

Genome-wide Association Studies

Kristel Van Steen & Andreas Ziegler

kristel.vansteen@ulg.ac.be & ziegler@imbs.uni-luebeck.de

Florianopolis, Brazil

IBC 2010

Content (Afternoon)

7 Curse of dimensionality and multiple testing

8 Missing data

9 Variable selection methods

10 Epistasis: a curse or a blessing?

11 Modeling epistasis

- Methods to detect epistasis: state of the art
- Focus on data dimensionality reduction methods
- The importance of adjusting for confounding factors or lower-order effects

12 Interpretation of identified interactions

- The value of entropy-based measures

Recapitulation Learning Outcomes

- Familiarize attendees with all stages of GWA analysis
- Able to
 - analyze basic GWA study,
 - identify significant main effects,
 - identify significant interaction effects.
- Aware of potential pitfalls in GWA studies, whether the focus is on
 - main effects,
 - interaction effects, or
 - both.
- Acquire essential background to overcome some of the hurdles involved in GWAS

Part 9

Variable selection methods

Why selecting variables?

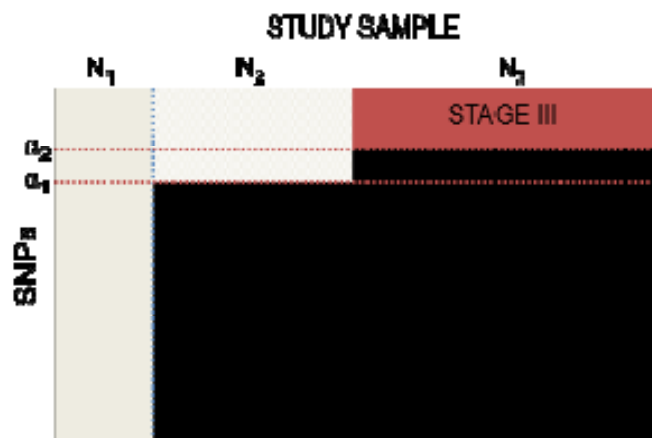
Introduction

- 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

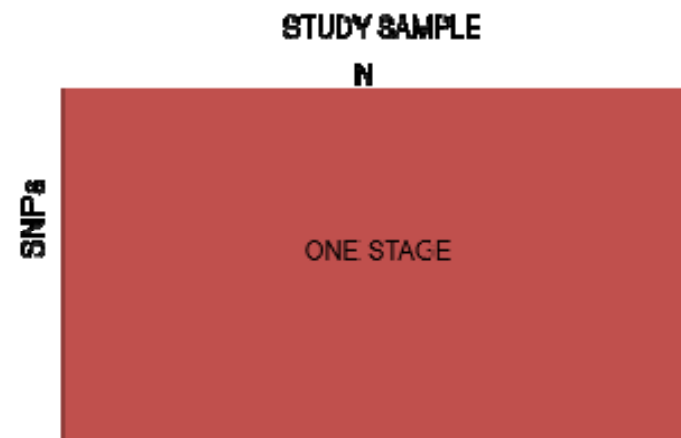
Multi-stage

- 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)

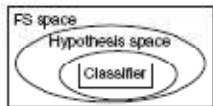
Feature selection methods

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 (Siedlecky 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

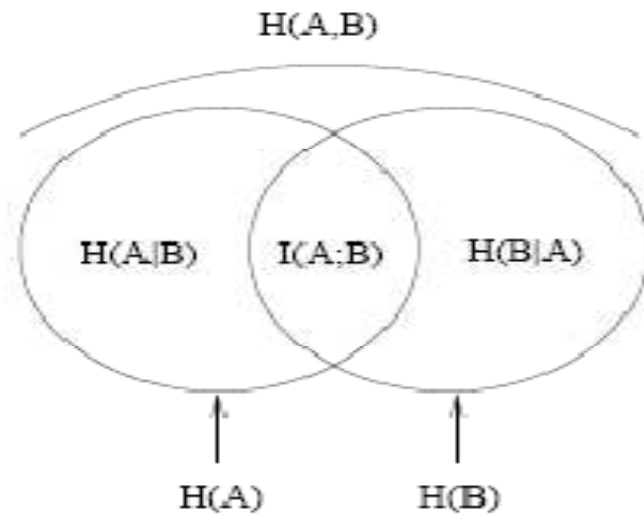
Conditional entropy

- 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)

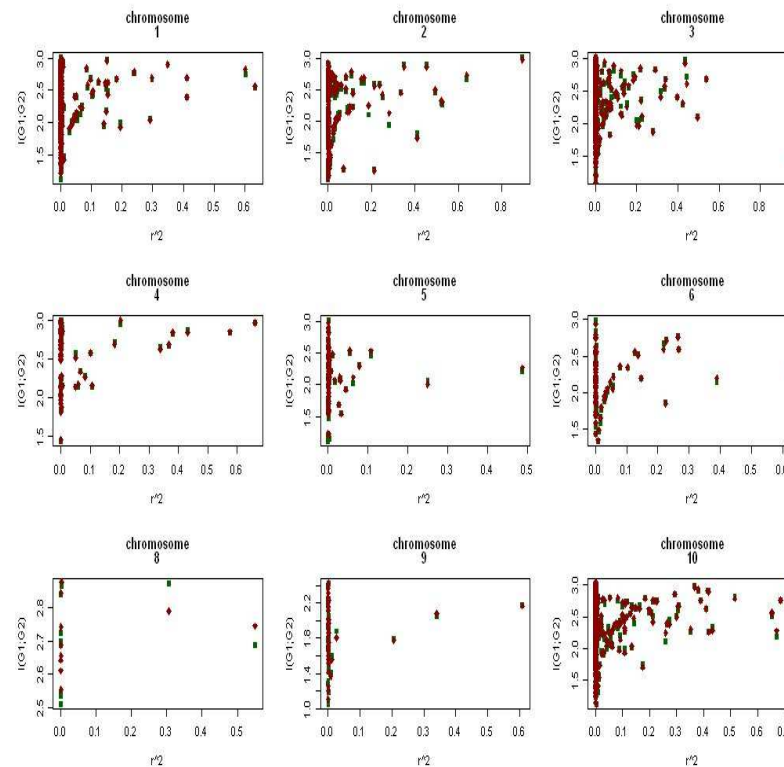
- Mutual information can be expressed as a Kullback-Leibler divergence, of the product $p_A(a)p_B(b)$ of the marginal distributions of the two random variables A and B , from the random variables' joint distribution
- $I(A; B)$ can also be understood as the expectation of the Kullback-Leibler divergence 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

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

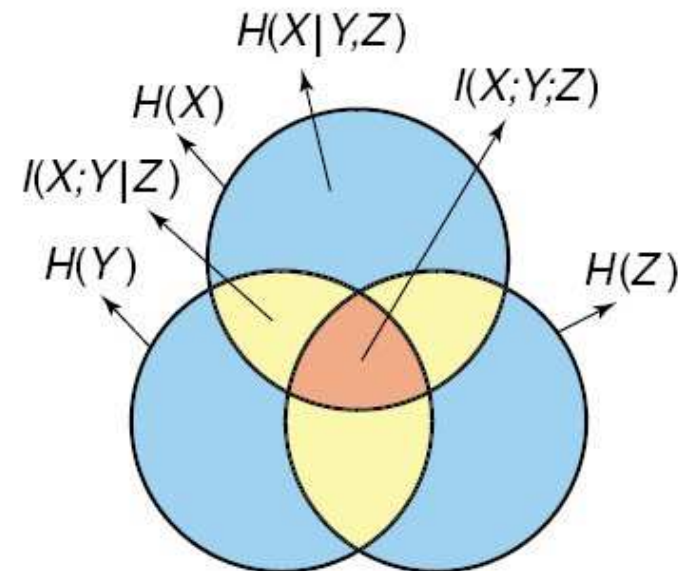
Multivariate mutual information

- 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 of information gain

- 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 2007)

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

Strategy 3: Data mining

Random Forests (RF)

(Breiman 2001)

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

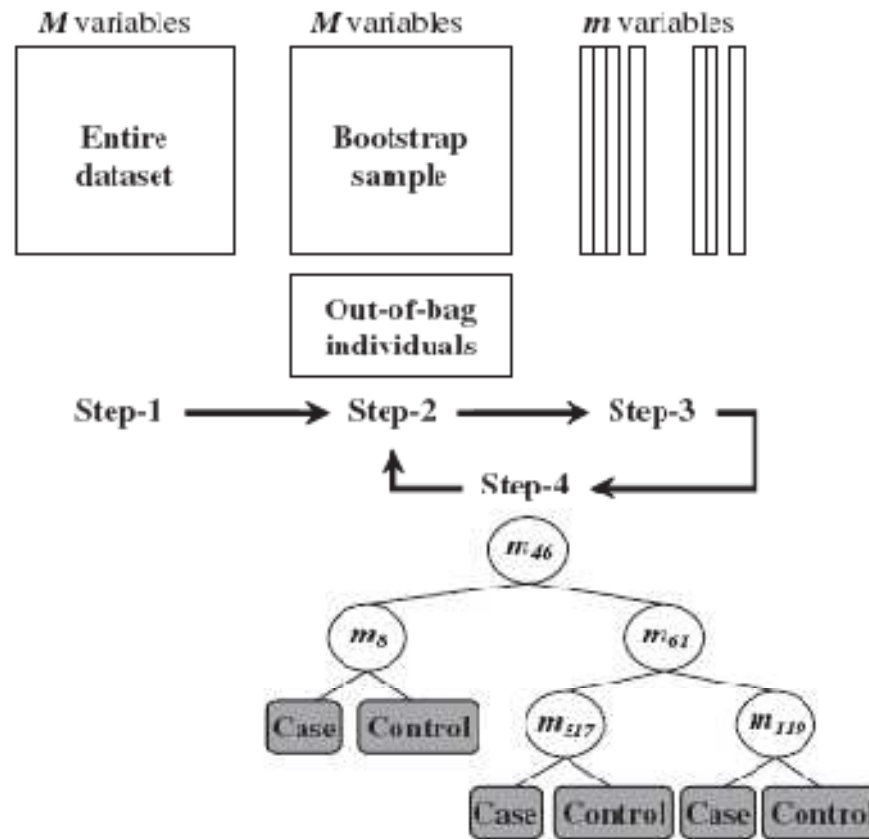
- Draw n_{tree} bootstrap samples from the original data.
- For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following specifications:
 - at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and 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; Nicodemus 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

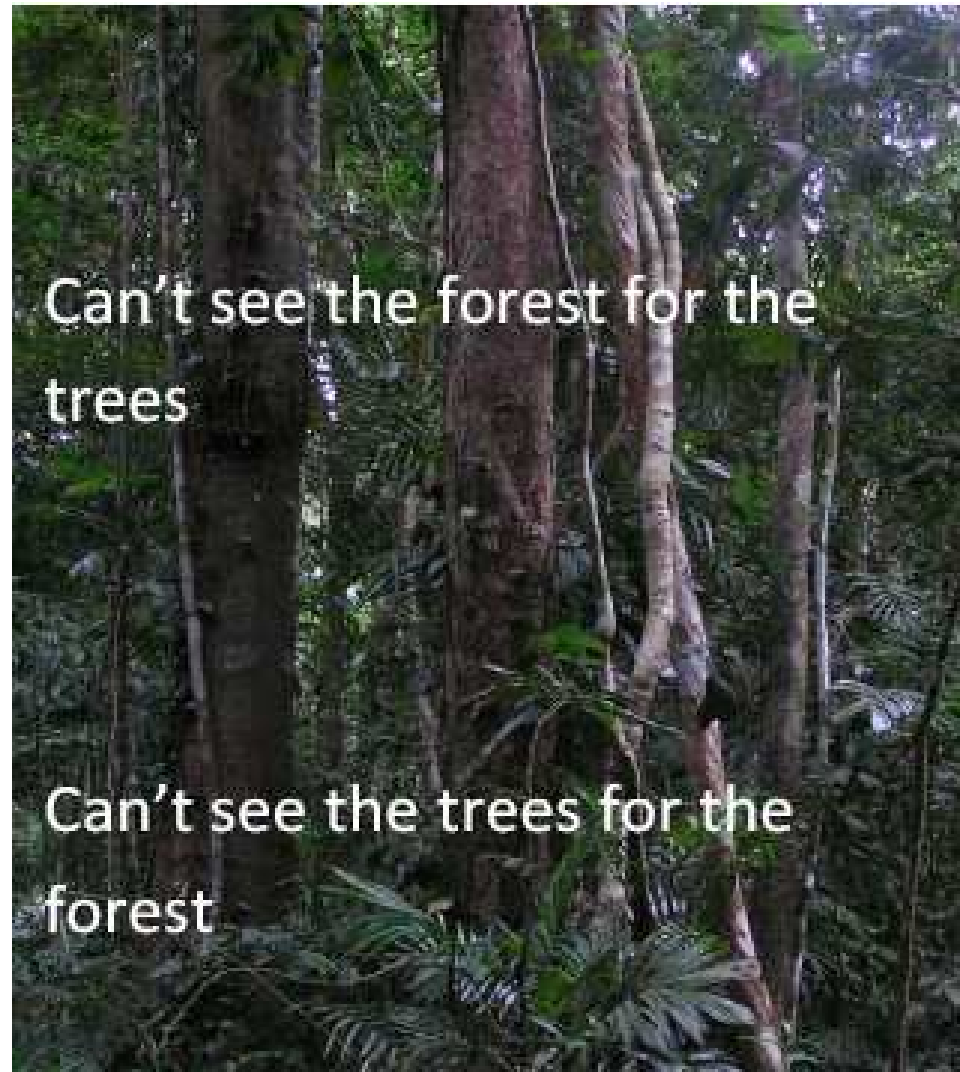
- New implementations of RF allow rapid analysis of highly dimensional data such as those generated for GWA studies (Schwarz et al 2010): Random Jungle (<http://www.randomjungle.org/rjungle/>)



Part 10

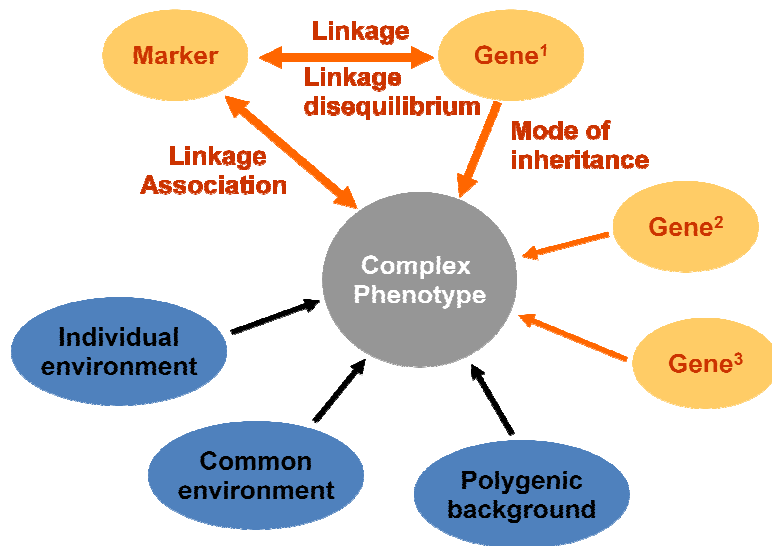
Epistasis: a curse or a blessing?

No need to get lost in the jungle ...



The nature of complex disease

How complex is complex?



(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 *nonlinear interactions* with *genetic and environmental* factors

(Moore 2008)

The quest for epistasis

- In the quest for genetic susceptibility factors and the search for “the missing heritability”, supplementary and complementary efforts to classical main effects GWAS have been undertaken:
 - the inclusion of several genetic inheritance assumptions in model development,
 - the consideration of different sources of information,
 - the acknowledgement of disease underlying pathways of networks, and
 - indirectly or directly testing for gene-gene interactions

The quest for epistasis

- 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 efforts of statistical epistasis to biological relevance,
 - and integration of -omics information sources

Definition: epistasis – what's in the name?

- **Interaction** is a kind of action that occurs as two or more objects have an effect upon one another. The idea of a two-way effect is essential in the concept of interaction, as opposed to a one-way causal effect. (Wikipedia)



(slide : C Amos)

Definition: epistasis – what's in the name?

- Distortions of Mendelian segregation ratios due to one gene masking the effects of another (William Bateson 1861-1926).

Genotype at locus B	Genotype at locus G		
	<i>g/g</i>	<i>g/G</i>	<i>G/G</i>
<i>b/b</i>	White	Grey	Grey
<i>b/B</i>	Black	Grey	Grey
<i>B/B</i>	Black	Grey	Grey

- Deviations from linearity in a statistical model (Ronald Fisher 1890-1962).
- “Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans” (Cordell 2002)

The origin of epistasis

Why is there epistasis?

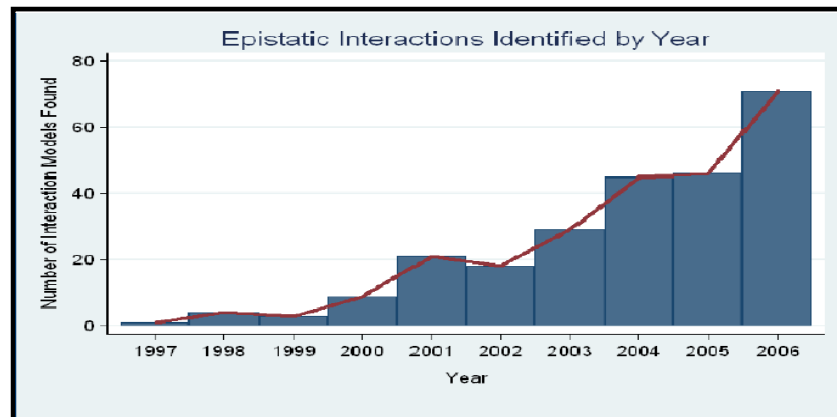
- To date it is unclear why epistasis exists and why it might be an important component of the genetic architecture of many biological complex traits.
- That epistasis plays a role in human genetics is without doubt, given the numerous discoveries of significant gene-gene interactions in model organisms, providing evidence for interactions in the presence and absence of important individual effects (Carlborg et al 2004) and given the insights gained in cell biology showing complex interactions between different types of biomolecules (Joyce et al 2006).

Why is there epistasis?

- Epistasis and genomic complexity are correlated, in the sense that in less complicated genomes mutational effects involved in epistasis tend to cancel each other out, whereas in more complex genomes mutational effects rather strengthen each other, leading to so-called synergetic epistasis (Sanjuan et al 2006, 2008).
- Also, dependencies among genes in networks, leading to epistasis, naturally arise when believing that the human system guards itself to negative evolutionary effects of mutations via redundancy and robustness (Moore 2005).

Gradual shift from main towards epistatis effects

- It is therefore not surprising that, with a growing tool-box of analysis techniques and approaches, the number of identified epistatic effects in humans, showing susceptibility to common complex human diseases, is gradually increasing (Emily et al 2009, Wu et al 2010).



(Motsinger et al 2007)

Different degrees of epistasis

One example of a two-locus model (dichotomous trait)

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

- Here, penetrance values are enumerated as 0 and 1 (i.e., fully penetrant – show-case example).
- There are $2^9=512$ possible models, not accounting for symmetries in the data

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(1L:R) 0 0 0 0 0 0 1 1 1	M10 0 0 0 0 0 1 0 1 0	M11 (T) 0 0 0 0 0 1 0 1 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	M18 0 0 0 0 1 0 0 1 0
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 (DD) 0 0 0 0 1 1 0 1 1	M28 0 0 0 0 1 1 1 0 0	M29 0 0 0 0 1 1 1 0 1
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(1L:I) 0 0 0 1 1 1 0 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	M70 0 0 1 0 0 0 1 1 0
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	M98 0 0 1 1 0 0 0 1 0
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	M170 0 1 0 1 0 1 0 1 0
M186 0 1 0 1 1 1 0 1 0						

Each model represents a group of equivalent models under permutations. The representative model is the one with the smallest model number.

Pure epistasis: An example

- $p(A)=p(B)=p(a)=p(b)=0.5$
- HWE (hence, $p(AA)=0.5^2=0.25, p(Aa)=2 \times 0.5^2=0.5$) and no LD
- Penetrances are given according to the table below

$P(\text{affected} | \text{genotype})$

Penetrance	bb	bB	BB	prob
aa	0	0	1	0.25
aA	0	0.50	0	0.25
AA	1	0	0	0.25
prob	0.25	0.25	0.25	

- Make use of the total law of probability to derive the $P(\text{affected} | aa) = 0.25 \times 0 + 0.5 \times 0 + 0.25 \times 1 = \mathbf{0.25}$

Pure epistasis: An example (continued)

- ...The marginal genotype distributions for cases and controls are the same: one-locus approaches will be powerless!

P(genotypes | affected)

	bb	bB	BB	prob
aa	0	0	0.25	0.25
aA	0	0.50	0	0.50
AA	0.25	0	0	0.25
prob	0.25	0.50	0.25	1

P(genotypes | unaffected)

	bb	bB	BB	prob
aa	0.083	0.167	0	0.25
aA	0.167	0.167	0.167	0.50
AA	0	0.167	0.083	0.25
prob	0.25	0.50	0.25	1

$$P(aa, BB | D) = p(D | aa, BB)p(aa, BB) / p(D)$$

$$= 1 \times 0.5^2 \times 0.5^2 / (1 \times 0.5^2 \times 0.5^2 + 0.5 \times 2 \times 0.5^2 \times 2 \times 0.5^2 + 1 \times 0.5^2 \times 0.5^2)$$

$$= \frac{1}{4} = \mathbf{0.25}$$

Part 11

Modeling epistasis

Main challenges in epistasis detection

- Variable selection (see Part 9)
- Modeling
- Interpretation
 - Making inferences about biological epistasis from statistical epistasis

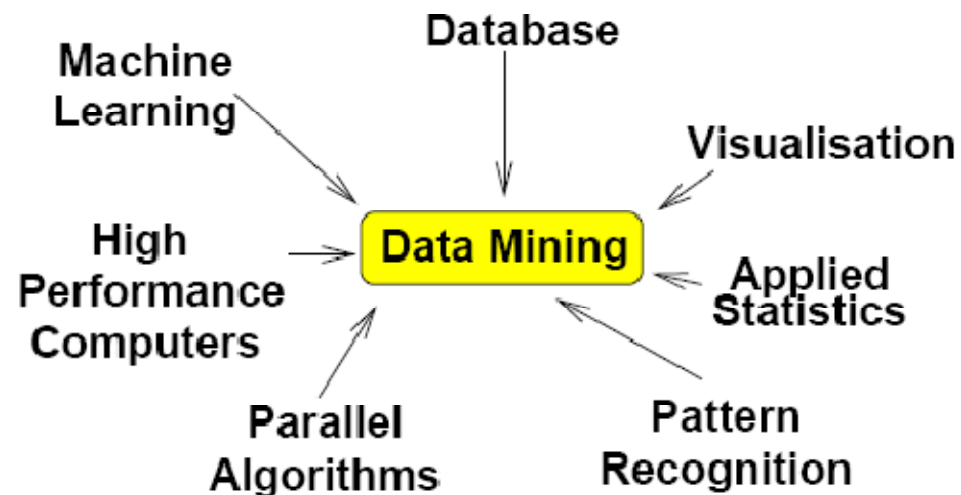
Classification of methods to detect epistasis

- The abundance of developed strategies in the context of epistasis detection complicates classifying these methods.
- Several criteria have been used to make such a classification:
 - whether the strategy is exploratory in nature or not,
 - whether modeling is the main aim, or rather testing,
 - whether the epistatic effect is tested indirectly or directly,
 - whether the approach is parametric or non-parametric,
 - whether 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.

Classification of methods to detect epistasis

(Thomas 2005)

- Exploratory data techniques / non-parametric techniques
 - Large overlap with “data mining” techniques



(Williams 1998)

- Exploratory data techniques / non-parametric techniques:
 - Data segmentation methods :
 - Clustering (InterClus)
 - Tree-based methods:
 - Recursive Partitioning (Helix Tree)
 - Random Forests (R, CART, Random Jungle)
 - Pattern recognition methods:
 - Symbolic Discriminant Analysis (SDA)
 - Mining association rules (MA)
 - Neural networks (NN)
 - Support vector machines (SVM)
 - Multidimensional reduction methods:
 - DICE (Detection of Informative Combined Effects)
 - MDR (Multifactor Dimensionality Reduction)
 - Logic regression (LR) and trees (Onkamo and Toivonen 2006)

Classification of methods to detect epistasis

- Techniques that allow putting more structure on the model to overcome curse of dimensionality / parametric techniques
 - For genetic association studies, a general paradigm is a (parametric) regression analysis
 - The validity of conclusions crucially depends on the underlying model assumptions.
 - The interpretation of the results crucially depends on the adopted coding scheme ... (see Part 12)

Classification of methods to detect epistasis

- These traditional methods often fail due to
 - the large number of genotyped polymorphisms requiring very small p-values for significance assessment,
 - the curse of dimensionality or 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 (exhaustive vs non-exhaustive search),
 - the presence of important interacting loci with relatively small marginal effects,
 - the abundance of rare (or absent) multi-locus genotype combinations with increasing dimensionality.

Classification of methods to detect epistasis

- Efforts in this area concentrate on dealing with one or more of these issues
- Efforts in this area may also distinguish between
 - aiming to explicitly test for interaction
 - aiming to test for a global multi-locus effect
 - aiming to develop an optimal prediction model

Classification of methods to detect epistasis

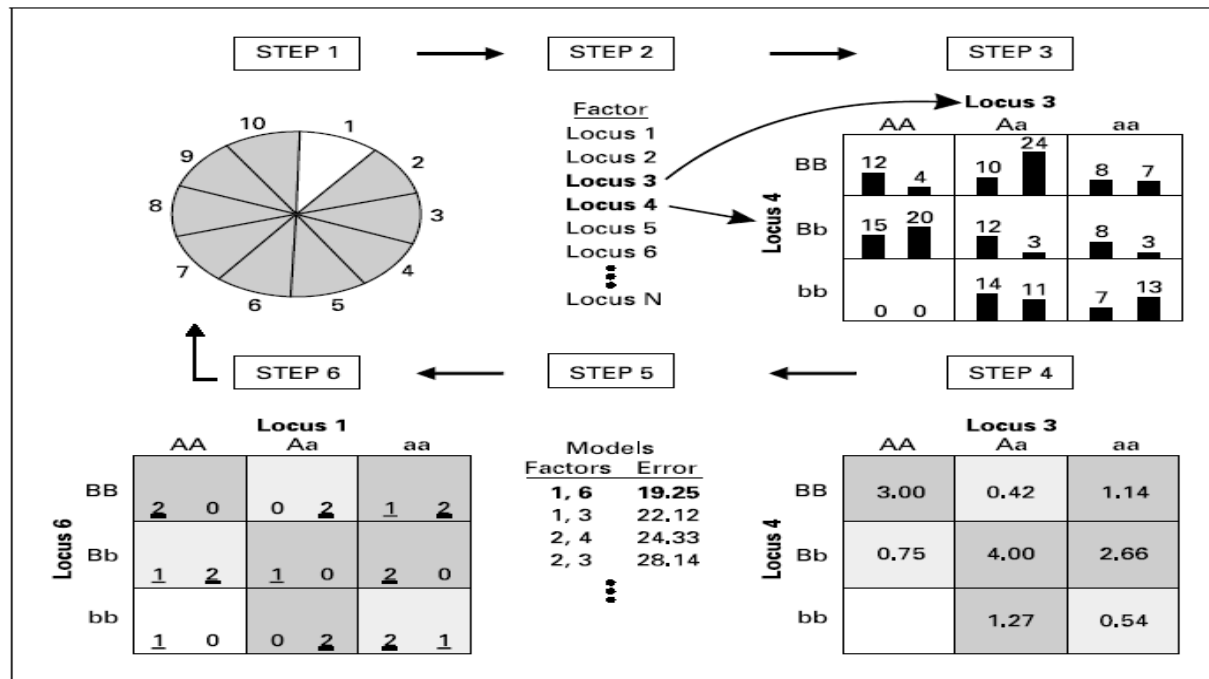
- An **exhaustive method** is a method that enumerates all possible k -way interactions for some k in order to identify the effect or effects which best predict/model phenotypic outcomes
 - Multifactor Dimensionality Reduction techniques (MDR - Ritchie et al 2001; MB-MDR - Calle et al 2008)
 - Full interaction parametric regression models (ITF - Millstein et al 2006)
 - Restricted search by sub-selecting all pairs on a property unrelated to the phenotype of interest, followed by an evaluation of the most promising pairs for significant interaction signals (COE – Zhang et al 2000)

- A non-exhaustive search performs a partial search of the interaction space to derive conclusions as quickly as possible
 - Greedy methods
perform filtering based on non-epistatic or lower-order interaction results to filter out markers displaying no main or lower-order effects → pure epistatic effects are likely to be missed
 - Two-stage approaches (Marchini et al 2005)
 - Trees and forests (Chen et al 2007 and Schwarz et al 2010)

- A non-exhaustive search performs a partial search of the interaction space to derive conclusions as quickly as possible
 - Stochastic methods
 - iteratively select a small number of loci and perform a thorough test for epistasis → success relies on the true interaction appearing in at least one iteration
 - Logic regression (LR – Schwender and Ickstadt 2008)
 - Random Jungle (Schwarz et al 2010)

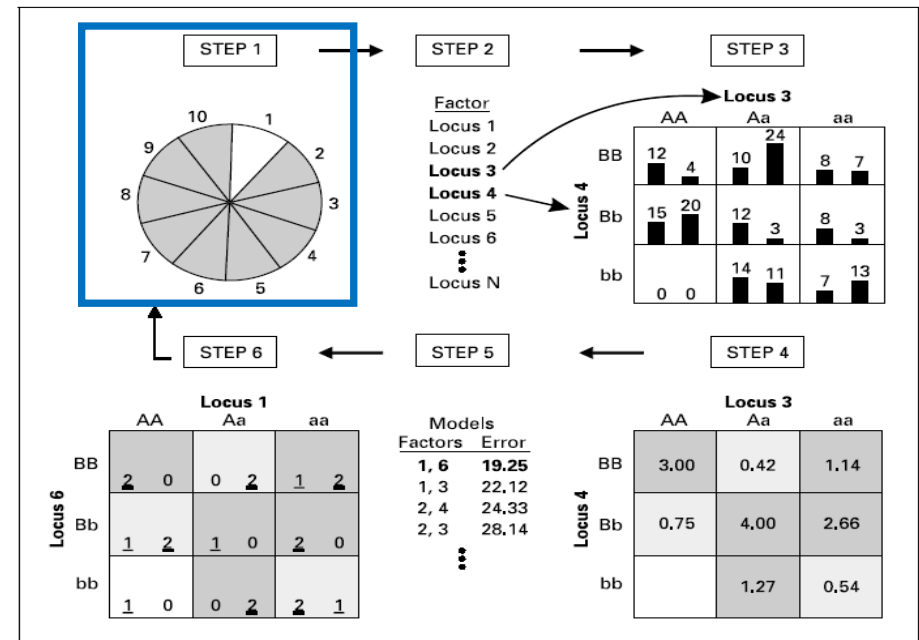
Multifactor Dimensionality Reduction (MDR)

The 6 steps of MDR



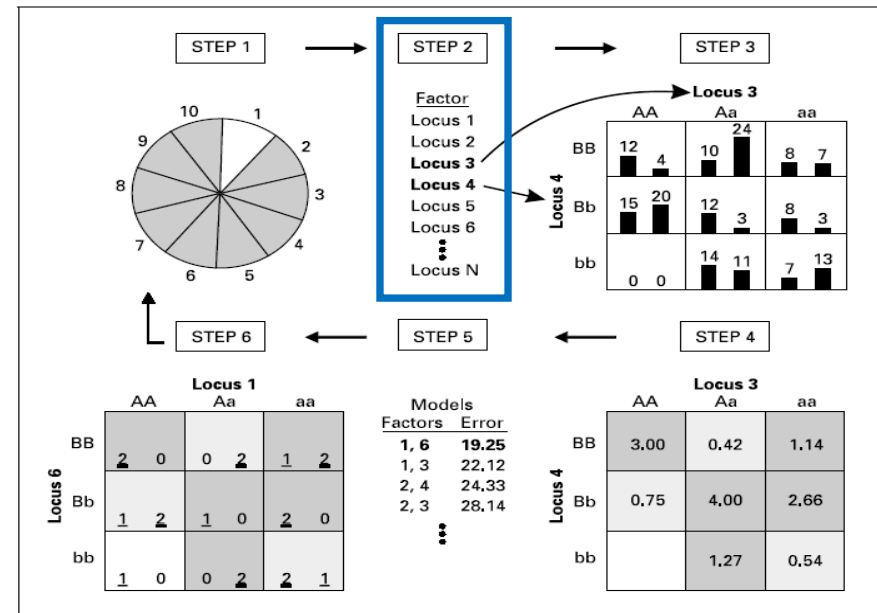
MDR Step 1

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



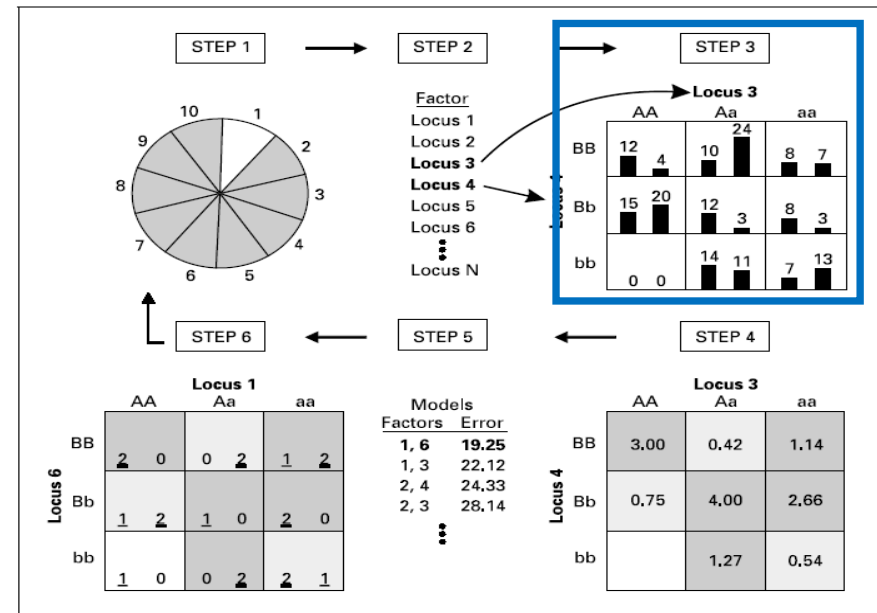
MDR Step 2

- 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



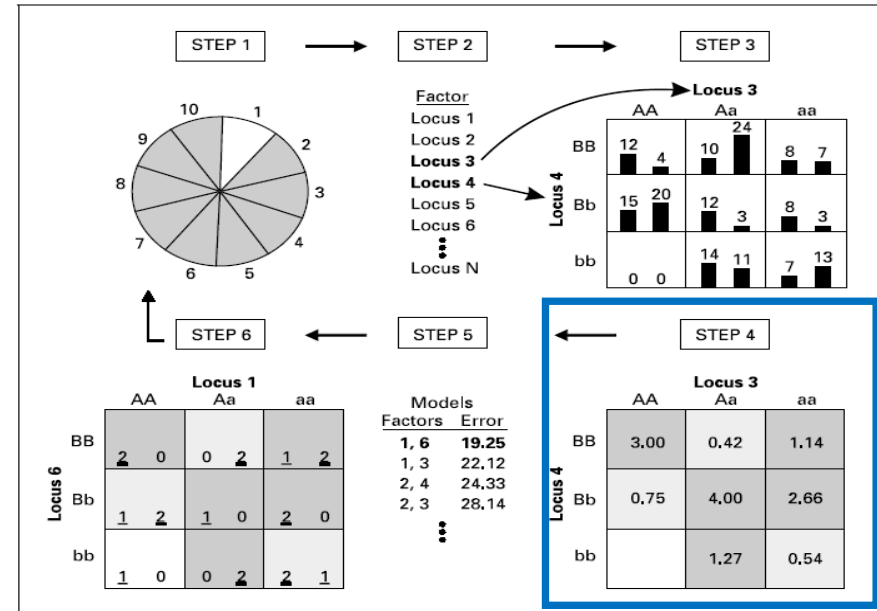
MDR Step 3

- 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:

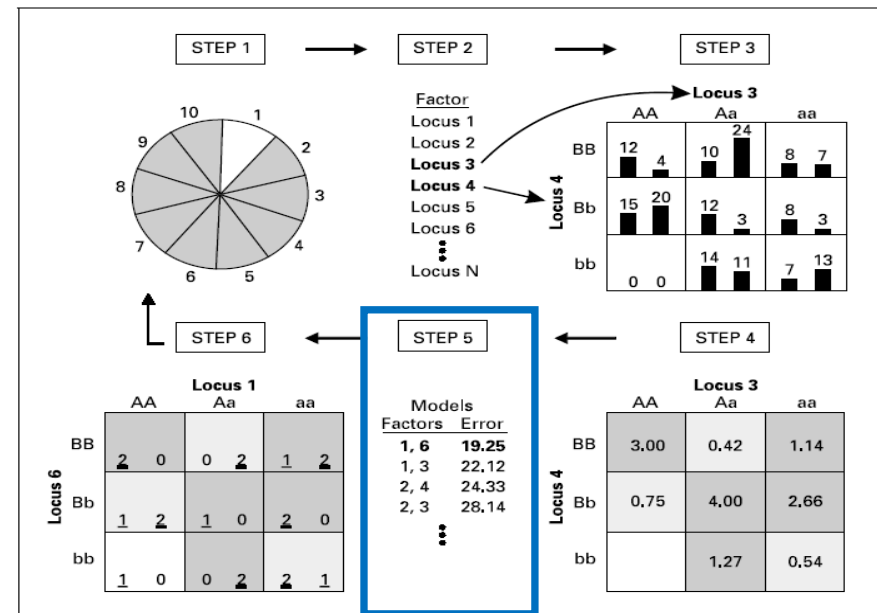


Reduces k -dimensional space to 1 dimension with 2 levels

MDR Step 5

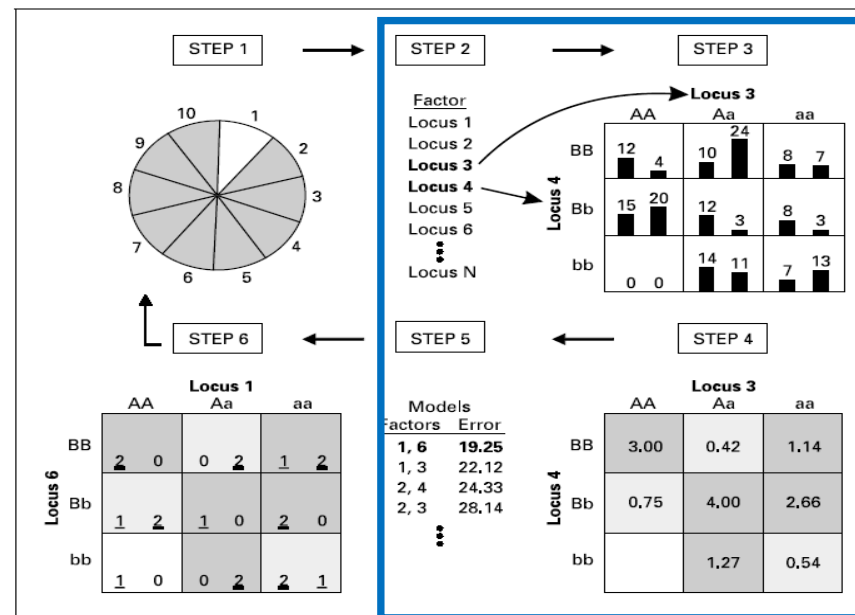
- To evaluate the developed model in Step 4, use labels to classify individuals as cases or controls, and calculate the misclassification error
- In fact: balanced accuracy are preferred (arithmetic mean between sensitivity and specificity), which IS mathematically equivalent to

classification accuracy when data are balanced



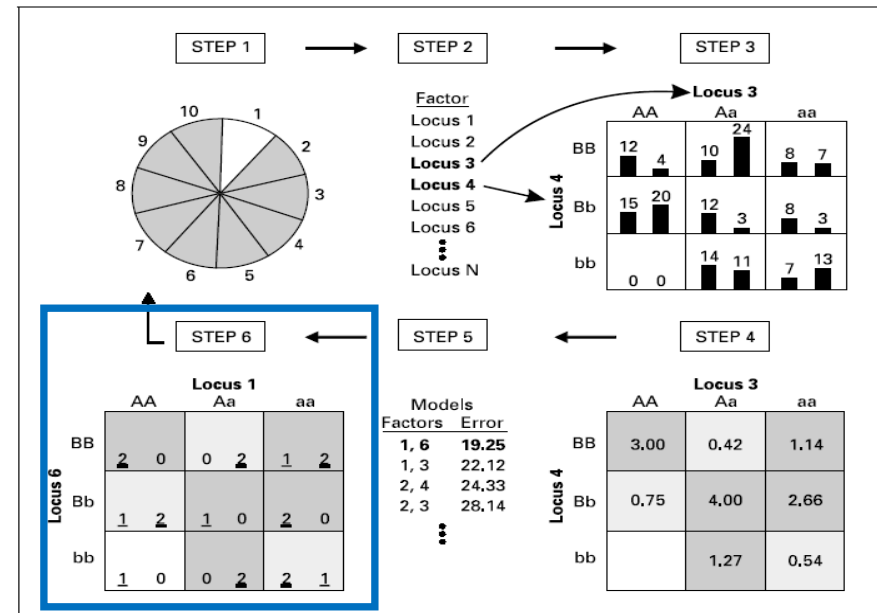
Repeat Steps 2 to 5

- 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

- An example

	Lowest TrainAcc for	TestAcc computed for	Best 2-order model
CV1	SNP1-SNP2	SNP1-SNP2	
CV2	SNP3-SNP10	SNP3-SNP10	
...			
CV10	SNP1-SNP2	SNP1-SNP2	
		Average of these values ↓ Average TestAcc	SNP3-SNP10 (because average TrainAcc over all CVs is lowest)

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

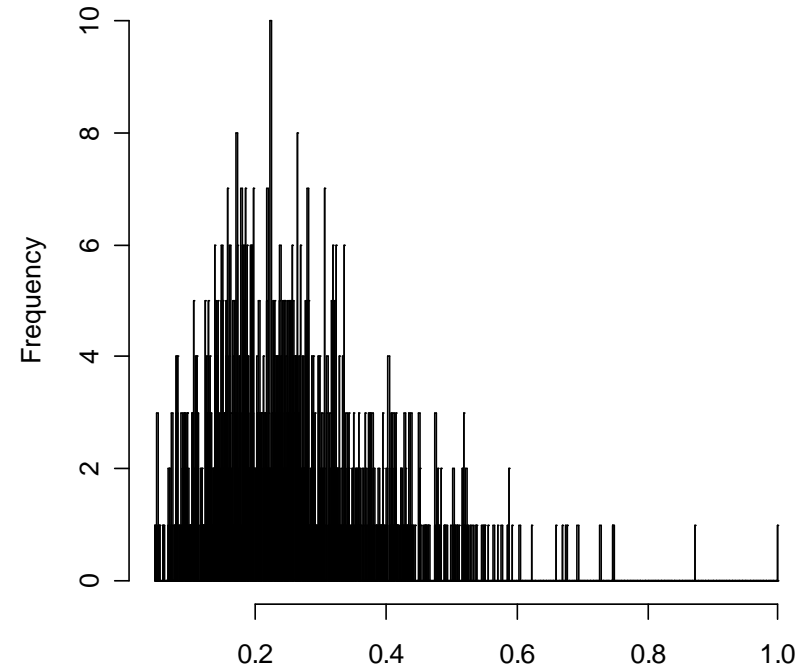
Hypothesis test of best model

- 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
- Important remark: Significance is no guarantee for the presence of epistasis!!!

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

Downloads

- Available from www.sourceforge.net
- 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).

More information can also be found at
<http://phg.mc.vanderbilt.edu/Software/MDR>

Required operating system software

Linux:

Linux (Fedora version Core 3):

Java(TM) 2 Runtime Environment, Standard Edition (build 1.4.2_06-b03)

Java HotSpot(TM) Client VM (build 1.4.2_06-b03, mixed mode)

Windows:

Windows (XP Professional and XP Home):

Java(TM) 2 Runtime Environment, Standard Edition (build v1.4.2_05)

Application to simulated data

- We simulated 200 cases and 200 controls using different multi-locus epistasis models (Evans 2006)
 - Scenario 1: 10 SNPs, adapted epistasis model M170, minor allele frequencies of disease susceptibility pair 0.5
 - Scenario 2: 10 SNPs, epistasis model M27, minor allele frequencies of disease susceptibility pair 0.25

M170

	0	1	2
0	0	0.1	0
1	0.1	0	0.1
2	0	0.1	0

M27

	0	1	2
0	0	0	0
1	0	0.1	0.1
2	0	0.1	0.1

- All markers were assumed to be in HWE. No LD between the markers.

Application to simulated data

Marginal distributions for the controls

M170	0	1	2	
0	0.07	0.12	0.07	0.25
1	0.12	0.26	0.12	0.50
2	0.07	0.12	0.07	0.25
	0.25	0.50	0.25	

M27	0	1	2	
0	0.15	0.29	0.15	0.58
1	0.10	0.17	0.09	0.36
2	0.02	0.03	0.01	0.06
	0.26	0.49	0.25	

Marginal distributions for the cases

M170	0	1	2	
0	0.00	0.25	0.00	0.25
1	0.25	0.00	0.25	0.50
2	0.00	0.25	0.00	0.25
	0.25	0.50	0.25	

M27	0	1	2	
0	0	0.00	0.00	0.00
1	0	0.57	0.29	0.86
2	0	0.10	0.05	0.14
	0.00	0.66	0.33	

Data format for MDR

- The definition of the format is as follows:
 - All fields are tab-delimited.
 - The first line contains a header row. This row assigns a label to each column of data. Labels should not contain whitespace.
 - Each following line contains a data row. Data values may be any string value which does not contain whitespace.
 - The right-most column of data is the class, or status, column. The data values for this column must be 1, to represent "Affected" or "Case" status, or 0, to represent "Unaffected" or "Control" status. No other values are allowed.

M170 case control data

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	Class
1	2	0	0	0	0	1	0	1	1	1	
1	2	1	1	0	2	0	0	0	1	1	
1	2	0	0	0	0	0	0	1	1	1	
2	1	0	0	0	0	2	2	1	0	1	
2	1	0	0	1	0	0	1	1	1	1	
...											
0	0	0	1	1	1	1	1	0	1	0	
1	1	0	0	1	2	0	1	0	0	0	
1	2	0	0	0	1	0	0	1	0	0	
2	2	0	0	0	0	1	0	2	0	0	
1	0	1	0	1	1	1	0	1	2	0	

Performing the MDR permutation test for M170 (no main effects)

	SNP5	SNP1-SNP2	SNP1-SNP2-SNP5
Testing BA (p-value)	0.5875 (0.0540)	0.7975 (<0.0010)	0.7950 (<0.0010)
CVC (p-value)	10 (0.2160)	10 (0.2160)	10 (0.2160)

Obtained from MDR summary table

Obtained from MDR Permutation Testing
p-value calculator

Performing the MDR permutation test for M170 (no main effects)

	SNP5	SNP1-SNP2	SNP1-SNP2-SNP5
Testing BA (p-value)	0.5875 (0.0540)	0.7975 (<0.0010)	0.7950 (<0.0010)
CVC (p-value)	10 (0.2160)	10 (0.2160)	10 (0.2160)

What do you think is going on???

Note:

- Testing accuracies generally go up as the order of the model increases and then start going down at some point due to false positives that are added to the model which hamper predictive ability

Performing the MDR permutation test for M27 (main effects exist)

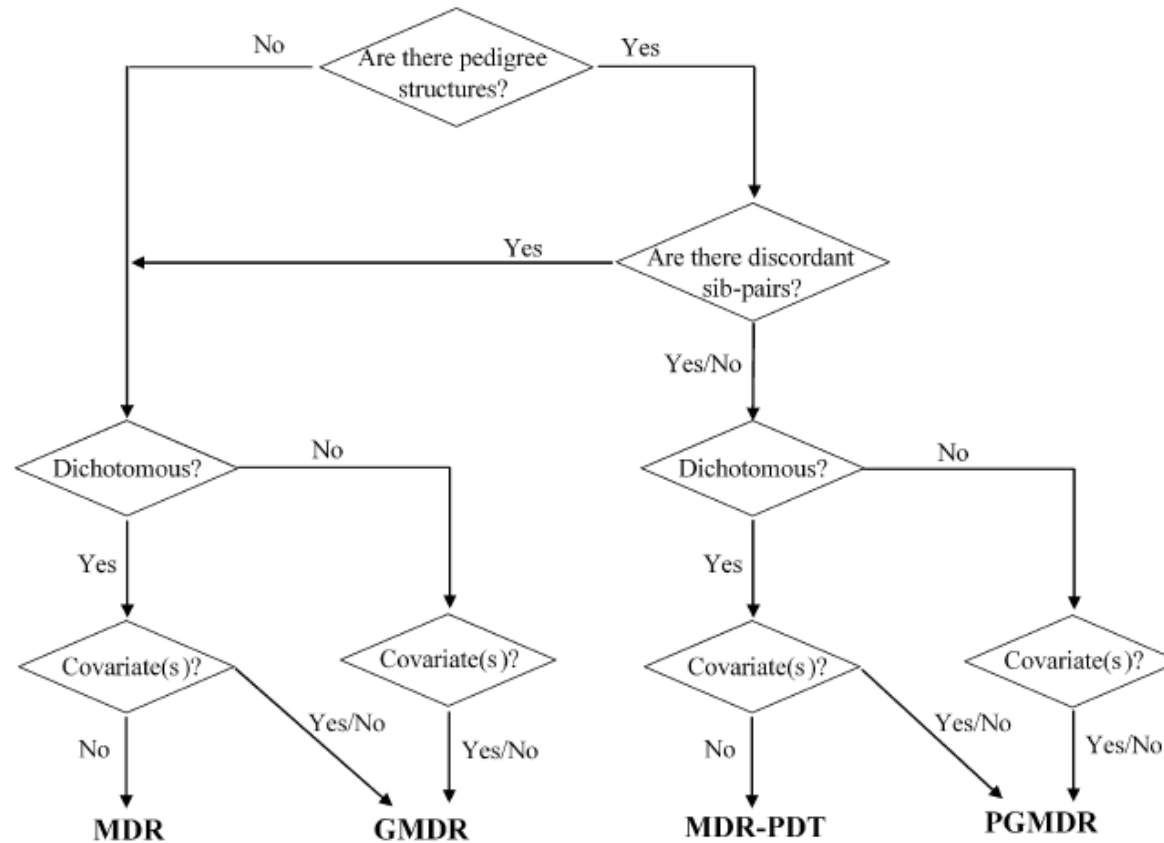
	SNP1	SNP1-SNP2	SNP1-SNP2-SNP4
Testing BA (p-value)	0.7875 (<0.0010)	0.8325 (<0.0010)	0.8600 (<0.0010)
CVC (p-value)	10 (0.1790)	10 (0.2310)	5 (0.9110)

- Maximizing CVC first and then looking at prediction accuracy highlights SNP1-SNP2. Maximizing prediction accuracy alone, would point towards SNP1-SNP2-SNP4.
- Only 1 best main effects model is outputted: what about SNP2?
- Wouldn't you rather want to correct for SNP1 when looking for 2nd order effects?

Some strengths of MDR

- Facilitates simultaneous detection and characterization of multiple genetic loci associated with a discrete clinical endpoint by reducing the dimensionality of the multi-locus data
- Non-parametric in nature, in that no model parameters are estimated
- Assumes no particular driving genetic model
- Minimal false-positive rates (assuming null data)

Continuing on the success of MDR – dealing with “issues”



(Lou et al 2008)

Some weaknesses of MDR (and aforementioned techniques, based on Cross-Validation and permutation testing)

- Computationally intensive (especially with >10 loci)
 - Parallel MDR (Bush et al 2006) is a redesign of the initial MDR algorithm to allow an unlimited number of study subjects, total variables and variable states, and to remove restrictions on the order of interactions being analyzed
 - The algorithm gives an approximate 150-fold decrease in runtime for equivalent analysis.
 - Use of extreme value distributions (Pattin et al 2009)
 - 50 times faster than 1000-fold permutation testing
 - MDR-GPU allows to run MDR on a genome-wide dataset and hence for statistically rigorous testing of epistasis (Greene et al 2010)

Some weaknesses of MDR

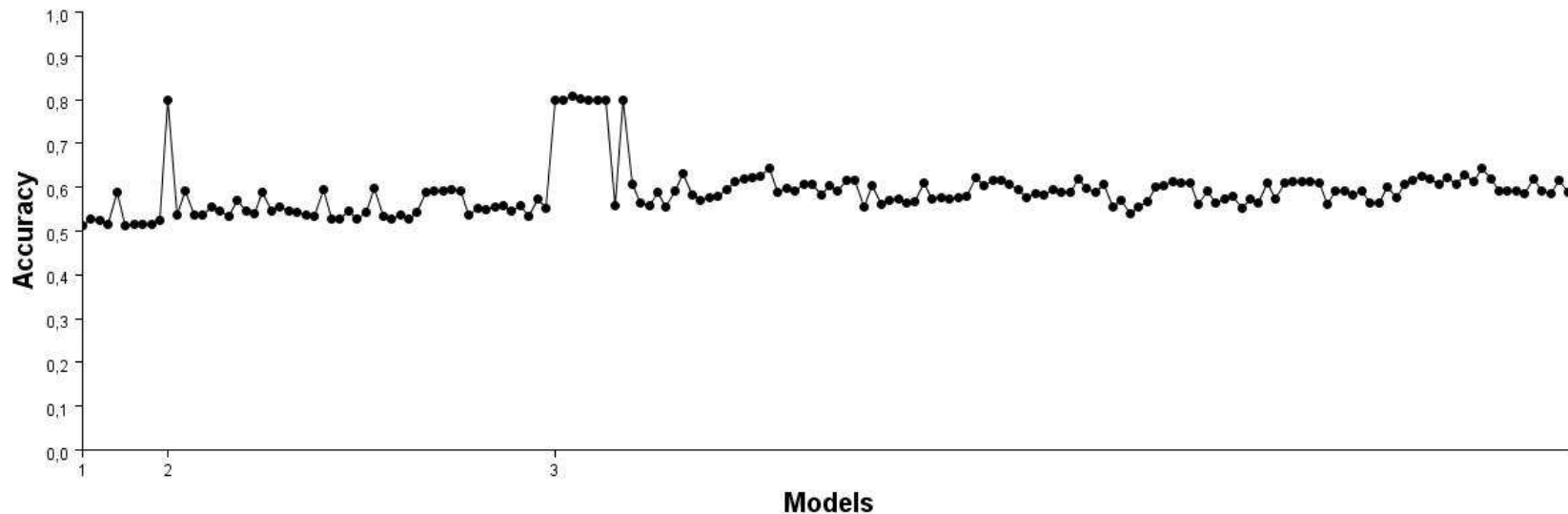
- Inability to adjust for important confounders in a flexible way
 - Implementation of rather simple, yet computationally efficient, sampling method to adjust for covariate effects in MDR (Gui et al 2010)
- Inability to readily distinguish between global and interaction-specific effects
 - Regression-based permutation test procedure that does not reject the null when only main effects are present (Edwards et al 2010).
 - Novel permutation test that allows the effects of nonlinear interactions between multiple genetic variants to be specifically tested in a manner that is not confounded by linear additive effects (Greene et al 2009)

Some weaknesses of MDR

- Inability to incorporate several outcomes (at once) and merge different study designs (family-based / population-based
 - Modifying of MDR's constructive induction algorithm to use the log-rank test, hereby accommodating survival type of outcomes (Gui et al 2010).
- Poor performance in the presence of genetic heterogeneity
 - Although the MDR authors claim that genetic heterogeneity may not be as threatening as you think (Ritchie et al 2007), even with the most optimal settings of MDR, the power of MDR suffers in the presence of locus heterogeneity (Cattaert et al 2010)

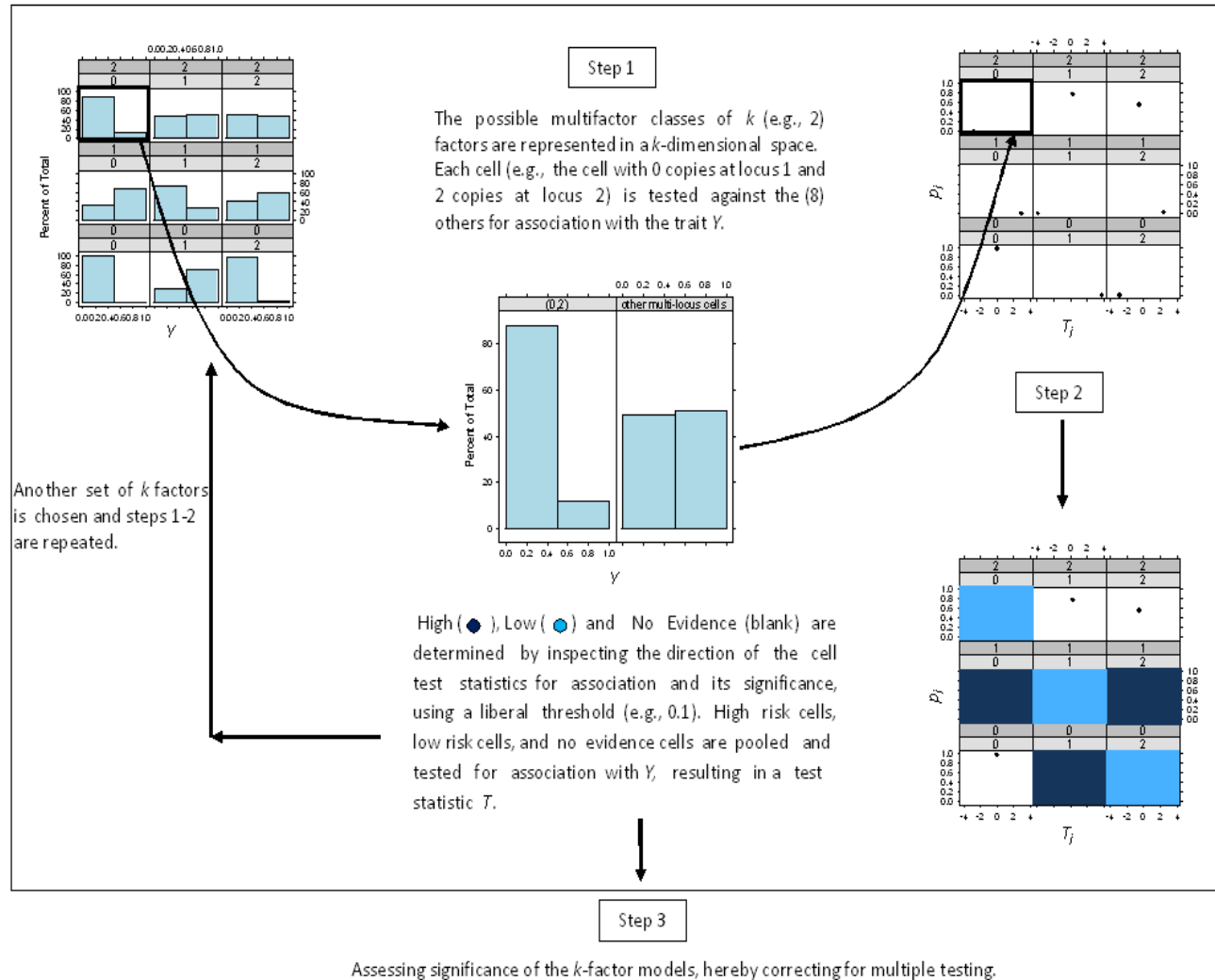
Some weaknesses of MDR

- Single best model, whereas in reality there might be several competing models present



Fitness landscape: The models produced are on the x-axis of the chart. The models on the x-axis are in the order in which they were generated (e.g., 1,2,3, ..., 12, 13, 14, ...). Training accuracy is shown on the y-axis.

MB-MDR



Characteristics of MB-MDR

- MB-MDR aims to identify the most significant associations (possibly more than one) between groups of markers and the trait of interest.
 - In contrast, MDR identifies a single best model on the basis of measures of prediction accuracy and cross-validation consistency.
- Besides making it possible to detect multiple models, the use of association models in MB-MDR, rather than prediction accuracy and cross-validation consistency as in MDR, seems to be beneficial in itself, in that it leads to a better performance, both in terms of controlling false positives and in terms of achieving adequate power, in most of the studied simulated settings.

False positive percentages under alternatives

(Cattaert et al 2010)

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

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

Motivation 1 for MB-MDR

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

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

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

H: High risk; L: Low risk

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

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

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

(Calle et al 2008)

Motivation 2 for MB-MDR

- MDR cannot deal with main effects / confounding factors / non-dichotomous outcomes

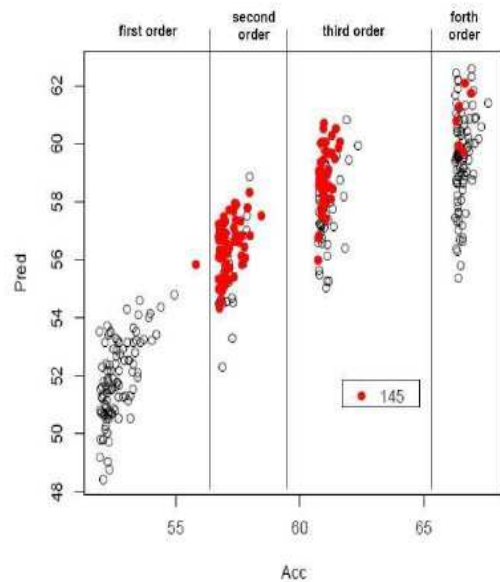


Fig. 1. Average Balanced Training accuracy (Acc) versus Average Balanced Predictive accuracy (Pred) for the 100 models with higher balanced training accuracy for the whole sample. First, second, third and fourth order interactions are considered.

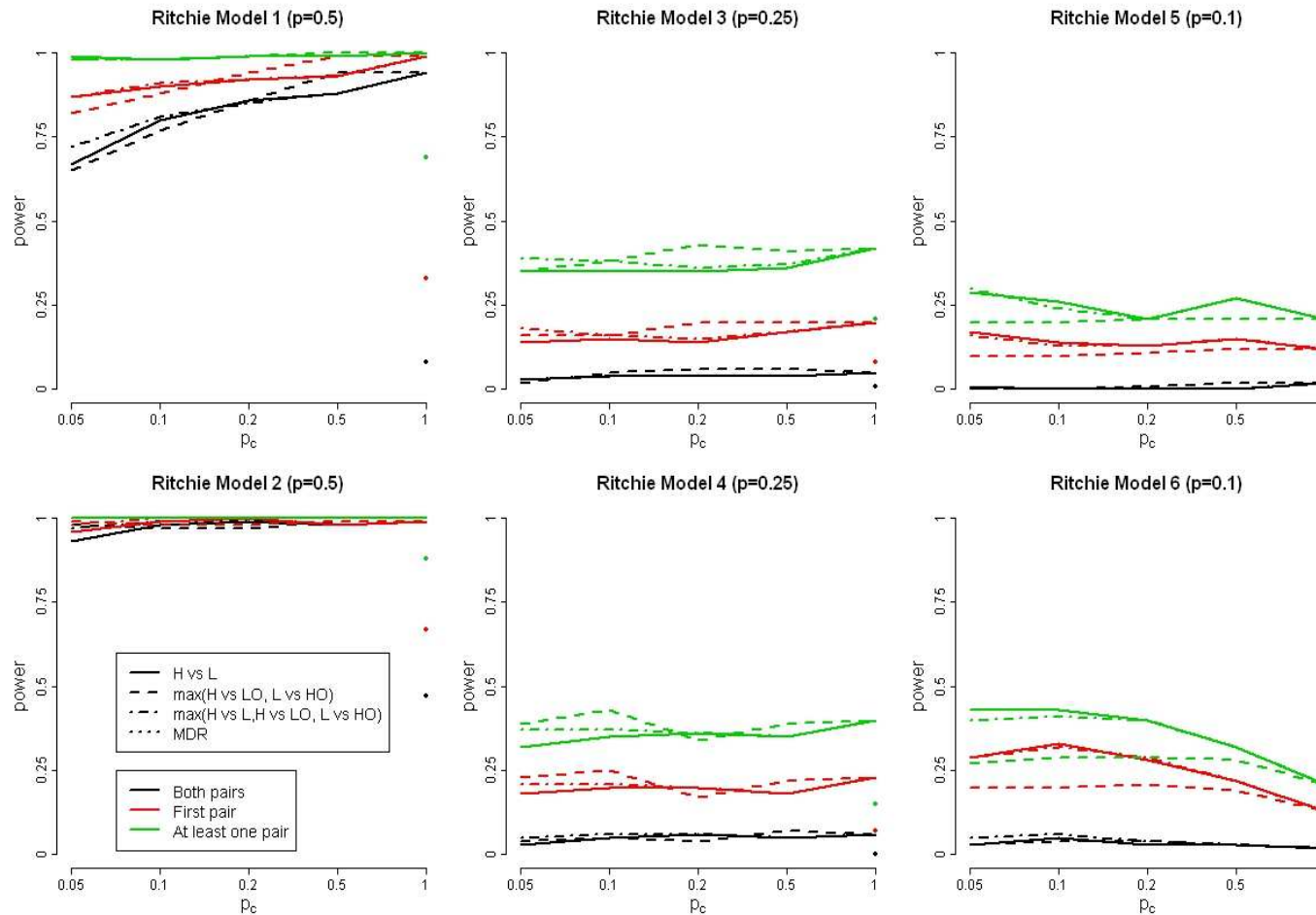
Table 2. First, second and third order significant interactions identified by MDR in the bladder cancer study

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

Motivation 3 for MB-MDR

(Cattaert et al 2010)

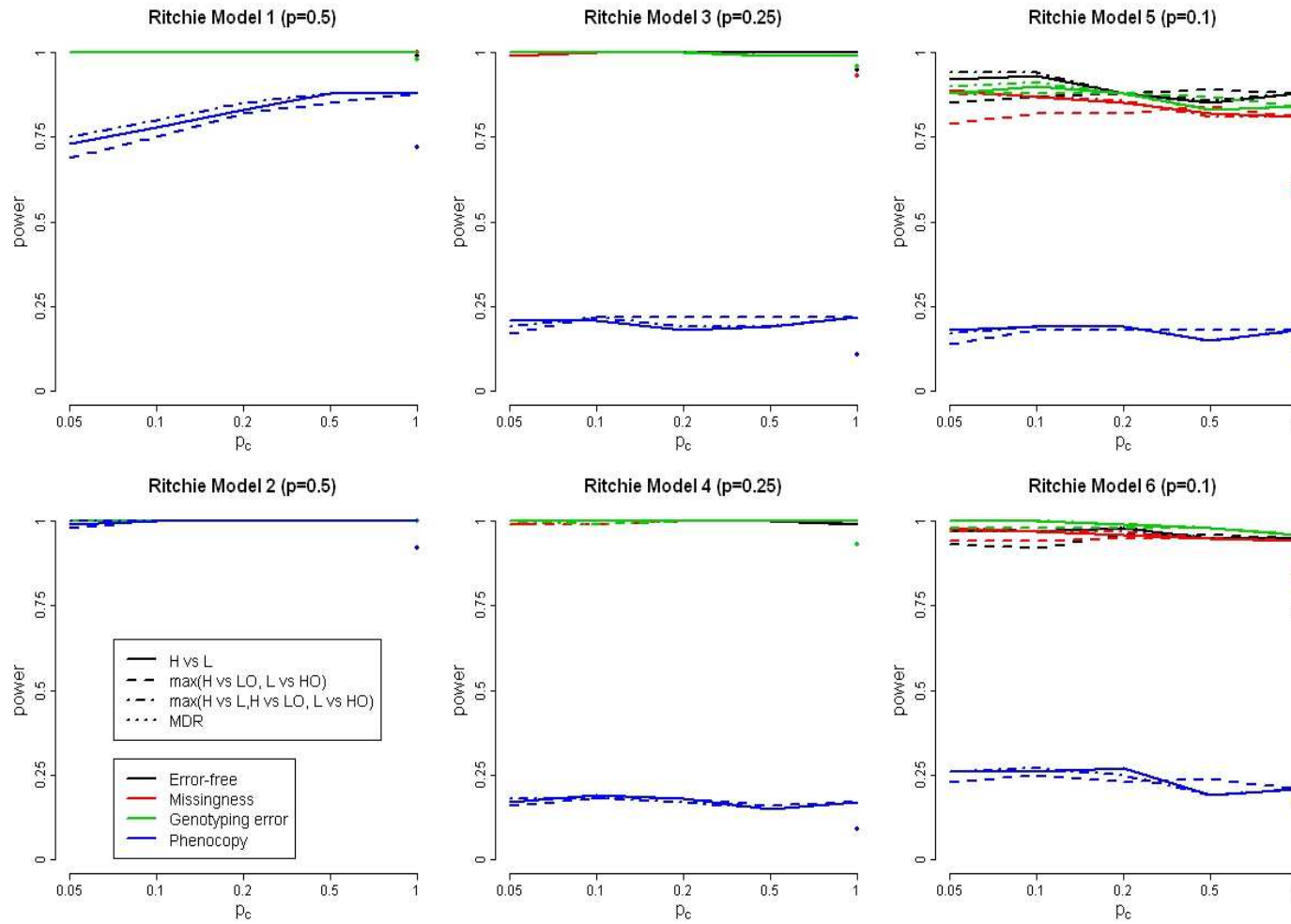
- MDR has low performance in the presence of genetic heterogeneity



Motivation 4 for MB-MDR

(Cattaert et al 2010)

- Maximize power for the already “difficult” epistasis screens



The MB-MDR Software

Downloads

- The MDR method is described in further detail by Calle et al (2008), Cattaert et al (2010a,b).
- A simplified version of MB-MDR is available in the free software R as an mbmdr package (<http://cran.r-project.org/>) and described in Calle et al (2010)
- A comprehensive MB-MDR executable file is available from Van Steen (kristel.vansteen@ulg.ac.be) or via www.statgen.be

Required operating system software

- Same as MDR, no JAVA requirements

Application to simulated data (introduced before)

- The required MB-MDR data format is as follows:
 - All fields are **space**-delimited.
 - The first line contains a header row. This row assigns a label to each column of data. Labels should not contain whitespace.
 - Each following line contains a data row. Data values may be any string value which does not contain whitespace.
 - The **left-most column** of data is the disease status column or continuous trait column.
 - For binary traits, the data values for this column must be 1 ("Affected"), or 0 ("Unaffected").
 - For continuous traits, the data values can be any real number.
 - Missing trait values are indicated by NA
 - Missing genotypes are indicated by -9.

M170 case control data for MB-MDR

Trait1 SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10

1	1	2	0	0	0	1	0	1	1	
1	1	2	1	1	0	2	0	0	0	1
1	1	2	0	0	0	0	0	0	1	1
1	2	1	0	0	0	0	2	2	1	0
1	2	1	0	0	1	0	0	1	1	1
1	0	1	0	0	1	0	1	1	0	1
1	2	1	0	0	0	0	0	0	1	0
1	1	0	1	0	0	0	2	1	1	1
1	1	2	0	0	0	0	1	0	1	1
1	0	1	1	1	0	1	2	1	1	1

MB-MDR (1 dimension): M170 (no main effects)

options="--maxT --sequential --binary --hlo-mode --one-cell-approach -c 0.1 --two-tests -d 1 -r 1969 -p 999 -n 3" # see MBMDR.cpp for the different possible options

SNP	Chi-square	pValue
SNP5	13.032	0.006
SNP2	0	1
SNP1	0	1

- Stepwise logistic regression (order 1): $\text{Trait1} \sim \text{SNP5}$
- Stepwise logistic regression (order 2): $\text{Trait1} \sim \text{SNP1} + \text{SNP3} + \text{SNP4} + \text{SNP5} + \text{SNP6} + \text{SNP7} + \text{SNP8} + \text{SNP9} + \text{SNP10} + \text{SNP1:SNP3} + \text{SNP1:SNP4} + \text{SNP1:SNP5} + \text{SNP1:SNP6} + \text{SNP4:SNP10} + \text{SNP5:SNP7} + \text{SNP6:SNP8} + \text{SNP7:SNP9}$

MB-MDR (1 dimension): M27 (main effects present)

options="--maxT --sequential --binary --hlo-mode --one-cell-approach -c 0.1 --two-tests -d 1 -r 1969 -p 999 -n 3" # see MBMDR.cpp for the different possible options

SNP	Chi-square	pValue
SNP1	161.404	0.001
SNP2	62.427	0.001
SNP7	6.300	0.212

- Stepwise logistic regression (order 1): Trait1 ~ SNP1 + SNP2 + SNP10
- Stepwise logistic regression (order 2): Trait1 ~ SNP1 + SNP2 + SNP4 + SNP5 + SNP7 + SNP8 + SNP9 + SNP10 + SNP1:SNP2 + SNP2:SNP4 + SNP2:SNP5 + SNP2:SNP10 + SNP5:SNP9 + SNP7:SNP8 + SNP7:SNP10

Remark

- There seems to be a tendency for logistic regression to be overly optimistic
- 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

MB-MDR (2 dimensions): M170 (no main effects)

options="--maxT --sequential --binary --hlo-mode --one-cell-approach -c 0.1 --two-tests -d 2 -r 1969 -p 999 -n 3" # see MBMDR.cpp for the different possible options

FirstSNP	SecondSNP	Chi-square	pValue
SNP1	SNP2	169.395	0.001
SNP4	SNP5	15.947	0.051
SNP5	SNP7	14.092	0.118

MB-MDR (2 dimensions): M27 (main effects present, but not accounted for)

FirstSNP	SecondSNP	Chi-square	pValue
SNP1	SNP2	199.251	0.001
SNP1	SNP10	161.404	0.001
SNP1	SNP9	161.404	0.001
SNP1	SNP6	161.404	0.001
SNP1	SNP7	161.404	0.001
SNP1	SNP5	161.404	0.001
SNP1	SNP8	161.404	0.001
SNP1	SNP4	161.404	0.001
SNP1	SNP3	159.441	0.001
SNP2	SNP9	62.427	0.001
SNP2	SNP10	62.427	0.001
SNP2	SNP3	62.427	0.001
SNP2	SNP4	62.427	0.001
SNP2	SNP5	62.427	0.001
SNP2	SNP8	62.427	0.001
SNP2	SNP6	61.095	0.001
SNP2	SNP7	59.770	0.001
SNP7	SNP10	11.470	0.204



Corrected MB-MDR (2 dimensions): M170 (no main effects)

```
options="--maxT --sequential --continuous --hlo-mode --one-cell-approach -c 0.1 --two-tests -d 2 -r 1969 -p 999 -n 20" # see MBMDR.cpp for the different possible options
```

- Does signal for interaction weaken after adjusting for SNP5?

FirstSNP	SecondSNP	<u>F-test</u>	pValue
SNP1	SNP2	261.815	0.001
SNP3	SNP8	8.608	0.71
SNP4	SNP10	6.215	0.958

Corrected MB-MDR (2 dimensions): M170 (no main effects)

```
options="--maxT --sequential --binary --ajust1-mode -c 0.1 --co-dominant -- two-tests -d 2 -r  
1969 -p 999 -n 20" # see MBMDR.cpp for the different possible options
```

- Does signal for interaction weaken after adjusting for the components in the pair we are looking at?

FirstSNP	SecondSNP	<u>Chi-square</u>	pValue
SNP1	SNP2	166.6	0.001
SNP8	SNP10	5.651	0.445
SNP4	SNP10	5.463	0.472

Corrected MB-MDR (2 dimensions): M27 (main effects present)

- Corrected with “genotype” coding for SNP1 and SNP2

FirstSNP	SecondSNP	F-test	pValue
SNP1	SNP6	21.360	0.009
SNP1	SNP4	19.295	0.025
SNP1	SNP2	19.178	0.028

- Corrected with “additive” coding for SNP1 and SNP2

FirstSNP	SecondSNP	F-test	pValue
SNP1	SNP6	158.125	0.001
SNP1	SNP7	109.945	0.002
SNP1	SNP9	106.667	0.002
SNP1	SNP3	99.847	0.002

Corrected MB-MDR (2 dimensions): M27 (main effects present)

options="--maxT --sequential --binary --ajust1-mode -c 0.1 --co-dominant -- two-tests -d 2 -r 1969 -p 999 -n 20" # see MBMDR.cpp for the different possible options

FirstSNP	SecondSNP	Chi-square	pValue
SNP5	SNP9	5.362	0.477
SNP4	SNP7	5.154	0.517
SNP7	SNP10	4.112	0.796
SNP4	SNP10	3.176	0.979
SNP7	SNP8	3.068	0.985

• Conclusions:

- It DOES matter how to correct for lower order effects – this motivates the use of “semi-parametric” techniques
- Correcting for lower-order effects need to be part of the entire epistasis screening process

Part 12

Interpretation of identified gene-gene interactions

Main challenges in epistasis detection

- Variable selection
- Modeling

- Interpretation

- Making inferences about biological epistasis from statistical epistasis

Technical and conceptual constraints

- The presence and magnitude of non-additivity are scale and model dependent, so that in principle, one strategy in the context of epistasis could be to remove any non-additivity by a transformation prior to data analysis, followed by a back-transformation to the original scale for easy interpretation (Wang et al 2010).
- There is a conceptual discrepancy between genetical and biological epistasis (both occurring at the individual level) and statistical epistasis (occurring as the result of differences in genetical and biological epistasis among individuals in a population) (Moore 2005).

Technical and conceptual constraints

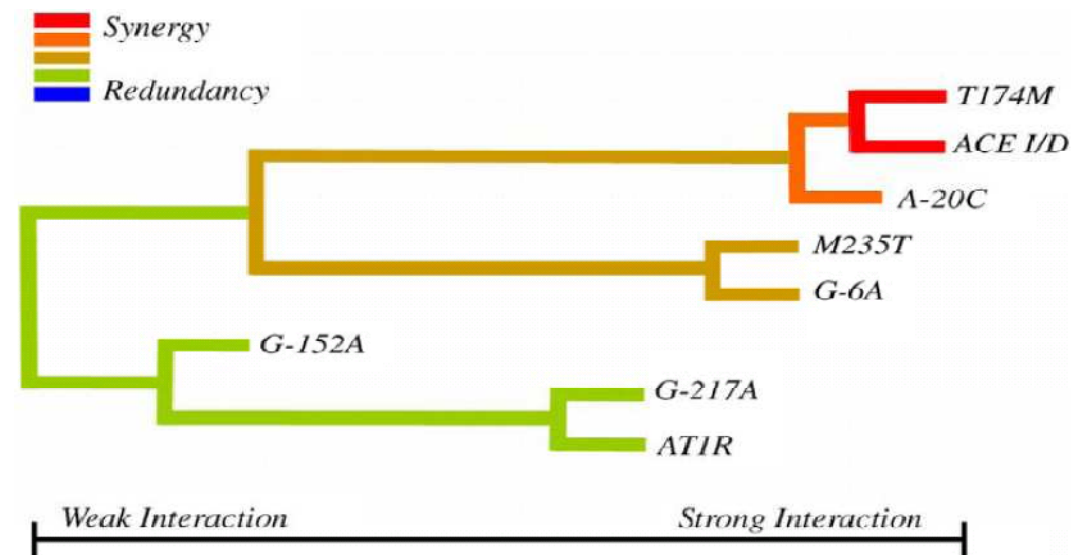
- In general, regression-based statistical tests for interaction are of limited use in detecting "epistasis" in the sense of masking (Cordell 2009).
- To this regard, the concept of "compositional epistasis" may be more useful (is said to be present when the effect of a genetic factor at one locus is masked by a variant at another locus - Phillips 2008).
 - VanderWeele (2010a,b,c) consider empirical tests for "compositional epistasis" under models for the joint effect of two genetic factors which place no restrictions on the main effects of each factor but constrain the interactive effects of the two factors so as to be captured by a single parameter in the model.

Advocating a flexible framework for analysis acknowledging interpretation capability

- The framework proposed by Moore and colleagues (2005) 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

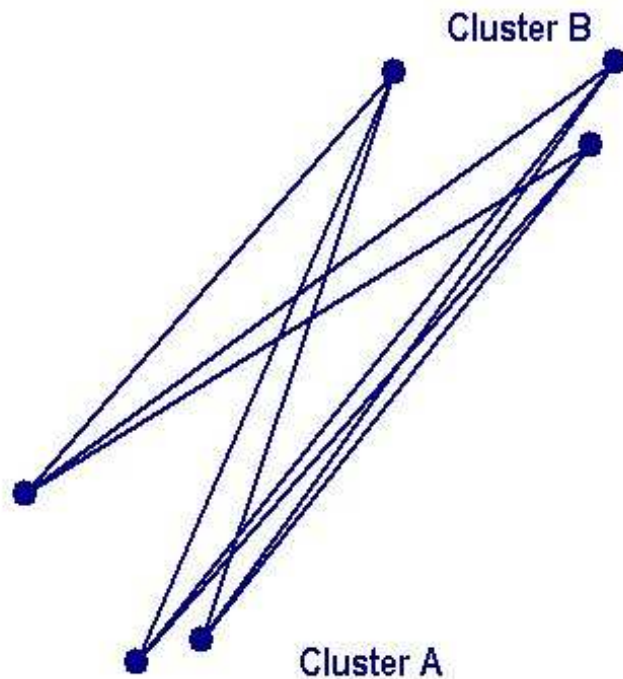
Example of a visual method: the interaction dendrogram

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



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)

In conclusion

- Inclusion of -omics data bases,
- assessment and incorporation of “optimal” scoring systems to accumulate evidence from these data bases,
- possibly allowing for uncertainty involved in the data source entries,
- acknowledging the complementary characteristics of each of the available data sources,
- and allowance for different assignment strategies from genetic variants to genes,

are only some of the components of a biology assistant-driven approach, incorporated in this epistasis analysis framework, that need careful thought, but will be hard to avoid in the future ...

In conclusion

- Never cease to plug and play ... within a reasonable context ...



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