

Genome-Wide Association Interaction (GWAI) Studies

Kristel Van Steen, PhD²

kristel.vansteen@ulg.ac.be

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

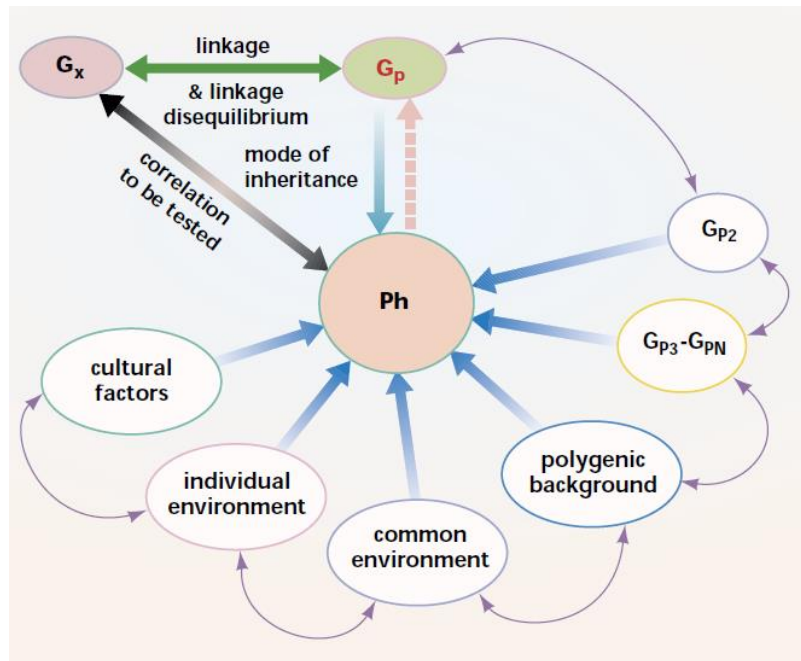
Bioinformatics and Modeling, GIGA-R, University of Liège, Avenue de l'Hôpital 1, 4000 Liège, Belgium

Outline

- The origin of “interactions”
- Travelling the world of interactions
- How to best build our working space
- Components of epistasis analysis
- Model-Based Multifactor Dimensionality Reduction
- An example on Alzheimer’s disease
- Validation and replication: An impossible task?
- Challenges and opportunities
- Proof of concept

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

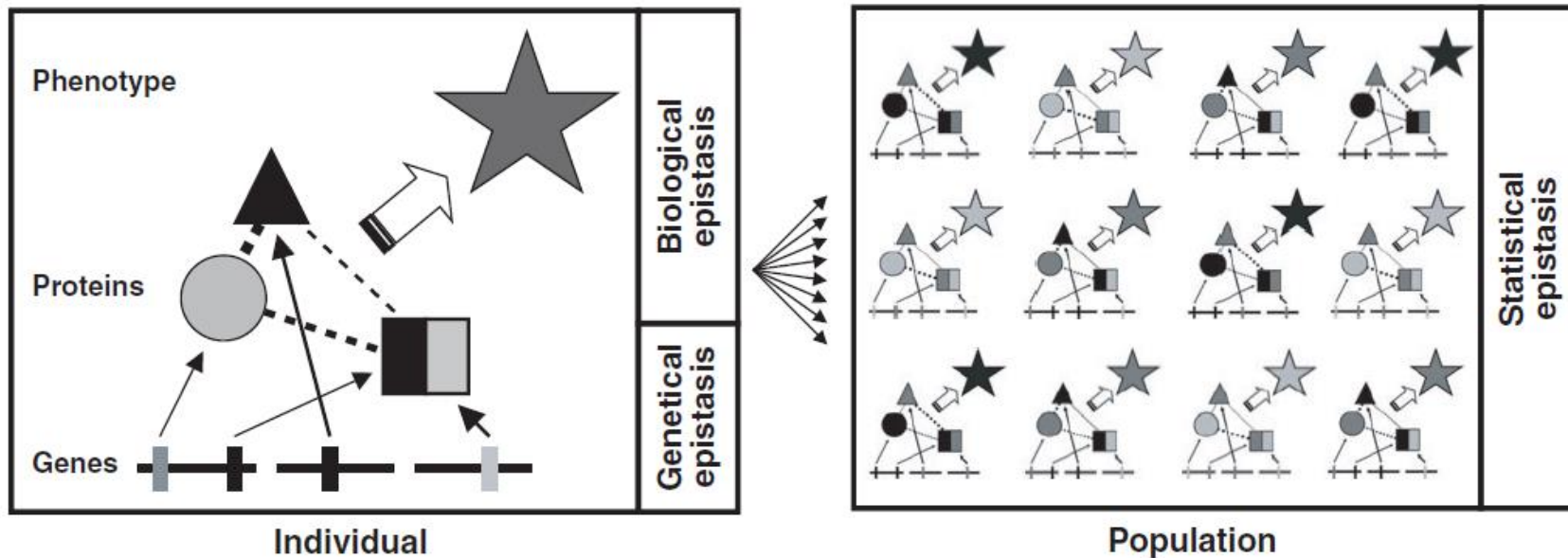
	Locus Heterogeneity	Trait Heterogeneity	Gene-Gene Interaction
Definition	when two or more DNA variations in distinct genetic loci are independently associated with the same trait	when a trait, or disease, has been defined with insufficient specificity such that it is actually two or more distinct underlying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

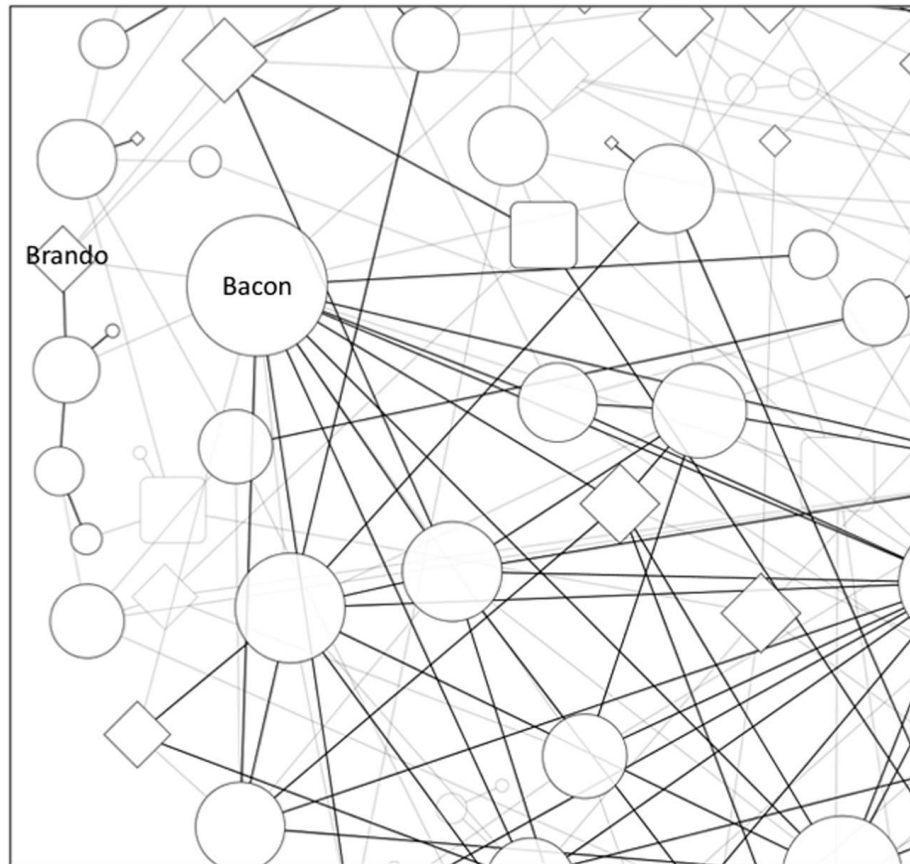
Great expectations

- 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

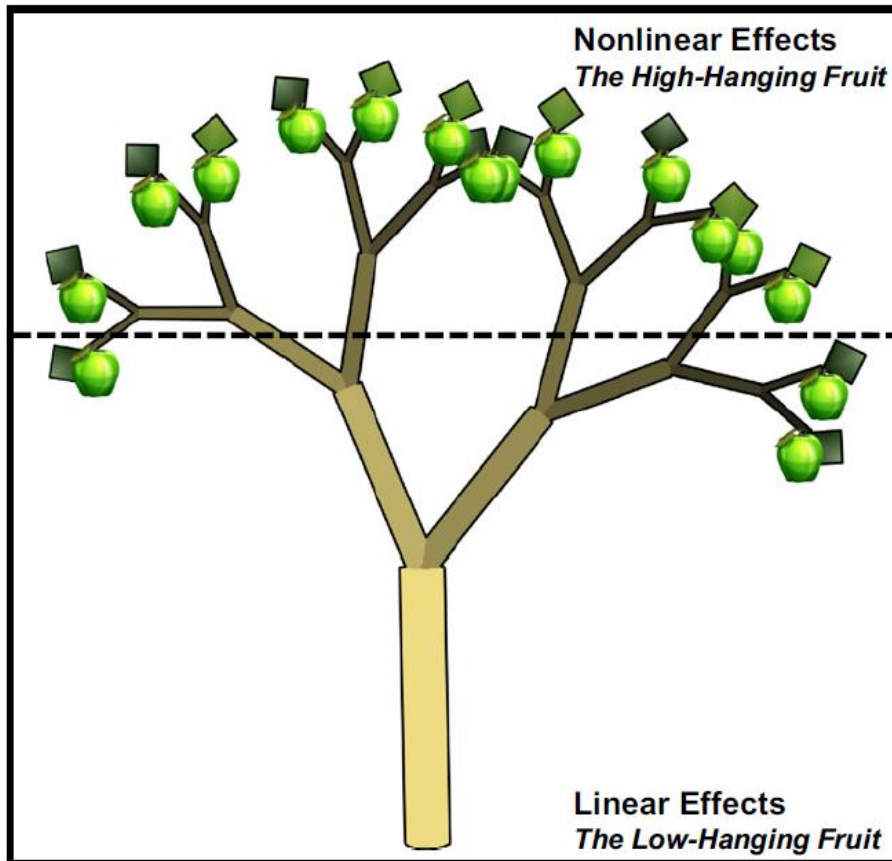
- Human genetics has been haunted by the mystery of “missing heritability” of common traits.
- Although studies have discovered >1,200 variants associated with common diseases and traits, these variants typically appear to explain only a minority of the heritability.
- The proportion of heritability 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





- 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.

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

(Moore and Williams 2009)

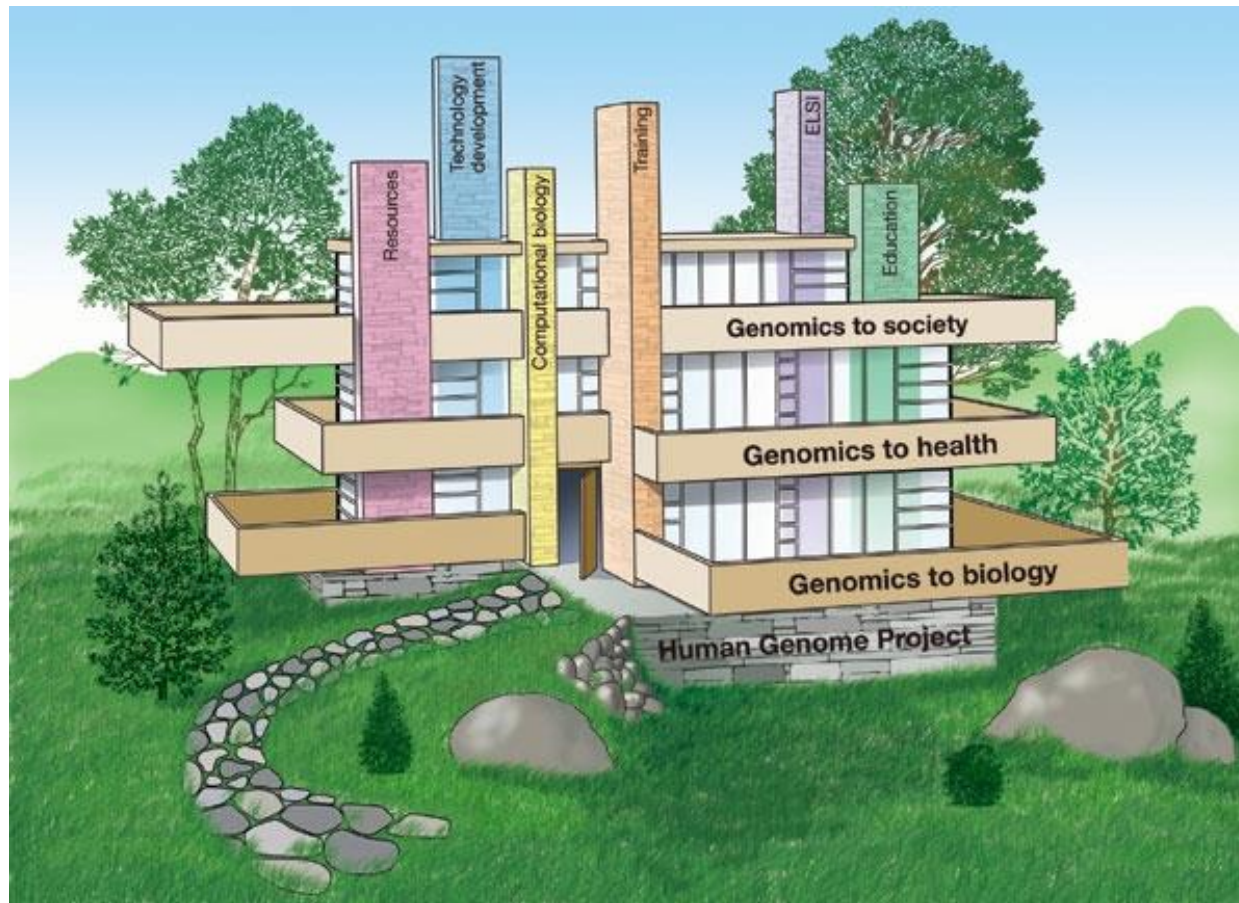
From GWA to GWAI studies ...

- 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”

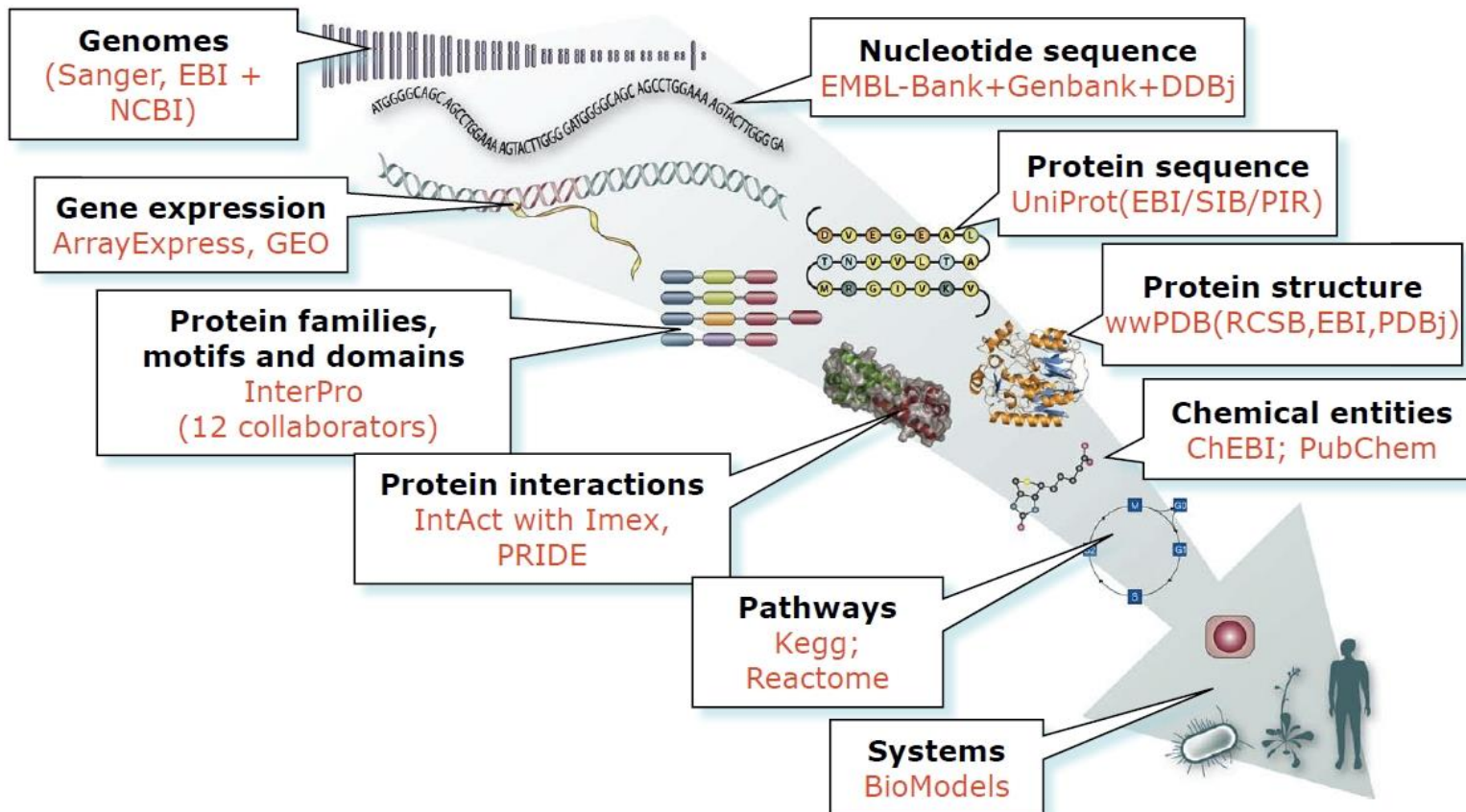


(<http://www.genome.gov>: the future of human genomics) + harmonization of biobanks

Creating an atmosphere of “integration”

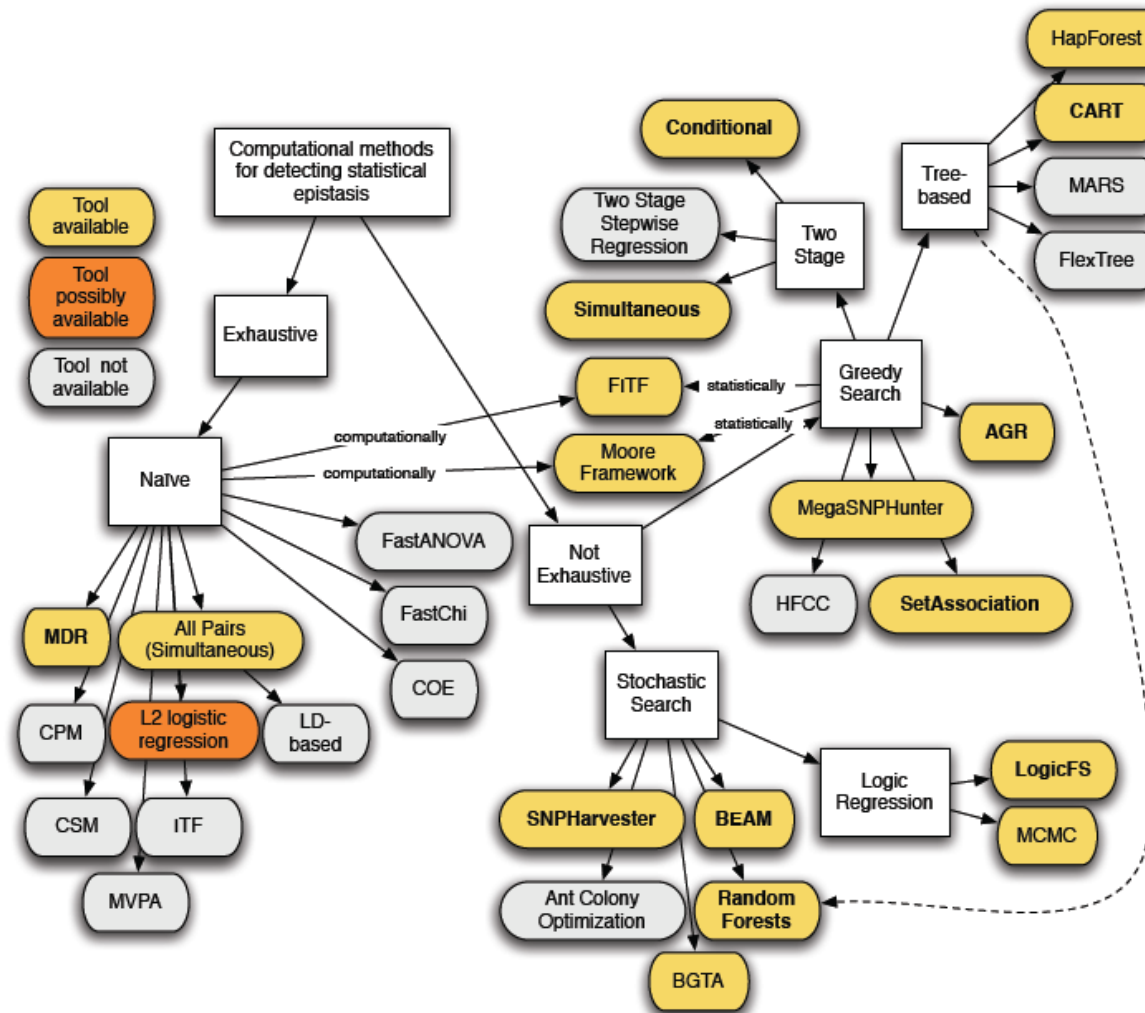
with HTP omics data

(J Thornton, EBI)



Extending the toolbox

(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 disease⁶³, diabetes⁶⁴, bipolar effective disorder⁶⁵, and autism⁶⁶.
 - 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), Alzheimer’s (Combarros et al 2009),

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 GWAs 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?

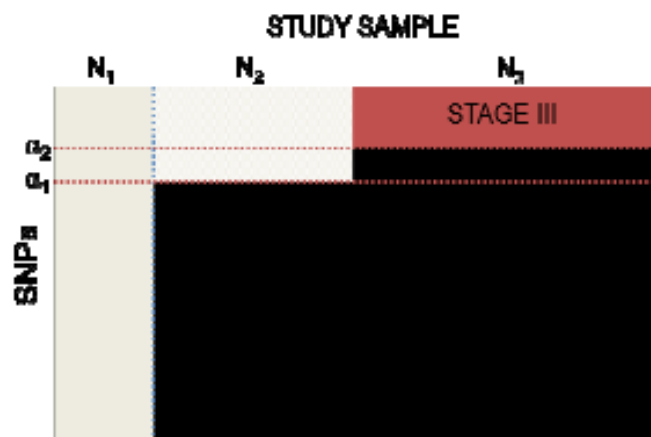
Introduction

- 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 main effects GWAS

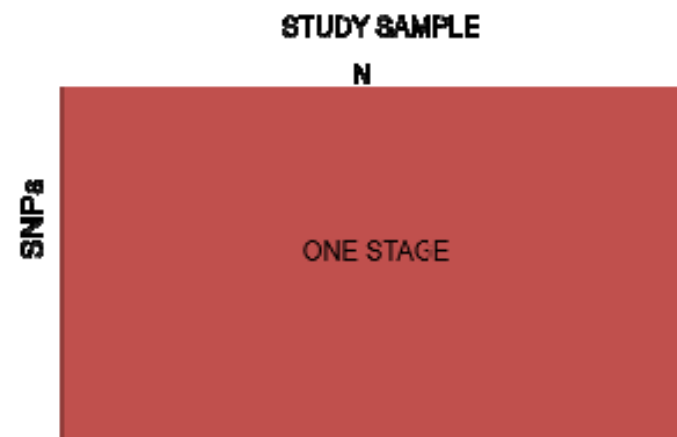
Multi-stage

- Less expensive
- More complicated
- Less powerful



Single-stage

- More expensive
- Less complicated
- More powerful



(slide: courtesy of McQueen)

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)

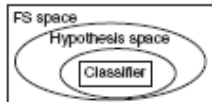
Feature selection methods

Model search	Advantages	Disadvantages	Examples
Filter	Univariate		
	Fast Scalable Independent of the classifier	Ignores feature dependencies Ignores interaction with the classifier	χ^2 Euclidean distance <i>i</i> -test Information gain, Gain ratio (Ben-Bassat, 1982)
	Multivariate		
	Models feature dependencies Independent of the classifier Better computational complexity than wrapper methods	Slower than univariate techniques Less scalable than univariate techniques Ignores interaction with the classifier	Correlation-based feature selection (CFS) (Hall, 1999) Markov blanket filter (MBF) (Koller and Sahami, 1996) Fast correlation-based feature selection (FCBF) (Yu and Liu, 2004)

(Saeys et al 2007)


Feature selection methods

Model search	Advantages	Disadvantages	Examples
Wrapper	Deterministic		
	Simple Interacts with the classifier Models feature dependencies Less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection	Sequential forward selection (SFS) (Kittler, 1978) Sequential backward elimination (SBE) (Kittler, 1978) Plus q take-away r (Ferri <i>et al.</i> , 1994) Beam search (Siedelecky and Sklansky, 1988)
	Randomized		
	Less prone to local optima Interacts with the classifier Models feature dependencies	Computationally intensive Classifier dependent selection Higher risk of overfitting than deterministic algorithms	Simulated annealing Randomized hill climbing (Skalak, 1994) Genetic algorithms (Holland, 1975) Estimation of distribution algorithms (Inza <i>et al.</i> , 2000)



(Saeys et al 2007)

Feature selection methods

Model search	Advantages	Disadvantages	Examples
Embedded 	Interacts with the classifier Better computational complexity than wrapper methods Models feature dependencies	Classifier dependent selection	Decision trees Weighted naive Bayes (Duda <i>et al.</i> , 2001) Feature selection using the weight vector of SVM (Guyon <i>et al.</i> , 2002; Weston <i>et al.</i> , 2003)

(Saeys et al 2007)

- In contrast: 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 a defined 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

Information Gain: Slide 10

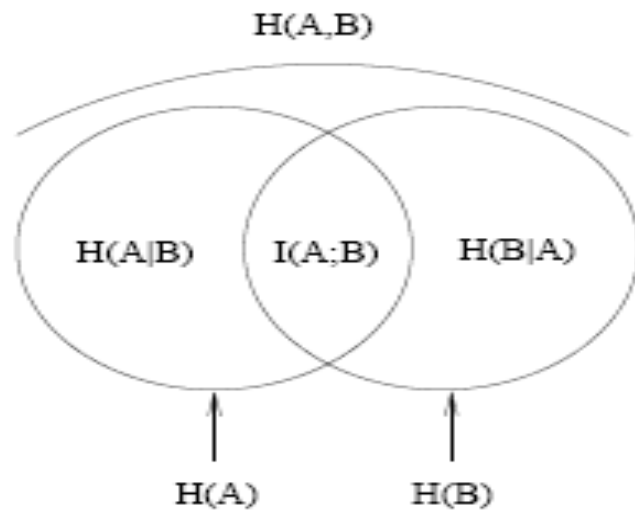
Conditional entropy

- The conditional entropy of two events A and B , taking on values a and b respectively, is defined as

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

- This quantity should be understood as the amount of randomness in the random variable A given that you know the value of B

Conditional entropy: interpretation



The surface area of a section corresponds to the labeled quantity

$H(A)$ = entropy of A

$I(A;B)$

= mutual information common to A and B

= the amount of information provided by A about B
(= non-negative!)

(Jakulin 2003)

Mutual information

- It can be shown that mutual information of two random variables A and B satisfies

$$I(A; B) = \sum_{a \in A, b \in B} p(a, b) \log_2 \frac{p(a, b)}{p_A(a)p_B(b)}$$

(Shannon 1948)

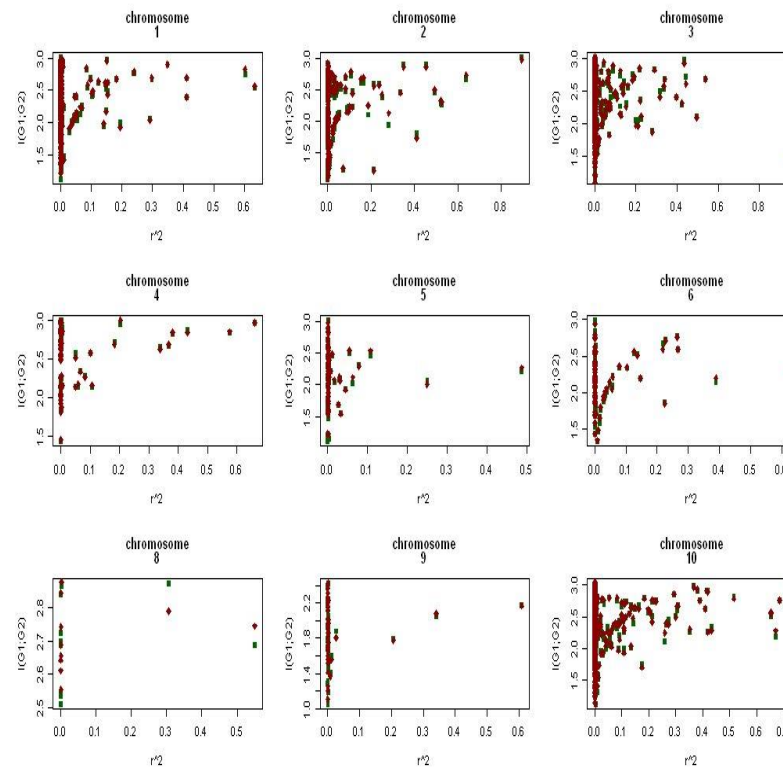
- Mutual information can be expressed as a Kullback-Leibler divergence, of the product $p_A(a)p_B(b)$ of the marginal distributions of the two random variables A and B , from the random variables' joint distribution
- $I(A; B)$ can also be understood as the expectation of the Kullback-Leibler divergence of the univariate distribution $p_A(a)$ of A from the conditional distribution $p_{A|B}(a|b)$ of A given B : the more different the distributions $p_{A|B}(a|b)$ and $p_A(a)$, the greater the **information gain**.

Mutual information: interpretation

- Intuitively, mutual information measures the information that A and B share: it measures how much knowing one of these variables reduces our uncertainty about the other.
 - For example, if A and B are independent, then knowing A will not give any information about B and vice versa, so their mutual information is zero.
 - At the other extreme, if A and B are identical, then all information conveyed by A is shared with B : knowing A determines the value of B and vice versa. As a result, in this case, the mutual information is the same as the uncertainty contained in A or B alone

Mutual information and r^2

- Mutual information $I(A ; B)$ as a function of r^2 (as a measure of LD between markers), for a subset of the Spanish Bladder Cancer data



(Van Steen et al - unpublished)

Mutual information and machine learning

- Suppose there is a message Y , that was sent through a communications channel, and we received the value X .
- We would like to decode the received value X , and recover the correct Y , hence perform a decoding operation $\hat{Y} = g(X)$
- In machine learning terms this translates to: Y is the original (unknown) class label distribution, X is the particular set of features chosen to represent the problem, and g is our predictor.
- The set of features chosen may or may not be sufficient to perfectly recover or predict Y :

$$\frac{H(Y) - I(X; Y) - 1}{\log(|Y|)} \leq p(g(X)) \leq \frac{1}{2}H(Y|X)$$

Fano 1961 Hellman & Raviv 1970)

Multivariate mutual information

- The multivariate form of Shannon's mutual information $I(X;Y)$ is often referred to as **Interaction Information** (McGill 1954), and accounts for dependencies among multiple variables (i.e. more than 2)
- To derive its expression, we first define the conditional mutual information between two variables X_1 and X_2 , after the value of Y is revealed

$$I(X_1; X_2|Y) = \sum_{y \in Y} p(y) \sum_{x_1 \in X_1; x_2 \in X_2} p(x_1 x_2 | y) \log \frac{p(x_1 x_2 | y)}{p(x_1 | y) p(x_2 | y)}$$

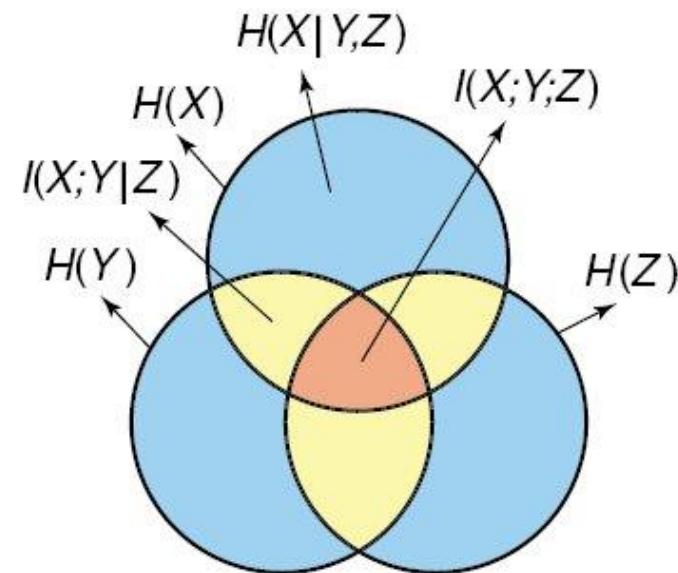
Multivariate mutual information

- 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



Multivariate mutual information

- McGill's interaction information is actually

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

- This coincides with a notion of **bivariate synergy**, comparing the joint contribution of X_1 and X_2 to X_3 with the additive contributions of each of them separately
- Bivariate synergy is defined as

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

- It can be shown, with this definition, that indeed

$$\text{Syn}(X_1, X_2; X_3) = -I(X_1; X_2; X_3)$$

(Anastassiou 2007)

Bivariate synergy: interpretation

- This quantity represents the additional information that both genetic factors jointly provide about the phenotype after removing the individual information provided by each genetic factor separately.
- Hence, in general, synergy is the additional contribution provided by the “whole” compared with the sum of the contributions of the “parts”.

(Varadan et al 2006)

- Or stated otherwise, since

$\text{Syn}(X_1, X_2; X_3) = I(X_1; X_2|X_3) - I(X_1; X_2)$, 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)

Bivariate synergy: interpretation

If $\text{Syn}(A,B;C) > 0$

Evidence for an attribute interaction that cannot be linearly decomposed

If $\text{Syn}(A,B;C) < 0$

The information between A and B is redundant

If $\text{Syn}(A,B;C) = 0$

Evidence of conditional independence or a mixture of synergy and redundancy

Attribute selection based on information gain: 2nd order effects

- Based on the definition of “synergy” and its equivalent expressions, we can now derive a rule for feature selection:
 - Compute the entropy-based measure $\text{Syn}(SNP1, SNP2; C)$, the synergy of $SNP1$ and $SNP2$ with respect to a class variable C , for each pair-wise combination of attributes $SNP1$ and $SNP2$
 - Pairs of attributes are sorted and those with the highest $\text{Syn}(SNP1, SNP2; C)$ are selected for further epistasis analysis

Highlight 2: Multivariate filtering

Attribute selection based on Relief

(Kira and Rendell 1992)

- For each instance, the closest instance of the same class (nearest hit) and the closest instance of a different class (nearest miss) are selected, through a type of nearest neighbor algorithm.
- The weight or score $S(i)$ of the i -th variable is computed as the average, over all instances, of magnitude of the difference between the distance to the nearest hit and the distance to the nearest miss, in projection on the i -th variable.

Attribute selection based on ReliefF

- ReliefF is an extension of the Relief algorithm and is more robust than the original because it selects a set of nearby hits and a set of nearby misses for every target sample and averages their distances (Kononenko 1994)
- This minimizes the effects of spurious samples.
- ReliefF also extends Relief to multi-class problems by defining a different set of “miss” samples for every category.

Attribute selection based on tuned ReliefF

- The advantage of the Relief and ReliefF algorithms to capture attribute interactions is also a disadvantage because the presence of many noisy attributes can reduce the signal the algorithm is trying to capture.
- The “tuned” ReliefF algorithm (TuRF) systematically removes attributes that have low quality estimates so that the ReliefF weights of the remaining attributes can be re-estimated.

(Moore and White 2008)

- Gear up to SURF ... (Spatially Uniform ReliefF) for computationally efficient filtering of gene-gene interactions (Greene et al 2009)

Strategy 3: Data mining as embedding technique

Random Forests (RF)

(Breiman 2001)

The random forests algorithm (for both classification and regression) is as follows:

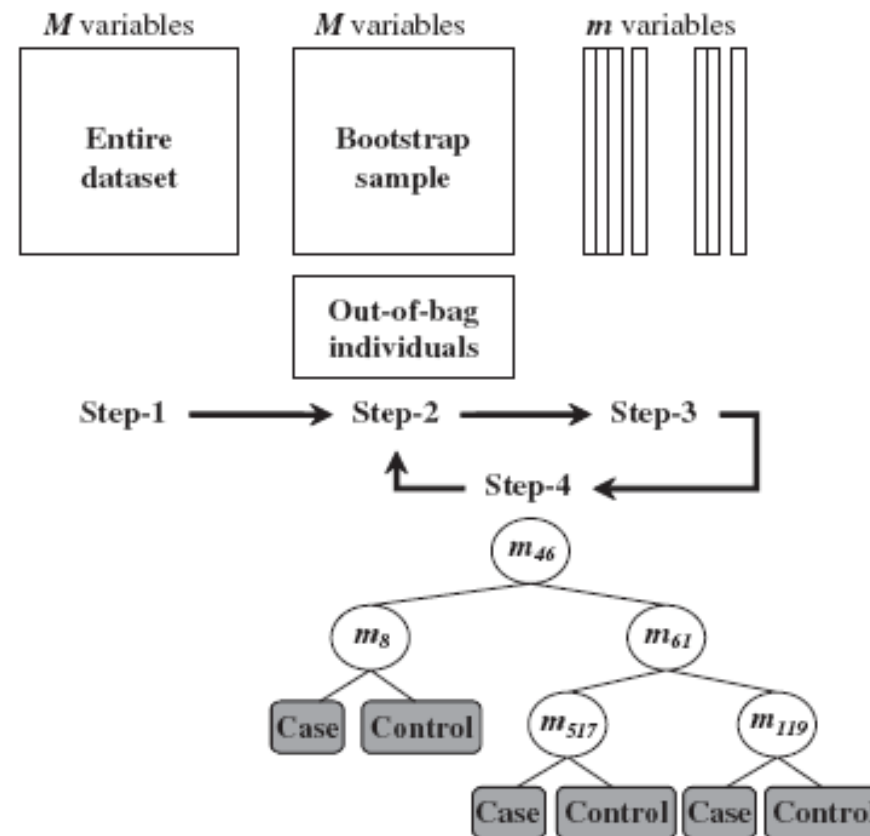
- Draw n_{tree} bootstrap samples from the original data.
- For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following specifications:
 - at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when $m_{try} = p$, the number of predictors)
 - Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification, average for regression).

Random Forests (RF)

- An estimate of the error rate can be obtained, based on the training data, by the following:
 - At each bootstrap iteration, predict the data not in the bootstrap sample (what Breiman calls “out-of-bag”, or OOB, data) using the tree grown with the bootstrap sample.
 - Aggregate the OOB predictions. (On the average, each data point would be out-of-bag around 36% of the times, so aggregate these predictions.)
 - Calculate the error rate, and call it the OOB estimate of error rate.

(Breiman 2001)

A schematic overview of the RF method



(Motsinger-Reif et al 2008)

Some advantages of the Random Forest method

- It estimates the relative importance of variables in determining classification, thus providing a metric for feature selection.
 - Beware: different RF importance measures have different stability properties and performance in the presence of highly correlated features ... (Calle and Urrea 2010; Nicodemos et al 2010)
- RF is fairly robust in the presence of heterogeneity and relatively high amounts of missing data (Lunetta et al., 2004).
- As the number of input variables increases, learning is fast and computation time is modest even for very large data sets (Robnik-Sikonja 2004).

Some advantages of the Random Forest method

- New implementations of RF allow rapid analysis of highly dimensional data such as those generated for GWA studies (Schwarz et al 2010): Random Jungle (<http://www.randomjungle.org/rjungle/>)



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.

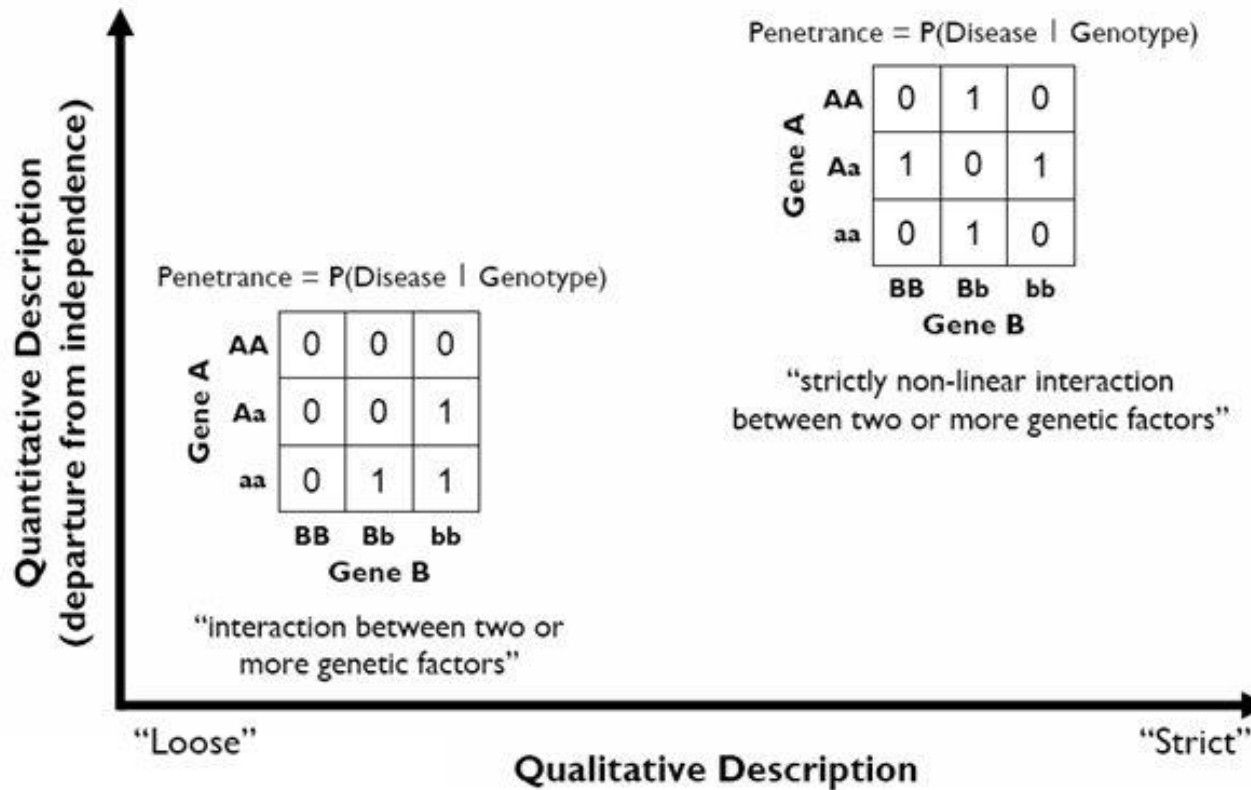
Enumeration of two-locus models

(Li and Reich 2000)

M1(RR) 0 0 0 0 0 0 0 0 1	M2 0 0 0 0 0 0 0 1 0	M3(RD) 0 0 0 0 0 0 0 1 1	M5 0 0 0 0 0 0 1 0 1	M7(IL:R) 0 0 0 0 0 0 1 1 1	M10 0 0 0 0 0 1 0 1 0	M11 0 0 0 0 0 1
M12 0 0 0 0 0 1 1 0 0	M13 0 0 0 0 0 1 1 0 1	M14 0 0 0 0 0 1 1 1 0	M15(Mod) 0 0 0 0 0 1 1 1 1	M16 0 0 0 0 1 0 0 0 0	M17 0 0 0 0 1 0 0 0 1	M1 0 0 0 1 0 1
M19 0 0 0 0 1 0 0 1 1	M21 0 0 0 0 1 0 1 0 1	M23 0 0 0 0 1 0 1 1 1	M26 0 0 0 0 1 1 0 1 0	M27 (ID) 0 0 0 0 1 1 0 1 1	M28 0 0 0 0 1 1 1 0 0	M2 0 0 0 1 1 0
M30 0 0 0 0 1 1 1 1 0	M40 0 0 0 1 0 1 0 0 0	M41 0 0 0 1 0 1 0 0 1	M42 0 0 0 1 0 1 0 1 0	M43 0 0 0 1 0 1 0 1 1	M45 0 0 0 1 0 1 1 0 1	M56(I) 0 0 1 1 0 0
M57 0 0 0 1 1 1 0 0 1	M58 0 0 0 1 1 1 0 1 0	M59 0 0 0 1 1 1 0 1 1	M61 0 0 0 1 1 1 1 0 1	M68 0 0 1 0 0 0 1 0 0	M69 0 0 1 0 0 0 1 0 1	M7 0 0 0 0 1 1
M78(XOR) 0 0 1 0 0 1 1 1 0	M84 0 0 1 0 1 0 1 0 0	M85 0 0 1 0 1 0 1 0 1	M86 0 0 1 0 1 0 1 1 0	M94 0 0 1 0 1 1 1 1 0	M97 0 0 1 1 0 0 0 0 1	M8 0 0 1 0 0 1
M99 0 0 1 1 0 0 0 1 1	M101 0 0 1 1 0 0 1 0 1	M106 0 0 1 1 0 1 0 1 0	M108 0 0 1 1 0 1 1 0 0	M113 0 0 1 1 1 0 0 0 1	M114 0 0 1 1 1 0 0 1 0	M9 0 1 1 0 0 1
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.

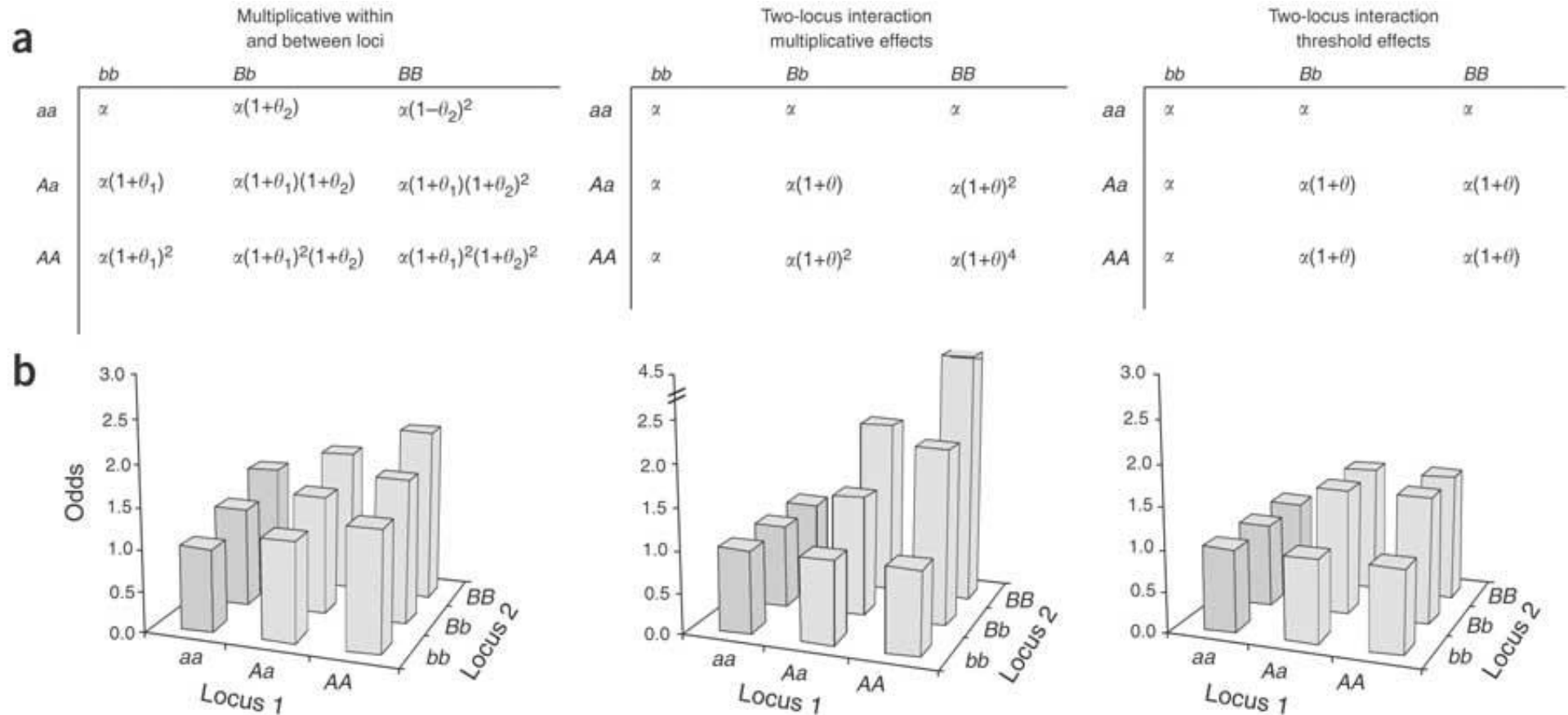
Different degrees of epistasis



(slide: Motsinger)

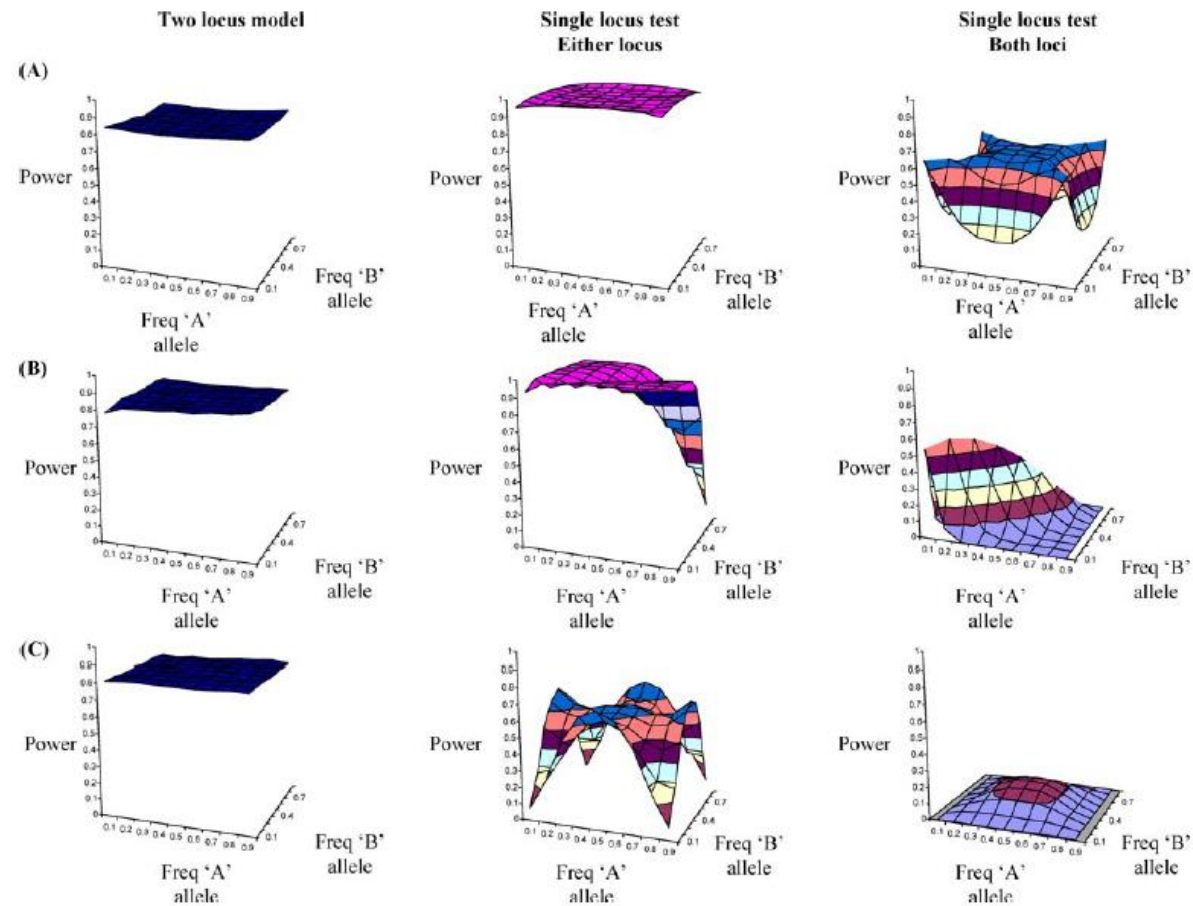
Incomplete penetrances

- Odds of disease for 2 loci under epistatic scenarios



(Marchini et al. 2005)

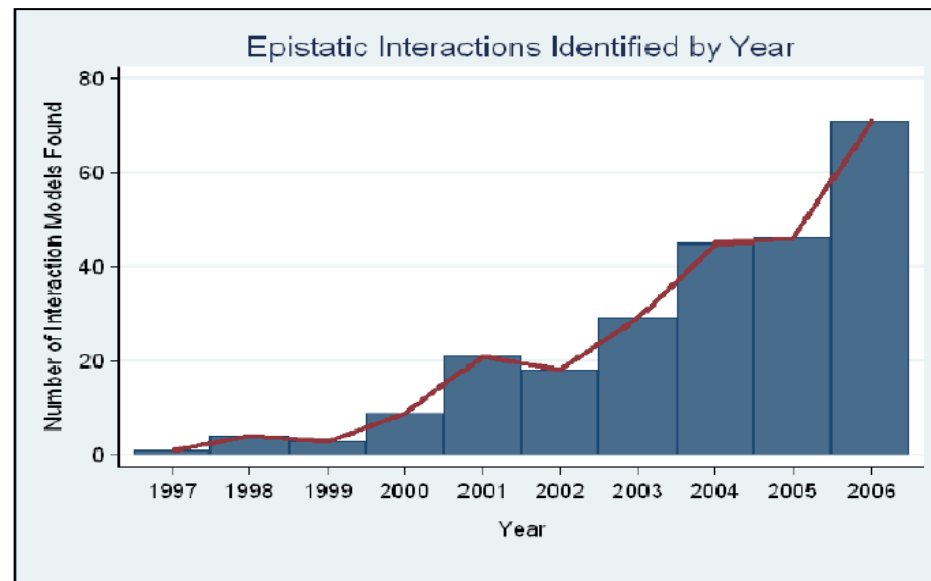
Power to Detect Association for 1,500 Individuals where Both Loci Are Responsible for 5% of the Trait Variance



(Evans et al 2006; A: no, B: M27, C: M16)

A growing toolbox

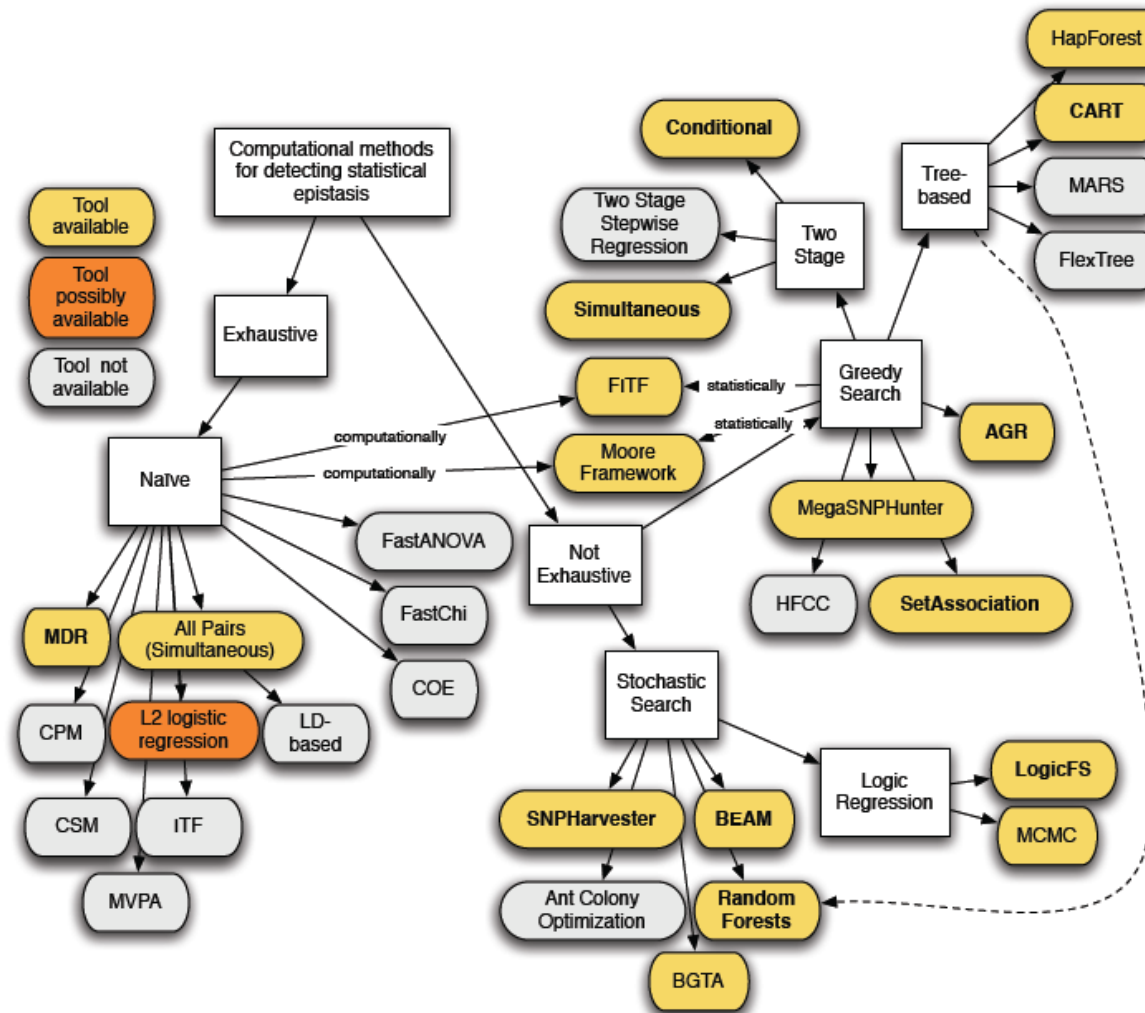
- The number of identified epistasis effects in humans, showing susceptibility to common complex human diseases, follows a steady growth curve (Emily et al 2009, Wu et al 2010), due to the growing number of toolbox methods and approaches.



(Motsinger et al. 2007)

Selection an epistasis detection method

(Kilpatrick 2009)



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*

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”.

Type	Example	Note
Exhaustive epistasis analysis methods	Multifactor dimensionality reduction (MDR, [59])	All possible interactions of the input variables When necessary, combined with variable reduction step, which may (cf. variable selection) or may not involve the phenotype of interest
	Model-based multifactor dimensionality reduction (MB-MDR, [48])	Non-parametric data mining method that aggregates multi-locus signals into 'risk' groups Semi-parametric data mining method that aggregates multilocus signals and orders them according to 'severity'
	(Penalized) Logistic regression [91–93], multivariate adaptive regression splines [94], adaptive group lasso [98], Mnets [95], partial least squares [96], Boolean operation-based screening and testing [97], interaction testing framework (ITF) [47] compositional epistasis [86–88], reconstructability analysis (RA, [105])	Parametric approach with regression-based foundation or overlap
Non-exhaustive epistasis analysis methods Greedy viewpoint	EPIBLASTER [106]	Contrasting measure of LD between markers Partial search among all possible interactions of the input variables Pre-select candidate interactions based on evidence for lower order effects
	Focused regression-based interaction screening approaches (thresholding combinations for interaction testing: focused interaction testing framework (FITF, [47])	
	Variable selection (filtering) followed-up by an exhaustive epistasis screening method	
Stochastic viewpoint	SNPHarvester [52]	Iteratively pre-select a subgroup of variables for full-blown epistasis analysis Interaction detection method merging ideas from k -means clustering and Markov chain Monte Carlo Decision tree-based methods
	Logic regression (LR) [35, 65, 107], MCMC logic regression [64], logic forest [68], random forests + MDR [50], random jungle (RJ), [51])	
	Bayesian epistasis association mapping (BEAM, [53])	Bayesian partitioning with posterior probabilities for epistatic markers

One popular method singled out

- 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.
- 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

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

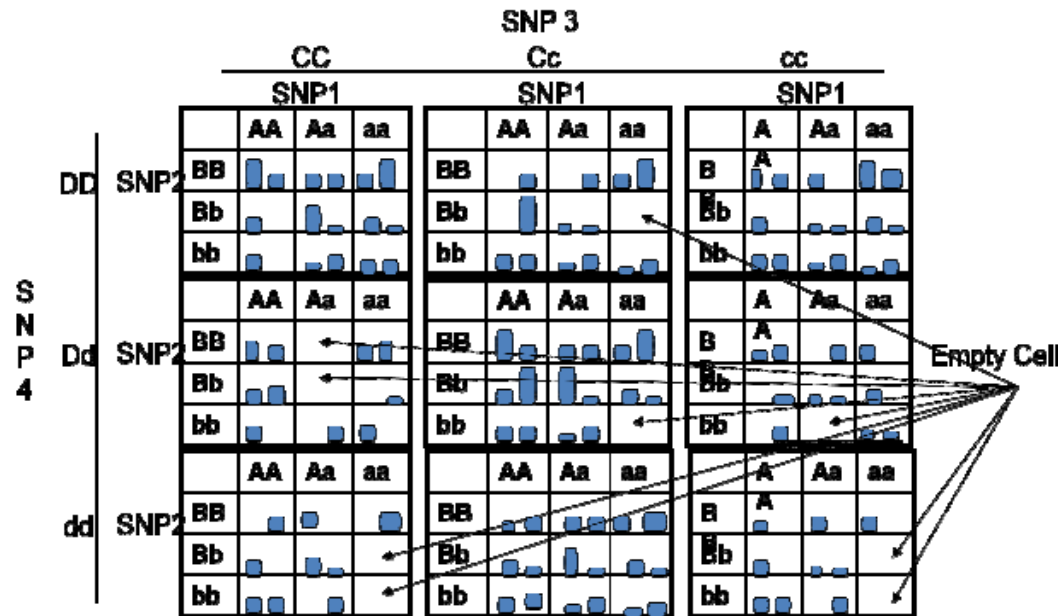
“Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders”

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 GWAs is needed”

A note aside

Missing data

Making most of available genotype information

- The idea is that data on a modest set of genetic variants measured in a number of related individuals can provide useful information about other genetic variants in those individuals
- This forms the theoretical underpinning of both genetic linkage mapping in pedigrees and haplotype mapping in founder populations.
- Genetic linkage implies that family members who share a region IBD will be more similar to each other than will family members with the same degree of relatedness who do not share the region IBD.

(Lander and Schork 1994 ; de la Chapelle and Wright 1998)

Making most of available genotype information

- In traditional genetic linkage and founder haplotype mapping studies:
 - Long stretches of shared chromosome inherited from a relatively recent common ancestor
- In GWAs with (apparently) unrelated individuals:
 - Relatively short stretches of shared chromosomes
- However, genotype imputation can use these short stretches to estimate with great precision the effects of many variants that are *not directly* genotyped

(Li et al 2009)

Causes for missing data

- Restricting to genotype data, missingness may be due to several reasons:
 - Quality of the genotyping
 - Limitations of current genotyping platforms and calling-algorithms:
 - Missing genotypes may not randomly distributed throughout the homozygous and heterozygous groups
 - Different coverage by different genotyping efforts
 - Relevant in the context of pooled data or data to be used for meta-analysis purposes

Causes for missing data

- Missingness may – in theory - be introduced via several mechanisms:
(Rubin taxonomy – Rubin 1976)
 - Missing completely at random
 - MCAR: missing data values are a simple random sample of all data values
 - Missing at random
 - MAR: the probability that an observation is missing depends on observed values but not on missing ones
 - Not missing at random
 - NMAR: the missingness depends on data that is not observed
 - Relies on unverifiable assumptions
 - PLINK is able to “test” whether or not genotypes are missing at random wrt the true (unobserved) genotype, based on the observed genotypes of nearby SNPs

Severity of having missing genotypes

- For single-SNP analyses, if a few genotypes are missing there is not much problem.
- For multipoint SNP analyses, missing data can be more problematic because many individuals might have one or more missing genotypes.

“Any bias in the missing data (e.g., different distributions in cases and controls or according to genotype groups) could create spurious results”

Solutions for dealing with missing genotypes

Imputation

- One convenient solution is data imputation
 - Data imputation involves replacing missing genotypes with predicted values that are based on the observed genotypes at neighboring SNPs (tightly linked markers).
- In the “early days” of addressing this problem, several studies on missing genotype data had been published, but many of these were family studies
- Authors such as Kistner & Weinberg (2004) used multiple imputation for missing genotype data but, since their studies consisted (partly) of related individuals, they adjusted the imputation method to avoid possible inconsistencies of imputed genotypes among family members.

(Souverein et al 2006)

Imputation

- In contrast, replacing missing genotypes with observed means or the most probable genotypes does not use linkage disequilibrium (LD) information from nearby markers, decreasing statistical efficiency and possibly causing bias.
- Estimation of missing genotypes can be a by-product of haplotype reconstruction, with either a maximum likelihood method implemented by the expectation maximization or Bayesian methods
 - While the maximum likelihood can lead to computer memory limitations, the Bayesian methods can take a longer time to converge. In both approaches, missing genotypes and missing phase are treated equivalently and inferred simultaneously.

Imputation

- Estimation of missing genotypes can also be achieved by inferring missing genotypes by iteratively estimating missing values and updating models that formulate the relationship between a SNP and its flanking markers
- Alternatively, parametric regression models or non-parametric (e.g., based on clustering, tree-building) methods are adopted, in order to select the relevant SNPs for imputation purposes
- Caution:
 - Population stratification
 - MNAR
 - Weak or strong LD between markers
 - Unrelated vs related individuals

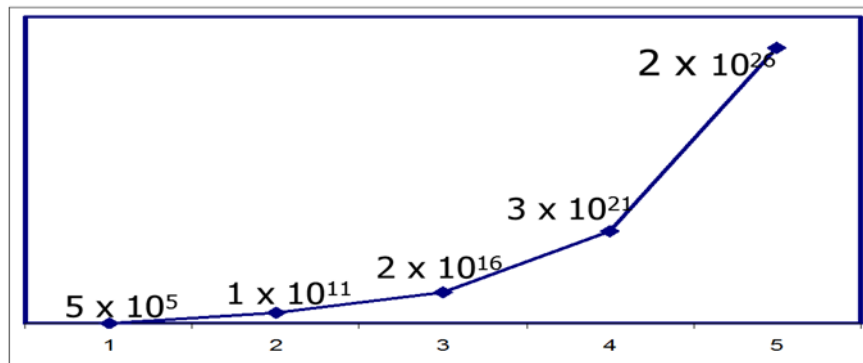
(Yu and Schaid 2007)

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 GWAs is hampered by undetected false positives”

Significance assessment

What is the general setting?

Introduction

- The genome is large and includes many polymorphic variants and many possible disease models, requiring a large number of tests to be performed.
- Any given variant (or set of variants) is highly unlikely, *a priori*, to be causally associated with any given phenotype under an assumed model, and strong evidence is required to overcome scepticism about an association.

(Balding 2006)

Introduction

- Even if a researcher only tests one SNP for one phenotype, if many other researchers do the same and the nominally significant associations are reported, there will be a problem of false positives.
- There is a need for statistical confidence measures associated with each discovery
- These may be stated in terms of:
 - P-values
 - False discovery rates
 - Q-values

Sources of multiple testing are multiple ...

Source	Example
Multiple outcomes	A cohort study looking at the incidence of breast cancer, colon cancer, and lung cancer
Multiple predictors	An observational study with 40 dietary predictors or a trial with 4 randomization groups
Subgroup analyses	A randomized trial that tests the efficacy of an intervention in 20 subgroups based on prognostic factors
Multiple definitions for the exposures and outcomes	An observational study where the data analyst tests multiple different definitions for "moderate drinking" (eg, 5 drinks per week, 1 drink per day, 1-2 drinks per day, etc.)
Multiple time points for the outcome (repeated measures)	A study where a walking test is administered at 1 month, 3 months, 6 months, and 1 year
Multiple looks at the data during sequential interim monitoring	A 2-year randomized trial where the efficacy of the treatment is evaluated by a Data Safety and Monitoring Board at 6 months, 1 year, and 18 months

(Sainani 2009)

The multiple testing problem further translated to GWAS

- Simultaneously test m null hypotheses, one for each SNP j
 H_{0j} : no association between SNP j and the trait
- Every statistical test comes with an inherent false positive, or type I error rate—which is equal to the threshold set for statistical significance, generally 0.05.
- However, this is just the error rate for one test. When more than one test is run, the overall type I error rate is much greater than 5%.

The multiple testing problem translated to GWAS

- Suppose 100 statistical tests are run when (1) there are no real effects and (2) these tests are independent, then the probability that no false positives occur in 100 tests is $0.95^{100} = 0.006$. So the probability that at least one false positive occurs is $1 - 0.006 = 0.994$ or 99.4%
- There is not a single measure to quantify false positives (Hochberg et al 1987).
- Several multiple testing corrections have been developed and curtailed to a genome-wide association context, when deemed necessary.

Measuring false-positives

- In general, false-positive controlling measures either control
 - the family-wise error rate (FWER), or the overall type I error rate,
 - the generalized family-wise error rate, $gFWER(k)$, the tail probability that the number of Type I errors exceeds a user-supplied integer k ,
 - the tail probability, $TPFP(q)$, that the proportion of Type I errors among the rejected hypotheses exceeds a user-supplied value $0 < q < 1$, and
 - the false discovery rate (FDR).

Measuring false-positives

- For discussions about the utility of the aforementioned and other multiple testing procedures [in genomics applications](#), we refer to Manly et al. (2004), Pollard et al (2004), Dudbridge et al (2006), Dudoit and van der Laan (2008), amongst others.

Measuring false-positives

	# non-rejected hypotheses	# rejected hypotheses	
# true null hypotheses (non-diff. genes)	U	V Type I error	m_0
# false null hypotheses (diff. genes)	T Type II error	S	m_1
	$m - R$	R	m

- $\text{FWER} = p(V \geq 1) \rightarrow$ Family-wise error rate
- $\text{FDR} = E(V/R) \rightarrow$ False discovery rate

Family-wise error rate (FWER)

$$\text{FWER} = p(V \geq 1)$$

- The frequentist paradigm of controlling the overall type-1 error rate sets a significance level α (often 5%), and states that all the tests that the investigator plans to conduct should together generate no more than probability α of a false positive.
- In complex study designs, which involve, for example, multiple stages and interim analyses, this can be difficult to implement
- Strong control of FWER at level α : FWER is upper-bounded by α regardless of the number of false null hypotheses ($m_1 > 0$)
- Weak control of FWER at level α : FWER is upper-bounded by α whenever all tested null hypotheses are true ($m_0 = m$)

The Bonferroni correction

- The most widely known multiple testing correction is the Bonferroni correction.
- If n SNPs are tested and the tests are approximately **independent**, the appropriate per-SNP significance level α' should satisfy

$$\alpha = 1 - (1 - \alpha')^n,$$

which leads to the Bonferroni correction $\alpha' \approx \alpha / n$.

- For example, to achieve $\alpha = 5\%$ over 1 million independent tests means that we must set $\alpha' = 5 \times 10^{-8}$.
- However, the *effective number* of independent tests in a genome-wide analysis depends on many factors, including sample size and the test that is carried out [see: Take-home messages – Part 7].

False discovery rate (FDR)

- The FDR refers to the proportion of false positive test results among all positives:
 - $FDR = E(V/R) \rightarrow$ What if no null hypotheses are rejected ($R=0$)?
 - $FDR = E[V/R \mid R>0] \cdot \text{prob}(R>0)$ (Benjamini and Hochberg 1995)
 - $pFDR = E[V/R \mid R>0]$ (Storey 2001)
- Hence, FDR measures come in different shapes and flavor.
 - Under the null hypothesis of no association, p -values should be uniformly distributed between 0 and 1;
 - FDR methods typically consider the actual distribution as a mixture of outcomes under the null (uniform distribution of p -values) and alternative (P -value distribution skewed towards zero) hypotheses.

FDR

- Rather than setting a fixed pFDR rate to control, Storey and colleagues suggest giving a value to each test that indicates what pFDR would result from declaring that test significant.
- The q-value associated with an individual test is defined as the minimum pFDR achieved when declaring all tests significant at the level of the test's pvalue.
- A q-value can be estimated for each test in a genome-wide experiment and follow-up tests selected from those with the lowest q-values.

Do these classical methods hold up in GWA settings?

Family-wise error rate (FWER) control using Bonferroni thresholds

- Bonferroni Threshold in the context of GWAS: $< 10^{-7}$, $< 10^{-8}$
- In the presence of too many tests, the Bonferroni threshold will be extremely low
- Moreover,
 - Bonferroni adjustments are conservative when statistical tests are not independent
 - Bonferroni adjustments control the error rate associated with the omnibus null hypothesis
 - The interpretation of a finding depends on how many statistical tests were performed

Do these classical methods hold up in GWA settings?

- FDR and variations thereof
 - Start to break down in GWAS settings with complex LD dependencies between markers (and therefore complex dependencies between test statistics)
 - The power over Bonferroni is minimal, especially when multiple signals are assumed to be present in the data and the aim is to identify most (if not all) of them in a single analysis run (e.g., Van Steen et al 2005)

Other popular ways to control false positives in GWA settings

FDR in Bayesian terms

- Suppose m **identical** hypothesis tests are performed with independent statistics T_1, \dots, T_m and rejection area C .
- Suppose that a null hypothesis is true with a-priori probability $\pi_0 = \text{Prob}(H = 0)$.
- Then

$$pFDR(C) = \frac{\pi_0 \cdot \text{Prob}(T \in C | H = 0)}{\text{Prob}(T \in C)} = \text{Prob}(H = 0 | T \in C).$$

using Bayes rule

(Storey 2001)

Bayesian methods

- The usual frequentist approach to multiple testing has a serious drawback in that researchers might be discouraged from carrying out additional analyses beyond single-SNP tests (read: epistasis screening)
- It is a matter of common sense that expensive and hard-won data should be investigated exhaustively for possible patterns of association.
- Under the Bayesian approach, there is no penalty for analysing data exhaustively because the prior probability of an association should not be affected by what tests the investigator chooses to carry out.

The false-positive report probability (FPFP)

- A further difficulty with FDR is that it says little about the individual tests. The most significant tests are most likely to be the true positives, but FDR and q-values ignore this in favour of averaging the error rate across all significant tests.
- The local FDR is computed as follows:

$$\frac{\pi_0 f_0(T)}{\pi_0 f_0(T) + (1 - \pi_0) f_1(T)}$$

where π_0 is the prior probability that the null hypothesis is true, T is the test statistic, and f_0 and f_1 are the probability densities of T under the null hypothesis and alternative hypothesis, respectively

(Efron et al 2001, 2002)

The false-positive report probability (FPFP)

(Wacholder et al 2004)

- The FPFP is the posterior probability that a null hypothesis is true, given a statistic at least as extreme as that observed
- It is defined as

$$\frac{\pi_0 F_0(T)}{\pi_0 F_0(T) + (1 - \pi_0) F_1(T)}$$

where now F_0 and F_1 are the cumulative distributions.

- For known π_0 and F_1 and large number of multiple tests, it can be shown that the FPRP is the same as the q-value, the main difference being one of context.

Do these methods hold up in GWA settings?

Bayesian methods

- Bayesian methods are believed to play an increasing role in genetic association analyses ... provided these methods can be made more accessible to a wider audience

Other popular ways to control false positives in GWA settings

Permutation-based control

- In **samples of unrelated individuals**, one simply swaps labels (assuming that individuals are interchangeable under the null) to provide a new dataset sampled under the null hypothesis.
 - Note that only the phenotype-genotype relationship is destroyed by permutation: the patterns of LD between SNPs will remain the same under the observed and permuted samples.

Permutation-based control

- For **family data**, it might be better (or in the case of affected-only designs such as the TDT, necessary) to perform gene-dropping permutation instead. In its most simple form this just involves flipping which allele is transmitted from parent to offspring with 50:50 probability.
 - This approach can extend to general pedigrees also, dropping genes from founders down the generations.

Permutation- based control

- Two sets of empirical significance values can then be calculated:
 - Pointwise estimates of an individual SNP's significance
 - A value that controls for the fact that thousands of other SNPs were tested, while comparing each observed test statistic against the maximum of all permuted statistics (i.e. over all SNPs) for each single replicate.
 - The p-value now controls the FWER, as the p-value reflects the chance of seeing a test statistic this large, given you've performed as many tests as you have.

Permutation- based control

- The accuracy of the permutation test can be improved by noting that the minimum p-value, sum statistic and truncated product can all be regarded as the extreme value of a large number of observations (Dudbridge et al 2004).
- Therefore, they should follow the extreme value distribution (Coles 2001) and by fitting the parameters of the distribution to the values observed in permutation replicates, more accurate significance levels are obtained.
- Equivalently, fewer replicates are needed to reach a given accuracy.

Do these classical methods hold up in GWA settings?

Permutation-based control

- The permutation method is conceptually simple but can be computationally demanding, particularly as it is specific to a particular data set and the whole procedure has to be repeated if other data are considered
- Particularly handy for rare genotypes, small studies, non-normal phenotypes, and tightly linked markers
 - In case-control data this is relatively straightforward
 - In family data this is not at all an easy task ... (see before)

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
- Methods that accommodate correlated hypothesis tests (e.g., due to LD structure between genetic variants) include:
 - applying a Bonferroni correction using effective sample size derived from principal components (Nyholt et al 2004, Moskvina et al 2008),
 - exploiting haplotype blocking algorithms (Nicodemus et al 2005),

Take-home messages (cnt-ed)

- adopting a framework for hidden Markov Model-dependent hypothesis testing (Sun and Cai 2009, Wei et al 2009).
- 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)

- The field is not yet saturated with time-efficient false-positive controlling methods.
- New promising tools, even in the presence of millions of correlated markers, are emerging as we speak, claiming to be as accurate as permutation-based testing.
 - One of these methods is 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)
 - Another one is PACT (**P** values **A**ddjusted for **C**orrelated **T**ests) (Conneely and Boehnke 2007)

How to compare 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

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

- 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”.

Towards alternative approaches

- The golden question:

*To what extent do methods based on
multifactor dimensionality reduction
accommodate the aforementioned issues?*

Interpretation

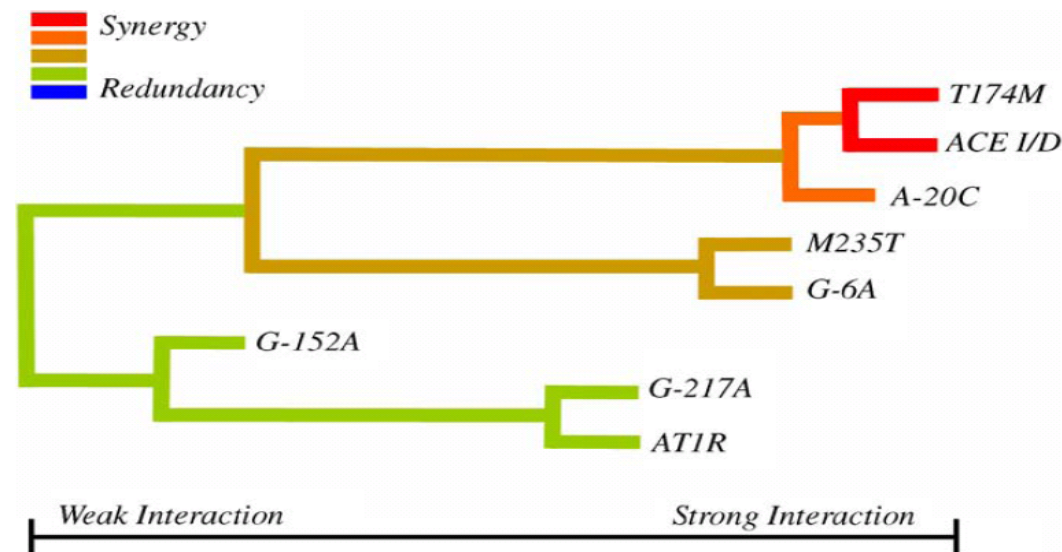
A flexible framework for analysis acknowledging interpretation capability

- 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.

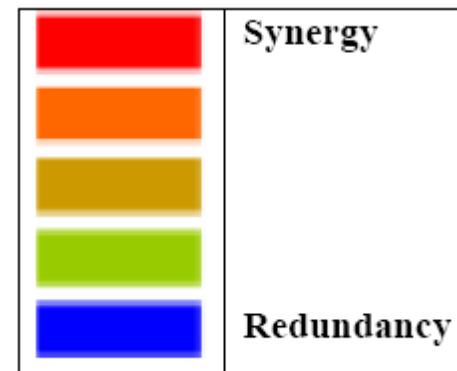


Interaction dendrogram

- 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.
- 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.

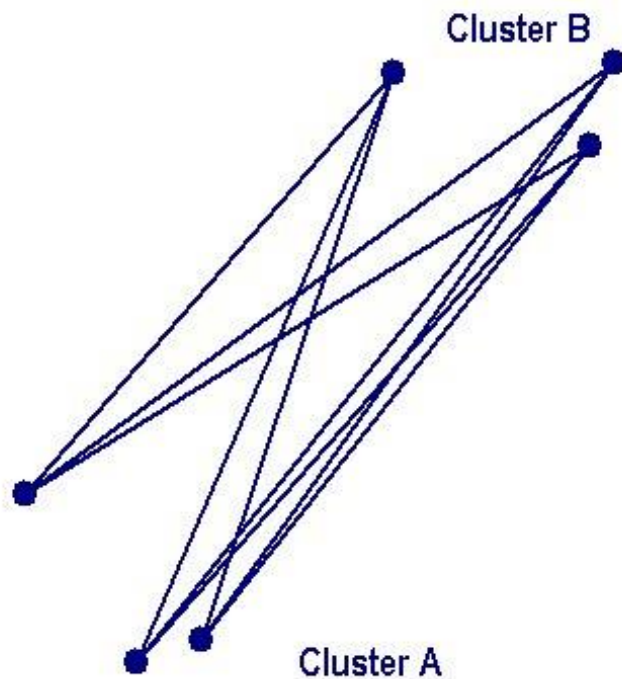
Synergy – The interaction between two attributes provides more information than the sum of the individual attributes.

Redundancy – The interaction between attributes provides redundant information.



Hierarchical clustering with average linkage

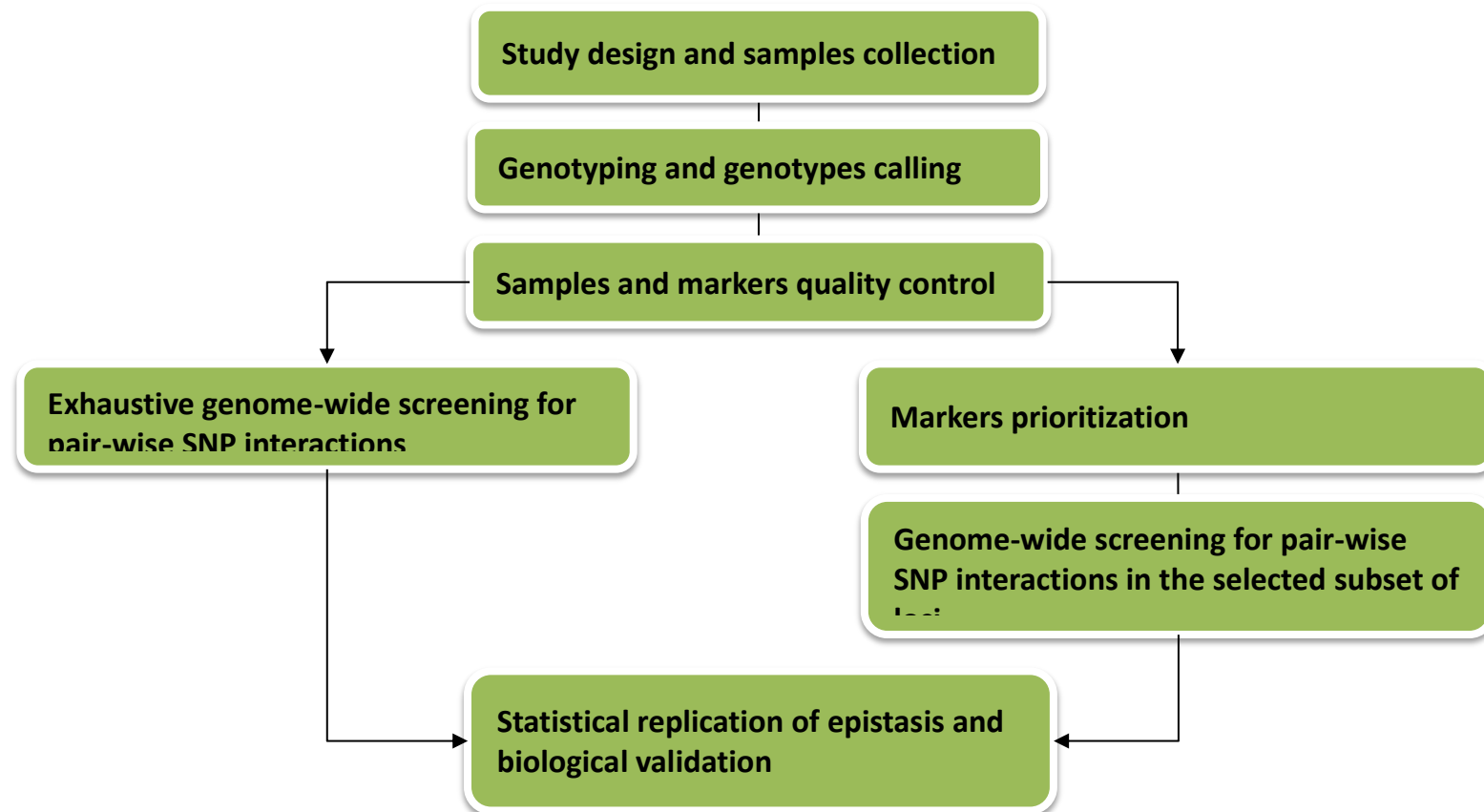
- Recall, here the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group



- 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)

Data Integration: a solution?!

- Where in 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 losing 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?

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!”

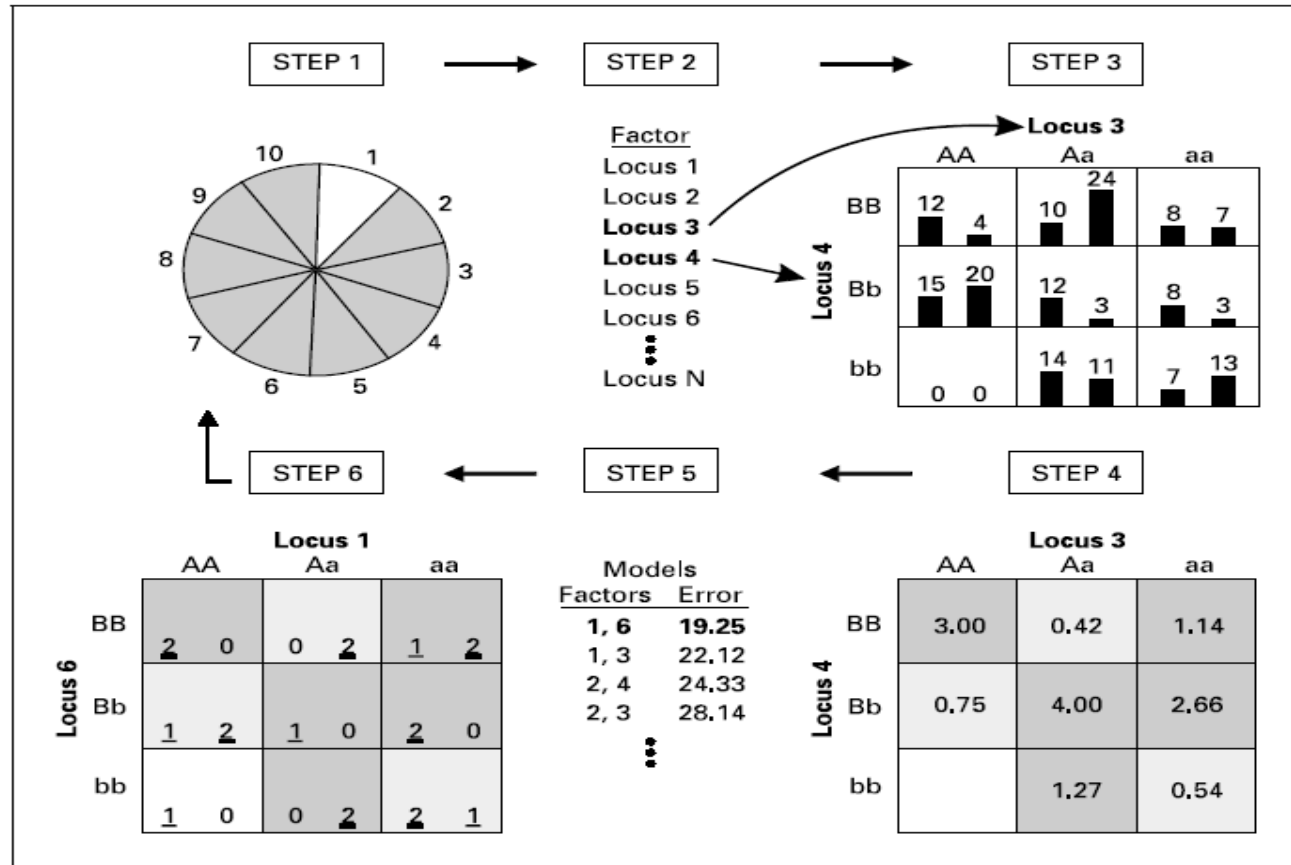
Model-Based Multifactor Dimensionality Reduction

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

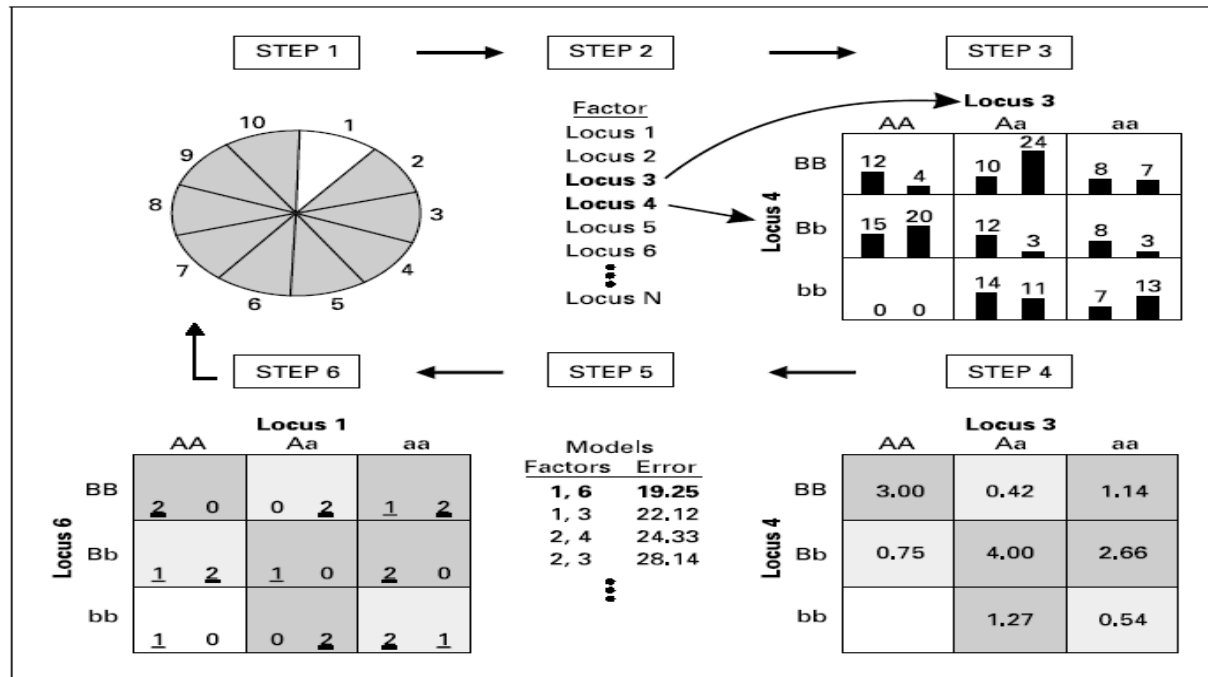
- Start: Multifactor Dimensionality Reduction by MD Ritchie et al (2001)



A note aside

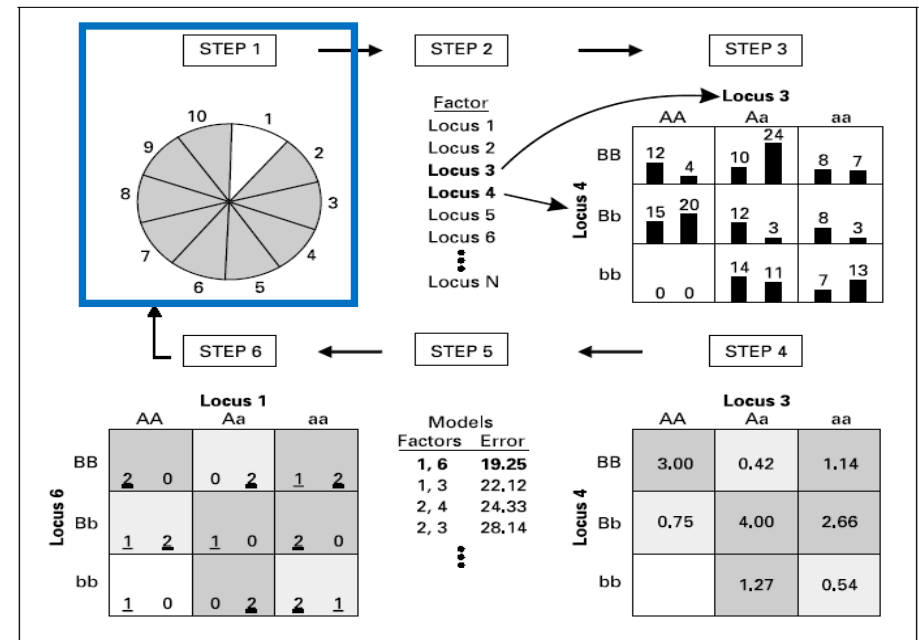
Multifactor Dimensionality Reduction (MDR)

The 6 steps of MDR



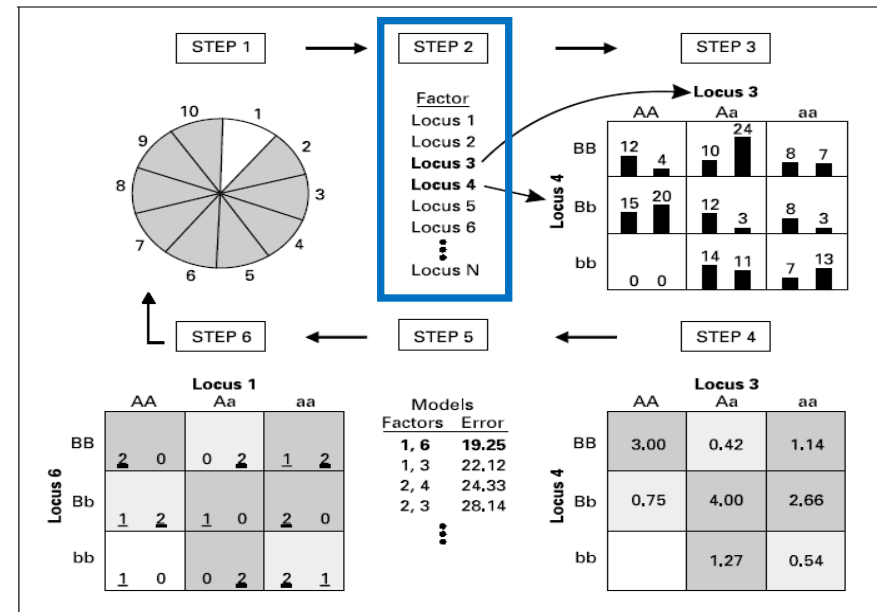
MDR Step 1

- Divide data (genotypes, discrete environmental factors, and affectation status) into 10 distinct subsets



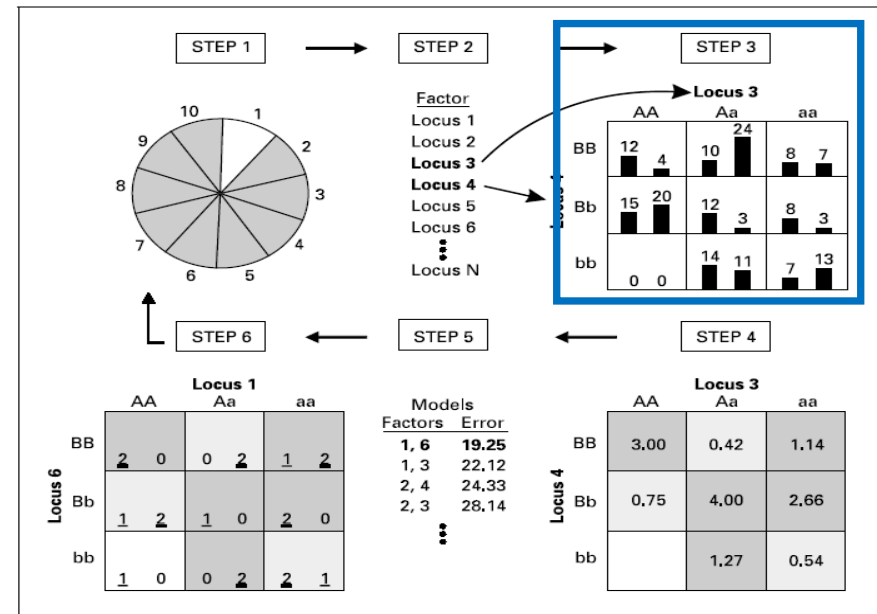
MDR Step 2

- Select a set of k genetic or environmental factors (which are suspected of epistasis together) from the set of all variables (N) in the training set



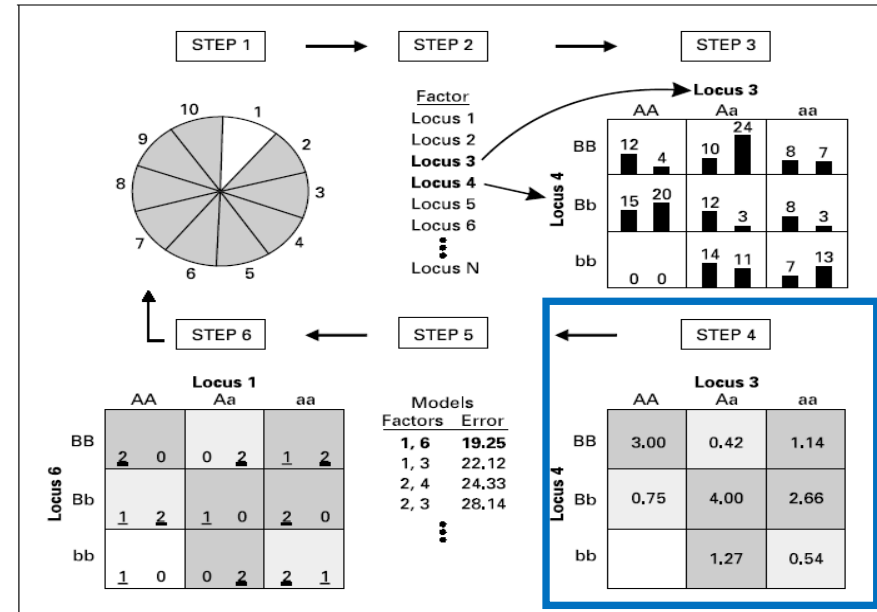
MDR Step 3

- Create a contingency table for these multi-locus genotypes, counting the number of affected and unaffected individuals with each multi-locus genotype



MDR Step 4

- Calculate the ratio of cases to controls for each multi-locus genotype
- Label each multi-locus genotype as “high-risk” or “low-risk”, depending on whether the case-control ratio is above a certain threshold
- This is the dimensionality reduction step:

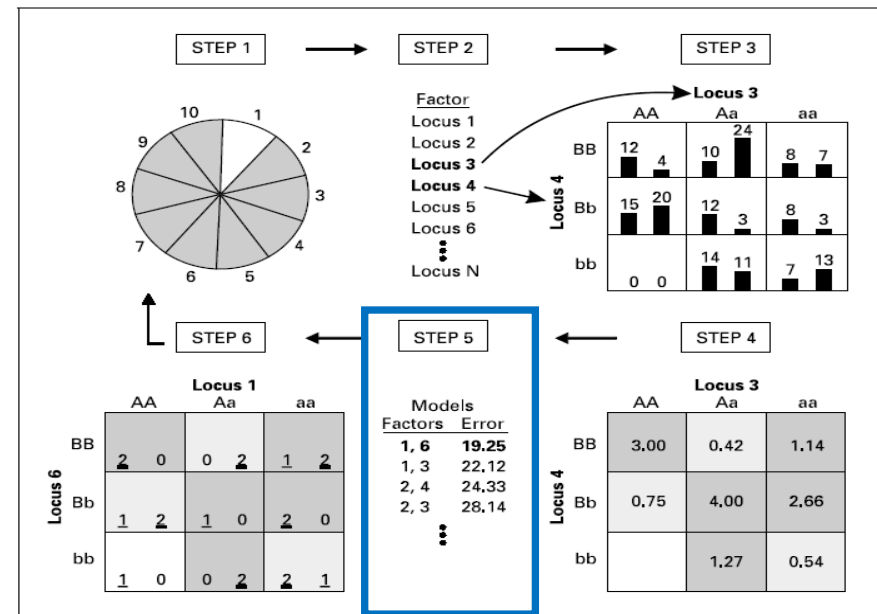


Reduces k -dimensional space to 1 dimension with 2 levels

MDR Step 5

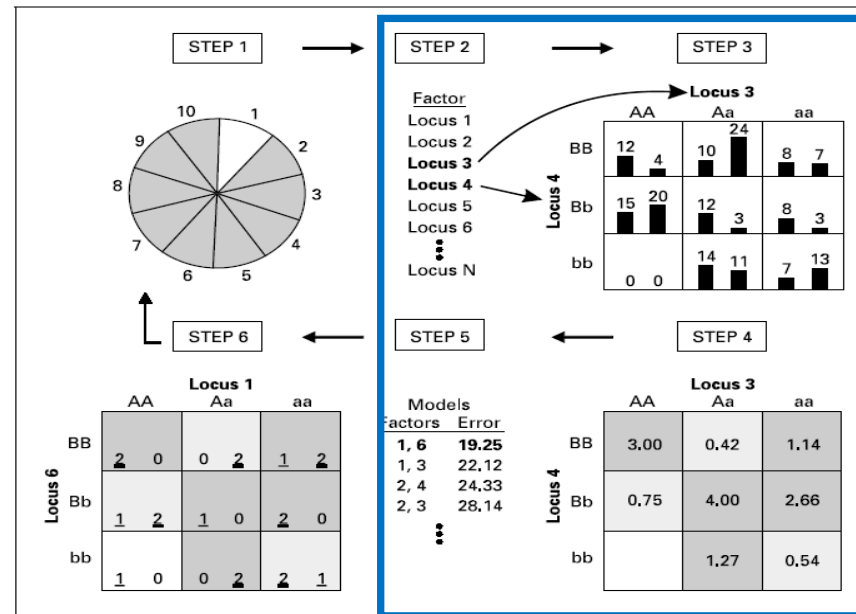
- To evaluate the developed model in Step 4, use labels to classify individuals as cases or controls, and calculate the misclassification error
- In fact: balanced accuracy are preferred (arithmetic mean between sensitivity and specificity), which IS mathematically equivalent to

classification accuracy when data are balanced



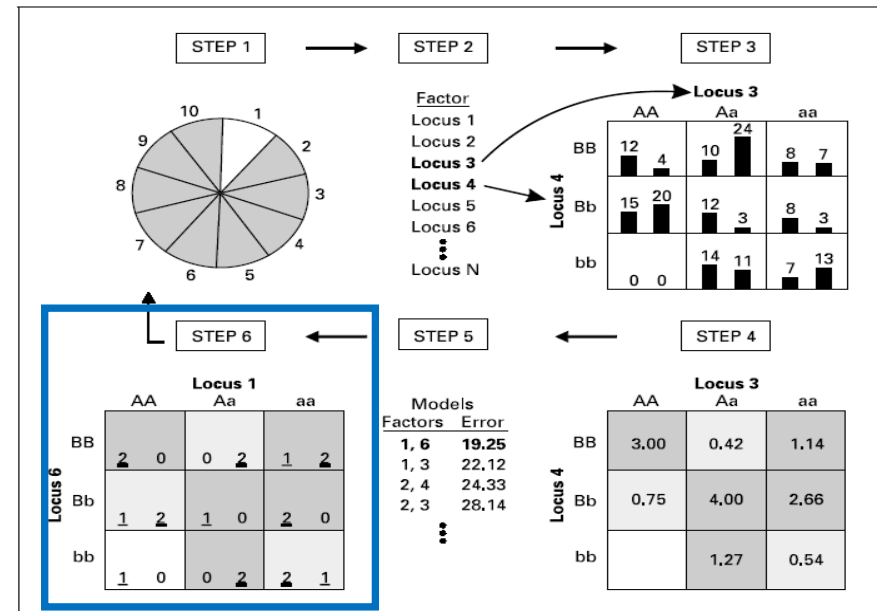
Repeat Steps 2 to 5

- All possible combinations of k factors are evaluated sequentially for their ability to classify affected and unaffected individuals in the training data, and the best k -factor model is selected in terms of minimal misclassification error



MDR Step 6

- The independent test data from the cross-validation are used to estimate the prediction error (testing accuracy) of the best k -order model selected



- **Towards final MDR:**
Repeat steps 1-6

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 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 $CVC < 10$
 - permutation-based null distribution (no association) !!!
 - Use accuracy measures that are not biased by the larger class
 - Use a threshold that is driven by the data at hand and naturally reflects the disproportion in cases and controls in the data

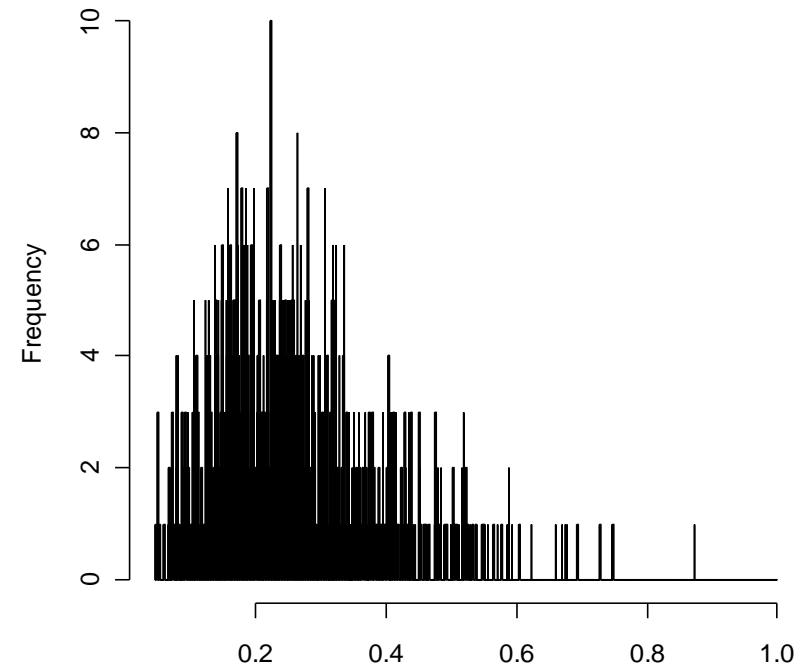
Hypothesis test of best model

- 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

Sample Quantiles

0%	0.045754
25%	0.168814
50%	0.237763
75%	0.321027
90%	0.423336
95%	0.489813
99%	0.623899
99.99%	0.872345
100%	1

An Example Empirical Distribution



The probability that we would see results as, or more, extreme than for instance 0.4500, simply by chance, is between 5% and 10%

(slide: L Mustavich)

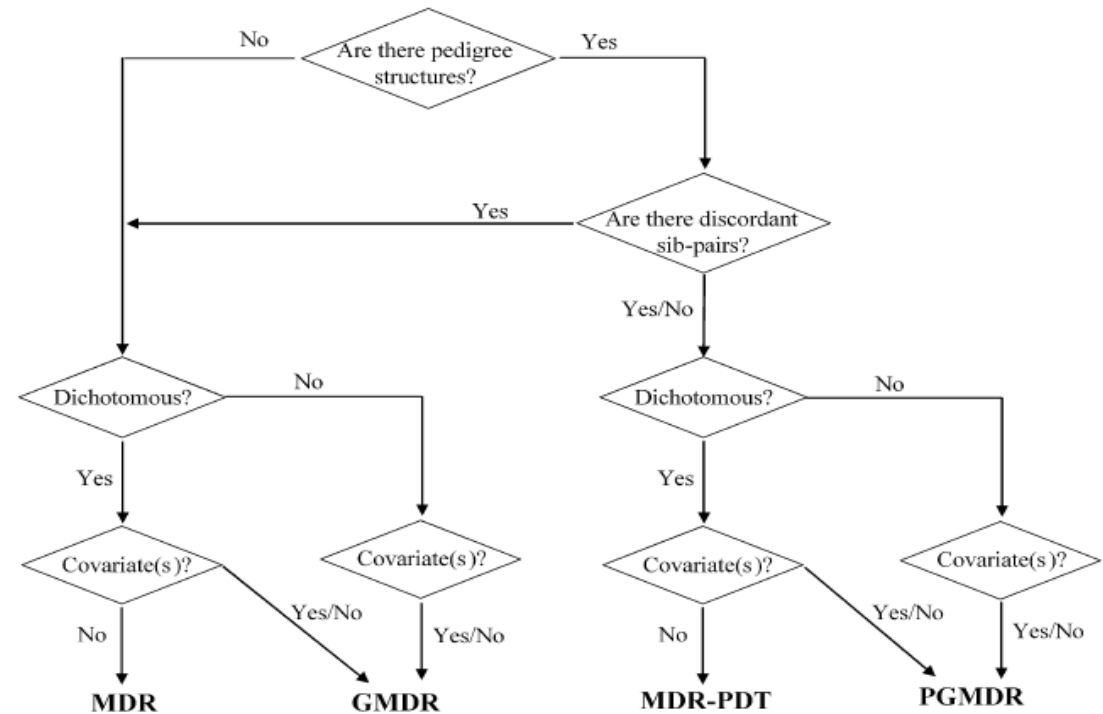
The MDR Software

- 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.multifactor dimensionality reduction.org/>

Historical notes about MB-MDR (cnt-ed)

- 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



Historical notes about MB-MDR

- 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”

Historical notes about MB-MDR

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

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

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

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

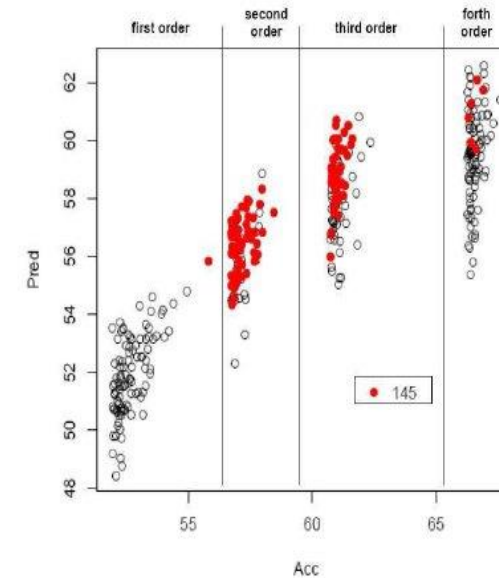
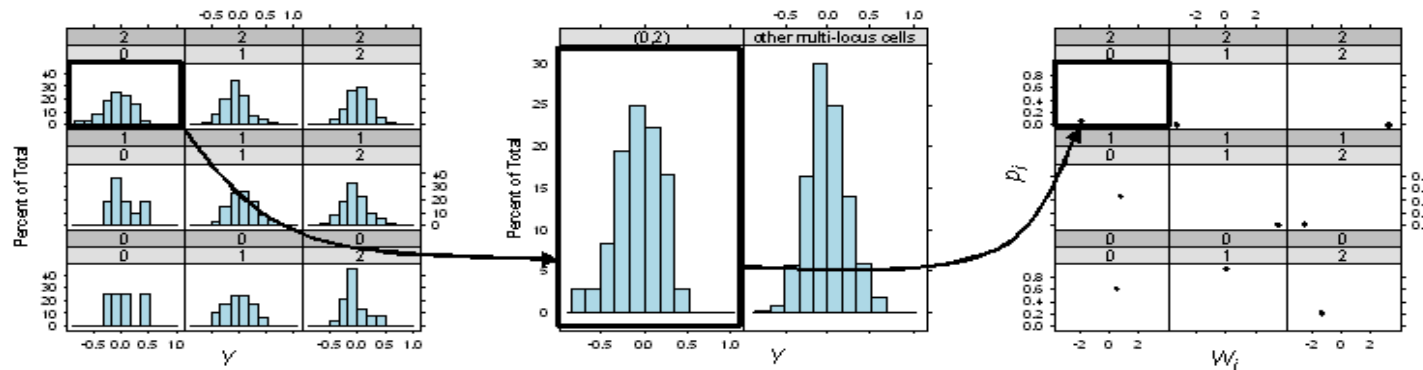


Fig. 1. Average Balanced Training accuracy (Acc) versus Average Balanced Predictive accuracy (Pred) for the 100 models with higher balanced training accuracy for the whole sample. First, second, third and fourth order interactions are considered.

Historical notes about MB-MDR

- 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), rather than Wald tests, and relying on MAF dependent reference distributions (Calle et al 2008)

Historical notes about MB-MDR

- Model-Based MDR by Cattaert et al (2011) – genetic heterogeneity

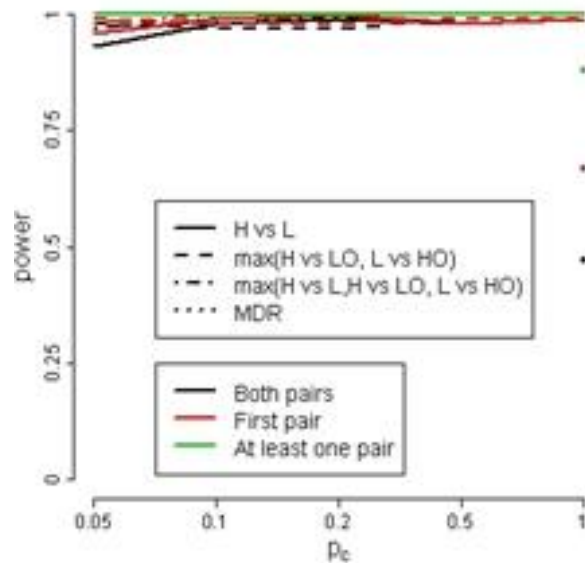
Model 2, $p = 0.5$

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

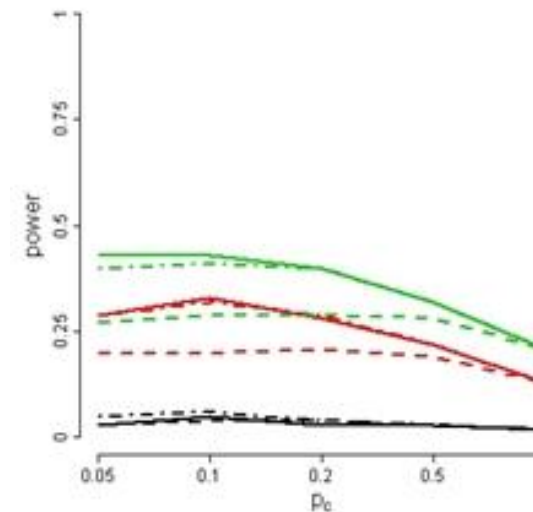
Model 6, $p = 0.1$

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

Ritchie Model 2 ($p=0.5$)

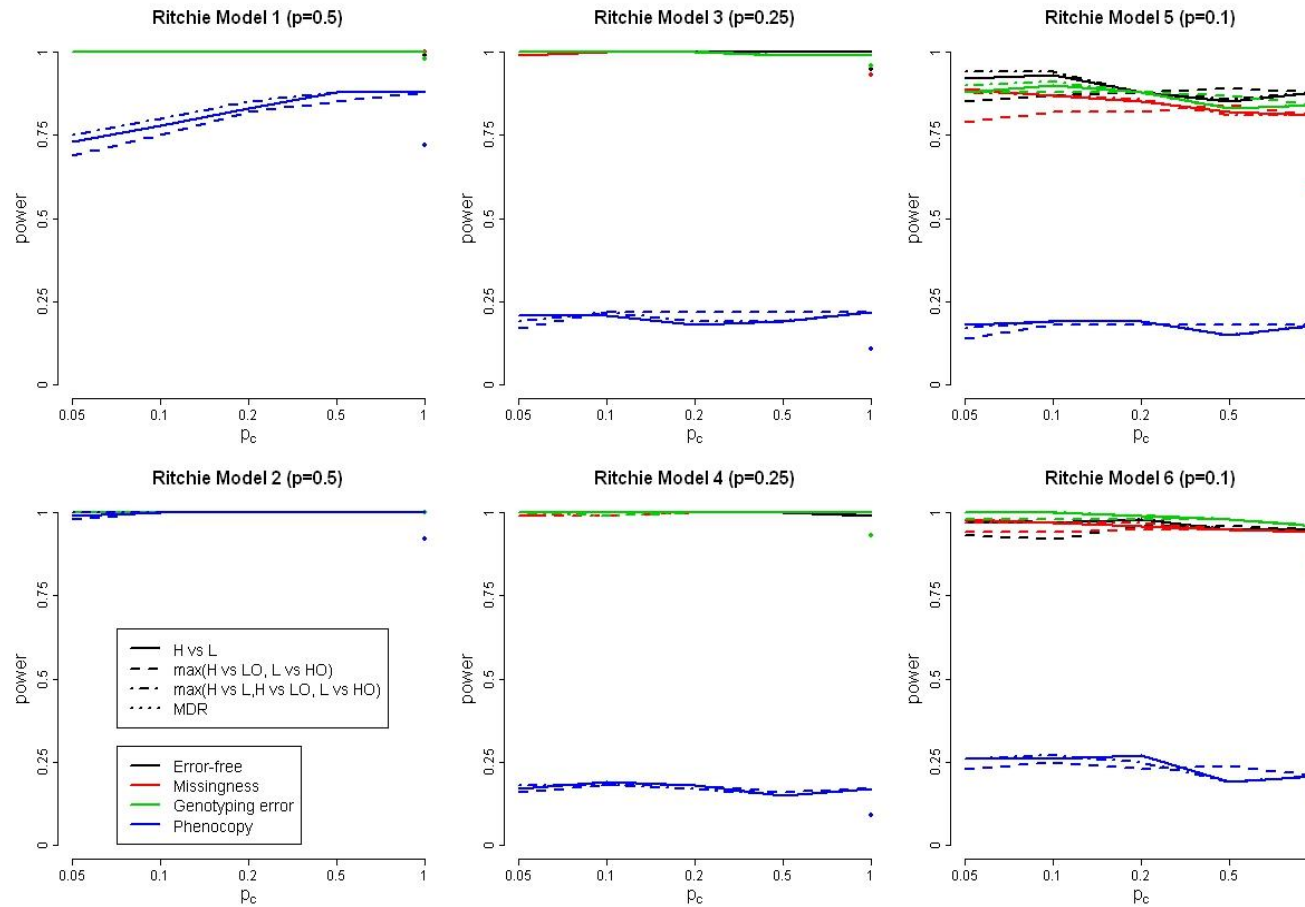


Ritchie Model 6 ($p=0.1$)



Historical notes about MB-MDR

- Model-Based MDR by Cattaert et al (2011) – maximal power



Historical notes about MB-MDR

- Model-Based MDR by Van Lishout et al (2012 – under review) – speed
 - MaxT algorithm ✓
 - Association test statistics (parametric and non-parametric) ✓ +

SNPs	<i>MBMDR-3.0.2</i> sequential execution Binary trait	<i>MBMDR-3.0.2</i> sequential execution Continuous trait	<i>MBMDR-3.0.2</i> parallel workflow Binary trait	<i>MBMDR-3.0.2</i> parallel workflow Continuous trait
100	45 sec	1 min 35 sec	<1sec	<1sec
1,000	1 hour 16 minutes	2 hours 39 minutes	38 sec	1 min 17 sec
10,000	5 days 13 hours	11 days 19 hours	1 hour 3 min	2 hours 14 min
100,000	≈ 1.5 year	≈ 3 years	4 days 9 hours	≈ 9 days

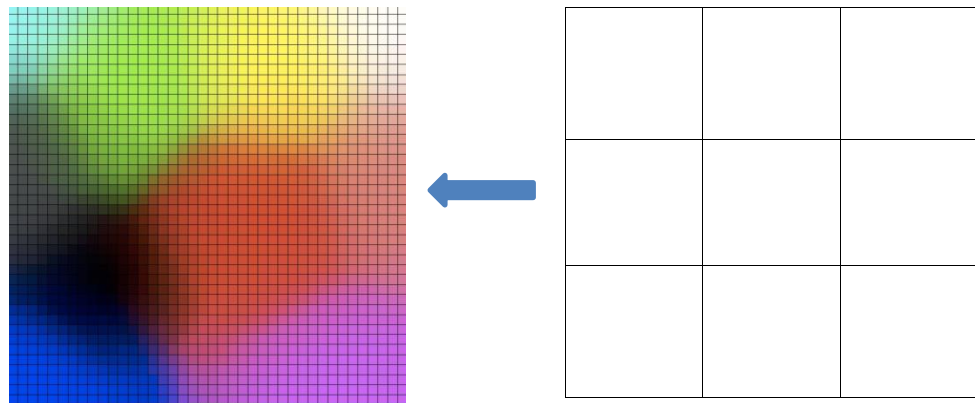
The parallel workflow was tested on a cluster composed of 10 blades, containing each four Quad-Core AMD Opteron(tm) Processor 2352 2.1 GHz.

The sequential executions were performed on a single core of this cluster.

The results prefixed by the symbol "≈" are extrapolated.

Historical notes about MB-MDR

- Model-Based MDR by Van Steen lab (2012 and +)
 - Lower order effects correction (omit at cell-labeling step) **v** **+**
 - Two-locus effect modifiers **v**
 - Different faces of “dimensions” in dimensionality reduction **+**



v: implemented

+: under construction or in beta-testing

Historical notes about MB-MDR

- Model-Based MDR by Van Steen lab (2012 and +)

**Human
Heredity**

Original Paper

Hum Hered 2004;58:82–92
DOI: [10.1159/000083029](https://doi.org/10.1159/000083029)

Received: June 30, 2004
Accepted after revision: September 23, 2004

MDR and PRP: A Comparison of Methods for High-Order Genotype-Phenotype Associations

L. Bastone^a M. Reilly^b D.J. Rader^b A.S. Foulkes^c

^aDivision of Biostatistics, ^bCardiovascular Division and Center for Experimental Therapeutics, University of Pennsylvania School of Medicine, Philadelphia, Pa., and ^cDepartment of Biostatistics, School of Public Health and Health Sciences, University of Massachusetts, Amherst, Mass., USA

Historical notes about MB-MDR

- Model-Based MDR by Van Steen lab (2012 and +)

Human
Heredit

Original Paper

Hum Hered 2004;58:82–92
DOI: [10.1159/000083029](https://doi.org/10.1159/000083029)

Received: June 30, 2004
Accepted after revision: September 23, 2004

MDR and PRP: A Comparison of Methods for High-Order Genotype-Phenotype Association

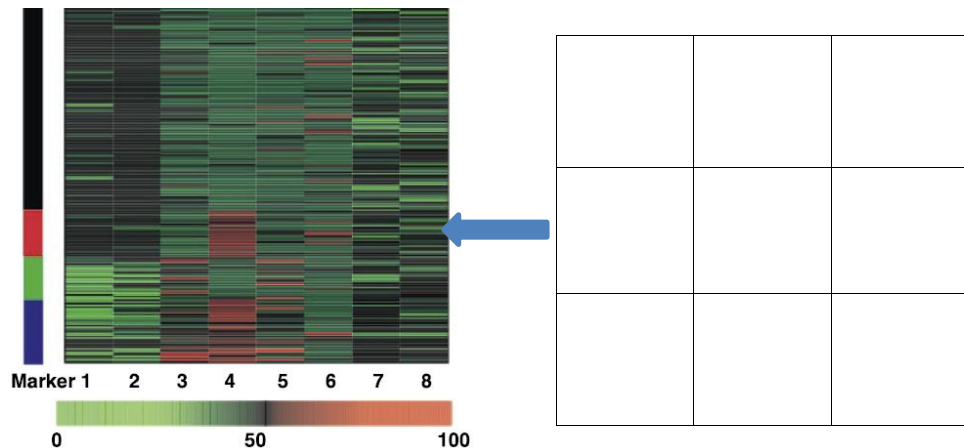
L. Bastone^a M. Reilly^b D.J. Rader^b A.S. Foulkes^c

^aDivision of Biostatistics, ^bCardiovascular Division and Center for Experimental Therapeutics, University of Pennsylvania School of Medicine, Philadelphia, Pa., and ^cDepartment of Biostatistics School of Public Health and Health Sciences, University of Massachusetts, Amherst, Mass., US

Statistical methods such as multifactor dimensionality reduction (MDR), the combinatorial partitioning method (CPM), recursive partitioning (RP), and patterning and recursive partitioning (PRP) are designed to uncover complex relationships without relying on a specific model for the interaction, and are therefore well-suited to this data setting. However, the theoretical overlap among these methods and their relative merits have not been well characterized. In this paper we demonstrate mathematically that MDR is a special case of RP.

Historical notes about MB-MDR

- Model-Based MDR by Van Steen lab (2012 and +)
 - Dimension (1,2) = (SNP1,SNP2) ✓
 - Dimension (1,2) = (SNP1, “categorized” continuous variable) ✓ +
 - Dimension (1,2) = (SNP1, genomic region with rare variants) +



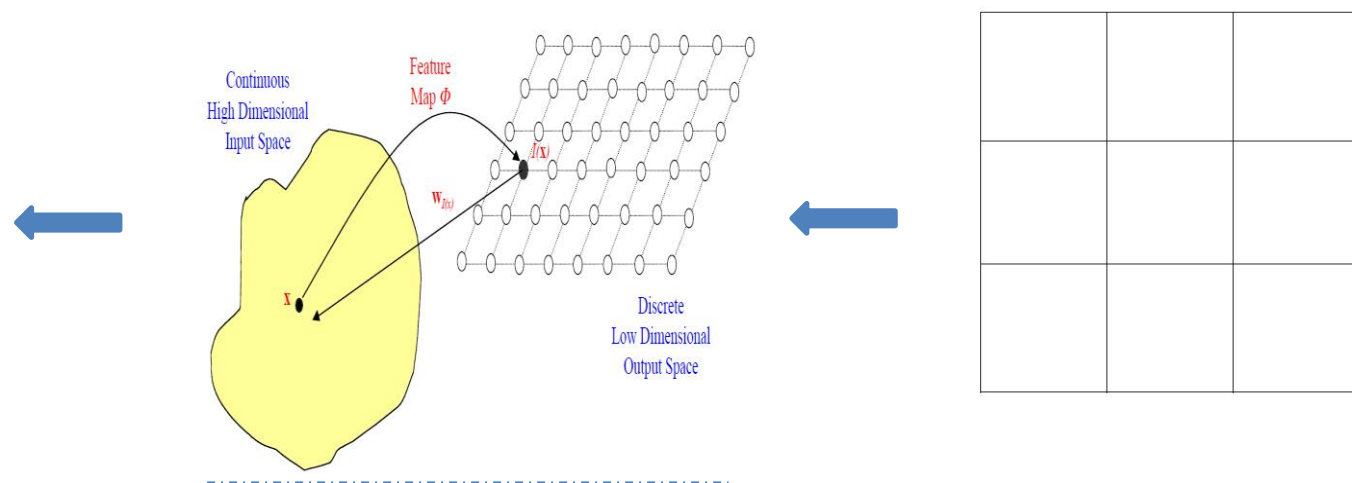
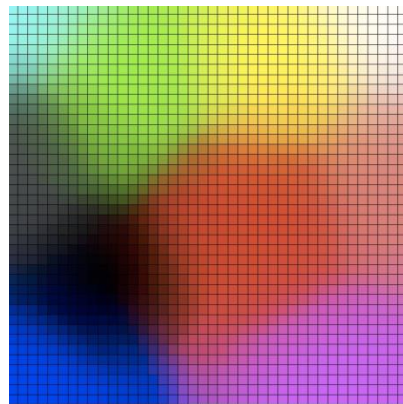
(Shi et al 2006, unsupervised clustering
with RFs)

✓: implemented

+ : under construction or in beta-testing

Historical notes about MB-MDR

- Model-Based MDR by Van Steen lab (2012 and +)
 - Dimension (1,2) = (pathway1, pathway2) +
 - Dimension (1,2) = +



OMs: Bullinaria 2004)

Key references about MB-MDR

Methodological papers

- **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**]
- **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 – under review [**recommendations on quantitative trait analysis**]
- **Van Lishout** F, Mahachie John JM, Gusareva ES, Urrea V, Cleyne I, Théâtre E, Charlotiaux B, Calle ML, Wehenkel L, Van Steen K (2012) An efficient algorithm to perform multiple testing in epistasis screening – under review [**C++ MB-MDR made faster!**]

Stay tuned for:

- + Applications of MB-MDR to screen for GxG interactions with a fixed Environmental or Genetic factor
- + Applications of MB-MDR to screen for genetic interactions involving genomic regions harboring rare variants
- + ... and much more!!!!

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

An example on Alzheimer's disease

First hurdle: Selection of most appropriate method

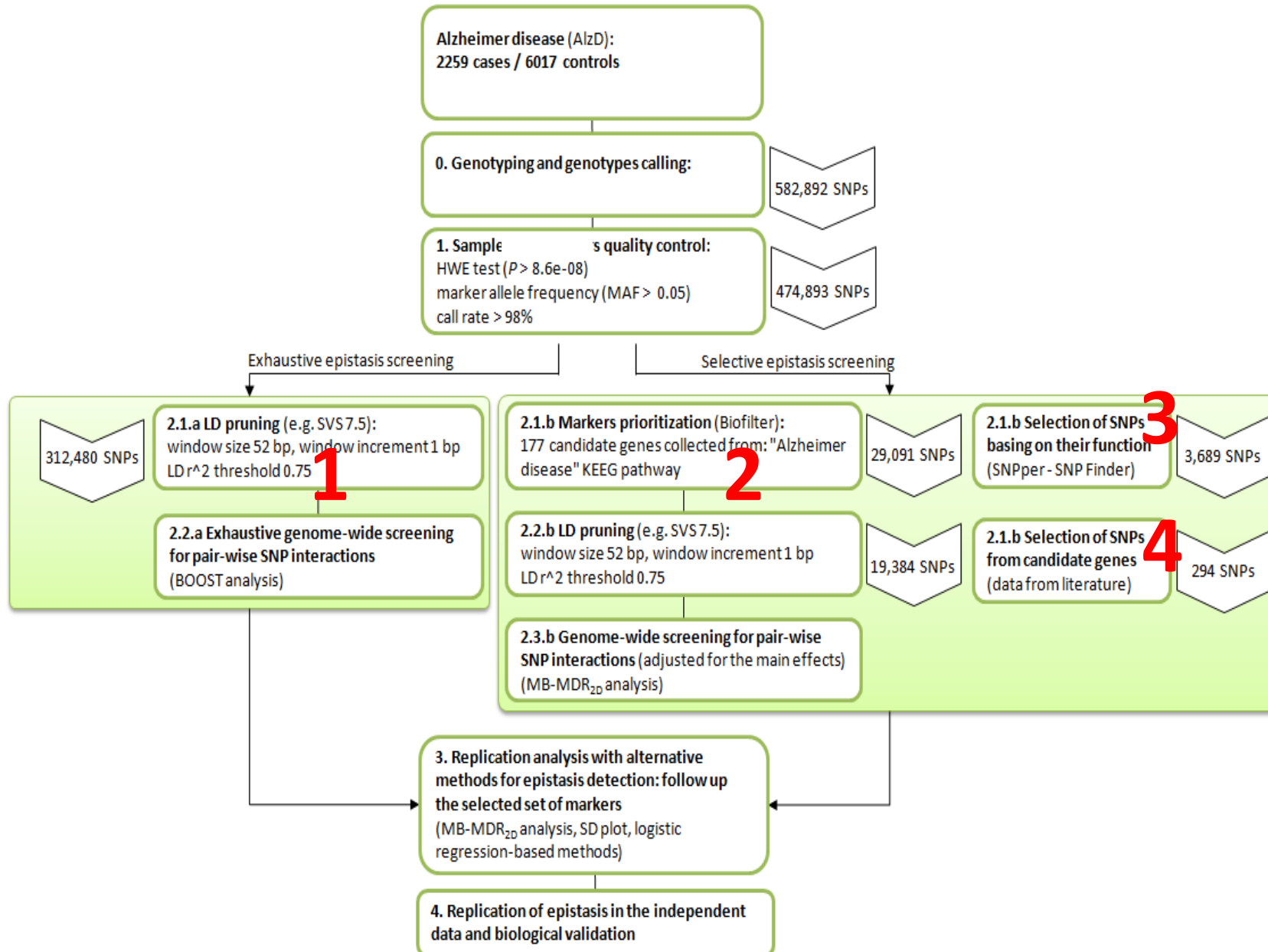
- Honest methods comparisons should / can highlight the “core” (**the ABC**) of each method:

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

Protocol for GWA analysis



Available “knowledge” about epistasis: Route 4

Gene	Gene name	Function	Location	Epistatic SNPs	Main effect for AlzD	Population (N cases/N controls)	Reference
<i>INS</i>	Insulin	Glucose metabolism	11p15.5	rs689	no	Germans (104/123)	Brune et al., 2003
<i>PPARA</i>	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs1800206	yes	Northern Europeans (336/2426)	Kölsch et al., 2012
<i>IL1A</i>	Interleukin 1 alfa	Inflammatory cytokine	2q13	rs3783550	no	Northern Europeans (336/2426)	Heun et al., 2012
<i>PPARA</i>	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs1800206	yes		
<i>IL1B</i>	Interleukin 1 beta	Inflammatory cytokine	2q13	rs16944	no	Northern Europeans (336/2426)	Heun et al., 2012
<i>PPARA</i>	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs1800206	yes		
<i>IL10</i>	Interleukin 10	Inflammatory cytokine	1q32.1	rs1800896	yes	Northern Europeans (336/2426)	Heun et al., 2012
<i>PPARA</i>	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs4253766	no		
<i>IL1A</i>	Interleukin 1 alfa	Inflammatory cytokine	2q13	rs1800587	no	Northern Europeans (336/2426)	Combarros et al., 2010
<i>DBH</i>	b-Hydroxylase	Onverts dopamine to norepinephrine in the synaptic vesicles of postganglionic sympathetic neurons	9q34.2	rs1611115	yes		
<i>TF</i>	Transferrin	Iron metabolism	3q22.1	rs1049296	no	UK (191/269)	Robson et al., 2004
<i>HFE</i>	Hemochromatosis		6p22.2	rs1800562	yes	Caucasians USA (1166/1404)	Kauwe et al., 2010
						North Europeans (336/2426)	Lehmann et al., 2012
<i>TF</i>	Transferrin	Iron metabolism	3q22.1	rs1130459	no	North Europeans (336/2426)	Lehmann et al., 2012
<i>HFE</i>	Hemochromatosis		6p22.2	rs1799945	yes		
<i>MTHFR</i>	Methylenetetrahydrofolate reductase	Homocysteine metabolism useful for normal brain functioning	1p36.22	rs1801131	yes	Indians (80/120)	Mansoori et al., 2012
<i>IL6</i>	Interleukin 6	Pro-inflammatory cytokine	7p15.3	rs1800795	no		
<i>IL10</i>	Interleukin 10	Limit inflammation in the brain	1q32.1	rs1800871	yes	North Spains (232/191)	Infante et al., 2004
<i>IL6</i>	Interleukin 6	Pro-inflammatory cytokine	7p15.3	rs2069837	yes	North Europeans (336/2426)	Combarros et al., 2009
<i>ABCA1</i>	ATP-binding cassette transporter A1	Intracellular cholesterol transport and maintance of cell cholesterol balance	9q31.1	rs2422493	no	Spanish (631/731)	Rodríguez-Rodríguez et al., 2010
<i>NPC1</i>	Niemann-Pick C1		18q11.2	rs18050810 rs4800488 rs2236707 rs2510344	no		

<i>LRP1</i>	low density lipoprotein receptor-related protein 1	Neuronal uptake of cholesterol	12q13.3	rs1799986	no	Spanish (246/237)	Vázquez-Higuera et al., 2009
<i>MAPT</i>	Microtubule-associated protein tau		17q21.33	rs2471738	no		
<i>GSK3B</i>	Glycogen synthase kinase-3 beta	Abnormal hyperphosphorylation of tau, neuronal uptake of cholesterol	3q13.33	rs334558	no	Spanish (246/237)	Vázquez-Higuera et al., 2009
<i>CDK5R1</i>	Cyclindependent kinase 5		17q11.2	rs735555			
<i>NR1H2</i>	Liver X receptor beta	Cholesterol metabolism	19q13.33	rs1052533 rs1405655	no	Spanish (414/442)	Infante et al., 2010
<i>HMOX1</i>	Heme oxygenase-1		22q12.3	rs2071746			

Different levels

- Genetic marker
- Locus
- Gene
- Window including either one of the previous
- Pathway

Revised analysis for candidate gene pairs

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

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

"+" - 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.

Replication and validation of GWAs: An impossible task?



(Mission Impossible @ google)

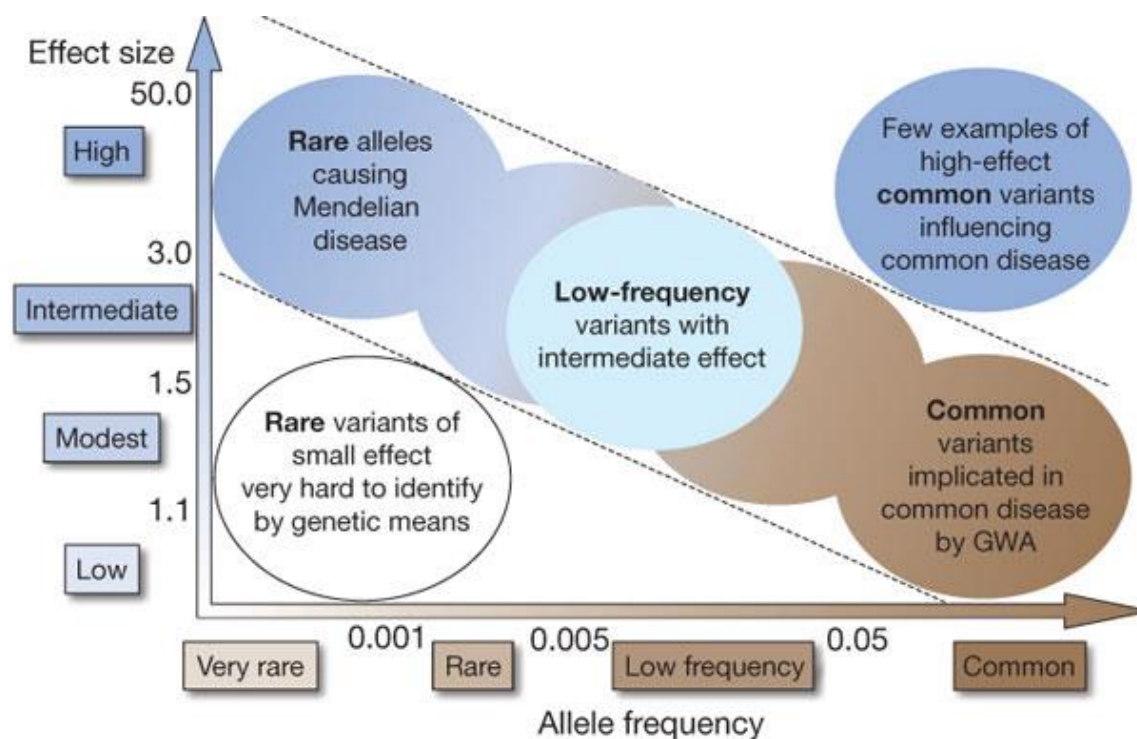
Replication

- 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

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

Optimal conditions for GWA (Interaction) replication

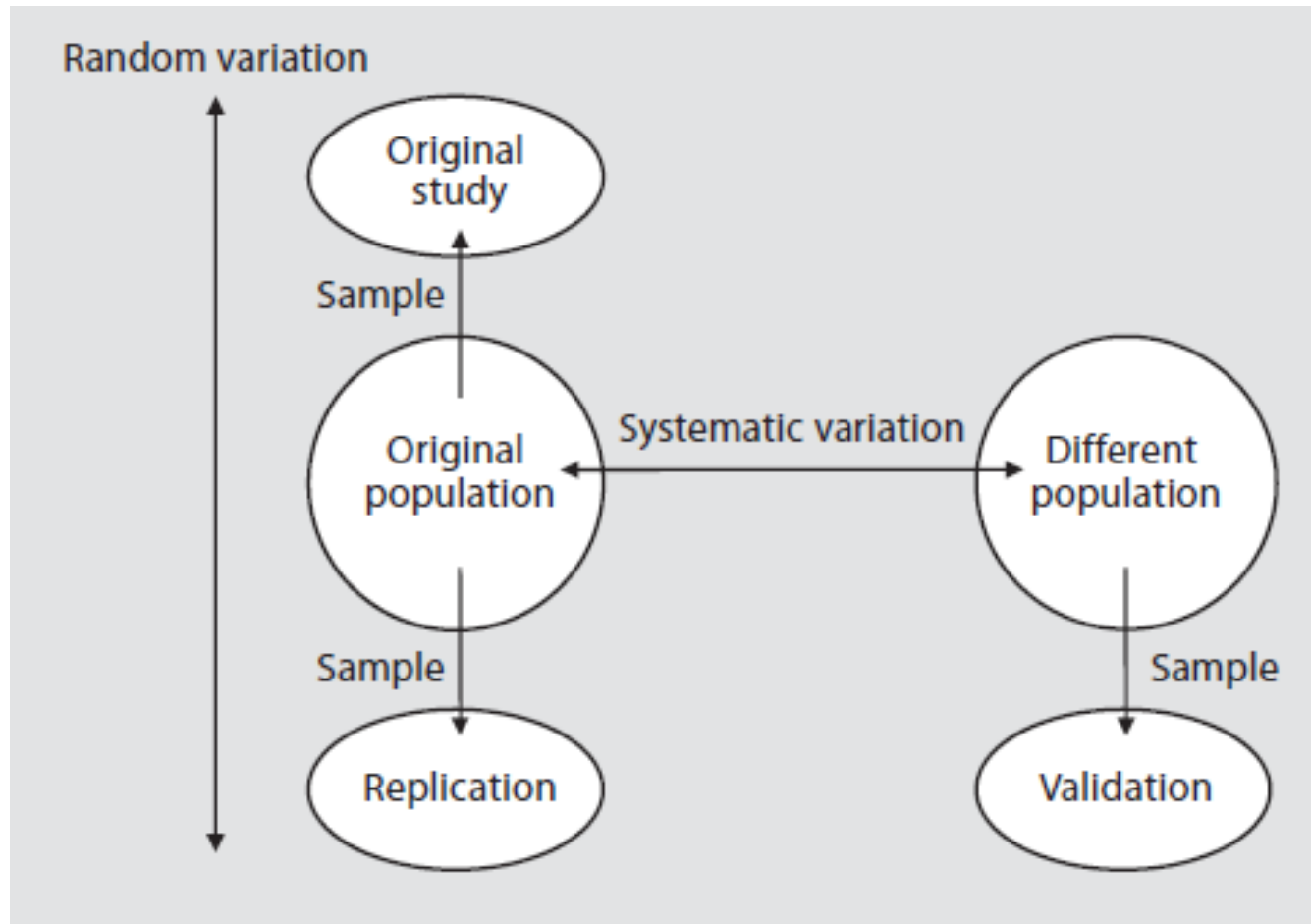
- Showing modest to strong statistical significance
- Having common minor allele frequency (>0.05)
- Modest to strong genetic effect sizes (parametric paradigms)



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

Validation

- Validation is not replication:



(Igl et al. 2009)

Challenges and opportunities

Which findings to pursue ~ replication / interpretation?

A selection of challenges:

- Restrict attention to the same chromosome as the hits or not?
- What are the LD-friends related to our pairs of interest?
- Target pairs that can be “replicated”?
 - Different steps in the GWAS process
 - Different approaches within one step
- Target pairs that can be mapped to underlying biological epistasis networks or pathways?

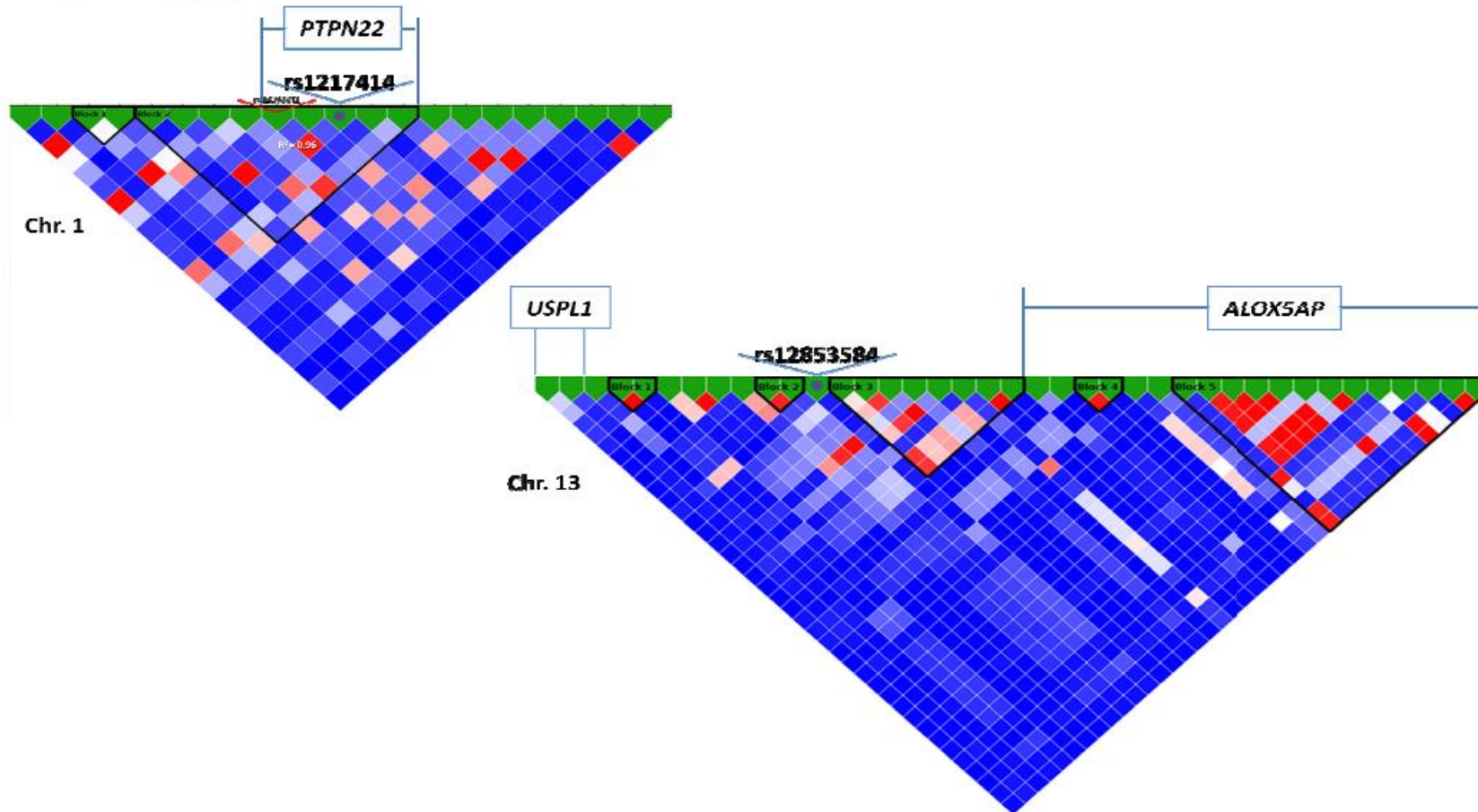
Challenge 1

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

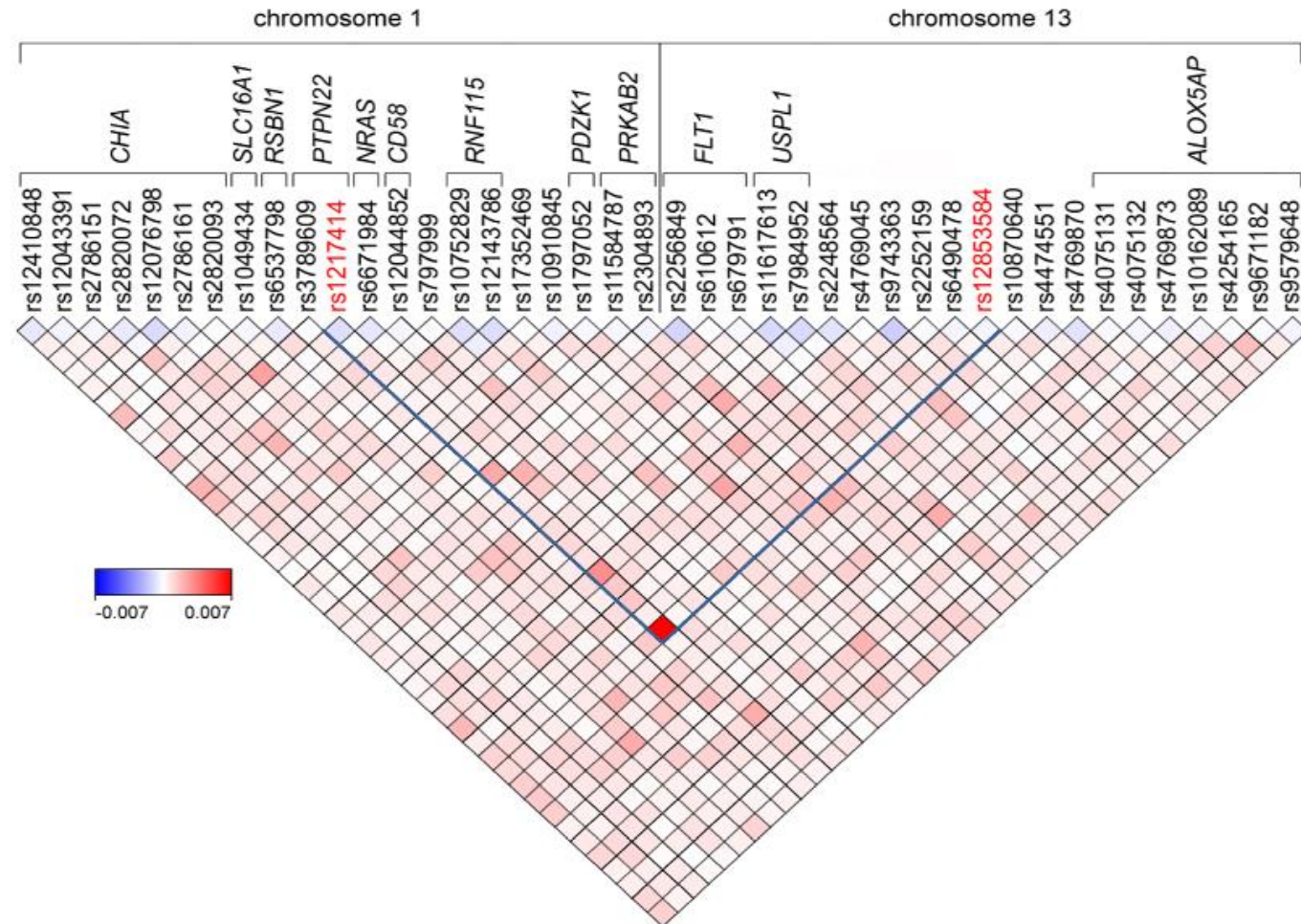
	SNP	SNP position	Gene	Main effect	MAF
$r^2 = 0.110$ $r^2 = 0.047$ $r^2 = 0.022$ $r^2 = 0.027$	rs17116117	chr11:113801591	HTR3B	0,001	0,052
	rs2513574	chr11:113681305	USP28	>0.05	0,123
	rs2519200	chr11:113684809	USP28	>0.05	0,238
	rs1713671	chr11:113674838	USP28	>0.05	0,416
	rs4938056	chr11:113786539	HTR3B	>0.05	0,400
$r^2 = 0.027$	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

Challenge 2

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



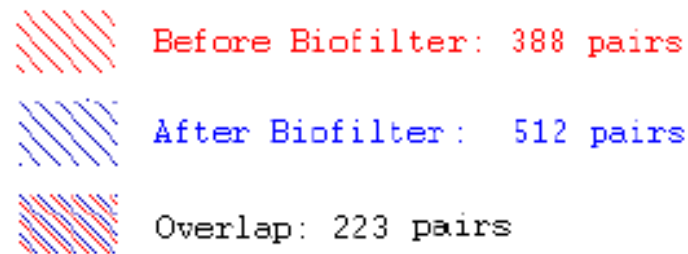
- Synergy Disequilibrium (SD) plots: LD \neq interaction



Challenge 3

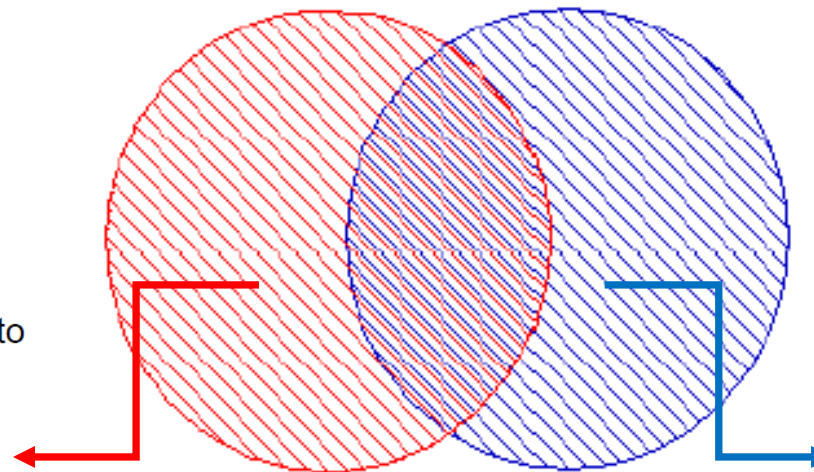
- What is replication?

Application of filtering on WTCCC Rheumatoid Arthritis (RA)



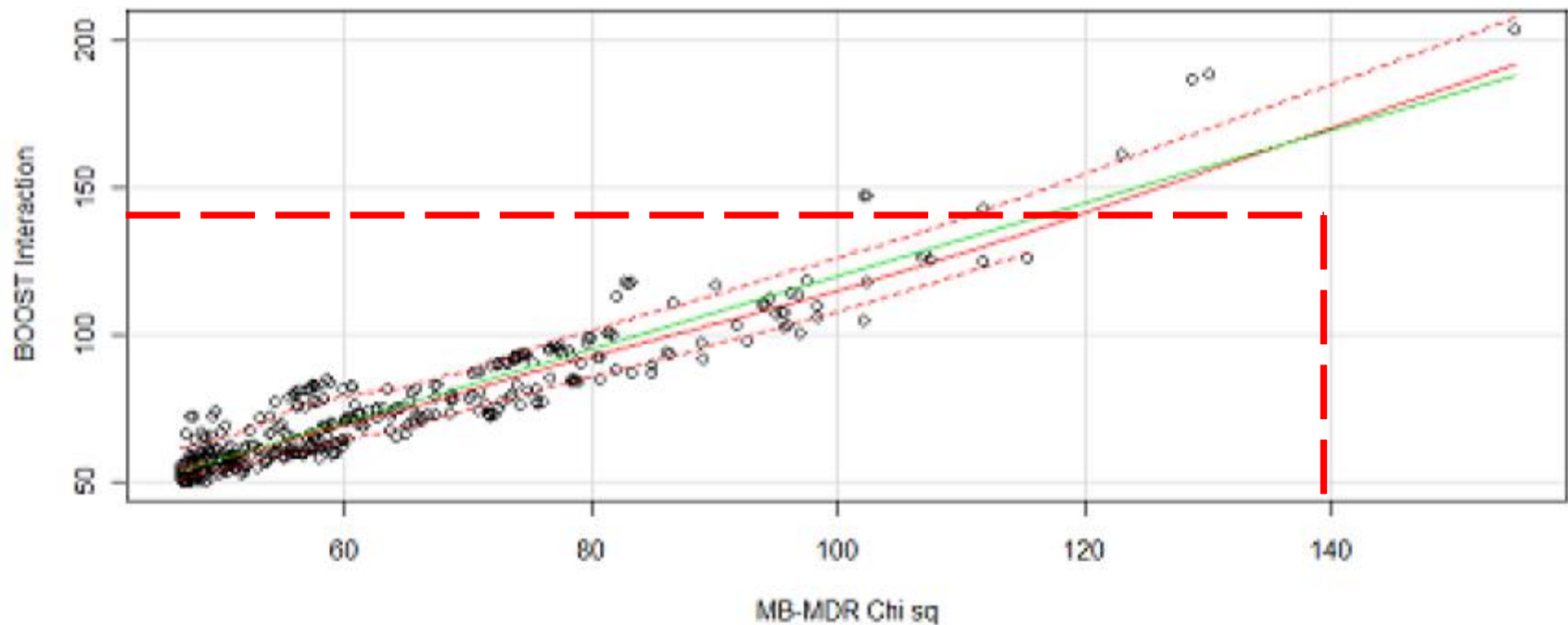
165 “lost” pairs
contain 191 SNPs:

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



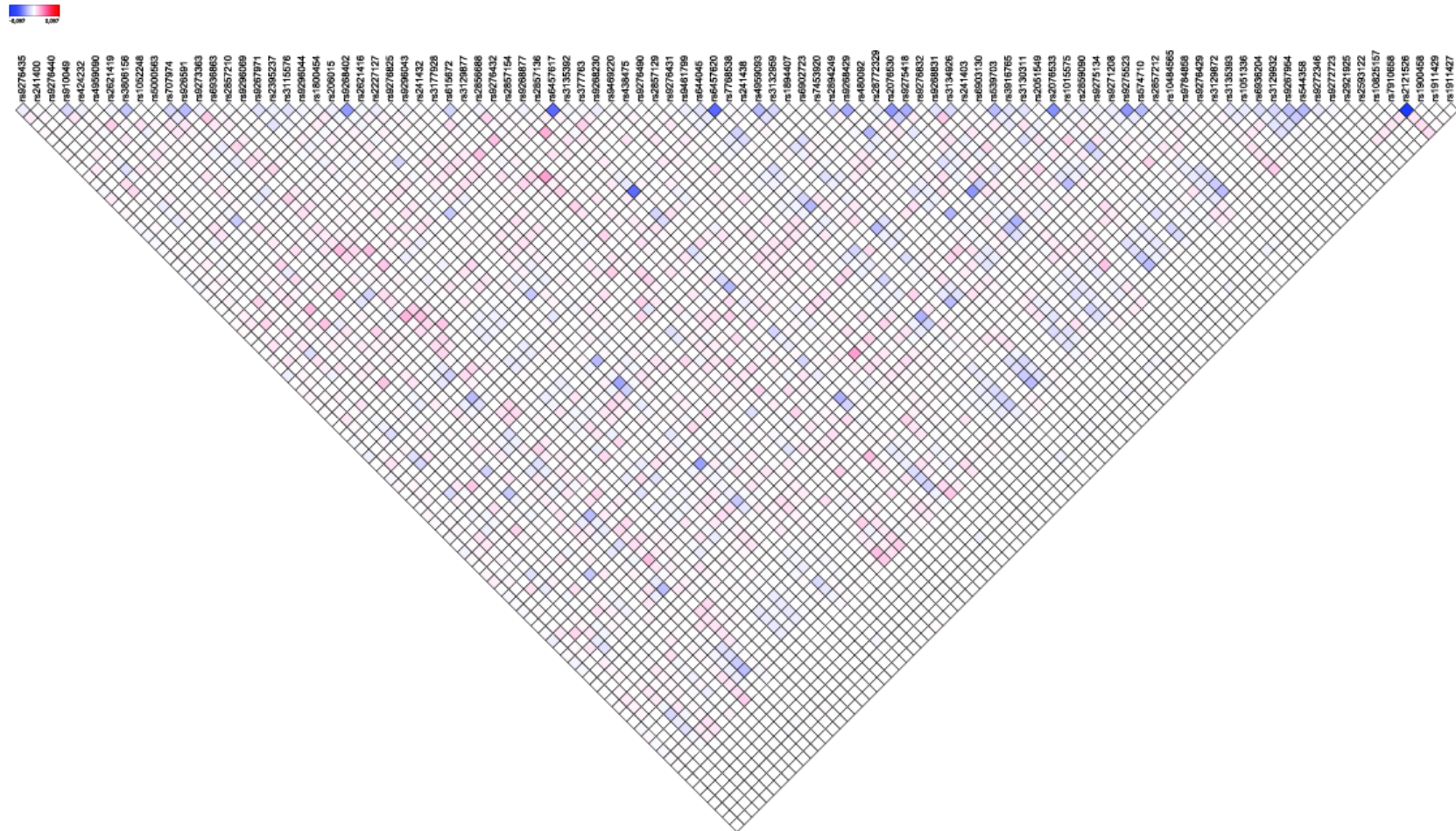
289 new pairs after
Biofilter: is Bonferroni
correction too severe?

- On the same Bio-filtered data, up-scaled logistic regression software (BOOST; 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!

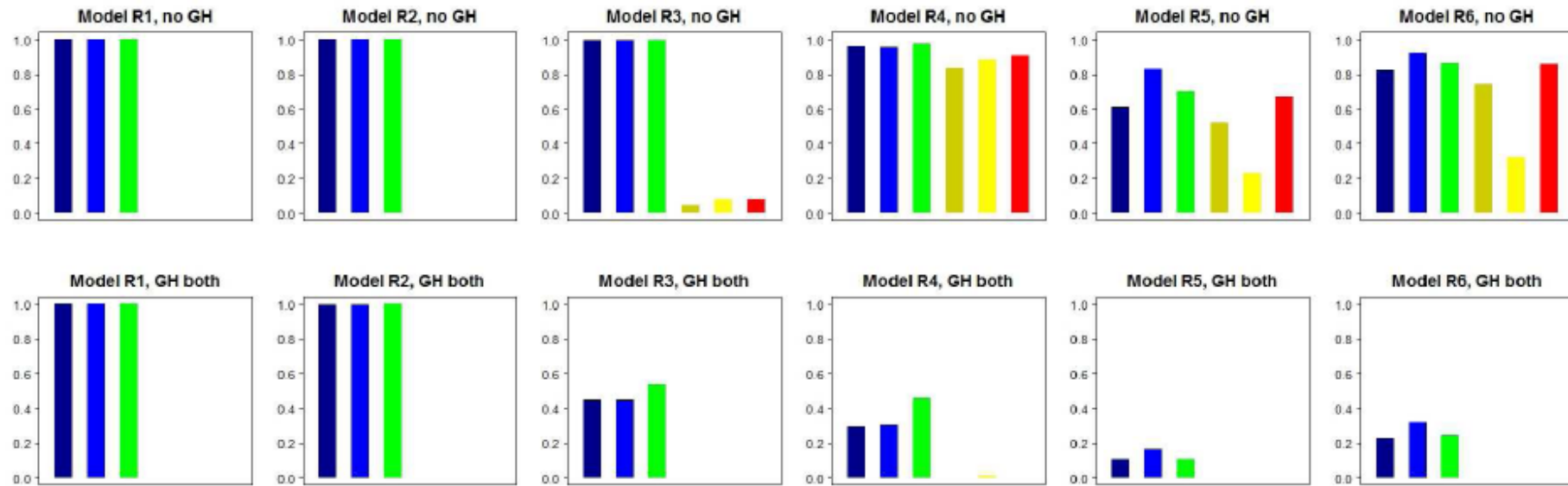
- SD between SNPs in pairs detected by BOOST but not by MB-MDR ...



- Different approaches exist within a single step of the GWAI process
 - 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 widely used 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)

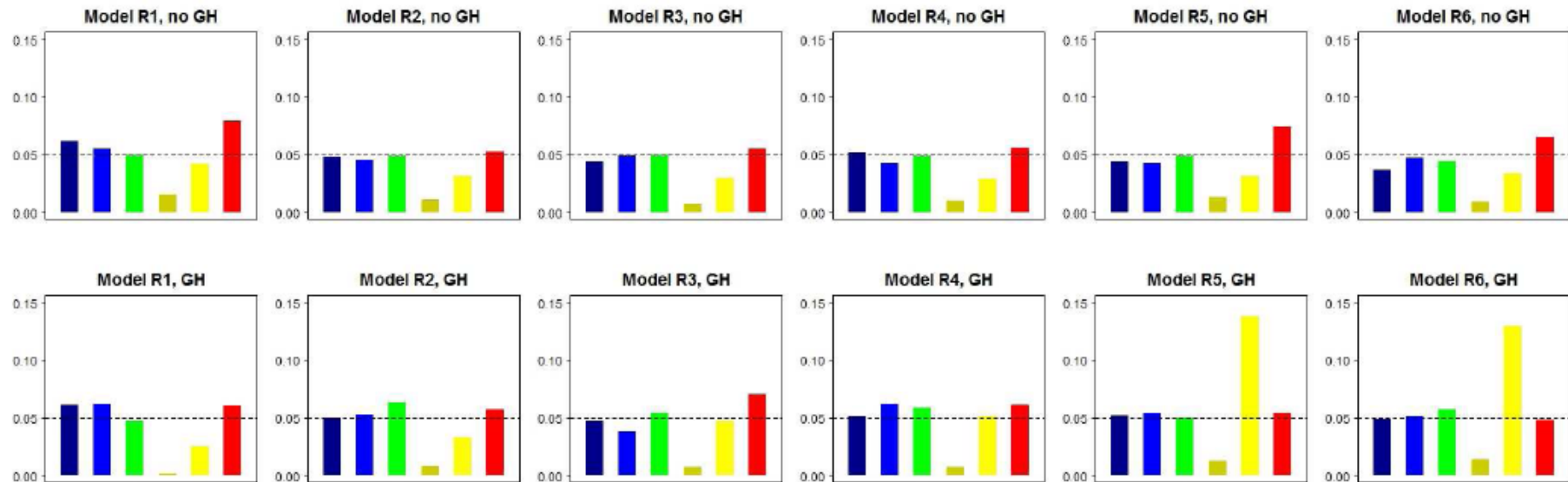
MB-MDR (green)

PLINK epistasis (dark yellow)

PLINK fast epistasis (light yellow)

EPIBLASTER (red)

False positives (pure epistasis scenario's)



BOOST (dark blue)

EpiCruncher optimal options (light blue)

MB-MDR (green)

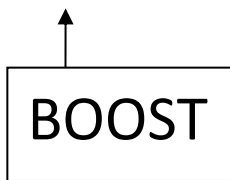
PLINK fast epistasis (light yellow)

EPIBLASTER (red)

PLINK epistasis (dark yellow)

- Concerns:
 - Are the methods 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			
rs17116117	rs1169722																			x
number significant		6	6	6	6	7	5	7	5	6	7	6	6	7	6	7	6	6	3	3



- There is a need for investigating the “information overlap” and “information complement” induced by different methodologies when applied to a variety of (reference?) data. This will allow the development of genuine “ensemble” methods (ongoing – Van Steen lab), will facilitate the interpretation and replication of findings.

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

SNP Pair	Epistasis Detection Method				
	MBMBDR	EpiCruncher	BOOST	PLINK	EpiBlaster
rs17116117rs2513574	1	1	1	1	1
rs17116117rs2519200	2	2	2	2	2
rs11936062rs13126272	3	3	3	179	100
rs17116117rs1713671	4	4	4	5	100
rs17116117rs4938056	5	5	5	3	100
rs1217414 rs128535846	6	7	251	100	
rs1169722 rs171161177	7	9	82	4	
rs17116117rs7126080	8	8	6	81	100
rs13126272rs4862419	9	9	8	198	100
rs1933641 rs6099309	10	309	308	297	100

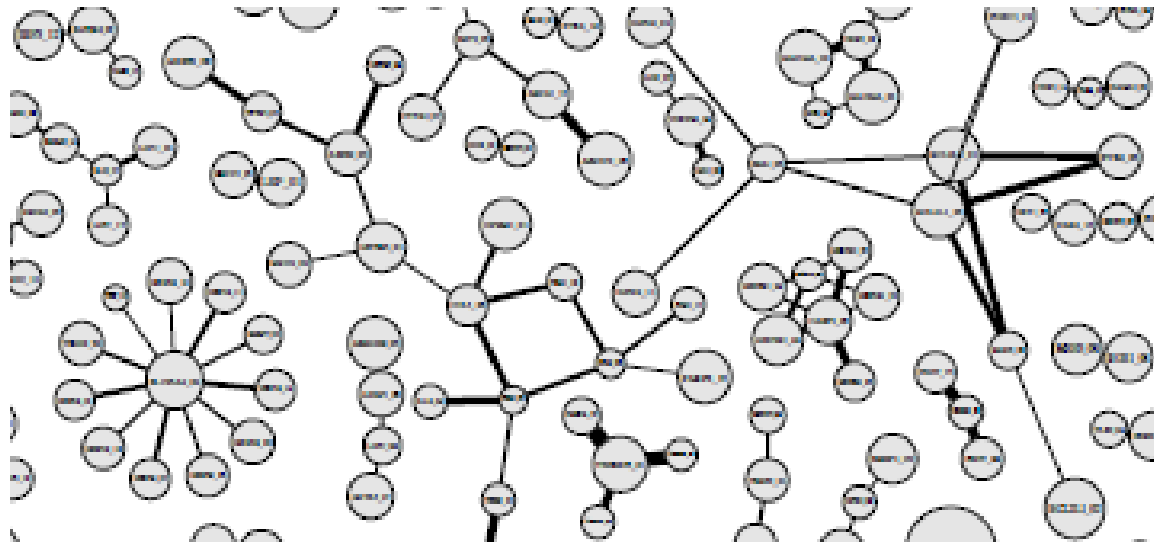
Challenge 4

- Target pairs that can be mapped to underlying biological epistasis networks or pathways?
- Relying on criteria for assessing the functional significance of each involved functional variant separately? (Rebbeck et al 2004)

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

- Relying on criteria for assessing the functional significance of gene-gene interaction patterns?

Would involve overlaying “statistical” epistasis networks with “biological” networks (e.g., linking hubs in “statistical” epistasis networks to functional groups or pathways)



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

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 GWAI.

Omics integrative approaches for GWAs and GWEIs

Example in GWAs

- 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

Proof of concept

Spondylitis



Interaction between *ERAP1* and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility

The Australo-Anglo-American Spondyloarthritis Consortium¹ (TASC) & the Wellcome Trust Case Control Consortium 2 (WTCCC2)¹

Ankylosing spondylitis is a common form of inflammatory arthritis predominantly affecting the spine and pelvis that occurs in approximately 5 out of 1,000 adults of European descent. Here we report the identification of three variants in the *RUNX3*, *LTBR-TNFRSF1A* and *IL12B* regions convincingly associated with ankylosing spondylitis ($P < 5 \times 10^{-8}$ in the combined discovery and replication datasets) and a further four loci at *PTGER4*, *TBKBP1*, *ANTXR2* and *CARD9* that show strong association across all our datasets ($P < 5 \times 10^{-6}$ overall, with support in each of the three datasets studied). We also show that polymorphisms of *ERAP1*, which encodes an endoplasmic reticulum aminopeptidase involved in peptide trimming before HLA class I presentation, only affect ankylosing spondylitis risk in HLA-B27-positive individuals. These findings provide strong evidence that HLA-B27 operates in ankylosing spondylitis through a mechanism involving aberrant processing of antigenic peptides.

Psoriasis

nature
genetics

A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between *HLA-C* and *ERAP1*

Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2^{1,2}

To identify new susceptibility loci for psoriasis, we undertook a genome-wide association study of 594,224 SNPs in 2,622 individuals with psoriasis and 5,667 controls. We identified associations at eight previously unreported genomic loci. Seven loci harbored genes with recognized immune functions (*IL28RA*, *REL*, *IFIH1*, *ERAP1*, *TRAF3IP2*, *NFKB1A* and *TYK2*). These associations were replicated in 9,079 European samples (six loci with a combined $P < 5 \times 10^{-8}$ and two loci with a combined $P < 5 \times 10^{-7}$). We also report compelling evidence for an interaction between the *HLA-C* and *ERAP1* loci (combined $P = 6.95 \times 10^{-6}$). *ERAP1* plays an important role in MHC class I peptide processing. *ERAP1* variants only influenced psoriasis susceptibility in individuals carrying the *HLA-C* risk allele. Our findings implicate pathways that integrate epidermal barrier dysfunction with innate and adaptive immune dysregulation in psoriasis pathogenesis.

Subjects for the GWAS discovery set were recruited from the UK and Ireland and were of self-reported European ancestry (Supplementary Table 1). Individuals with psoriasis (cases) were genotyped on the Illumina Human660W-Quad, and controls were genotyped on the Illumina custom Human1.2M-Duo (Supplementary Note), with a primary analysis performed on the overlapping set of SNPs. We performed stringent data quality control procedures (Online Methods), resulting in a GWAS dataset comprising 2,178 individuals with psoriasis and 5,175 controls genotyped at 535,475 SNPs. Principal components analysis of the study data showed the first principal component stratified individuals by population origin (Supplementary Fig. 1b). We performed single SNP analysis using score tests under a logistic regression model which assumed multiplicative effects, including the first principal component as a covariate and we accounted for uncertainty in genotype calls as implemented in SNPTEST (see URLs). After removal of known and replicated psoriasis association loci, the overdispersion factor¹⁰ of association test statistics (λ_{GC}) was 1.045

Acknowledgments (MB-MDR)



Acknowledgments (Alzheimer's)

Kristel Van Steen (PI)

Montefiore Institute / GIGA-R,
University of Liege,
Belgium



Nilufer Taner

Mayo Clinic Rochester, Departments of
Neurology and Neuroscience, Rochester,
USA



Elena Gusareva

Montefiore Institute / GIGA-R,
University of Liege,
Belgium

Kristel Slegers

Neurodegenerative Brain Diseases Group,
Department of Molecular Genetics, VIB,
University of Antwerp,
Belgium



Jean-Charles Lambert

INSERM U744, Lille, France
Institut Pasteur de Lille, Lille, France
Universite de Lille Nord de France, Lille,
France

