Topics in Bioinformatics

Kristel Van Steen, PhD^2

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

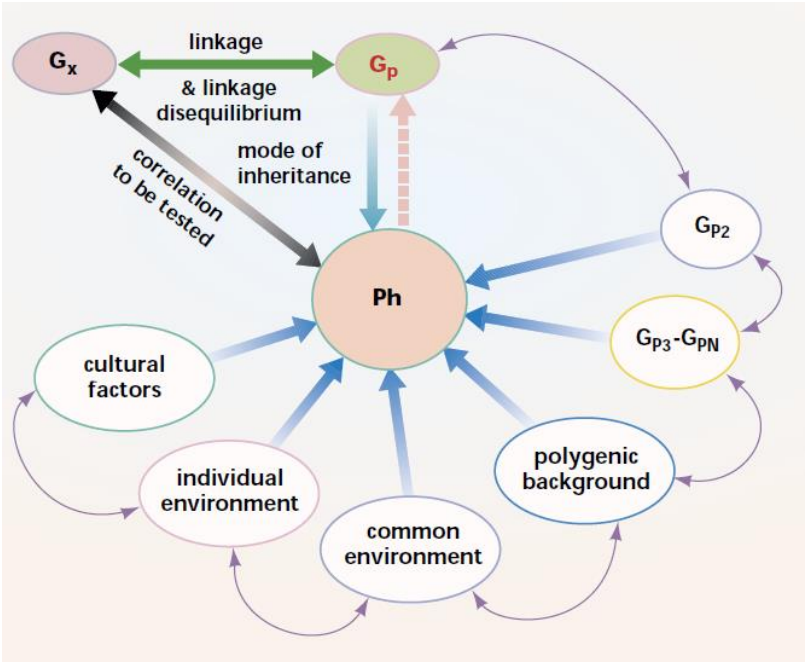
kristel.vansteen@ulg.ac.be

Lecture 5: Genome-wide association “I”nteraction (GWAI) studies

1. The origin of “interactions”
2. Travelling the world of interactions
3. How to best build our working space
4. Components of epistasis analysis
5. Model-Based Multifactor Dimensionality Reduction
6. GWAIs in practice

The origin of interactions

The complexity of complex diseases

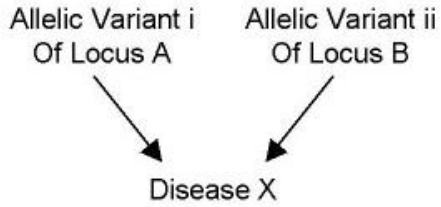
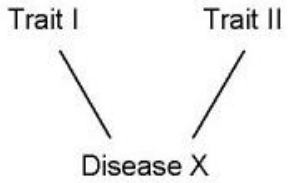
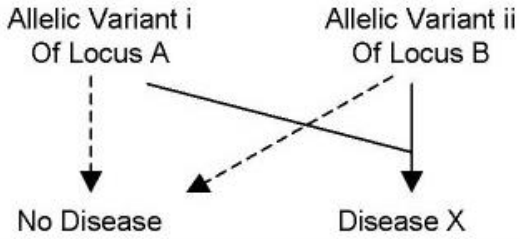


(Weiss and Terwilliger 2000)

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)

Factors complicating analysis of complex genetic disease

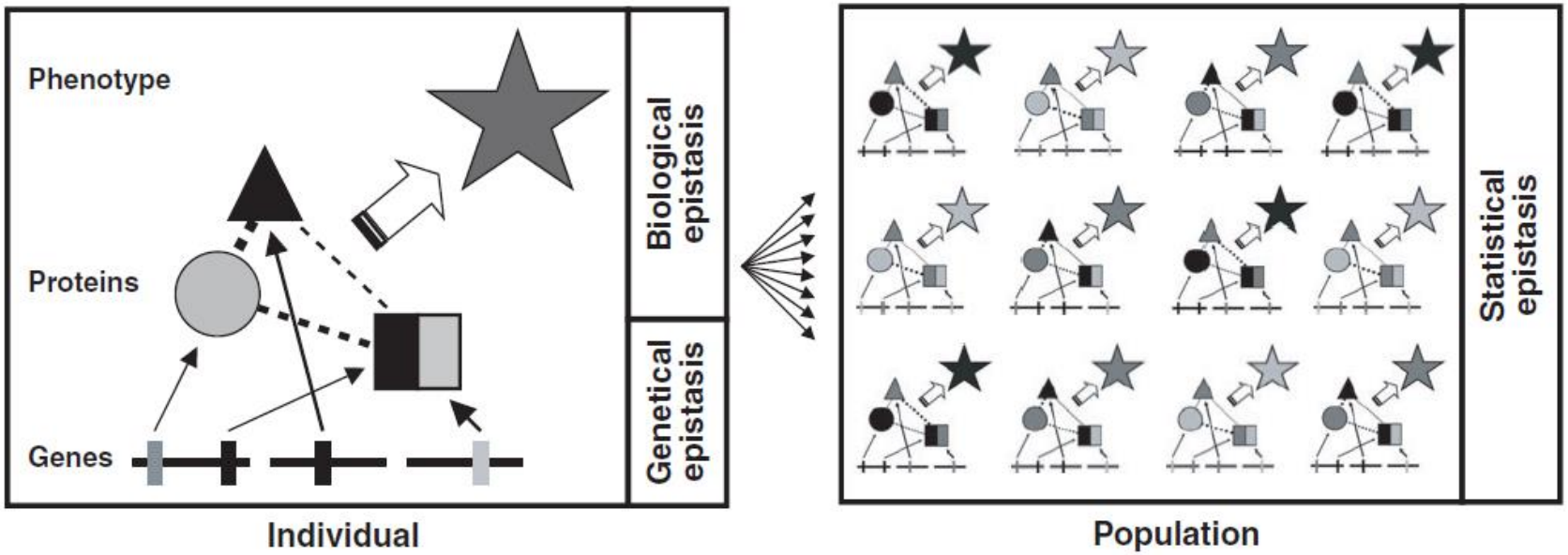
	Locus Heterogeneity	Trait Heterogeneity	Gene-Gene Interaction
Definition	when two or more DNA variations in distinct genetic loci are independently associated with the same trait	when a trait, or disease, has been defined with insufficient specificity such that it is actually two or more distinct underlying 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			
Example	**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)

Factors complicating analysis of complex genetic disease

Gene-gene interactions

… when two or more DNA variations 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)

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.

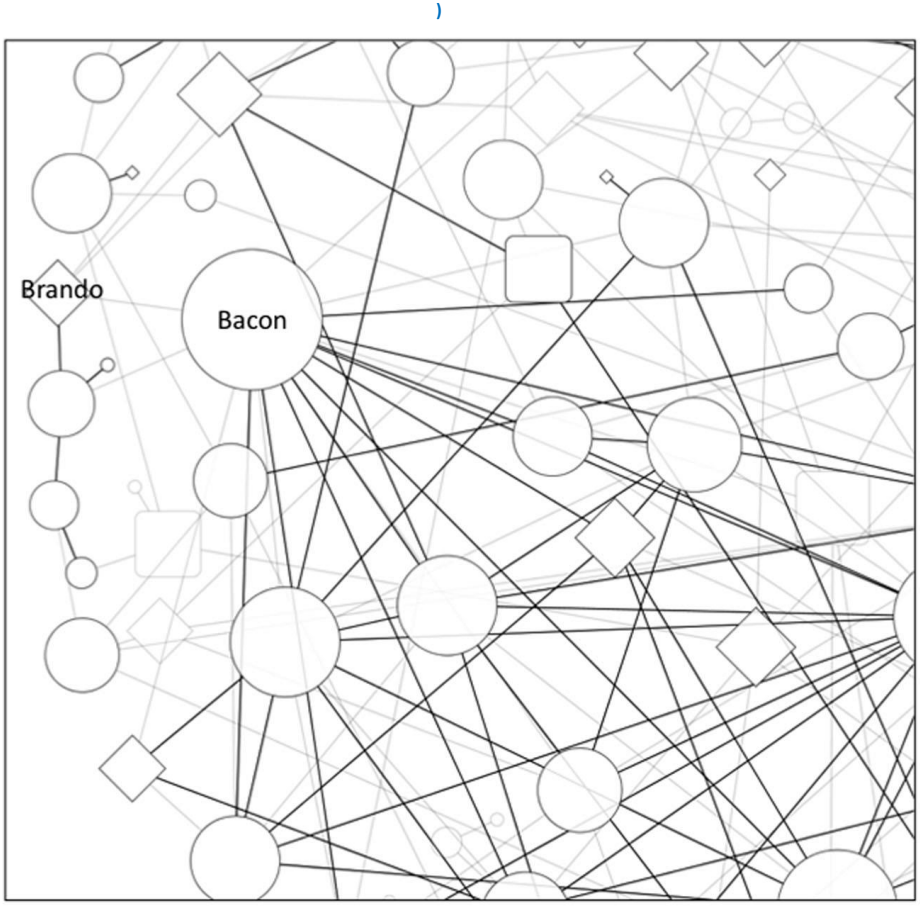
Great expectations

- 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).

Great expectations - empowering personal genomics

- Considering the epic complexity of the transcriptions process, the genetics of gene expression seems just as likely to harbor epistasis as biological pathways.
- When examining HapMap genotypes and gene expression levels from corresponding cell lines to look for cis-epistasis, over 75 genes pop up where SNP pairs in the gene's regulatory region can interact to influence the gene's expression.
- What is perhaps most interesting is that there are often large distances between the two interacting SNPs (with minimal LD between them), meaning that most haplotype and sliding window approaches would miss these effects. (Turner and Bush 2011)

Complementing insights from GWA studies



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)

Epistasis and phantom heritability

(Maher 2008)

Epistasis and phantom heritability

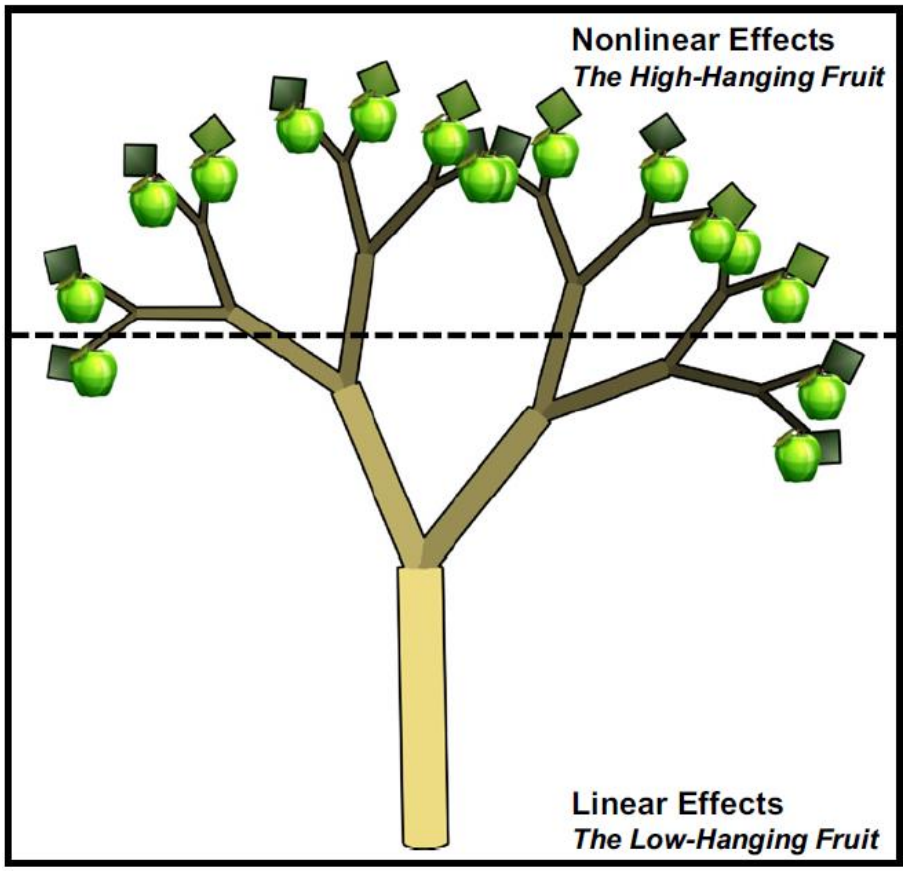
- Human genetics has been haunted by the mystery of “missing heritability” of common traits.
- Although studies have discovered >1,200 variants associated with common diseases and traits, these variants typically appear to explain only a minority of the heritability.
- The proportion of heritability 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.
- The prevailing view has been that the explanation for missing heritability lies in the numerator – variants still to identify

Epistasis and phantom heritability

- Overestimation of the total heritability can create “phantom heritability.”
  - estimates of total heritability implicitly assume the trait involves no genetic interactions (epistasis) among loci
  - this assumption is not justified
  - under such models, the total heritability may be much smaller and thus the proportion of heritability explained much larger.
- For example, 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)

Traveling the world of interactions


You can
You have to



- Few SNPs with moderate to large independent and additive main effects

- Most SNPs of interest will only be found by embracing the complexity of the genotype-to-phenotype mapping relationship that is likely to be characterized by nonlinear gene-gene interactions, gene-environment interaction and locus heterogeneity.

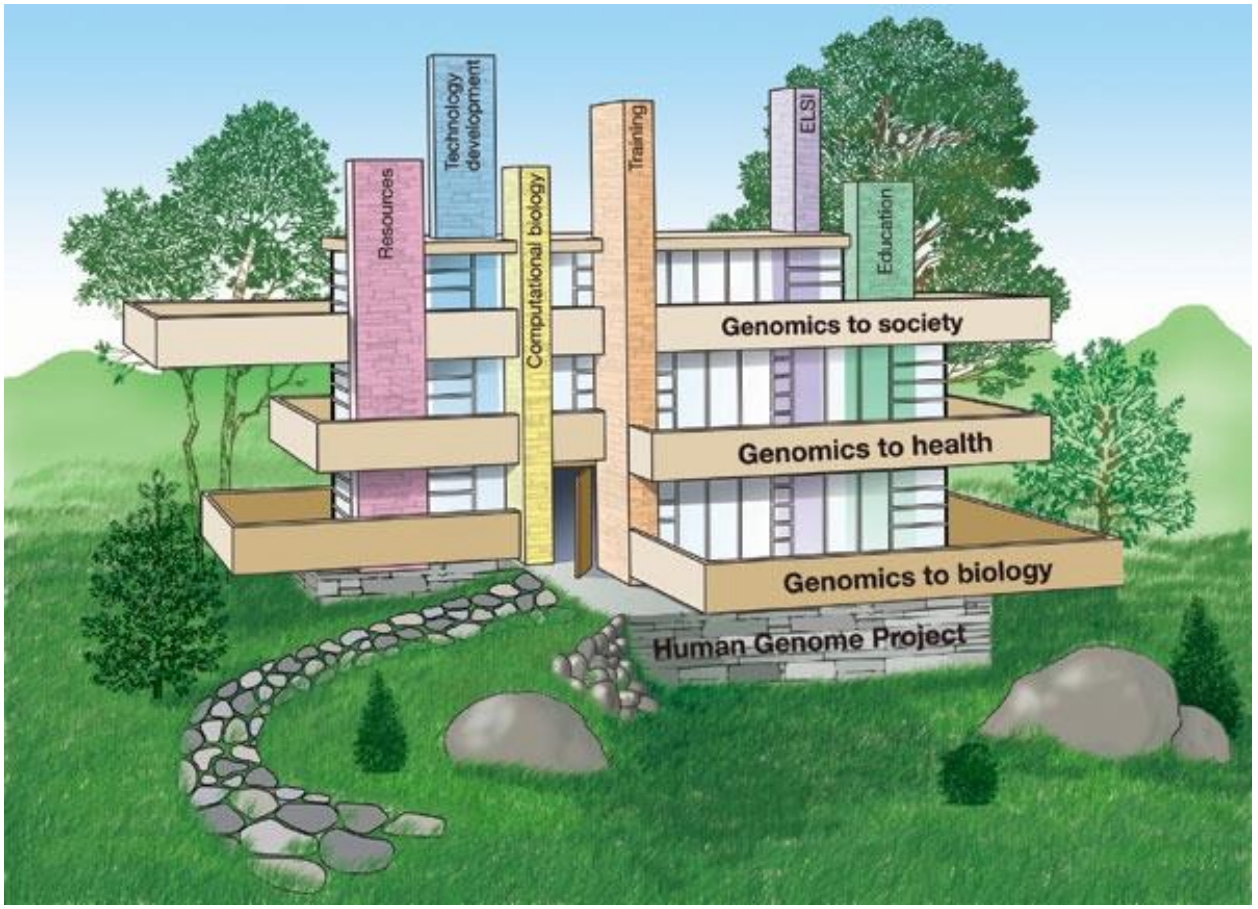
(Moore and Williams 2009)

From GWA to GWAI studies …

- 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 best build our working space

Creating an atmosphere of “interdisciplinarity”

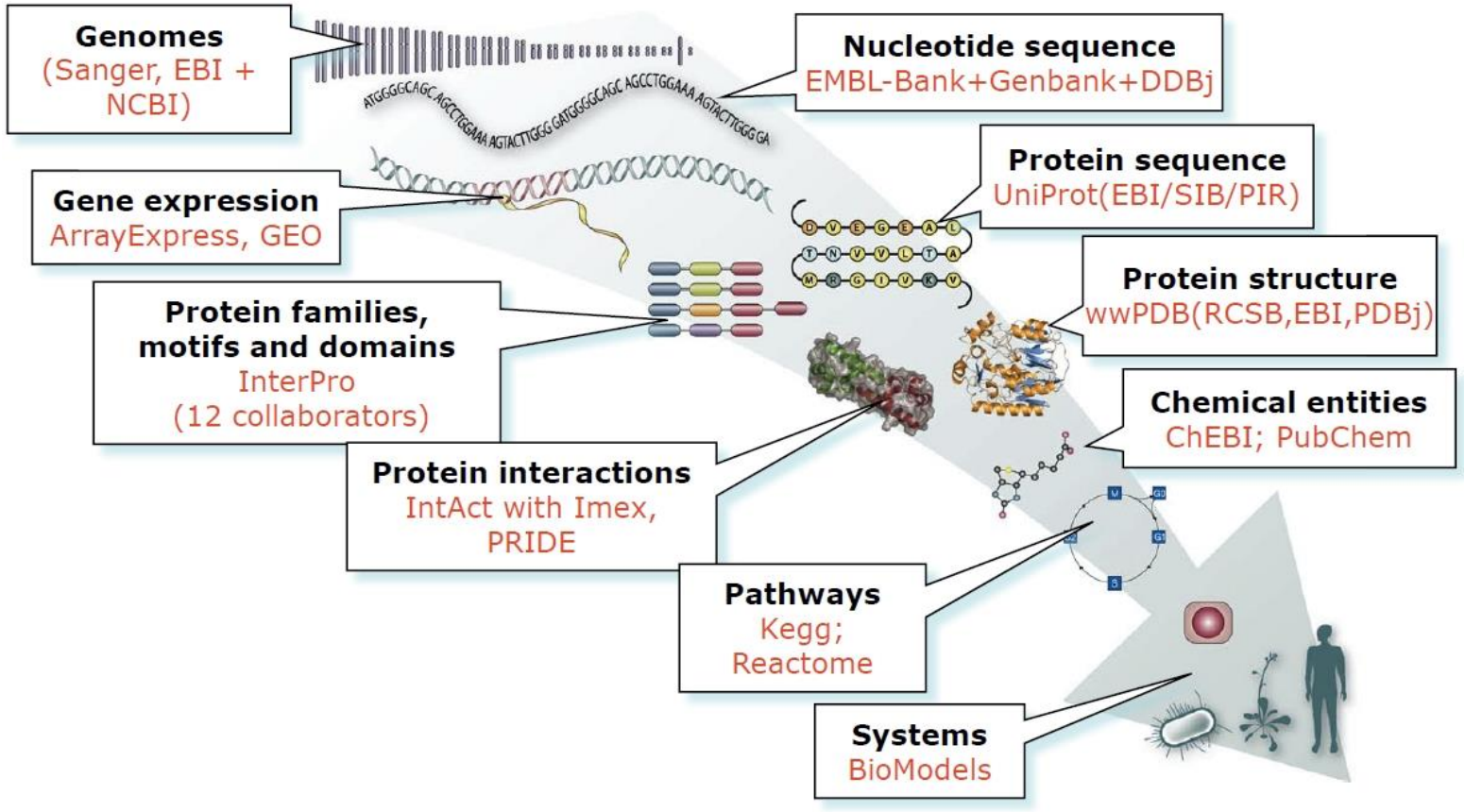


(http://www.genome.gov: the future of human genomics) + harmonization of biobanks

Creating an atmosphere of “integration”

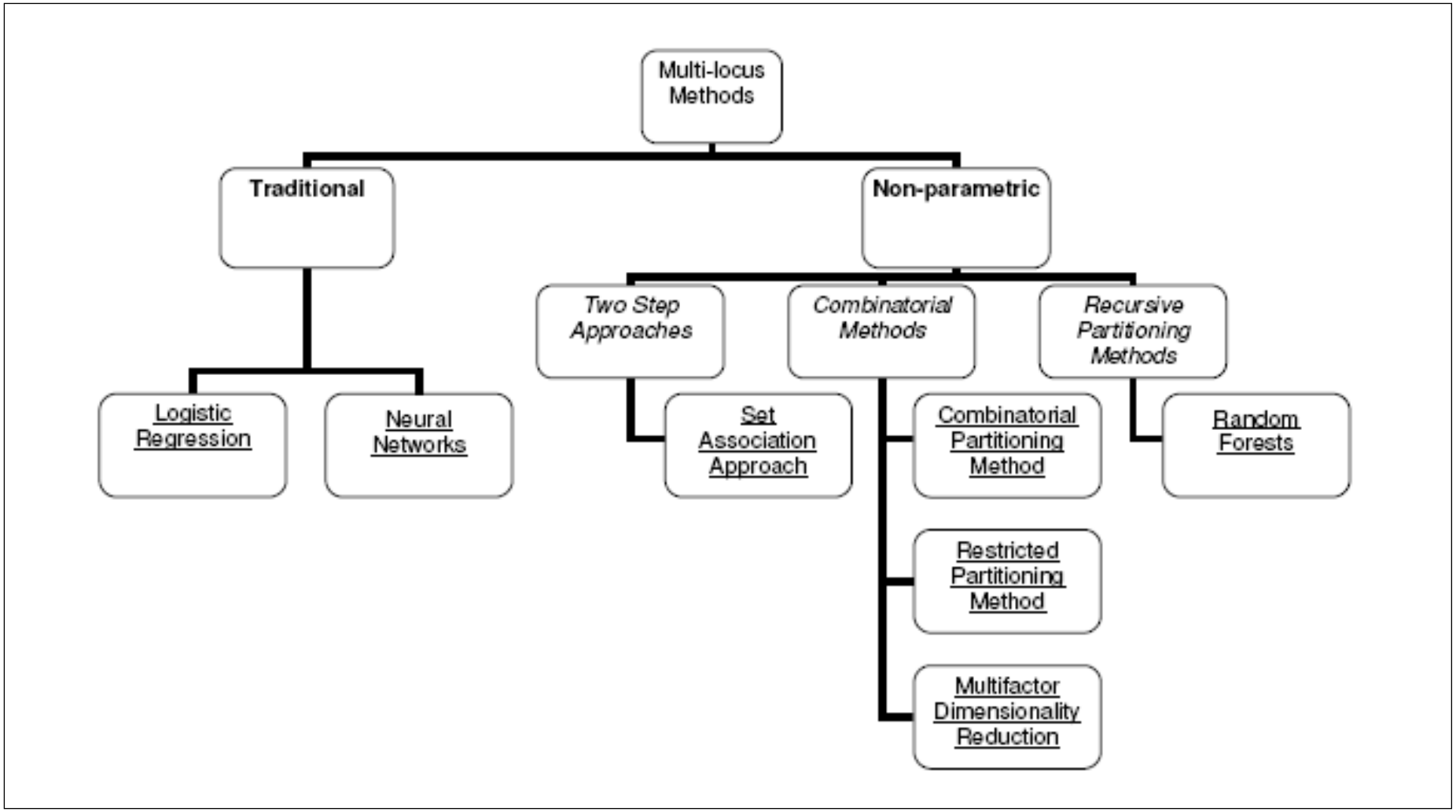
with HTP omics data

(J Thornton, EBI)



Extending the toolbox

(Heidema et al. 2006)

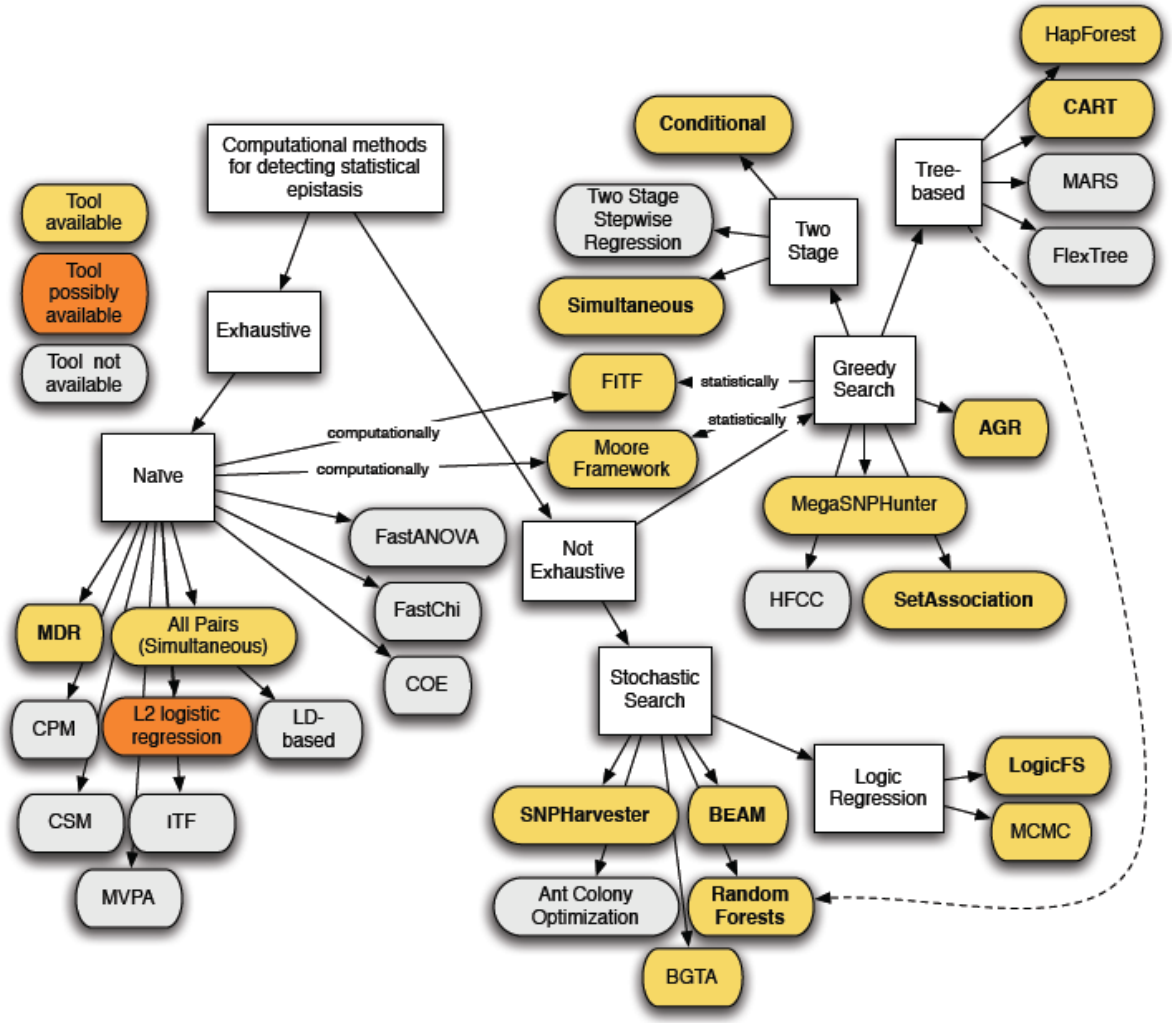


"Although there is growing appreciation that attempting to map genetic interactions in humans may be a fruitful endeavor, there is no consensus as to the best strategy for their detection, particularly in the case of genome-wide association where the number of potential comparisons is enormous"

(Evans et al. 2006)

Extending the toolbox

(Kilpatrick 2009)



Extending the toolbox

- Why?
  - LD between markers
  - Long-distance between-marker associations
  - Missing data handling
  - Multi-stage designs: marker selection and subsequent testing
  - Multiple testing handling
  - Population stratification and admixture
  - Meta-analysis
  - …

Extending the toolbox

- Comes with a caveat: need for thorough comparison studies using reference data sets!
- Several criteria exist to classify epistasis detection methods:
  - Exploratory versus non-exploratory
  - Testing versus Modeling
  - Direct versus Indirect testing
  - Parametric versus non-parametric
  - Exhaustive versus non-exhaustive search algorithms
  - ... (Van Steen et al 2011)

The “observed” occurrences of epistasis – humans

- Phillips et al (2008):
  - There are several 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).

The “observed” occurrences of epistasis – humans

- More recent examples include:
  - Alzheimer’s disease (Combarros et al 2009),
  - psoriasis (WTCCC2 2010),
  - breast cancer (Ashworth et al. 2011),
  - ankylosing spondylitis (WTCCC 2011),
  - total IgE (Choi et al. 2012)
  - High-Density Lipoprotein Cholesterol Levels (Ma et al. 2012)
- So far, only for some of the reported findings additional support could be provided by functional analysis or could be “replicated” (see also later)

Taking it a few steps back … What's in a name?

- 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 …

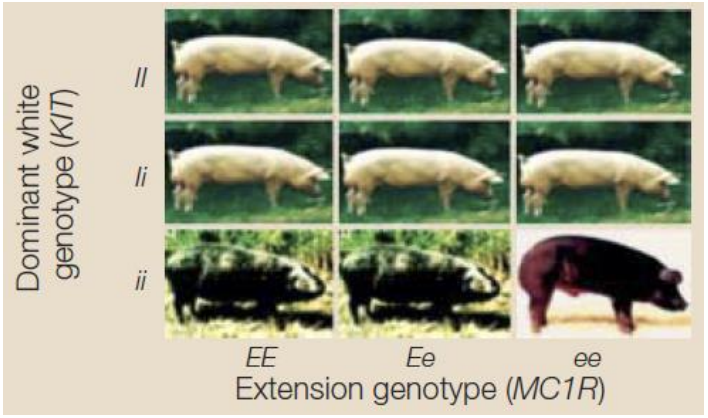
… Studying genetic interactions can reveal gene function, the nature of the mutations, functional redundancy, and protein interactions. Because protein complexes are responsible for most biological functions, genetic interactions are a powerful tool …

Taking it a few steps back … What's in a name?

- Our ability to detect epistasis depends on what we mean by epistasis

"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).



(Carlborg and Haley 2004)

Compositional epistasis

- 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)

Taking it a few steps back … What's in a name?

"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).

- It seems that the interpretation of GWAIs is hampered by undetected false positives

Components of an Epistasis Analysis

Any epistasis analysis is characterized by at least 2 of the following components

- Variable selection
- Modeling / testing
- Significance assessment
- Interpretation

Variable Selection

Why selecting variables?

Introduction

- The aim is to make “clever” selections of markers or marker combinations to look at in the association analysis
- This may not only aid in the interpretation of analysis results, but also reduced the burden of multiple testing and the computational burden

Variable selection in interaction effects GWAS

Several strategies can be adopted to select the number of genetic variants to be used for epistasis screening.

- **Strategy I** involves performing an exhaustive search

  
  Address several computational issues and confront a severe multiple testing problem.

- **Strategy II** involves selecting genetic markers based on the statistical significance or strength of their singular main effects (Kooperberg et al 2008).

  
  Address the difficulty in finding gene-gene interactions when the underlying disease model is purely epistatic.

Variable selection in interaction effects GWAS

- **Strategy III** involves prioritizing sets of genetic markers based on feature selection methods.


Address finding your way into the jungle of different possible feature selection methods and algorithms

- **Strategy IV** involves prioritizing sets of genetic markers based on (prior) expert knowledge


Address biasing of findings towards “what is already known”.

Feature selection methods

- In contrast to other dimensionality reduction techniques like those based on projection (e.g., principal components analysis), feature selection techniques do not change the original presentation of the variables
- Hence, feature selection does not only reduce the burden of multiple testing, but also aids in the interpretation of analysis results

Feature selection methods

- Filter techniques assess the relevance of features by looking only at the intrinsic properties of the data. In most cases a feature relevance score is calculated, and low-scoring features are removed.
- Wrapper techniques involve a search procedure in the space of possible feature subsets, and an evaluation of specific subsets of features. The evaluation of a specific subset of features is obtained by training and testing a specific classification model.
- Embedded techniques involve a search in the combined space of feature subsets and hypotheses. Hence, the search for an optimal subset of features is built into the classifier construction.

(Saeys et al 2007)

Feature selection methods

- **Remark**: When screening and testing involve two separate steps, and these steps are not independent, then proper accounting should be made for this dependence, in order to avoid overly optimistic test results

Highlight 1: entropy-based filtering

Raw entropy values

- Entropy is basically defined as a measure of randomness or disorder within a system.
- Let us assume an attribute, *A*. We have observed its probability distribution, $p_A(a)$.
- Shannon's entropy measured in bits is a measure of predictability of an attribute and is defined as:

$$H(A) \stackrel{\text{def}}{=} -\sum_{a \in A} p(a)\, log_2\, p(a)$$

Raw entropy values: interpretation

- We can understand *H(A*) as the amount of uncertainty about *A*, as estimated from its probability distribution
- The higher the entropy *H(A),* the less reliable are our predictions about *A*.
- The lower the entropy values *H(A)* are, the higher the likelihood that the "system" is in a "more stable state".

Low Entropy

..the values (locations of soup) sampled entirely from within the soup bowl

High Entropy

..the values (locations of soup) unpredictable... almost uniformly sampled throughout our dining room

Copyright © 2001, 2003, Andrew W. Moore

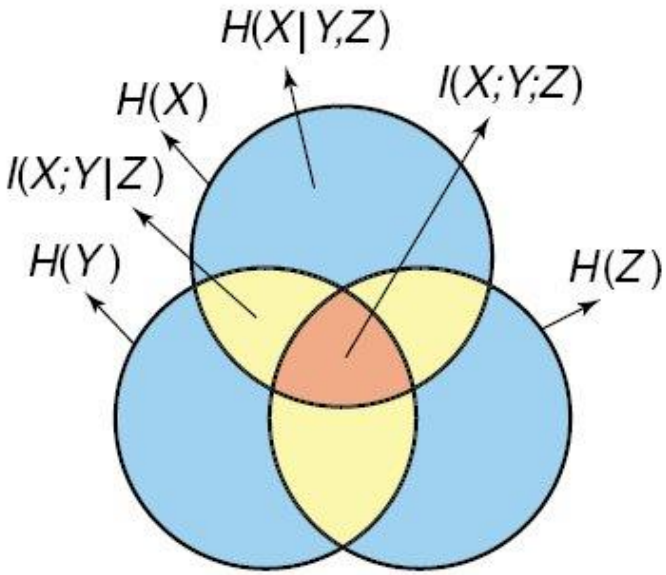
Multivariate mutual information

- For 3 random variables, the mutual information is

$$I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3),$$

  the difference between the simple mutual information and the conditional mutual information

- For higher dimensions, interaction information is defined recursively



Bivariate synergy

$$\mathrm{Syn}(X_1, X_2; X_3) = I(X_1, X_2; X_3) - [I(X_1; X_3) + I(X_2; X_3)]$$

- I.e., the additional contribution provided by the “whole” compared with the sum of the contributions of the “parts”. (Varadan et al 2006)

- The synergy of 2 of the variables with respect to the third is the gain in the mutual information of 2 of the variables, due to knowledge of the third. (Anastassiou 2007)

- Strictly positive synergy is seen as evidence that one has to go beyond a linear decomposition; Strictly negative synergy is seen as evidence for redundancy.

The challenge of detecting epistasis (G x G interactions): Genetic Analysis Workshop 16.

An P, Mukherjee O, Chanda P, Yao L, Engelman CD, Huang CH, Zheng T, Kovac IP, Dubé MP, Liang X, Li J, de Andrade M, Culverhouse R, Malzahn D, Manning AK, Clarke GM, Ju

Genet Epidemiol. 2009;33 Suppl 1:S58-67. doi: 10.1002/gepi.20474

PMID: 19924703 [PubMed - indexed for MEDLINE]

Related citations

Comparison of information-theoretic to statistical methods for gene-gene interactions in the presence of genetic heterogeneity.

Sucheston L, Chanda P, Zhang A, Tritchler D, Ramanathan M.

BMC Genomics. 2010 Sep 3;11:487. doi: 10.1186/1471-2164-11-487.

PMID: 20815886 [PubMed - indexed for MEDLINE] Free PMC Article

Related citations

Information-theoretic gene-gene and gene-environment interaction analysis of quantitative traits.

Chanda P, Sucheston L, Liu S, Zhang A, Ramanathan M.

BMC Genomics. 2009 Nov 4;10:509. doi: 10.1186/1471-2164-10-509.

PMID: 19889230 [PubMed - indexed for MEDLINE] Free PMC Article

Related citations

Information metrics in genetic epidemiology.

Tritchler DL, Sucheston L, Chanda P, Ramanathan M.

Stat Appl Genet Mol Biol. 2011;10:Article 12.

PMID: 21381437 [PubMed - indexed for MEDLINE] Free PMC Article

The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors.

Chanda P, Sucheston L, Zhang A, Ramanathan M.

Eur J Hum Genet. 2009 Oct;17(10):1274-86. doi: 10.1038/ejhg.2009.38. Epub 2009 Mar 18.

PMID: 19293841 [PubMed - indexed for MEDLINE] Free PMC Article

Related citations

AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes.

Chanda P, Sucheston L, Zhang A, Brazeau D, Freudenheim JL, Ambrosone C, Ramanathan M.

Genetics. 2008 Oct;180(2):1191-210. doi: 10.1534/genetics.108.088542. Epub 2008 Sep 9.

PMID: 18780753 [PubMed - indexed for MEDLINE] Free PMC Article

Related citations

Information-theoretic metrics for visualizing gene-environment interactions.

Chanda P, Zhang A, Brazeau D, Sucheston L, Freudenheim JL, Ambrosone C, Ramanathan M.

Am J Hum Genet. 2007 Nov;81(5):939-63. Epub 2007 Oct 3.

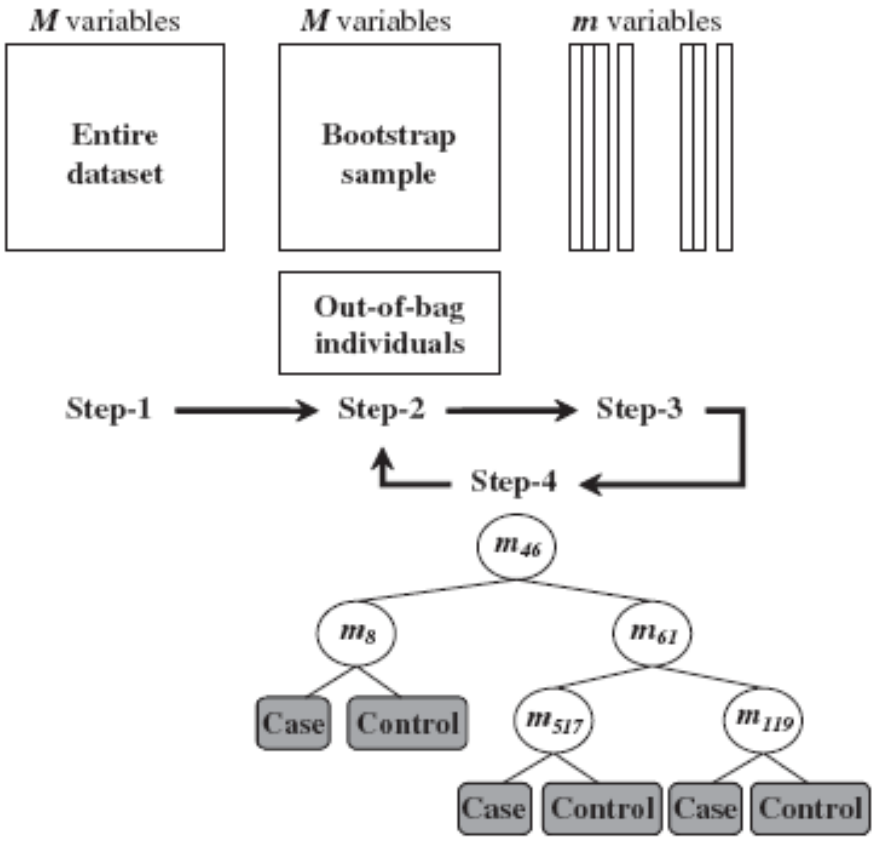
PMID: 17924337 [PubMed - indexed for MEDLINE] Free PMC Article

Related citations

Strategy 2: Data mining as embedding technique

Random Forests (RF)

(Breiman 2001)



(Motsinger-Reif et al 2008)

Modeling / Testing

What do we want to model/test?

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

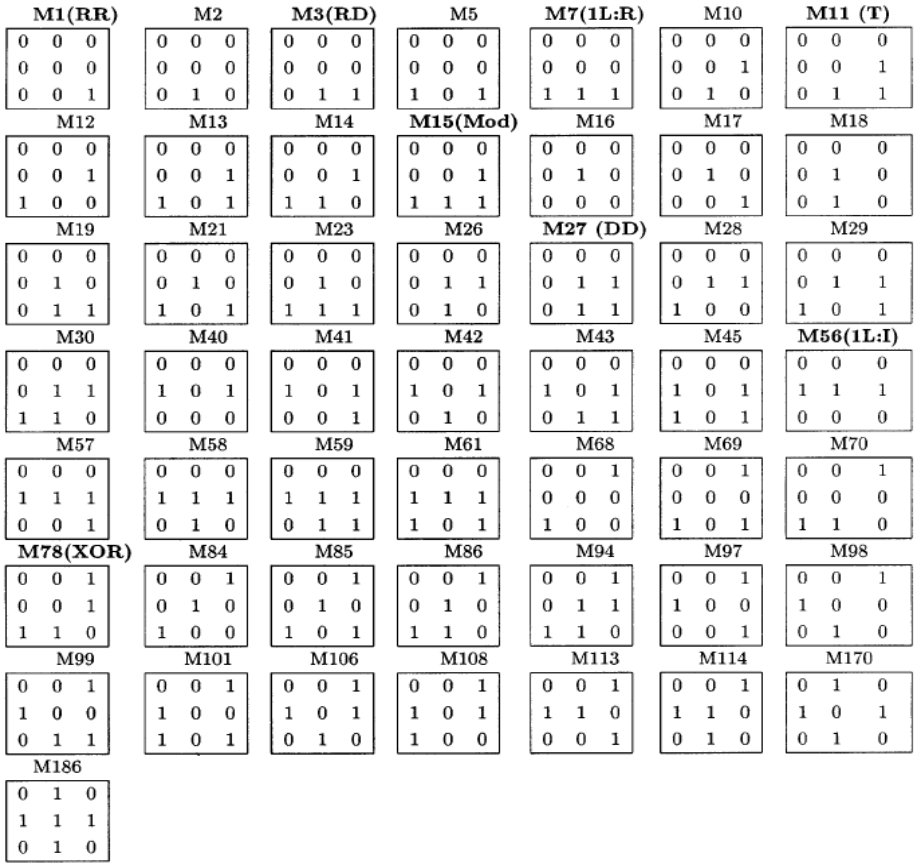
Genotype	bb	bB	BB
aa	0	0	0
aA	0	1	1
AA	0	1	1

(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.

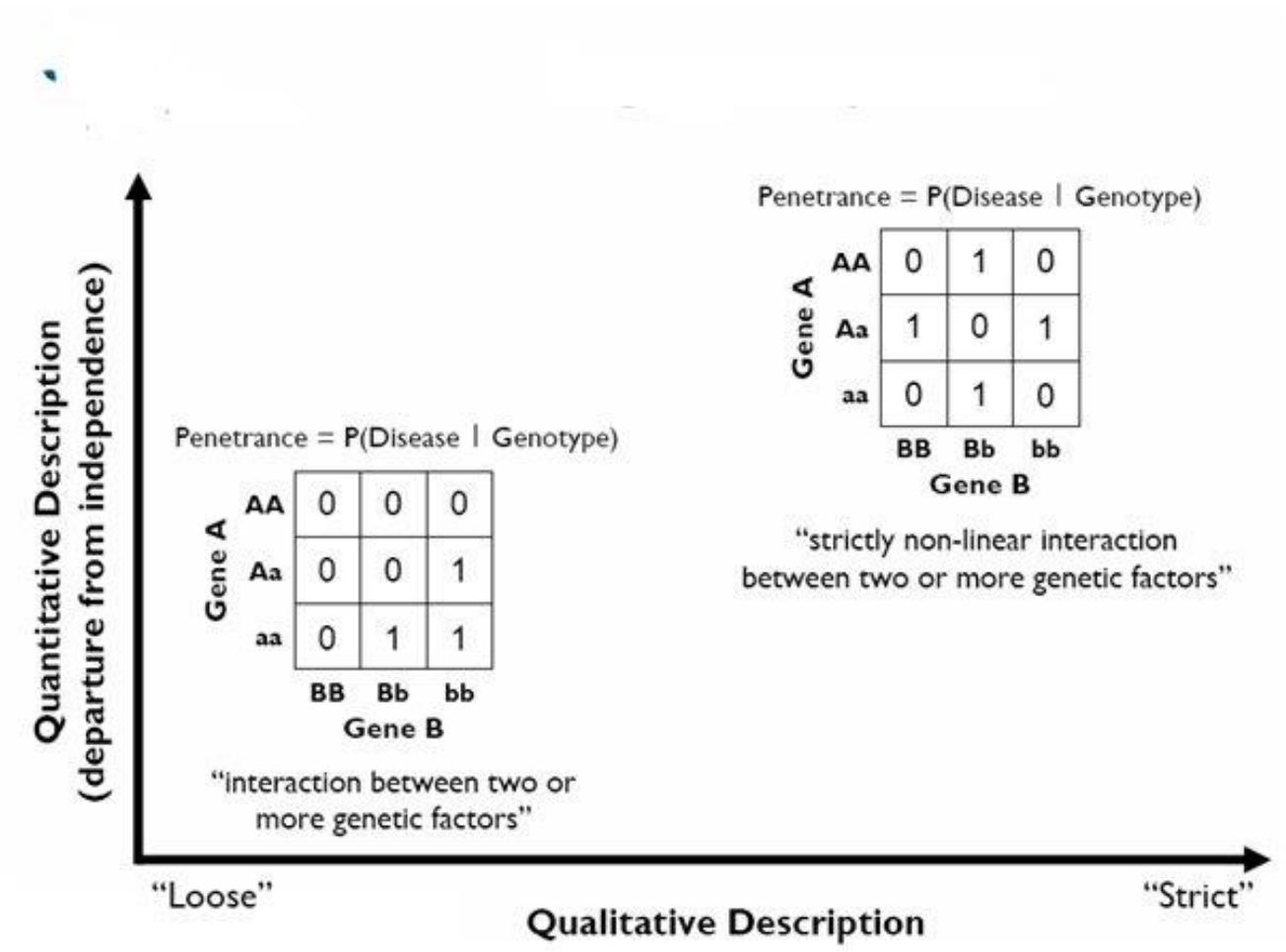
Enumeration of two-locus models

(Li and Reich 2000)



- 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 (‘1L’) – the recessive (R) and the interference (I) model.

Different degrees of epistasis



(slide: Motsinger)

Incomplete penetrance

- 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	$\alpha(1+\theta_1)^2$	$\alpha(1+\theta_1)^2(1+\theta_2)$	$\alpha(1+\theta_1)^2(1+\theta_2)^2$

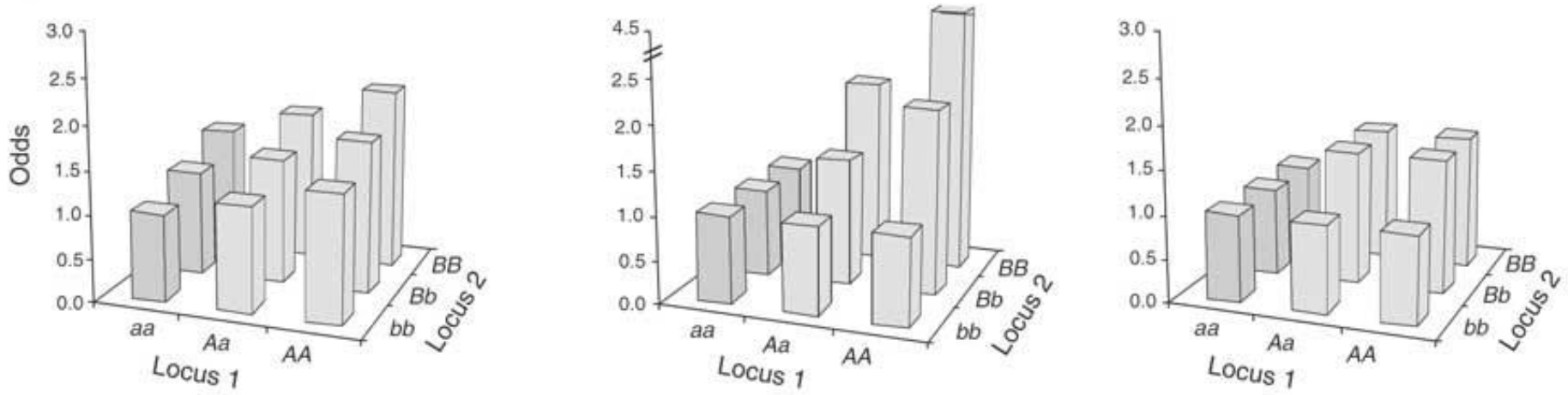
Two-locus interaction multiplicative effects

	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	$\alpha(1+\theta)$	$\alpha(1+\theta)$

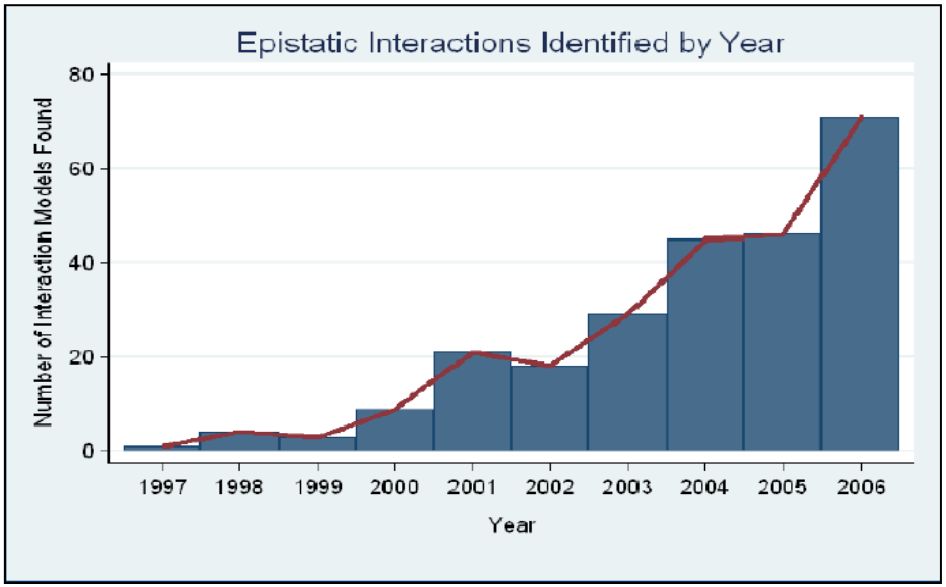
b



(Marchini et al. 2005)

A growing toolbox

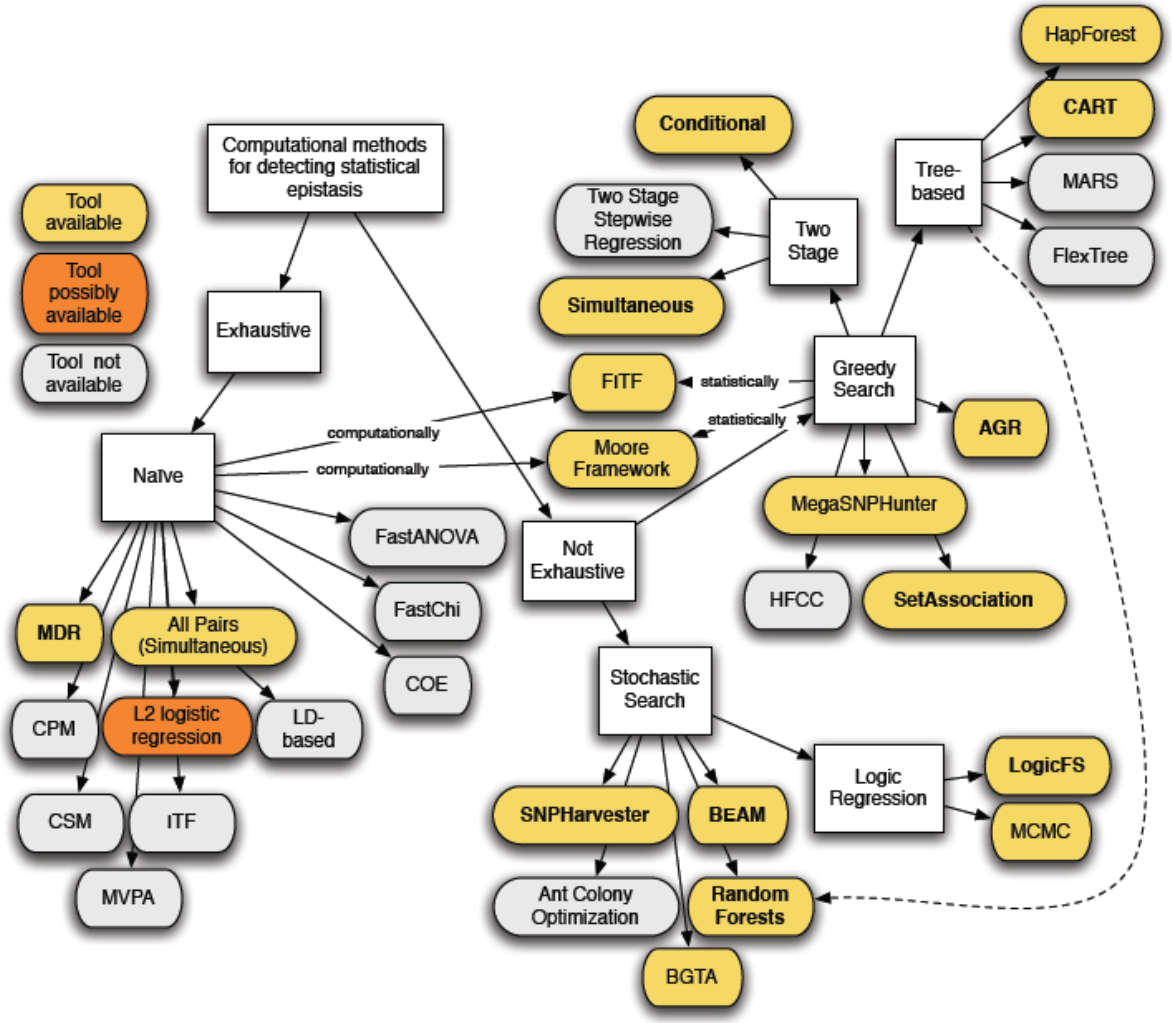
- The number of identified epistasis effects in humans, showing susceptibility to common complex human diseases, follows a steady growth curve (Emily et al 2009, Wu et al 2010), due to the growing number of toolbox methods and approaches.



(Motsinger et al. 2007)

Selecting an epistasis detection method

(Kilpatrick 2009)



Briefings in Bioinformatics Advance Access published March 26, 2011
BRIEFINGS IN BIOINFORMATICS. page 1 of 19 doi:10.1093/bib/bbr012

Travelling the world of gene–gene interactions

Kristel Van Steen

Submitted: 22nd December 2010; Received (in revised form): 13th February 2011

Abstract

Over the last few years, main effect genetic association analysis has proven to be a successful tool to unravel genetic risk components to a variety of complex diseases. In the quest for disease susceptibility factors and the search for the 'missing heritability', supplementary and complementary efforts have been undertaken. These include the inclusion of several genetic inheritance assumptions in model development, the consideration of different sources of information, and the acknowledgement of disease underlying pathways of networks. The search for epistasis or gene–gene interaction effects on traits of interest is marked by an exponential growth, not only in terms of methodological development, but also in terms of practical applications, translation of statistical epistasis to biological epistasis and integration of omics information sources. The current popularity of the field, as well as its attraction to interdisciplinary teams, each making valuable contributions with sometimes rather unique viewpoints, renders it impossible to give an exhaustive review of to-date available approaches for epistasis screening. The purpose of this work is to give a perspective view on a selection of currently active analysis strategies and concerns in the context of epistasis detection, and to provide an eye to the future of gene–gene interaction analysis.

Keywords: *gene–gene interaction; variable selection; controlling false positives; translational medicine*

Hum Genet (2012) 131:1591–1613
DOI 10.1007/s00439-012-1192-0

REVIEW PAPER

Challenges and opportunities in genome-wide environmental interaction (GWEI) studies

Hugues Aschard · Sharon Lutz · Bärbel Maus · Eric J. Duell · Tasha E. Fingerlin · Nilanjan Chatterjee · Peter Kraft · Kristel Van Steen

Received: 1 March 2012 / Accepted: 11 June 2012 / Published online: 4 July 2012
© Springer-Verlag 2012

Abstract The interest in performing gene–environment interaction studies has seen a significant increase with the increase of advanced molecular genetics techniques. Practically, it became possible to investigate the role of environmental factors in disease risk and hence to investigate their role as genetic effect modifiers. The understanding that genetics is important in the uptake and metabolism of toxic substances is an example of how genetic profiles can modify important environmental risk factors to disease. Several rationales exist to set up gene–environment interaction studies and the technical challenges related to these studies—when the number of environmental or genetic risk factors is relatively small—has been described before. In the post-genomic era, it is now possible to study thousands of genes and their interaction with the environment. This brings along a whole range of new challenges and opportunities. Despite a

S. Lutz and B. Maus contributed equally to this work.

Are all methods equal?

- Several criteria have been used to make 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".

One popular method singled out

- North et al (2005) showed that in some instances the inclusion of interaction parameters - within a regression framework - is advantageous but that there is no direct correspondence between the interactive effects in the logistic regression models and the underlying penetrance based models displaying some kind of epistasis effect
- Vermeulen et al (2007) re-confirmed that regression approaches suffer from inflated findings of 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.

One popular method singled out

- 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)

… Targeting 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)

…Targeting 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.

Towards alternative approaches

- What do we know?
  - Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders
  - Small *n* big *p* problems may give rise to curse of dimensionality problems (Bellman 1961); sparse cells issues
  - A lot more knowledge needs to be discovered, naturally giving rise to “data mining” type of strategies
- To keep in mind:
  - Data snooping: statistical bias due to inappr. use of data mining!
  - Biological knowledge integration

The curse of dimensionality in GWAI studies

- 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

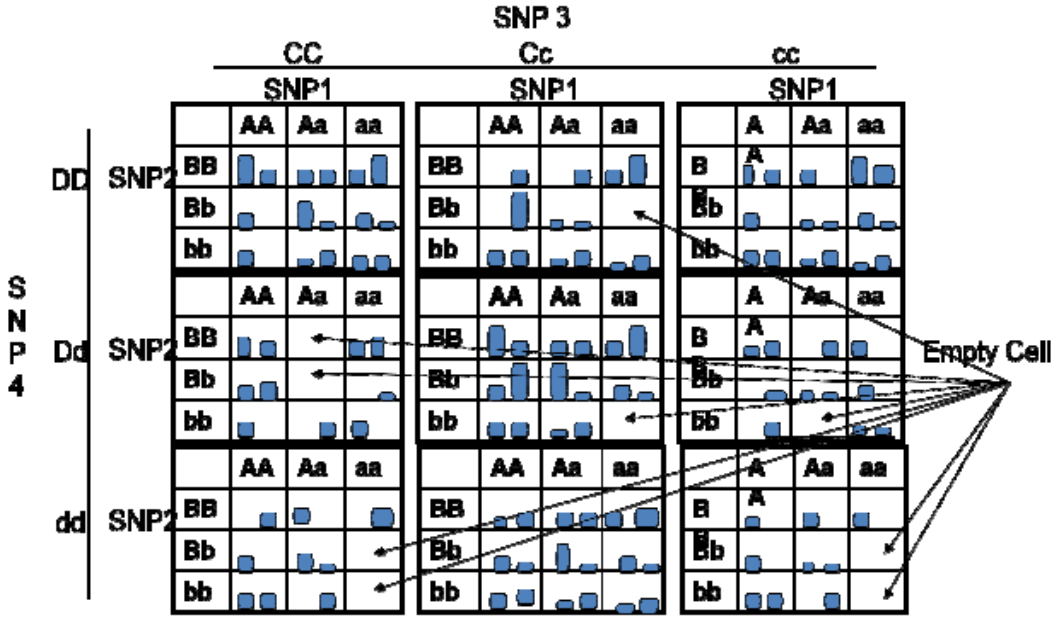
"Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders"

Towards alternative approaches

- What do we know?
  - Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders
  - Small *n* big *p* problems may give rise to curse of dimensionality problems (Bellman 1961); sparse cells issues
  - A lot more knowledge needs to be discovered, naturally giving rise to “data mining” type of strategies
- To keep in mind:
  - Data snooping: statistical bias due to inappr. use of data mining!
  - Biological knowledge integration

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)

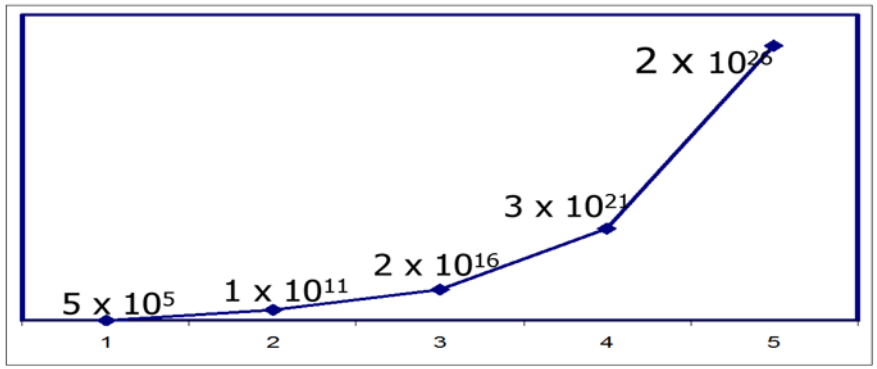
"A revision of LD based imputation strategies for GWAIs is needed"

Towards alternative approaches

- What do we know?
  - Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders
  - Small *n* big *p* problems may give rise to curse of dimensionality problems (Bellman 1961); sparse cells issues
  - A lot more knowledge needs to be discovered, naturally giving rise to “data mining” type of strategies
- To keep in mind:
  - Data snooping: statistical bias due to inappr. use of data mining!
  - Biological knowledge integration

The multiple testing problem ~ significance assessment

- 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"

Significance assessment

Take-home messages

- It is important to verify the validity of the assumptions that underlie each corrective method for multiple testing, in order to select the most optimal corrective method for the data at hand.
- Several methods have been developed to curtail “classical” methods to GWAS settings
- For instance, methods that accommodate correlated hypothesis tests (e.g., due to LD structure between genetic variants):
  - Bonferroni correction using effective sample size derived from principal components (Nyholt et al 2004, Moskvina et al 2008)
  - Haplotype blocking algorithms (Nicodemus et al 2005)
  - Hidden Markov Model-dependent hypothesis testing (Sun and Cai 2009, Wei et al 2009).

Take-home messages (cnt-ed)

- The permutation test is widely considered the gold standard for accurate multiple testing correction, but it is often computationally impractical for these large datasets

- Several variations of permutation-based methods have been worked out, including those based on:
  - deriving an early-evidence stopping rule (Doerge and Churchill 1996)
  - approximating the tail distribution by generalized extreme value distributions (Knijnenburg et al 2009 → in the context of main effects GWAS, Pattin et al 2009 → in the context of epistasis)

Take-home messages (cnt-ed)

- Alternatives to permutation-based testing, even in the presence of millions of correlated markers, claiming to have similar performance as permutation-based approaches:

  - SLIDE (a **S**liding-window Monte-Carlo approach for **L**ocally **I**nter-correlated markers with asymptotic **D**istribution **E**rrors corrected ; Han et al 2009)
  - PACT (**P** values **A**djusted for **C**orrelated **T**ests) (Conneely and Boehnke 2007)

- What about the utility of these methods in the context of GWAI studies?

Towards alternative approaches

- What do we know?
  - Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders
  - Small *n* big *p* problems may give rise to curse of dimensionality problems (Bellman 1961); sparse cells issues
  - A lot more knowledge needs to be discovered, naturally giving rise to “data mining” type of strategies
- To keep in mind:
  - Data snooping: statistical bias due to inappr. use of data mining!
  - Biological knowledge integration

Data Integration: a solution?!

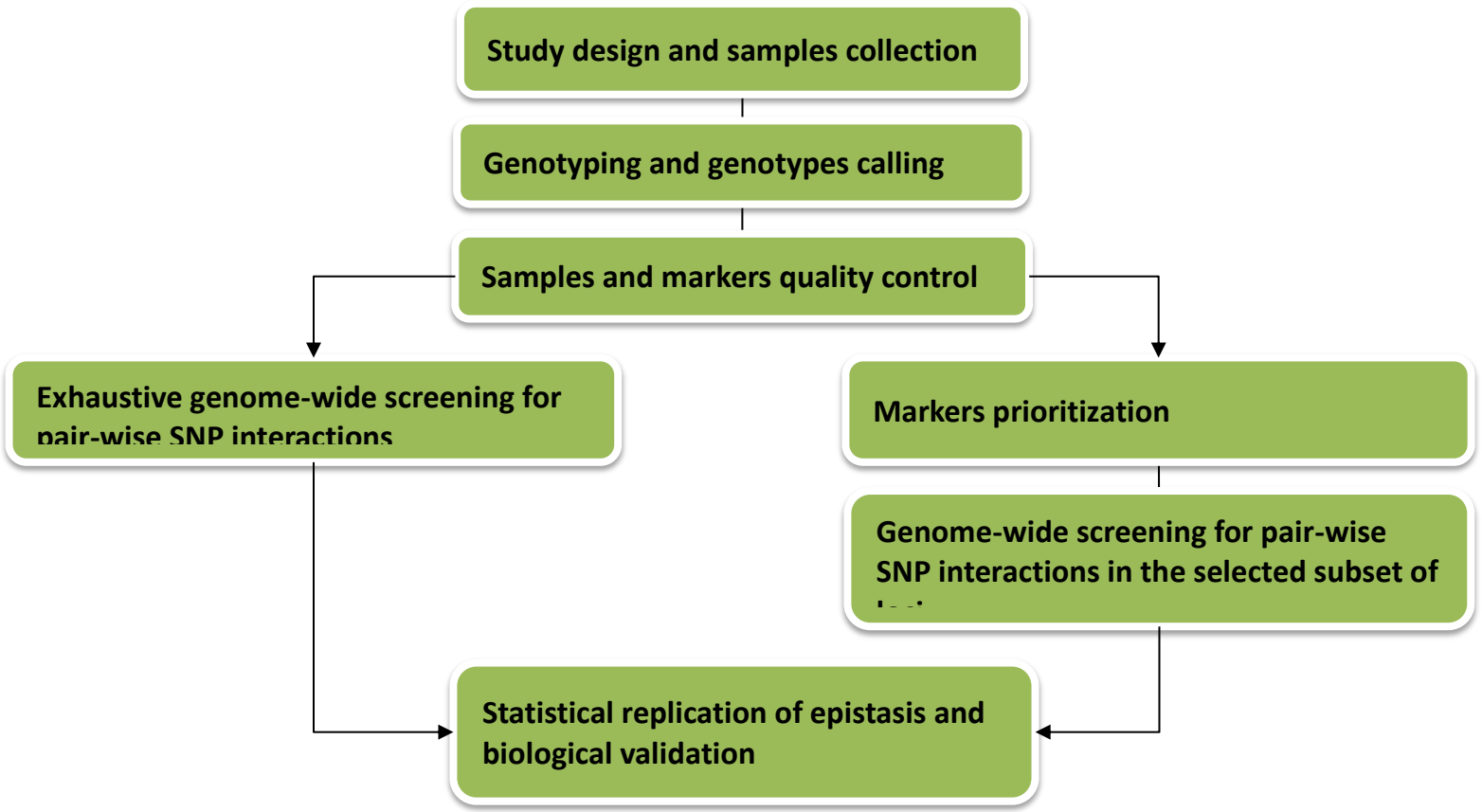
"The genome on its own has turned out to be a relatively poor source of explanation for the differences between cells or between people"

(Bains 2001)

- **Broad definition in the context of GWAI studies** (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: Where during the GWAI process?



(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?

Interpretation

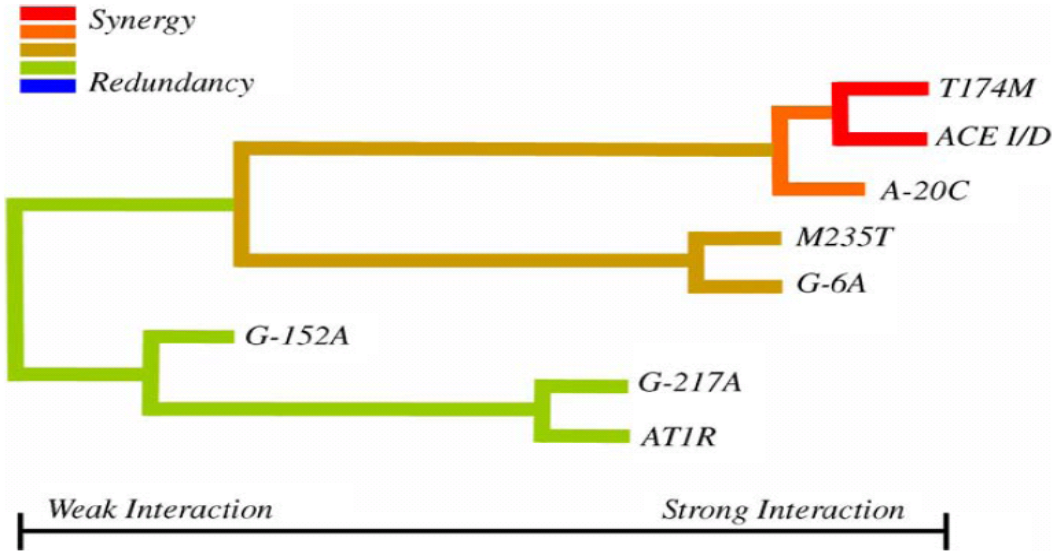
A flexible framework for analysis acknowledging interpretation capability

- The framework contains four steps to detect, characterize, and interpret epistasis
  - Select interesting combinations of SNPs
  - Construct new attributes from those selected
  - Develop and evaluate a classification model using the newly constructed attribute(s)
  - Interpret the final epistasis model using visual methods

(Moore et al 2005)

Example of a visual method: the interaction dendrogram

- Hierarchical clustering is used to build a dendrogram that places strongly interacting attributes close together at the leaves of the tree.

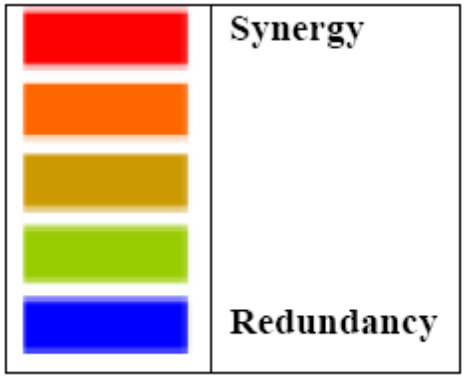


Interaction dendrogram

- The colors range from red representing a high degree of synergy (positive information gain), orange a lesser degree, and gold representing the midway point between synergy and redundancy.

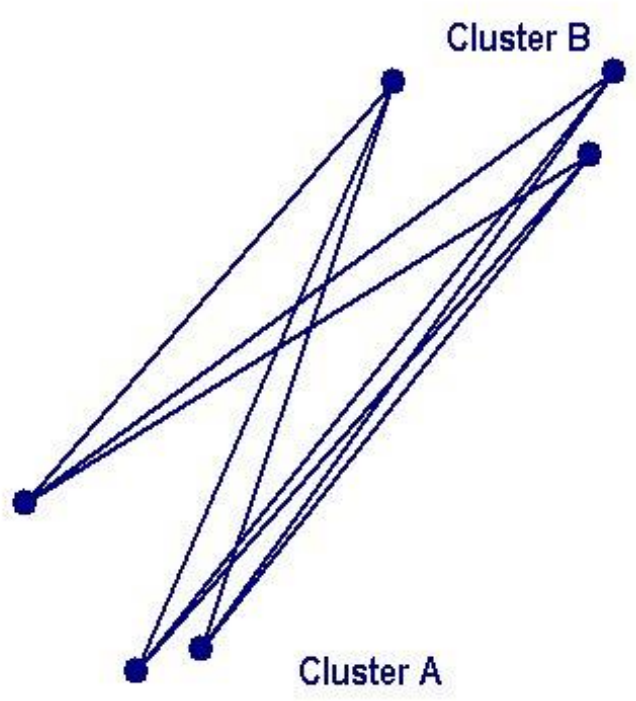
Synergy – The interaction between two attributes provides more information than the sum of the individual attributes.
Redundancy – The interaction between attributes provides redundant information.

- On the redundancy end of the spectrum, the highest degree is represented by the blue color (negative information gain) with a lesser degree represented by green.



Hierarchical clustering with average linkage

- Recall, here the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group



- The distance matrix used by the cluster analysis is constructed by calculating the information gained by constructing two attributes (Moore et al 2006, Jakulin and Bratko 2003, Jakulin et al 2003)

Towards alternative approaches - Plug and play

- The best advice towards success is to adopt different viewpoints to approach the biological problem (see later: example on Alzheimer)
- Plug and play … but not carelessly!

"If you consider the wind-chill factor, adjust for inflation and score on a curve, I only weigh 98 pounds!"

Plug and play … but not carelessly!

How to <u>compare</u> methods… Is this truly a basic question?

- Power
- Type I error / False positives

		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**

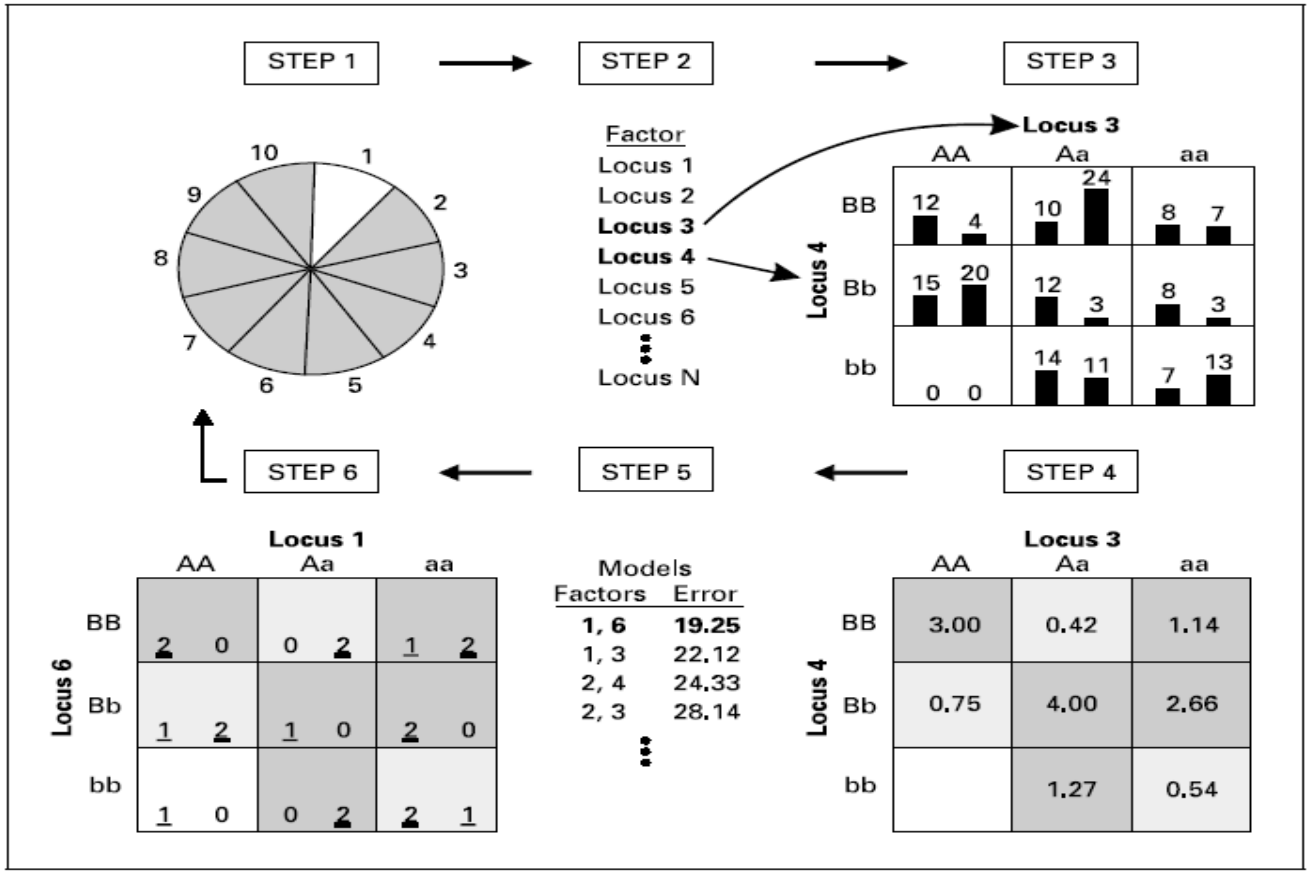
Model-Based Multifactor Dimensionality Reduction

Historical notes about MB-MDR

- Knowledge:
  - Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders
  - Small *n* big *p* problems may give rise to curse of dimensionality problems (Bellman 1961)
  - A lot more knowledge needs to be discovered, naturally giving rise to “data mining” type of strategies
- To keep in mind:
  - Data snooping: statistical bias due to inappr. use of data mining!
  - Biological knowledge integration

Historical notes about MB-MDR

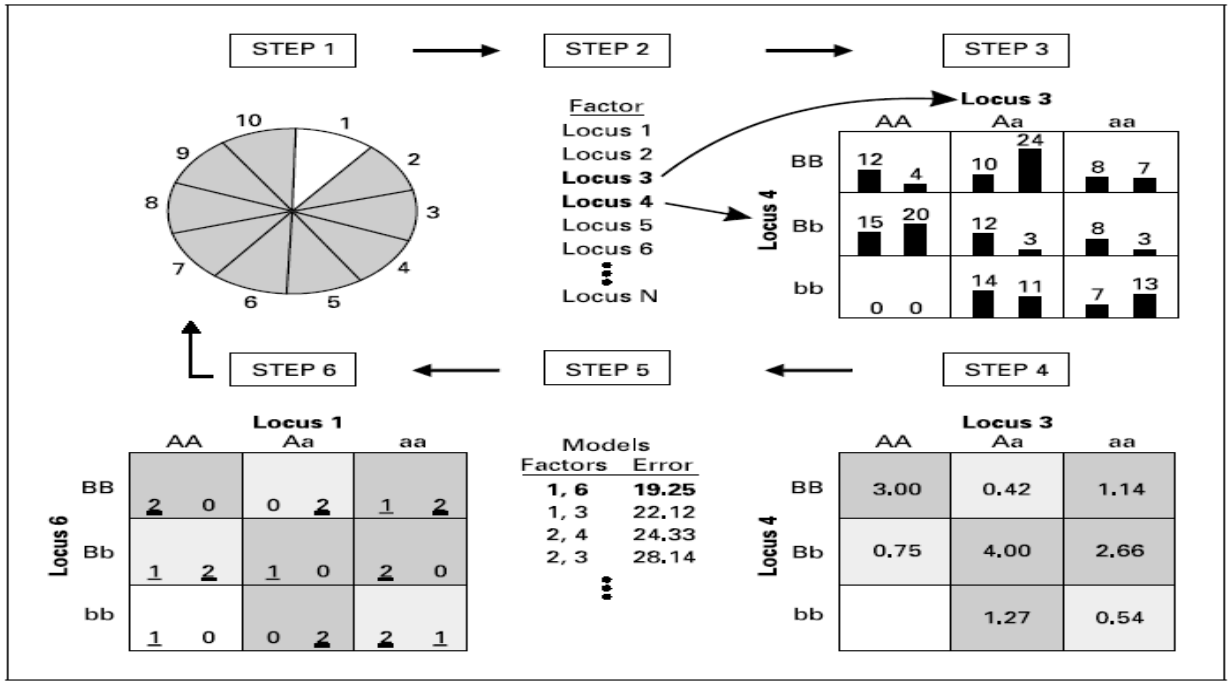
- Start: Multifactor Dimensionality Reduction by MD Ritchie et al (2001)



A note aside

Multifactor Dimensionality Reduction (MDR)

The 6 steps of MDR



Towards MDR Final

- The best model across all 10 training and testing sets is selected on the basis of the criterion:
  - Maximizing the average training accuracy across the 10 cross-validation intervals, within an "interaction order *k*" of interest
    - Order *k*=2: best model with highest average training accuracy
    - Order *k*=3: best model with highest average training accuracy
    - …
  - The best model for each CV interval is applied to the testing proportion of the data and the testing accuracy is derived.
    - The average testing accuracy can be used to pick the best model among 2, 3, ... order "best" models derived before

(Ritchie et al 2001, Ritchie et al 2003, Hahn et al 2003)

Towards MDR Final

- Several improvements:
  - Use accuracy measures that are not biased by the larger class
  - Use a threshold for dimensionality reduction that is driven by the data at hand and naturally reflects the disproportion in cases and controls in the data
  - Use of cross validation consistency (CVC) measure, which records the number of times MDR finds the same model as the data are divided in different segments
    - Useful when average testing accuracies for different “best” higher order models are the same
    - Average testing accuracy estimates are biased when CVC < 10

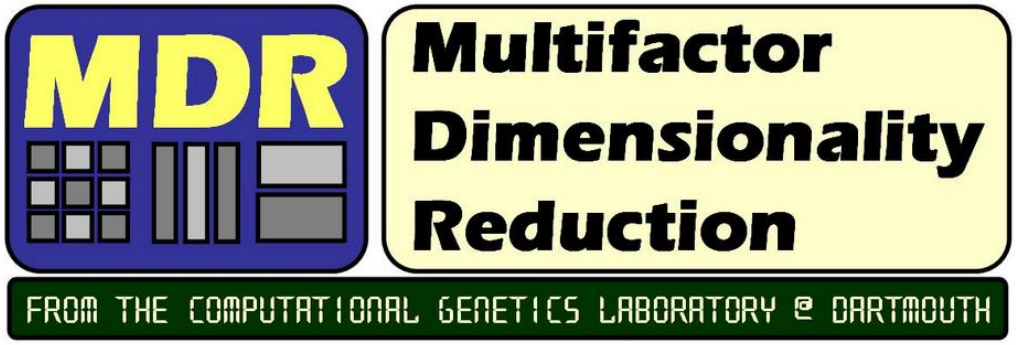
Hypothesis test of best model

- In particular, derive the empirical distribution of the average balanced testing accuracy for the best model:
  - Randomize disease labels
  - Repeat MDR analysis several times (1000?) to obtain the null distribution of cross-validation consistencies and prediction errors

The MDR Software

- The MDR method is described in further detail by Ritchie et al. (2001) and reviewed by Moore and Williams (2002).
- An MDR software package is available from the authors by request, and is described in detail by Hahn et al. (2003).

- Download information and much more can be found at http://www.multifactordimensionalityreduction.org/



Welcome to the new homepage of the Multifactor Dimensionality Reduction (MDR) open-source software package.

MDR is a data mining strategy for detecting and characterizing nonlinear interactions among discrete attributes (e.g. SNPs, smoking, gender, etc.) that are predictive of a discrete outcome (e.g. case-control status). The MDR software combines attribute selection, attribute construction and classification with cross-validation to provide a powerful approach to modeling interactions.

Click here to download the latest version of the free open-source MDR software.

See Epistasis Blog for an MDR 101 tutorial (Nov. 12 - Dec. 4, 2006).

See the Wikipedia entry for a conceptual introduction to MDR.

See our recent paper in the Journal of Theoretical Biology for a scientific review of MDR.

Click here to see a list of papers I know about that have applied MDR to real data.

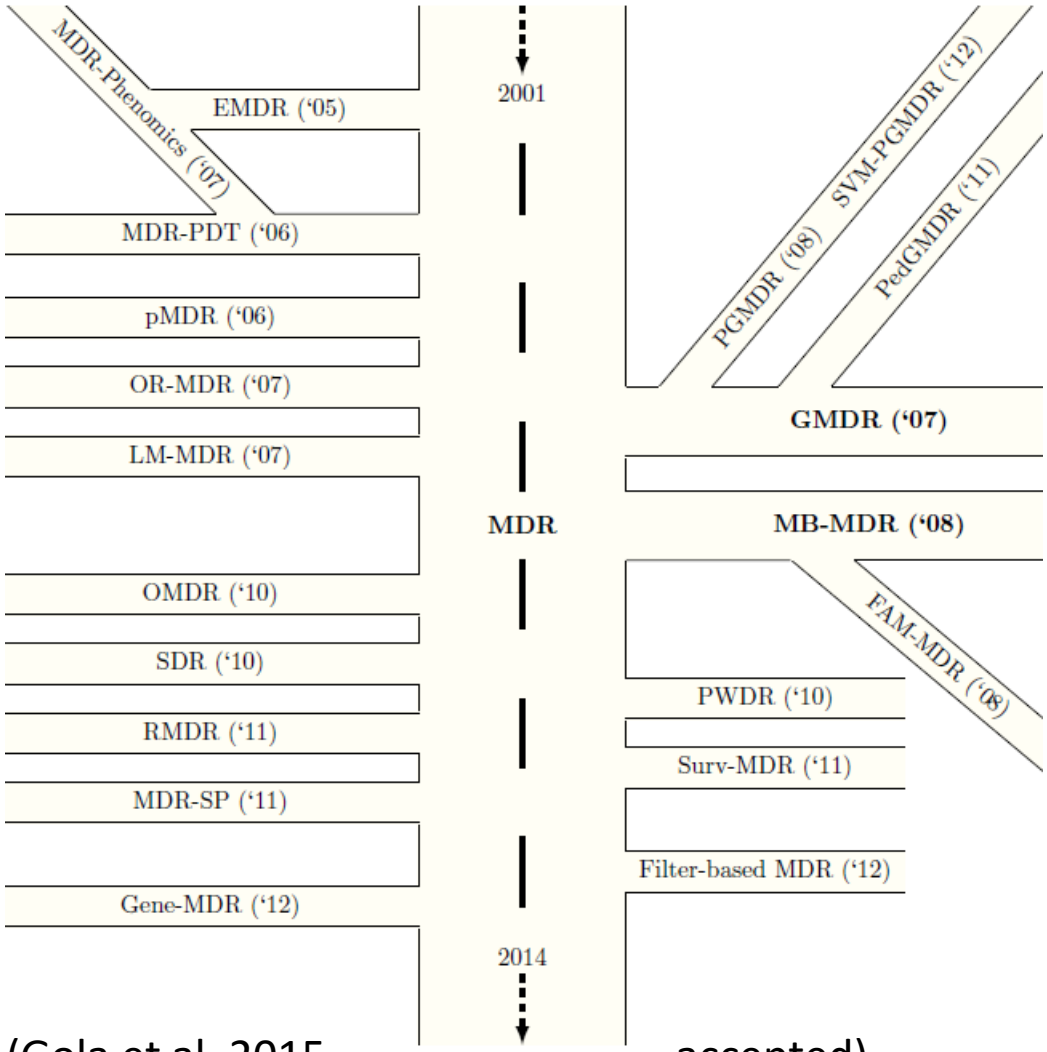
Click here to carry out a PubMed search for MDR papers.

Click here to Google MDR.

Click here to Google Scholar MDR.

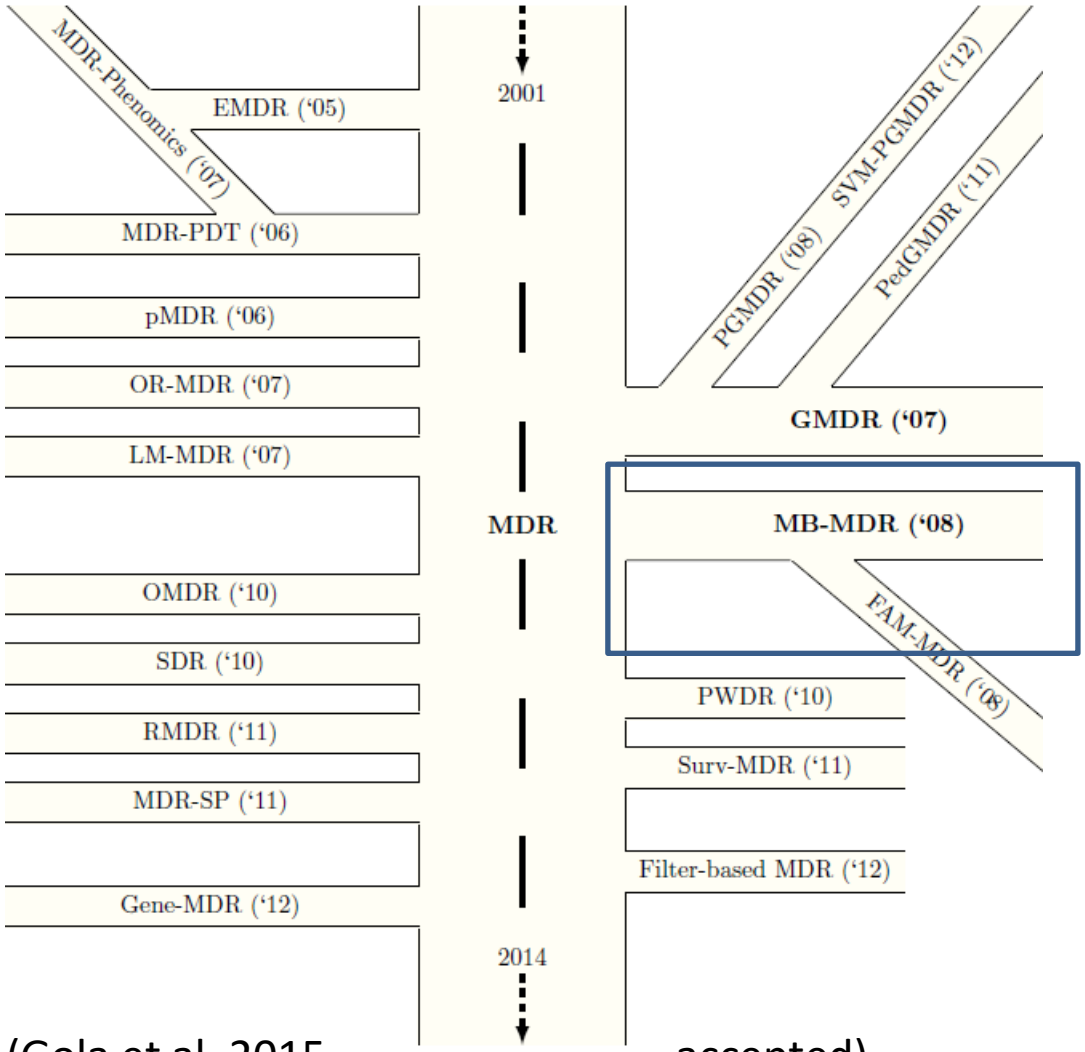
For more information about MDR please see Epistasis.org or Epistasis Blog.

Several MDR roads lead to Rome



(Gola et al. 2015 – accepted)

Several MDR roads lead to Rome

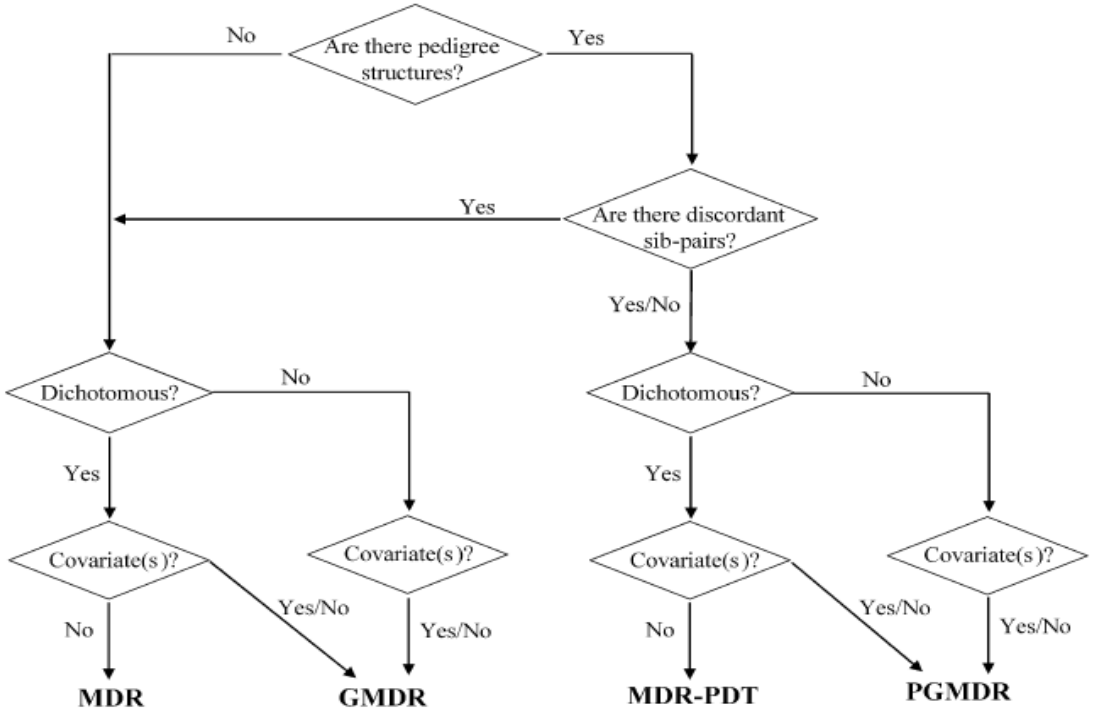


(Gola et al. 2015 – accepted)

Several MDR roads lead to Rome

- Follow-up: Model-Based MDR by Calle et al (2007)

Unlike other MDR-like methods (right), MB-MDR breaks with the tradition of cross-validation to select optimal multilocus models with significant accuracy estimates



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

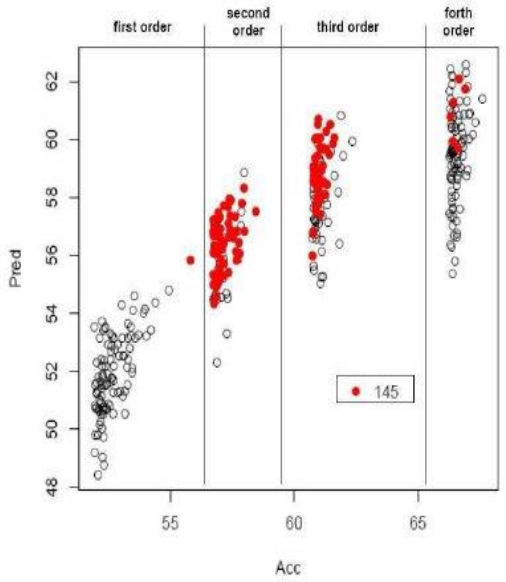
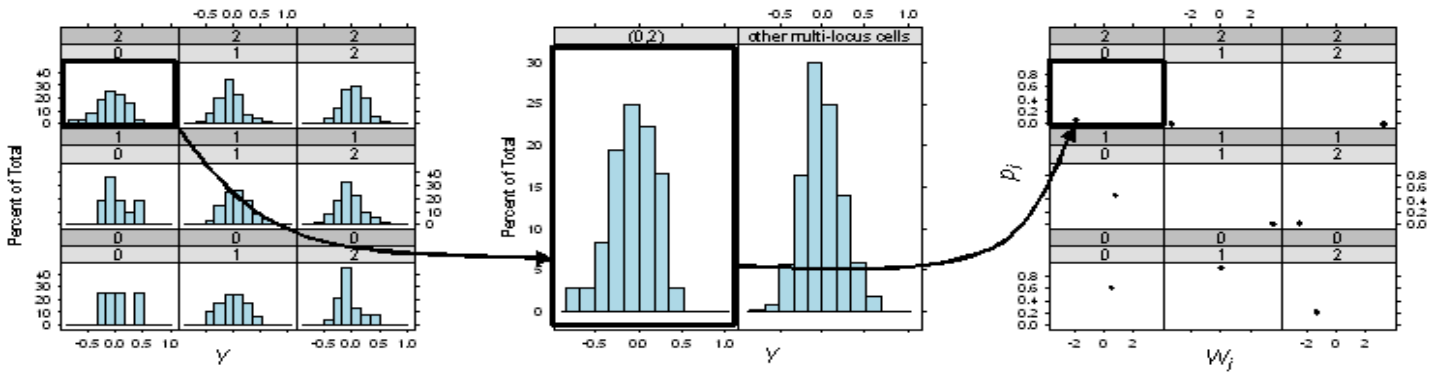


Fig. 1. Average Balanced Training accuracy (Acc) versus Average Balanced Predictive accuracy (Pred) for the 100 models with higher balanced training accuracy for the whole sample. First, second, third and forth order interactions are considered.

Statistical optimization of MB-MDR (binary traits)

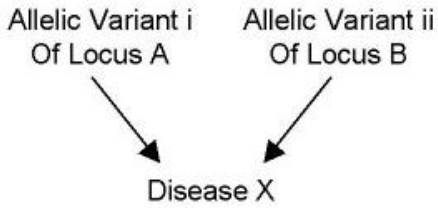
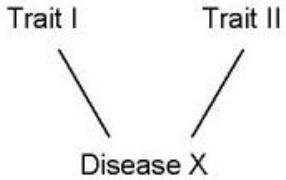
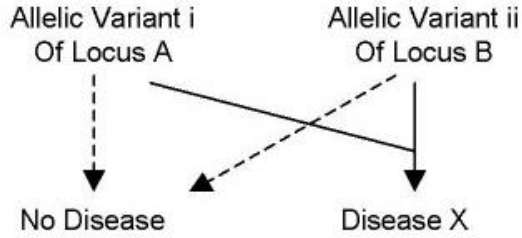
- Model-Based MDR by Cattaert et al (2010) – fine-tuning MB-MDR



- Pooling "alike" (for instance, all low-risk and all high-risk) multilocus genotypes leads to statistic distribution that is different from the theoretical distribution (data snooping)
- Stable score tests, one multilocus p-value and permutation-based strategy (Cattaert et al 2010), as an alternative to Wald tests and relying on MAF dependent reference distributions (Calle et al 2008)

Dealing with heterogeneity

- Complicating factors when analyzing complex disease genetics

	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			
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]

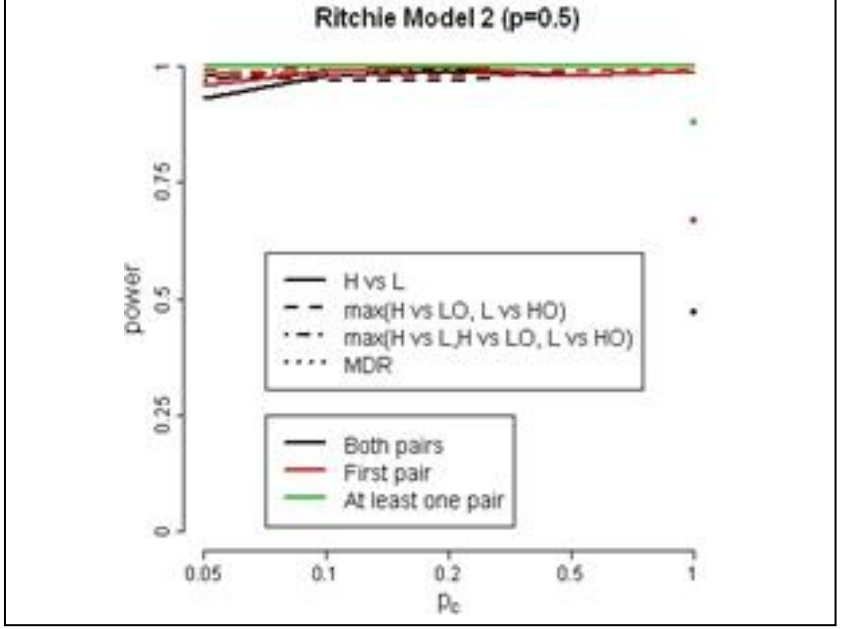
(Thornton-Wells et al. 2006)

Performance in the presence of 2-locus genetic heterogeneity

- Model-Based MDR by Cattaert et al. 2011

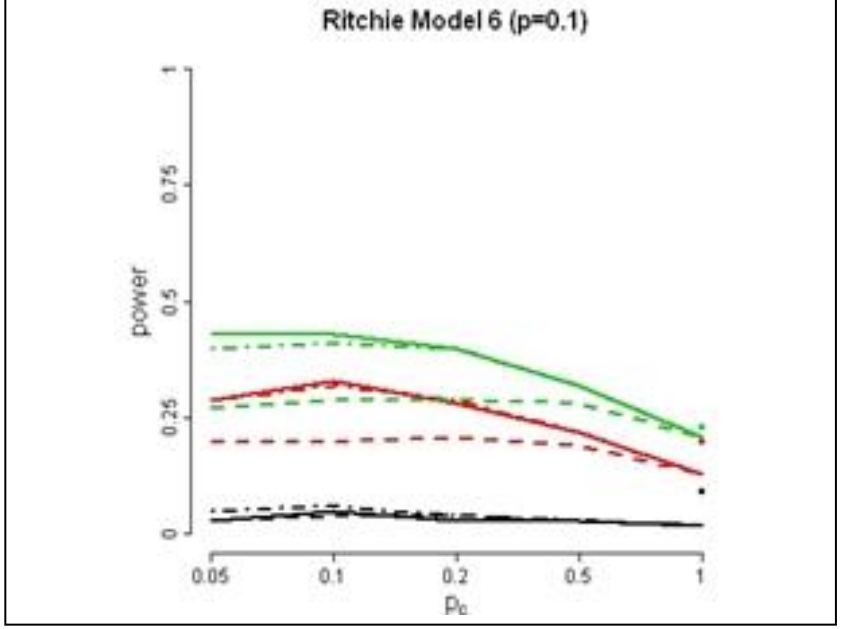
Model 2, $p = 0.5$

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



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



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]**

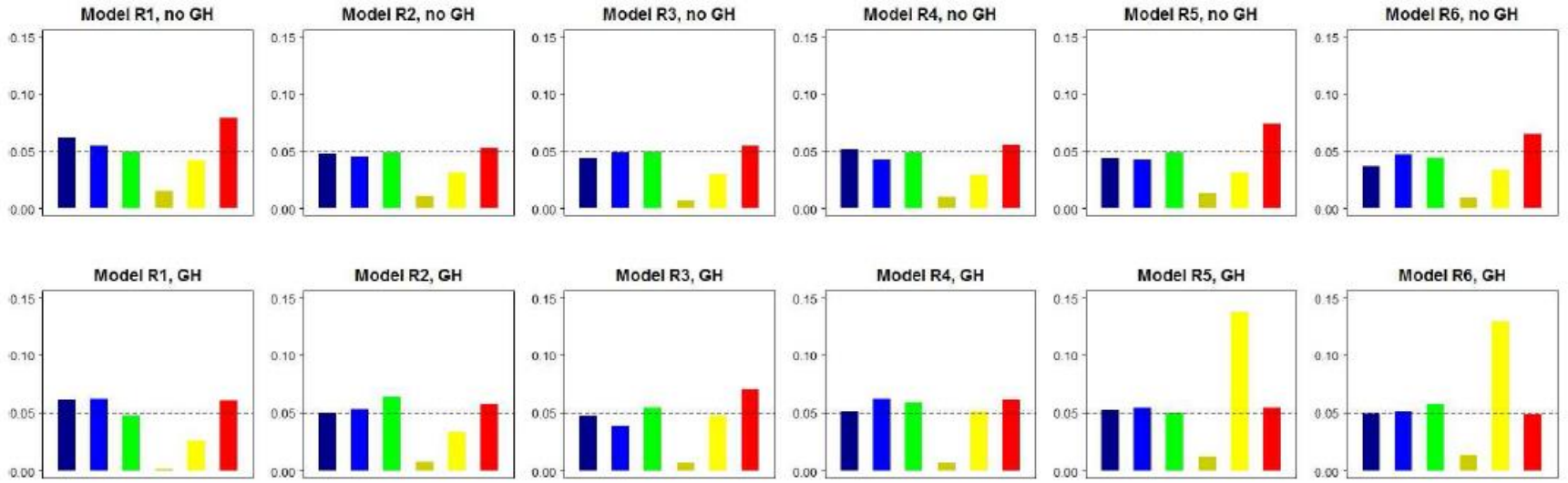
- **Mahachie John** JM, Van Lishout F, Gusareva ES, Van Steen K (2012) A Robustness Study of Parametric and Non-parametric Tests in Model-Based Multifactor Dimensionality Reduction for Epistasis Detection. BioData Min. 2013 Apr 25;6(1):9**[recommendations on QT analysis]**

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

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)

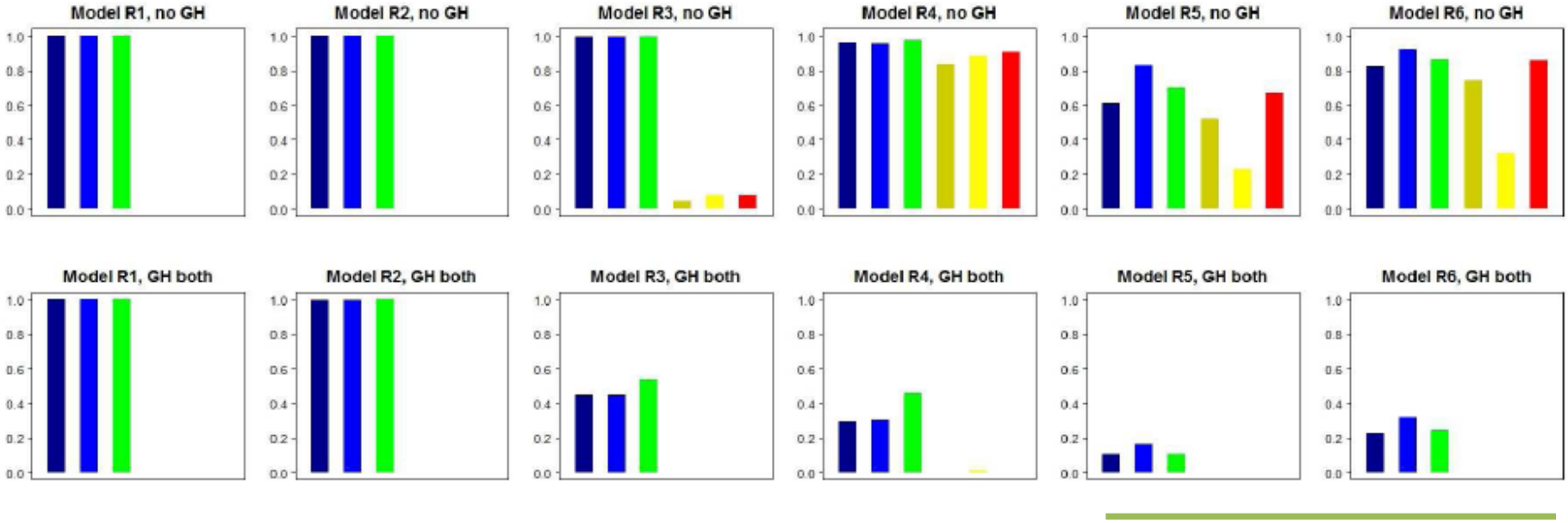
PLINK fast epistasis (light yellow)

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)

Refinements …. Scaling up

- From sequential to parallel workflows and a better usage of "null data"

SNPs	*MBMDR-4.2.2* Binary trait sequential execution	*MBMDR-4.2.2* Binary trait parallel workflow	*MBMDR-4.2.2* Continuous trait sequential execution	*MBMDR-4.2.2* Continuous trait parallel workflow
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	≈ 270 days	25 hours 12 min	≈ 290 days	24 hours 06 min

(results prefixed by "≈" are extrapolated; 29 min 28 sec used to be approx. 9 days before; 256-core computer cluster Intel L5420 2.5 GHz)

- **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!]**
- **Van Lishout F,** Gadaleta F, Moore JH, Wehenkel L, Van Steen K (2015) gammaMAXT: a fast multiple-testing correction algorithm – submitted **[C++ MB-MDR made SUPER-fast]**

Refinements ….Towards a GWAI protocol

- **Gusareva ES,** Van Steen K (2014) Practical aspects of genome-wide association interaction analysis. Hum Genet 133(11):1343-58 **[GWAI analysis protocol]**

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

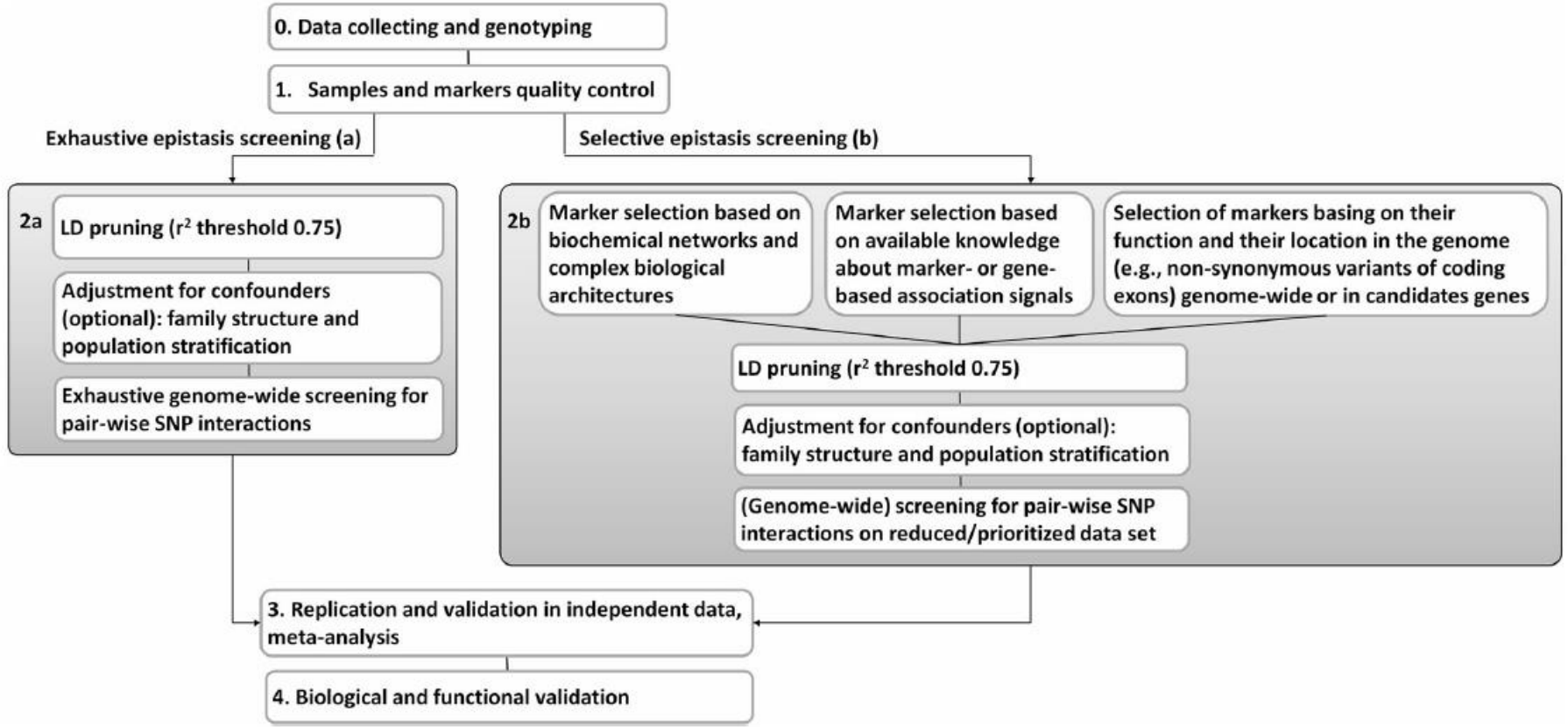
Received: 21 May 2014 / Accepted: 18 August 2014 / Published online: 28 August 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract Large-scale epistasis studies can give new clues to system-level genetic mechanisms and a better understanding of the underlying biology of human complex disease traits. Though many novel methods have been proposed to carry out such studies, so far only a few of them have demonstrated replicable results. Here, we propose a

Introduction

Genome-wide association (GWA) studies have been very successful in identifying predisposing genetic variants to a variety of complex traits (e.g., GWAS Diagram Browser for exploring GWA studies at http://www.ebi.ac.uk/fgpt/gwas/

A GWAI protocol consists of several components

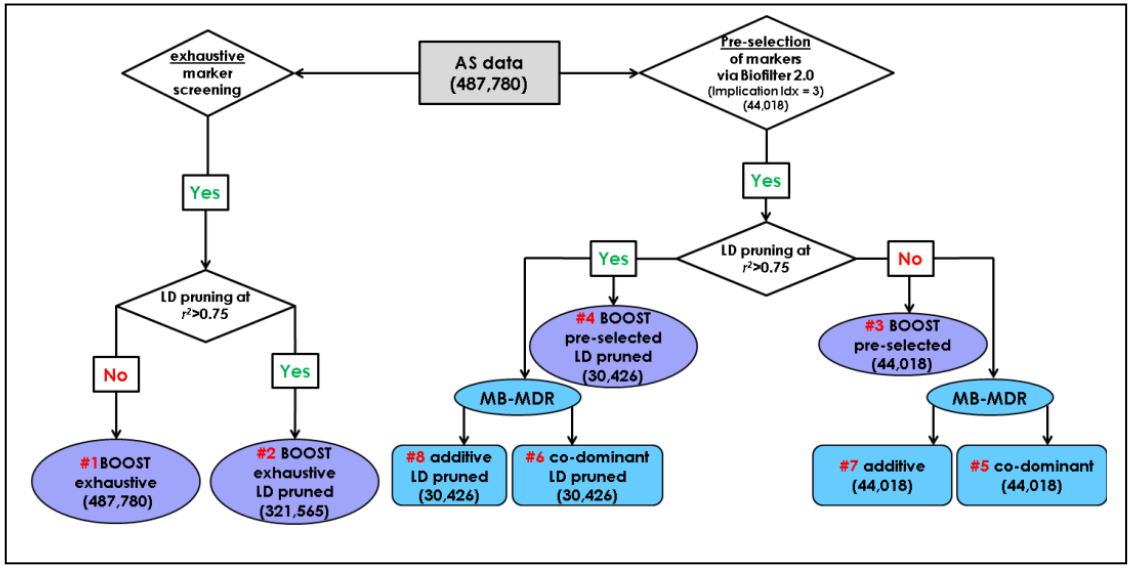


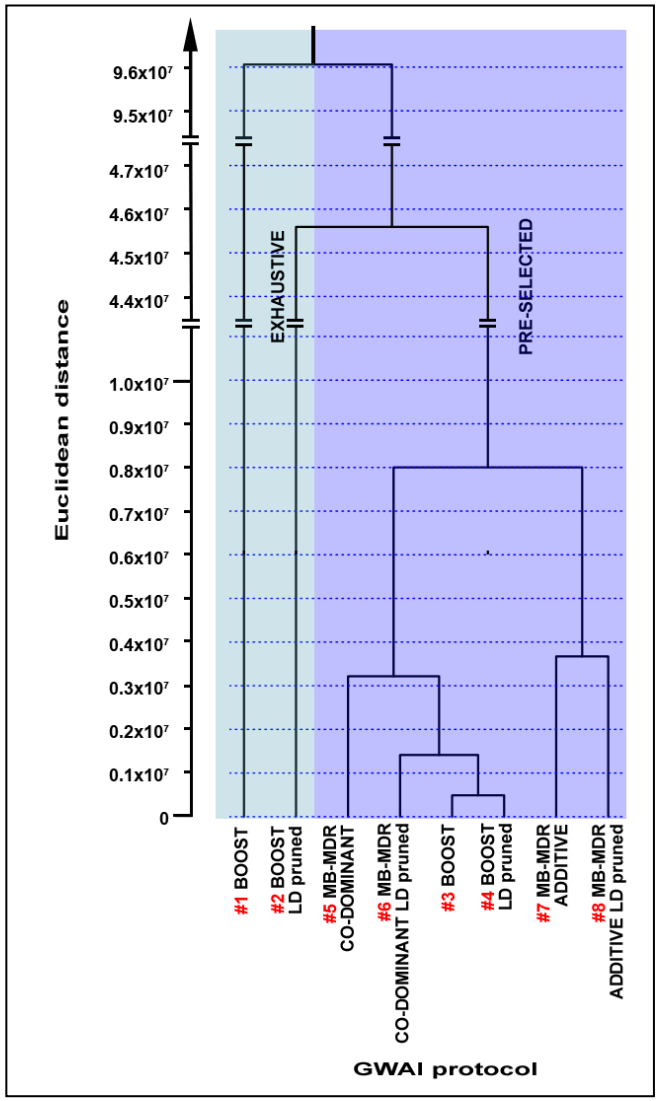
(Gusareva et al. 2014)

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

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]



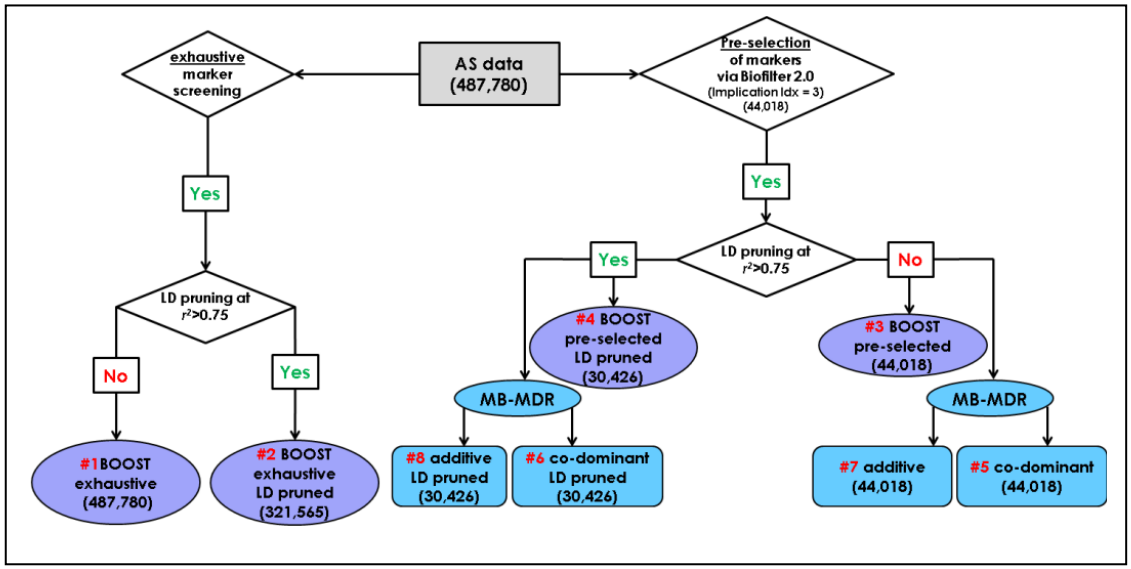


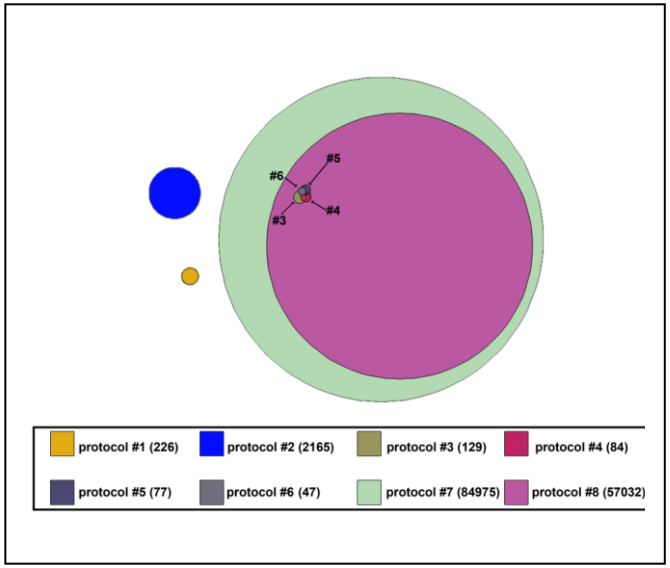
Slight protocol changes may lead to huge differences in results

- **Bessonov K, Gusareva ES, Van Steen K (2015)**

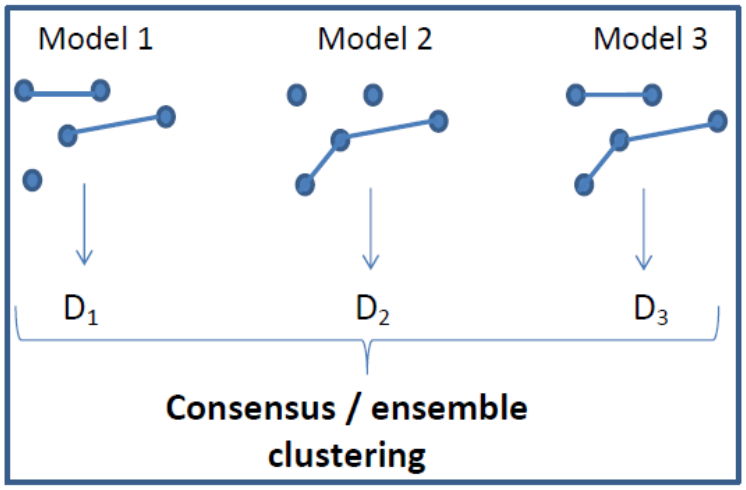
  A cautionary note on the impact of protocol changes for Genome-Wide Association SNP x SNP Interaction studies: an example on ankylosing spondylitis. Hum Genet - accepted

  [non-robustness of GWAI analysis protocols]



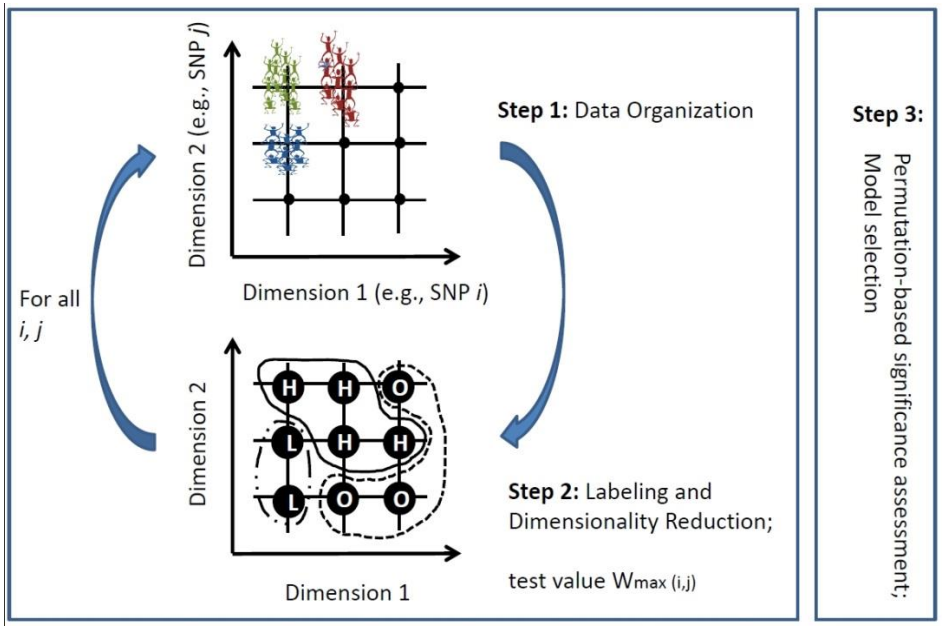


Refinements …. Consensus versus non-consensus interactions



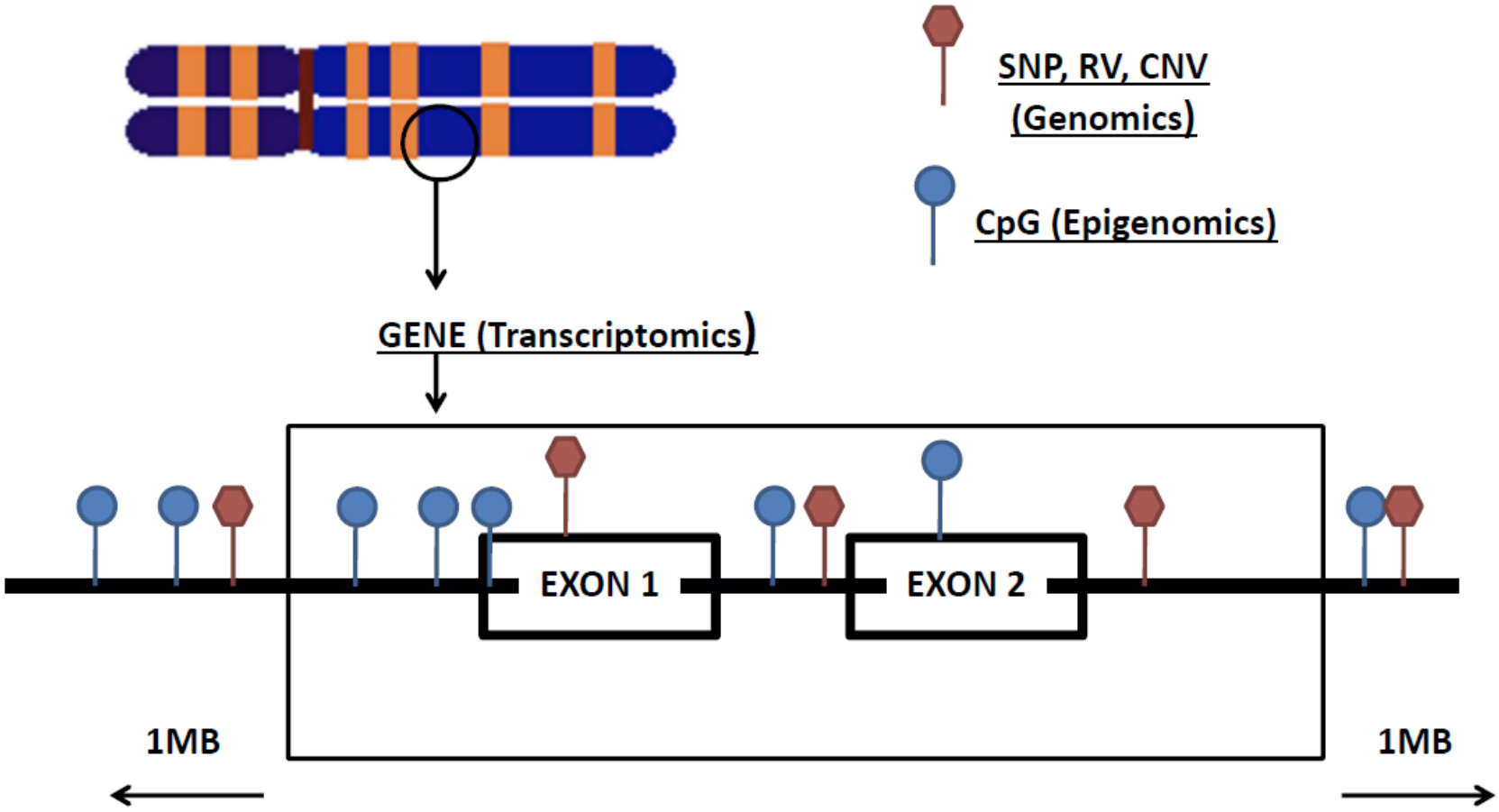
- Partitionings derived from undirected weighted graphs may be significantly different due to analytic heterogeneity
- Consensus clustering + meta clustering = Multiple Consensus Clustering (MCC: Zhang and Li 2011)
- Similarity Network Fusion (Wang et al. 2014) and independent power estimates

Context matters



- Structured populations (genomic control – Van Lishout)
- Meta-analysis (non-parametric – Gusareva)
- Replication and functional interpretation (Gusareva)
- Set-based (Fouladi)

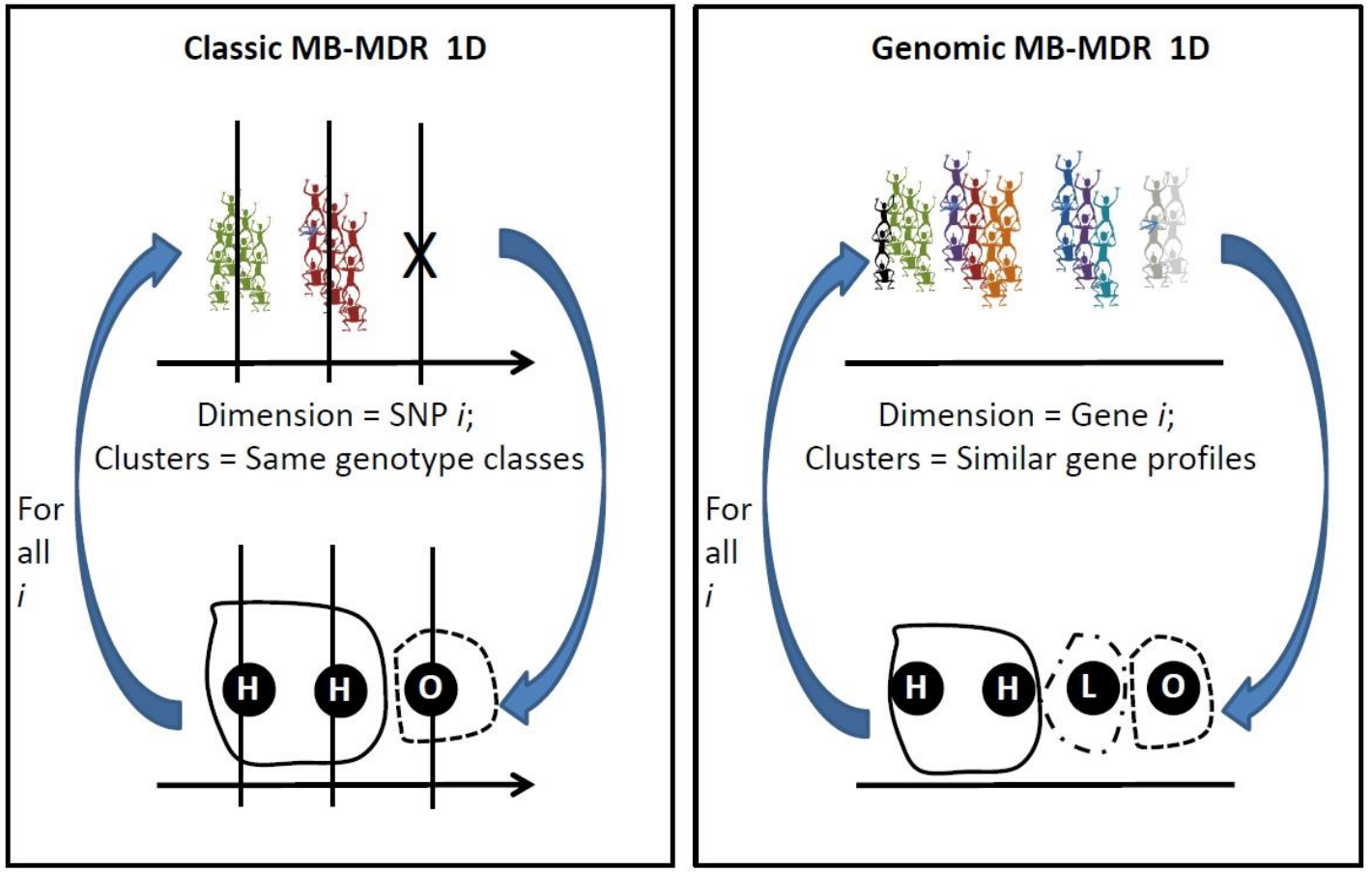
Refinements …. Higher-level interactions - Genes have different faces



(Slide S Pineda – lab meeting 2014)

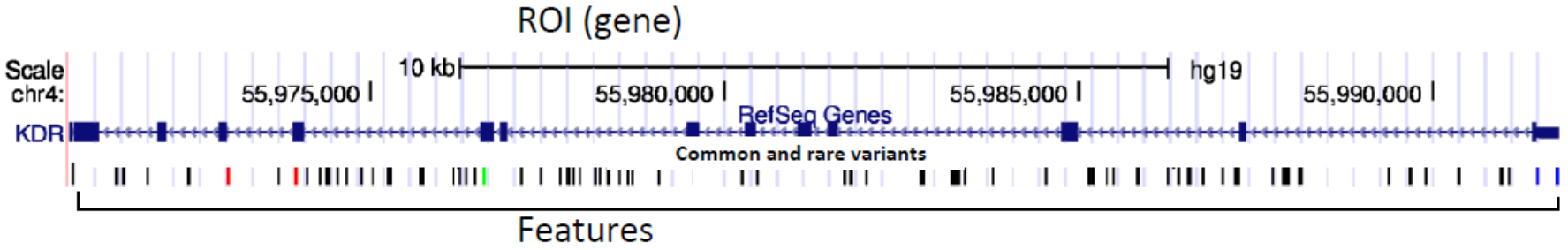
MB-MDR (SNPxSNP) → Genomic MB-MDR

- **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]**

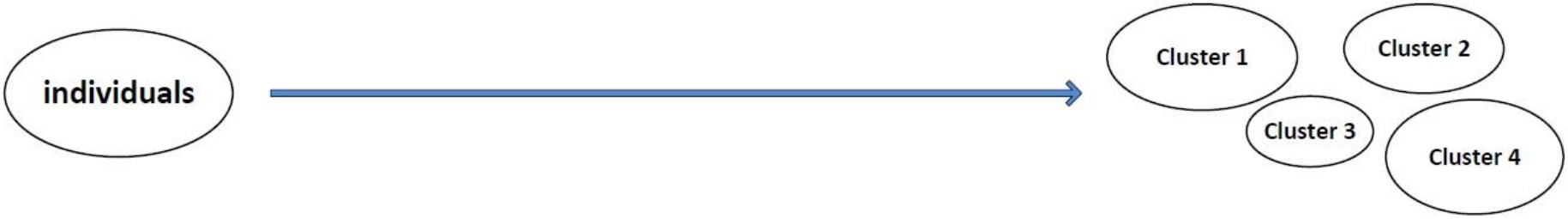


The genomic MB-MDR framework (Fouladi et al. 2015 – DNA-seq)

- **Phase 1:** Select sets of interest (ROI) / Prepare the data



- **Phase 2:** Clustering individuals according to features (e.g., common and rare variants, epigenetic markers, … and kernel methodology)



- **Phase 3:** Application of classic MB-MDR on new constructs

Genomic MB-MDR facilitates replication efforts in GWAI studies

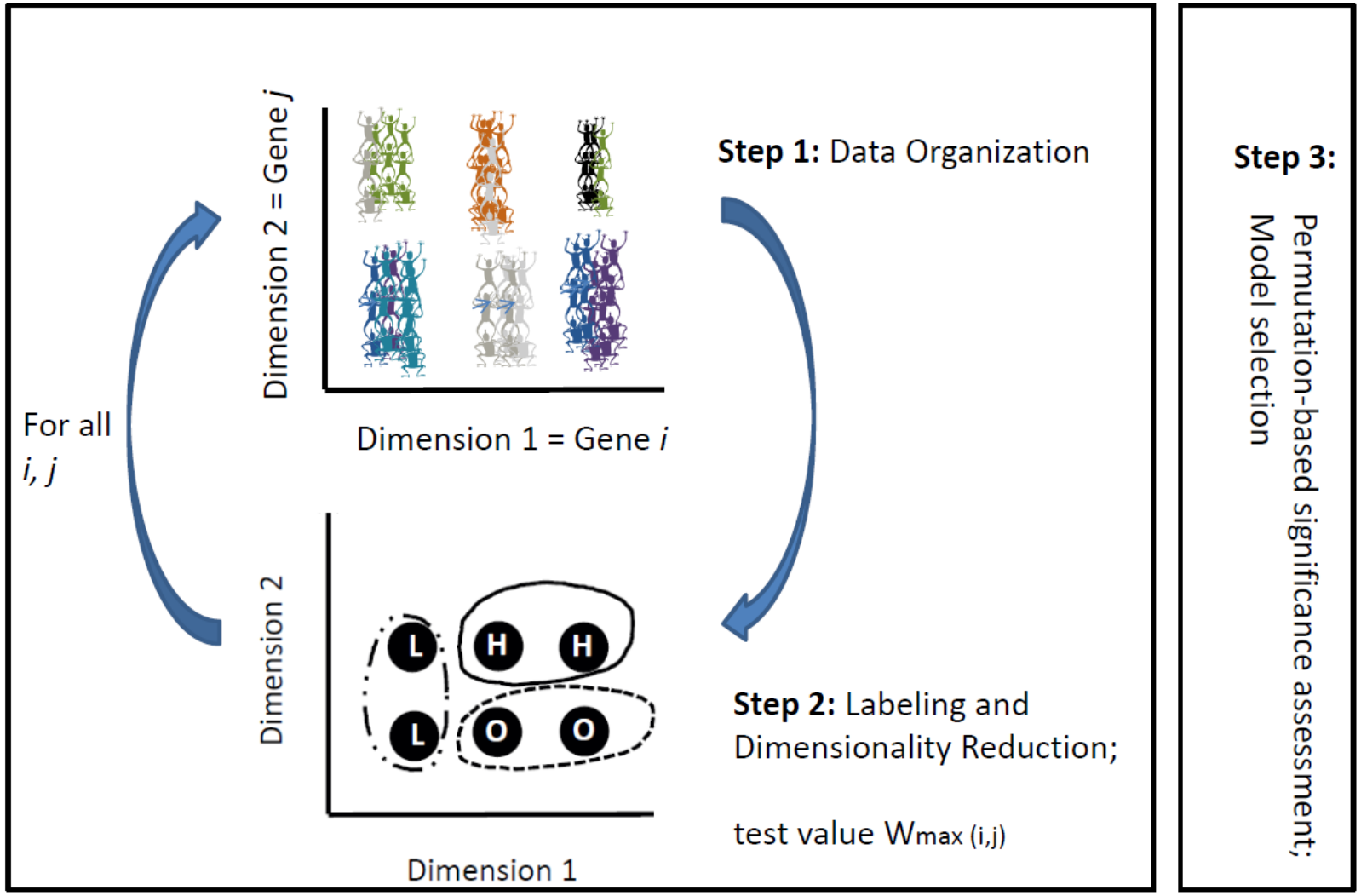
"Leaving aside for the moment **what replication means** or should mean in the context of GWAIS, even for the currently so-called replicated genetic interactions it is unclear to what extent **a false positive has been replicated** due to the adopted methodological strategy itself or whether the replication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

"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 2015 – under review)

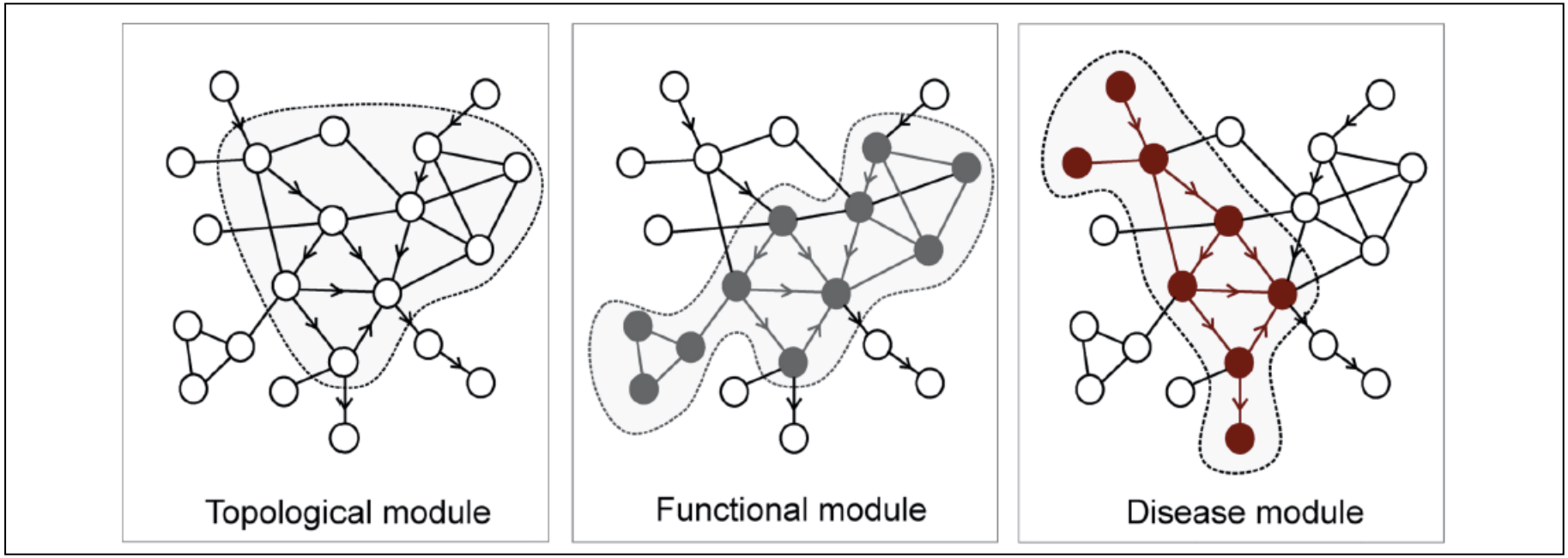
- Define the **(higher) level** that is common to studies (e.g., gene-level).

MB-MDR (set x set)



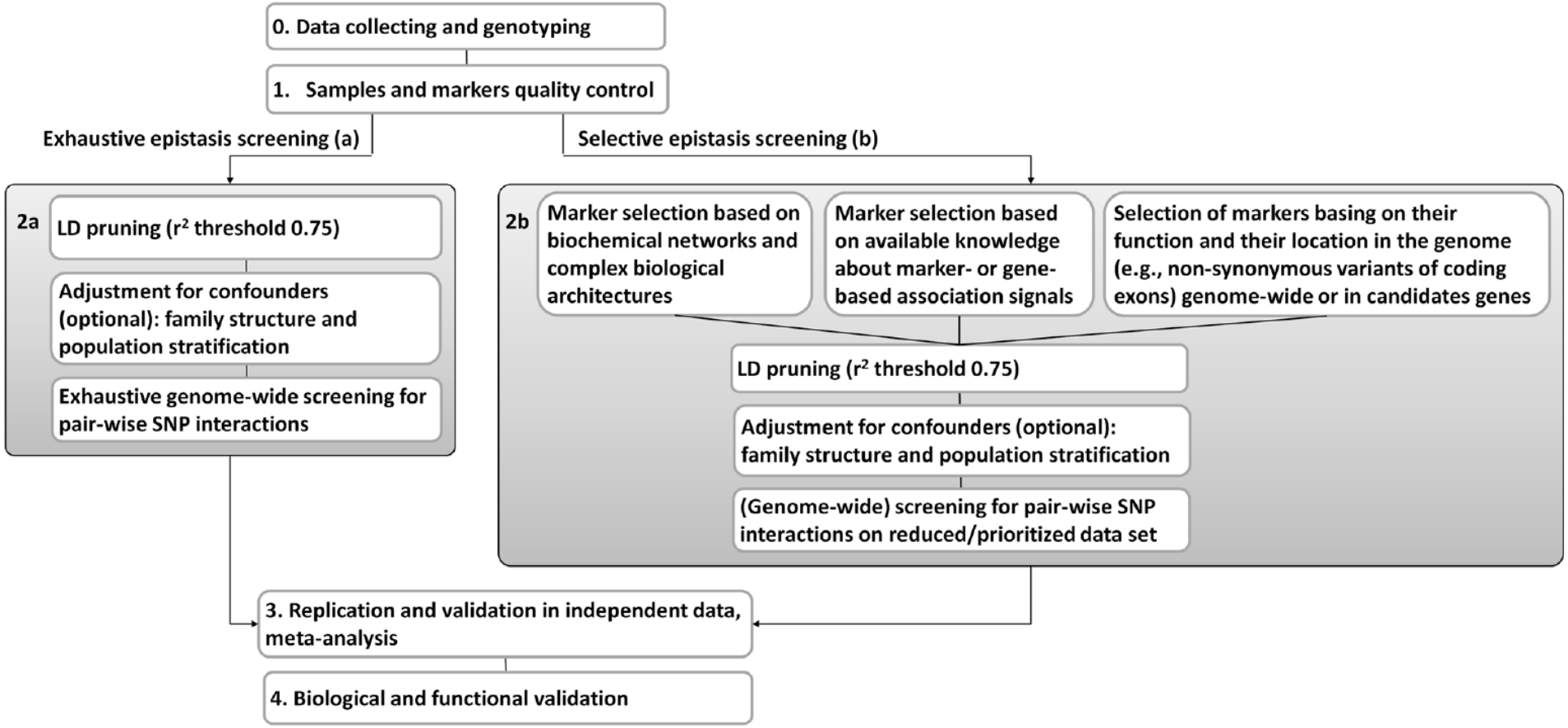
Genomic MB-MDR facilitates network medicine

- The underlying assumption is that the topological, functional, and disease modules overlap so that functional modules correspond to topological modules and a disease can be viewed as the breakdown of a functional module (Barabási et al. 2011)



GWAIs in practice

Protocol for GWAI analysis



(Gusareva and Van Steen 2014)

First hurdle: Selection of most appropriate method

- Honest methods comparisons should / can highlight the "core" (**the ABC**) of each method:

  A: Pre-processing (screening); **B**: core; **C**: multiple testing

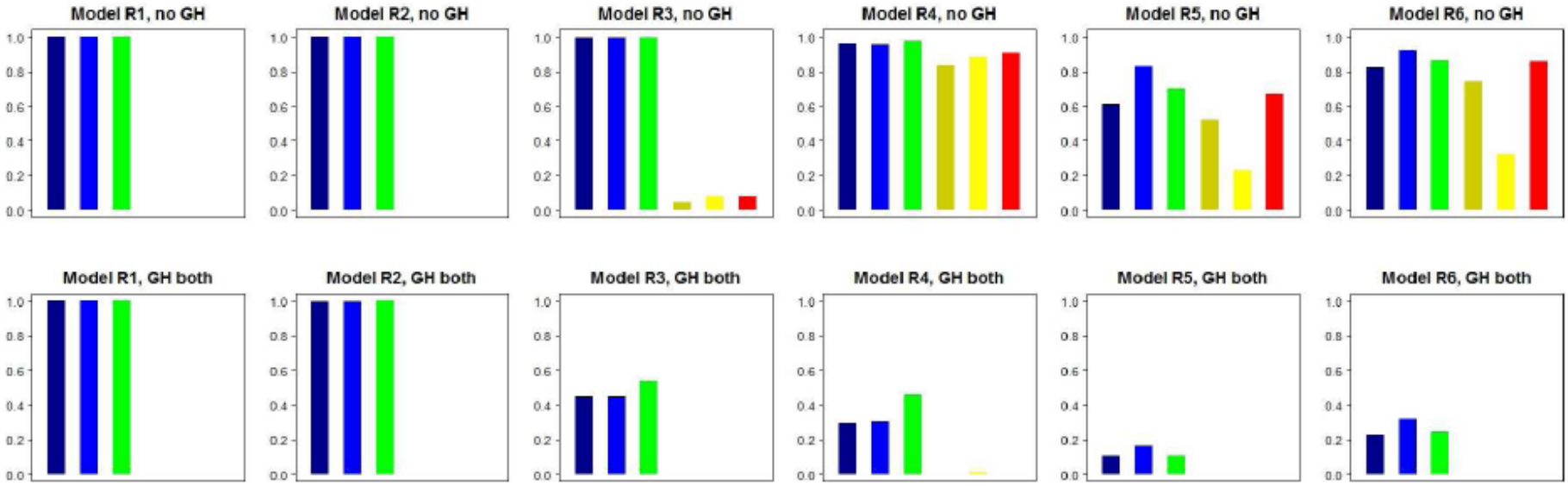
(Van Steen lab: in preparation)

		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**

Extending the toolbox for “interaction” detection

- Why? *Huge variability in power performance*

(example: pure epistasis scenario’s)



BOOST (dark blue)

EpiCruncher optimal options (light blue)

MB-MDR (green)

PLINK epistasis (dark yellow)

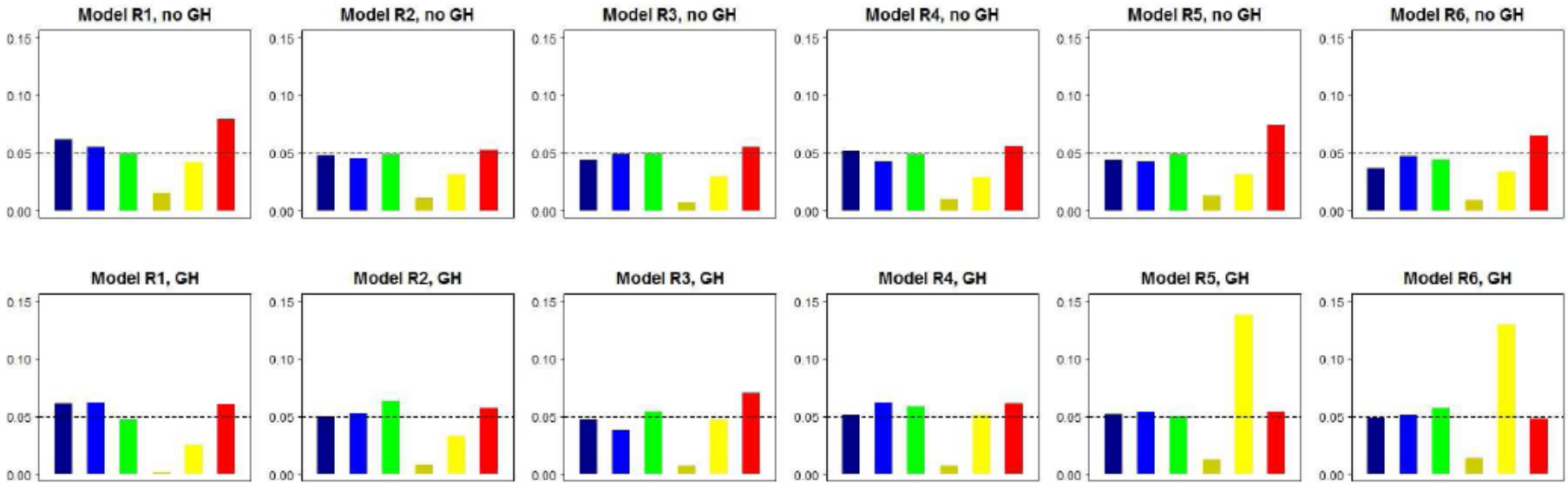
PLINK fast epistasis (light yellow)

EPIBLASTER (red)

Extending the toolbox for “interaction” detection

- Why? *Huge variability in false positive control*

(example: 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)

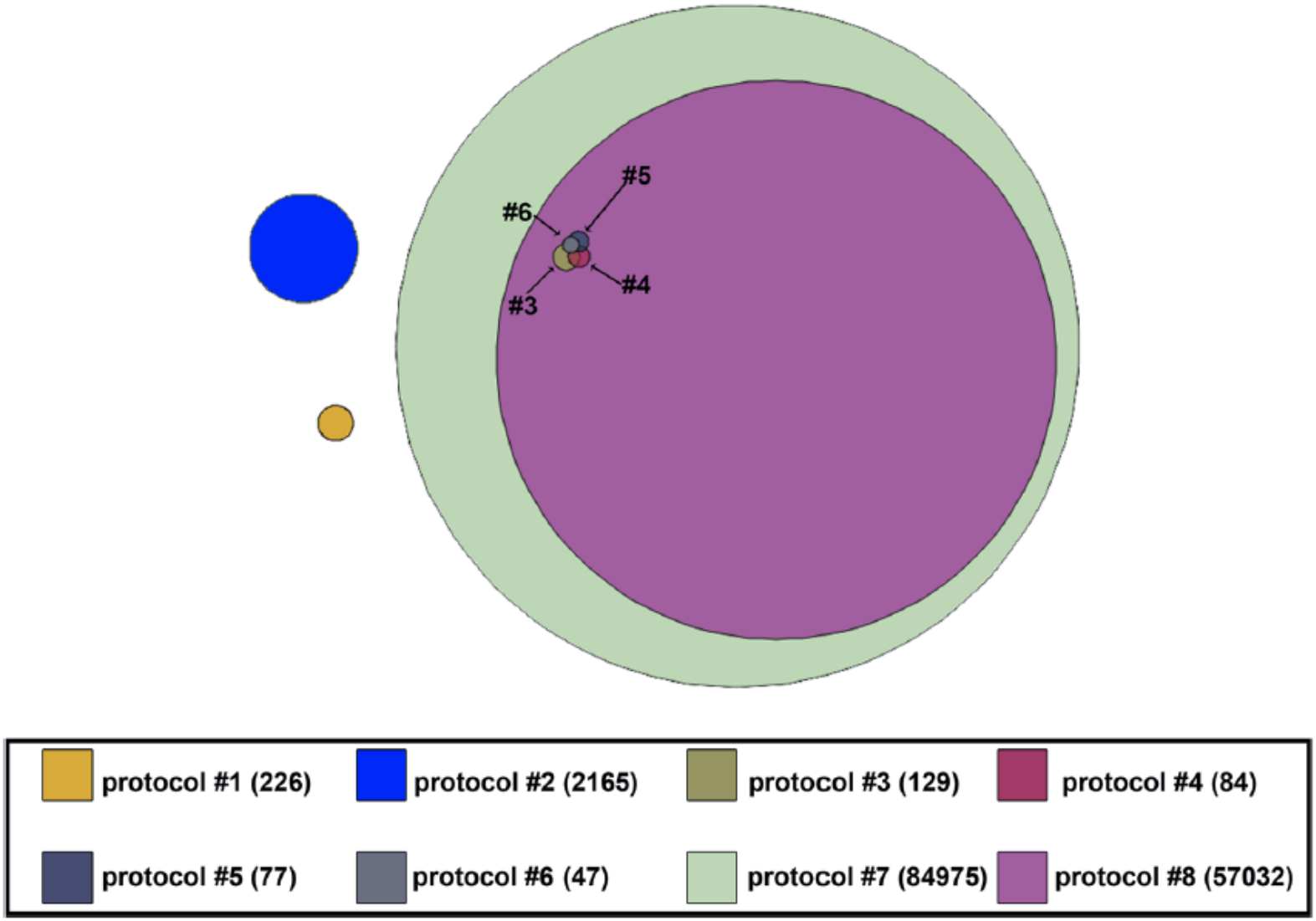
Extending the toolbox for “interaction” detection

- Recall:
  - LD between markers and long-distance between-marker associations
  - Missing data handling
  - Multi-stage designs incl marker selection
  - Multiple testing handling
  - Population stratification and admixture
  - Meta-analysis and replication
  - Multi-locus models that can improve the proportion of heritability explained

Second hurdle: (bio-) data filtering and LD handling

(K Bessonov et al. 2015)

	biofiltering	pruning	additive encoding
BOOST			
#1	N	N	N
#2	N	Y	N
#3	Y	N	N
#4	Y	Y	N
MB-MDR			
#5	Y	N	N
#6	Y	Y	N
#7	Y	N	Y
#8	Y	Y	Y


#5
#6
#4
#3
protocol #1 (226)
protocol #2 (2165)
protocol #3 (129)
protocol #4 (84)
protocol #5 (77)
protocol #6 (47)
protocol #7 (84975)
protocol #8 (57032)

Third hurdle: Level of detail – SNPS, genes, pathways, …

- MB-MDR analysis: 294 SNPs selected from France_AlzD panel of SNPs

	MTHFR	IL10	IL1A	IL1B	TF	HFE	IL6	ABCA1	DBH	INS	LRP1	CDK5R1	MAPT	NPC1	NR1H2	HMOX1	PPARA
MTHFR		+	ns	+	+	+	+	+	+	+	+	ns	+	+	+	ns	+
IL10			+	+	+	ns	ns	+	+	ns	+	ns	+	ns	ns	+	+
IL1A				ns	+	+	+	+	ns	+	ns	ns	+	ns	ns	ns	+
IL1B					+	ns	ns	+	ns	ns	+	ns	+	+	ns	ns	ns
TF						+	+	+	+	ns	+	ns	+	+	+	+	+
HFE							+	+	ns	+	+	ns	+	+	+	ns	+
IL6								+	+	ns	ns	ns	+	+	+	+	+
ABCA1									+	+	+	ns	+	+	+	+	+
DBH										+	+	ns	+	+	ns	+	+
INS											ns	ns	+	ns	ns	+	+
LRP1												ns	+	ns	ns	+	+
CDK5R1													ns	ns	ns	ns	ns
MAPT														+	ns	+	+
NPC1															ns	ns	+
NR1H2																ns	ns
HMOX1																	+
PPARA																	

"+" - at least one SNP pair from the corresponding genes was associated with AlzD

(the marginal *p*-value < 0.05 for the MB-MDR$_{2D}$ analysis)

“Replication” is highlighted by green; no replication is highlighted by red.

Fourth hurdle: Replication

(Mission Impossible @ google)

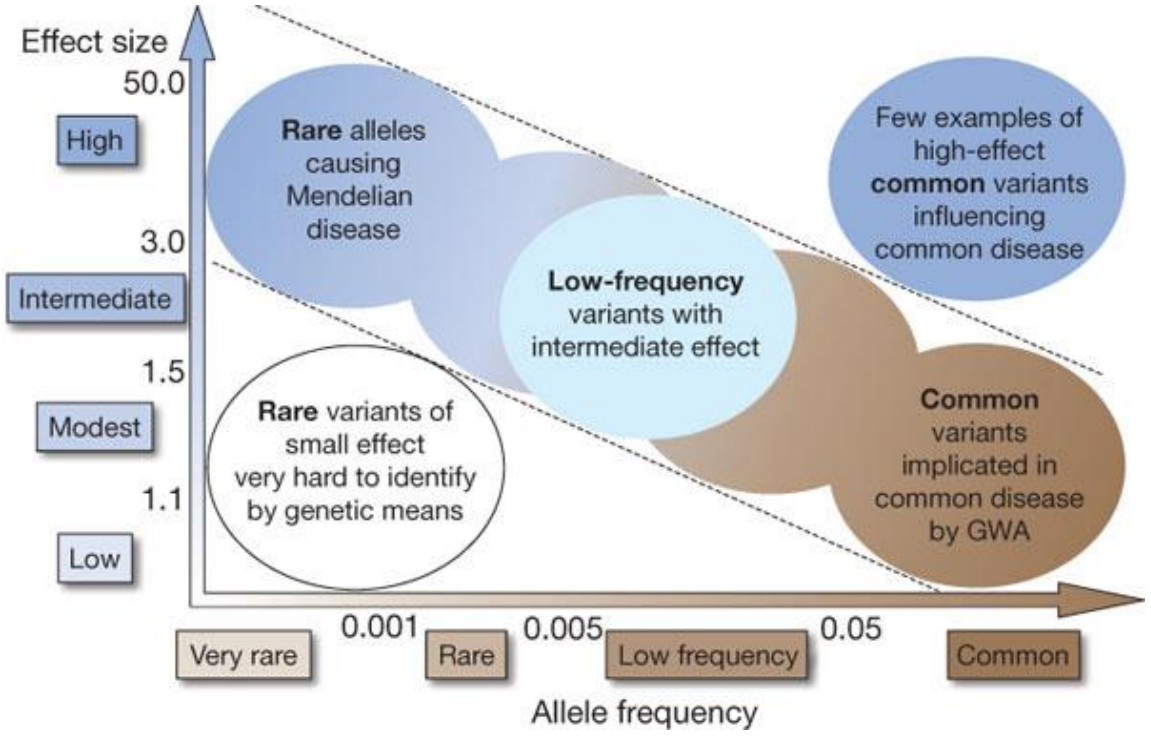
Replication

- 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 ….

“Replication as a concept should be revised in the context of GWAI studies”

Optimal conditions for GWA (Interaction) replication

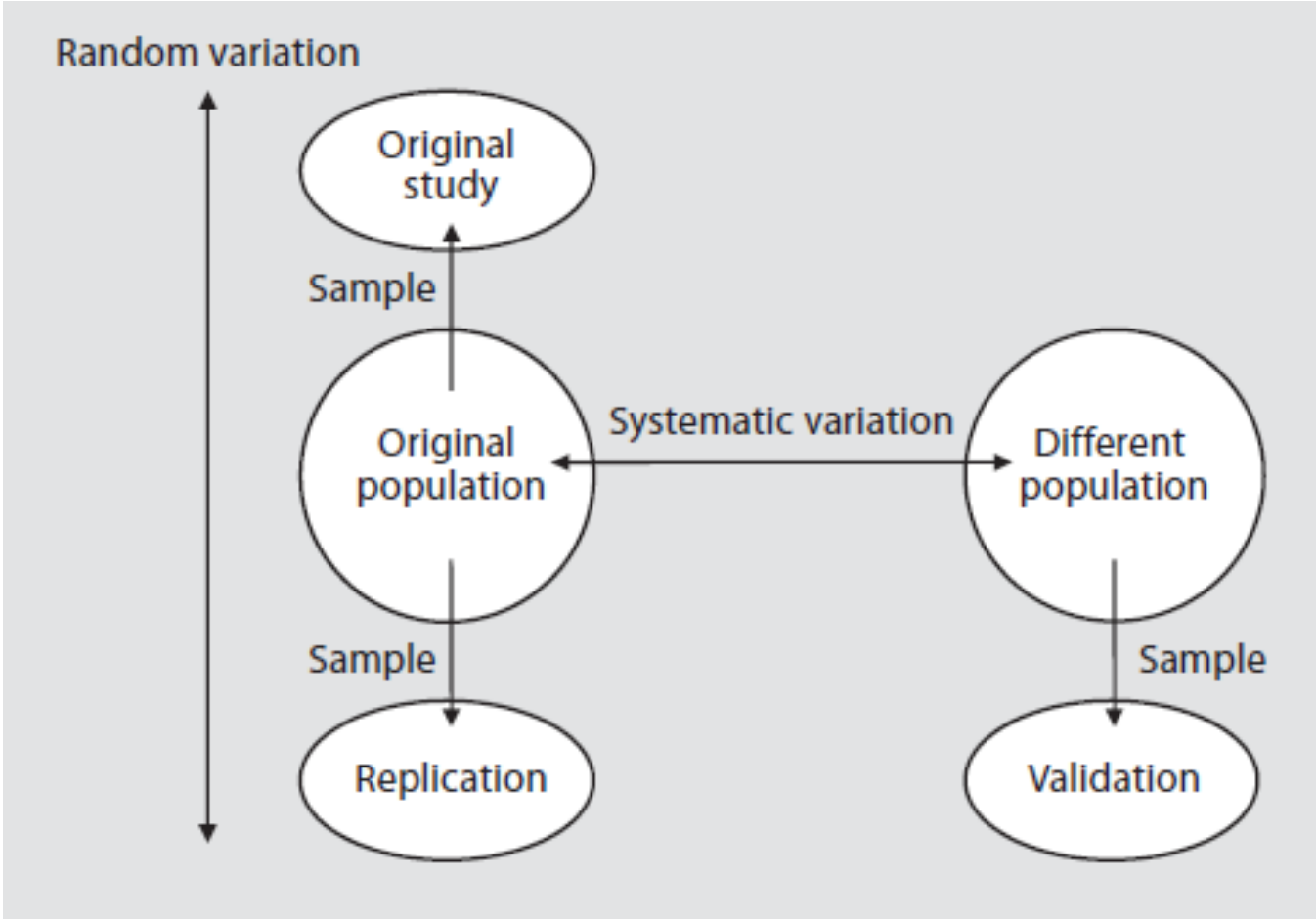
- Showing modest to strong statistical significance
- Having common minor allele frequency (>0.05)
- Modest to strong genetic effect sizes (parametric paradigms)



Compare to the diagonal focus region of GWAs
(Manolio et al. 2009)

Validation

- Validation is not replication:



(Igl et al. 2009)

Concluding Remarks

Human complex disease (gen-)omics

- Complex disease research aims/should aim to find answers to particular questions that are of interest to "people"
  - Requires an interdisciplinary approach (goes beyond multidisciplinary)
  - Requires an international approach (http://www.cost.eu/domains_actions/bmbs/Actions/BM1204
  - Requires an atmosphere of openness and complementarity (rather than competitiveness)

"Everything, however finely spun, finally comes to the sun"

(nothing can be hidden forever)

Pieter Bruegel the Elder, c. 1525 – 1569: “the Tower of Babel”)