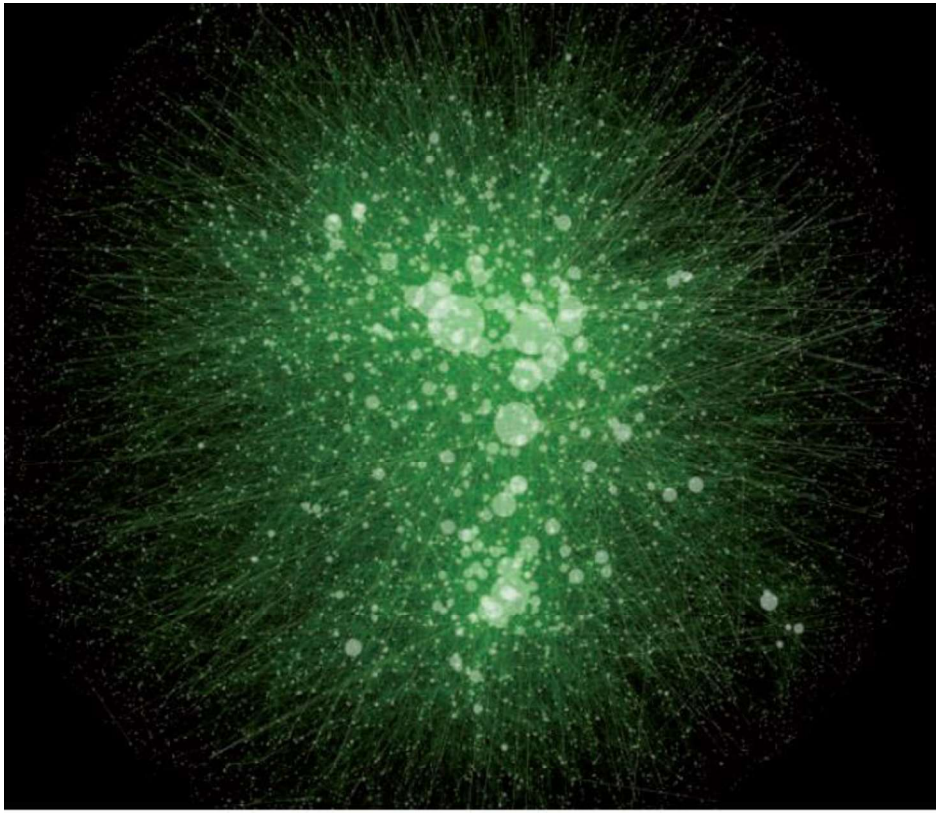
Differences between human genomes

- Any two people plucked at random off the street are on average 99.9 percent the same, DNA-wise (> 3 million positional differences)
- Most genome variations are relatively small and simple, involving only a few bases—an A substituted for a T here, a G left out there, a short sequence such as CG added somewhere else



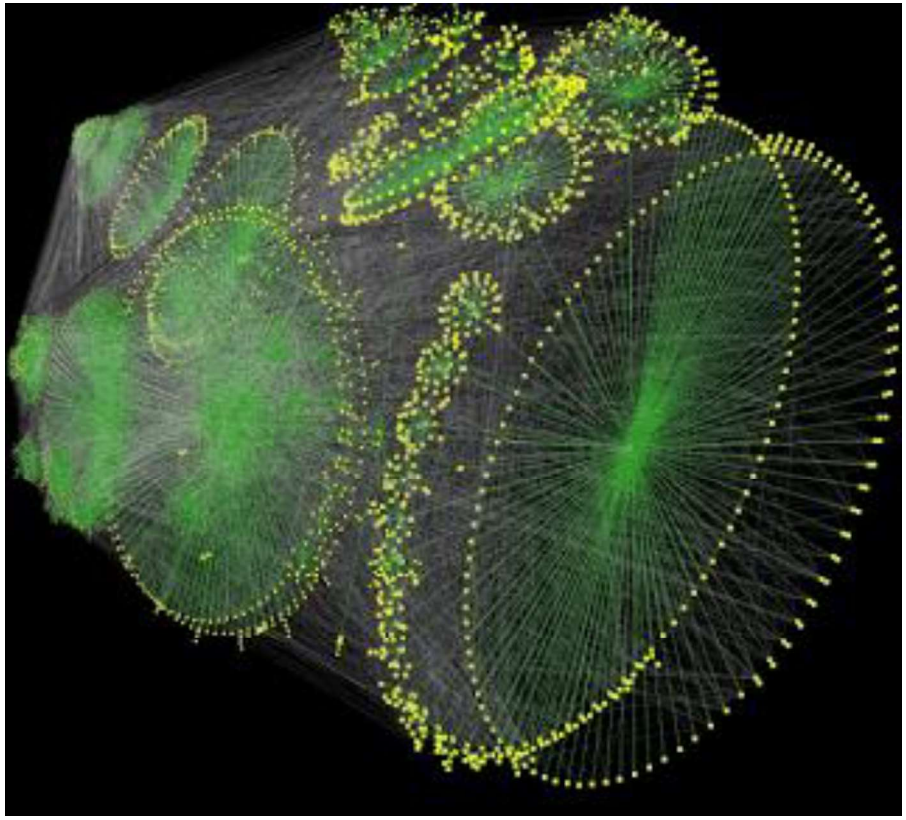
(U.S. National Library of Medicine)

Interactome differences between organisms



Human interactome (PPI)

(Bonetta 2010)



Fruit fly interactome

(owww.molgen.mpg.de)

Human interactome differences and complex diseases

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

(Moore and Williams 2009)

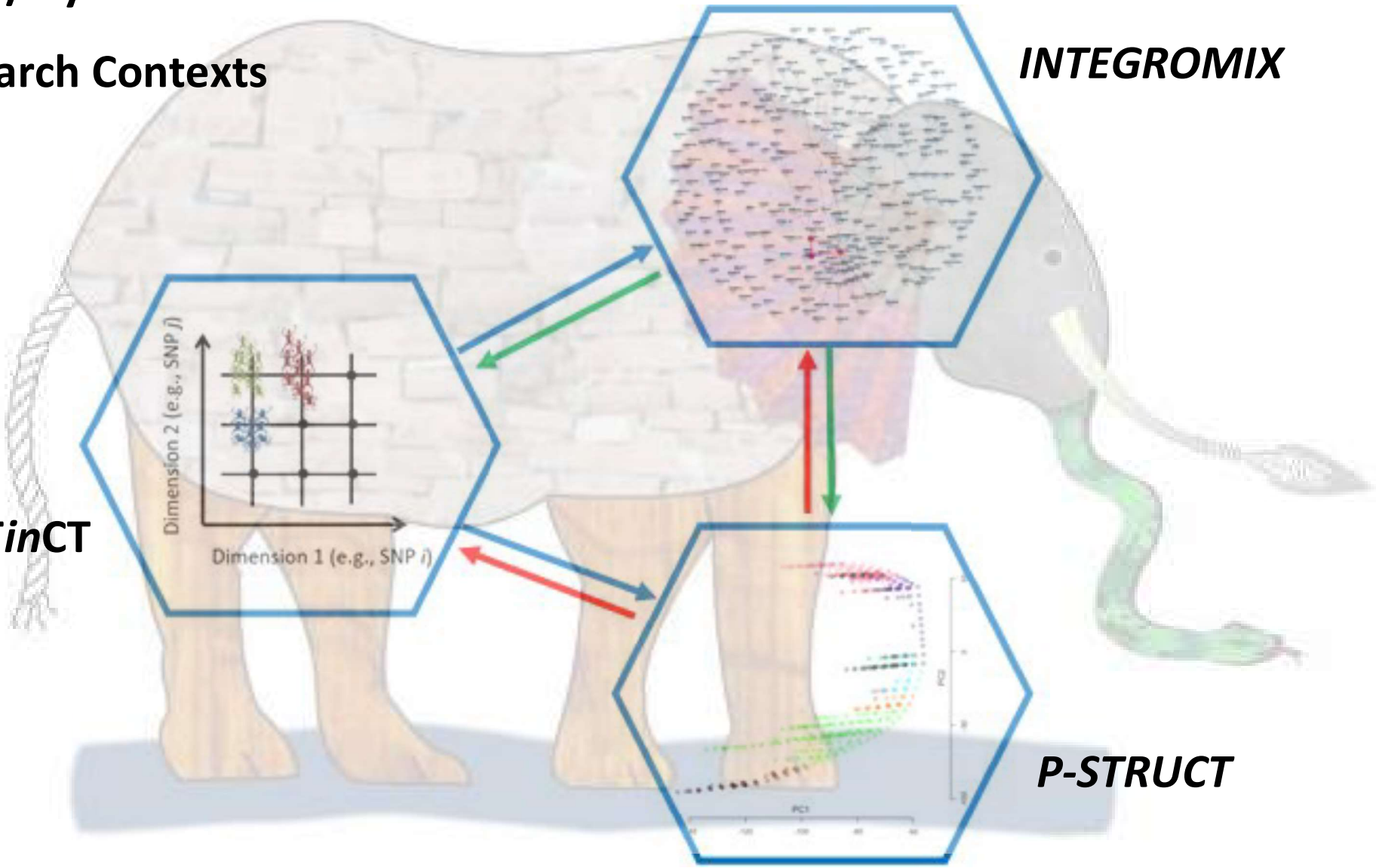
- The consequence is an underlying genetic architecture that is comprised of networks of genes that are redundant and robust

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***.

BIO3 / Systems Medicine

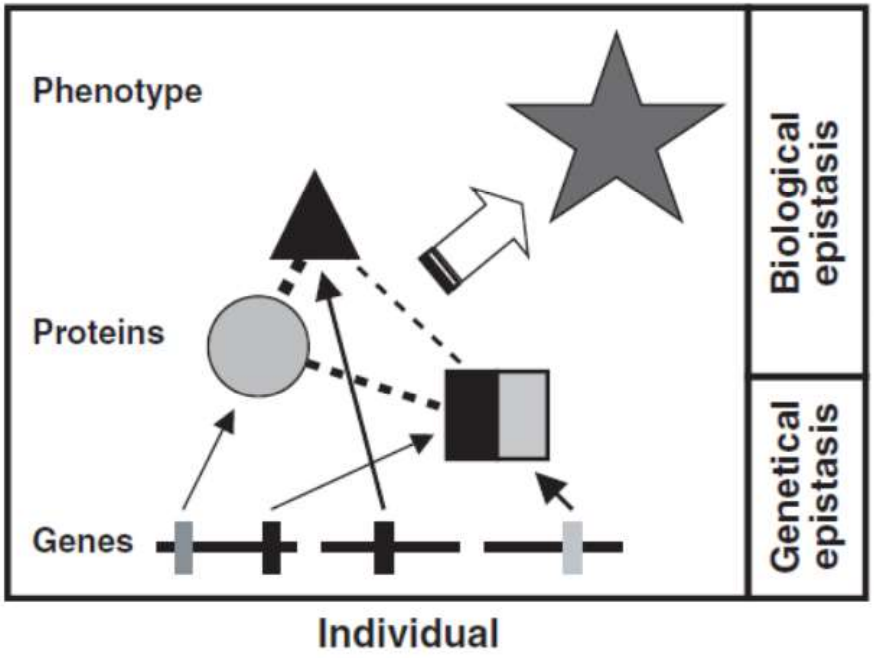
Research Contexts



How to identify interactions?

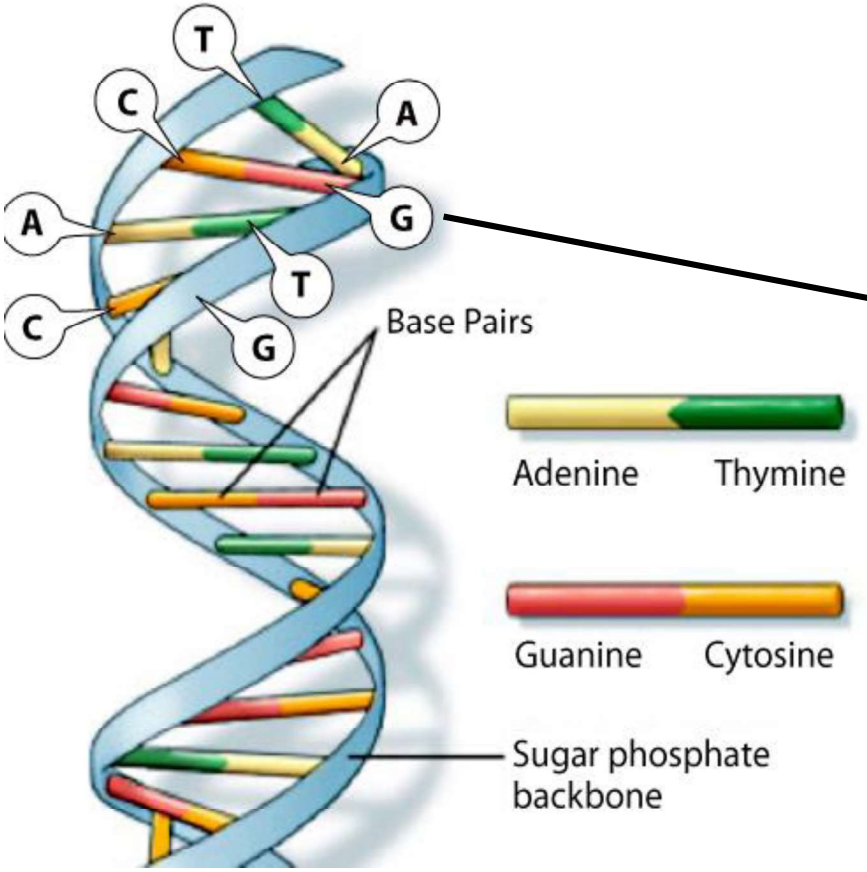
DNA-DNA interactions: biological viewpoint

- 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)

Common genetic variations



Single Nucleotide Polymorphisms (SNPs)	Frequency in general population
G	95%
A	5% > 1%

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

(Heather Cordell –CSCDA2016)

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)

(Heather Cordell –CSCDA2016)

(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 interaction 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

(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	
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

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

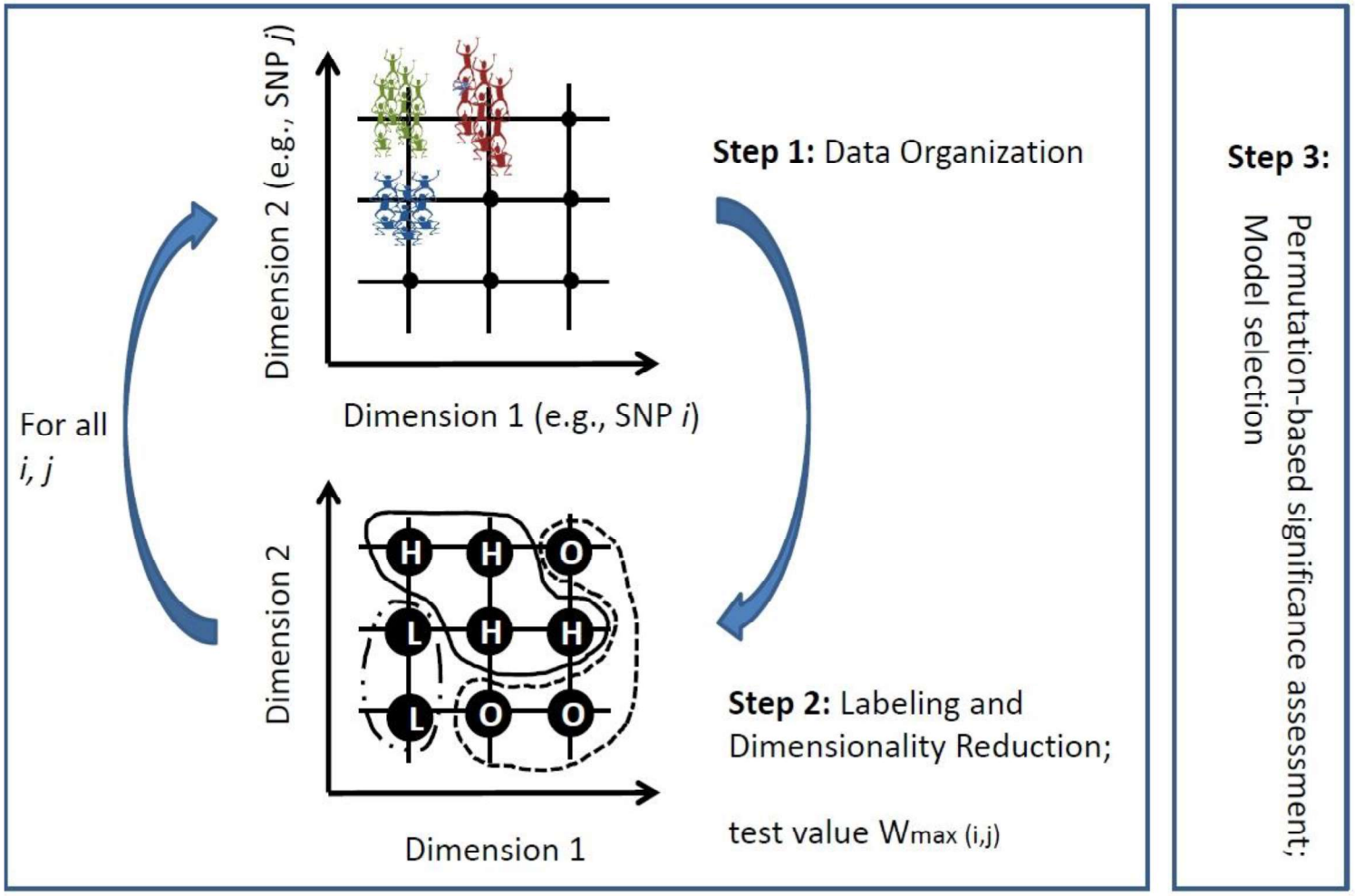
and **dosage encoding for X1 and X2**.

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)

Model-Based Multifactor Dimensionality Reduction

Model-Based Multifactor Dimensionality Reduction (MB-MDR)



MB-MDR and MDR are conceptually different (BIO3 team – 2010+)

- 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%

Performance

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

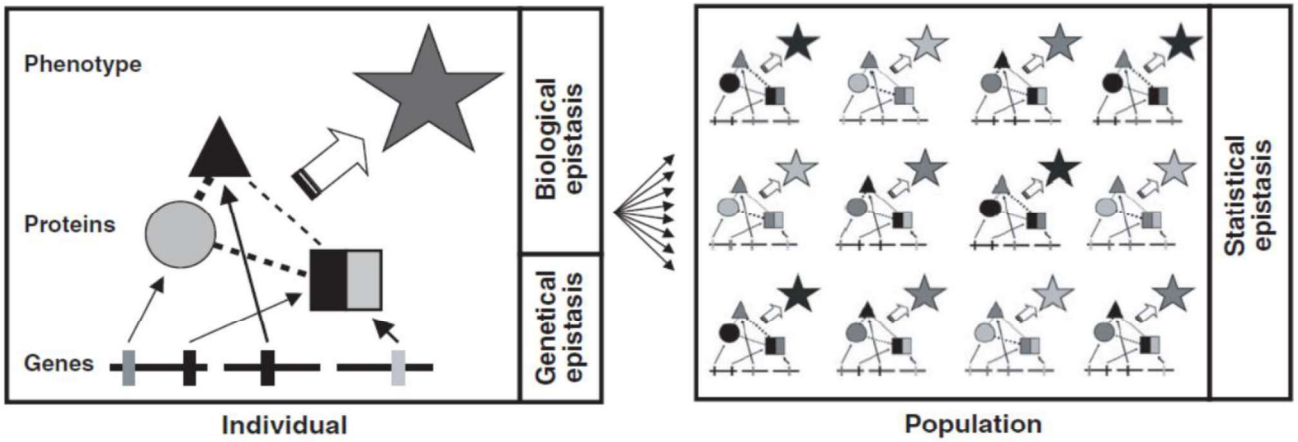
Model 6, $p = 0.1$

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

Challenges

DNA-DNA interactions: BIOLOGICAL VS STATISTICAL VIEWPOINT

- 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)

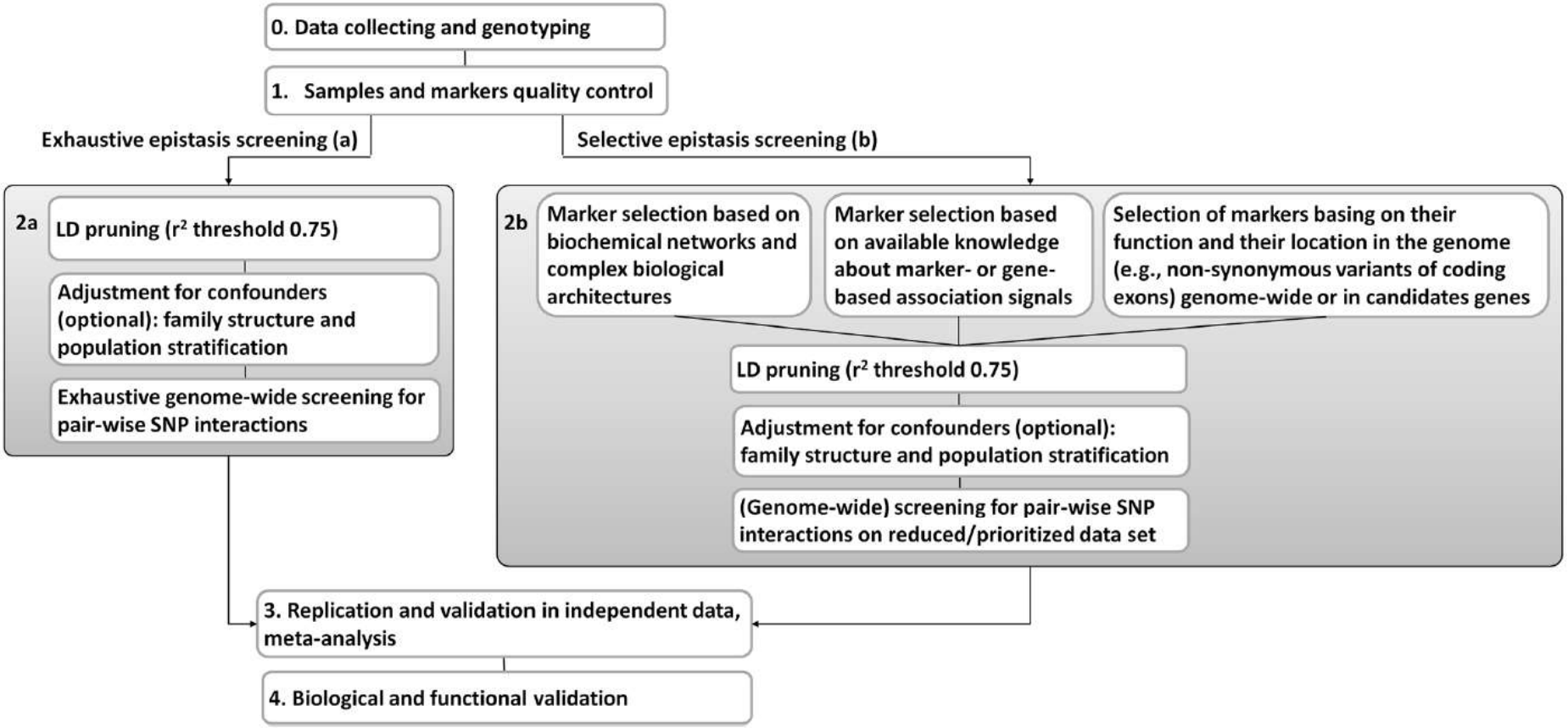
Lack of obvious correspondence

- From the literature:
  - 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**

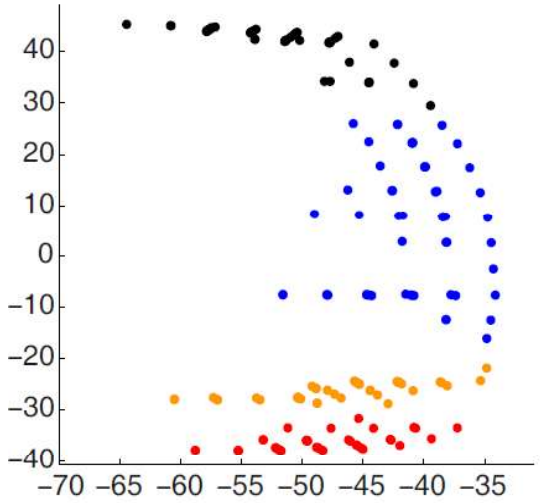
- Statistical interactions DO imply joint involvement

REMEDY: Towards a consensus GWAIs protocol



(Gusareva et al. 2014)

Confounding: SHARED GENETIC ANCESTRY

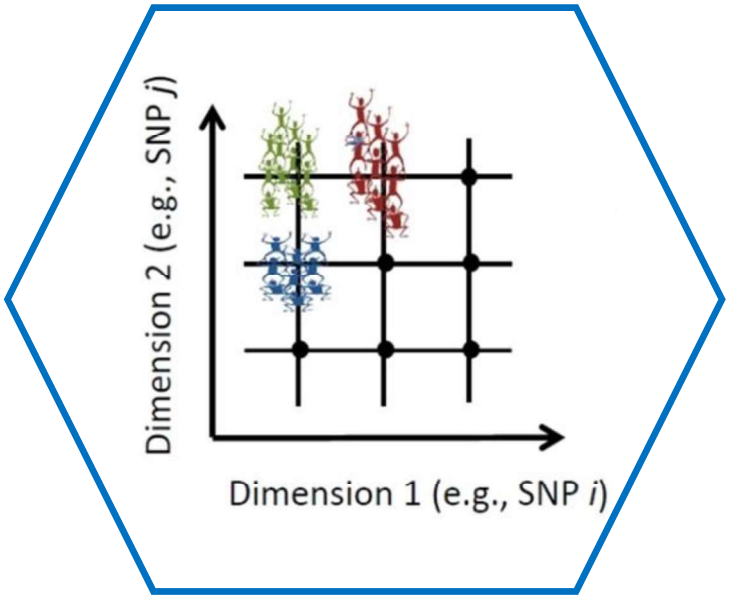


AA, BB	AA, Bb	AA, bb
Aa, BB	Aa, Bb	Aa, bb
aa, BB	aa, Bb	Aa, bb

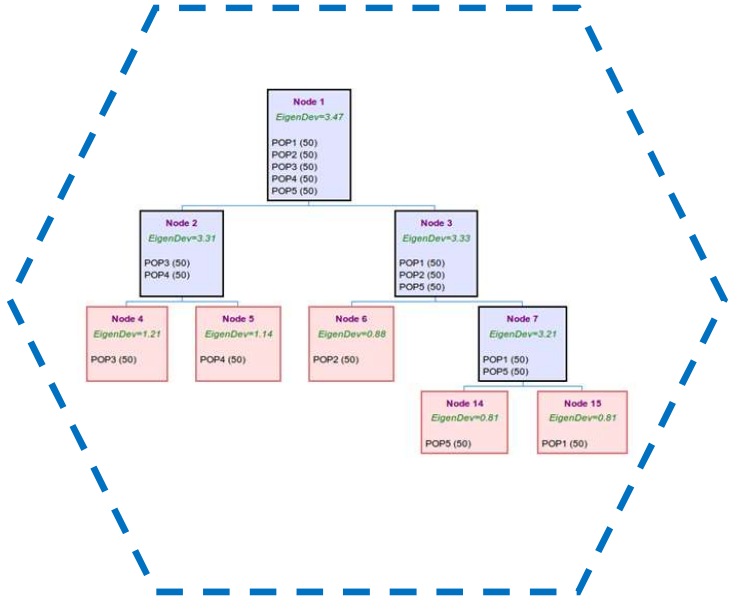
Multilocus Genotype Exposure

Disease Outcome

P-STRUCT



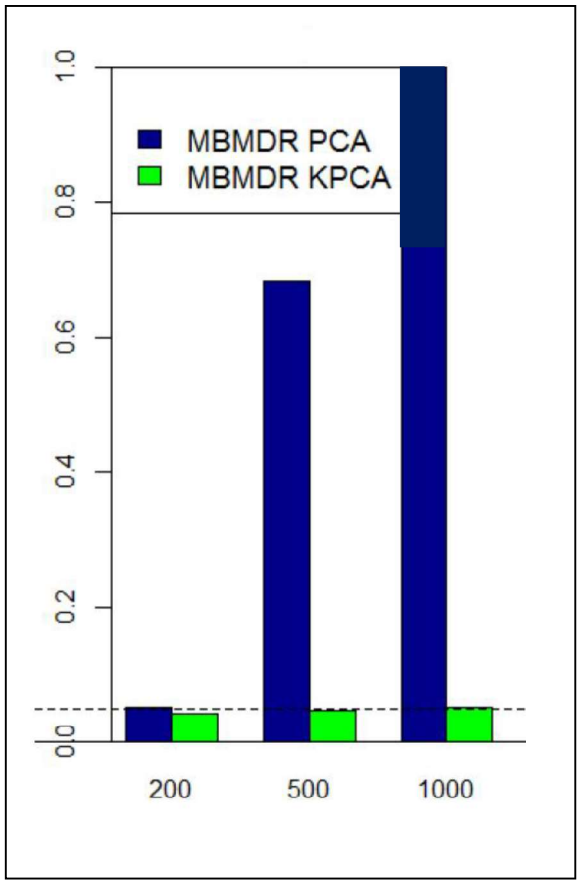
MB-MDR for structured populations

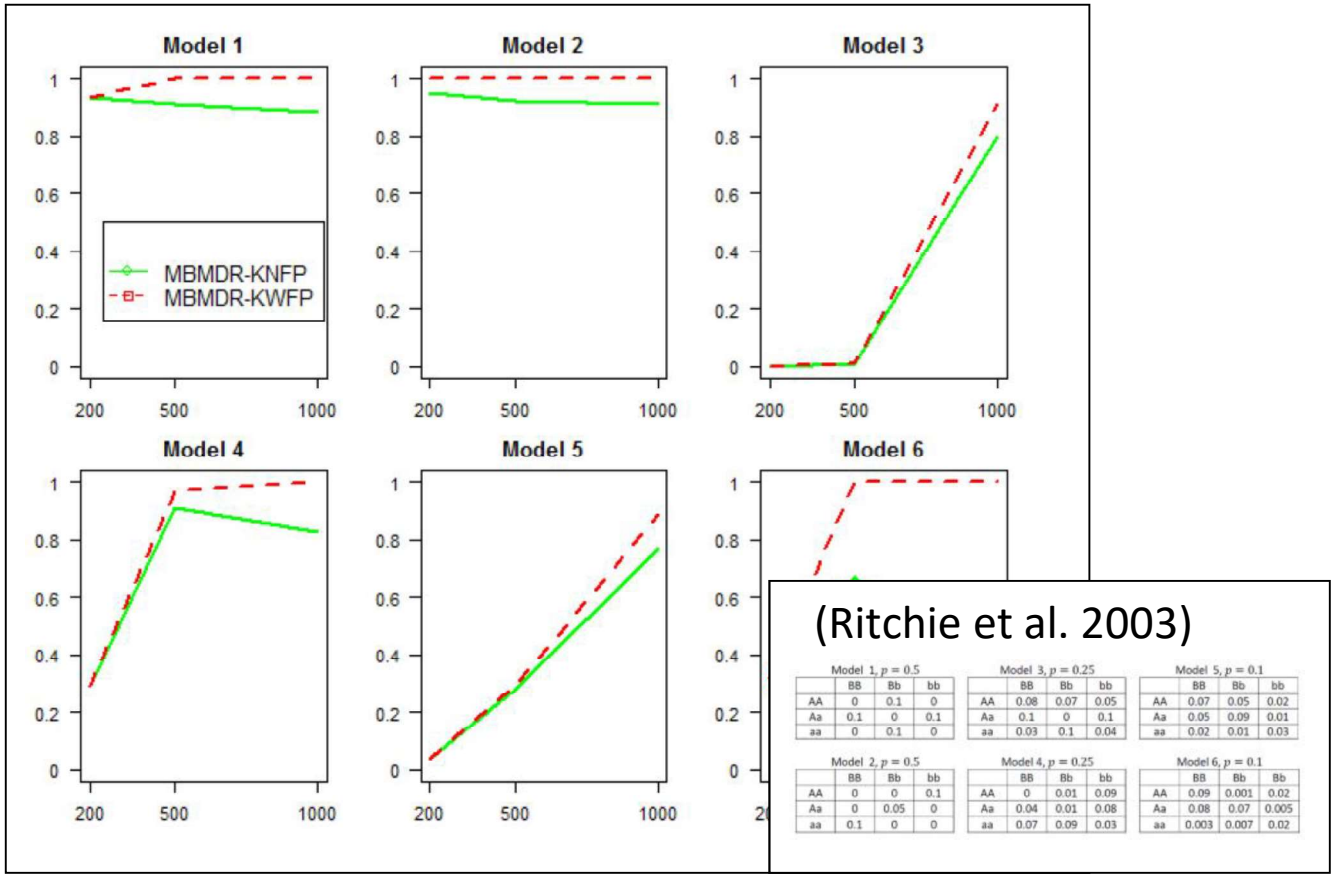


- Continuous axes of confounding
- IPCAPS
- Hypothesis-specific genomic control

(Chaichoompu et al. 2016+ ;
Abegaz et al., 2016+)

Remedy: Kernels (Fouladi et al. 2016+ ; Abegaz et al. 2016+)





(Ritchie et al. 2003)

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

Above : 60/40 CC ratio, structural epistasis according to corresponding full penetrance Rtichie epistasis model ; Below : 50/50 (200+200)

	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

(Cattaert et al. 2011)

Testing hypothesis: GLOBAL VERSUS INTERACTION SPECIFIC

- MDR (Ritchie et al. 2001+)

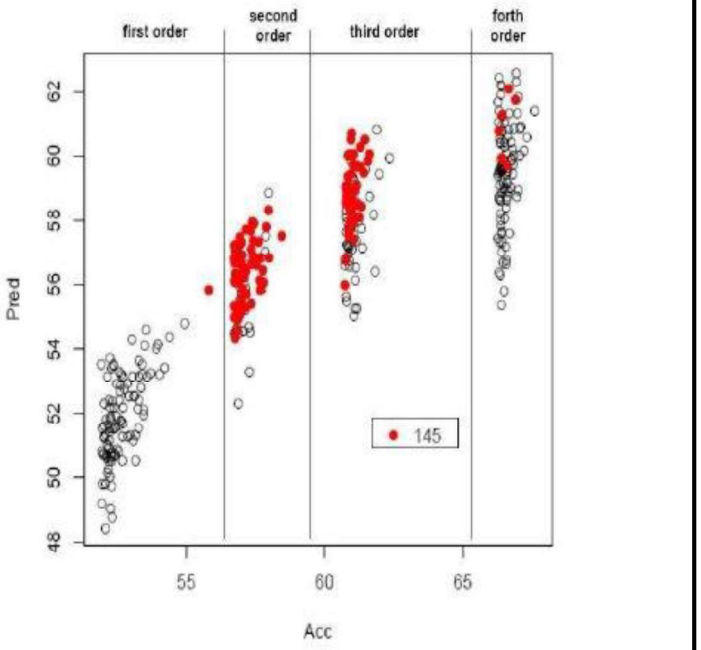
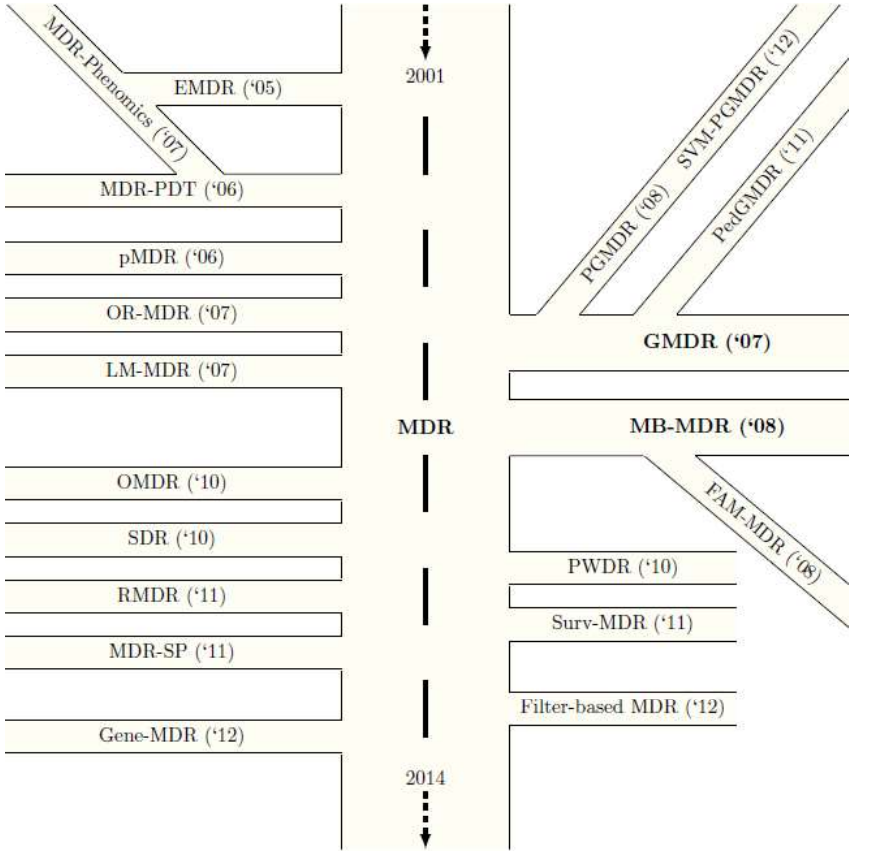
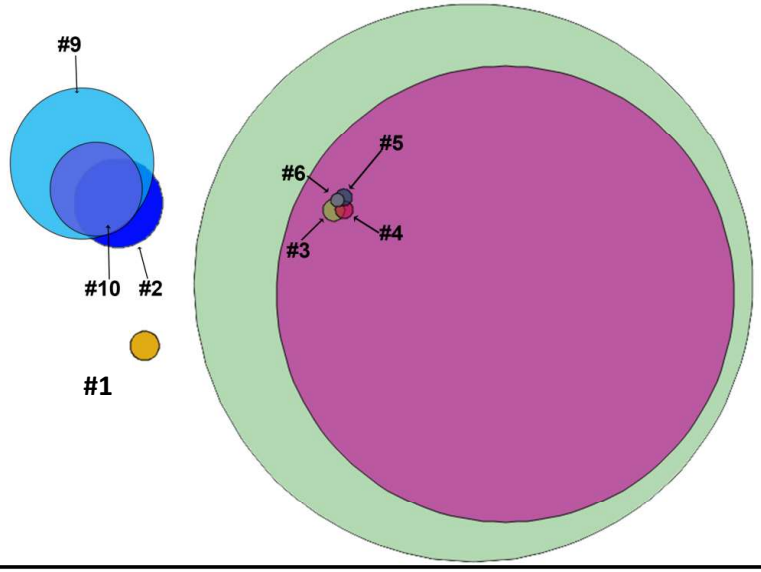


Fig. 1. Average Balanced Training accuracy (Acc) versus Average Balanced Predictive accuracy (Pred) for the 100 models with higher balanced training accuracy for the whole sample. First, second, third and forth order interactions are considered.

MDR alike tools (Gola et al. 2015)



REMEDY: Encoding lower order effects – "MB" in MB-MDR

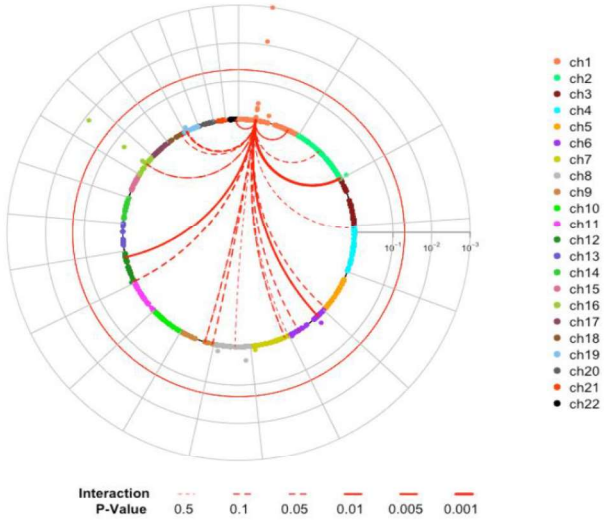


protocol #1 (226) protocol #2 (2,165) protocol #3 (129) protocol #4 (84)

protocol #5 (77) protocol #6 (47) protocol #7 (84,975) protocol #8 (57,032)

protocol #9 (6,589) protocol #10 (2,340)

(extended from Bessonov et al. 2016)



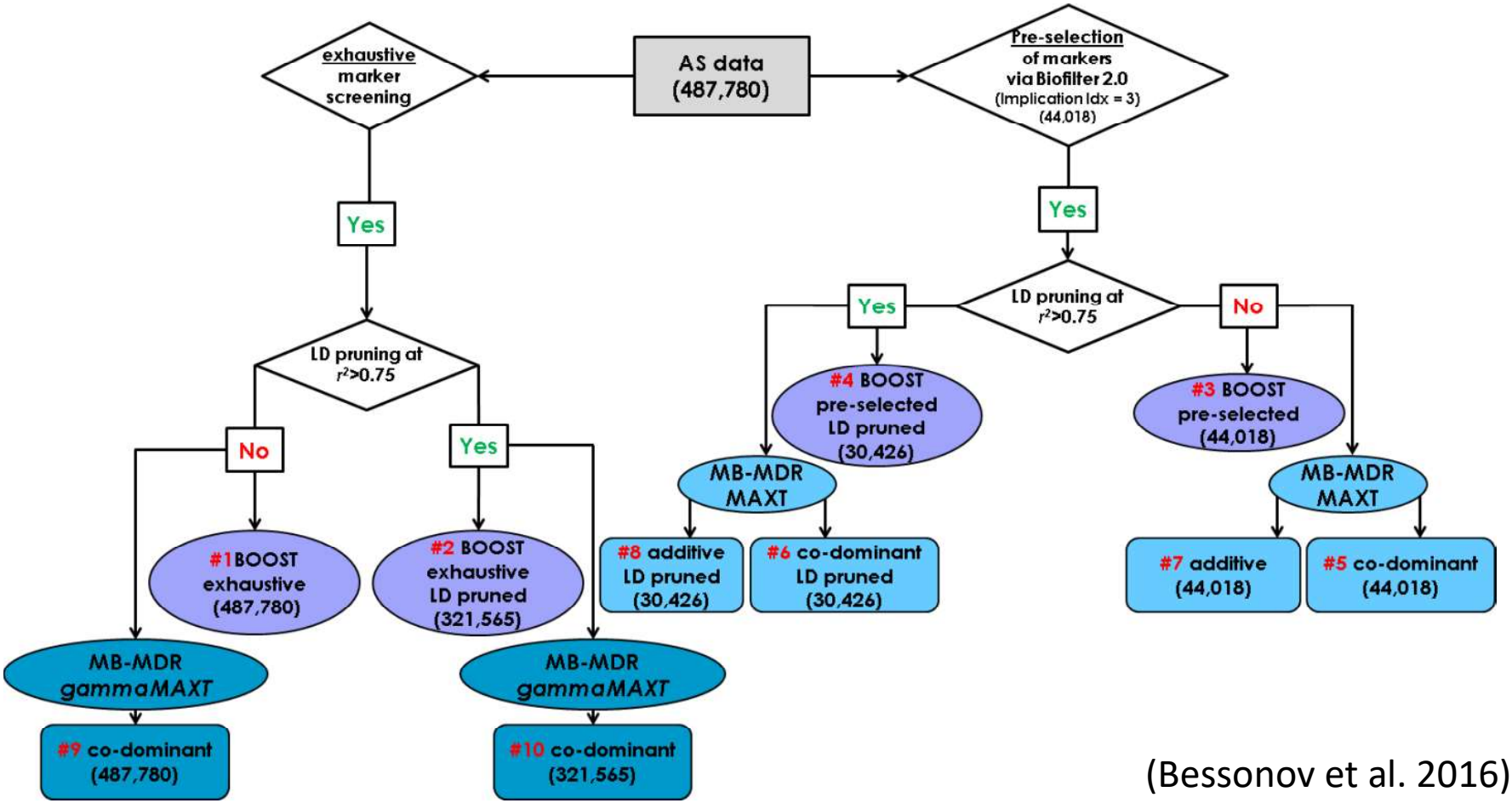
MOGPLOT:

Manhattan plot (main effects) over

GRAIL plot (interactions)

(Van Lishout et al. 2016)

Stability of results: REPLICATION



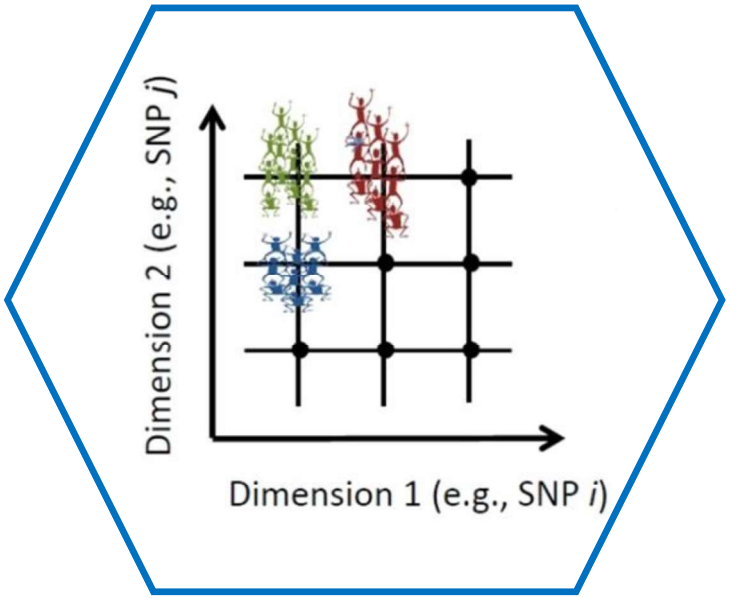
(Bessonov et al. 2016)

What should GWAI replication mean?

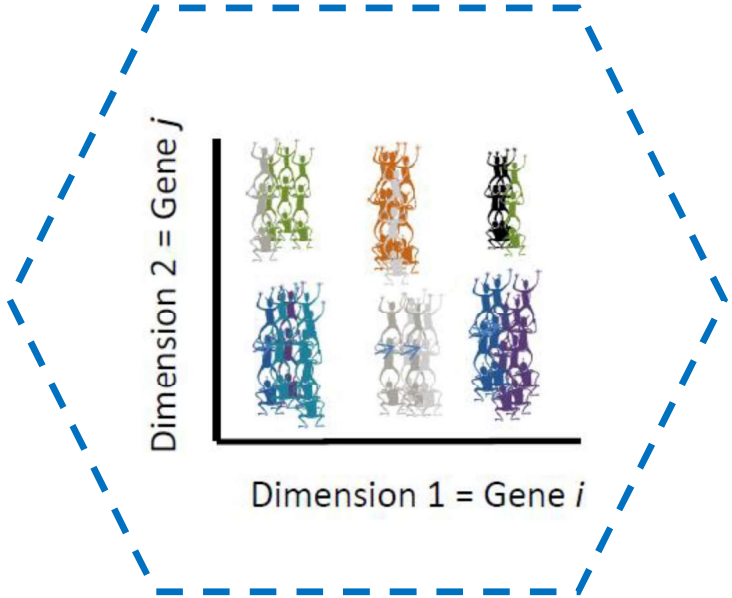
*Even for so-called replicated genetic interactions it is unclear to what extent **a false positive has been replicated** or to what extent main effects are responsible for the epistasis signal.*

(Ritchie and Van Steen, 2017 – under review)

REMEDY: Gene-based omics integration - INTEGROMIX



MB-MDR in integrative context



- Component-based
- Kernel-based
- Network-based

(Fouladi et al. 2015 ; 2016+)

Practical feasibility: SPEED

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	$\approx$ 1.5 year	$\approx$ 3 years	4 days 9 hours	$\approx$ 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 "$\approx$" are extrapolated.

(Van Lishout et al. 2013)

REMEDY: approximation

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)

Computational challenge … not for long!

- 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)

Take-Home Messages


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]**
- **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]**

The only source of knowledge is experience – A. Einstein

Published in final edited form as:

Neurobiol Aging. 2014 November ; 35(11): 2436–2443. doi:10.1016/j.neurobiolaging.2014.05.014.

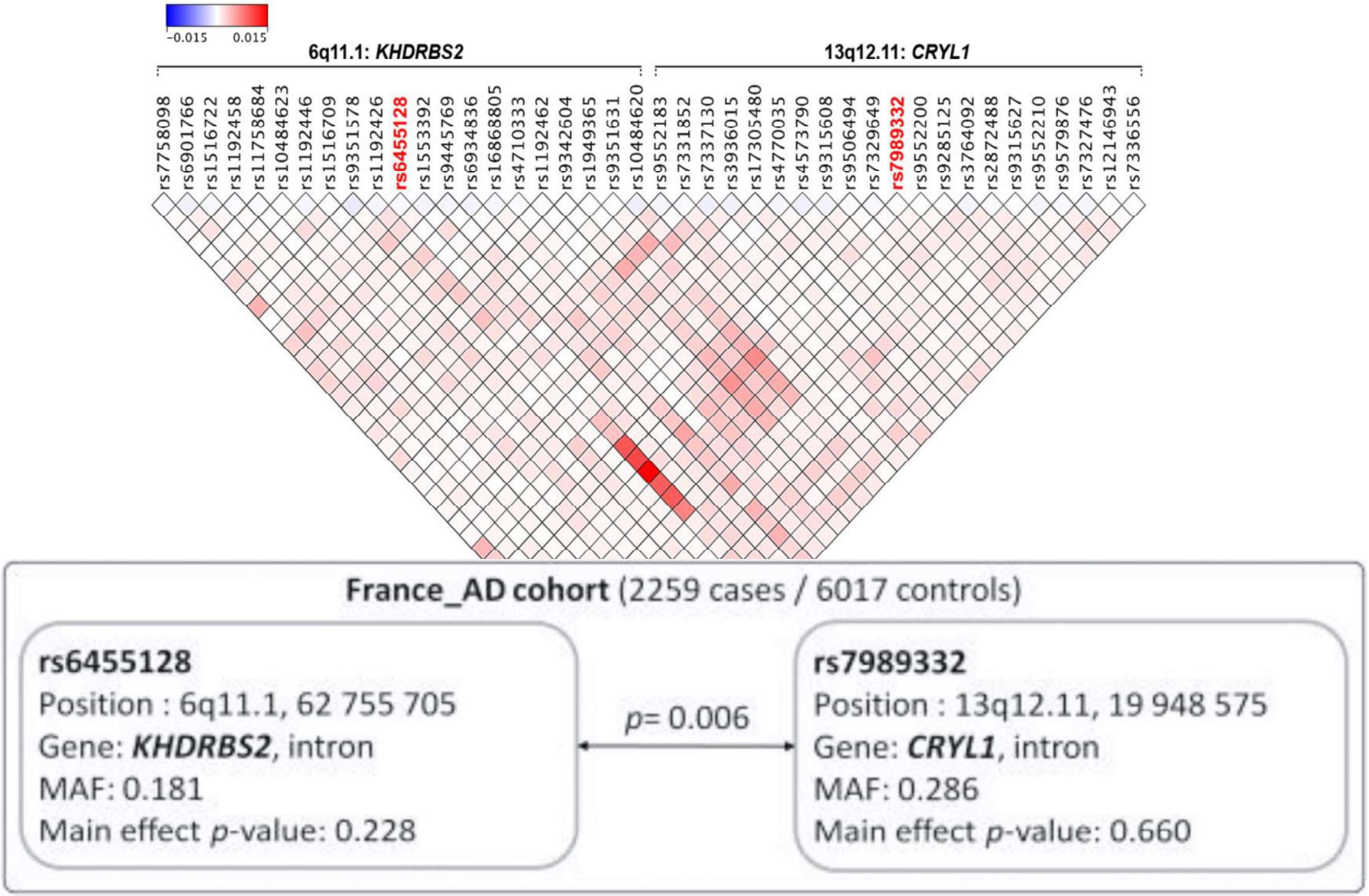
Genome-wide association interaction analysis for Alzheimer's disease

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]

[1]Systems and Modeling Unit, Montefiore Institute, University of Liege, Belgium

[2]Bioinformatics and Modeling, GIGA-R, University of Liege, Belgium

The only source of knowledge is experience – A. Einstein



Sex-specific interactions for AD

ANALYTICAL BLOCK

Discovery cohort EADI1: 3150 males (788 cases), 5110 females (1455 cases)
582,982 SNPs

1. Samples and markers quality control: HWE test in controls ($P > 1*10^{-4}$), call rate > 98%, marker allele frequency (MAF > 0.05)

474,020 SNPs

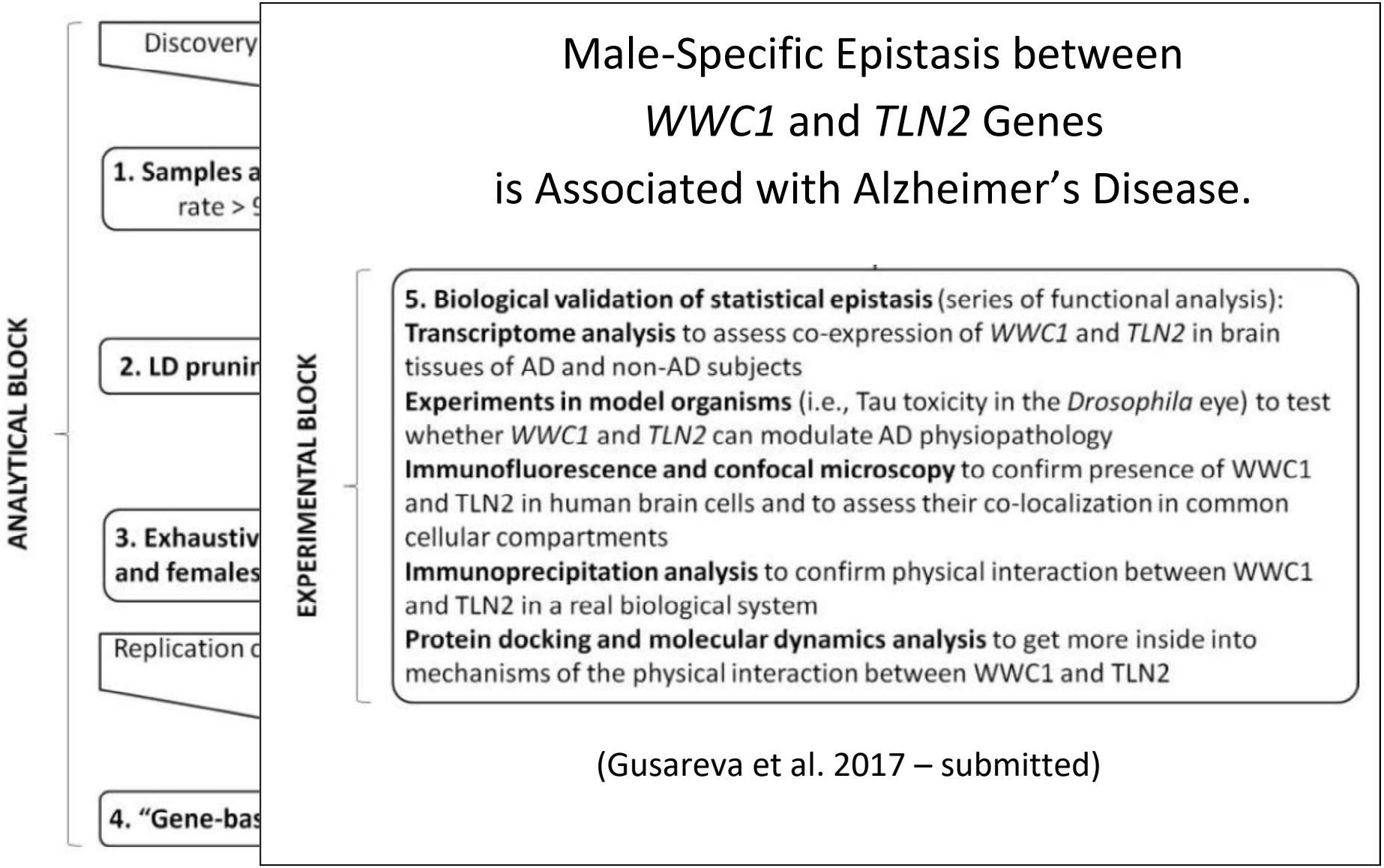
2. LD pruning: Window size 50 bp, window increment 1 bp, LD r^2 threshold 0.75

312,064 SNPs

3. Exhaustive genome-wide screening for pair-wise SNP interactions in males and females: BOOST, MB-MDR and regression modeling in R

Replication cohorts: GERAD1 3929 males (988 cases), RS 2376 males (264 cases), ADGC 6149 males (2584 cases)
68 SNPs - *WWC1*, 98 SNPs - *TLN2*

4. "Gene-based" replication analysis and meta-analysis

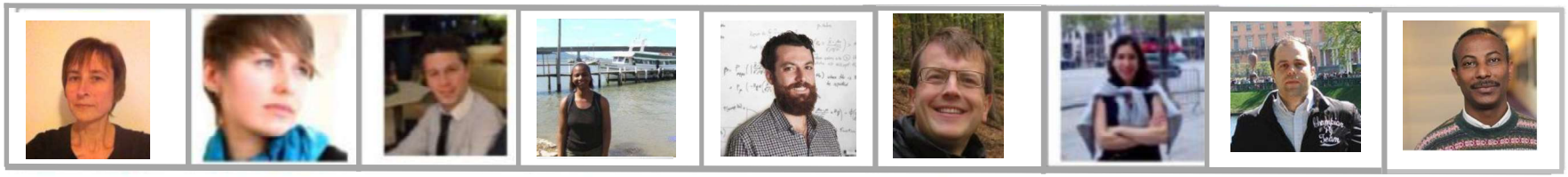


Male-Specific Epistasis between *WWC1* and *TLN2* Genes is Associated with Alzheimer's Disease.

(Gusareva et al. 2017 – submitted)

Acknowledgments

BIO3: Biostatistics, Biomedicine, Bioinformatics *(interactions)*



Machine Learning for Personalized Medicine

LA LIBERTÉ DE CHERCHER

GIGA-R, Medical Genomics Thematic Research Unit, Liège, Belgium