Biological Interpretation of Data and Interactions

Kristel Van Steen, PhD[2]

kristel.vansteen@ulg.ac.be

Systems and Modeling Unit, Montefiore Institute, University of Liège, Grande Traverse 10, 4000 Liège, Belgium

Bioinformatics and Modeling, GIGA-R, University of Liège, Avenue de l'Hôpital 1, 4000 Liège, Belgium

Université de Liège

Outline

- Setting the pace
- What’s in a name?
- Why should we bother?
- How to detect interactions?
    - Are all methods equally useful?
    - Interactions: A curse or a blessing?
    - Gearing up to GWAI and GWEI studies
- A minimal GWAIs protocol
- Validation and replication: An impossible task?
- Through the looking-glass

Setting the pace

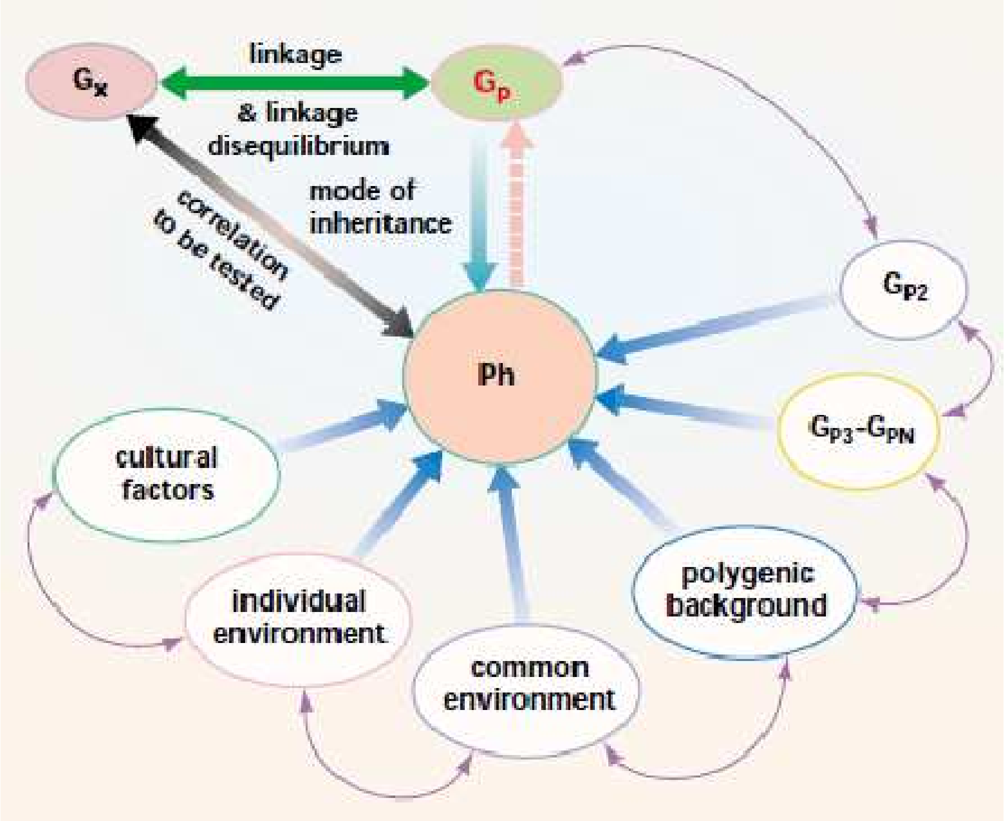
Université
de Liège

Genetic architecture of complex diseases

- Goal in statistical genetics / genetic epidemiology:
  - Unravel the biological mechanism underlying complex diseases
  - We hope to improve public health or to get closer to personalized medicine
- Achieving this goal is only possible with "appropriate tools" to capture the "genetic architecture" of the disease
- Genetic architecture:
  - The number of genes that impact disease susceptibility
  - The distribution of alleles and genotypes at those genes
  - The manner in which the alleles and genotypes impact disease susceptibility

(Weiss 1993)

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)

Université de Liège

What’s in a name?

Genetic associations

A genetic association refers to statistical relationships in a population between an individual's phenotype and their genotype at a genetic locus.

- Phenotypes:
  - Dichotomous
  - Measured
  - Time-to-onset
- Genotypes:
  - Known mutation in a gene (CKR5-deletion heterozygotes progress slower to AIDS, APOE ϵ4 allele predicts faster cognitive decline)
  - Marker or SNP with/without known effects on coding

Gene-gene interactions defined ?

- Wikipedia (23/04/2012)

In genetics, **epistasis** is the phenomenon where the effects of one gene are modified by one or several other genes, which are sometimes called **modifier genes**. The gene whose phenotype is expressed is called **epistatic** … Epistasis is often studied in relation to Quantitative Trait Loci (QTL) and polygenic inheritance…

… Epistasis and genetic interaction refer to different aspects of the same phenomenon …

… Studying genetic interactions can reveal gene function, the nature of the mutations, functional redundancy, and protein interactions. Because protein complexes are responsible for most biological functions, genetic interactions are a powerful tool …

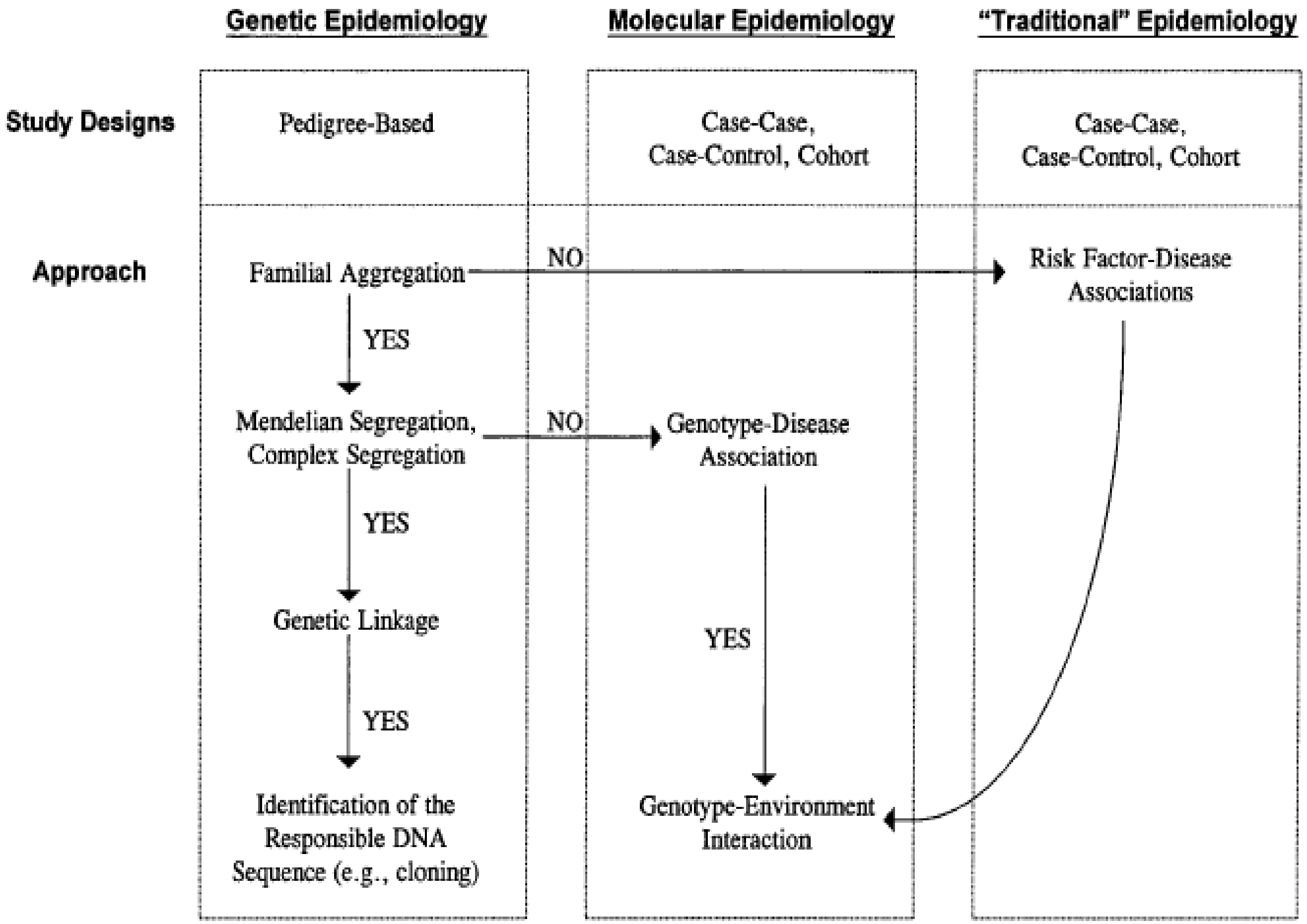
Gene-gene interactions defined ?

(Photo: J. Murken via A Ziegler)

© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com
They banned soda pop in Los Angeles schools 'cuz they say it makes kids fat!
What's next... Nintendo?

(Via presentation C Amos)

X – epidemiology



Genetic Epidemiology

Molecular Epidemiology

"Traditional" Epidemiology

Study Designs

Pedigree-Based

Case-Case,
Case-Control, Cohort

Case-Case,
Case-Control, Cohort

Approach

Familial Aggregation

NO

Risk Factor-Disease
Associations

YES

Mendelian Segregation,
Complex Segregation

NO

Genotype-Disease
Association

YES

Genetic Linkage

YES

YES

Identification of the
Responsible DNA
Sequence (e.g., cloning)

Genotype-Environment
Interaction

(Rebbeck TR, *Cancer*, 1999)

Genetic epidemiology

- Aim of genetic epidemiology is to **detect the inheritance pattern** of a particular disease, to **localize** the gene and to **find** a marker associated with **disease susceptibility**

- Genetic epidemiology is highly dependent on the direct incorporation of family structure and biology.
    - The structure of families and chromosomes leads to major dependencies between the data and thus to customized models and tests.
    - In many studies only indirect evidence can be used, since the disease-related gene, or more precisely the functionally relevant DNA variant of a gene, is not directly observable.

Gene-gene interactions defined: “compositional epistasis”

- The original definition (**driven by biology**) refers to distortions of Mendelian segregation ratios due to one gene masking the effects of another; a variant or allele at one locus prevents the variant at another locus from manifesting its effect (William Bateson 1861-1926).
- Example of phenotypes (e.g. hair colour) from different genotypes at 2 loci interacting epistatically under Bateson's (1909) definition:

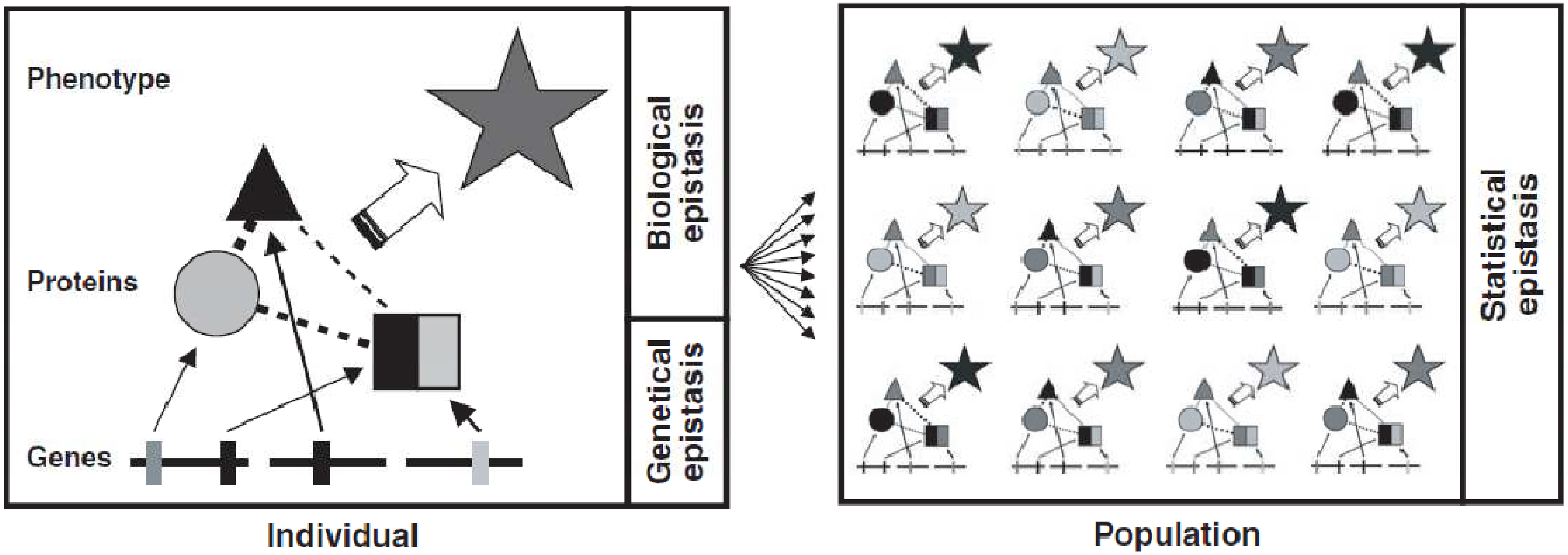
Genotype at locus B/G	**gg**	**gG**	**GG**
bb	White	Grey	Grey
bB	Black	Grey	Grey
BB	Black	Grey	Grey

The effect at locus B is masked by that of locus G: locus G is epistatic to locus B.

(Cordell 2002)

Gene-gene interactions defined: “statistical epistasis”

- A later definition of 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).



(Moore 2005)

A slightly more complicated two-locus model

- Example of penetrance table for two loci interacting epistatically in a general sense (fully penetrant: either 0 or 1)

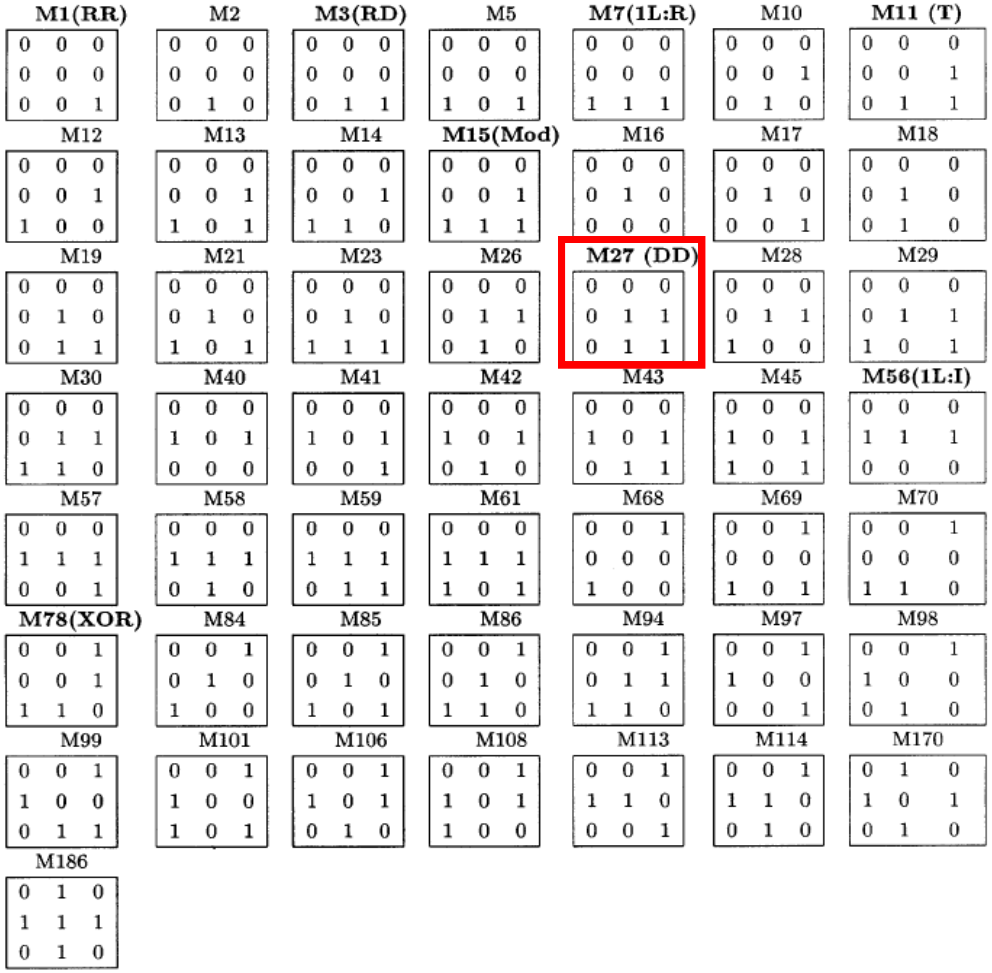
Genotype	bb	bB	BB
aa	0	0	0
aA	0	1	1
AA	0	1	1

(Cordell 2002)

- Enumeration of two-locus models:
  - Although there are 2^9=512 possible models, because of symmetries in the data, only 50 of these are unique.

Enumeration of two-locus models

(Li and Reich 2000)



M1(RR)	M2	**M3(RD)**	M5	**M7(1L:R)**	M10	**M11 (T)**
0 0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 1 0	0 0 0 0 0 0 0 1 1	0 0 0 0 0 0 1 0 1	0 0 0 0 0 0 1 1 1	0 0 0 0 0 1 0 1 0	0 0 0 0 0 1 0 1 1
M12	M13	M14	**M15(Mod)**	M16	M17	M18
0 0 0 0 0 1 1 0 0	0 0 0 0 0 1 1 0 1	0 0 0 0 0 1 1 1 0	0 0 0 0 0 1 1 1 1	0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 1	0 0 0 0 1 0 0 1 0
M19	M21	M23	M26	**M27 (DD)**	M28	M29
0 0 0 0 1 0 0 1 1	0 0 0 0 1 0 1 0 1	0 0 0 0 1 0 1 1 1	0 0 0 0 1 1 0 1 0	0 0 0 0 1 1 0 1 1	0 0 0 0 1 1 1 0 0	0 0 0 0 1 1 1 0 1
M30	M40	M41	M42	M43	M45	**M56(1L:I)**
0 0 0 0 1 1 1 1 0	0 0 0 1 0 1 0 0 0	0 0 0 1 0 1 0 0 1	0 0 0 1 0 1 0 1 0	0 0 0 1 0 1 0 1 1	0 0 0 1 0 1 1 0 1	0 0 0 1 1 1 0 0 0
M57	M58	M59	M61	M68	M69	M70
0 0 0 1 1 1 0 0 1	0 0 0 1 1 1 0 1 0	0 0 0 1 1 1 0 1 1	0 0 0 1 1 1 1 0 1	0 0 1 0 0 0 1 0 0	0 0 1 0 0 0 1 0 1	0 0 1 0 0 0 1 1 0
M78(XOR)	M84	M85	M86	M94	M97	M98
0 0 1 0 0 1 1 1 0	0 0 1 0 1 0 1 0 0	0 0 1 0 1 0 1 0 1	0 0 1 0 1 0 1 1 0	0 0 1 0 1 1 1 1 0	0 0 1 1 0 0 0 0 1	0 0 1 1 0 0 0 1 0
M99	M101	M106	M108	M113	M114	M170
0 0 1 1 0 0 0 1 1	0 0 1 1 0 0 1 0 1	0 0 1 1 0 1 0 1 0	0 0 1 1 0 1 1 0 0	0 0 1 1 1 0 0 0 1	0 0 1 1 1 0 0 1 0	0 1 0 1 0 1 0 1 0
M186						
0 1 0 1 1 1 0 1 0						

- Each model represents a group of equivalent models under permutations. The representative model is the one with the smallest model number.
- Two single-locus models (‘IL’) – the recessive (R) and the interference (I) model.

Note 1: Heterogeneity

- Example of penetrance table for two loci acting together in a heterogeneity model

Genotype	bb	bB	BB
aa	0	0	1
aA	0	0	1
AA	1	1	1

(Cordell 2002)

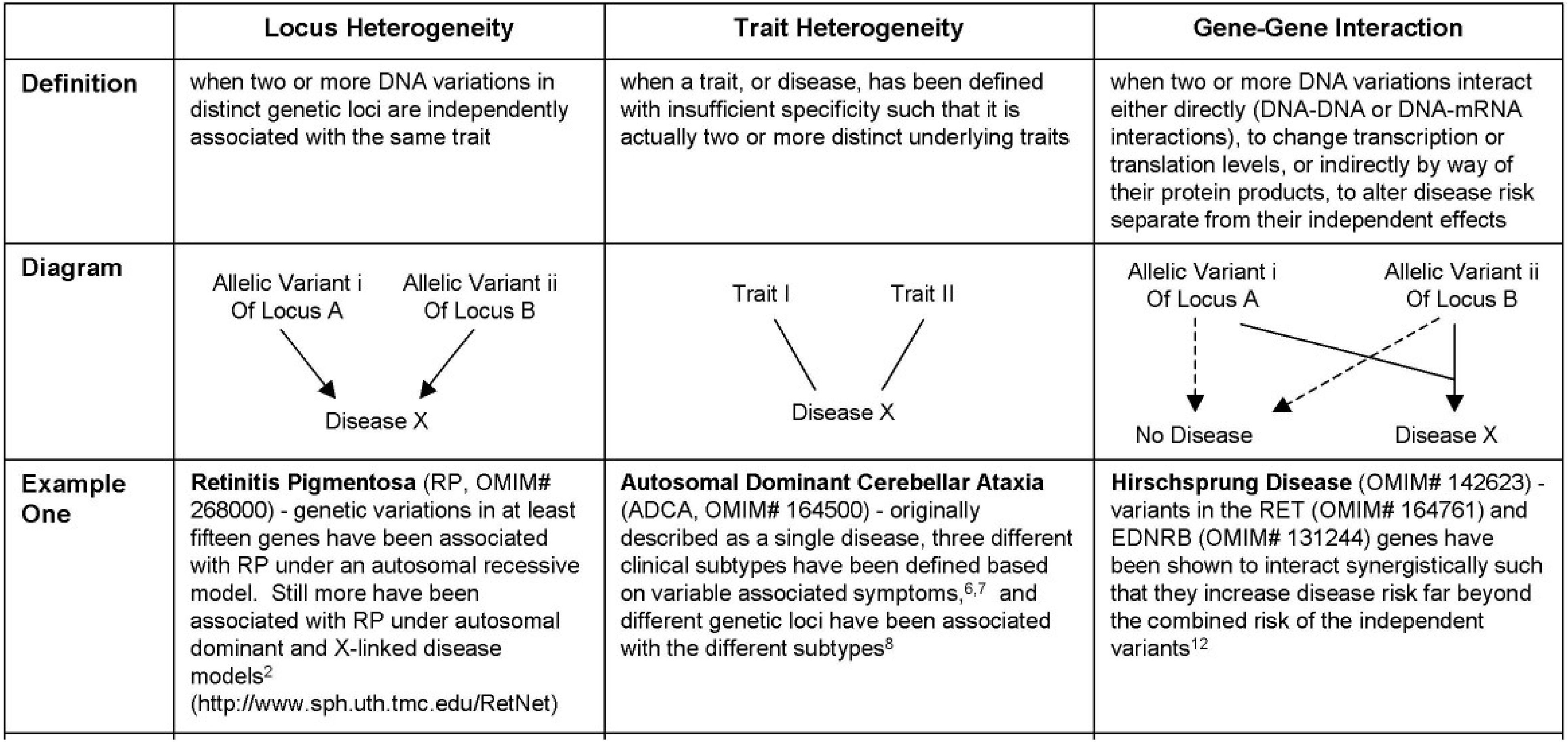
- Compare to model M27:

Genotype	bb	bB	BB
aa	0	0	0
aA	0	1	1
AA	0	1	1

(Li and Reich 2000)

Note 1: Heterogeneity

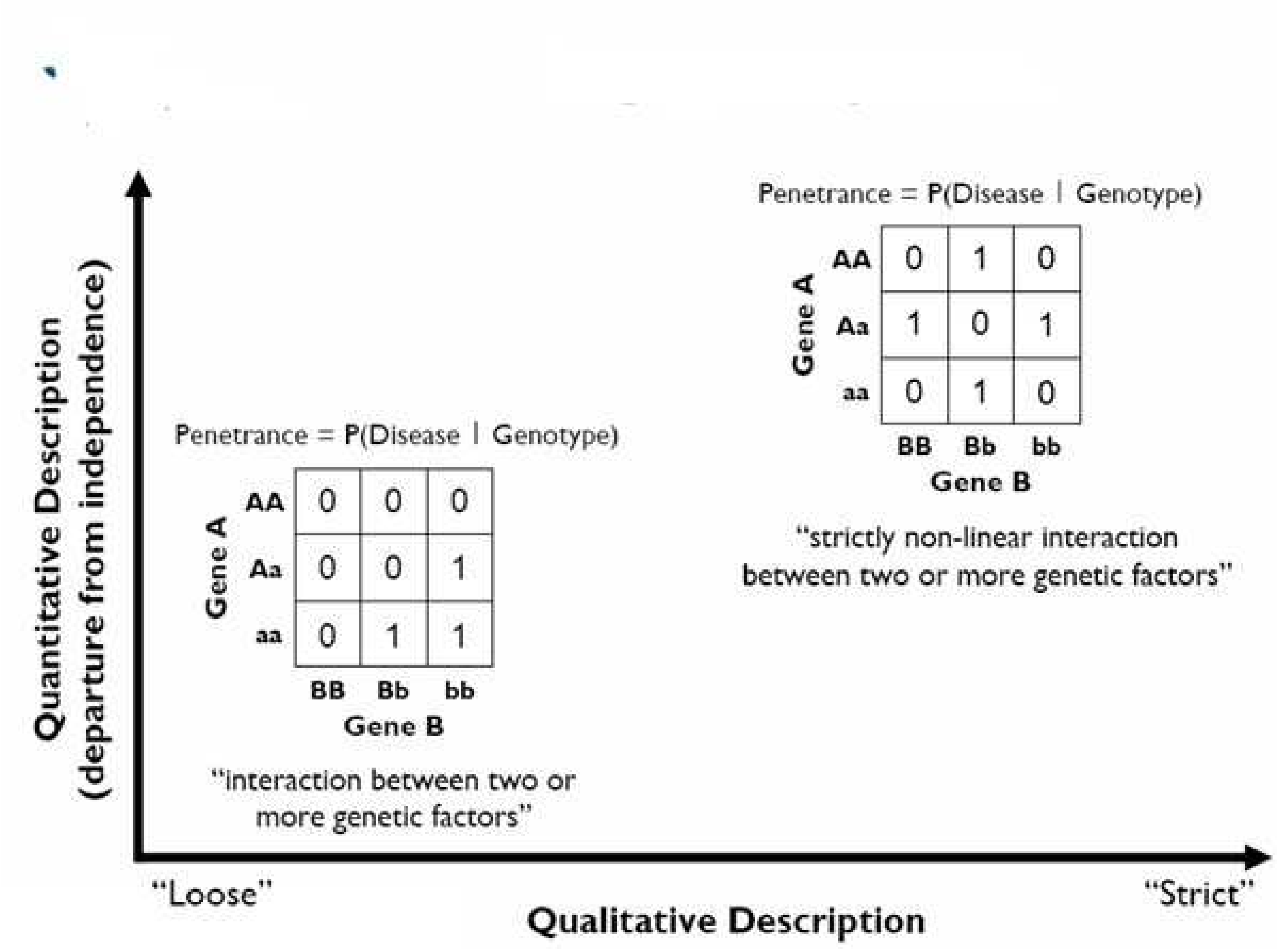
- Dissecting trait heterogeneity



	Locus Heterogeneity	Trait Heterogeneity	Gene-Gene Interaction
Definition	when two or more DNA variations in distinct genetic loci are independently associated with the same trait	when a trait, or disease, has been defined with insufficient specificity such that it is actually two or more distinct underlying traits	when two or more DNA variations interact either directly (DNA-DNA or DNA-mRNA interactions), to change transcription or translation levels, or indirectly by way of their protein products, to alter disease risk separate from their independent effects
Diagram	Allelic Variant i Of Locus A Allelic Variant ii Of Locus B Disease X	Trait I Trait II Disease X	Allelic Variant i Of Locus A Allelic Variant ii Of Locus B No Disease Disease X
Example One	**Retinitis Pigmentosa** (RP, OMIM# 268000) - genetic variations in at least fifteen genes have been associated with RP under an autosomal recessive model. Still more have been associated with RP under autosomal dominant and X-linked disease models[2] (http://www.sph.uth.tmc.edu/RetNet)	**Autosomal Dominant Cerebellar Ataxia** (ADCA, OMIM# 164500) - originally described as a single disease, three different clinical subtypes have been defined based on variable associated symptoms,[6,7] and different genetic loci have been associated with the different subtypes[8]	**Hirschsprung Disease** (OMIM# 142623) - variants in the RET (OMIM# 164761) and EDNRB (OMIM# 131244) genes have been shown to interact synergistically such that they increase disease risk far beyond the combined risk of the independent variants[12]

(Thornton-Wells et al. 2006)

Note 2: Different degrees of epistasis



Penetrance = P(Disease | Genotype)

Gene A \ Gene B	BB	Bb	bb
AA	0	0	0
Aa	0	0	1
aa	0	1	1

"interaction between two or more genetic factors"

Penetrance = P(Disease | Genotype)

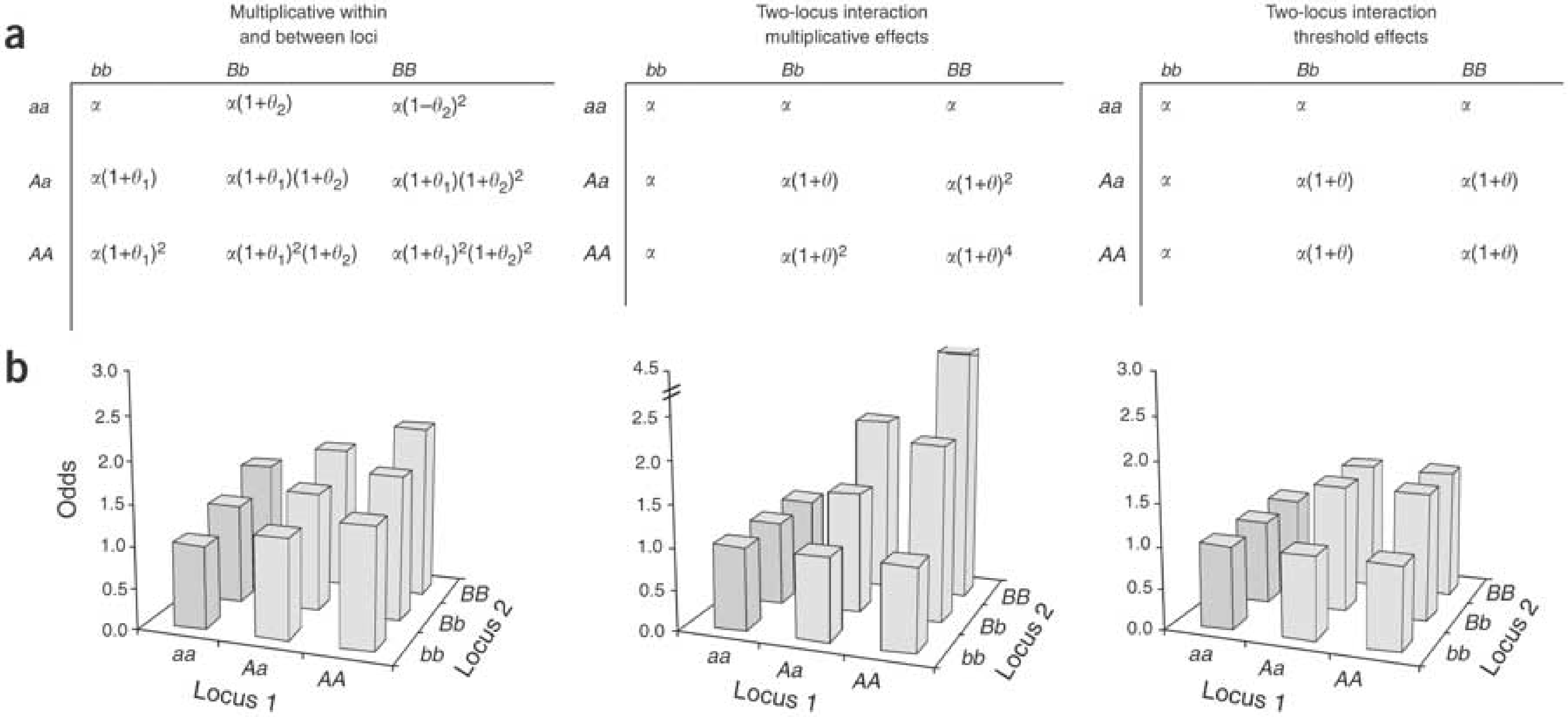
Gene A \ Gene B	BB	Bb	bb
AA	0	1	0
Aa	1	0	1
aa	0	1	0

"strictly non-linear interaction between two or more genetic factors"

(slide: Motsinger)

Note 3: Incomplete penetrances

- Odds of disease for 2 loci under epistatic scenarios



a

Multiplicative within and between loci

	bb	*Bb*	*BB*
aa	α	$\alpha(1+\theta_2)$	$\alpha(1-\theta_2)^2$
Aa	$\alpha(1+\theta_1)$	$\alpha(1+\theta_1)(1+\theta_2)$	$\alpha(1+\theta_1)(1+\theta_2)^2$
AA	$\alpha(1+\theta_1)^2$	$\alpha(1+\theta_1)^2(1+\theta_2)$	$\alpha(1+\theta_1)^2(1+\theta_2)^2$

Two-locus interaction multiplicative effects

	bb	*Bb*	*BB*
aa	α	α	α
Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
AA	α	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^4$

Two-locus interaction threshold effects

	bb	*Bb*	*BB*
aa	α	α	α
Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
AA	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$

(Marchini et al. 2005)

Why should we bother?

The true occurrences of epistasis

- 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 epistasis.
- Does 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).

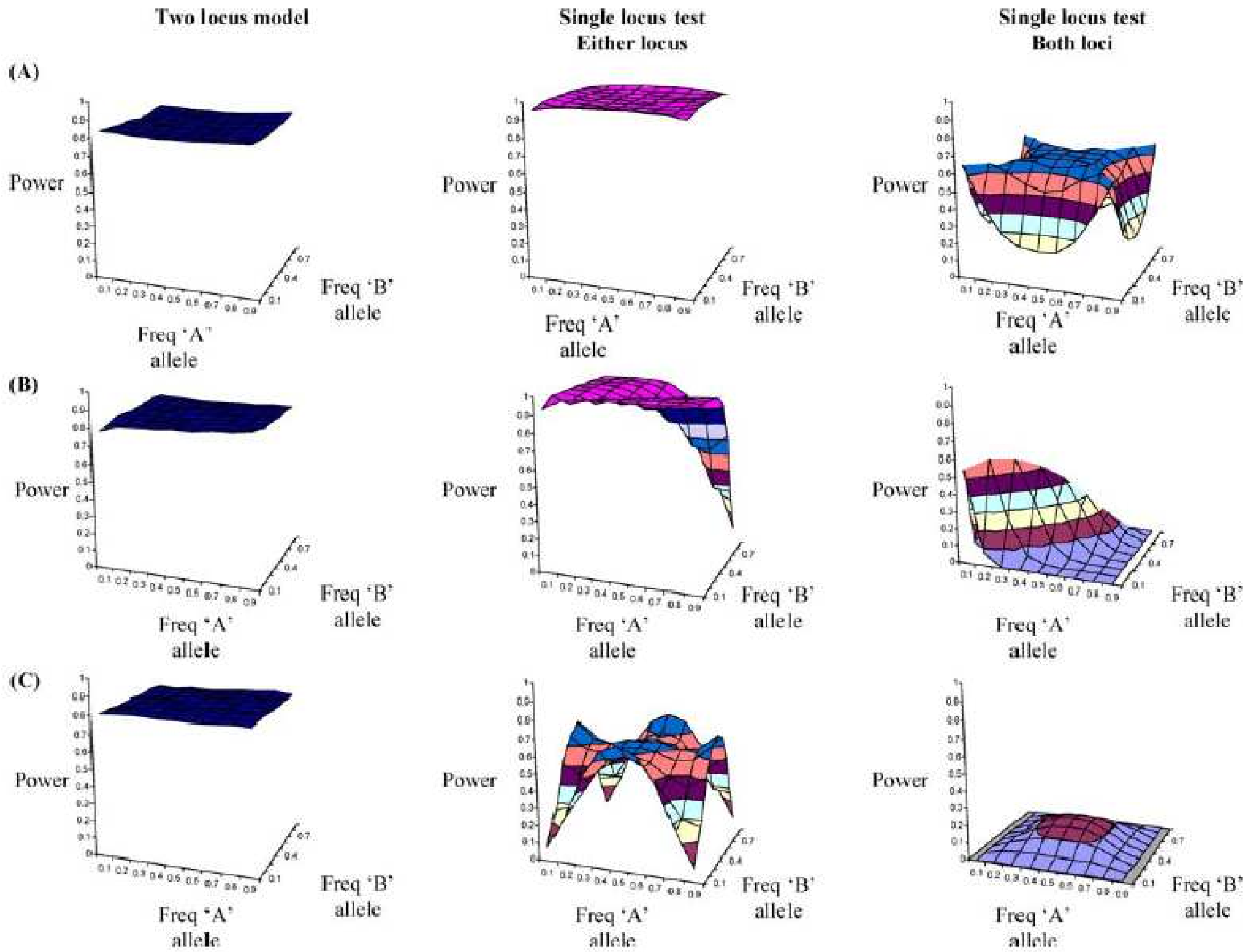
The observed occurrences of epistasis – model organisms

- Carlborg and Haley (2004):
  - Epistatic QTLs without individual effects have been found in various organisms, such as birds[26,27], mammals[28–32], Drosophila melanogaster[33] and plants[18,34].
  - However, 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].
  - This clearly 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 “observed” occurrences of epistasis – humans

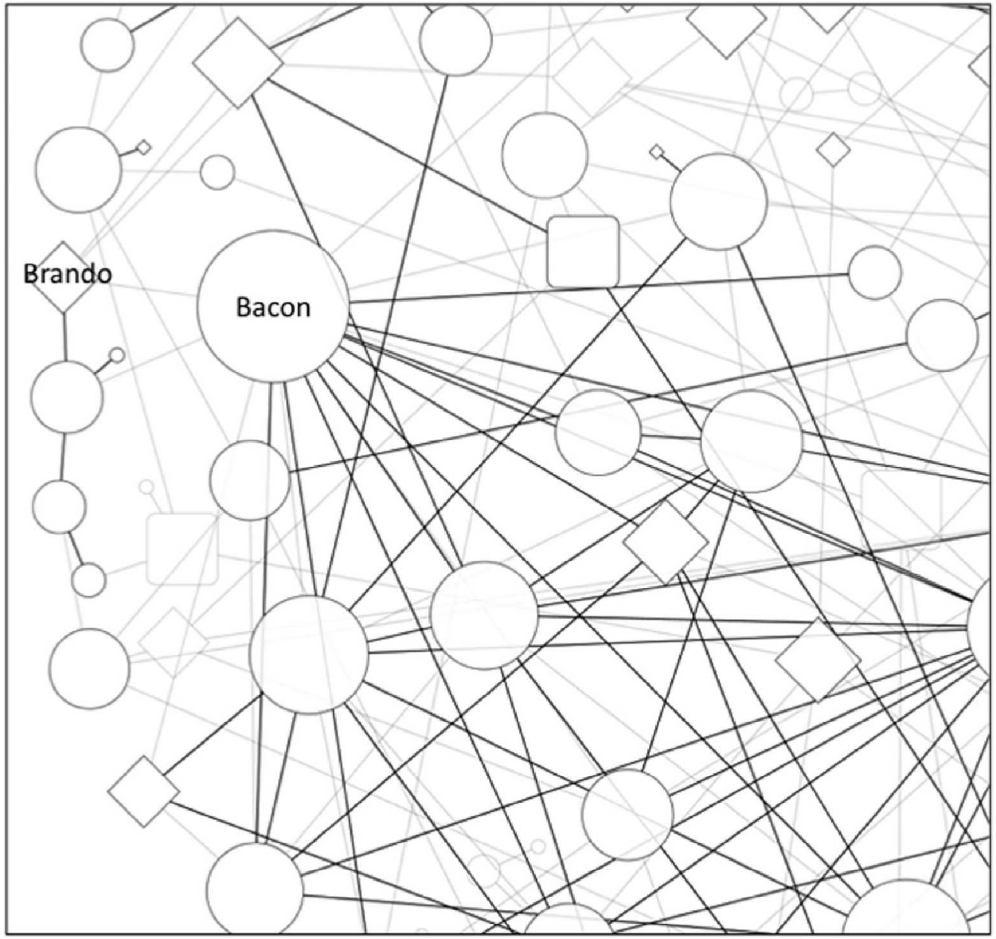
- Phillips et al (2008):
  - There are numerous cases of epistasis appearing as a statistical feature of association studies of human disease.
  - A few recent examples include coronary artery disease[63], diabetes[64], bipolar effective disorder[65], and autism[66].
  - So far, only for some of the reported findings additional support could be provided by functional analysis, as was the case for multiple sclerosis (Gregersen et al 2006).
- More recent examples, e.g., breast cancer (Ashworth et al. 2011)

Power to Detect Association for 1,500 Individuals where Both Loci Are Responsible for 5% of the Trait Variance



Epistasis network from a hypothetical GWAS

(McKinney et al 2012)



)

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**

Epistasis as a source of missing heritability?

(Maher 2008)

From GWAs to GWAIs

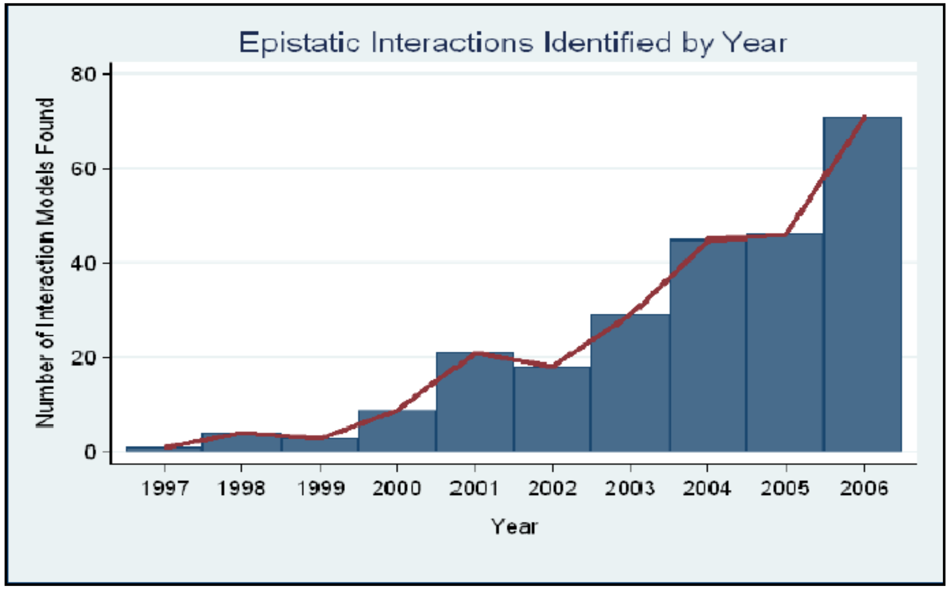
- Genome-Wide Association Interaction (GWAI) studies have not been as successful as GWA studies:
    - **Possible negligible role** of epistatic variance in a population? (Davierwala et al 2005)
    - Consequence of **not yet available** powerful epistasis detection **methods or approaches**?

      “ Gene-gene interactions are commonly found when properly investigated ” (Templeton 2000)

How to detect interactions?

A growing toolbox

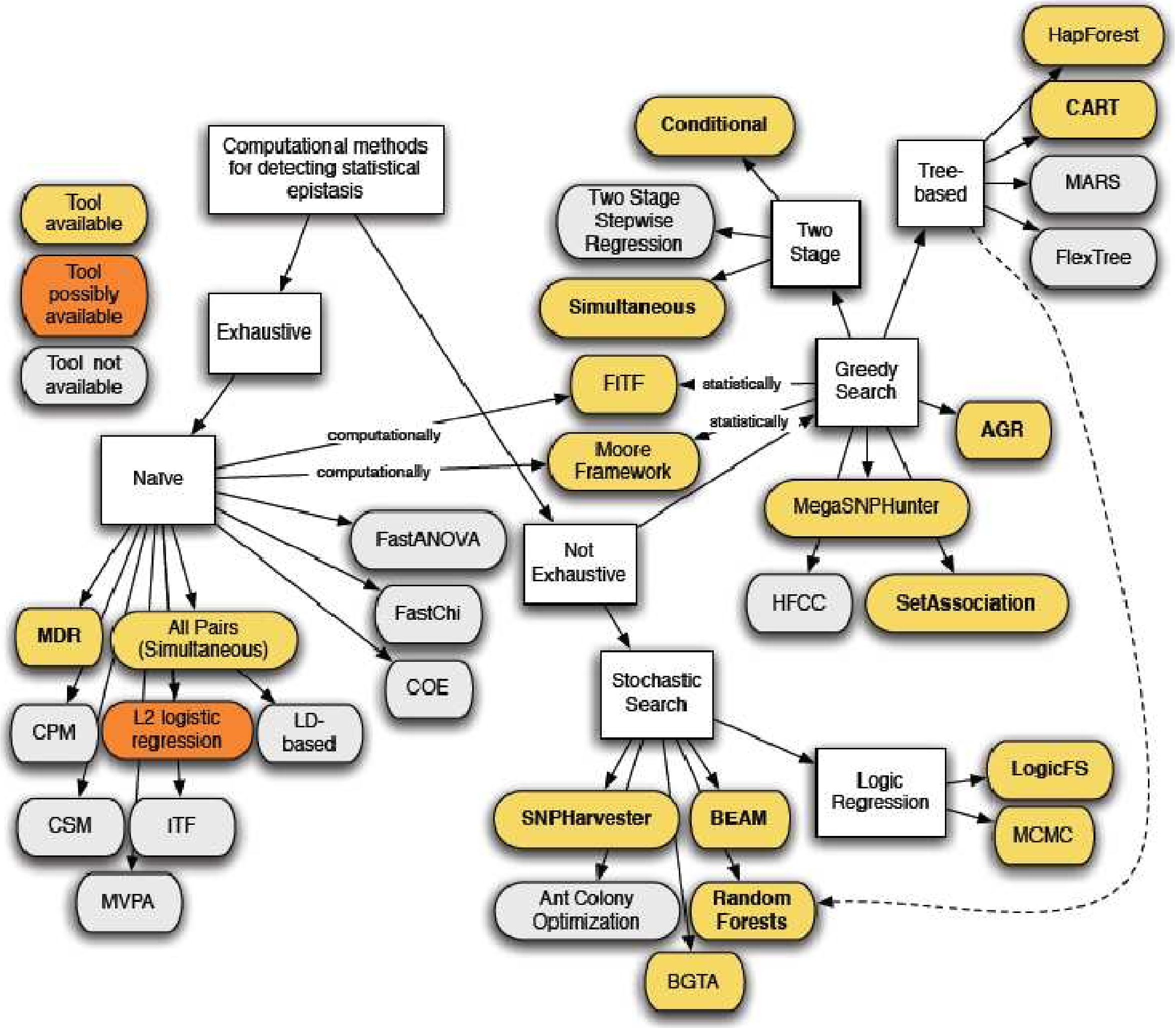
- The number of identified epistasis effects in humans, showing susceptibility to common complex human diseases, follows a steady growth curve (Emily et al 2009, Wu et al 2010), due to the growing number of toolbox methods and approaches.



(Motsinger et al. 2007)

Classification of epistasis detection methods

(Kilpatrick 2009)



Are all methods equally useful?

- Several criteria have been used to make such a classification:
  - the strategy is exploratory in nature or not,
  - modeling is the main aim, or rather testing,
  - the epistatic effect is tested indirectly or directly,
  - the approach is parametric or non-parametric,
  - the strategy uses exhaustive search algorithms or takes a reduced set of input-data, that may be derived from
    - prior expert knowledge or
    - some filtering approach

"These criteria show the diversity of methods and approaches and complicates making honest comparisons".

Epistasis : a curse or a blessing ?

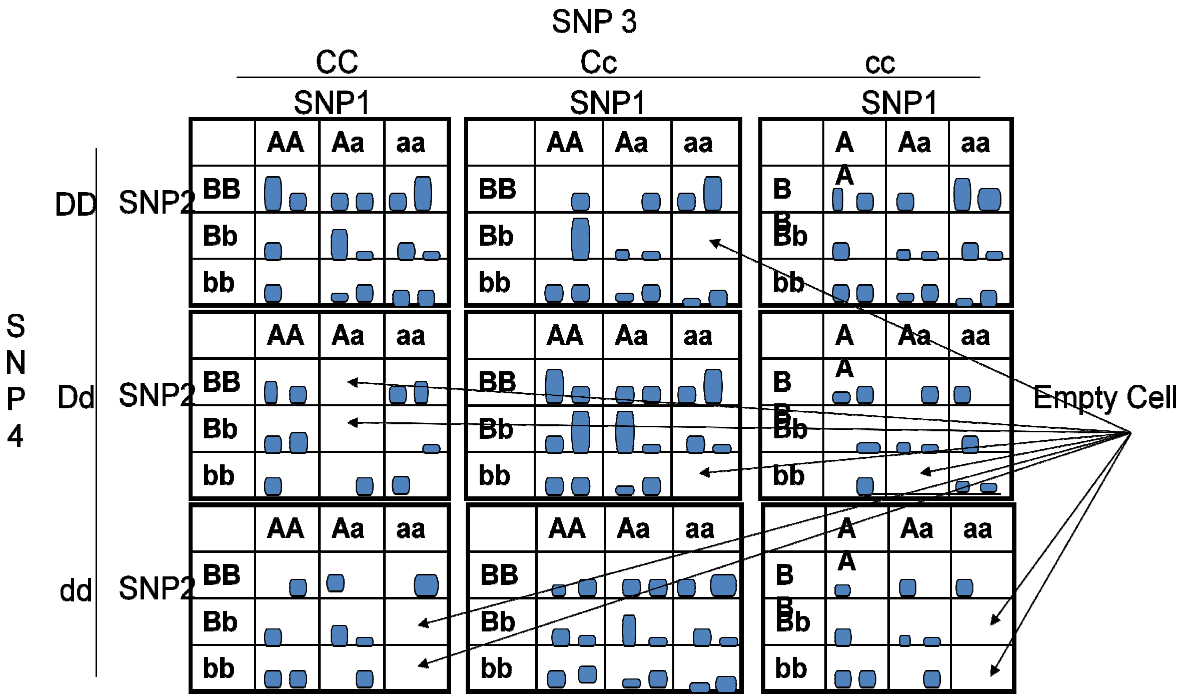
The curse of dimensionality

- The curse of dimensionality refers to the fact that the convergence of any parametric model estimator to the true value of a smooth function defined on a space of high dimension is very slow (Bellman and Kalaba 1959).
- This is already a problem for main effects GWAS, when trying to assess those SNPs that are jointly most predictive for the disease or trait of interest, but is compounded when epistasis screenings are envisaged

"Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders"

Missing data

- For 4 SNPs, there are 81 possible combinations with even more parameters to potentially model and more possible empty cells …

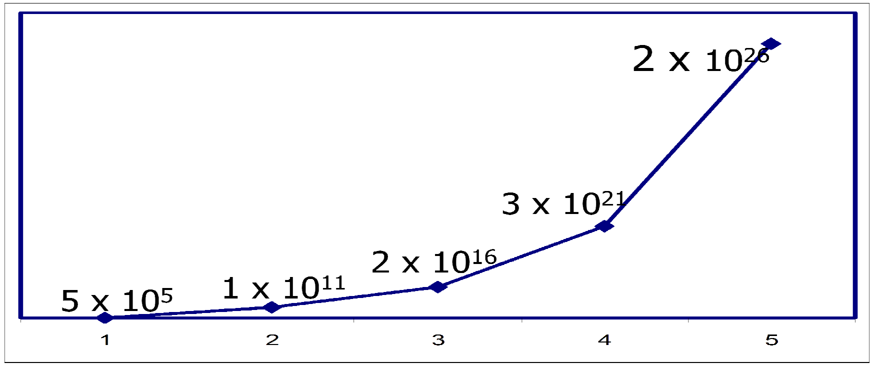


(slide: C Amos)

“A revision of LD based imputation strategies for GWAIs is needed”

The multiple testing problem

- The genome is large and includes many polymorphic variants and many possible disease models, requiring a large number of tests to be performed.
- This poses a “statistical” problem: a large number of genetic markers will be highlighted as significant signals or contributing factors, whereas in reality they are not (i.e. false positives).



~500,000 SNPs span 80% of common variation (HapMap)

“The interpretation of GWAIs is hampered by undetected false positives”

Université de Liège

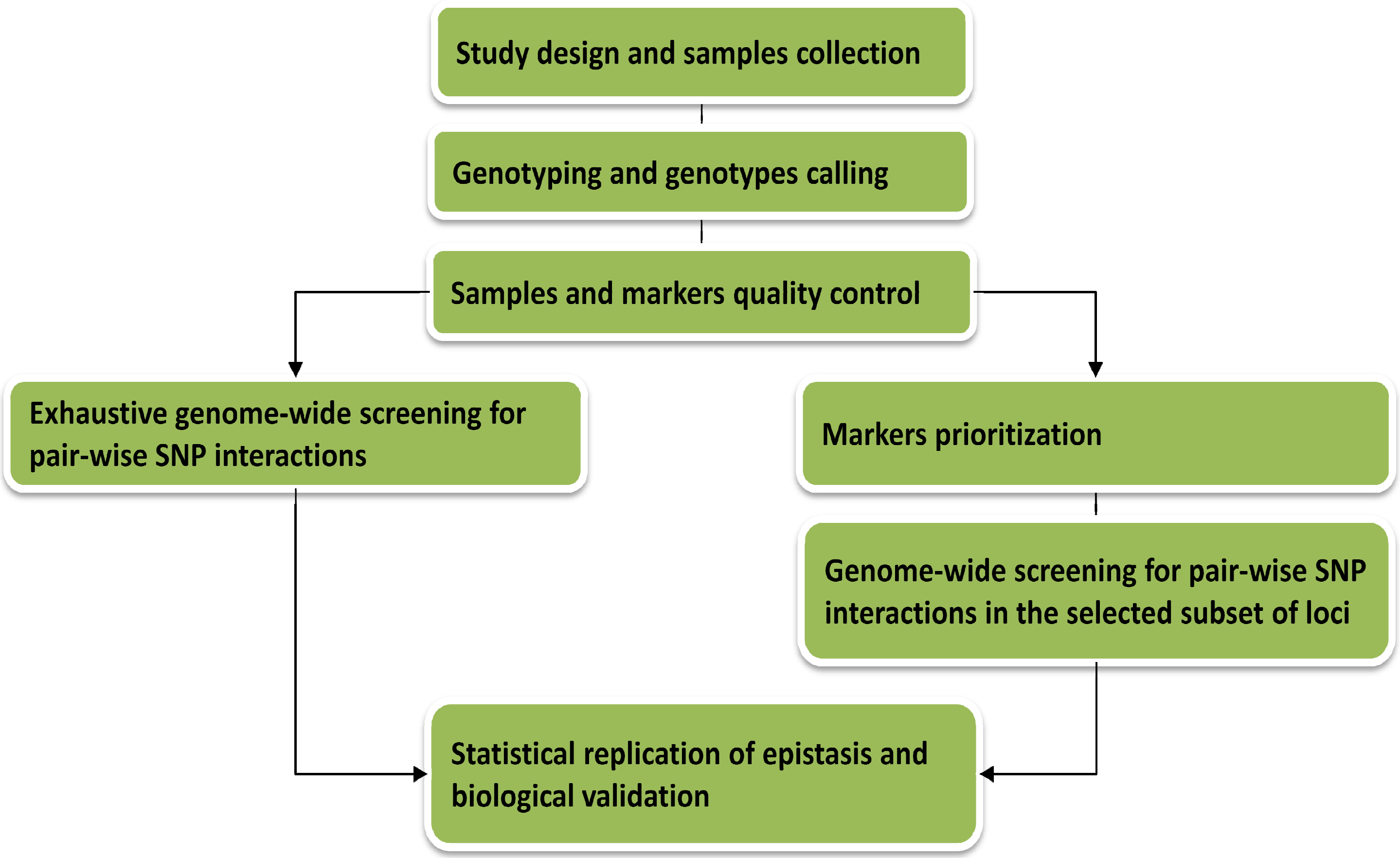
Data Integration: a solution?!

- The genome on its own has turned out to be a relatively poor source of explanation for the differences between cells or between people (Bains 2001)
- **Broad definition** (Van Steen):

  "Combining evidences from different data resources, as well as data fusion with biological domain knowledge, using a variety of statistical, bioinformatics and computational tools".

Data Integration: a solution?!

- Where in the GWAI process?



Study design and samples collection

Genotyping and genotypes calling

Samples and markers quality control

Exhaustive genome-wide screening for pair-wise SNP interactions

Markers prioritization

Genome-wide screening for pair-wise SNP interactions in the selected subset of loci

Statistical replication of epistasis and biological validation

(slide: E Gusareva)

Data Integration: a solution?!

Where?	**How?**	**Comments**
Data preparation / Quality control	Impute using different data resources	Filling in the gaps or inducing LD-driven interactions?
Variable selection	Use a priori knowledge about networks and genetical / biological interactions (e.g., Biofilter)	Feature selection (dimensionality reduction) or loosing information?
Modeling	"Integrative" analysis	Obtaining a multi-dimensional perspective or combining/merging data in a single analysis?
Interpretation (validation)	Use a posteriori knowledge (e.g., Gene Ontology Analysis, Biofilter – Bush et al. 2009)	Targeting known interactions or ruling out possibly relevant unknown interactions?

Gearing up to GWAIs and GWEIs

- Interactions are commonly assessed by regressing on the product between both ‘exposures’ (genes / environment)

$$E[Y|G_1, G_2, X) = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_X X + \beta G_1 G_2$$

with X a possibly high-dimensional collection of confounders.

- There are at least 2 concerns about this approach:
  - Model misspecification → we need a robust method
  - Capturing statistical versus mechanistic interaction → guard against high-dimensional (genetic or environmental) confounding)

(adapted from slide: S Vansteelandt)

Stijn

Mechanistic interactions

- Tests for **sufficient cause interactions** to identify mechanistic interactions aim to signal the presence of individuals for whom the outcome (e.g., disease) would occur if both exposures were "present", but not if only one of the two were present.

(Rothman 1976, VanderWeele and Robins 2007)

- For $E[Y|G_1, G_2, X) = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_X X + \beta G_1 G_2$ a sufficient cause interaction is present if

$$\beta > \beta_0.$$

- When both exposures have monotonic effects on the outcome, this can be strengthened to

$$\beta > 0.$$

(X suffices to control for confounding of the estimation of G_1, G_2 effects)

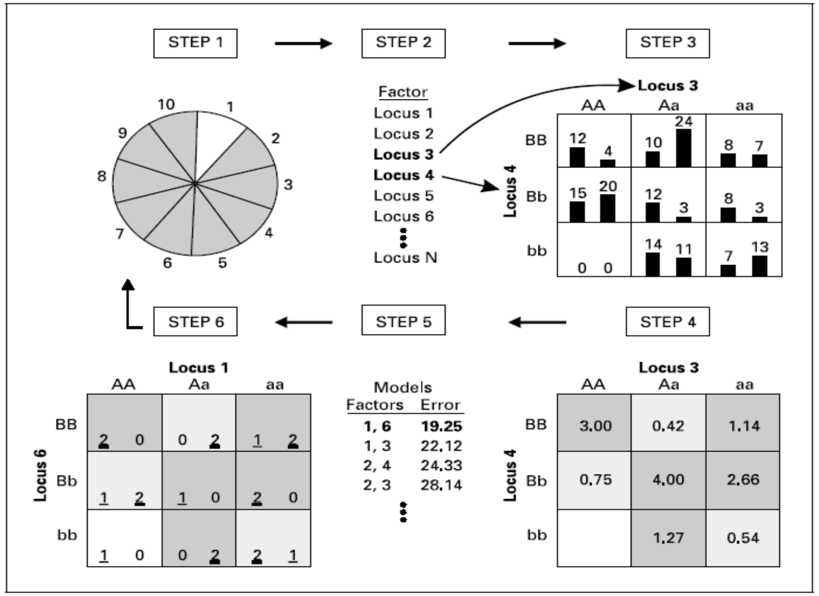
Mechanistic interactions

(adapted from slide: S Vansteelandt)

- Issues:
  - Tests for sufficient cause interactions involve testing on the risk difference scale
  - Reality may show high-dimensional confounding
  - Estimators and tests for interactions are needed that are robust to model misspecification
- Possible solution:
  - Semi-parametric interaction models that attempt to estimate statistical interactions without modeling the main effects
- Comment: already hard in the case of two SNPs, using a theory of causality that is not widely accessible.

Multifactor Dimensionality Reduction (MDR) (Ritchie et al 2001)

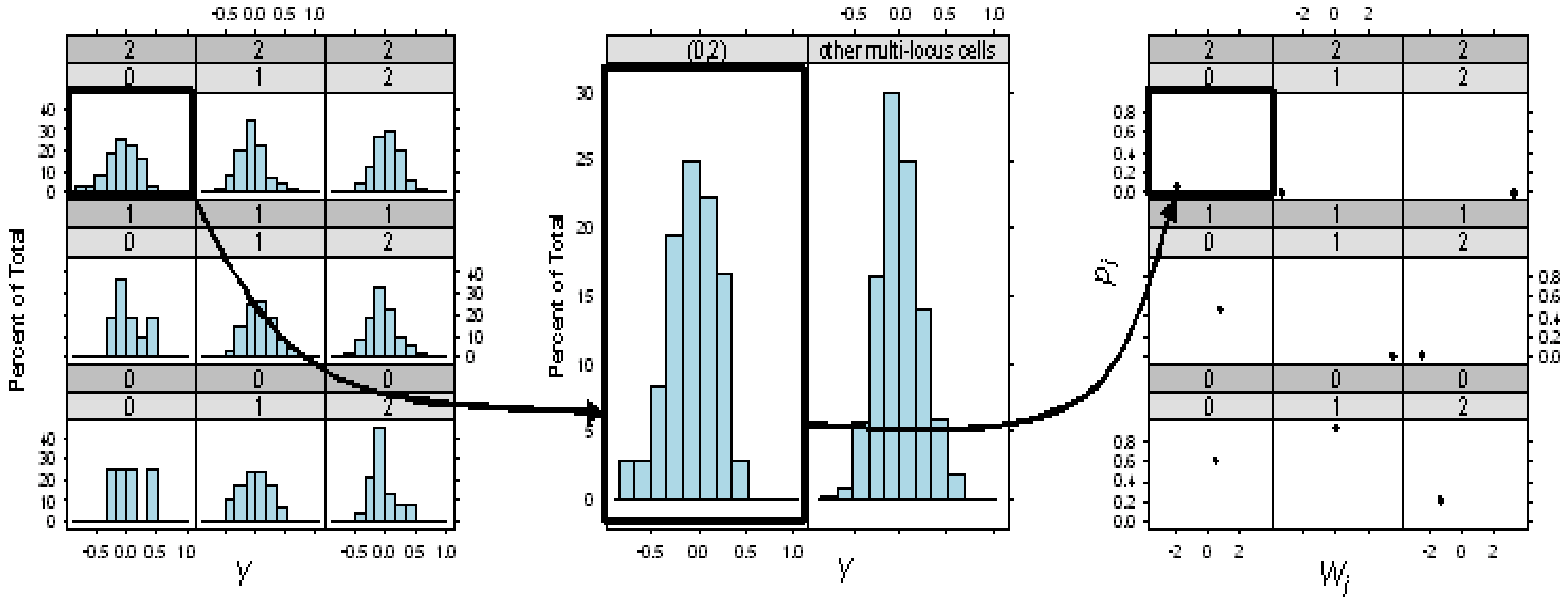
- A model-free and non-parametric approach to epistasis detection
- Was proposed to overcome the problem that the type of encoding of SNPs affects the results in generalized linear models; does not assume a specific genetic model
- Measures the association between SNPs and disease risk using prediction accuracy of selected multifactor models (relies on CV!!!).



Université de Liège

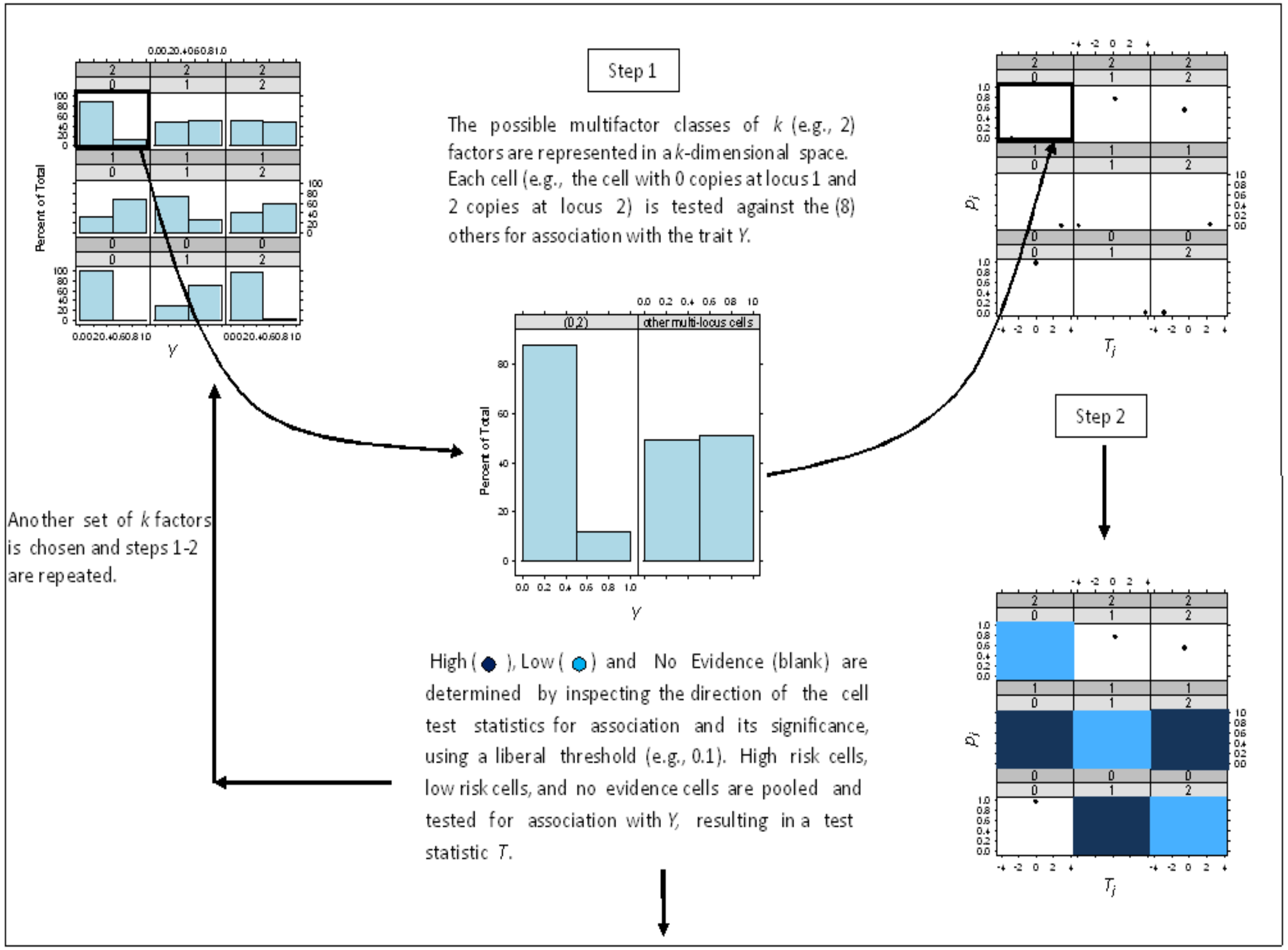
Model-Based Multifactor Dimensionality Reduction (MB-MDR)

- Graphical workflow



(Calle et al 2008, Cattaert et al 2010)

Model-Based Multifactor Dimensionality Reduction (MB-MDR)



Step 1

The possible multifactor classes of *k* (e.g., 2) factors are represented in a *k*-dimensional space. Each cell (e.g., the cell with 0 copies at locus 1 and 2 copies at locus 2) is tested against the (8) others for association with the trait *Y*.

Step 2

High (●), Low (●) and No Evidence (blank) are determined by inspecting the direction of the cell test statistics for association and its significance, using a liberal threshold (e.g., 0.1). High risk cells, low risk cells, and no evidence cells are pooled and tested for association with *Y*, resulting in a test statistic *T*.

Another set of *k* factors is chosen and steps 1-2 are repeated.

Step 3

Assessing significance of the *k*-factor models, hereby correcting for multiple testing.

MB-MDR advantage 1

- Some important interactions could be missed by MDR due to pooling too many cells together

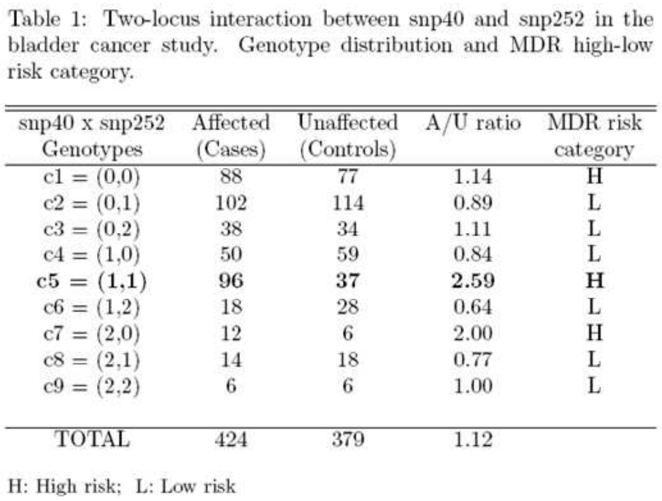


Table 1: Two-locus interaction between snp40 and snp252 in the bladder cancer study. Genotype distribution and MDR high-low risk category.

snp40 x snp252 Genotypes	Affected (Cases)	Unaffected (Controls)	A/U ratio	MDR risk category
c1 = (0,0)	88	77	1.14	H
c2 = (0,1)	102	114	0.89	L
c3 = (0,2)	38	34	1.11	L
c4 = (1,0)	50	59	0.84	L
c5 = (1,1)	**96**	**37**	**2.59**	**H**
c6 = (1,2)	18	28	0.64	L
c7 = (2,0)	12	6	2.00	H
c8 = (2,1)	14	18	0.77	L
c9 = (2,2)	6	6	1.00	L
TOTAL	424	379	1.12	

H: High risk; L: Low risk

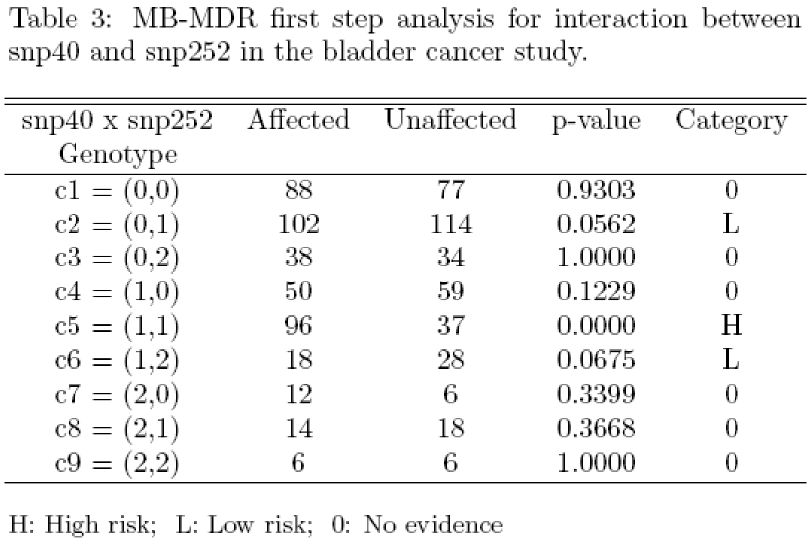


Table 3: MB-MDR first step analysis for interaction between snp40 and snp252 in the bladder cancer study.

snp40 x snp252 Genotype	Affected	Unaffected	p-value	Category
c1 = (0,0)	88	77	0.9303	0
c2 = (0,1)	102	114	0.0562	L
c3 = (0,2)	38	34	1.0000	0
c4 = (1,0)	50	59	0.1229	0
c5 = (1,1)	96	37	0.0000	H
c6 = (1,2)	18	28	0.0675	L
c7 = (2,0)	12	6	0.3399	0
c8 = (2,1)	14	18	0.3668	0
c9 = (2,2)	6	6	1.0000	0

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

(Calle et al 2008)

MB-MDR advantage 2

- MDR has difficulties with main effects and confounding factors corrections, as well as non-dichotomous outcomes

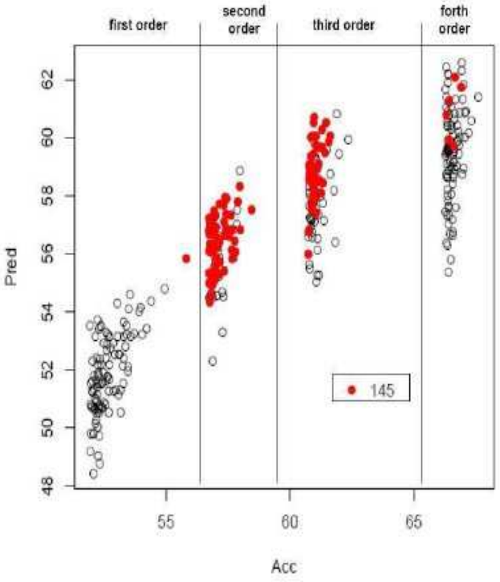


Fig. 1. Average Balanced Training accuracy (Acc) versus Average Balanced Predictive accuracy (Pred) for the 100 models with higher balanced training accuracy for the whole sample. First, second, third and forth order interactions are considered.

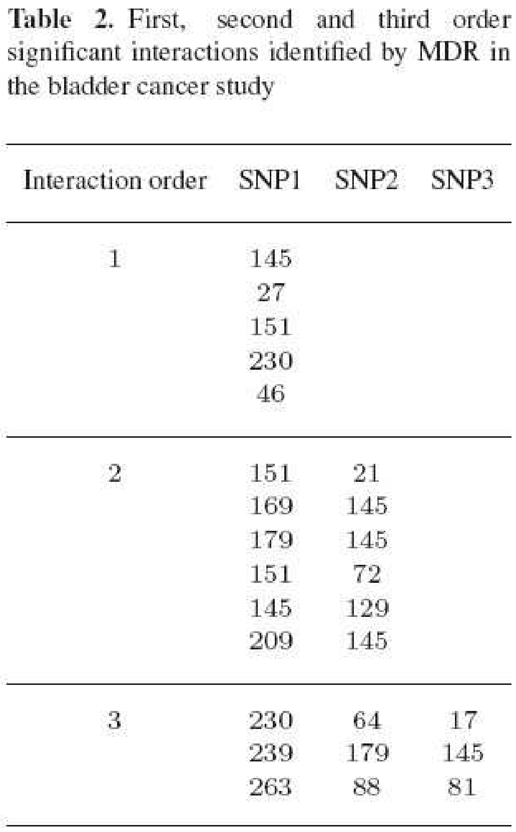


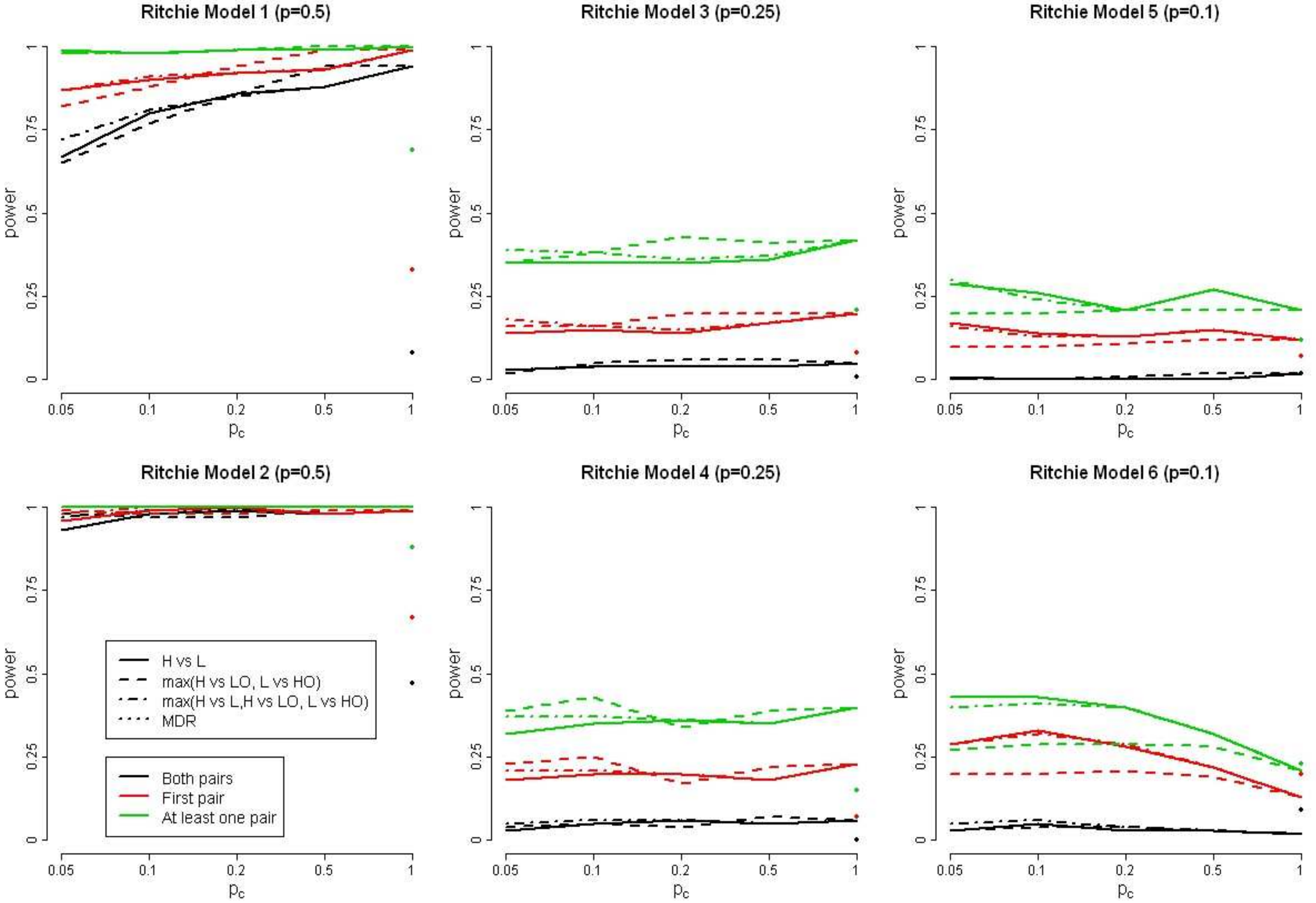
Table 2. First, second and third order significant interactions identified by MDR in the bladder cancer study

Interaction order	SNP1	SNP2	SNP3
1	145		
	27		
	151		
	230		
	46		
2	151	21	
	169	145	
	179	145	
	151	72	
	145	129	
	209	145	
3	230	64	17
	239	179	145
	263	88	81

MB-MDR advantage 3

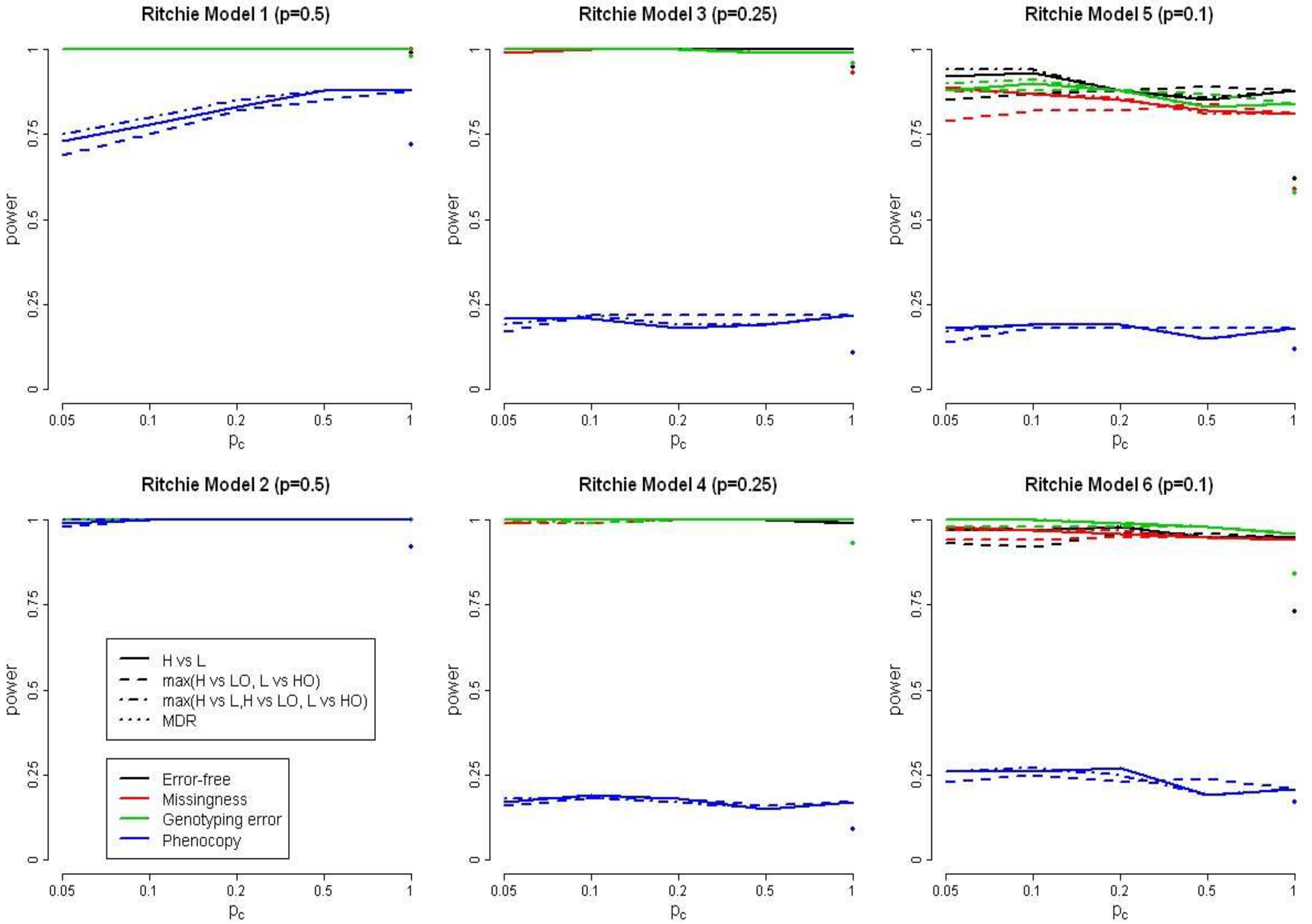
(Cattaert et al 2010)

- MDR has low performance in the presence of genetic heterogeneity



MB-MDR advantage 4 (Cattaert et al 2010)

- Maximize power for the already "difficult" epistasis screens



MB-MDR advantage 5

- False positive percentages under alternatives (Cattaert et al 2010)

	Model 1		**Model 6**	
Error	**MB-MDR**	**MDR**	**MB-MDR**	**MDR**
None	6	9	5	23
Genotyping Error	2	14	4	23
Genetic Heterogeneity	4	7	2	17
Phenocopies	6	8	3	11
Missing Genotypes	7	16	7	24

Family-wise error rates (FWER) are shown for MB-MDR (MB) with $p_c = 0.1$ using the $T = |T_{H/L}|$ test approach and MaxT multiple testing correction and for MDR screening first-to-fifth-order models. Model 1: pure epistasis, MAF=0.5; Model 6: pure epistasis, MAF=0.10

The MB-MDR Software

Downloads

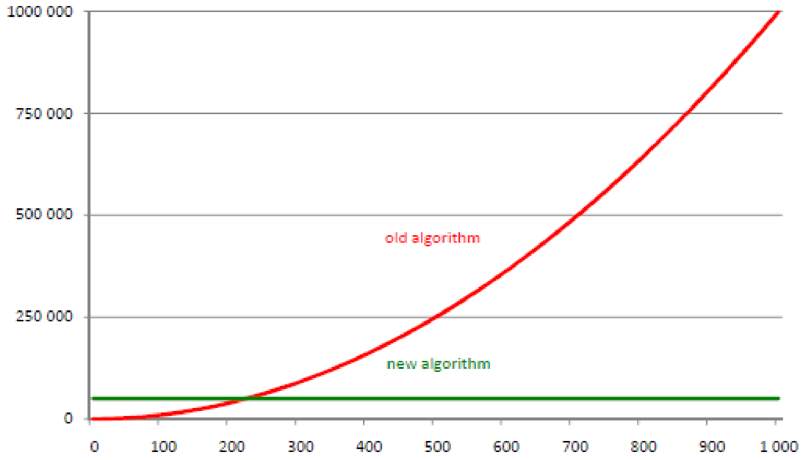
- A simplified version of MB-MDR is available in the free software R as an mbmdr package (http://cran.r-project.org/) and described in Calle et al (2010)
- A comprehensive MB-MDR executable file of an efficient C++ implementation is available from K Van Steen (kristel.vansteen@ulg.ac.be) or via www.statgen.be

Features

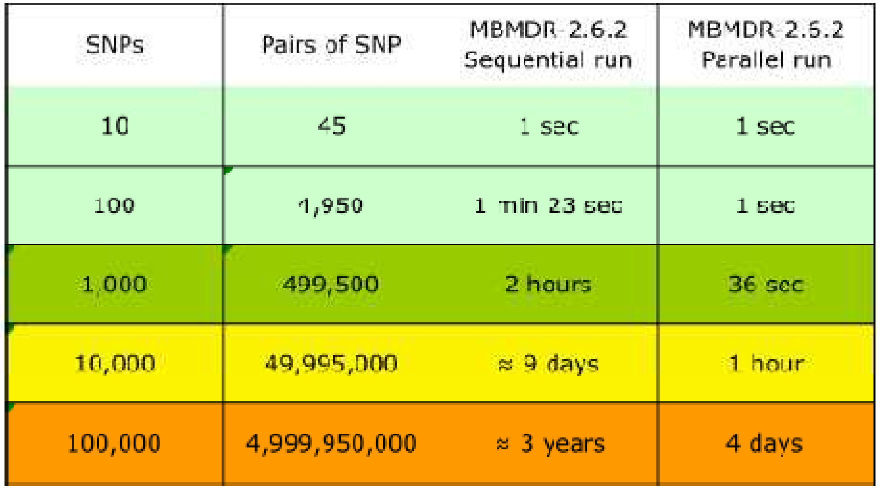
- Continuous, dichotomous, censored; univariate and multivariate
- Covariate correction on-the-fly
- Population-based and family-based designs

The MB-MDR Software

- Multiple testing (memory usage)



- Parallel run on 50 quad-core AMD opteron 2.1 GHz

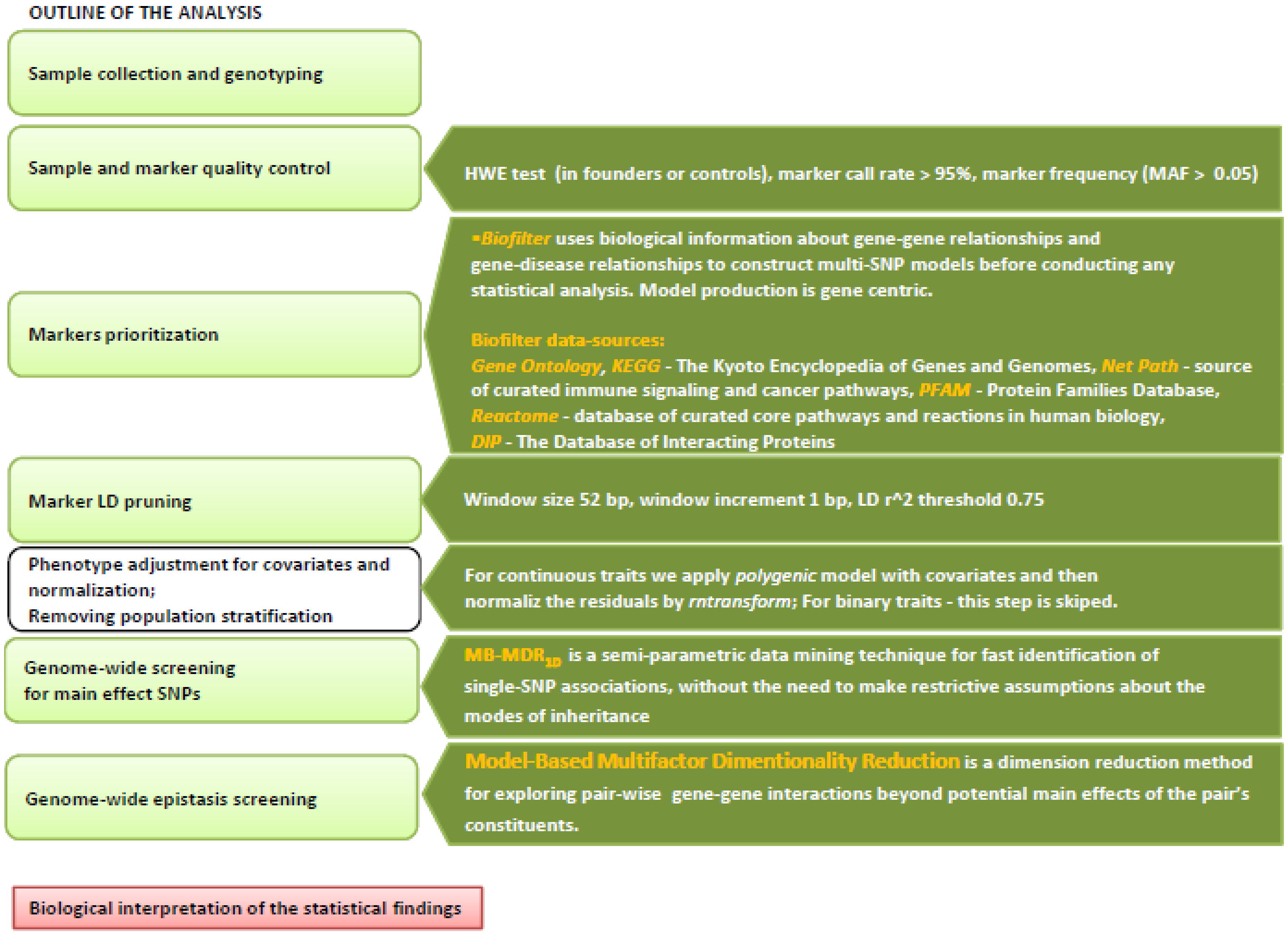


SNPs	Pairs of SNP	MBMDR-2.6.2 Sequential run	MBMDR-2.6.2 Parallel run
10	45	1 sec	1 sec
100	4,950	1 min 23 sec	1 sec
1,000	499,500	2 hours	36 sec
10,000	49,995,000	≈ 9 days	1 hour
100,000	4,999,950,000	≈ 3 years	4 days

A minimal GWAIs protocol

Université
de Liège

GWAIs protocol



OUTLINE OF THE ANALYSIS

Sample collection and genotyping

Sample and marker quality control

HWE test (in founders or controls), marker call rate > 95%, marker frequency (MAF > 0.05)

Markers prioritization

▪*Biofilter* uses biological information about gene-gene relationships and gene-disease relationships to construct multi-SNP models before conducting any statistical analysis. Model production is gene centric.

Biofilter data-sources:
Gene Ontology, *KEGG* - The Kyoto Encyclopedia of Genes and Genomes, *Net Path* - source of curated immune signaling and cancer pathways, *PFAM* - Protein Families Database, *Reactome* - database of curated core pathways and reactions in human biology, *DIP* - The Database of Interacting Proteins

Marker LD pruning

Window size 52 bp, window increment 1 bp, LD r^2 threshold 0.75

Phenotype adjustment for covariates and normalization;
Removing population stratification

For continuous traits we apply *polygenic* model with covariates and then normaliz the residuals by *rntransform*; For binary traits - this step is skiped.

Genome-wide screening for main effect SNPs

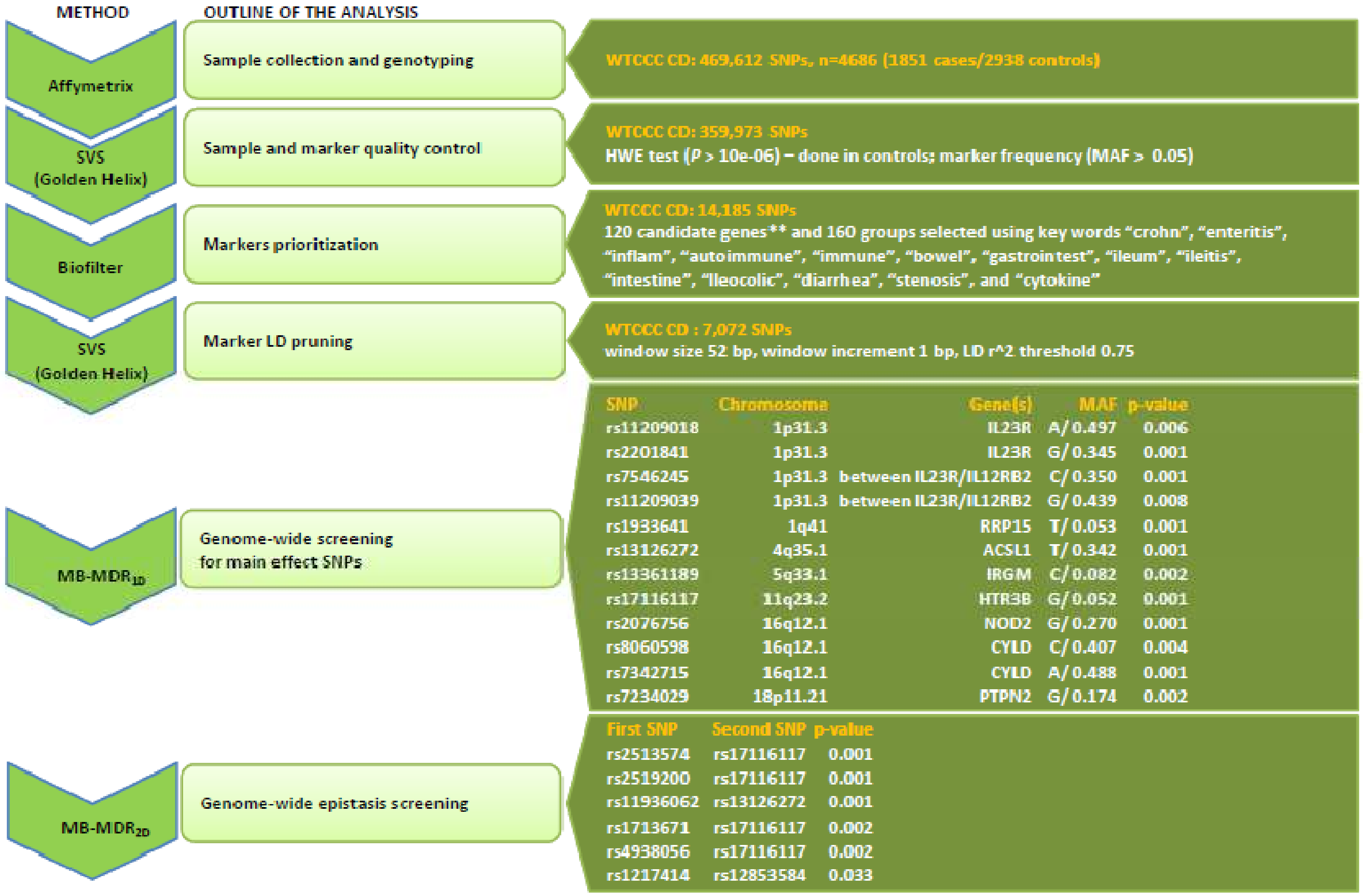
MB-MDR_{1D} is a semi-parametric data mining technique for fast identification of single-SNP associations, without the need to make restrictive assumptions about the modes of inheritance

Genome-wide epistasis screening

Model-Based Multifactor Dimentionality Reduction is a dimension reduction method for exploring pair-wise gene-gene interactions beyond potential main effects of the pair's constituents.

Biological interpretation of the statistical findings

A GWAIs protocol in action



METHOD

OUTLINE OF THE ANALYSIS

Affymetrix

Sample collection and genotyping

WTCCC CD: 469,612 SNPs, n=4686 (1851 cases/2938 controls)

SVS (Golden Helix)

Sample and marker quality control

WTCCC CD: 359,973 SNPs

HWE test (P > 10e-06) – done in controls; marker frequency (MAF > 0.05)

Biofilter

Markers prioritization

WTCCC CD: 14,185 SNPs

120 candidate genes** and 160 groups selected using key words “crohn”, “enteritis”, “inflam”, “autoimmune”, “immune”, “bowel”, “gastrointest”, “ileum”, “ileitis”, “intestine”, “Ileocolic”, “diarrhea”, “stenosis”, and “cytokine”

SVS (Golden Helix)

Marker LD pruning

WTCCC CD : 7,072 SNPs

window size 52 bp, window increment 1 bp, LD r^2 threshold 0.75

MB-MDR_{1D}

Genome-wide screening for main effect SNPs

SNP	Chromosome	Gene(s)	MAF	p-value
rs11209018	1p31.3	IL23R	A/ 0.497	0.006
rs2201841	1p31.3	IL23R	G/ 0.345	0.001
rs7546245	1p31.3	between IL23R/IL12RB2	C/ 0.350	0.001
rs11209039	1p31.3	between IL23R/IL12RB2	G/ 0.439	0.008
rs1933641	1q41	RRP15	T/ 0.053	0.001
rs13126272	4q35.1	ACSL1	T/ 0.342	0.001
rs13361189	5q33.1	IRGM	C/ 0.082	0.002
rs17116117	11q23.2	HTR3B	G/ 0.052	0.001
rs2076756	16q12.1	NOD2	G/ 0.270	0.001
rs8060598	16q12.1	CYLD	C/ 0.407	0.004
rs7342715	16q12.1	CYLD	A/ 0.488	0.001
rs7234029	18p11.21	PTPN2	G/ 0.174	0.002

MB-MDR_{2D}

Genome-wide epistasis screening

First SNP	Second SNP	p-value
rs2513574	rs17116117	0.001
rs2519200	rs17116117	0.001
rs11936062	rs13126272	0.001
rs1713671	rs17116117	0.002
rs4938056	rs17116117	0.002
rs1217414	rs12853584	0.033

Main challenge: Assess which findings to pursue ~ interpretation

- Challenges:
  1. Same chromosome or not?

  2. What are the LD-friends related to our pairs of interest?

  3. Target pairs that can be replicated by different methodologies?
     - Different steps in the GWAI process
     - Different approaches within one step

  4. Target pairs that can be mapped to underlying biological epistasis networks or pathways?

Challenge 1

- Same chromosome or not? (Composites in LD → haplotype analysis)

r^2=0.110

r^2=0.047

r^2=0.022

r^2=0.027

r^2=0.027

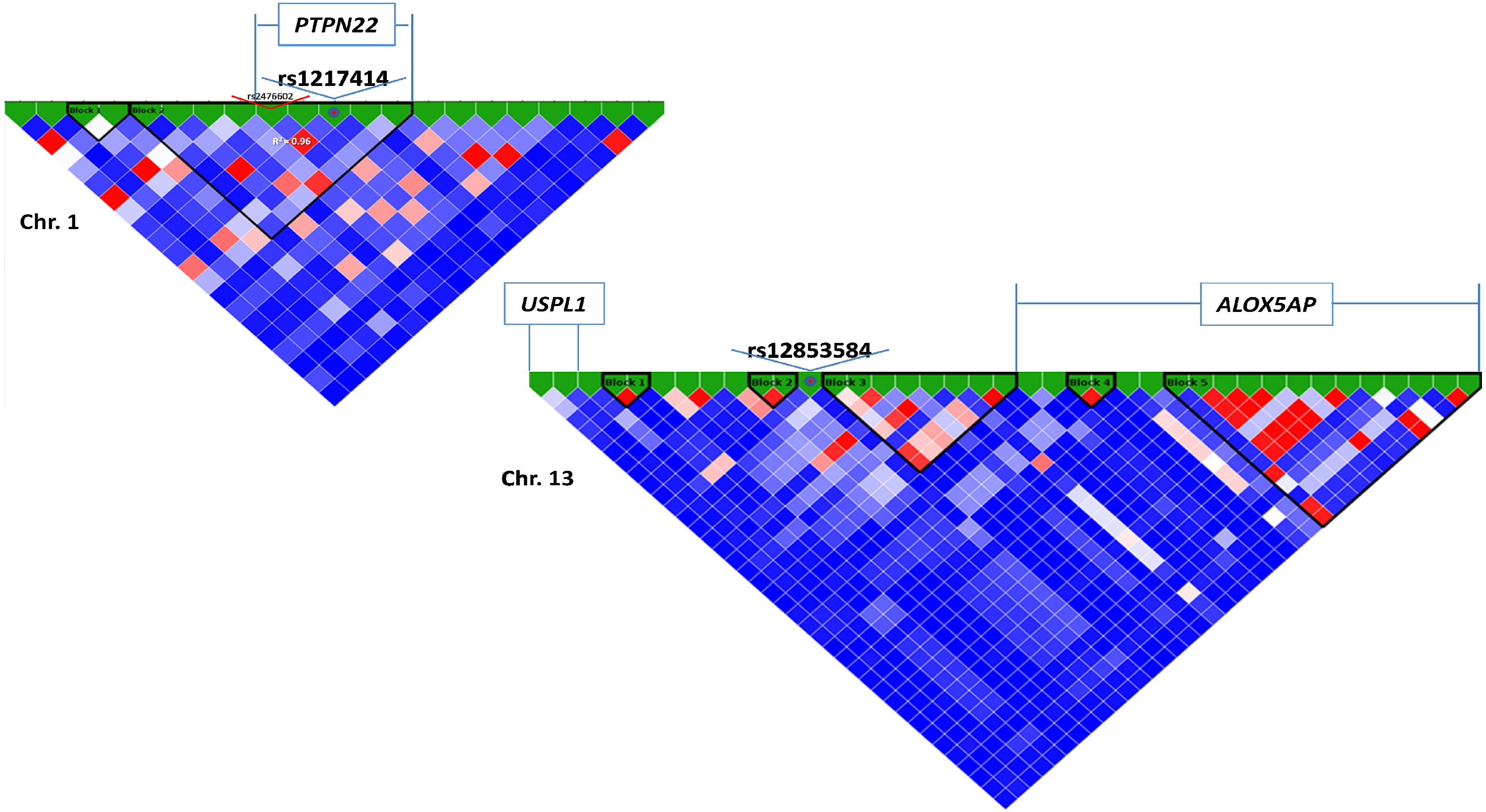
SNP	SNP position	Gene	Main effect	MAF
rs17116117	chr11:113801591	HTR3B	**0,001**	0,052
rs2513574	chr11:113681305	USP28	>0.05	0,123
rs2519200	chr11:113684809	USP28	>0.05	0,238
rs1713671	chr11:113674838	USP28	>0.05	0,416
rs4938056	chr11:113786539	HTR3B	>0.05	0,400
rs11936062	chr4:185721370	SLED1	>0.05	0,165
rs13126272	chr4:185731940	ACSL1	**0,001**	0,342
rs1217414	chr1:114412667	PTPN22	>0.05	0,273
rs12853584	chr13:31279946	between USPL1/ALOX5AP	>0.05	0,272

Université de Liège

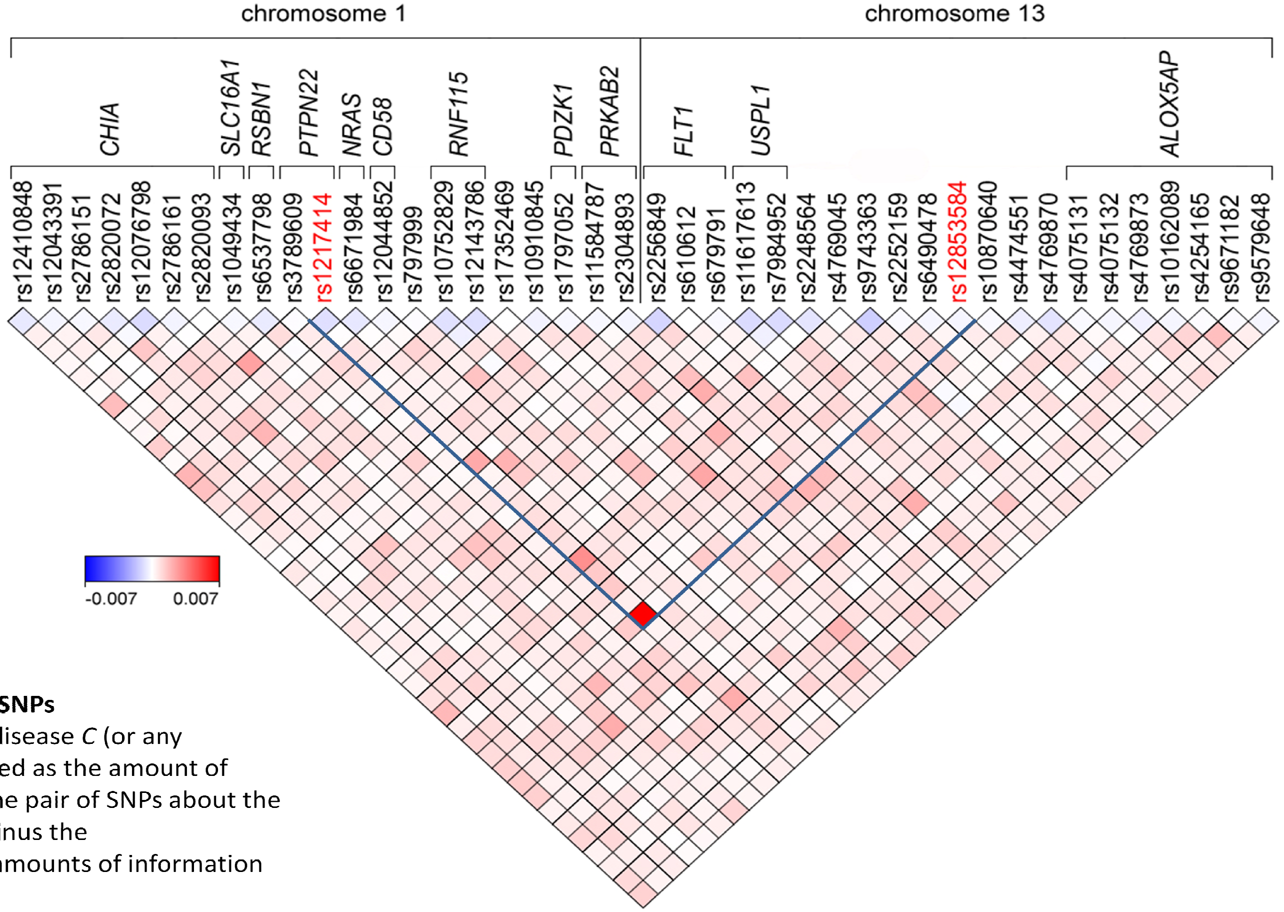
Challenge 2

- What are the LD-friends related to our pairs of interest?

LD plots (r^2) – before LD pruning:



Synergy Disequilibrium (SD) plots: LD ≠ interaction



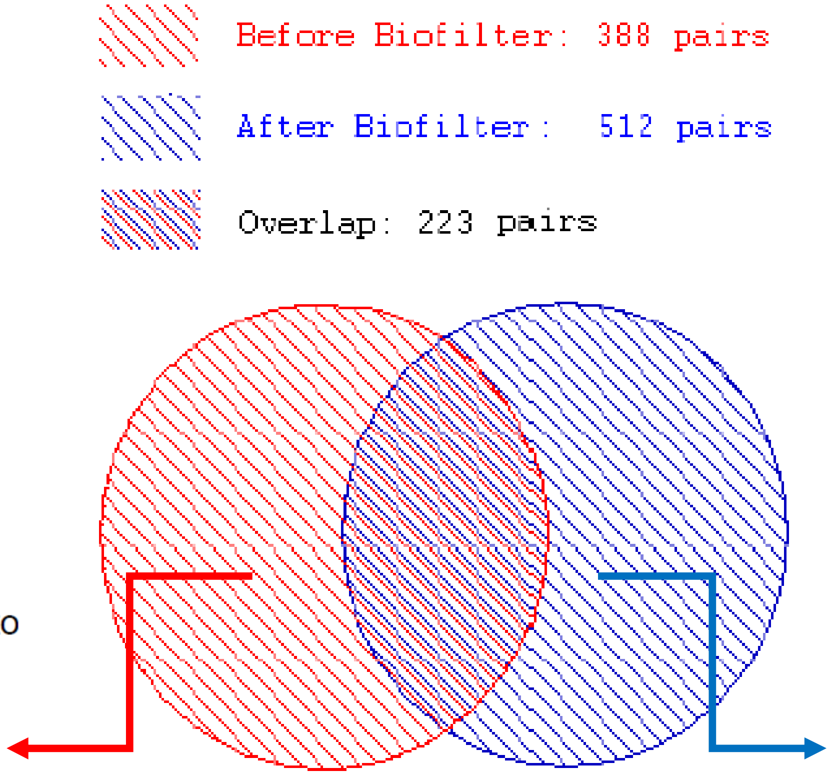
The synergy between two SNPs *Si* and *Sj* with respect to a disease *C* (or any phenotype or trait) is defined as the amount of information conveyed by the pair of SNPs about the presence of the disease, minus the sum of the corresponding amounts of information conveyed by each SNP:

$I(Si,Sj;C)–I(Si;C)+I(Sj;C)$

Challenge 3

- ***Different steps in the GWAI process***
  - What is the danger / benefit of filtering?

Application on WTCCC Rheumatoid Arthritis (RA)



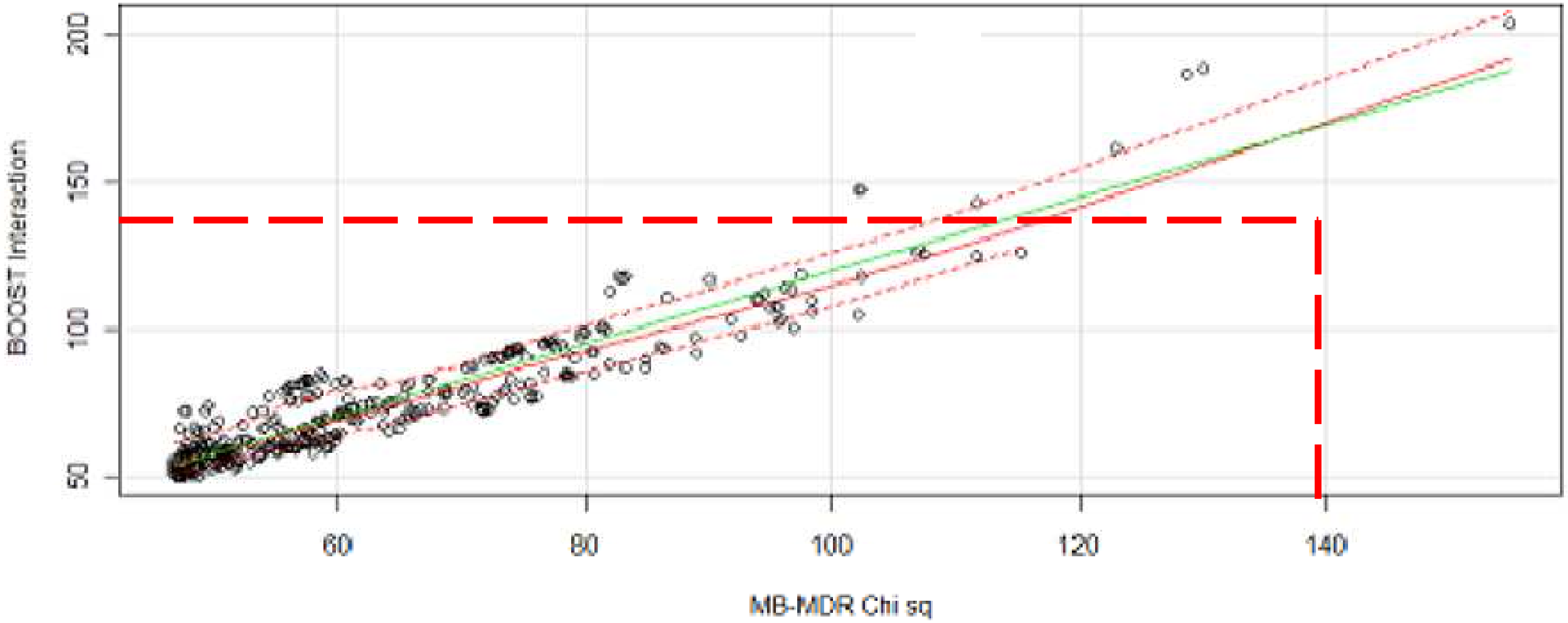
165 "lost" pairs contain 191 SNPs:

- 18 of them passed the Biofilter.
- 173 did not:
  - 55 can be mapped to genes.
  - 118 in intergenic regions.

289 new pairs after Biofilter: is Bonferroni correction too severe?

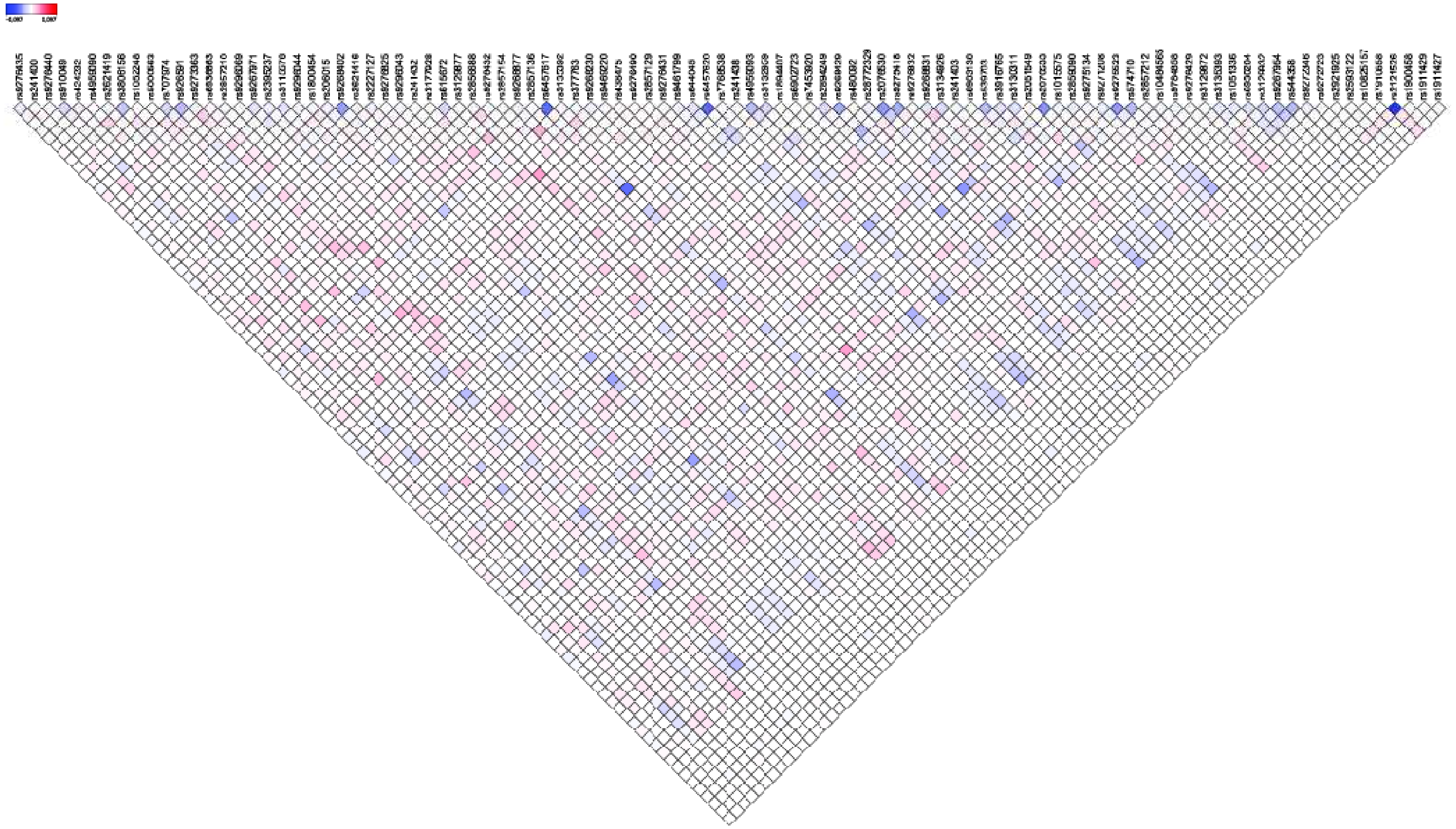
Université de Liège

- ***Different approaches within a single step of the GWAI process***
  - On the same Bio-filtered data, up-scaled logistic regression software (Wan et al. 2010) reports 512 significant pairs and MB-MDR 401: 395 significant pairs in common for RA …

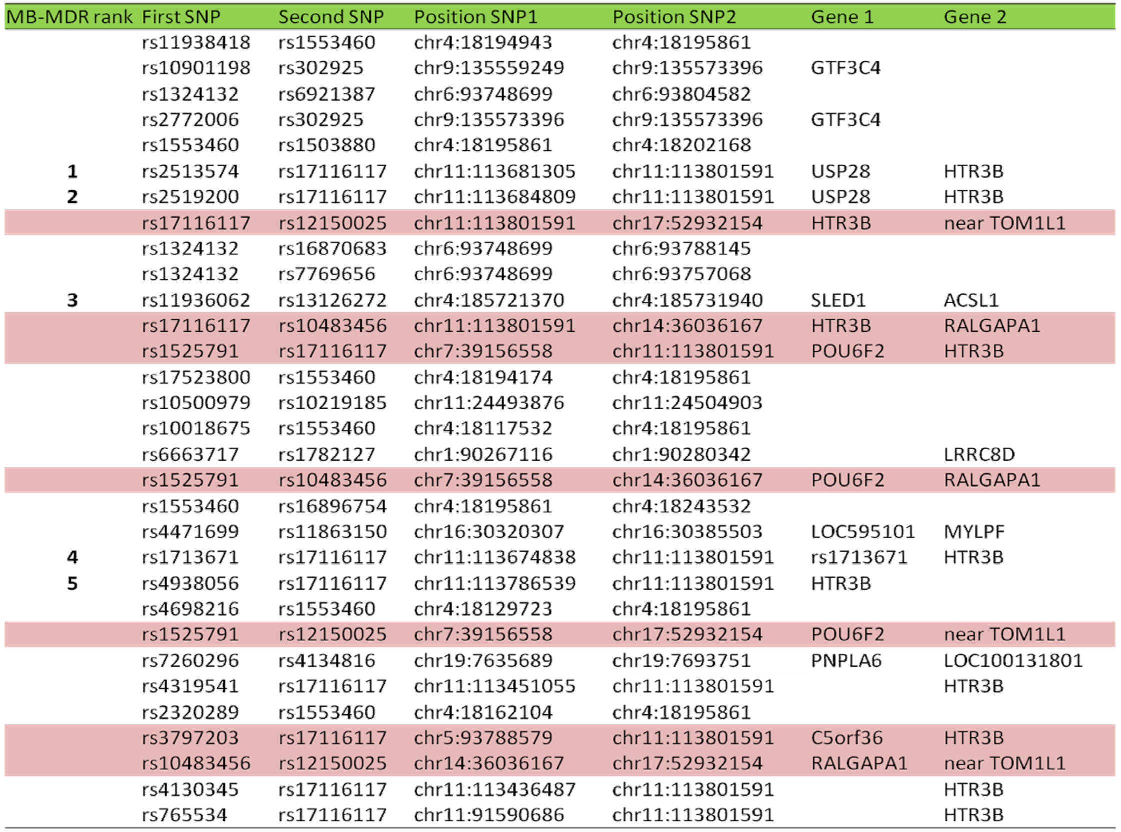


117 pairs detected by BOOST but not by MB-MDR!

- SD between SNPs in pairs detected by both BOOST only: More false positives by regression approaches?



- For the aforementioned **unfiltered** CD data, BOOST finds 26 additional significant pairs, compared to MB-MDR on **Bio-filtered** data: What to believe?



MB-MDR rank	First SNP	Second SNP	Position SNP1	Position SNP2	Gene 1	Gene 2
	rs11938418	rs1553460	chr4:18194943	chr4:18195861		
	rs10901198	rs302925	chr9:135559249	chr9:135573396	GTF3C4	
	rs1324132	rs6921387	chr6:93748699	chr6:93804582		
	rs2772006	rs302925	chr9:135573396	chr9:135573396	GTF3C4	
	rs1553460	rs1503880	chr4:18195861	chr4:18202168		
1	rs2513574	rs17116117	chr11:113681305	chr11:113801591	USP28	HTR3B
2	rs2519200	rs17116117	chr11:113684809	chr11:113801591	USP28	HTR3B
	rs17116117	rs12150025	chr11:113801591	chr17:52932154	HTR3B	near TOM1L1
	rs1324132	rs16870683	chr6:93748699	chr6:93788145		
	rs1324132	rs7769656	chr6:93748699	chr6:93757068		
3	rs11936062	rs13126272	chr4:185721370	chr4:185731940	SLED1	ACSL1
	rs17116117	rs10483456	chr11:113801591	chr14:36036167	HTR3B	RALGAPA1
	rs1525791	rs17116117	chr7:39156558	chr11:113801591	POU6F2	HTR3B
	rs17523800	rs1553460	chr4:18194174	chr4:18195861		
	rs10500979	rs10219185	chr11:24493876	chr11:24504903		
	rs10018675	rs1553460	chr4:18117532	chr4:18195861		
	rs6663717	rs1782127	chr1:90267116	chr1:90280342		LRRC8D
	rs1525791	rs10483456	chr7:39156558	chr14:36036167	POU6F2	RALGAPA1
	rs1553460	rs16896754	chr4:18195861	chr4:18243532		
	rs4471699	rs11863150	chr16:30320307	chr16:30385503	LOC595101	MYLPF
4	rs1713671	rs17116117	chr11:113674838	chr11:113801591	rs1713671	HTR3B
5	rs4938056	rs17116117	chr11:113786539	chr11:113801591	HTR3B	
	rs4698216	rs1553460	chr4:18129723	chr4:18195861		
	rs1525791	rs12150025	chr7:39156558	chr17:52932154	POU6F2	near TOM1L1
	rs7260296	rs4134816	chr19:7635689	chr19:7693751	PNPLA6	LOC100131801
	rs4319541	rs17116117	chr11:113451055	chr11:113801591		HTR3B
	rs2320289	rs1553460	chr4:18162104	chr4:18195861		
	rs3797203	rs17116117	chr5:93788579	chr11:113801591	C5orf36	HTR3B
	rs10483456	rs12150025	chr14:36036167	chr17:52932154	RALGAPA1	near TOM1L1
	rs4130345	rs17116117	chr11:113436487	chr11:113801591		HTR3B
	rs765534	rs17116117	chr11:91590686	chr11:113801591		HTR3B

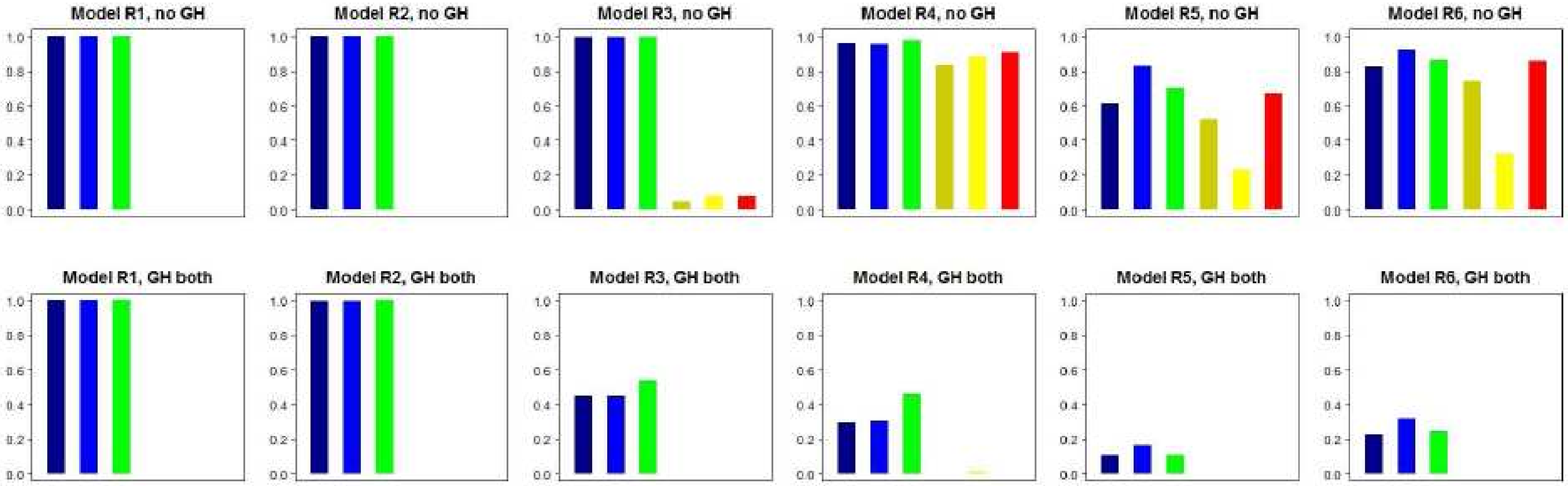
BOOST analysis:
32 SNP pairs
p-value < 0.05
(Bonferroni)

Université de Liège

Different approaches within a single step of the GWAI process (continued)

- Which epistasis detection method to choose?
- We have chosen MB-MDR and BOOST but there is an abundance of epistasis methods (Van Steen 2011) and even a larger compendium of "comparison papers" is available … Was our choice a clever one?
- Two criteria that help making a choice are:
  - power
  - Type I error (false positive rate)

Power (pure epistasis scenario's)



BOOST (dark blue)

EpiCruncher optimal options (light blue)

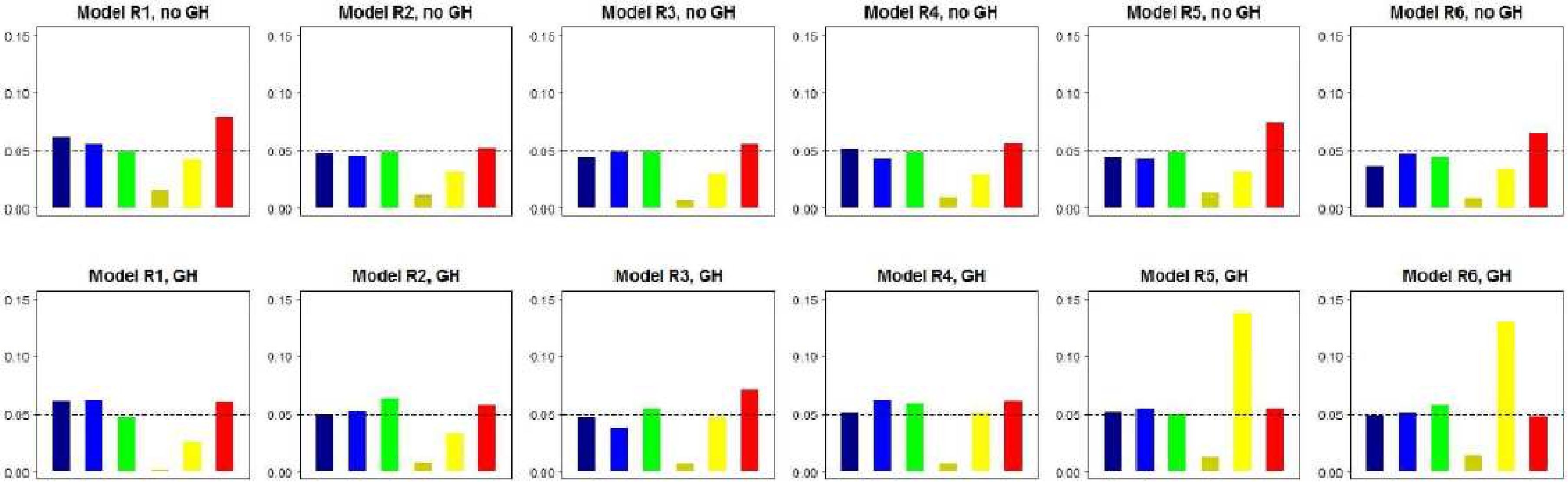
MB-MDR (green)

PLINK epistasis (dark yellow)

PLINK fast epistasis (light yellow)

EPIBLASTER (red)

Type I Error (pure epistasis scenario’s)



BOOST (dark blue)

EpiCruncher optimal options (light blue)

MB-MDR (green)

PLINK fast epistasis (light yellow)

EPIBLASTER (red)

PLINK epistasis (dark yellow)

- Concerns:
  - Are the comparisons “honest”?
  - What is the “core” (**the ABC**) of the method?
    - **A**: Pre-processing (screening); **B**: core; **C**: multiple testing

		EpiCruncher																MB-MDR	PLINK	EPIBLASTER
		Bonferroni								Permutations										
		LR test				Score test				LR test				Score test						
		Test statistic		P-value		Test statistic		P-value		Test statistic		P-value		Test statistic		P-value				
		M=1	M=5	M=1	M=5	M=1	M=5	M=1	M=5	M=1	M=5	M=1	M=5	M=1	M=5	M=1	M=5			
rs17116117	rs2513574	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rs17116117	rs2519200	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rs17116117	rs4938056	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
rs17116117	rs1713671	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
rs13126272	rs11936062	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
rs17116117	rs7126080	x	x	x	x					x	x	x	x							
rs3770132	rs1933641					x		x						x		x				
rs12339163	rs1933641					x		x						x		x				
rs12853584	rs1217414	🟥									x				x		x	x		
rs17116117	rs1169722																			x
number significant		**6**	**6**	**6**	**6**	**7**	**5**	**7**	**5**	**6**	**7**	**6**	**6**	**7**	**6**	**7**	**6**	**6**	**3**	**3**

BOOST

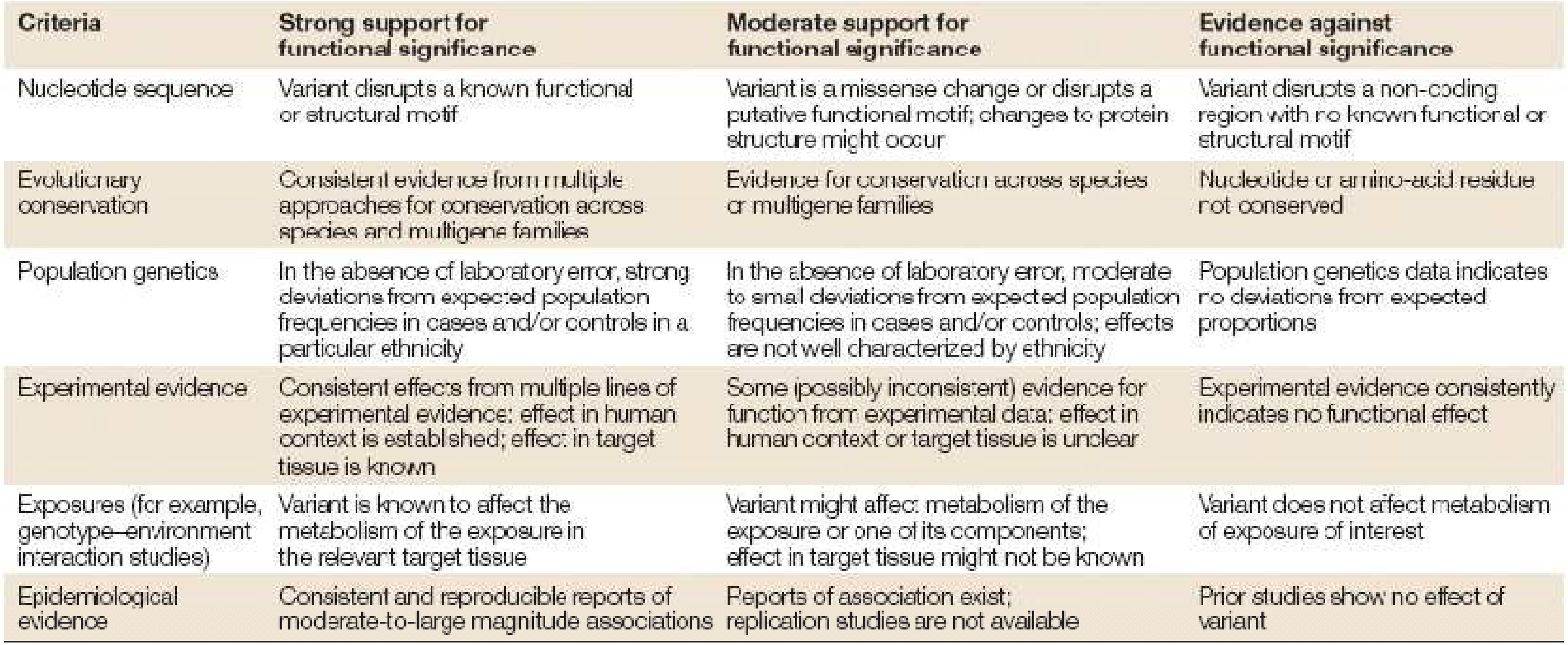
- Only by investigating the “information overlap” and “information complement” induced by different methodologies applied to the same data, one is able to either “interpret” different findings using different methods as a “pain” or a “confirmation”.

Ranks – same input WTCCC CD dataset based on 7,072 SNPs

		Epistasis Detection Method				
SNP Pair		**MBMBDR**	**EpiCruncher**	**BOOST**	**PLINK**	**EpiBlaster**
rs17116117	rs2513574	1	1	1	1	1
rs17116117	rs2519200	2	2	2	2	2
rs11936062	rs13126272	3	3	3	179	100
rs17116117	rs1713671	4	4	4	5	100
rs17116117	rs4938056	5	5	5	3	100
rs1217414	rs12853584	6	6	7	251	100
rs1169722	rs17116117	7	7	9	82	4
rs17116117	rs7126080	8	8	6	81	100
rs13126272	rs4862419	9	9	8	198	100
rs1933641	rs6099309	10	309	308	297	100

Challenge 4

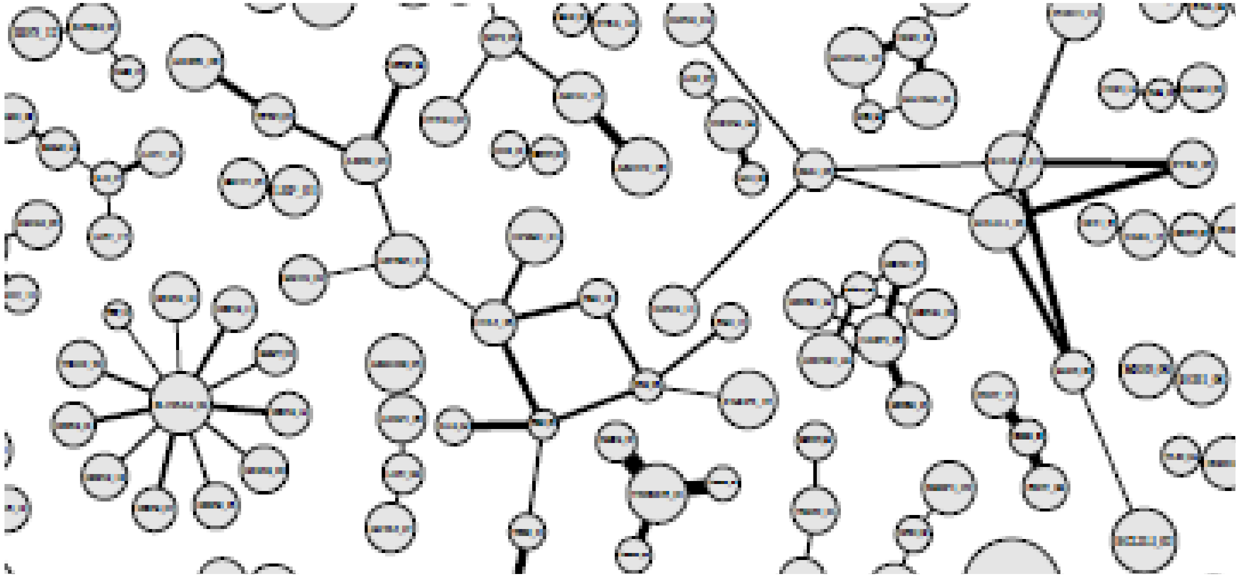
- Target pairs that can be mapped to underlying biological epistasis networks or pathways?
  - Criteria for assessing the functional significance of a variant



Criteria	Strong support for functional significance	Moderate support for functional significance	Evidence against functional significance
Nucleotide sequence	Variant disrupts a known functional or structural motif	Variant is a missense change or disrupts a putative functional motif; changes to protein structure might occur	Variant disrupts a non-coding region with no known functional or structural motif
Evolutionary conservation	Consistent evidence from multiple approaches for conservation across species and multigene families	Evidence for conservation across species or multigene families	Nucleotide or amino-acid residue not conserved
Population genetics	In the absence of laboratory error, strong deviations from expected population frequencies in cases and/or controls in a particular ethnicity	In the absence of laboratory error, moderate to small deviations from expected population frequencies in cases and/or controls; effects are not well characterized by ethnicity	Population genetics data indicates no deviations from expected proportions
Experimental evidence	Consistent effects from multiple lines of experimental evidence; effect in human context is established; effect in target tissue is known	Some (possibly inconsistent) evidence for function from experimental data; effect in human context or target tissue is unclear	Experimental evidence consistently indicates no functional effect
Exposures (for example, genotype–environment interaction studies)	Variant is known to affect the metabolism of the exposure in the relevant target tissue	Variant might affect metabolism of the exposure or one of its components; effect in target tissue might not be known	Variant does not affect metabolism of exposure of interest
Epidemiological evidence	Consistent and reproducible reports of moderate-to-large magnitude associations	Reports of association exist; replication studies are not available	Prior studies show no effect of variant

(Rebbeck et al. 2004)

- Criteria for assessing the functional significance of gene-gene interaction patterns are largely lacking
  - Would involve overlaying "statistical" epistasis networks with "biological" networks
  - Would involve linking hubs in "statistical" epistasis networks to functional groups or pathways



(Statistical epistasis network adapted from Hu et al. 2011)

Replication and validation of GWAIs: An impossible task?

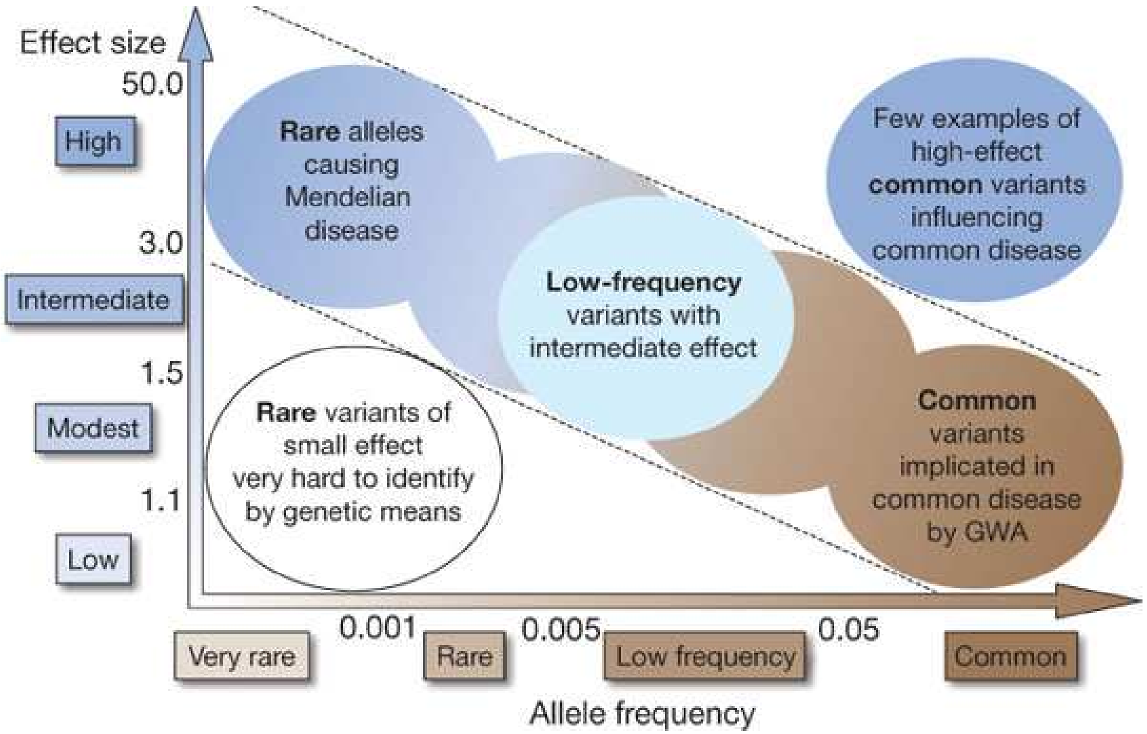
(Mission Impossible @ google)

Replication

- Replicating an association is the “gold standard” for “proving” an association is genuine
- Most epistasis signals underlying complex diseases will not be of large effect. It is unlikely that a single study will unequivocally establish an association without the need for replication
- Guidelines for replication studies include that these should be of sufficient size to demonstrate the effect … and should involve the same SNPs for testing ….

Optimal conditions for interaction replication

- Showing modest to strong statistical significance
- Having common minor allele frequency (>0.05)
- Modest to strong genetic effect sizes (parametric paradigms)

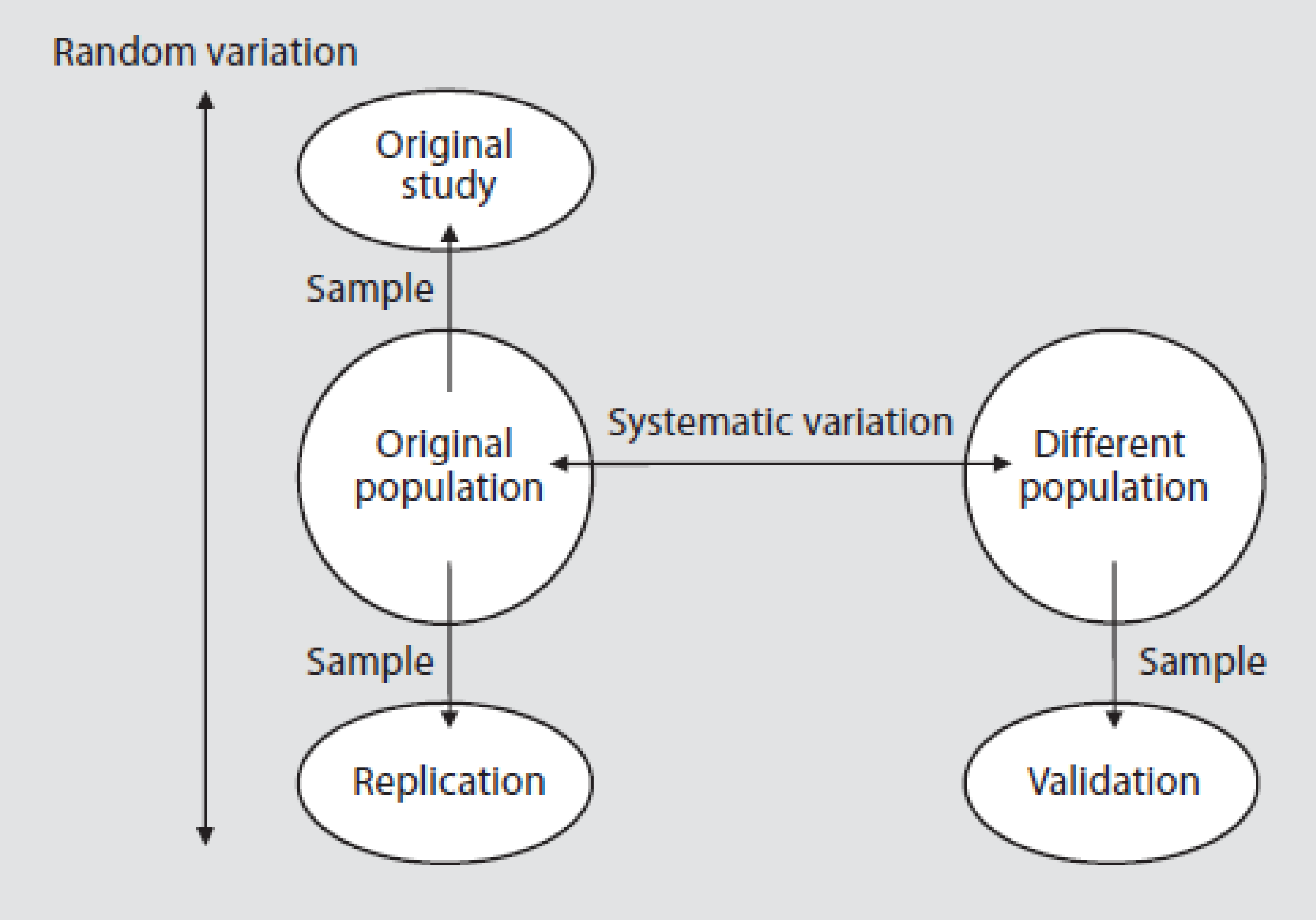

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

Compare to the diagonal focus region of GWAs

(Manolio et al. 2009)

Validation

- Validation is not replication:



(Igl et al. 2009)

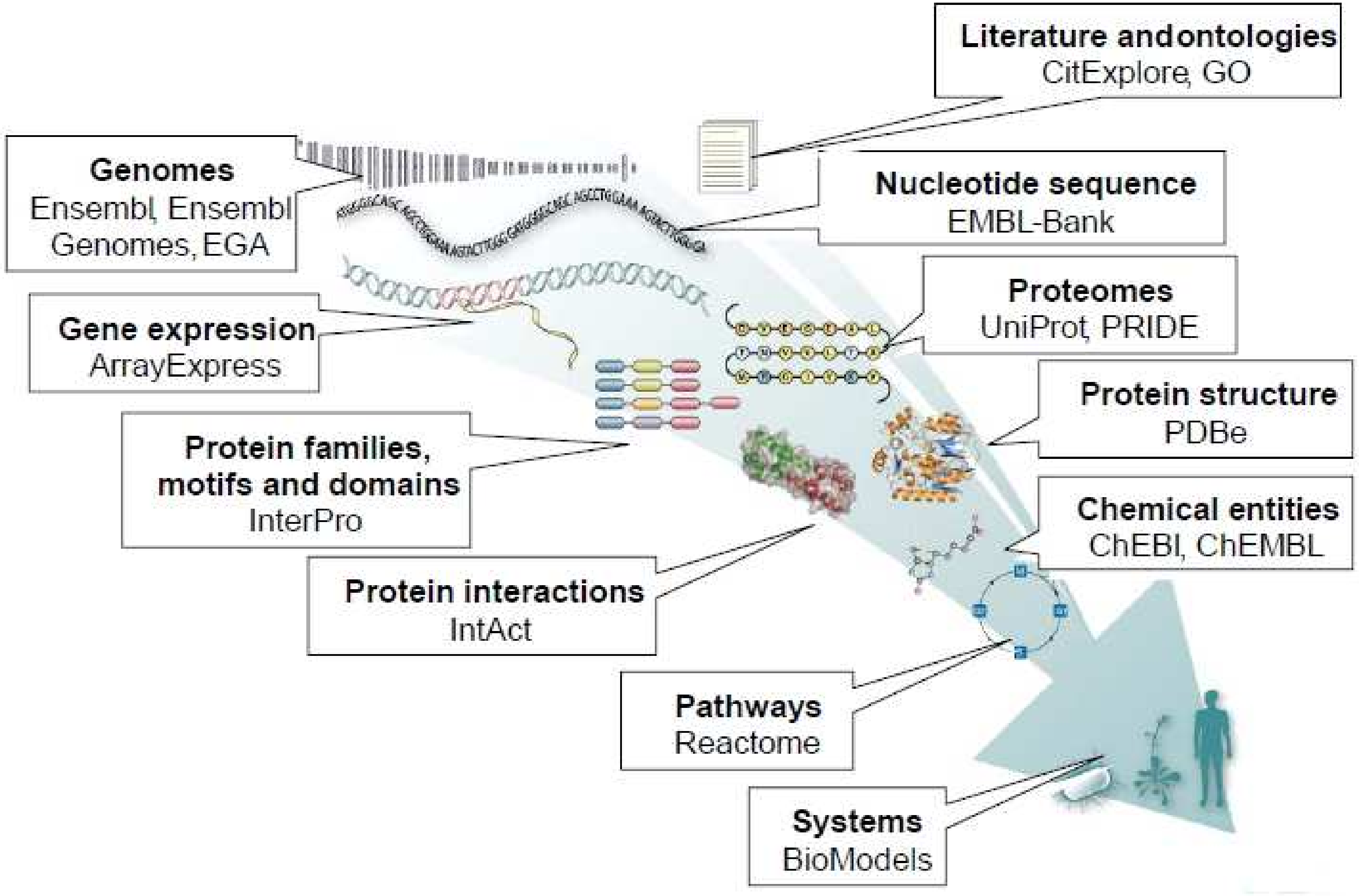
Université de Liège

Through the looking-glass

Meta-GWAI studies

- Given the availability of a comprehensive meta-analysis toolbox, it may be surprising that hardly any meta-GWAIs have been published as the core topic of the publication.
- This may in part be explained by the absence of strict guidelines or best practices for epistasis analysis, and the fact that new epistasis screening approaches arise every day.
- Additional complicating factors include:
  - Traditional meta-analysis methods in genetic association studies usually assume a specific genetic model of action to summarize the effect of genetic markers on a phenotype.
  - GWA imputation strategies ensure that different data sets are made comparable, but most be revised in the context of GWAIs.

Population based registries integrated with HTP omics



(www.elixir-europe.org 2010)

Omics integrative approaches for GWAIs and GWEIs

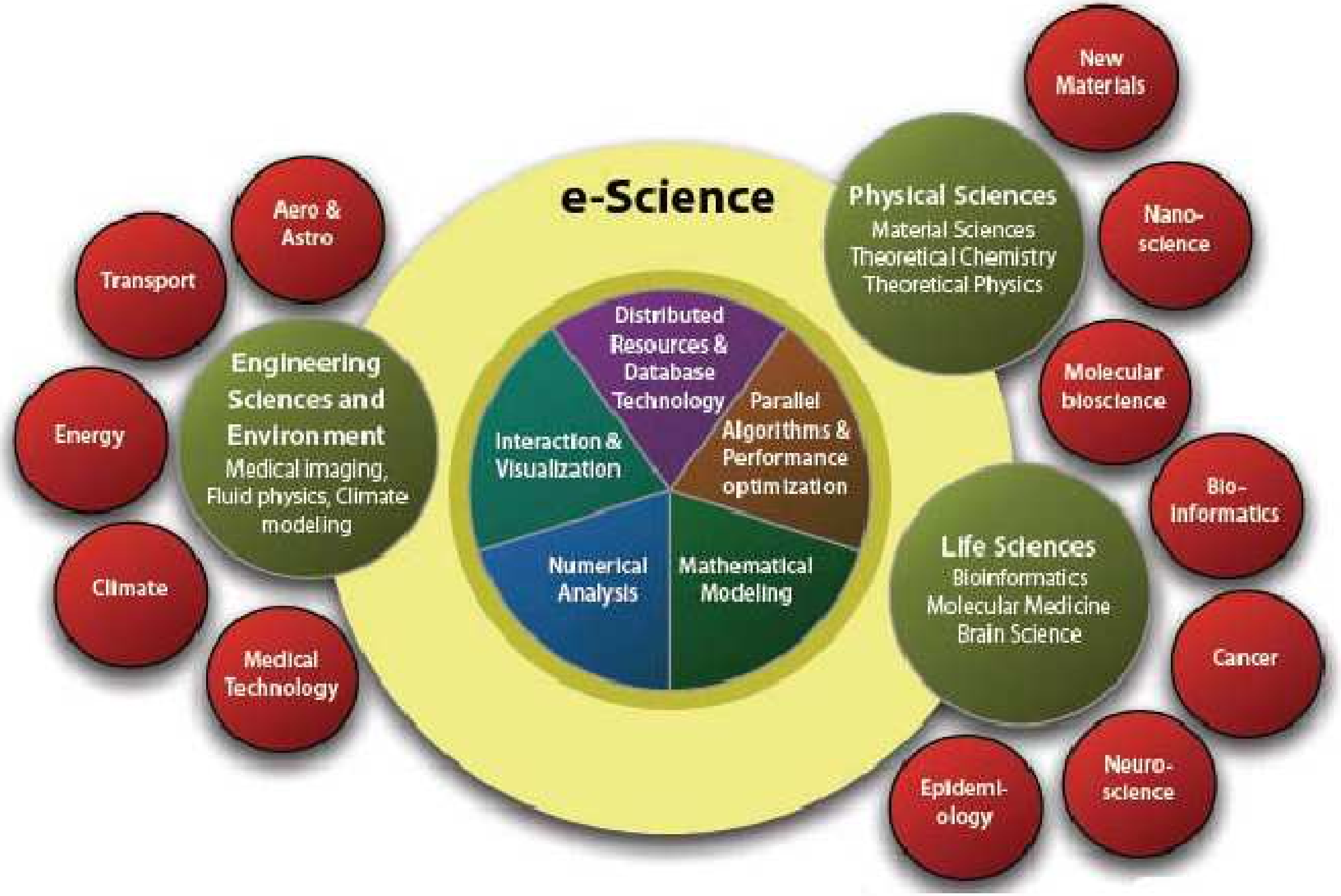
Example in GWAIs

- Before and after modeling using e.g. Biofilter
  - Assess and incorporate "optimal" scoring systems to accumulate evidence from these data bases
  - Allow for uncertainty involved in the data source entries
  - Acknowledge the complementary characteristics of each of the available data sources
  - Allow for different assignment strategies from genetic variants to genes

Example in GWEIs

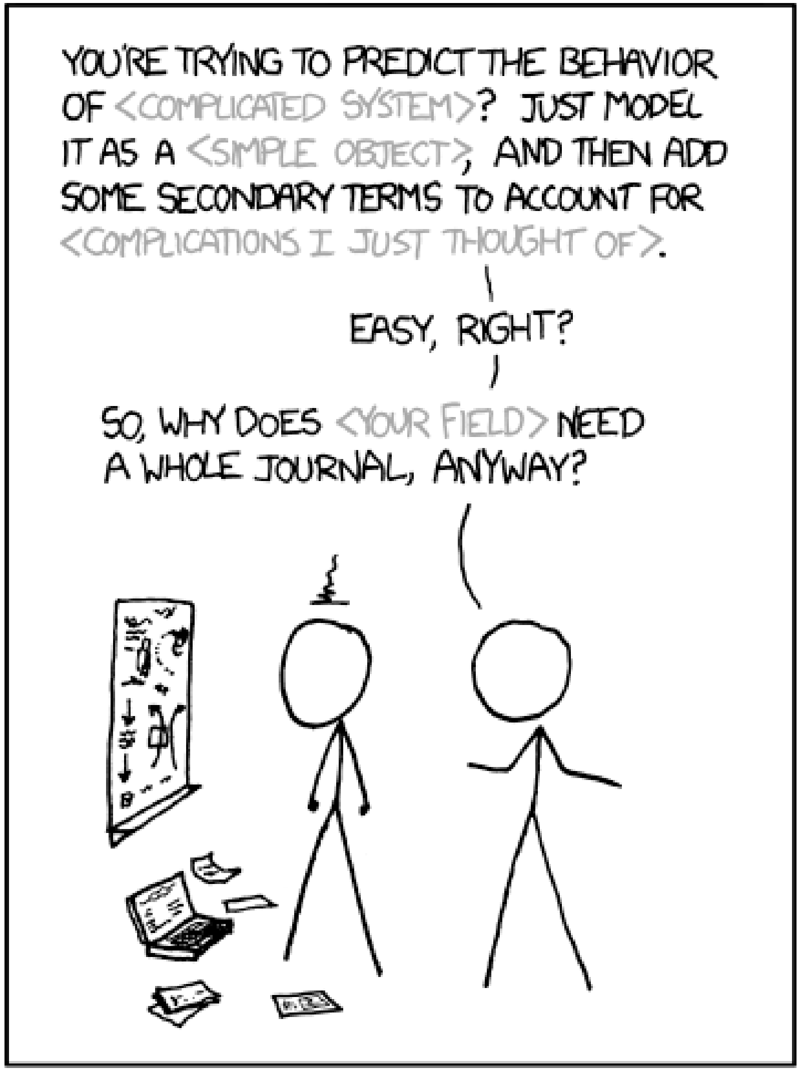
- When environmental epigenetic effects are operating, a heavily biology assistant-driven approach is required

Integration of technologies



(Harmonising biobank research – Brussels 2009)

THE


YOU'RE TRYING TO PREDICT THE BEHAVIOR OF <COMPLICATED SYSTEM>? JUST MODEL IT AS A <SIMPLE OBJECT>, AND THEN ADD SOME SECONDARY TERMS TO ACCOUNT FOR <COMPLICATIONS I JUST THOUGHT OF>.
EASY, RIGHT?
SO, WHY DOES <YOUR FIELD> NEED A WHOLE JOURNAL, ANYWAY?

LIBERAL-ARTS MAJORS MAY BE ANNOYING SOMETIMES, BUT THERE'S *NOTHING* MORE OBNOXIOUS THAN A PHYSICIST FIRST ENCOUNTERING A NEW SUBJECT.

END

Acknowledgments