Genome-wide Association Studies

Kristel Van Steen & Andreas Ziegler

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

Florianopolis, Brazil

IBC 2010

Université de Liège

UNIVERSITÄT ZU LÜBECK

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

- Data dimensionality reduction methods (with emphasis on MDR)
- Tree-based methods (with emphasis on random forests and random jungle)
- Adjustment for confounding factors

12 Interpretation of identified interactions (entropy-based interaction graphs)

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
- Acquired essential background to overcome some of the hurdles

Part 7

Curse of dimensionality and multiple testing

What is the general setting?

Introduction

- Multiple testing is a thorny issue, the bane of statistical genetics. The problem is not really the number of tests that are carried out: 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.

  (Balding 2006)

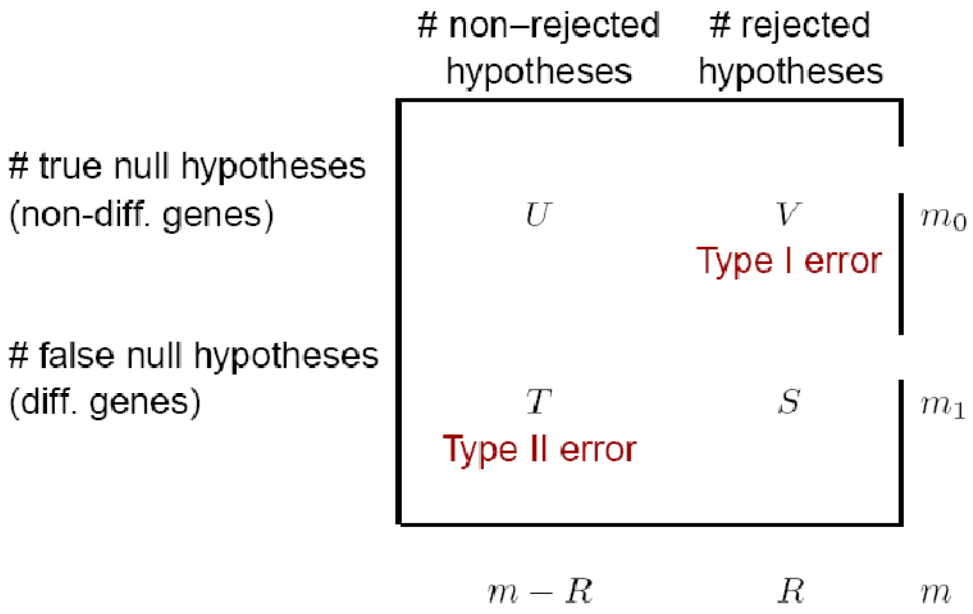
- The genome is large and includes many polymorphic variants and many possible disease models. Therefore, any given variant (or set of variants) is highly unlikely, *a priori*, to be causally associated with any given phenotype under the assumed model. Strong evidence is required to overcome the appropriate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

The multiple testing problem

- Simultaneously test G null hypotheses, one for each SNP *j*

  H_j: no association between SNP *j* and the trait
- Because GWAs simultaneously harbor SNP effects of thousands of genes, there is a large multiplicity issue
- We would like some sense of how 'surprising' the observed results are

False positive rates



- PFER = E(V) → Per-family error rate (note that it is not a rate …)
- PCER = E(V)/m → Per-comparison error rate
- FWER = p(V ≥ 1) → Family-wise error rate
- FDR = E(Q), where Q = V/R if R > 0; Q = 0 if R = 0

Popular ways to control type I error in GWA settings?

- Family-wise error rate (FWER)
- Permutation data sets
- False discovery rate (FDR) and variations thereof
- Bayesian methods such as false-positive report probability (FPRP)

Family-wise error rate (FWER)

- The frequentist paradigm of controlling the overall type-1 error rate sets a significance level α (often 5%), and 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, in part because it was the analysis that was planned by the investigator that matters, not only the analyses that were actually conducted.

The Bonferroni correction

- In simple settings the frequentist approach gives a practical prescription:
  - 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.

Permutation data sets

- For tightly linked SNPs, the Bonferroni correction is conservative.
- A practical alternative is to approximate the type-I error rate using a permutation procedure.
- 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 data sets

- 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.
- 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.

Permutation based control

- If 1000 permutations are specified, then all 1000 will be performed, for all SNPs.
- Two sets of empirical significance values can then be calculated
  - pointwise estimates of an individual SNPs significance,
  - a value that controls for that 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 step down max(T) procedure

1. Permute the n columns of the data matrix X.
2. Compute test statistics $t_{1,b}, \ldots, t_{m,b}$ for each hypothesis.
3. Next, compute successive maxima of the test statistics

$$u_{m,b} = |t_{r_m,b}|,$$
$$u_{j,b} = \max\left(u_{j+1,b}, |t_{r_j,b}|\right) \qquad \text{for } j = m-1, \ldots, 1,$$

where r_j denotes the ordering of the *observed* test statistics such that $|t_{r_1}| \geq |t_{r_2}| \geq \cdots \geq |t_{r_m}|$.

The adjusted p-values are estimated by

$$\tilde{p}^*_{r_j} = \frac{\sum_{b=1}^{B} I(u_{j,b} \geq |t_{r_j}|)}{B},$$

Université de Liège

UNIVERSITÄT ZU LÜBECK

The 5% magic percentage

- Although the 5% global error rate is widely used in science, it is inappropriately conservative for large-scale SNP-association studies:
  - Most researchers would accept a higher risk of a false positive in return for greater power.
- There is no “rule” saying that the 5% value cannot be relaxed
- Another approach is to monitor the false discovery rate (FDR) instead
- The FDR refers to the *proportion of false positive test results among all positives*.
  - FDR = E(Q), where Q = V/R if R > 0; Q = 0 if R = 0
  - FDR = E[V/R | R>0] . prob(R>0) (Benjamini and Hochberg 1995)
  - pFDR = E[V/R | R>0] (Storey 2001)

False discovary rate (FDR)

- Hence, FDR measures come in different shapes and flavor.
  - But 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.
  - Assumptions about the alternative hypothesis might be required for the most powerful methods, but the simplest procedures avoid making these explicit assumptions.

FDR in Bayesian terms

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

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

Algorithms for calculating $\widehat{FDR}$ and $\widehat{pFDR}$ in Storey (2001b).

(slide Stefanie Scheid 2002)

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, even though these might reveal interesting associations, because all their analyses would then suffer a multiple-testing penalty.
- It is a matter of common sense that expensive and hard-won data should be investigated exhaustively for possible patterns of association.

Bayesian methods

- Although the frequentist paradigm is convenient in simple settings, strict adherence to it can be dangerous: true associations may be missed!
- 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.

Do these classical methods hold up in GWA settings?

- Family-wise error rate (FWER)
  - Bonferroni Threshold: $< 10^{-7}$
  - In the presence of too many tests, the Bonferroni threshold will be extremely low:
    - 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?

- Permutation data sets
  - Enough compute capacity?
  - 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 …

- False discovery rate (FDR) and variations thereof
  - Start to break down …
  - The power over Bonferroni is minimal (e.g. see Van Steen et al 2005)

Do these classical methods hold up in GWA settings?

- Bayesian methods such as false-positive report probability (FPRP)
  - In general, Bayesian approaches do not yet have a big role in genetic association analyses, possibly because of computational burden and/or choice of prior?
  - Works though, but for now not extremely well documented

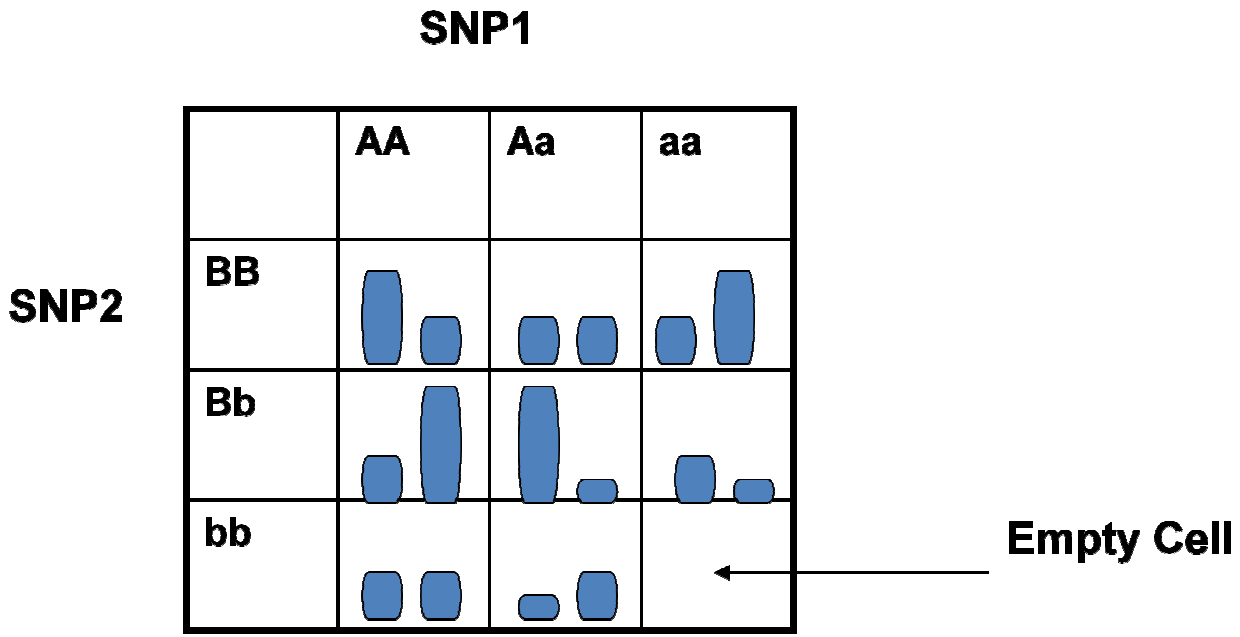
(Balding 2006; Lucke 2008)

  - FPRP = the probability of no true association between a genetic variant and disease given a statistically significant finding
  - FRPP depends not only on the observed p-value but also on both the prior probability that the association between the genetic variant and the disease is real and the statistical power of the test

(Wacholder et al 2004)

The curse of dimensionality

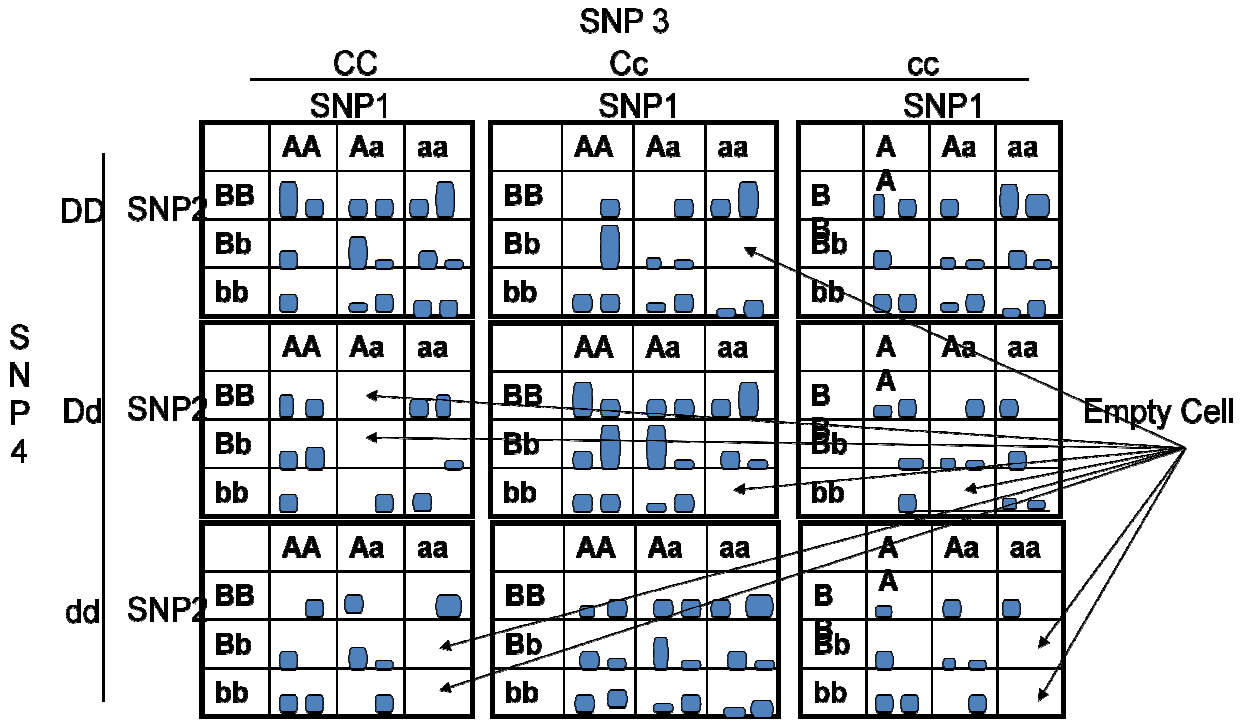
- For 2 SNPs, there are $9 = 3^2$ possible two locus genotype combinations.
- If the alleles are rare (MAF≤10%), then some cells will be empty



(slide: C Amos)

The curse of dimensionality when looking for interaction effects

- For 4 SNPs, there are 81 possible combinations with more possible empty cells …



(slide: C Amos)

Part 8

Missing data

How to deal with missing genotypes?

Introduction

- 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.

Imputation

- One convenient solution is data imputation
  - Data imputation involves replacing missing genotypes with predicted values that are based on the observed genotypes at neighbouring SNPs.
- For tightly linked markers data imputation can be reliable, can simplify analyses and allows better use of the observed data.
- For untightly linked markers?

Imputation

- Imputation methods either seek a best prediction of a missing genotype, such as a
  - maximum-likelihood estimate (single imputation), or
  - randomly select it from a probability distribution (multiple imputations).

- The advantage of the latter approach is that repetitions of the random selection can allow averaging of results or investigation of the effects of the imputation on resulting analyses.

Can improper missing genotype handling induce bias?

- Yes !!!

- Beware of settings in which cases are collected differently from controls. These can lead to differential rates of missingness even if genotyping is carried out blind to case-control status.
  - One way to check differential missingness rates is to code all observed genotypes as 1 and unobserved genotypes as 0 and to test for association of this variable with case-control status …

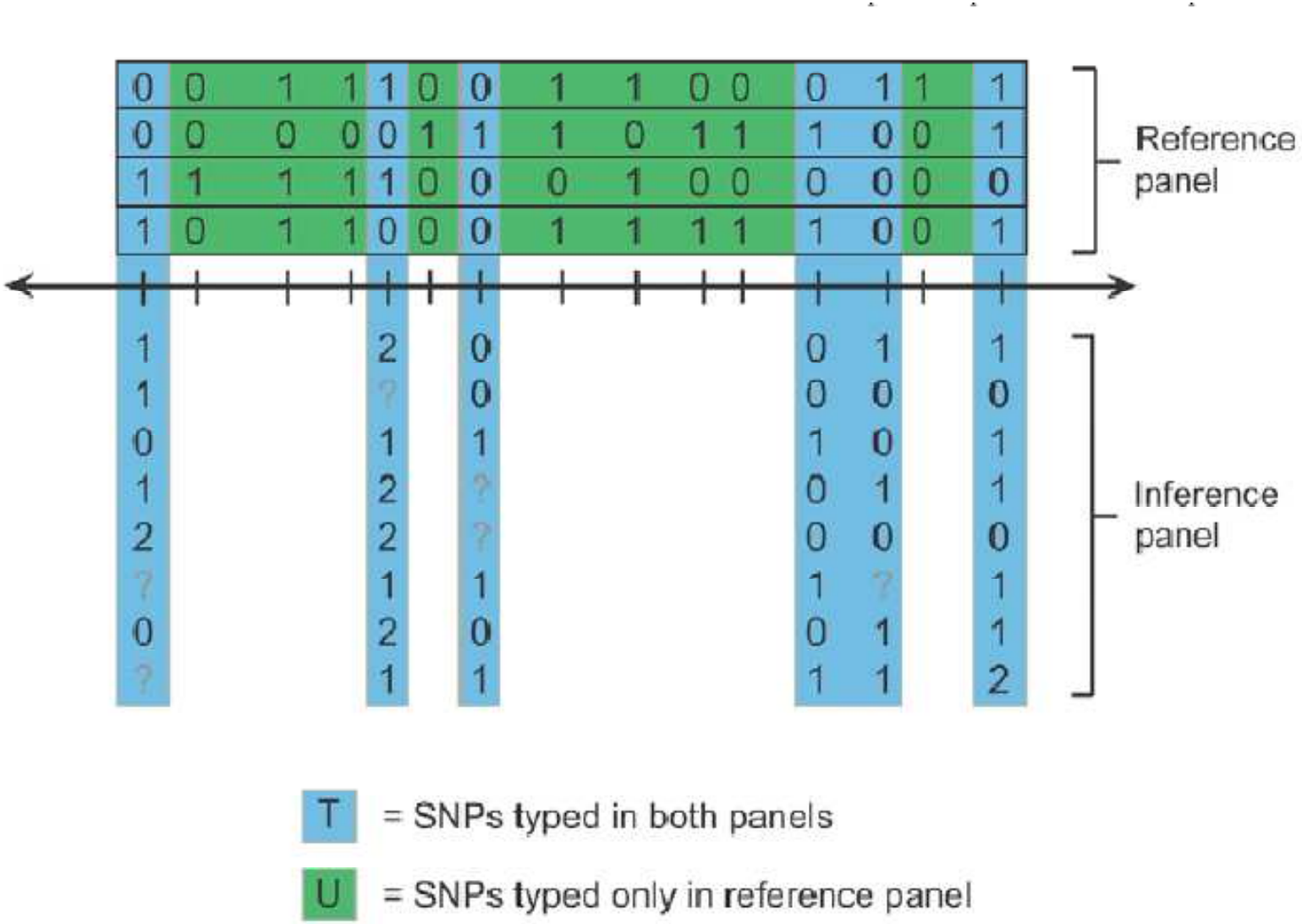
Software packages for imputation

within a sample of unrelated individuals

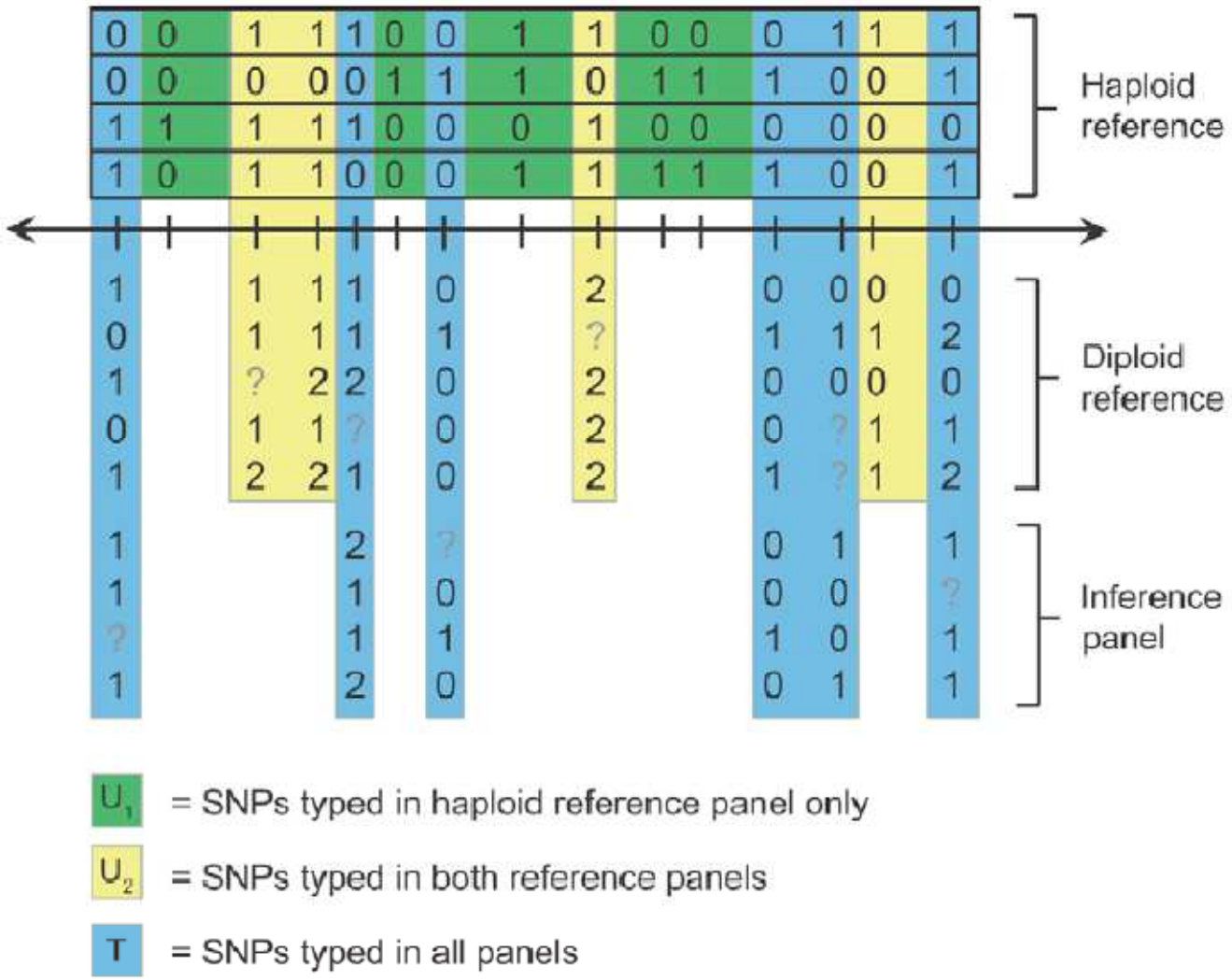
1. Impute	Oxford University
2. Plink	Massachusetts General Hospital / Broad Institute
3. Mach	University of Michigan
4. Beagle	University of Auckland

1. http://www.sph.umich.edu/csg/abecasis/MaCH/tour/imputation.html
2. http://pngu.mgh.harvard.edu/~purcell/plink/
3. http://www.sph.umich.edu/csg/abecasis/MaCH/tour/imputation.html
4. http://www.stat.auckland.ac.nz/~bbrowning/beagle/beagle.html

(Illumina technical note 2010)



(IMPUTE_v2: Howie et al 2009)



(IMPUTE_v2: Howie et al 2009)

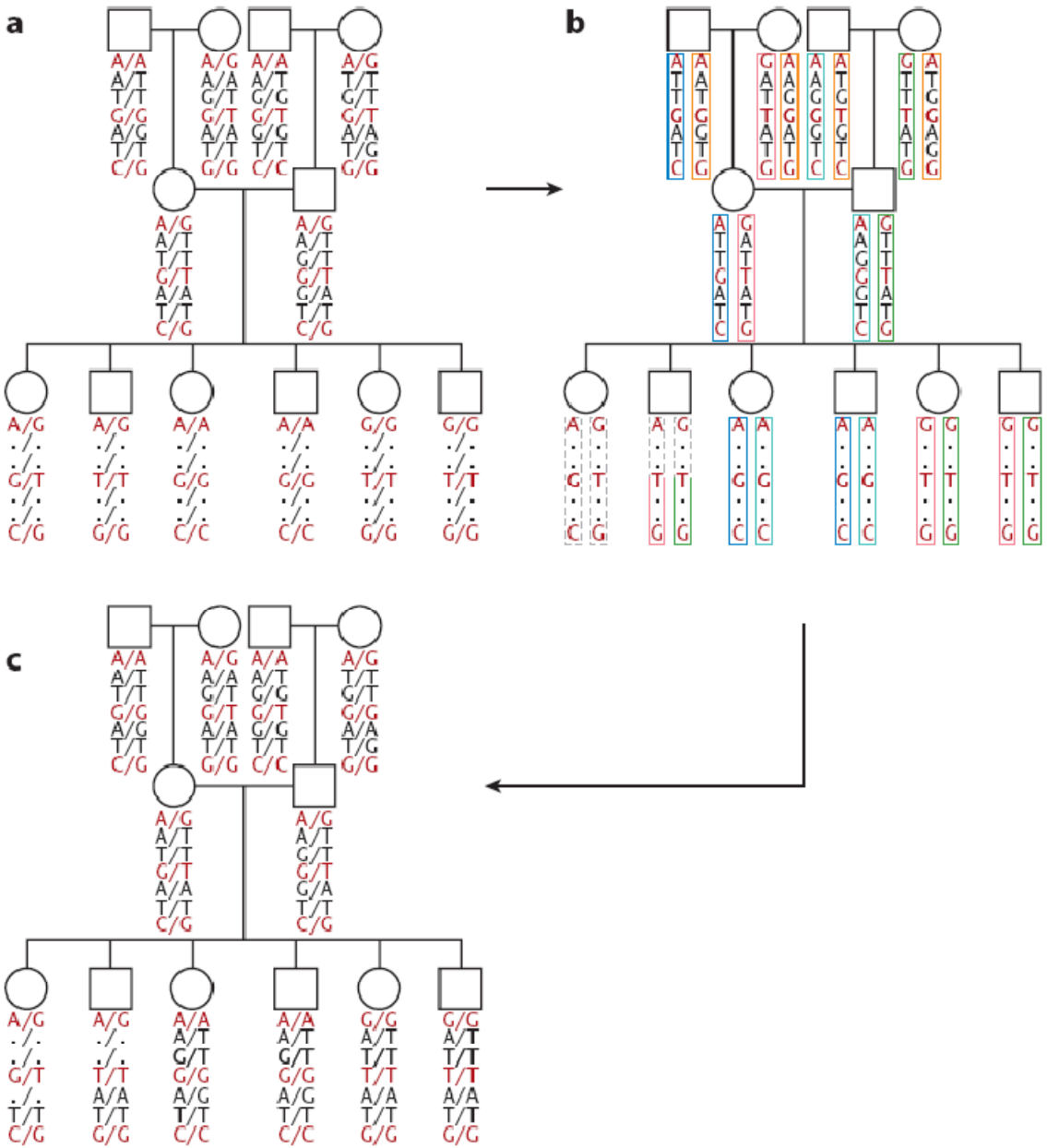
Software packages for imputation (relateds)

within a sample of related individuals

5. Merlin	University of Michigan
6. Mendel	University of California

5. http://www.sph.umich.edu/csg/abecasis/merlin/tour/assoc.html
6. http://www.genetics.ucla.edu/software/mendel

(Li et al 2009)

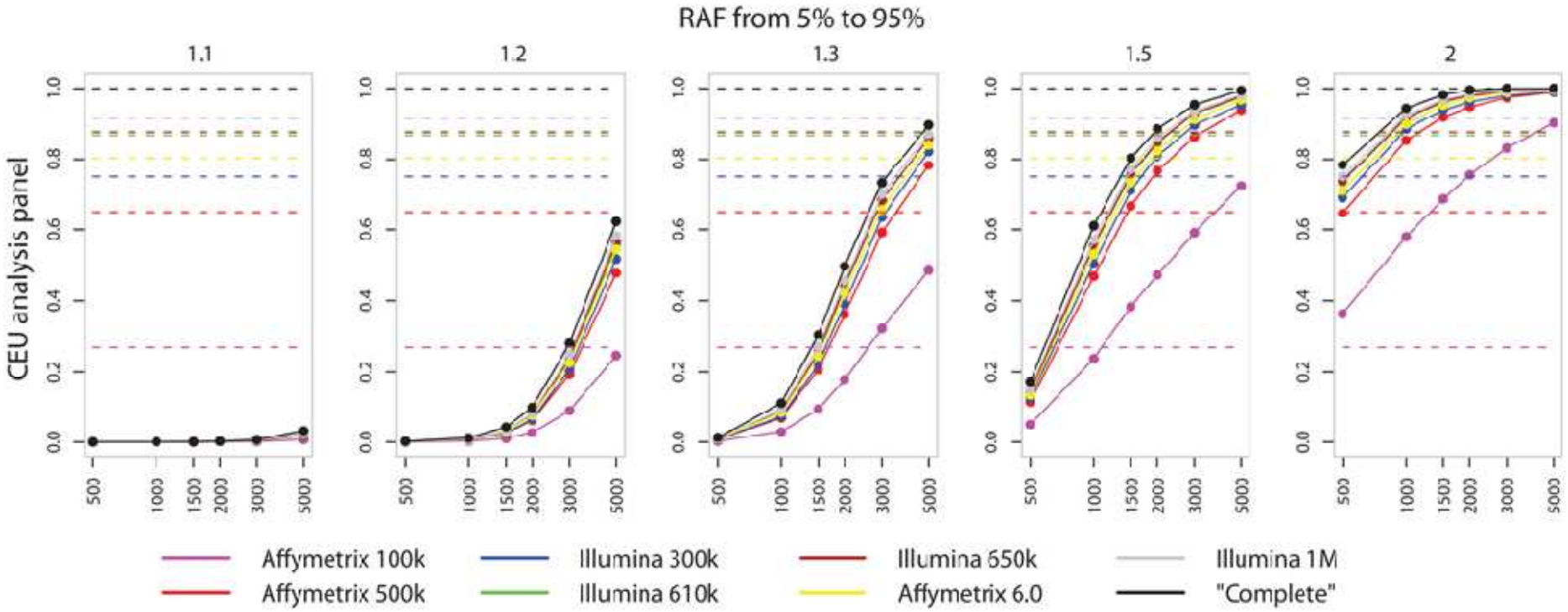


(Li et al 2009)

Software packages for imputation

Technical reports from companies providing genotyping platforms (E.g., ILLUMINA) usually contain lots of information on computational requirements to perform imputations….

Does the power of your GWA increase when imputing?



(Spencer et al 2009)

Part 9

Variable selection methods

Why selecting variables?

Introduction

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

What are different flavors of variable selection?

- Identify linkage disequilibrium blocks according to some criterion and infer and analyze haplotypes within each block, while retaining for individual analysis those SNPs that do not lie within a block

- Multi-stage designs …

- Pre-screening for subsequent testing:
  - Independent screening and testing step (PBAT screening)
  - Dependent screening and testing step

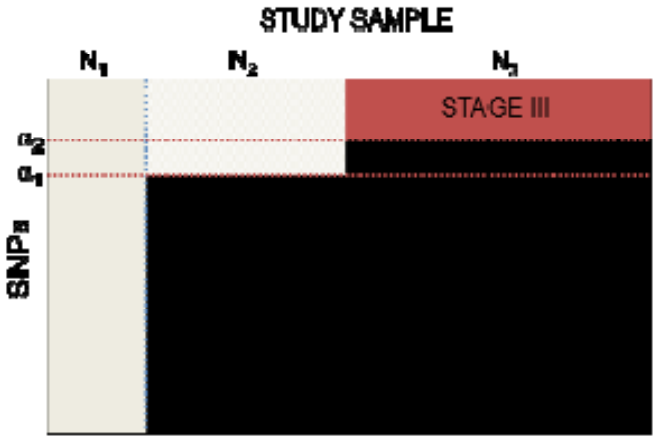
Multi-stage designs

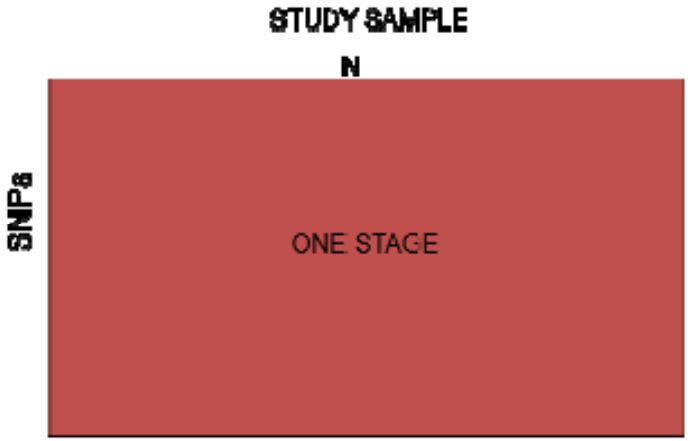
Multi-stage

- Less expensive
- More complicated
- Less powerful

Single-stage

- More expensive
- Less complicated
- More powerful





(slide: courtesy of McQueen)

Strategy 1: entropy-based

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 is defined as:

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

Raw entropy values: interpretation

- The higher the entropy H(Y), the less reliable are our predictions about Y.
- We can understand H(Y) as the amount of uncertainty about Y, as estimated from its probability distribution

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

Information Gain: Slide 10

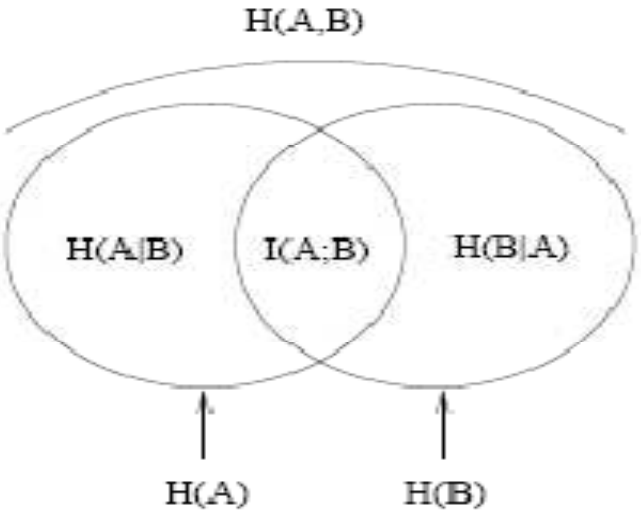
Conditional entropy

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

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

- 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



$$
\begin{aligned}
I(X,Y) = & \quad H(X) - H(X|Y) \\
= & \quad H(Y) - H(Y|X) \\
- & \quad II(X) + II(Y) - II(Y,X)
\end{aligned}
$$

The surface area of a section corresponds to the labeled quantity

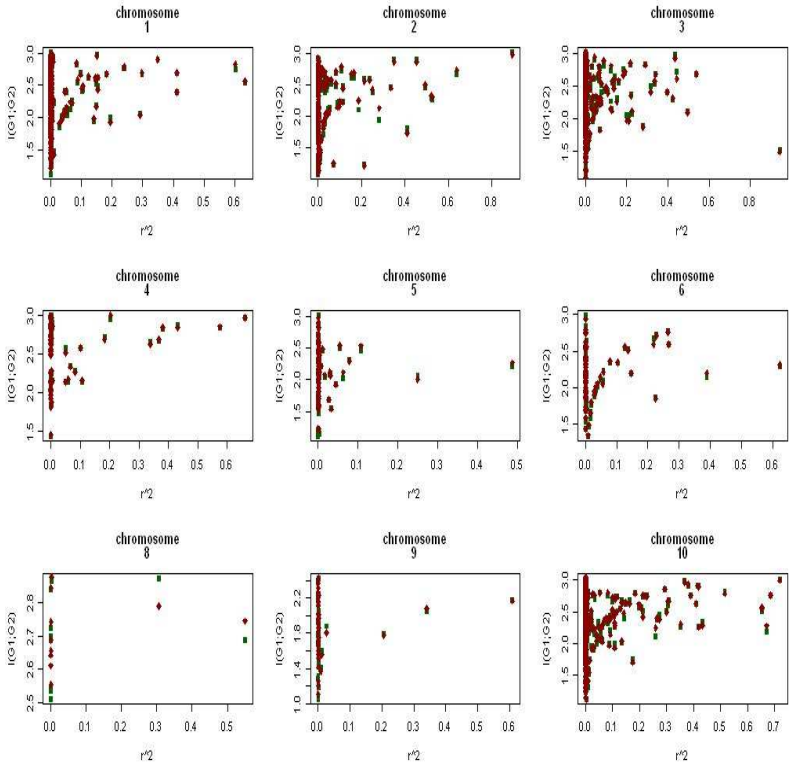
(Jakulin 2003)

H(A) = entropy of A

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

Mutual information

- Mutual information I(A ;B) as a function of r2 (as a measure of LD between markers), for a subset of the Spanish Bladder Cancer data



Mutual information: interpretation

- Mutual information I(A;B) is the reduction of uncertainty of A due to knowledge of B (or vice versa). It is therefore also referred to as the information gain of the given attribute A given B

- Mutual information I(A;B) can also be understood as the expectation of the Kullback-Leibler divergence of the univariate distribution p(a) of A from the conditional distribution p(a|b) of A given B
- In other words, the more different the distributions p(a|b) and p(a), the greater the information gain.

Bivariate synergy

- The bivariate synergy compares the joint contribution with the additive contributions of the individual factors
- It is defined as

$$I(A;B;C) = I(A,B;C) - I(A;C) - I(B;C)$$

- 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. The synergy may also be normalized by dividing it by H(C), in which case it is a quantity between -1 and +1.

(Varadan et al 2006)

Bivariate synergy: interpretation

If I(A;B;C) > 0

Evidence for an attribute interaction that cannot be linearly decomposed

If I(A;B;C) < 0

The information between A and B is redundant

If I(A;B;C) = 0

Evidence of conditional independence or a mixture of synergy and redundancy

- Assume that we are uncertain about the value of C, but we have information about A and B.
  - Knowledge of A alone eliminates I(A;C) bits of uncertainty from C.
  - Knowledge of B alone eliminates I(B;C) bits of uncertainty from C.
  - However, the joint knowledge of A and B eliminates I(A,B;C) bits of uncertainty.

Multivariate synergy

- In general

$$Syn(G_1, \ldots, G_n; C) = I(G_1, \ldots, G_n; C) - \max_{\substack{\text{all partitions} \\ \{S_j\} \text{ of } S}} \sum_j I(S_j; C).$$

- For the special case of 3 contributing variables, the synergy is equal to:

$$\mathrm{Syn}(G_1, G_2, G_3; C) = I(G_1, G_2, G_3; C) - \max \begin{cases} I(G_1; C) + I(G_2; C) + I(G_3; C) \\ I(G_1; C) + I(G_2, G_3; C) \\ I(G_2; C) + I(G_1, G_3; C) \\ I(G_3; C) + I(G_1, G_2; C) \end{cases}$$

(Varadan et al 2006)

Attribute selection based on information gain (IG): 2nd order effects

- Compute I(A;B;C), the synergy of A and B wrt C, or the information gain for attribute (A) or attribute (B) given class (C)
- Entropy-based IG is estimated for each pairwise combination of attributes A and B(i.e. SNP pairs).
- Pairs of attributes are sorted and those with the highest IG, or percentage of entropy in the class removed, are selected for further consideration

(slide: Chen 2007)

Strategy 2: Multivariate filtering

Attribute selection based on reliefF

- The Relief statistic was developed by the computer science community as a powerful method for determining the quality or relevance of an attribute (i.e. variable) for predicting a discrete endpoint or class variable (Kira and Rendell 1992, Konenko 1994, Robnik-Sikonja and Kononenko 2003).
- Relief is especially useful when there is an interaction between two or more attributes and the discrete class variable.
- It is thus superior to univariate filters such as a chi-square test of independence (see later) when interactions are present.

Attribute selection based on reliefF

- In particular, Relief estimates the quality of attributes through a type of nearest neighbor algorithm that selects neighbors (instances) from the same class and from the different class based on the vector of values across attributes.
- Weights (W) or quality estimates for each attribute (A) are estimated based on whether the nearest neighbor (nearest hit, H) of a randomly selected instance (R) from the same class and the nearest neighbor from the other class (nearest miss, M) have the same or different values.
- This process of adjusting weights is repeated for m instances.
- The algorithm produces weights for each attribute ranging from -1 (worst) to +1 (best).

Attribute selection based on reliefF

- ReliefF is able to capture attribute interactions because it selects nearest neighbors using the entire vector of values across all attributes.
- However, this advantage 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 values if the remaining attributes can be re-estimated.

(Moore and White 2008)

Attribute selection based on univariate filtering

- A simple chi-square test of independence is an example of a univariate filter.
  - The manual specifies that this filter should be used to condition your MDR analysis on those attributes that have an independent main effect.
  - However, the MDR software itself does not give you a lot of options to actually perform this conditioning …
- The ReliefF filter will be more useful for capturing those attributes that are likely to be involved in an interaction.

Strategy 3: Data mining

Random Forests (RF)

- Machine-learning technique that builds a forest of classification trees wherein each component tree is grown from a bootstrap sample of the data, and the variable at each tree node is selected from a random subset of all variables in the data (Breiman, 2001). The final classification of an individual is determined by voting over all trees in the forest.
- RF models may uncover interactions among factors that do not exhibit strong marginal effects, without demanding a pre-specified model (McKinney et al., 2006).
- Well-suited to dealing with certain types of genetic heterogeneity, since splits near the root node define separate model subsets in the data.

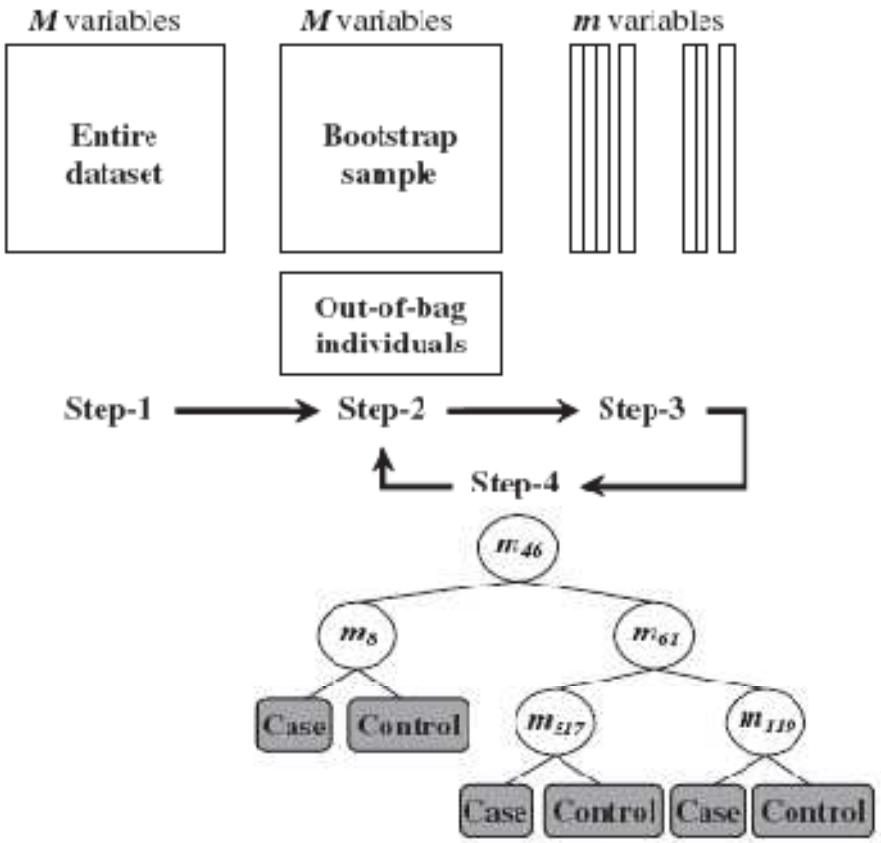
(Motsinger-Reif et al 2008)

Random Forests (RF)

- Each tree in the forest is constructed as follows from data having N individuals and M explanatory variables:
  - Choose a training sample by selecting N individuals, with replacement, from the entire data set.
  - At each node in the tree, randomly select m variables from the entire set of M variables in the data. The absolute magnitude of m is a function of the number of variables in the data set and remains constant throughout the forest building process.
  - Choose the best split at the current node from among the subset of m variables selected above.
  - Iterate the second and third steps until the tree is fully grown (no pruning).

(Motsinger-Reif et al 2008)

A schematic overview of the RF method



(Motsinger-Reif et al 2008)

Advantages of the Random Forest method

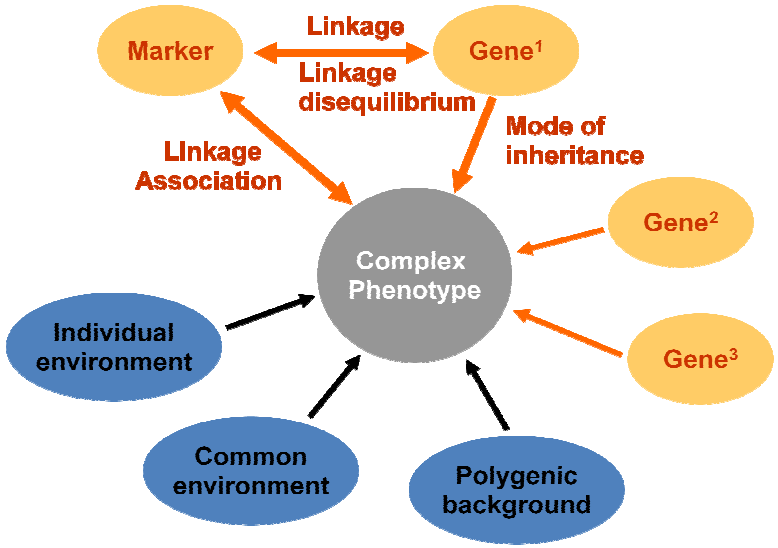
- It can handle a large number of input variables.
- It estimates the relative importance of variables in determining classification, thus providing a metric for feature selection.
- RF produces a highly accurate classifier with an internal unbiased estimate of generalizability during the forest building process.
- RF is fairly robust in the presence of etiological heterogeneity and relatively high amounts of missing data (Lunetta et al., 2004).
- Finally, and of increasing importance as the number of input variables increases, learning is fast and computation time is modest even for very large data sets (Robnik-Sikonja, 2004).

(Motsinger-Reif et al 2008)

Part 10

Epistasis: a curse or a blessing?

The nature of complex disease



(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

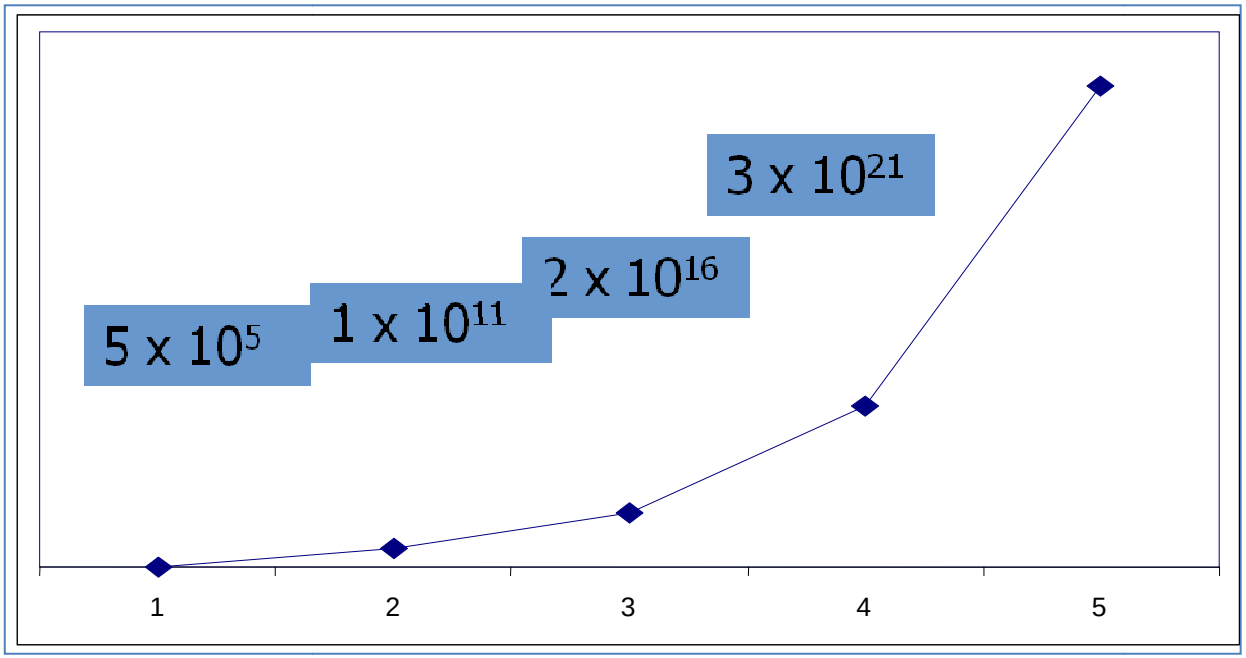
.

(Moore

Beyond main effects

Dealing with multiplicity

- Multiple testing explosion: ~500,000 SNPs span 80% of common variation in genome (HapMap)



n-th order interaction

Ways to handle multiplicity

Recall that several strategies can be adopted, including:

- clever multiple corrective procedures
- pre-screening strategies,
- multi-stage designs,
- haplotype tests
- multi-locus tests or gene-based tests

Which of these approaches are more powerful is still under heavy debate…

Epistasis: What's in a 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)

Epistasis: What's in a name?

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

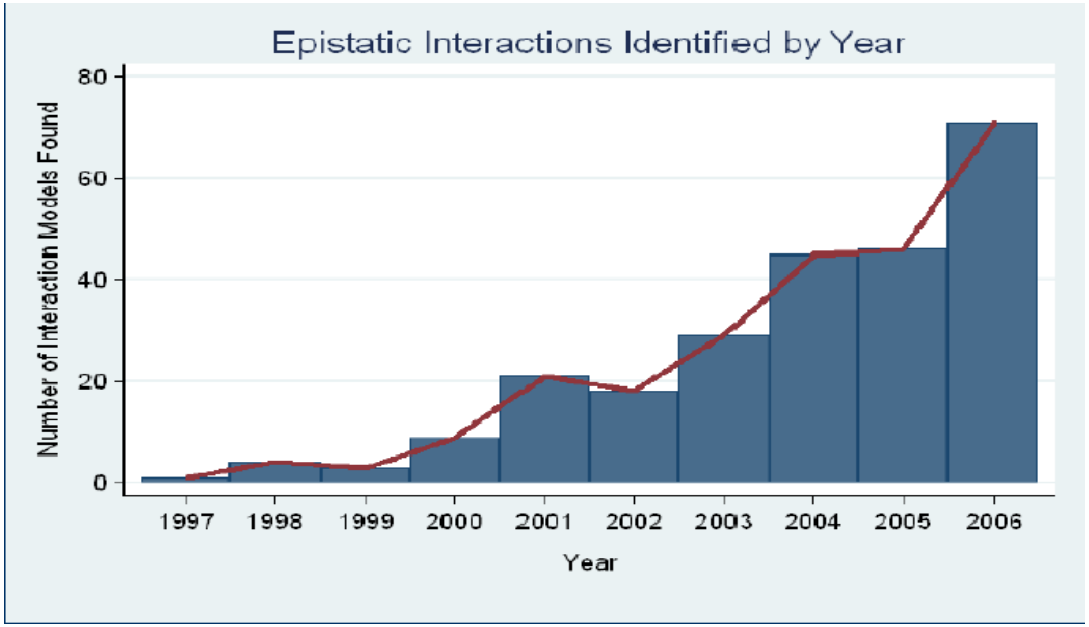
(Cordell 2002)

Why is there epistasis?

- From an evolutionary biology perspective, for a phenotype to be buffered against the effects of mutations, it must have an underlying genetic architecture that is comprised of networks of genes that are redundant and robust.
- Complexity of gene regulation and biochemical networks (Gibson 1996; Templeton 2000)
- This creates dependencies among the genes in the network and is realized as epistasis.
- Single gene results don't replicate (Hirschhorn et al. 2002) and gene-gene interactions are commonly found when properly investigated (Templeton 2000)

Slow shift from main towards epistatis effects

- Working hypothesis:
  Single gene studies don't replicate because gene-gene interactions are more important (Moore and Williams 2002)
  (Moore 2003)



(Motsinger et al 2007)

Power of a gene-gene interaction analysis

- There is a vast literature on power considerations. Most of this literature strengthen their beliefs by extensive simulation studies
- There is a need for user-friendly software tools that allow the user to perform hands-on power calculations
- Main package targeting interaction analyses is QUANTO (v1.2.1):
  - Available study designs for a disease (binary) outcome include the unmatched case-control, matched case-control, case-sibling, case-parent, and case-only designs. Study designs for a quantitative trait include independent individuals and case parent designs.

Gauderman (2000a), Gauderman (2000b), Gauderman (2003)

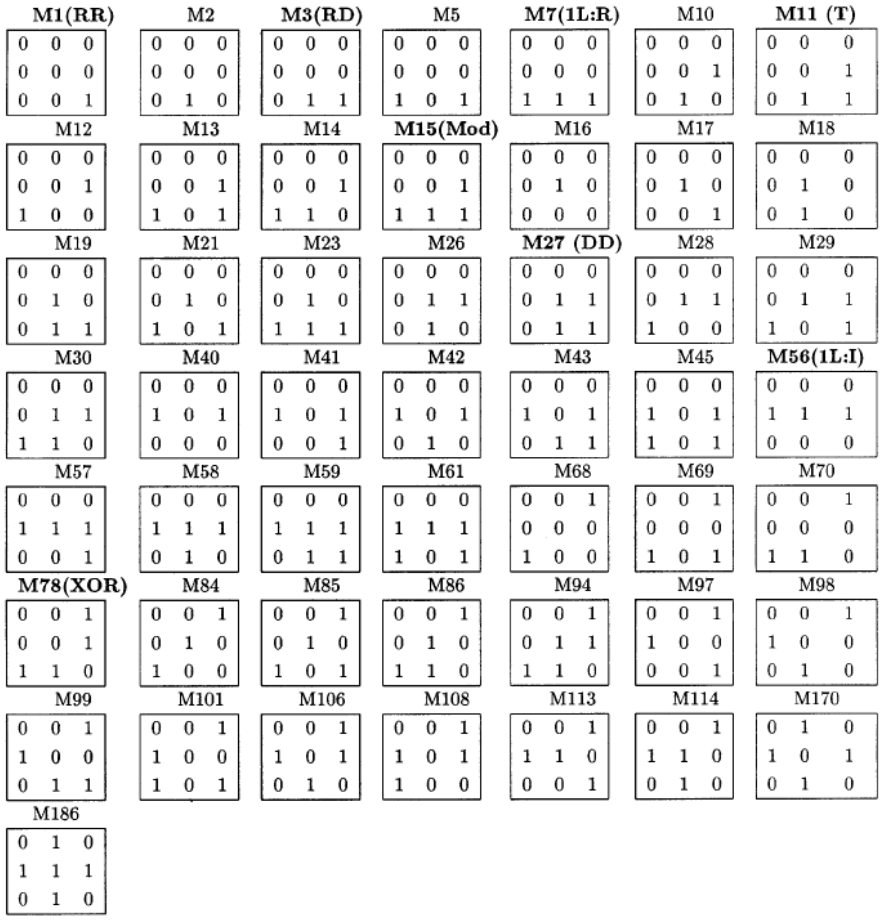
Different degrees of epistasis

An example of a two-locus model

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

- Although there are 2^9=512 possible models, because of symmetries in the data, only 50 of these are unique.
- Enumeration allows 0 and 1 only for penetrance values (‘fully penetrant’; i.e., “show” example).

Enumeration of two-locus models



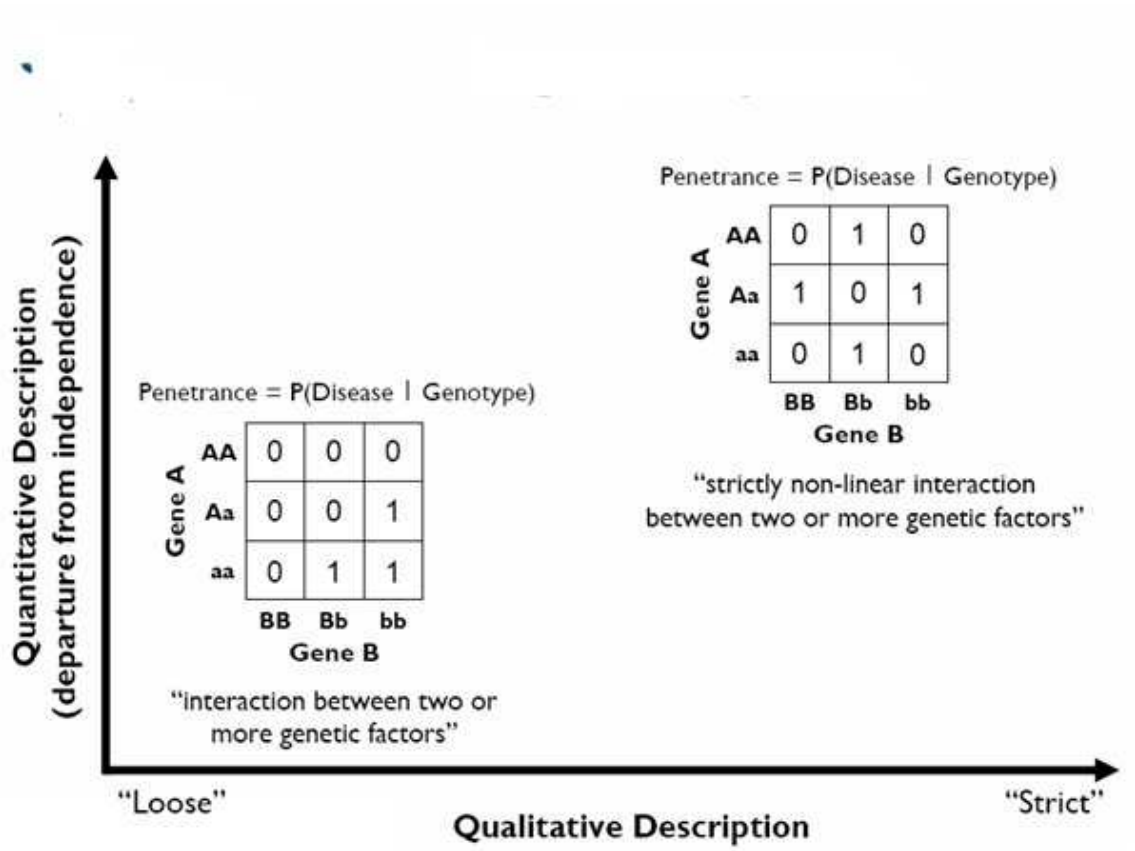
(Li and Reich 2000)

- Each model represents a group of equivalent models under permutations. The representative model is the one with the smallest model number.
- The six models studied in Neuman and Rice [67] ('RR, RD, DD, T, Mod, XOR'), as well as two single-locus models ('IL') – the recessive (R) and the interference (I) model, are marked.

Université de Liège

UNIVERSITÄT ZU LÜBECK

Different degrees of epistasis



(slide: Motsinger)

Pure epistasis

An example

- p(A)=p(B)=p(a)=p(b)=0.5
- HWE (hence, $p(AA)=0.5^2=0.25$,$p(Aa)=2\times0.5^2=0.5$) and no LD
- penetrances are given according to the table below

P(affected|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(affected| aa) = 0.25x0+0.5x0+0.25x1 = **0.25**

Pure epistasis model for dichotomous traits

- …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**

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

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

$= ¼ =$ **0.25**

Université de Liège

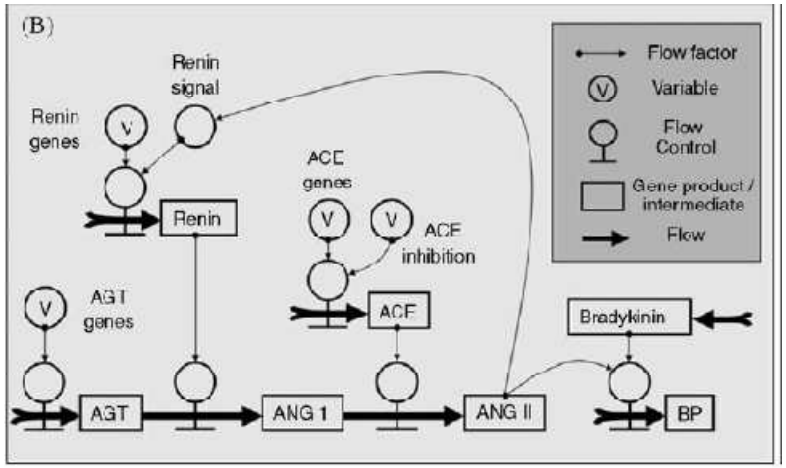
UNIVERSITÄT ZU LÜBECK

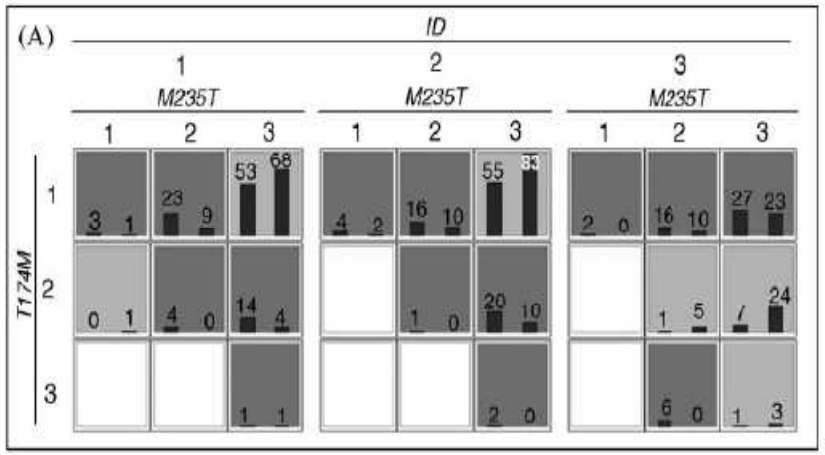
Part 11

Modeling epistasis

Main challenges in epistasis detection

- Variable selection
- Modeling
- Interpretation
  - Making inferences about biological epistasis from statistical epistasis





(slide Chen 2007)

Two frameworks for multi-locus approaches

- Parametric methods:
  - Regression
  - Logistic or (Bagged) logic regression
- Non-parametric methods:
  - Tree-based methods:
    - Random Forests (R, CART, Random Jungle)
  - Pattern recognition methods:
    - Neural networks (NN)
    - Support vector machines (SVM)
  - Data reduction methods:
    - DICE (Detection of Informative Combined Effects)
    - MDR (Multifactor Dimensionality Reduction)

(Onkamo and Toivonen 2006)

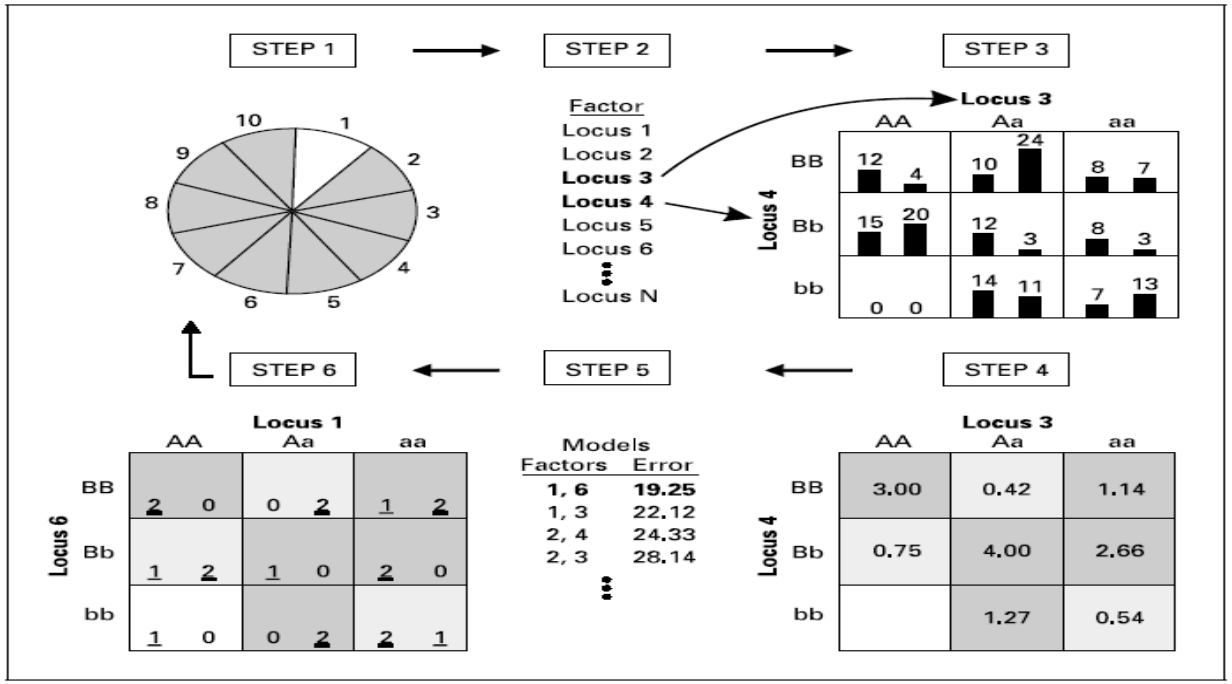
Multifactor Dimensionality Reduction(MDR)

What is MDR?

- A data mining approach to identify interactions among discrete variables that influence a binary outcome
- A nonparametric alternative to traditional statistical methods such as logistic regression
- Driven by the need to improve the power to detect gene-gene interactions

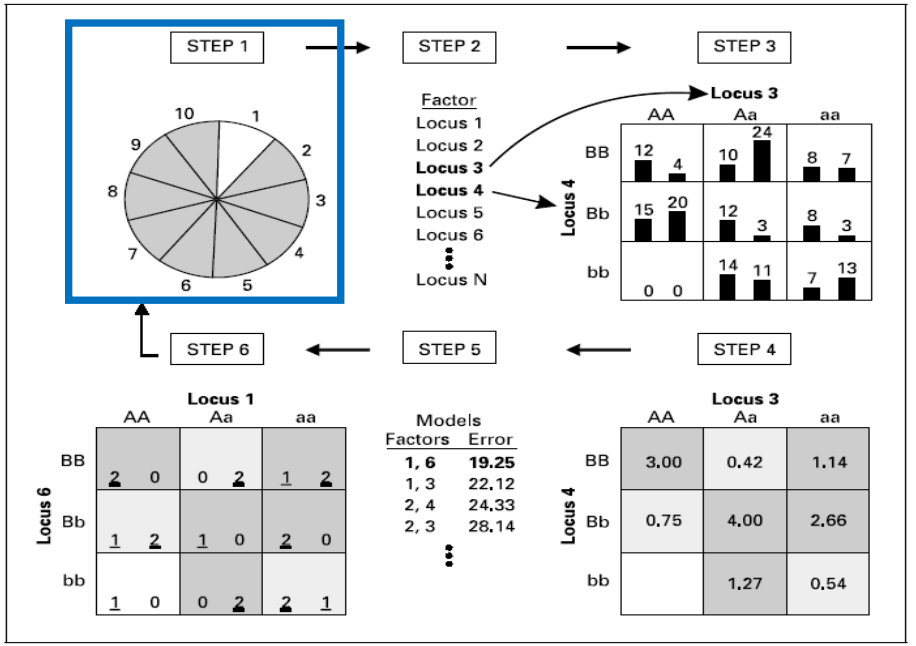
(slide: L Mustavich)

The 6 steps of MDR



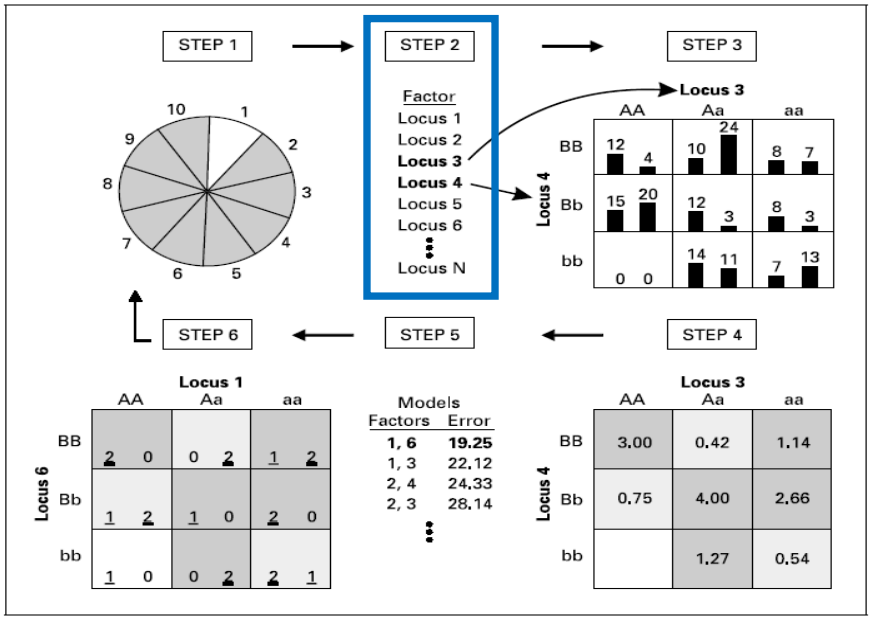
MDR Step 1

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



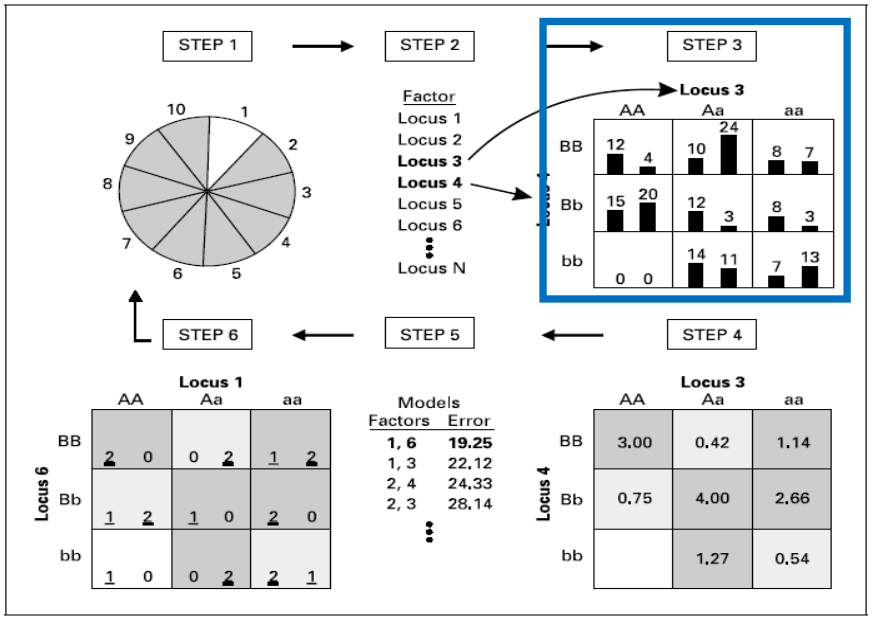
MDR Step 2

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



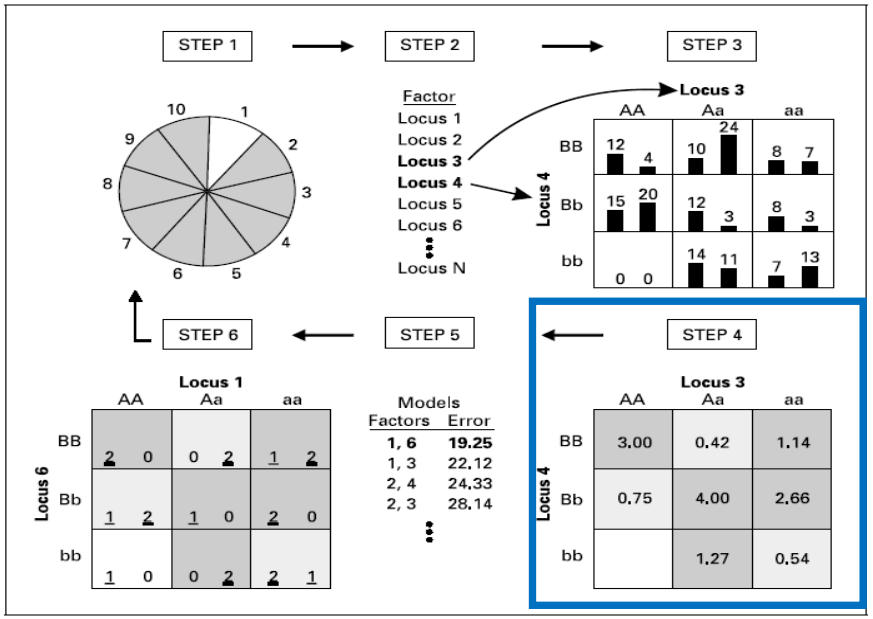
MDR Step 3

- 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

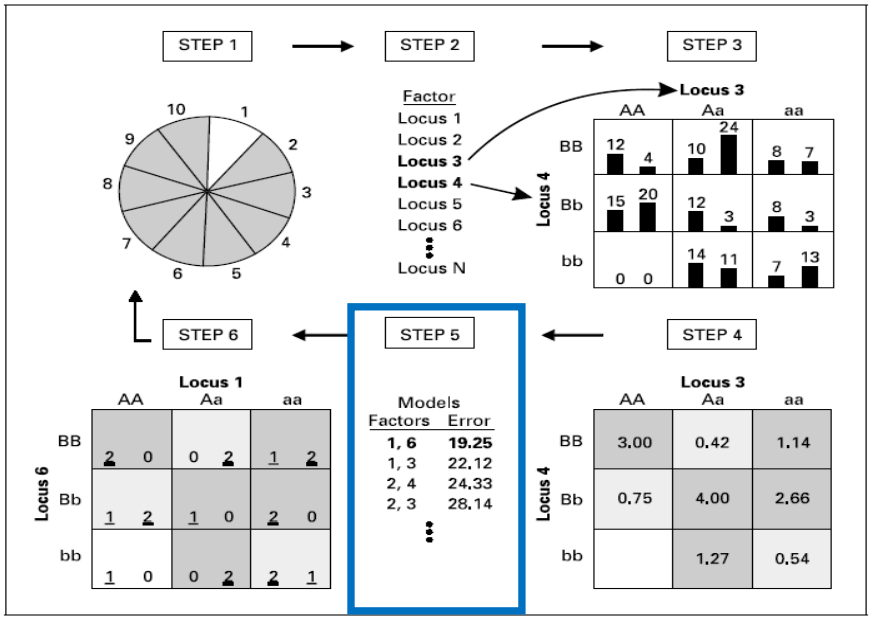
- Calculate the ratio of cases to controls for each multi-locus genotype
- Label each multi-locus genotype as “high-risk” or “low-risk”, depending on whether the case-control ratio is above a certain threshold
- This is the dimensionality reduction step:



Reduces n-dimensional space to 1 dimension with 2 levels

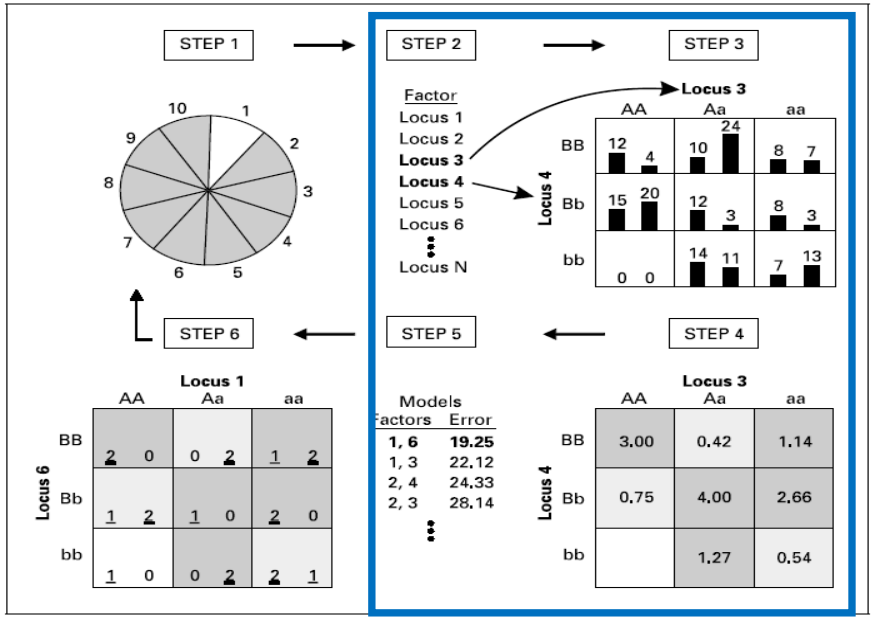
MDR Step 5

- 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 is used (arithmetic mean between sensitivity and specificity), which IS mathematically equivalent to classification accuracy when data are balanced



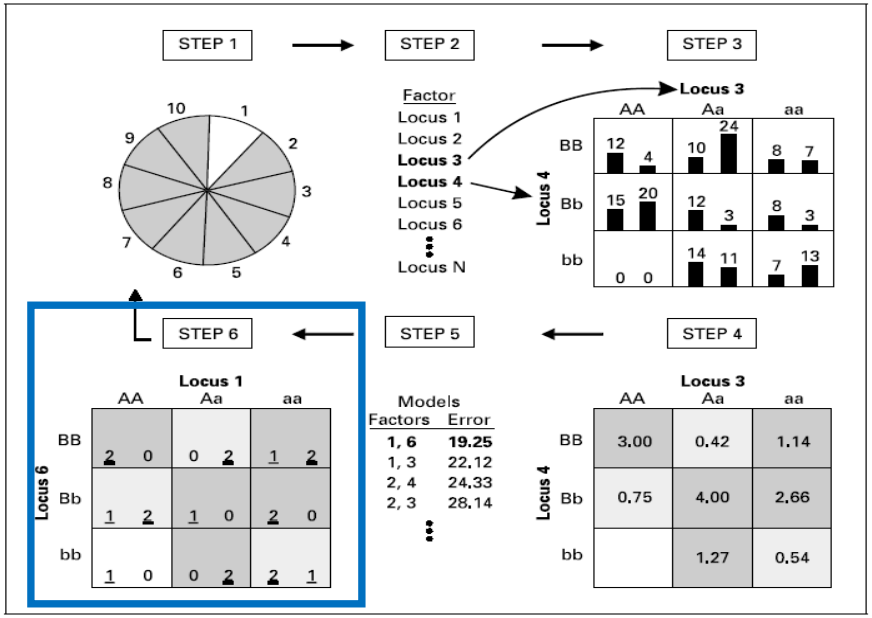
Repeat Steps 2 to 5

- All possible combinations of n factors are evaluated sequentially for their ability to classify affected and unaffected individuals in the training data, and the best n-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 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” of interest
    - Order 2: best model with highest average training accuracy
    - Order 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 measure, which records the number of times MDR finds the same model as the data are divided in different segments
    - Third criteria when testing accuracies for different “best” higher order models are the same
    - Can be biased though!!! → permutation null distribution !!!
  - Using accuracy measures that are not biased by the larger class
  - Using a threshold that is driven by the data at hand and naturally reflects the disproportion in cases and controls in the data

Several measures of fitness to compare models

Balanced accuracy

- Balanced accuracy(BA) weighs the classification accuracy of the two classes equally and it is thought to be more powerful than using accuracy alone when data are imbalanced, or when the counts of cases and controls are not equal (Velez et al 2007)
  - BA is calculated from a 2 × 2 table relating exposure to status by [(sensitivity+specificity)/2].

	Real case	Real control
Model case	TP	FP
Model control	FN	TN

When #cases = #controls, then

TP+FN = FP+TN and

BA = (TP+TN)/(2*#cases)

= TP+TN/(total sample size)

Several measures of fitness to compare models

Model-adjusted balanced accuracy

- Model-adjusted balanced accuracy uses in addition a different threshold in the MDR modeling, one that is based on the actual counts of case and control samples in the data.
  - When individuals have missing data, it accounts for the precise number of individuals with complete data for that particular multi-locus combination
  - This makes MDR robust to class imbalances (Velez et al 2007)

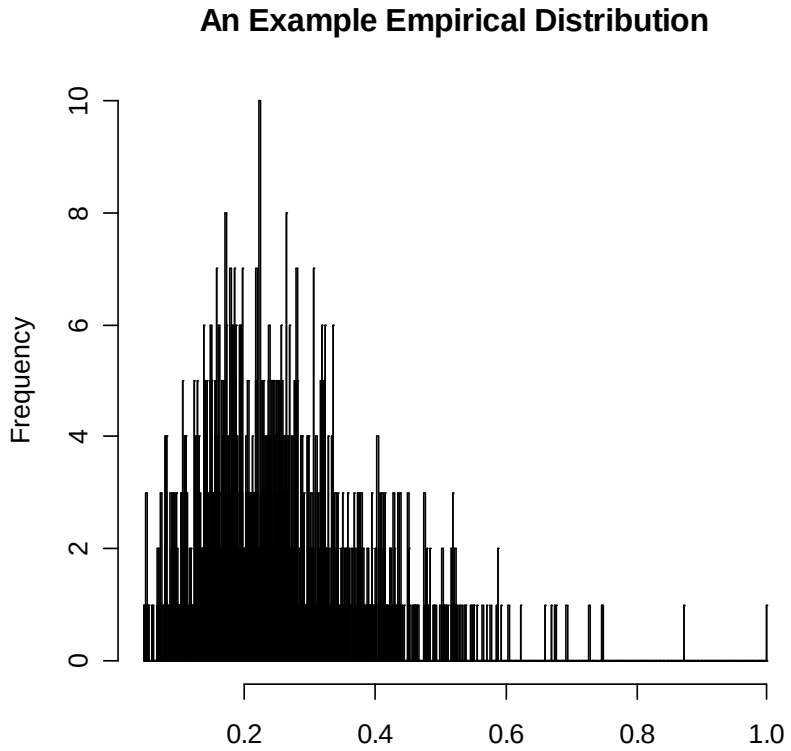
Hypothesis test of best model

- Evaluate magnitude of cross-validation consistency and prediction error estimates by adopting a permutation strategy
- In particular:
  - Randomize disease labels
  - Repeat MDR analysis several times (1000?) to get distribution of cross-validation consistencies and prediction errors
  - Use distributions to derive the p-values for the actual cross-validation consistencies and prediction errors
- Important remark:

  Can give info about whether or not your overall best model is significant, but does NOT provide direct evidence for interaction!!!

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



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)

Minimum system requirements

- 1 GHz Processor
- 256 MB Ram
- 800x600 screen resolution

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

	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

Perm null distr for best k=1-3 models	**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

Perm null distr for best k=1-3 models	**SNP1-SNP2**	**SNP1-SNP2-SNP4**
Testing BA (p-value)	0.8325 (<0.0010)	0.8600 (<0.0010)
CVC (p-value)	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.

Performing the MDR permutation test for M27

- Using permutation null distributions per k-locus setting, the following results are obtained:

Perm null distr for best k-locus model (hence 3 distr)	**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 (0.0620)	5 (0.9110)

- Wouldn't it be natural to correct for SNP1 when looking for interactions?
- What if more than one main effect is present in the data?

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 – no values are estimated
- Assumes no particular genetic model
- Minimal false-positive rates

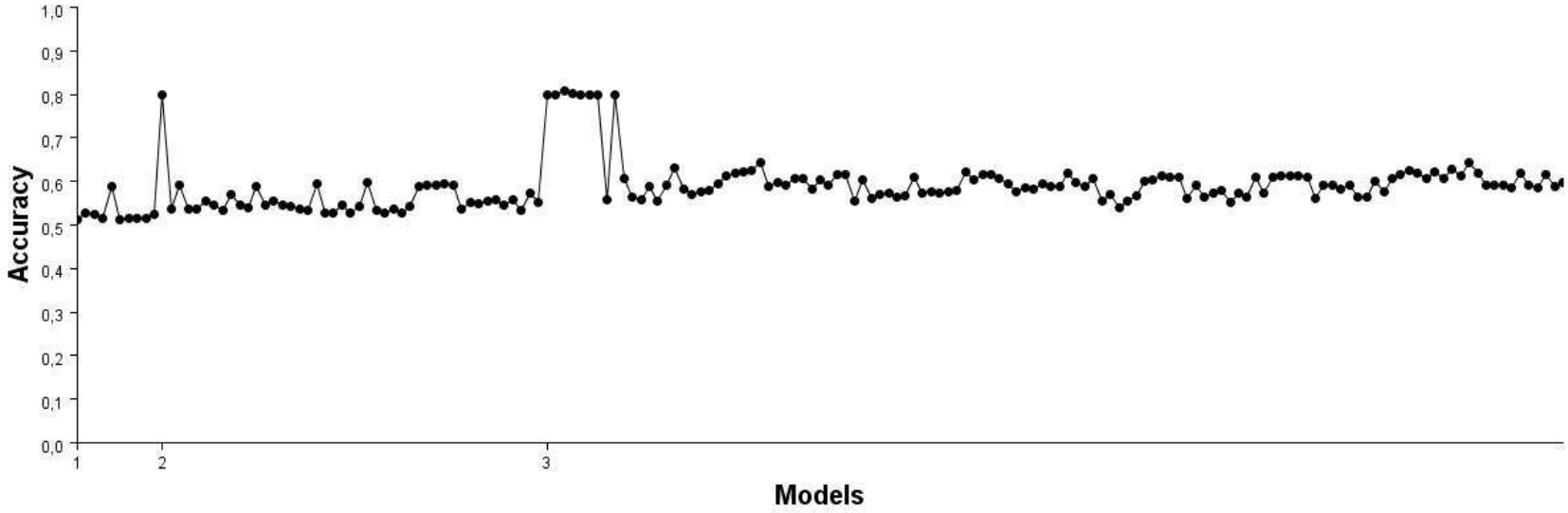
Some weaknesses of MDR

- Computationally intensive (especially with >10 loci)
  - The original MDR software supports disease models with up to 15 factors at a time from a list of up to 1000 total factors and a maximum sample size of about 4,000 – 5,000 subjects.
  - 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 analyses.

- The curse of dimensionality: decreased predictive ability with high dimensionality and small sample due to cells with no data

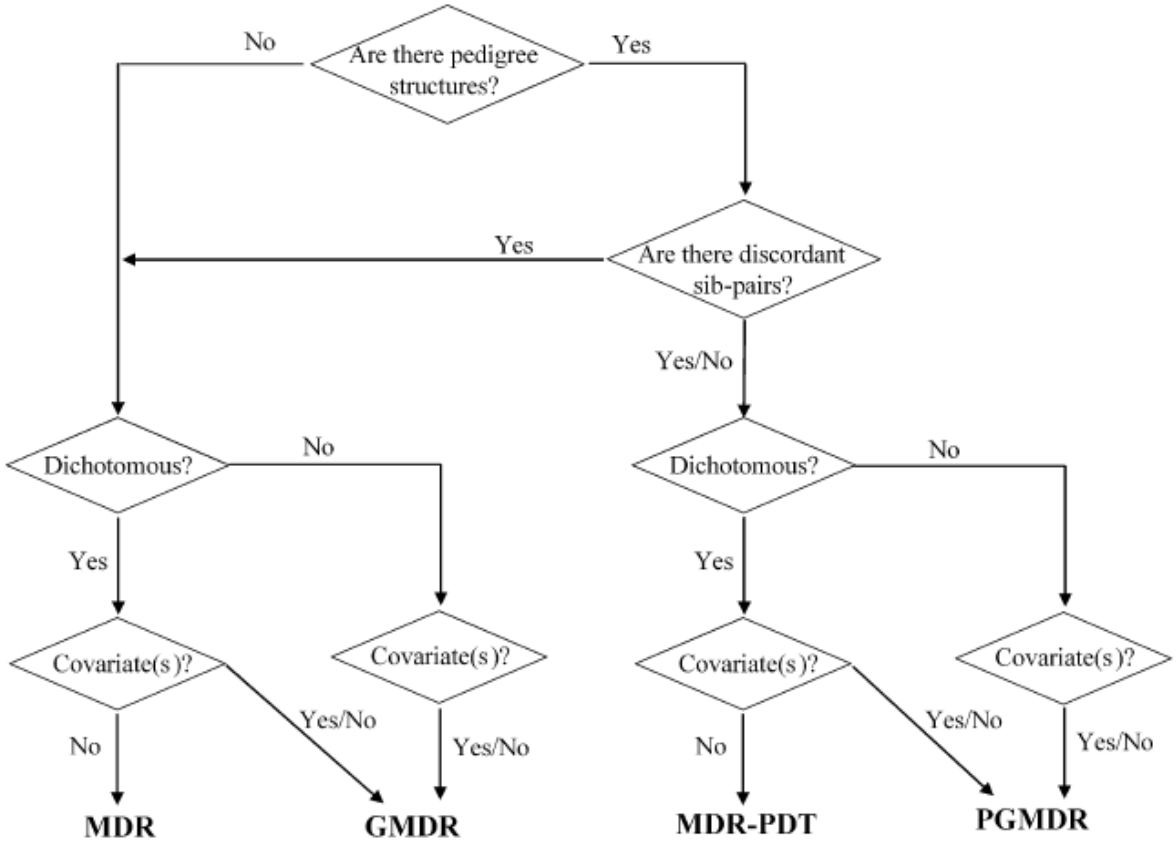
Some weaknesses of MDR

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



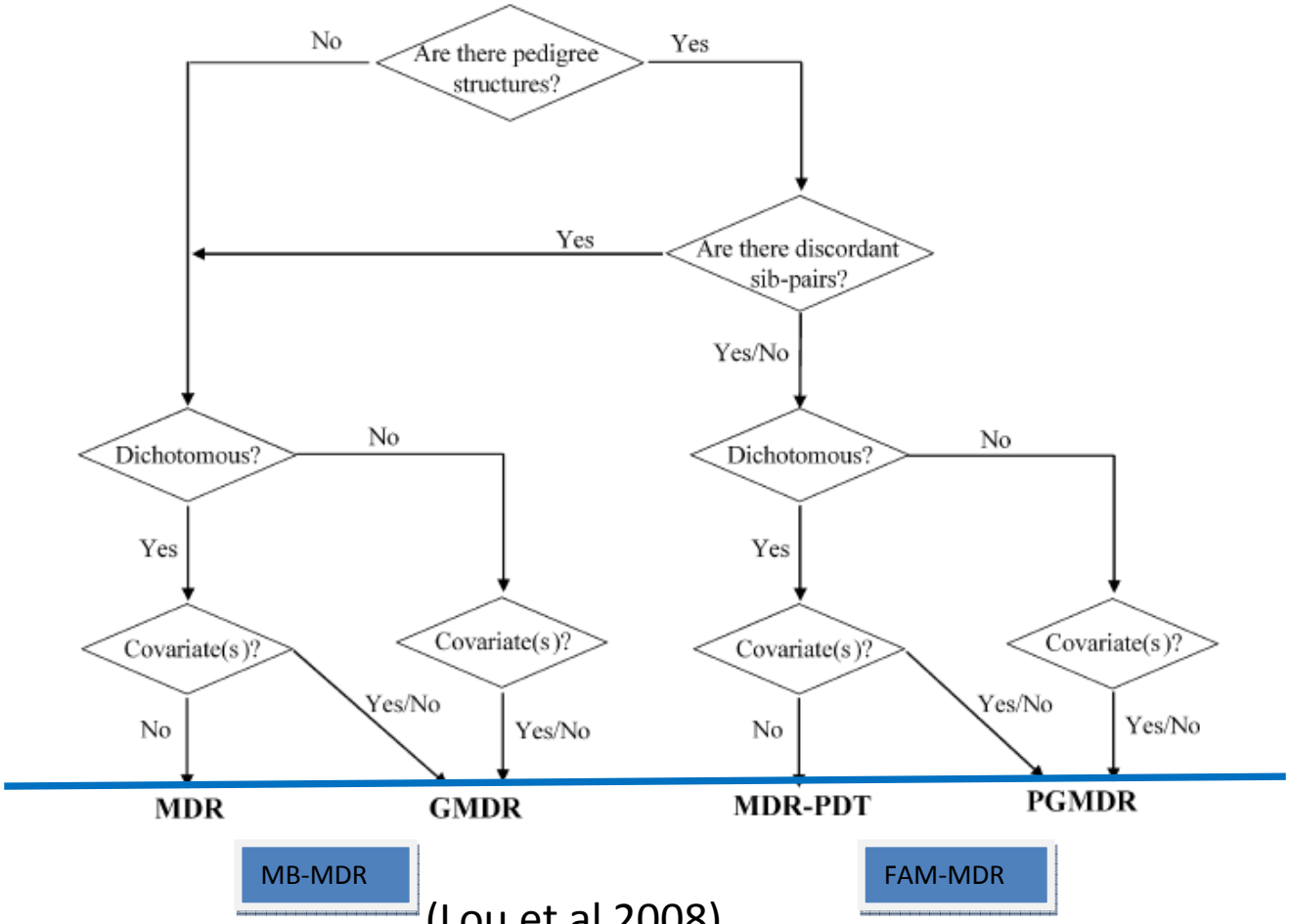
Fitness landscape: The models produced are on the x-axis of the chart. The models on the x-axis are in the order in which they were generated (e.g., 1,2,3, …, 12, 13, 14, …). Training accuracy is shown on the y-axis.

Several (other) extensions to the MDR paradigm (CV based)



(Lou et al 2008)

Towards an easy-to- adapt framework



(Lou et al 2008)

MB-MDR as a semi-parametric approach for unrelateds

- Step 1: New risk cell identification via association test on each genotype cell c_j
  - Parametric or non-parametric test of association
- Step 2: Test one-dimensional "genetic" construct X on Y
- Step 3: assess significance
  - W = [b/se(b)]2, b=ln(OR)
  - Derive correct null distribution for W

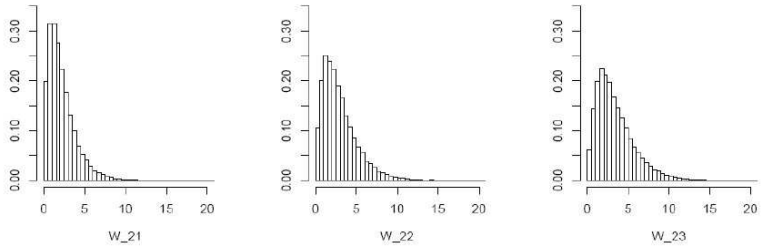


Fig. 2. Permutation null distributions, W_{jk}, for second-order interactions, $j = 2$, conditional on the number of combined cells, $k = 1, 2$ and 3.

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

(Calle et al 2007, Calle et al 2008)

Université de Liège

UNIVERSITÄT ZU LÜBECK

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

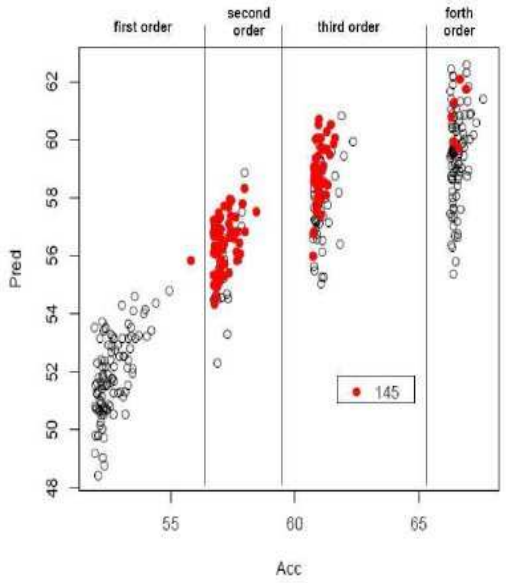


Fig. 1. Average Balanced Training accuracy (Acc) versus Average Balanced Predictive accuracy (Pred) for the 100 models with higher balanced training accuracy for the whole sample. First, second, third and forth order interactions are considered.

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

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

Université de Liège

UNIVERSITÄT ZU LÜBECK

Motivation 3 for MB-MDR

- MDR has low performance in the presence of genetic heterogeneity

		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
No Error	MDR	100	100	99	99	82	84
	MB-MDR	99	100	100	96	95	98
GE	MDR	100	100	100	97	80	92
	MB-MDR	99	100	100	94	96	99
PC	MDR	90	99	45	32	30	32
	MB-MDR	90	100	56	45	43	49
GH	MDR	3	41	2	3	4	4
	MB-MDR One interaction	100	100	84	81	64	77
	MB-MDR Both interactions	93	99	39	27	20	28

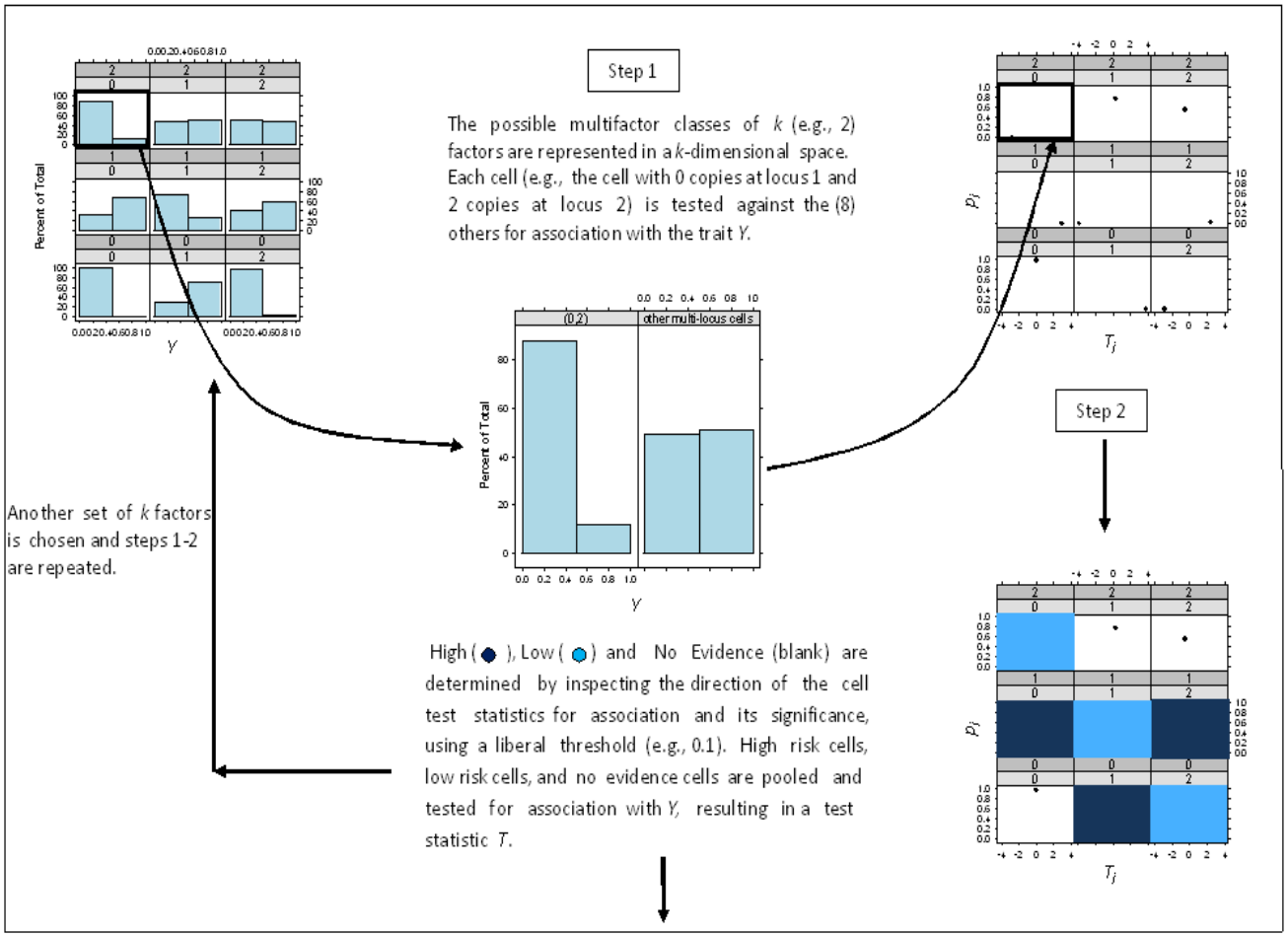
GE: Genotyping error; PC: Phenocopy; GH: Genetic heterogeneity

(Calle et al 2008)

Motivation 4 for MB-MDR

- Both RF and stepwise Logistic Regression models are unable to detect purely epistatic models
- Potential reason:
  - Both methods require marginal main effects to perform variable selection tasks
- Ideally, the variable selection process captures pure interactions (Bureau et al., 2005) – or combined pure and non-pure interactions !!!

MB-MDR



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

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 considered simulated settings.

Characteristics of MB-MDR

- Different disease traits can be accommodated within the same framework offered by MB-MDR
- Confounding factors and lower-order genetic effects can be accounted for as well.
- Allowing a “no evidence” (O) category is particularly relevant for those epistasis models with low MAFs and GH present (see later)
- MDR and MB-MDR inherently assume that the analysis is carried out in a sufficiently homogeneous population

Data format for MB-MDR

- The definition of the 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.

```
//NAME
//    mbmdr -- Model-Based Multifactor Dimentionality Reduction
//
//SYNOPSIS
//    mbmdr --plink2mbmdr -ped 'plinkPedFile' -map 'plinkMapFile' -o 'mbmdrFile' -tr
'traductFile'
//    mbmdr [options] 'mbmdrFile'
//
//DESCRIPTION
//    mbmdr implements Model Based Multifactor Dimension Reduction proposed by Calle et al.
     (2008)
//
//    The first synopsis form converts a ped and a map PLINK file into a file in the internal
      representation:
//      T1  T2  ... Tn  S1  S2  ... Sm
//      X11 X12 ... X1n Y11 Y12 ... Y1m
//      ... ... ... ... ... ... ... ...
//      Xk1 Xk2 ... Xkn Yk1 Yk2 ... Ykm
//      where Ti are the names of the traits and Si the names of the markers
//      special characters can be used but may cause alignement problems in the output file
```

```
//    The program also generates a traduction file of the form:
//    S1 N11 L11
//    S1 N12 L12
//    ...
//    Sm Nm1 Lm1
//
//    where Si are the names of the snp's
//    Nij are the different genotypes of the ith snp (sorted from the most to the less frequent
      ones)
//    Lij is the label chosen for the jth genotype of the ith snp (for exemple CC=0, CT=1 and
      TT=2)
```

```
//OPTIONS
//
//   EXECUTION
//     --sequential          use the sequential version
//     --parallel            use the parallel version
//
...
//   ALGORITHM
//     --maxT                use the max-T step-down permutation algorithm
//     --minP                use the min-P step-down permutation algorithm
//     --rawP                use the raw-P permutation algorithm
//
...
//   TYPE OF DATA
//     --binary              the input file contains only one trait with binary values
//     --continuous          the input file contains only one trait with continuous values
//     --multi-traits -t INT the input file contains multiple continuous traits (use -t to specify
                             the number of  traits)
```

```
//   TEST-STATISTIC COMPUTATION
//    --hlo-mode            uses the HLO method
//
//      STEP I: RISK CELL PRIORITIZATION
//      --one-cell-approach    generates the HLO matrix using prioritization by cell tests
//       [-c DOUBLE]           sets the p-value cut-off used in the cell tests     (default: 0.1)
//      --hlo-ranking          generates the HLO matrix using prioritization by ranking
//       [-h DOUBLE]           sets the target fraction of individuals in H and L  (default: 0.3)
//
//      STEP II: HLO CONSTRUCT ASSOCATION TEST
//      --h-vs-l               analyses the HLO matrix using the H vs L technique
//      --two-tests            analyses the HLO matrix using the TWO TESTS technique
//      --three-tests          analyses the HLO matrix using the THREE TESTS technique
```

```
//   TEST-STATISTIC COMPUTATION
//     --ajust1-mode      HLO method with prioritization by cell tests performing built-in main
                           effects orrections in both steps
//       [-c DOUBLE]       sets the p-value cut-off used in the cell tests      (default: 0.1)
//       --co-dominant    co-dominant main effects correction
//       --additive       additive main effects correction
//       --h-vs-l         analyses the HLO matrix using the H vs L technique
//       --two-tests      analyses the HLO matrix using the TWO TESTS technique
//       --three-tests    analyses the HLO matrix using the THREE TESTS technique

...


//   PARAMETERS
//     -d INT             sets the dimension (order of multi-locus model)
//     -n INT             sets the number of pairs in the result
//     -p INT             sets the number of permutations
```

M170 case control data for MB-MDR

Trait1 SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10

1 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 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

```
SNP        Chi-square pValue
```

MB-MDR (1 dimension) : M170

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

SNP5 13.032 0.006
SNP2 0 1
SNP1 0 1

- Stepwise logistic regression (order 1): Trait1 ~ SNP5
- Stepwise logistic regression (order 2): Trait1 ~ SNP1 + SNP3 + SNP4 + SNP5 + SNP6 + SNP7 + SNP8 + SNP9 + SNP10 + SNP1:SNP3 + SNP1:SNP4 + SNP1:SNP5 + SNP1:SNP6 + SNP4:SNP10 + SNP5:SNP7 + SNP6:SNP8 + SNP7:SNP9

MB-MDR (1 dimension) : M27

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): rait1 ~ 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 optimisitic
- This has been demonstrated by several authors in numerous simulation studies, e.g. Vermeulen et al
- As the targeted order of the interactions increases, the gain of using MDR-like approaches over regression based approaches becomes more apparent

MB-MDR (2 dimensions) :M170

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

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
SNP4	SNP7	10.530	0.279

?

Remarks

- Model M170: pure epistatic effect of SNP1-SNP2
  - The pair SNP 1- SNP2 is highlighted as a significant interaction
  - The significance is independent from the significant main effect SNP5
- Model M27: main effects SNP1 and SNP2 present
  - Too many pairs are marked as significant (increased type I error rate)
  - The significant pairs all involve SNP1 or SNP2
  - There is a need to correct for main effects of the pair under consideration …

Correct up front for SNP5 in M170 – does signal for interaction weaken?

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

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

Correct up front for SNP1 and SNP2 in M27

- Corrected with “genotype” coding

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

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
SNP1	SNP5	99.149	0.002

MB-MDR (2 dimensions), corrected for significant main effects : M170

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

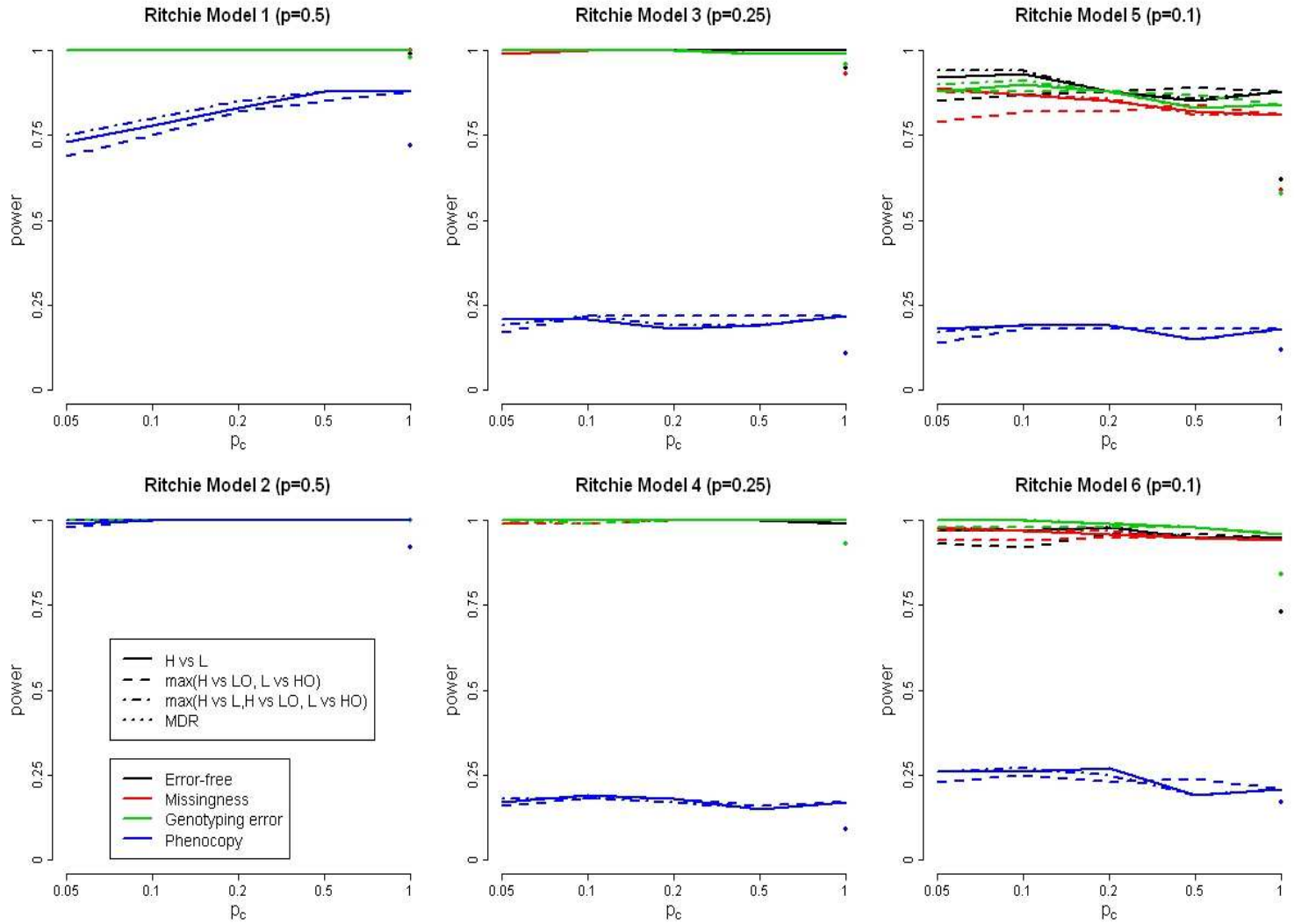
FirstSNP	SecondSNP	Chi-square	pValue
SNP1	SNP2	166.6	0.001
SNP8	SNP10	5.651	0.445
SNP4	SNP10	5.463	0.472
SNP1	SNP10	4.585	0.688

MB-MDR (2 dimensions), corrected for significant main effects : M27

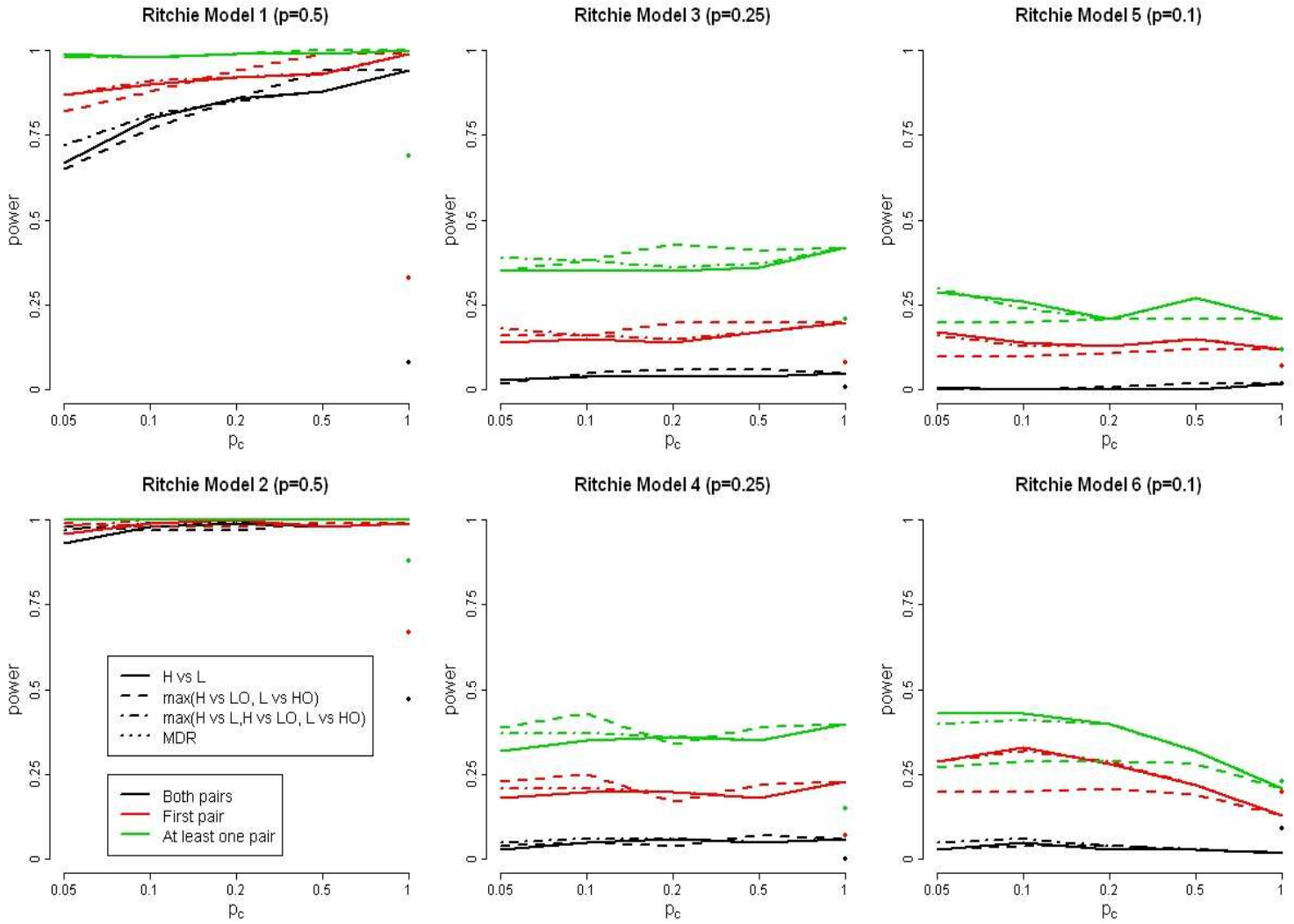
options="--maxT --sequential --binary --ajust1-mode -c 0.1 --co-dominant -- two-tests -d 2 -r 1969 -p 999 -n 20" # see MBMDR.cpp for the different possible options

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

Power in the absence of genetic heterogeneity



Power in the presence of genetic heterogeneity



False positive percentages under alternative hypotheses

Error	Model 1		Model 6	
	MB	MDR	MB	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 pc = 0.1 using the T = |TH/L| test approach and MaxT multiple testing correction and for MDR screening first-to-fifth-order models.

Required operating system software

Linux:

Linux (Fedora version Core 3):

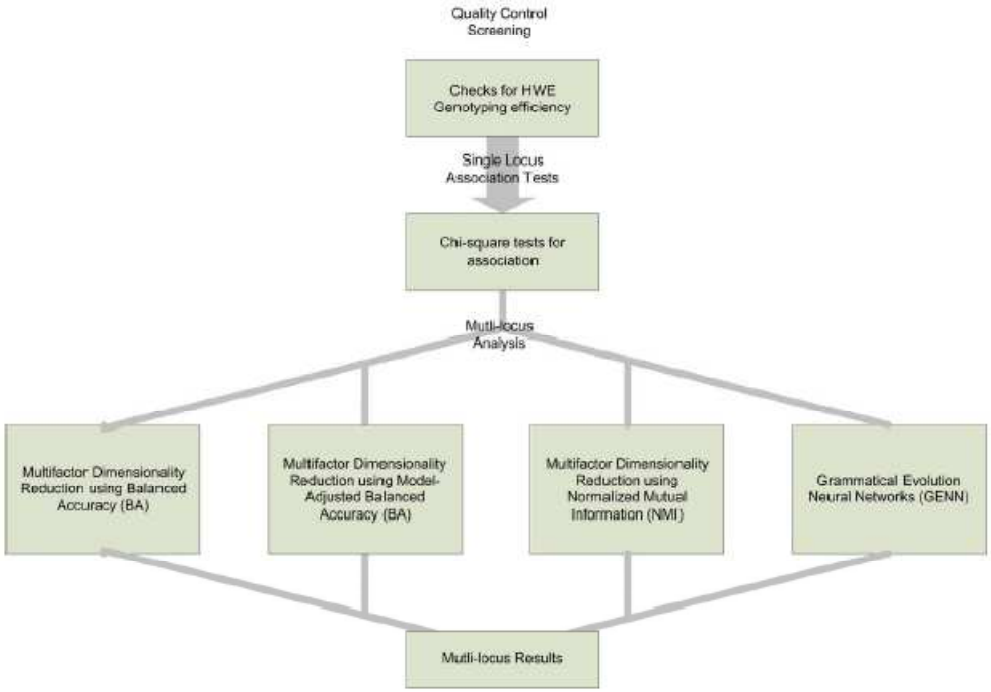
Minimum system requirements

Part 12

Interpretation of identified gene-gene interactions

Interpreting interactions

- It is always a good idea to use several model selection criterions before “interpreting”



(Ritchie et al 2007)

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)

Flexible framework Step 1

- Attribute selection
  - Use entropy-based measures of information gain (IG) and interaction
  - Evaluate the gain in information about a class variable (e.g. case-control status) from merging two attributes together
  - This measure of IG allows us to gauge the benefit of considering two (or more) attributes as one unit

(slide: Chen 2007)

Flexible framework Step 2

- Constructive induction, for instance MDR-like approaches
  - A multi-locus genotype combination is considered high-risk if the ratio of cases to controls exceeds given threshold T, else it is considered low-risk
  - Genotype combinations considered to be **high-risk** are labeled G1 while those considered **low-risk** are labeled G0.
  - This process constructs a new one-dimensional attribute with levels G0 and G1

(adapted from slide: Chen 2007)

Flexible framework Step 3

- Classification and machine learning
  - The single attribute obtained in Step 2 can be modeled using machine learning and classification techniques
- Bayes classifiers as one technique
  - Mitchell (1997) defines the naive Bayes classifier as

$$\arg\max_{v_j \epsilon V} P(v_j) \prod_{i=1}^{n} P(a_i|v_j)$$

  - where v_j is one of a set of V classes and a_i is one of n attributes describing an event or data element. The class associated with a specific attribute list is the one, which maximizes the probability of the class and the probability of each attribute value given the specified class.

Flexible framework Step 3

- The standard way to apply the naive Bayes classifier to genotype data would be to use the genotype information for each individual as a list of attributes to distinguish between the two hypotheses ‘‘The subject is high-risk’’ and ‘‘The subject is low-risk’’.

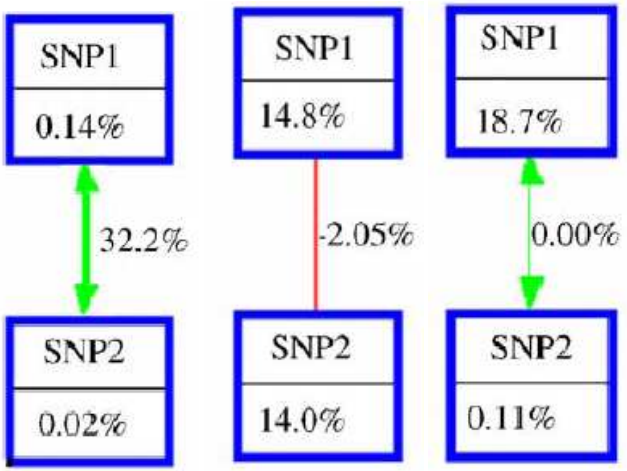
- Alternatively, an odds ratio for the single multilocus attribute can also be estimated using logistic regression to facilitate a traditional epidemiological analysis and interpretation.
  - Evaluation of the predictor can be carried out using cross-validation (Hastie et al., 2001) and permutation testing (Good, 2000), for example.

(Moore et al 2006)

Flexible framework Step 4

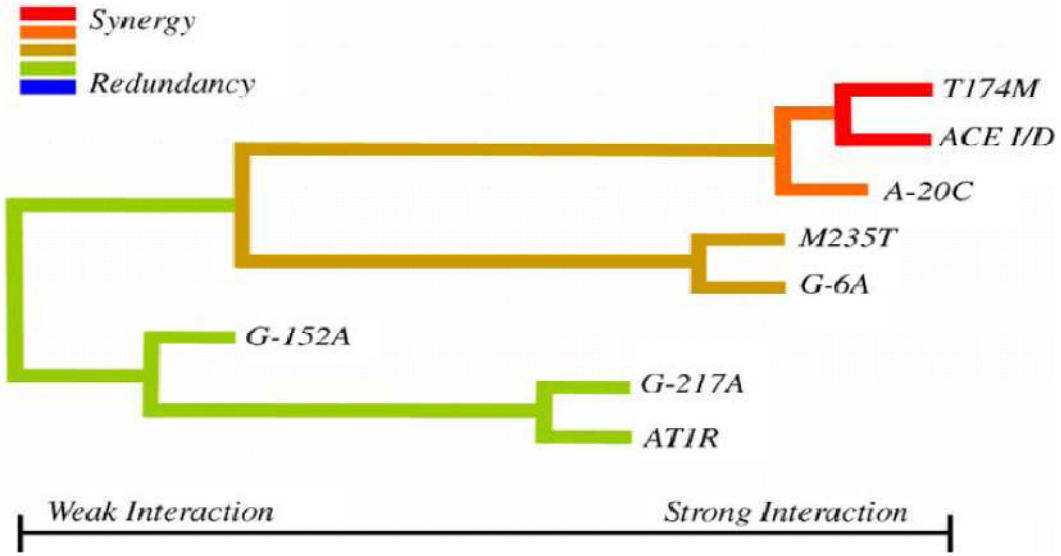
- Interpretation –interaction graphs
  - Comprised of a node for each attribute with pairwise connections between them.
  - Each node is labeled the percentage of entropy removed (i.e. IG) by each attribute.

- Each connection is labeled the percentage of entropy removed for each pairwise Cartesian product of attributes.



Flexible framework Step 4

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



Interaction Dendrogram

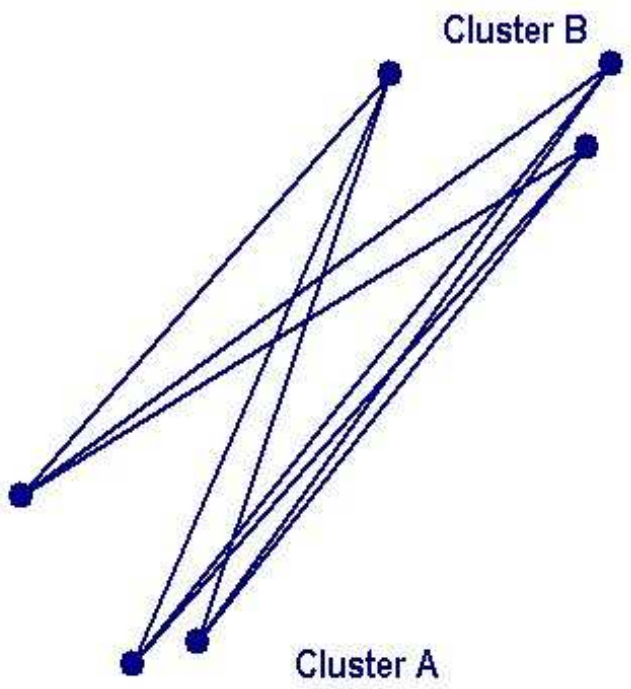
- The interaction dendrogram provides a graphical representation of the interactions between attributes
- The purpose of the interaction dendrogram is to assist the user with determining the nature of the interactions (redundant, additive, or synergistic).

Interaction Dendrogram

- The dendrogram can be constructed using hierarchical cluster analysis with average-linking (distance between two items x and y is the mean of all pairwise distances between items contained in x and y).
- 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)

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

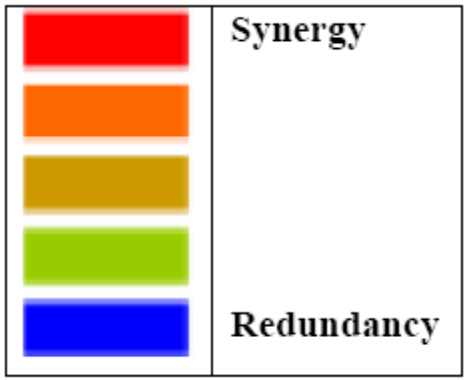


Interaction dendrogram

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

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

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



Flexible framework

- The flexibility of this framework is the ability to plug and play …
  - Different attribute selection methods
    other than the entropy-based
  - Different constructive induction algorithms
    other than the MDR
  - Different machine learning strategies
    other than a naïve Bayes classifier

(slide: Chen 2007)


off the mark
by Mark Parisi
www.offthemark.com
UH, YES... I WAS ABLE TO TRACE YOUR ANCESTRY... BUT REMEMBER... I'M ONLY THE MESSENGER HERE...
BLUE MAN GROUP