INTRODUCTION TO GENETIC EPIDEMIOLOGY

(1012GENEP1)

Prof. Dr. Dr. K. Van Steen
GENOME-WIDE ASSOCIATION INTERACTION (GWAI) STUDIES: Mission impossible?

Outline

• The origin of “interactions”
• Travelling the world of interactions
• How to best build our working space
• Components of epistasis analysis
• Model-Based Multifactor Dimensionality Reduction
• GWAI in practice
The origin of interactions
The complexity of complex diseases

There are likely to be many susceptibility genes each with combinations of rare and common alleles and genotypes that impact disease susceptibility primarily through non-linear interactions with genetic and environmental factors

(Weiss and Terwilliger 2000)
### Factors complicating analysis of complex genetic disease

<table>
<thead>
<tr>
<th></th>
<th>Locus Heterogeneity</th>
<th>Trait Heterogeneity</th>
<th>Gene-Gene Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>when two or more DNA variations in distinct genetic loci are independently associated with the same trait</td>
<td>when a trait, or disease, has been defined with insufficient specificity such that it is actually two or more distinct underlying traits</td>
<td>when two or more DNA variations interact either directly (DNA-DNA or DNA-mRNA interactions), to change transcription or translation levels, or indirectly by way of their protein products, to alter disease risk separate from their independent effects</td>
</tr>
<tr>
<td><strong>Diagram</strong></td>
<td><img src="image1.png" alt="Diagram of Locus Heterogeneity" /></td>
<td><img src="image2.png" alt="Diagram of Trait Heterogeneity" /></td>
<td><img src="image3.png" alt="Diagram of Gene-Gene Interaction" /></td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td><strong>Retinitis Pigmentosa</strong> (RP, OMIM# 268000) - genetic variations in at least fifteen genes have been associated with RP under an autosomal recessive model. Still more have been associated with RP under autosomal dominant and X-linked disease models&lt;sup&gt;2&lt;/sup&gt; (<a href="http://www.sph.uth.tmc.edu/RetNet">http://www.sph.uth.tmc.edu/RetNet</a>)</td>
<td><strong>Autosomal Dominant Cerebellar Ataxia</strong> (ADCA, OMIM# 164500) - originally described as a single disease, three different clinical subtypes have been defined based on variable associated symptoms,&lt;sup&gt;6&lt;/sup&gt;&lt;sup&gt;,7&lt;/sup&gt; and different genetic loci have been associated with the different subtypes&lt;sup&gt;8&lt;/sup&gt;</td>
<td><strong>Hirschsprung Disease</strong> (OMIM# 142623) - variants in the RET (OMIM# 164761) and EDNRB (OMIM# 131244) genes have been shown to interact synergistically such that they increase disease risk far beyond the combined risk of the independent variants&lt;sup&gt;12&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

(Thornton-Wells et al. 2006)
Factors complicating analysis of complex genetic disease

Gene-gene interactions

... when two or more DNA variations interact either directly to change transcription or translation levels, or indirectly by way of their protein product, to alter disease risk separate from their independent effects ...

(Moore 2005)
The “observed” occurrences of epistasis – model organisms

- Carlborg and Haley (2004):
  - Epistatic QTLs without individual effects have been found in various organisms, such as birds\textsuperscript{26,27}, mammals\textsuperscript{28–32}, Drosophila melanogaster\textsuperscript{33} and plants\textsuperscript{18,34}.
  - However, other similar studies have reported only low levels of epistasis or no epistasis at all, despite being thorough and involving large sample sizes\textsuperscript{35–37}.

This clearly indicates the complexity with which multifactorial traits are regulated; no single mode of inheritance can be expected to be the rule in all populations and traits.
Great expectations

• From an evolutionary biology perspective, for a phenotype to be buffered against the effects of mutations, it must have an underlying genetic architecture that is comprised of networks of genes that are redundant and robust.

• The existence of these networks creates dependencies among the genes in the network and is realized as gene-gene interactions or (trans-) epistasis.

• This suggests that epistasis is not only important in determining variation in natural and human populations, but should also be more widespread than initially thought (rather than being a limited phenomenon).
Great expectations - empowering personal genomics

- Considering the epic complexity of the transcriptions process, the genetics of gene expression seems just as likely to harbor epistasis as biological pathways.

- When examining HapMap genotypes and gene expression levels from corresponding cell lines to look for cis-epistasis, over 75 genes pop up where SNP pairs in the gene's regulatory region can interact to influence the gene's expression.

- What is perhaps most interesting is that there are often large distances between the two interacting SNPs (with minimal LD between them), meaning that most haplotype and sliding window approaches would miss these effects.  
  (Turner and Bush 2011)
Complementing insights from GWA studies

Edges represent small gene–gene interactions between SNPs. Gray nodes and edges have weaker interactions. Circle nodes represent SNPs that do not have a significant main effect. The diamond nodes represent significant main effect association. The size of the node is proportional to the number of connections.

(McKinney et al 2012)
Epistasis and phantom heritability

(Maher 2008)
Epistasis and phantom heritability

- Human genetics has been haunted by the mystery of “missing heritability” of common traits.
- Although studies have discovered >1,200 variants associated with common diseases and traits, these variants typically appear to explain only a minority of the heritability.
- The proportion of heritability explained by a set of variants is the ratio of (i) the heritability due to these variants (numerator), estimated directly from their observed effects, to (ii) the total heritability (denominator), inferred indirectly from population data.
- The prevailing view has been that the explanation for missing heritability lies in the numerator – variants still to identify.
Epistasis and phantom heritability

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

- For example, 80% of the currently missing heritability for Crohn's disease could be due to genetic interactions, if the disease involves interaction among three pathways.  
  (Zuk et al 2012)
Traveling the world of interactions
• Most SNPs of interest will only be found by embracing the complexity of the genotype-to-phenotype mapping relationship that is likely to be characterized by nonlinear gene-gene interactions, gene-environment interaction and locus heterogeneity.

(Moore and Williams 2009)

• Few SNPs with moderate to large independent and additive main effects
From GWA to GWAI studies ...

- Genome-Wide Association Interaction (GWAI) studies have not been as successful as GWA studies:
  - Possible negligible role of epistatic variance in a population? (Davierwala et al 2005)
  - Consequence of not yet available powerful epistasis detection methods or approaches?
    “Gene-gene interactions are commonly found when properly investigated” (Templeton 2000)
How to best build our working space
Creating an atmosphere of “interdisciplinarity”

Creating an atmosphere of “integration”

with HTP omics data

(J Thornton, EBI)
Extending the toolbox

(Kilpatrick 2009)
Extending the toolbox

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

Extending the toolbox

• Comes with a caveat: need for thorough comparison studies using reference data sets!

• Several criteria exist to classify epistasis detection methods:
  - Exploratory versus non-exploratory
  - Testing versus Modeling
  - Direct versus Indirect testing
  - Parametric versus non-parametric
  - Exhaustive versus non-exhaustive search algorithms
  - … (Van Steen et al 2011)
The “observed” occurrences of epistasis – humans

  - There are several cases of epistasis appearing as a statistical feature of association studies of human disease.
  - A few recent examples include coronary artery disease\textsuperscript{63}, diabetes\textsuperscript{64}, bipolar effective disorder\textsuperscript{65}, and autism\textsuperscript{66}.
  - So far, only for some of the reported findings additional support could be provided by functional analysis, as was the case for multiple sclerosis (Gregersen et al 2006).
The “observed” occurrences of epistasis – humans

• More recent examples include:
  - Alzheimer’s disease (Combarros et al 2009),
  - psoriasis (WTCCC2 2010),
  - breast cancer (Ashworth et al. 2011),
  - ankylosing spondylitis (WTCCC 2011),
  - total IgE (Choi et al. 2012)
  - High-Density Lipoprotein Cholesterol Levels (Ma et al. 2012)

• So far, only for some of the reported findings additional support could be provided by functional analysis or could be “replicated” (see also later)
Taking it a few steps back ... What’s in a name?

• Wikipedia (23/04/2012)

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

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

... Studying genetic interactions can reveal gene function, the nature of the mutations, functional redundancy, and protein interactions. Because protein complexes are responsible for most biological functions, genetic interactions are a powerful tool ...
Taking it a few steps back ... What’s in a name?

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

- The original definition (driven by biology) refers to distortions of Mendelian segregation ratios due to one gene masking the effects of another; a variant or allele at one locus prevents the variant at another locus from manifesting its effect (William Bateson 1861-1926).

(Carlberg and Haley 2004)
Compositional epistasis

- Example of phenotypes (e.g. hair colour) from different genotypes at 2 loci interacting epistatically under Bateson's (1909) definition:

<table>
<thead>
<tr>
<th>Genotype at locus B/G</th>
<th>gg</th>
<th>gG</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td>bb</td>
<td>White</td>
<td>Grey</td>
<td>Grey</td>
</tr>
<tr>
<td>bB</td>
<td>Black</td>
<td>Grey</td>
<td>Grey</td>
</tr>
<tr>
<td>BB</td>
<td>Black</td>
<td>Grey</td>
<td>Grey</td>
</tr>
</tbody>
</table>

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

(Cordell 2002)
Taking it a few steps back ... What’s in a name?

“statistical epistasis”

- A later definition of epistasis (driven by statistics) is expressed in terms of deviations from a model of additive multiple effects.
- This might be on either a linear or logarithmic scale, which implies different definitions (Ronald Fisher 1890-1962).

- It seems that the interpretation of GWAIIs is hampered by undetected false positives
Components of an Epistasis Analysis
Any epistasis analysis is characterized by at least 2 of the following components

- Variable selection
- Modeling / testing
- Significance assessment
- Interpretation
Variable Selection
Why selecting variables?

Introduction

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

Multi-stage

- Less expensive
- More complicated
- Less powerful

Single-stage

- More expensive
- Less complicated
- More powerful

(slide: courtesy of McQueen)
Variable selection in interaction effects GWAS

- Several strategies can be adopted to select the number of genetic variants to be used for epistasis screening.
- Strategy I involves performing an exhaustive search
  
  Address several computational issues and confront a severe multiple testing problem.
- Strategy II involves selecting genetic markers based on the statistical significance or strength of their singular main effects (Kooperberg et al 2008).

  Address the difficulty in finding gene-gene interactions when the underlying disease model is purely epistatic.
Variable selection in interaction effects GWAS

- Strategy III involves prioritizing sets of genetic markers based on feature selection methods.
  
  🎨 Address finding your way into the jungle of different possible feature selection methods and algorithms

- Strategy IV involves prioritizing sets of genetic markers based on (prior) expert knowledge
  
  🎨 Address biasing of findings towards “what is already known”.
Feature selection methods

• In contrast to other dimensionality reduction techniques like those based on projection (e.g., principal components analysis), feature selection techniques do not change the original presentation of the variables

• Hence, feature selection does not only reduce the burden of multiple testing, but also aids in the interpretation of analysis results
Feature selection methods

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

(Saeys et al 2007)
# Feature selection methods

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

(Saeys et al 2007)
Feature selection methods

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

(Saeys et al 2007)
Feature selection methods

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

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

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

Copyright © 2001, 2003, Andrew W. Moore

High Entropy

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

Information Gain: Slide 10
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

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

the additional contribution provided by the “whole” compared with the sum of the contributions of the “parts”. \(^{(\text{Varadan et al 2006})}\)

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

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

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. \(^{(\text{Anastassiou 2007})}\)
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
The challenge of detecting epistasis (G x G interactions); Genetic Analysis Workshop 16.
PMD: 19924703 [PubMed - indexed for MEDLINE]
Related citations

Information-theoretic gene-gene and gene-environment interaction analysis of quantitative traits.
Chanda P, Sucheston L, Liu S, Zhang A, Ramanathan M.
Related citations

The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors.
Chanda P, Sucheston L, Zhang A, Ramanathan M.
Related citations

AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes.
Chanda P, Sucheston L, Zhang A, Brazeau D, Freudenheim JL, Ambroson C, Ramanathan M.
Related citations

Information-theoretic metrics for visualizing gene-environment interactions.
Chanda P, Zhang A, Brazeau D, Sucheston L, Freudenheim JL, Ambrosone C, Ramanathan M.
Related citations
Strategy 2: Data mining as embedding technique

Random Forests (RF)  

(Breiman 2001)

(Motsinger-Reif et al 2008)
Introduction to Genetic Epidemiology

Genetic Association Interaction Studies

Modeling / Testing
What do we want to model/test?

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

<table>
<thead>
<tr>
<th>Genotype</th>
<th>bb</th>
<th>bB</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aA</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AA</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

(Cordell 2002)

- Enumeration of two-locus models:
  - Although there are $2^9 = 512$ possible models, because of symmetries in the data, only 50 of these are unique.
Enumeration of two-locus models
(Li and Reich 2000)

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

Penetration = $P(\text{Disease} \mid \text{Genotype})$

```
<table>
<thead>
<tr>
<th>Gene A</th>
<th>Gene B</th>
<th>Penetration</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>BB</td>
<td>0</td>
</tr>
<tr>
<td>AA</td>
<td>Bb</td>
<td>1</td>
</tr>
<tr>
<td>AA</td>
<td>bb</td>
<td>0</td>
</tr>
<tr>
<td>Aa</td>
<td>BB</td>
<td>0</td>
</tr>
<tr>
<td>Aa</td>
<td>Bb</td>
<td>1</td>
</tr>
<tr>
<td>Aa</td>
<td>bb</td>
<td>0</td>
</tr>
<tr>
<td>aa</td>
<td>BB</td>
<td>0</td>
</tr>
<tr>
<td>aa</td>
<td>Bb</td>
<td>1</td>
</tr>
<tr>
<td>aa</td>
<td>bb</td>
<td>0</td>
</tr>
</tbody>
</table>
```

“Strictly non-linear interaction between two or more genetic factors”

“Interaction between two or more genetic factors”

“Loose” vs. “Strict”

(degree of independence)

(slide: Motsinger)
Incomplete penetrances

- Odds of disease for 2 loci under epistatic scenarios

\[\begin{array}{ccc}
\text{aa} & x & x(1+\theta_2) \\
\text{Aa} & x(1+\theta_1) & x(1+\theta_1)(1+\theta_2) \\
\text{AA} & x(1+\theta_1)^2 & x(1+\theta_1)^2(1+\theta_2) \\
\text{bb} & x & x(1-\theta_2)^2 \\
\text{Bb} & x(1+\theta_1) & x(1+\theta_1)(1+\theta_2)^2 \\
\text{BB} & x(1+\theta_1)^2 & x(1+\theta_1)^2(1+\theta_2)^2 \\
\end{array}\]

(Marchini et al. 2005)
Power to Detect Association for 1,500 Individuals where Both Loci Are Responsible for 5% of the Trait Variance

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

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

(Motsinger et al. 2007)
Selection an epistasis detection method

(Kilpatrick 2009)
Travelling the world of gene–gene interactions

Kristel Van Steen

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

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

Keywords: gene–gene interaction; variable selection; controlling false positives; translational medicine
Are all methods equal?

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

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

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

- Vermeulen et al (2007) re-confirmed that regression approaches suffer from inflated findings of false positives, and diminished power caused by the presence of sparse data and multiple testing problems, even in small simulated data sets only including 10 SNPS.
One popular method singled out

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

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

with X a possibly high-dimensional collection of confounders.

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

(adapted from slide: S Vansteelandt)
... Targeting mechanistic interactions

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

  (Rothman 1976, VanderWeele and Robins 2007)

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

- When both exposures have monotonic effects on the outcome, this can be strengthened to 
  \[ \beta > 0. \]
  (X suffices to control for confounding of the estimation of \(G_1, G_2\) effects)
... **Targeting mechanistic interactions**

(adapted from slide: S Vansteelandt)

- **Issues:**
  - Tests for sufficient cause interactions involve testing on the risk difference scale
  - Reality may show high-dimensional confounding
  - Estimators and tests for interactions are needed that are robust to model misspecification

- **Possible solution:**
  - Semi-parametric interaction models that attempt to estimate statistical interactions without modeling the main effects

- **Comment:** already hard in the case of two SNPs, using a theory of causality that is not widely accessible.
Towards alternative approaches

• What do we know?
  - Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders
  - Small $n$ big $p$ problems may give rise to curse of dimensionality problems (Bellman 1961); sparse cells issues
  - A lot more knowledge needs to be discovered, naturally giving rise to “data mining” type of strategies

• To keep in mind:
  - Data snooping: statistical bias due to inappr. use of data mining!
  - Biological knowledge integration
The curse of dimensionality in GWAI studies

• The curse of dimensionality refers to the fact that the convergence of any parametric model estimator to the true value of a smooth function defined on a space of high dimension is very slow (Bellman and Kalaba 1959).

• This is already a problem for main effects GWAS, when trying to assess those SNPs that are jointly most predictive for the disease or trait of interest, but is compounded when epistasis screenings are envisaged.

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

• What do we know?
  - Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders
  - Small $n$ big $p$ problems may give rise to curse of dimensionality problems (Bellman 1961); sparse cells issues
  - A lot more knowledge needs to be discovered, naturally giving rise to “data mining” type of strategies

• To keep in mind:
  - Data snooping: statistical bias due to inappr. use of data mining!
  - Biological knowledge integration
**Missing data**

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

“**A revision of LD based imputation strategies for GWAIIs is needed**”
Towards alternative approaches

• What do we know?
  - Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders
  - Small $n$ big $p$ problems may give rise to curse of dimensionality problems (Bellman 1961); sparse cells issues
  - A lot more knowledge needs to be discovered, naturally giving rise to “data mining” type of strategies

• To keep in mind:
  - Data snooping: statistical bias due to inappr. use of data mining
  - Biological knowledge integration
The multiple testing problem ~ significance assessment

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

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

“The interpretation of GWAIIs is hampered by undetected false positives”
Towards alternative approaches

How to compare methods... Is this truly a basic question?

- Power
- Type I error / False positives

<table>
<thead>
<tr>
<th></th>
<th>rs17116117</th>
<th>rs2513574</th>
<th>rs2519200</th>
<th>rs4938056</th>
<th>rs1713671</th>
<th>rs11936062</th>
<th>rs7126080</th>
<th>rs3770132</th>
<th>rs1933641</th>
<th>rs12339163</th>
<th>rs1933641</th>
<th>rs12853584</th>
<th>rs1217414</th>
<th>rs17116117</th>
<th>rs1169722</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M=1</strong></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td><strong>M=5</strong></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

**Number significant**: 6 6 6 6 7 5 7 5 6 7 6 6 7 6 7 6 3 3
Towards alternative approaches

To what extent do methods based on multifactor dimensionality reduction accommodate the aforementioned issues?
Significance assessment
What is the general setting?

Introduction

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

(Balding 2006)

- This is certainly the case in the context of genetic interaction studies.
Take-home messages

• It is important to verify the validity of the assumptions that underlie each corrective method for multiple testing, in order to select the most optimal corrective method for the data at hand.
• Several methods have been developed to curtail “classical” methods to GWAS settings
• Methods that accommodate correlated hypothesis tests (e.g., due to LD structure between genetic variants) include:
  - applying a Bonferroni correction using effective sample size derived from principal components (Nyholt et al 2004, Moskvina et al 2008),
  - exploiting haplotype blocking algorithms (Nicodemus et al 2005),
Take-home messages (cnt-ed)

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

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

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

- The field is not yet saturated with time-efficient false-positive controlling methods.
- New promising tools, even in the presence of millions of correlated markers, are emerging as we speak, claiming to be as accurate as permutation-based testing.
  - One of these methods is SLIDE (a Sliding-window Monte-Carlo approach for Locally Inter-correlated markers with asymptotic Distribution Errors corrected; Han et al 2009)
  - Another one is PACT (P values Adjusted for Correlated Tests) (Conneely and Boehnke 2007)
Towards alternative approaches

• What do we know?
  - Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders
  - Small $n$ big $p$ problems may give rise to curse of dimensionality problems (Bellman 1961); sparse cells issues
  - A lot more knowledge needs to be discovered, naturally giving rise to "data mining" type of strategies

• To keep in mind:
  - Data snooping: statistical bias due to inappr. use of data mining!
  - Biological knowledge integration
Data Integration

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

- **Broad definition** (Van Steen):

  “Combining evidences from different data resources, as well as data fusion with biological domain knowledge, using a variety of statistical, bioinformatics and computational tools”. 
Interpretation
A flexible framework for analysis acknowledging interpretation capability

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

(Moore et al 2005)
Example of a visual method: the interaction dendrogram

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

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

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

**Redundancy** – The interaction between attributes provides redundant information.

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

<table>
<thead>
<tr>
<th>Synergy</th>
<th>Redundancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>Blue</td>
</tr>
<tr>
<td>Orange</td>
<td>Green</td>
</tr>
<tr>
<td>Gold</td>
<td></td>
</tr>
</tbody>
</table>

---

*Université de Liège]*
Hierarchical clustering with average linkage

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

- The distance matrix used by the cluster analysis is constructed by calculating the information gained by constructing two attributes (Moore et al 2006, Jakulin and Bratko 2003, Jakulin et al 2003).
Data Integration: a solution?!

- Where in the GWAI process?

- Study design and samples collection
- Genotyping and genotypes calling
- Samples and markers quality control
- Exhaustive genome-wide screening for pair-wise SNP interactions
- Markers prioritization
- Genome-wide screening for pair-wise SNP interactions in the selected subset of loci
- Statistical replication of epistasis and biological validation

(slide: E Gusareva)
# Data Integration: a solution?!

<table>
<thead>
<tr>
<th>Where?</th>
<th>How?</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data preparation / Quality</td>
<td>Impute using different data control</td>
<td>Filling in the gaps or inducing LD-driven interactions?</td>
</tr>
<tr>
<td>Variable selection</td>
<td>Use a priori knowledge about networks and genetical / biological</td>
<td>Feature selection (dimensionality reduction) or loosing information?</td>
</tr>
<tr>
<td>Modeling</td>
<td>“Integrative” analysis</td>
<td>Obtaining a multi-dimensional perspective or combining/merging data in a single analysis?</td>
</tr>
<tr>
<td>Interpretation (validation)</td>
<td>Use a posteriori knowledge (e.g., Gene Ontology Analysis, Biofilter</td>
<td>Targeting known interactions or ruling out possibly relevant unknown interactions?</td>
</tr>
</tbody>
</table>
Plug and play

- The best advice towards success is to adopt different viewpoints to approach the biological problem (see later: example on Alzheimer)
- Plug and play … but not carelessly!
Model-Based Multifactor Dimensionality Reduction
Historical notes about MB-MDR

• Knowledge:
  - Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders
  - Small $n$ big $p$ problems may give rise to curse of dimensionality problems (Bellman 1961)
  - A lot more knowledge needs to be discovered, naturally giving rise to “data mining” type of strategies

• To keep in mind:
  - Data snooping: statistical bias due to inappr. use of data mining!
  - Biological knowledge integration
Historical notes about MB-MDR

- Start: Multifactor Dimensionality Reduction by MD Ritchie et al (2001)
A note aside
Multifactor Dimensionality Reduction (MDR)

The 6 steps of MDR
Towards MDR Final

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

Towards MDR Final

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

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

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0.045754</td>
</tr>
<tr>
<td>25%</td>
<td>0.168814</td>
</tr>
<tr>
<td>50%</td>
<td>0.237763</td>
</tr>
<tr>
<td>75%</td>
<td>0.321027</td>
</tr>
<tr>
<td>90%</td>
<td>0.423336</td>
</tr>
<tr>
<td>95%</td>
<td>0.489813</td>
</tr>
<tr>
<td>99%</td>
<td>0.623899</td>
</tr>
<tr>
<td>99.99%</td>
<td>0.872345</td>
</tr>
<tr>
<td>100%</td>
<td>1</td>
</tr>
</tbody>
</table>

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

(slide: L Mustavich)
The MDR Software

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

- Download information and much more can be found at http://www.multifactordimensionalityreduction.org/
Historical notes about MB-MDR (cnt-ed)

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

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

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

  - Rather, computation time is invested in optimal **association tests** to prioritize multilocus genotype combinations and in statistically valid permutation-based methods to assess **joint statistical significance**
  
  - Results of association tests are used to “label” multilocus genotype cells (for instance: increased / **no evidence**/ reduced risk, based on sign of “effect”) and to “quantify” the multilocus signal wrt the trait of interest, **“above and beyond lower order signals”**
Historical notes about MB-MDR

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

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

<table>
<thead>
<tr>
<th>SNP 40 x SNP 252 genotypes</th>
<th>Cases</th>
<th>Controls</th>
<th>OR</th>
<th>p-value</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1 = (0,0)</td>
<td>88</td>
<td>77</td>
<td>1.01</td>
<td>0.9303</td>
<td>0</td>
</tr>
<tr>
<td>c2 = (0,1)</td>
<td>102</td>
<td>114</td>
<td>0.73</td>
<td>0.0562</td>
<td>L</td>
</tr>
<tr>
<td>c3 = (0,2)</td>
<td>38</td>
<td>34</td>
<td>0.08</td>
<td>1.0000</td>
<td>0</td>
</tr>
<tr>
<td>c4 = (1,0)</td>
<td>50</td>
<td>59</td>
<td>0.76</td>
<td>0.1229</td>
<td>0</td>
</tr>
<tr>
<td>c5 = (1,1)</td>
<td>96</td>
<td>37</td>
<td>2.68</td>
<td>0.0000</td>
<td>H</td>
</tr>
<tr>
<td>c6 = (1,2)</td>
<td>18</td>
<td>28</td>
<td>0.55</td>
<td>0.0675</td>
<td>L</td>
</tr>
<tr>
<td>c7 = (2,0)</td>
<td>12</td>
<td>6</td>
<td>1.99</td>
<td>0.3399</td>
<td>0</td>
</tr>
<tr>
<td>c8 = (2,1)</td>
<td>14</td>
<td>18</td>
<td>0.67</td>
<td>0.3668</td>
<td>0</td>
</tr>
<tr>
<td>c9 = (2,2)</td>
<td>6</td>
<td>6</td>
<td>0.84</td>
<td>1.0000</td>
<td>0</td>
</tr>
</tbody>
</table>

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

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

- Model-Based MDR by Cattaert et al (2010) – fine-tuning MB-MDR
  
  - Pooling “alike” (for instance, all low-risk and all high-risk) multilocus genotypes leads to statistic distribution that is different from the theoretical distribution (data snooping)
  - Stable score tests, one multilocus p-value and permutation-based strategy (Cattaert et al 2010), rather than Wald tests, and relying on MAF dependent reference distributions (Calle et al 2008)
Historical notes about MB-MDR

- Model-Based MDR by Cattaert et al (2011) – genetic heterogeneity
Historical notes about MB-MDR

- Model-Based MDR by Cattaert et al (2011) – maximal power
### Historical notes about MB-MDR

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

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

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

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

The results prefixed by the symbol "≈" are extrapolated.
Historical notes about MB-MDR

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

\( \checkmark \): implemented

\( + \): under construction or in beta-testing
Historical notes about MB-MDR

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

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

L. Bastone\(^a\), M. Reilly\(^b\), D.J. Rader\(^b\), A.S. Foulkes\(^c\)

\(^a\)Division of Biostatistics, \(^b\)Cardiovascular Division and Center for Experimental Therapeutics, University of Pennsylvania School of Medicine, Philadelphia, Pa., and \(^c\)Department of Biostatistics, School of Public Health and Health Sciences, University of Massachusetts, Amherst, Mass., USA

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

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

(Shi et al 2006, unsupervised clustering with RFs)

\( \checkmark \): implemented

\( + \): under construction or in beta-testing
Historical notes about MB-MDR

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

OMs: Bullinaria 2004)
Key references about MB-MDR

Methodological papers

Introduction to Genetic Epidemiology

Genetic Association Interaction Studies

correction and improved association tests + recommendations on handling family-based designs]


- **Mahachie John JM, Cattaert T, Van Lishout F, Van Steen K (2011)** Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data. European Journal of Human Genetics 19, 696-703. [detailed study of C++ MB-MDR performance with quantitative traits]


- **Mahachie John JM, Cattaert T, Van Lishout F, Gusareva ES, Van Steen K (2012)** Lower-Order Effects Adjustment in Quantitative Traits Model-Based Multifactor Dimensionality


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

**Contact:** f.vanlishout@ulg.ac.be (C++ MB-MDR software engineer)
GWAI in practice
Protocol for GWAI analysis

1. LD pruning (e.g., SVS 7.5): window size 52 bp, window increment 1 bp LD r^2 threshold 0.75

2. Exhaustive genome-wide screening for pair-wise SNP interactions (BOOST analysis)

3. Markers prioritization (Biofilter): 177 candidate genes collected from "Alzheimer disease" KEGG pathway

4. Selection of SNPs basing on their function (SNPper - SNP Finder)

5. Replication analysis with alternative methods for epistasis detection: follow up the selected set of markers (MB-MDRm analysis, SD plot, logistic regression-based methods)

6. Replication of epistasis in the independent data and biological validation
First hurdle: Selection of most appropriate method

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

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

(Van Steen lab: in preparation)

<table>
<thead>
<tr>
<th></th>
<th>Bonferroni</th>
<th>EpiCruncher</th>
<th>Permutations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LR test</td>
<td>Score test</td>
<td>LR test</td>
</tr>
<tr>
<td>Test statistic</td>
<td></td>
<td>Test statistic</td>
<td>P-value</td>
</tr>
<tr>
<td></td>
<td>M=1</td>
<td>M=5</td>
<td>M=1</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>rs17116117</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>number significant</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

(Van Steen lab: in preparation)
Second hurdle: Level of detail – SNPS, genes, pathways, ...

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

<table>
<thead>
<tr>
<th>MTHFR</th>
<th>IL10</th>
<th>IL1A</th>
<th>IL1B</th>
<th>TF</th>
<th>HFE</th>
<th>IL6</th>
<th>ABCA1</th>
<th>DBH</th>
<th>INS</th>
<th>LRP1</th>
<th>CDK5R1</th>
<th>MAPT</th>
<th>NPC1</th>
<th>NR1H2</th>
<th>HMOX1</th>
<th>PPARA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>ns</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>ns</td>
</tr>
<tr>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ns</td>
<td>+</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>i</td>
<td>ns</td>
<td>ns</td>
<td>i</td>
<td>ns</td>
<td>ns</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>ns</td>
<td>i</td>
<td>i</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>ns</td>
</tr>
<tr>
<td>+</td>
<td>i</td>
<td>+</td>
<td>i</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>ns</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>ns</td>
<td>ns</td>
<td>+</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
</tbody>
</table>

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

(the marginal p-value < 0.05 for the MB-MDR_{2D} analysis)

Replication is highlighted by green; no replication is highlighted by red.
Third hurdle: Replication

(Mission Impossible @ google)
Replication

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

“Replication as a concept should be revised in the context of GWAI studies”
Optimal conditions for GWA (Interaction) replication

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

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

- Validation is not replication:
Challenges and opportunities

Contact:

kristel.vansteent@ulg.ac.be
Acknowledgments (MB-MDR)