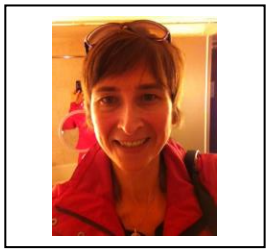
# Data Integration in Interaction Analysis

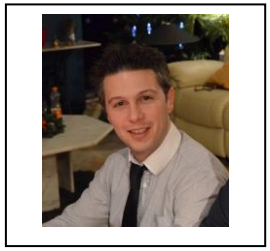
**Kristel Van Steen, PhD[2] (*)**

kristel.vansteen@ulg.ac.be

(*) Systems and Modeling Unit, Montefiore Institute, University of Liège, Belgium

(*) Bioinformatics and Modeling, GIGA-R, University of Liège, Belgium



BIO

X

X

# Bio$^3$: **Bio**statistics – **Bio**medicine - **Bio**informatics

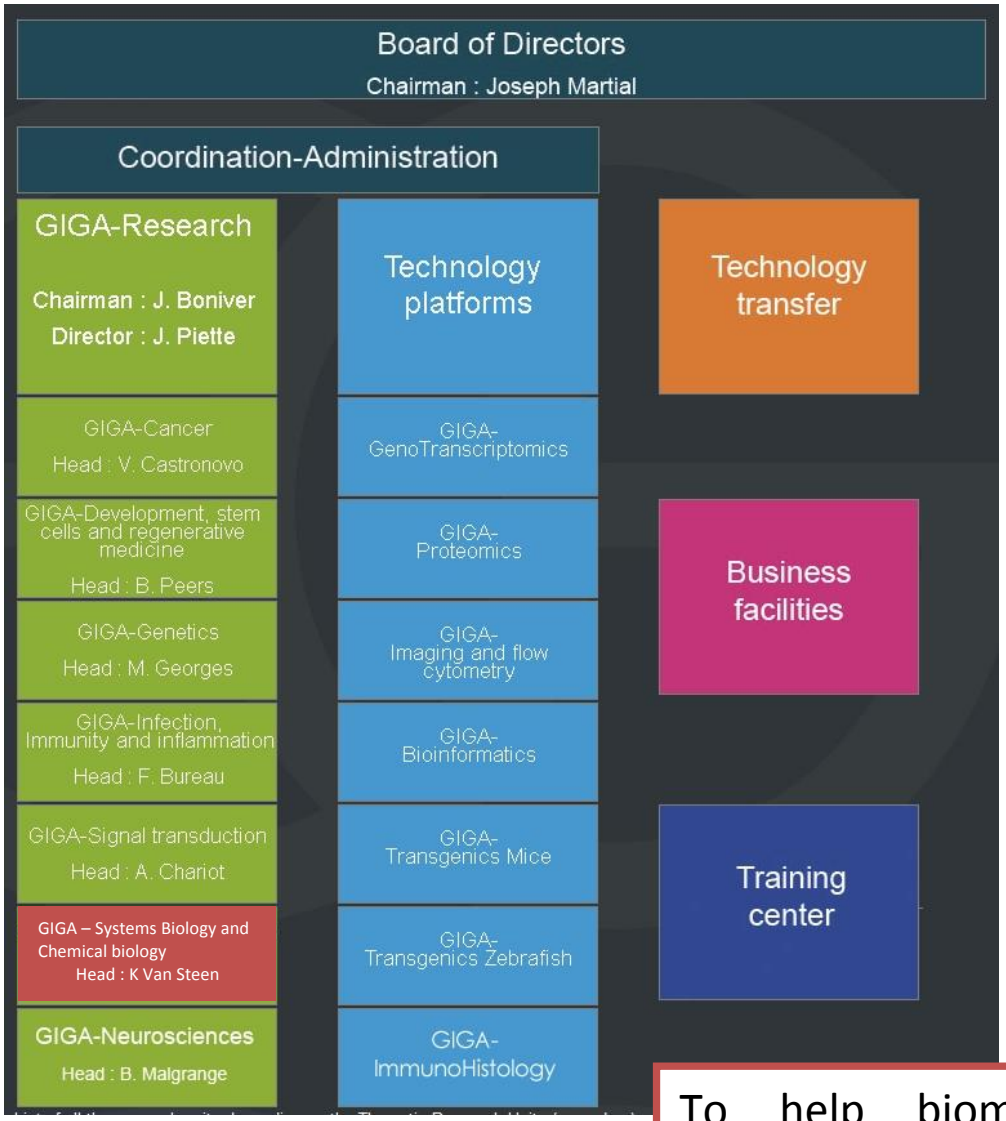



# OUTLINE

- Bio$^3$: Biomedicine, Biostatistics and Bioinformatics
- Motive and Opportunity for Integration
- (Up-scaling) Interaction Analyses
- The Rare Variant's Perspective
- Integration to enhance Biological Network Construction
- In Conclusion

# At the intersection of

# Biostatistics, Biomedicine and Bioinformatics

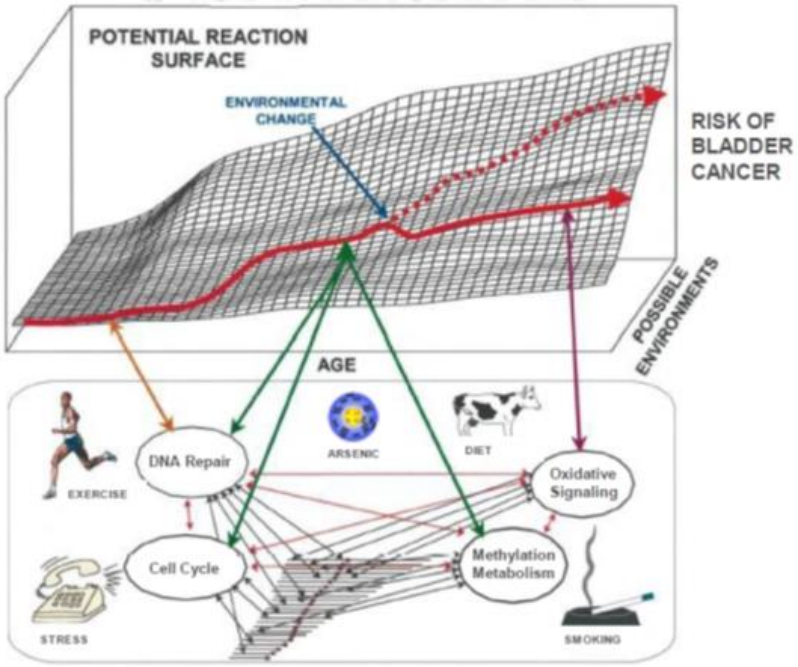



- Laboratory of molecular engineering and genetic engineering
- Laboratory of histology and mammalian cell culture
- Laboratory of mass spectrometry
- **Research unit of systems and modelling**

To help biomedical researchers carry out their investigations and analyze their data, as well as designing new statistical methods whenever they are needed.

## Mission

- To develop algorithms and methodologies for improved public health and personalized medicine
- To develop pipelines and software tools for human complex disease (gen-)omics




(adapted from Sing et al. 2003)

## Main Strategy

Find **motive** → Seek **opportunity** → Develop the **means**

# Motive and Opportunity for Integration

NEWS IN FOCUS

PROSPECTS

# New year, new science

Nature *looks at key findings and events that could emerge from the research world in 2011.*

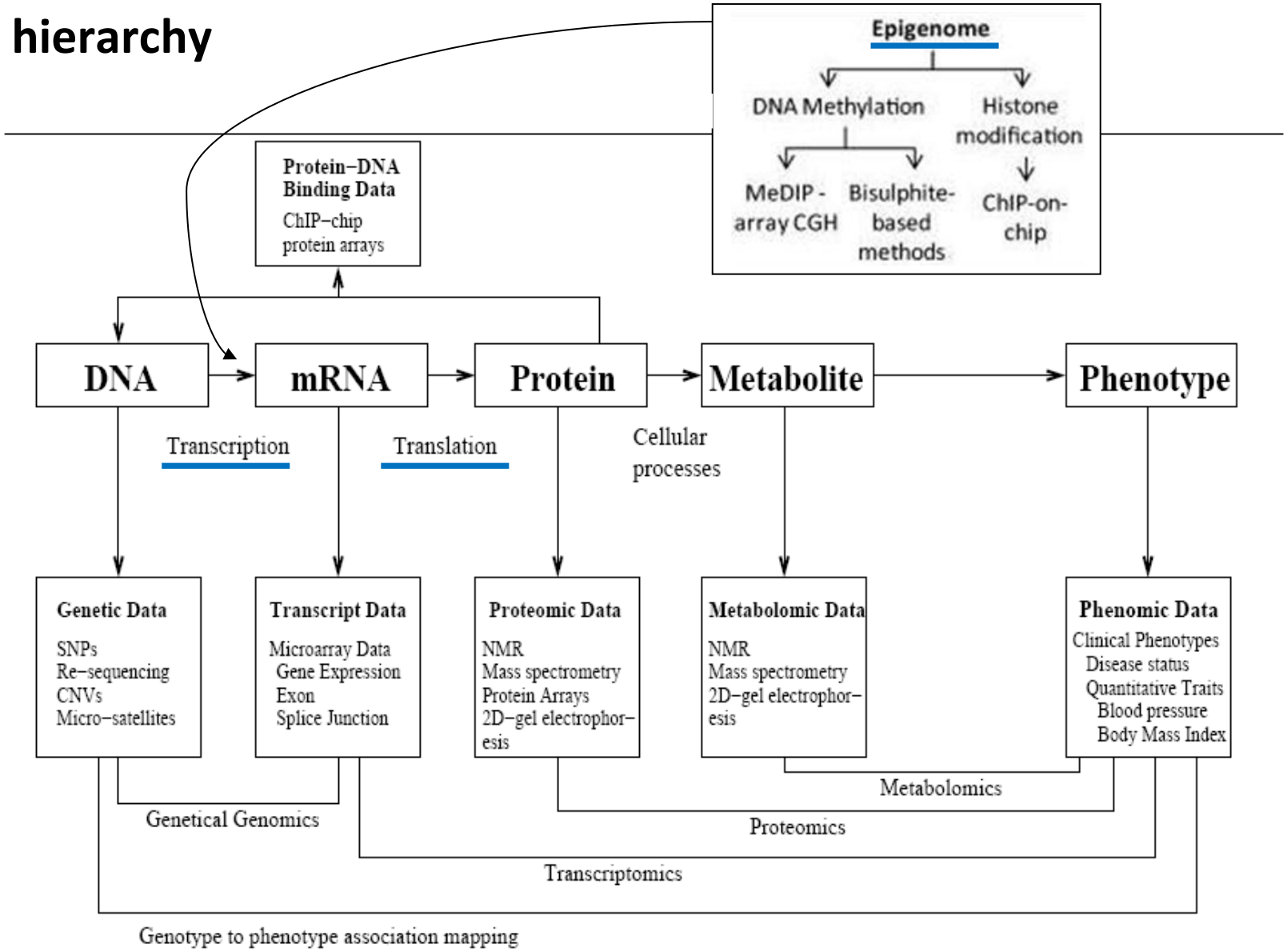
## GWAS PROVE THEIR WORTH

Genome-wide association studies (GWAS) have uncovered plenty of links between diseases and particular regions of the genome, but frustratingly haven't revealed much about the biochemistry behind these associations. In 2011, expect to see real mechanistic insights explaining how genes, and non-coding regions, affect the medical conditions they have been linked with. Metabolism, obesity and diabetes are among the hottest targets.

# Availability of different data resources

- Increased storage capacities and joint efforts to establish data banks with easy data access make available huge amounts of clinical, environmental, demographic data
- Rapid technological advances lead to various types of -omics data:
  - Genome
  - Epigenome
  - Transcriptome
  - Proteome
  - Metabolome
  - Phenome
  - ...

# Data hierarchy

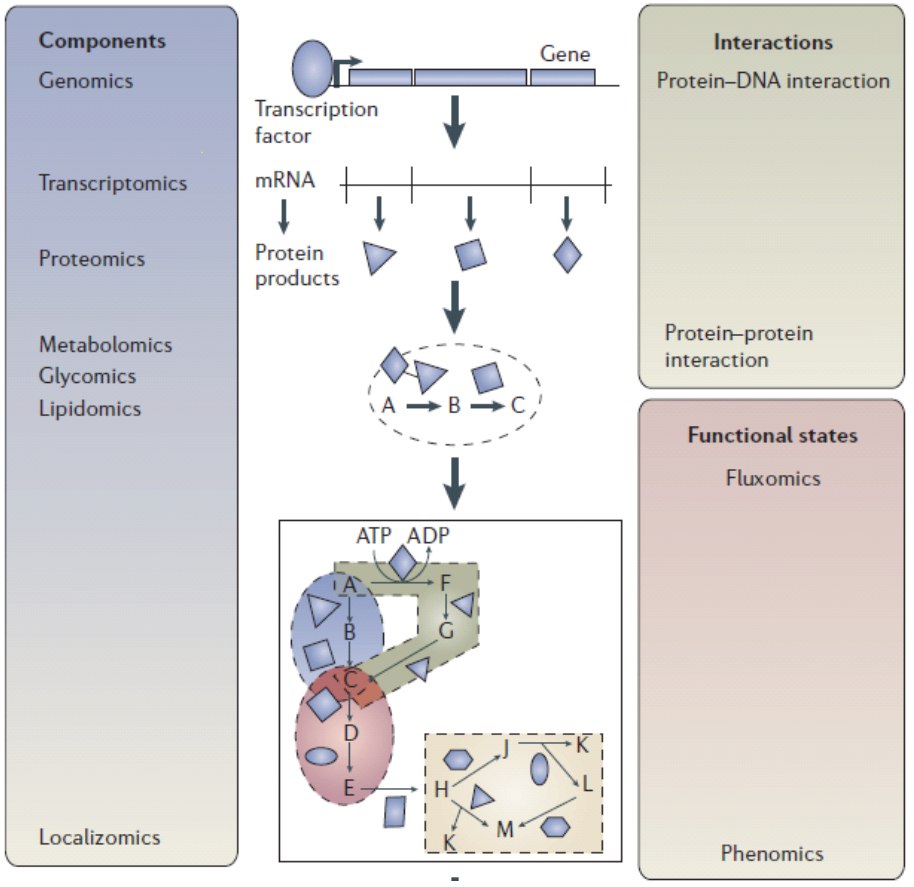



(adapted from: Davies et al 2009, Integrative genomics and functional explanation)

# 1+1 is more than 2 …

- Omics data provide comprehensive descriptions of nearly all components and interactions within the cell.

- Three data categories:
  - Components
  - Interactions
  - Functional-states

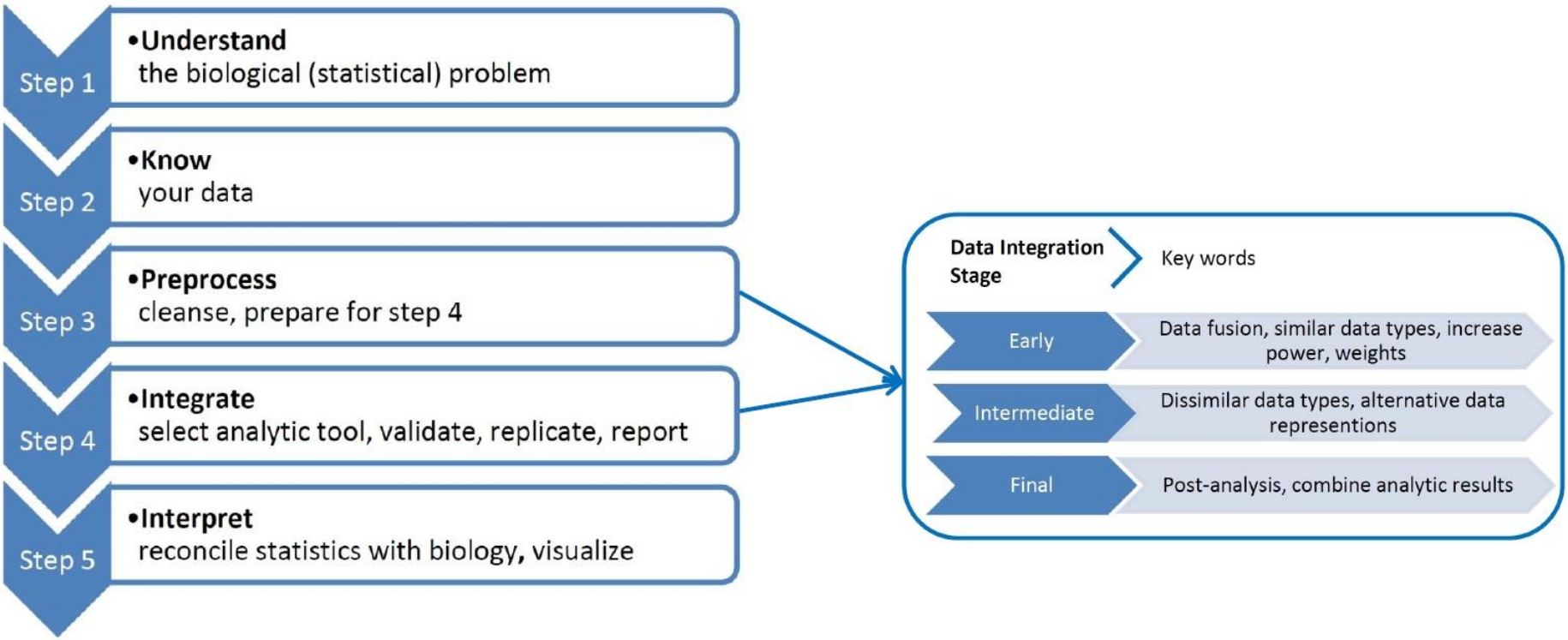



(Joyce and Palsson 2006)

## Functional genomics

- Full functional understanding of the etiology of a complex phenotype involves:
  - identifying the genetic, molecular, and environmental attributes that influence the phenotype, and
  - elucidating the biological pathway that fully defines the influence and describes how it occurs.
- The epigenome as the interface between the genome and the transcriptome, controlling long-term gene expression and integrating environmental signals, needs to be “integrated” into the functional genomics analysis.

# Key components of (statistical) data integration



# What's in a name?

- **Narrow:** "Process of statistically combining data from different sources to provide a unified view of the whole genome and make large-scale statistical inference" (Lu et al 2005).
- **Broad:** "Combining evidences from different data resources, as well as data fusion with biological domain knowledge, using a variety of statistical, bioinformatics and computational tools" (Van Steen)

→ Fusion or integration?

# What's in a name?

- Data fusion refers to fusing records on the same entity into a single file, and involves putting measures in place to detect and remove erroneous or conflicting data (Wang et al., 2014).
- Some definitions for "data fusion" use "data integration" in their definition. However, although some data integration efforts will rely on data fusion processes, data fusion and data integration are not equivalent.
- Oxley and Thorsen (Oxley & Thorsen, 2004) concluded that fusion can be defined as the process of optimally mapping several objects into a single object. In contrast, integration is the process of connecting systems (which may have fusion in them) into a larger system (Oxley & Thorsen, 2004).

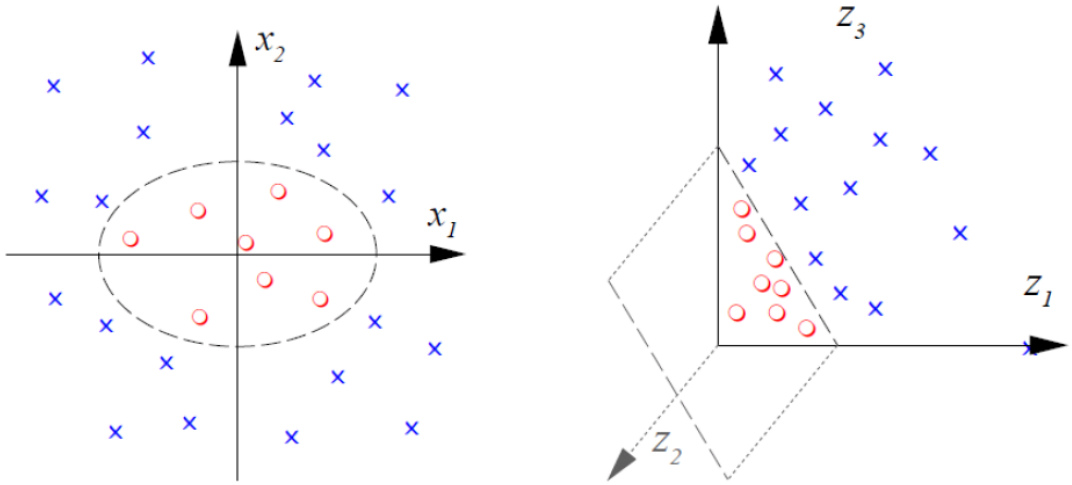
# Methodologies to “combine” data

## Crude division:

- Kernels
- Networks
- Components

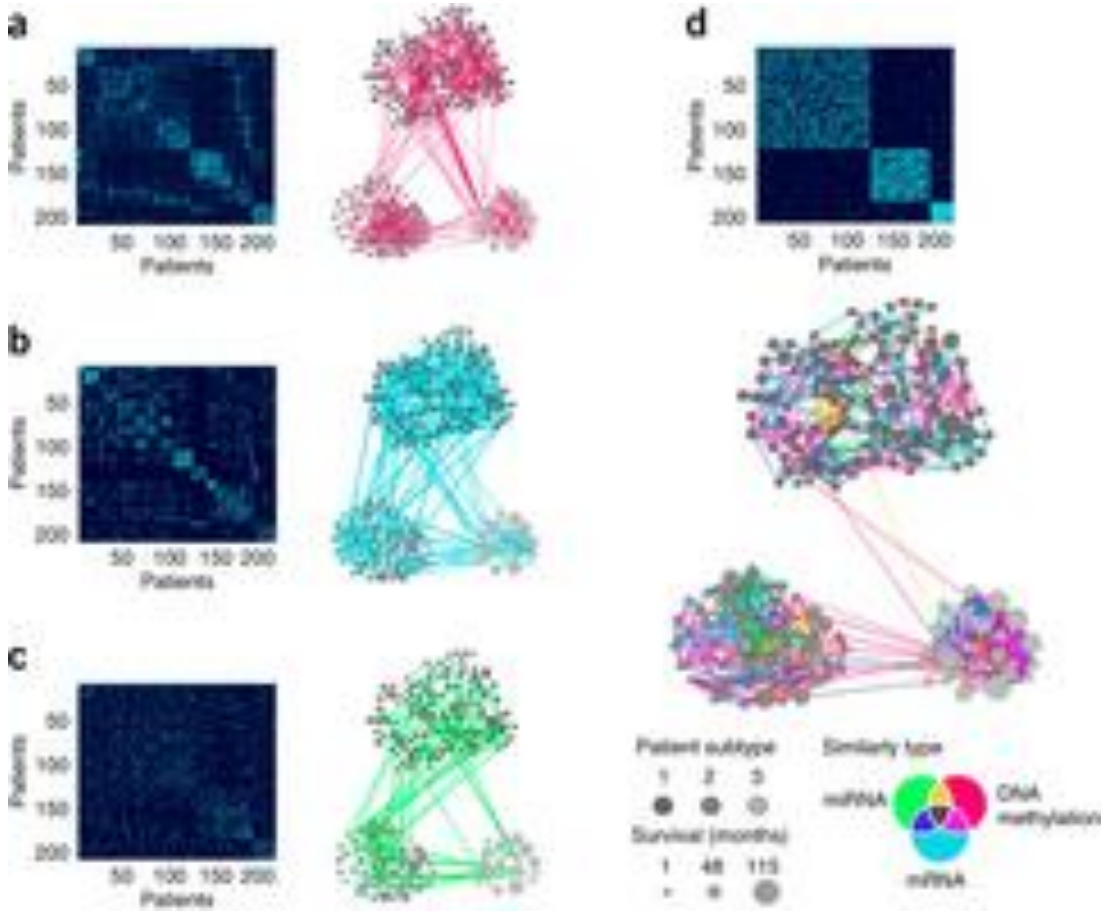
# Kernels

$$\Phi : \mathbf{R}^2 \rightarrow \mathbf{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$



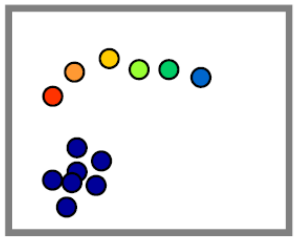

(http://www.ipam.ucla.edu/publications/ccstut/ccstut_9744.pdf)

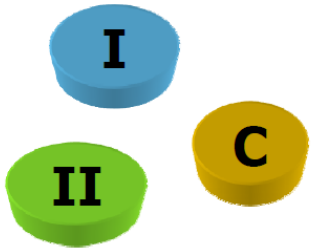
# Networks

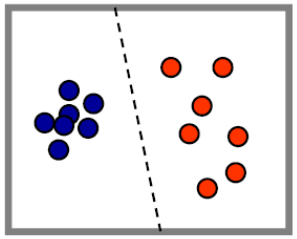


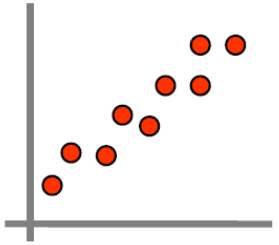

(http://www.nature.com/nmeth/journal/v11/n3/full/nmeth.2810.html)

# Components





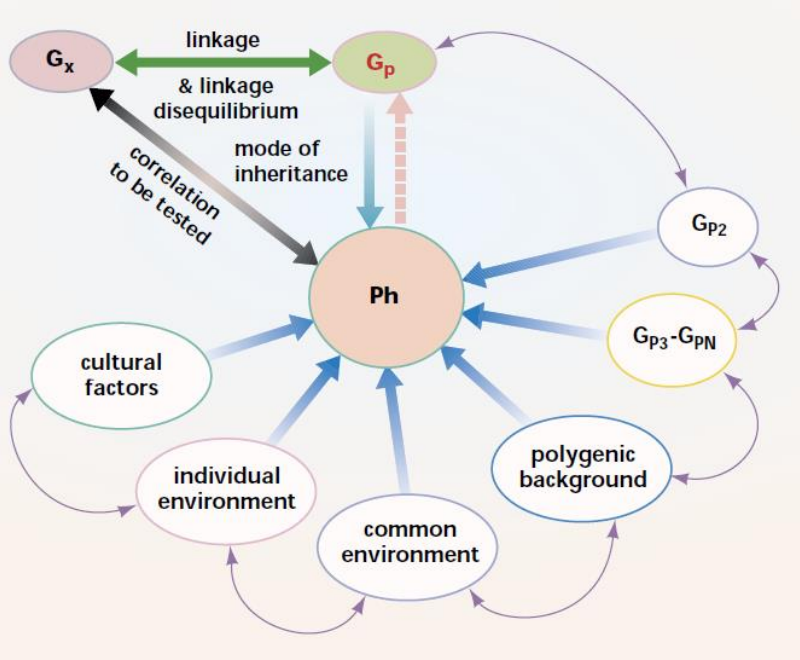


| Overview | Classification | Discrimination | Regression |
|---|---|---|---|
| Trends<br>Outliers<br>Quality Control<br>Biological Diversity<br>Patient Monitoring | Pattern Recognition<br>Diagnostics<br>Healthy/Diseased<br>Toxicity mechanisms<br>Disease progression | Discriminating between groups<br>Biomarker candidates<br>Comparing studies or instrumentation | Comparing blocks of omics data<br>Metab vs Proteomic vs Genomic<br>Correlation spectroscopy (STOCSY) |
| **PCA** | **SIMCA** | **PLS-DA**<br>**OPLS-DA** | **O2-PLS** |

(http://www.metabolomics.se/Courses/MVA/MVA%20in%20Omics_Handouts_Exercises_Solutions_Thu-Fri.pdf)

# Interaction Analysis

# The complexity of complex diseases

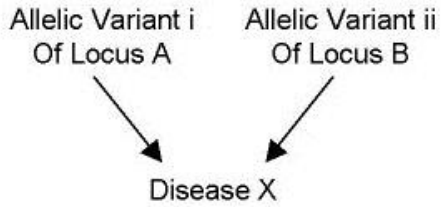
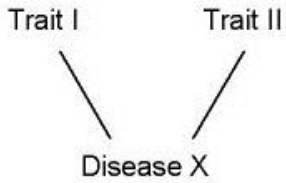
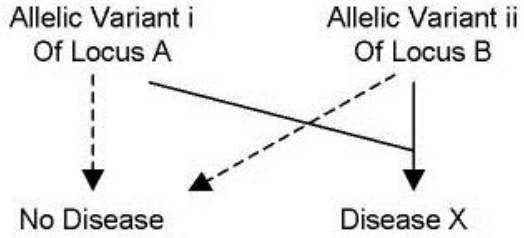



(Weiss and Terwilliger 2000)

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

(Moore 2008)

# Factors complicating analysis of complex genetic disease

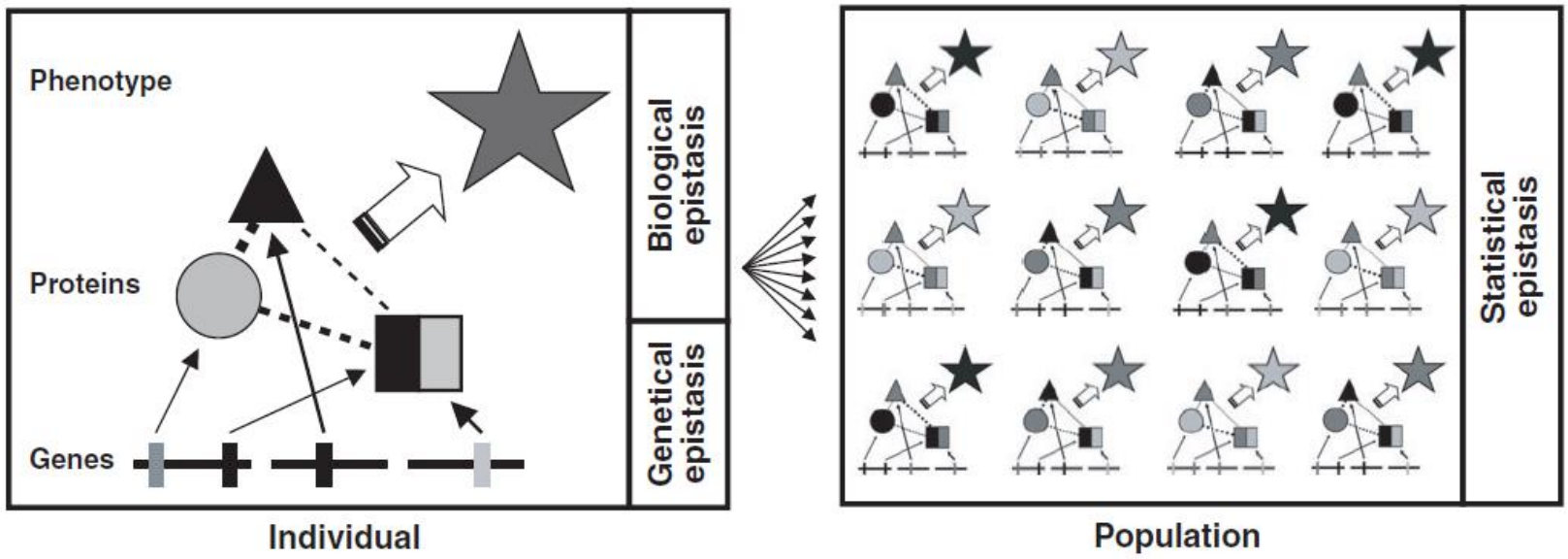
| | Locus Heterogeneity | Trait Heterogeneity | Gene-Gene Interaction |
|---|---|---|---|
| **Definition** | when two or more DNA variations in distinct genetic loci are independently associated with the same trait | when a trait, or disease, has been defined with insufficient specificity such that it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

(Thornton-Wells et al. 2006)

# Factors complicating analysis of complex genetic disease

## Gene-gene interactions

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




(Moore 2005)

# What's in a name?

- Wikipedia (26/05/2014)

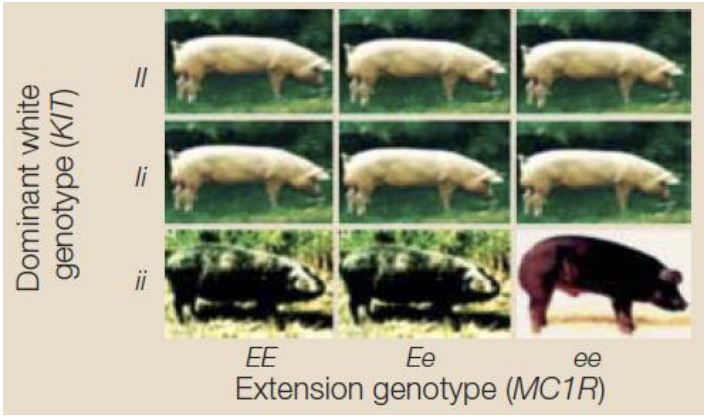
**"Epistasis** when the effect of one gene depends on the presence of one or more 'modifier genes' (genetic background). Similarly, epistatic mutations have different effects in combination than individually. It was originally a concept from genetics but is now used in biochemistry, population genetics, computational biology and evolutionary biology. It arises due to interactions, either between genes, or within them leading to non-additive effects. Epistasis has a large influence on the shape of evolutionary landscapes which leads to profound consequences for evolution and evolvability of traits."

# What's in a name?

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

**"compositional epistasis"**

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




(Carlborg and Haley 2004)

# Compositional epistasis

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

| Genotype at locus B/G | gg | gG | GG |
|---|---|---|---|
| **bb** | White | Grey | Grey |
| **bB** | Black | Grey | Grey |
| **BB** | Black | Grey | Grey |

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

(Cordell 2002)

# 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, see also VanderWeele 2009).

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

# Effect modification or interaction …

- In statistics: interaction = deviations from a model of additive multiple effects
- *"Interaction"* often confused with "*(effect) modification"*
- Miettinen (1985): modification = measure of association between a SNP and a trait is not constant across another characteristic

  (e.g., population strata)

  Such a characteristic typically changes the effect of the variate of interest to the trait: this phenomenon is often referred to in the literature as *effect modification*

  (some epidemiology text books reserve the term effect modification if the modification is linked to a *causal mechanism* and use the reduced term *modification* otherwise)

## Effect modification or interaction

- Interaction is defined in terms of the effects of 2 interventions
- Effect modification is defined in terms of the effect of one intervention varying across strata of a second variable.

  ✓ Effect modification can be present with no interaction.
  ✓ Interaction can be present with no effect modification.

- There are settings in which it is possible to assess effect modification but not interaction, or to assess interaction but not effect modification.
- There are settings in which effect modification and interaction coincide. (VanderWeele 2009)

# Theory and practice

- **Theory:**
  - understanding the principles of epistasis / effect modification

- **Practice:**
  - formulate the research question
  - decide upon a study design (e.g., families or not, GWAs SNP chip array or exome sequences)
  - collect samples (e.g; model organisms or individuals)
  - apply a methodology
  - answer the initial research question

# The “observed” occurrences of epistasis – model organisms

- Carlborg and Haley (2004):
  - Epistatic QTLs without individual effects have been found in various organisms, such as birds[26,27], mammals[28–32], Drosophila melanogaster[33] and plants[18,34].
  - However, other similar studies have reported only low levels of epistasis or no epistasis at all, despite being thorough and involving large sample sizes[35–37].
- This clearly indicates the complexity with which multifactorial traits are regulated; no single mode of inheritance can be expected to be the rule in all populations and traits.

## Great expectations

- From an evolutionary biology perspective, for a phenotype to be buffered against the effects of mutations, it must have an underlying genetic architecture that is comprised of networks of genes that are redundant and robust.
- The existence of these networks creates dependencies among the genes in the network and is realized as gene-gene interactions or (*trans*-) epistasis.
- This suggests that epistasis is not only important in determining variation in natural and human populations, but should also be more widespread than initially thought (rather than being a limited phenomenon).

# The “observed” occurrences of epistasis – humans

- Phillips et al (2008):
  - There are several cases of epistasis appearing as a statistical feature of association studies of human disease.
  - A few recent examples include coronary artery disease[63], diabetes[64], bipolar effective disorder[65], and autism[66].
  - So far, only for some of the reported findings additional support could be provided by functional analysis, as was the case for multiple sclerosis (Gregersen et al 2006).

# The “observed” occurrences of epistasis – humans

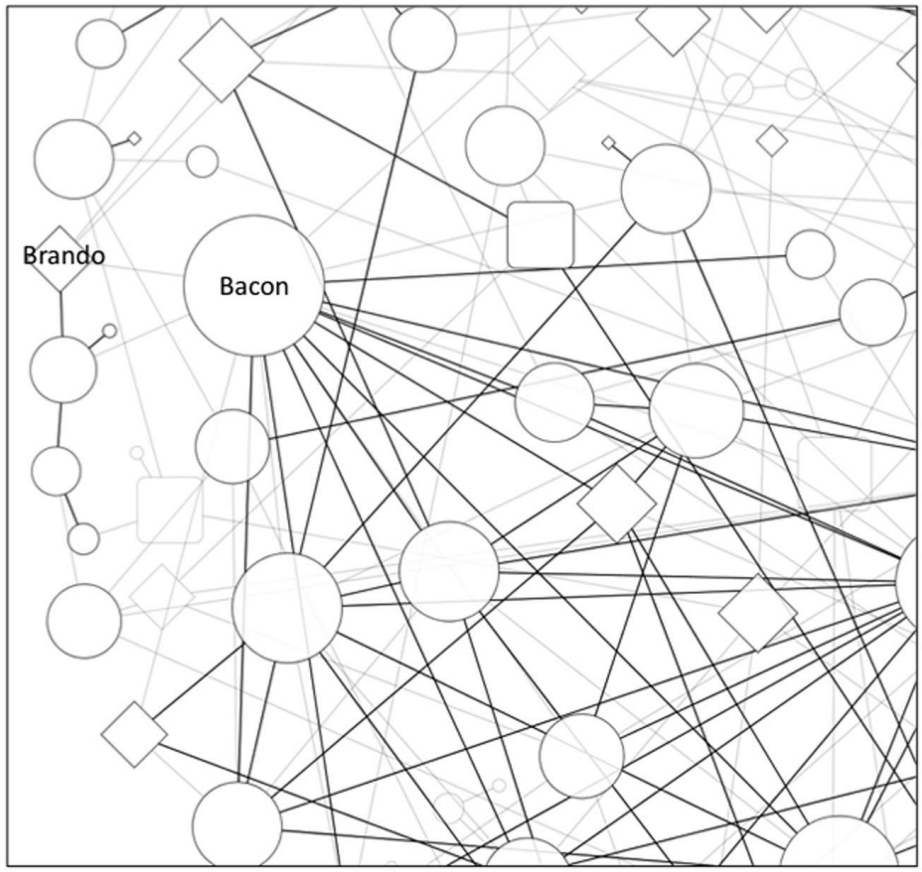
- More recent examples include:
  - Alzheimer’s disease (Combarros et al 2009),
  - psoriasis (WTCCC2 2010),
  - breast cancer (Ashworth et al. 2011),
  - ankylosing spondylitis (WTCCC 2011),
  - total IgE (Choi et al. 2012)
  - High-Density Lipoprotein Cholesterol Levels (Ma et al. 2012)
- So far, only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

## Great expectations - empowering personal genomics

- Considering the epic complexity of the transcriptions process, the genetics of gene expression seems just as likely to harbor epistasis as biological pathways.
- When examining HapMap genotypes and gene expression levels from corresponding cell lines to look for cis-epistasis, over 75 genes pop up where SNP pairs in the gene's regulatory region can interact to influence the gene's expression.

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

# Complementing insights from GWA studies




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

(McKinney et al 2012)

# Epistasis and phantom heritability

(Maher 2008)

## Epistasis and phantom heritability

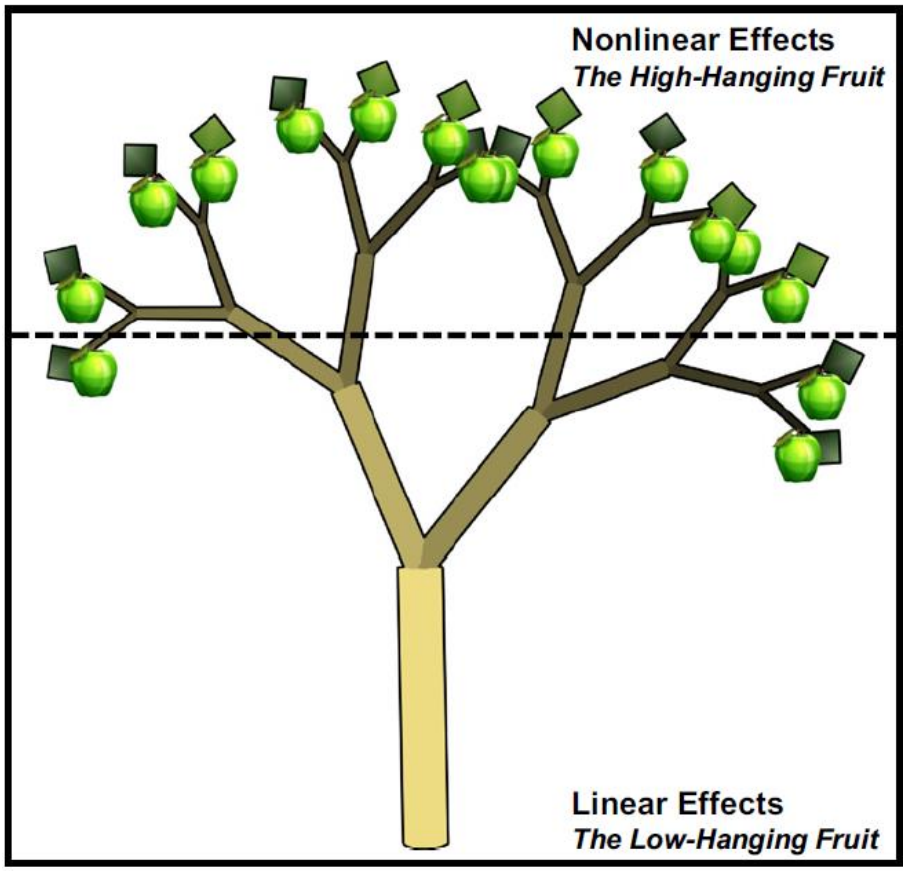
- Human genetics has been haunted by the mystery of “missing heritability” of common traits.
- Although studies have discovered >1,200 variants associated with common diseases and traits, these variants typically appear to explain only a minority of the heritability.
- The proportion of heritability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

## Epistasis and phantom heritability

- Overestimation of the total heritability can create “phantom heritability.”
  - estimates of total heritability implicitly assume the trait involves no genetic interactions (epistasis) among loci
  - this assumption is not justified
  - under such models, the total heritability may be much smaller and thus the proportion of heritability explained much larger.
- For example, 80% of the currently missing heritability for Crohn's disease could be due to genetic interactions, if the disease involves interaction among three pathways. (Zuk et al 2012)

# Traveling the world of interactions


You can
You have to



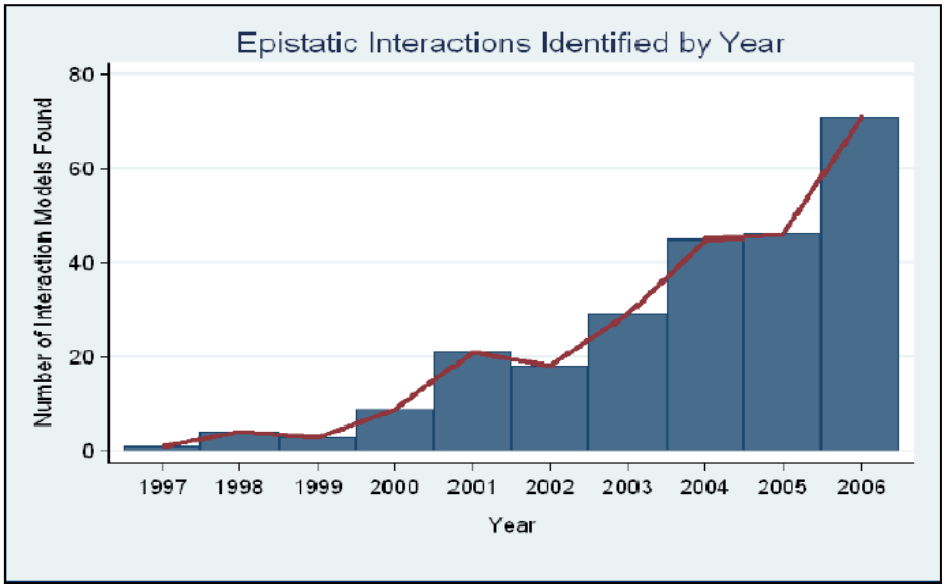

- Few SNPs with moderate to large independent and additive main effects

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

(Moore and Williams 2009)

# A growing toolbox

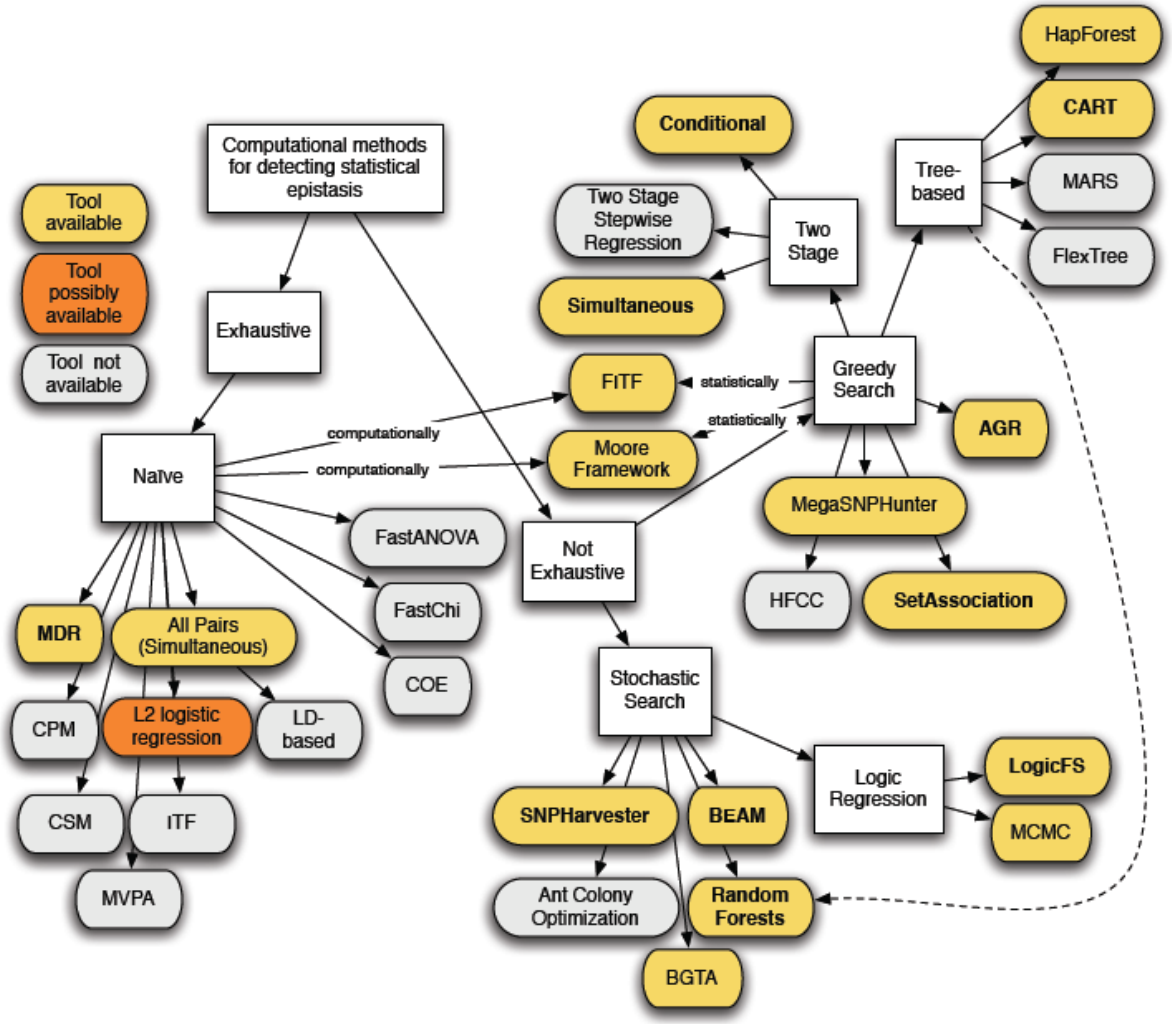
- The number of identified epistasis effects in humans, showing susceptibility to common complex human diseases, follows a steady growth curve (Emily et al 2009, Wu et al 2010), due to the growing number of toolbox methods and approaches.




(Motsinger et al. 2007)

# Our 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)

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

# Travelling the world of gene–gene interactions

*Kristel Van Steen*

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

## Abstract

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

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

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

REVIEW PAPER

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

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

Received: 1 March 2012 / Accepted: 11 June 2012 / Published online: 4 July 2012


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

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

## Are all methods equal?

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

**"These criteria show the diversity of methods and approaches and complicates making honest comparisons".**

## One popular method singled out

- North et al (2005) showed that in some instances the inclusion of interaction parameters - within a regression framework - is advantageous but that there is no direct correspondence between the interactive effects in the logistic regression models and the underlying penetrance based models displaying some kind of epistasis effect
- Vermeulen et al (2007) re-confirmed that regression approaches suffer from inflated findings of false positives, and diminished power caused by the presence of sparse data and multiple testing problems, even in small simulated data sets only including 10 SNPS.

## One popular method singled out

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

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

with X a possibly high-dimensional collection of confounders.

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

(adapted from slide: S Vansteelandt)

## … Targeting mechanistic interactions

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

(Rothman 1976, VanderWeele and Robins 2007)

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

$$\beta > \beta_0.$$

- When both exposures have monotonic effects on the outcome, this can be strengthened to

$$\beta > 0.$$

(X suffices to control for confounding of the estimation of $G_1, G_2$ effects)

## …Targeting mechanistic interactions (adapted from slide: S Vansteelandt)

- Issues:
  - Tests for sufficient cause interactions involve testing on the risk difference scale
  - Reality may show high-dimensional confounding
  - Estimators and tests for interactions are needed that are robust to model misspecification
- Possible solution:
  - Semi-parametric interaction models that attempt to estimate statistical interactions without modeling the main effects
- Comment: already hard in the case of two SNPs, using a theory of causality that is not widely accessible.

# Towards alternative approaches

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

## The curse of dimensionality in GWAI studies

- The curse of dimensionality refers to the fact that the convergence of any parametric model estimator to the true value of a smooth function defined on a space of high dimension is very slow (Bellman and Kalaba 1959).
- This is already a problem for main effects GWAS, when trying to assess those SNPs that are jointly most predictive for the disease or trait of interest, but is compounded when epistasis screenings are envisaged

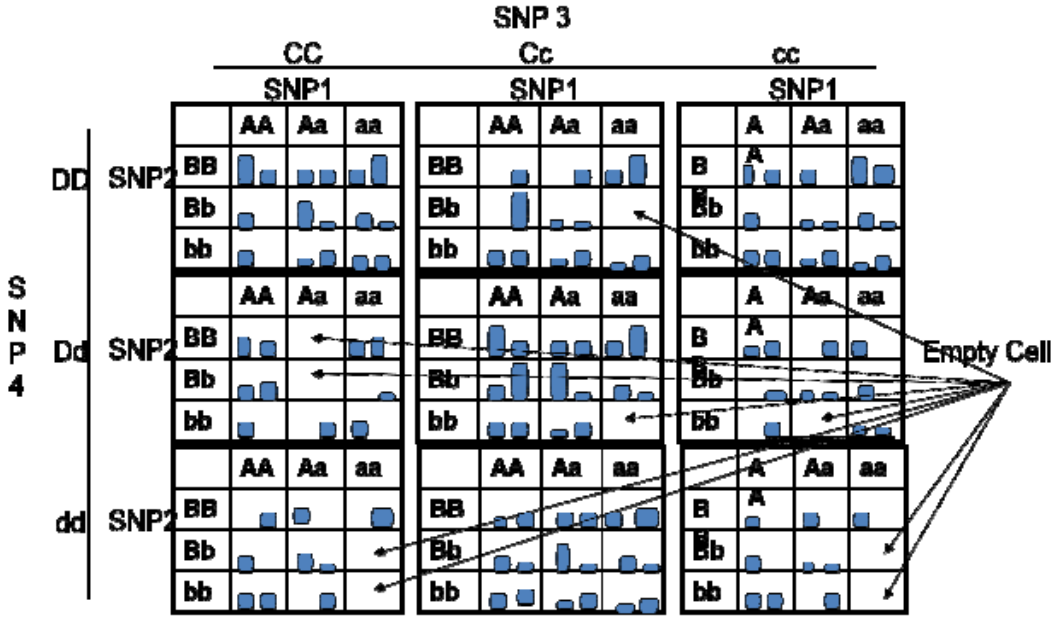
**"Parametric model (mis)specification is of major concern, especially in the presence of high-dimensional confounders"**

# Towards alternative approaches

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

# Missing data

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




(slide: C Amos)

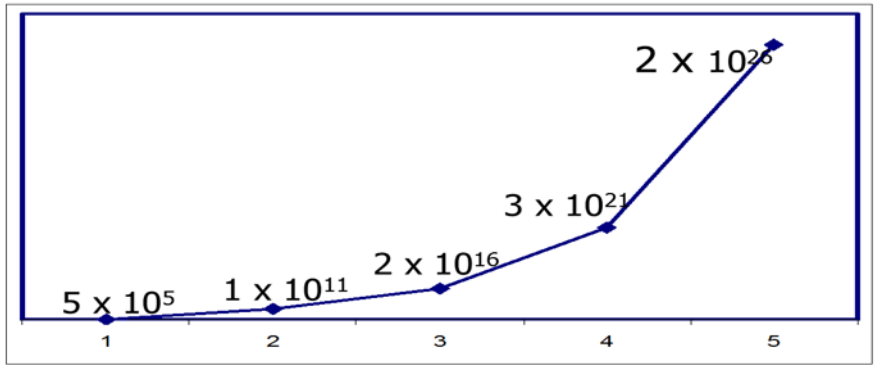
**"A revision of LD based imputation strategies for GWAIs is needed"**

# Towards alternative approaches

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

## The multiple testing problem ~ significance assessment

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

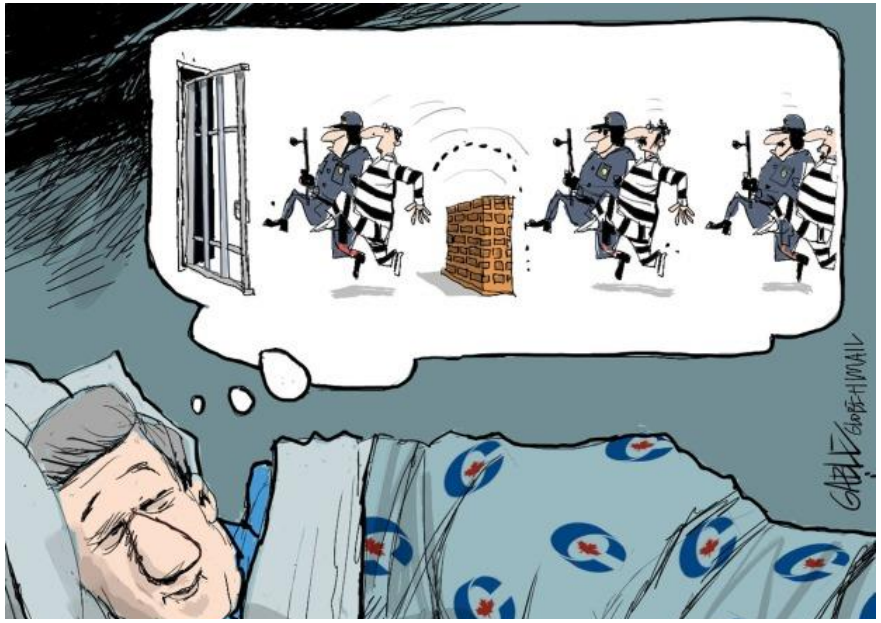



~500,000 SNPs span 80% of common variation (HapMap)

**"The interpretation of GWAIs is hampered by undetected false positives"**

# BIO3 and Effect Modification / Interactions

- **Thematic objectives:**
  - Identify genetic factors associated with complex traits of interest,
  - Identify subject characteristics that may alter the aforementioned association (descriptive modifiers),
  - Explain observed modifications of the determinant-outcome association (causal modifiers)



(Boston Globe)

# A new method under the sun

Choice 1
- **Variable selection**
(entropy-based filtering, mutual info, biological knowledge)

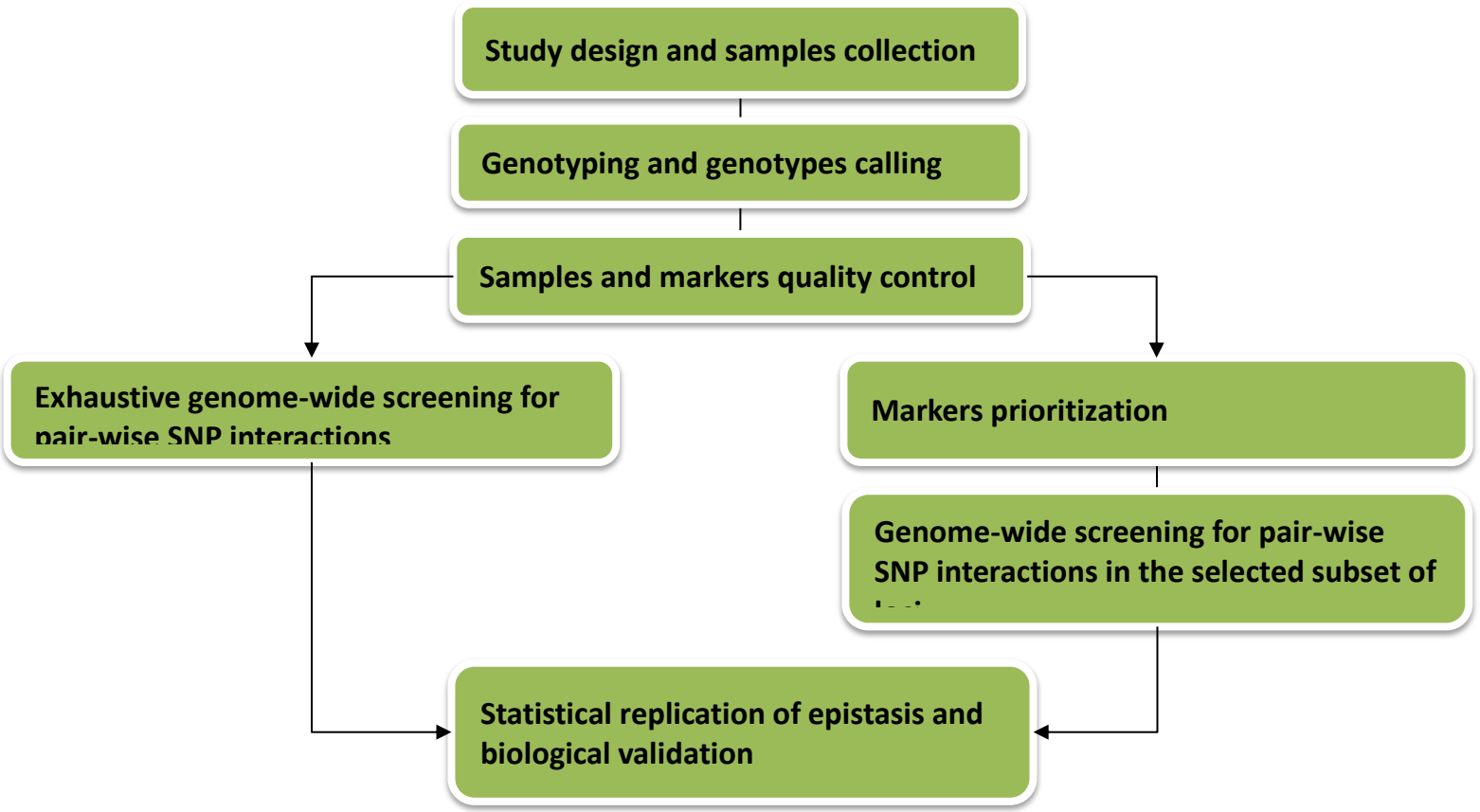
Choice 2
- **Modelling/ testing**
(balance between what is available and what is feasible ~ IT-infrastructure)

Choice 3
- **Significance assessment**
(permutation based is wholy grail? accounting for dependent tests)

Choice 4
- **Interpretation**
(integration, visualization, statistical epistasis networks)

# Data Integration: a solution?!

- Where in the GWAI process?

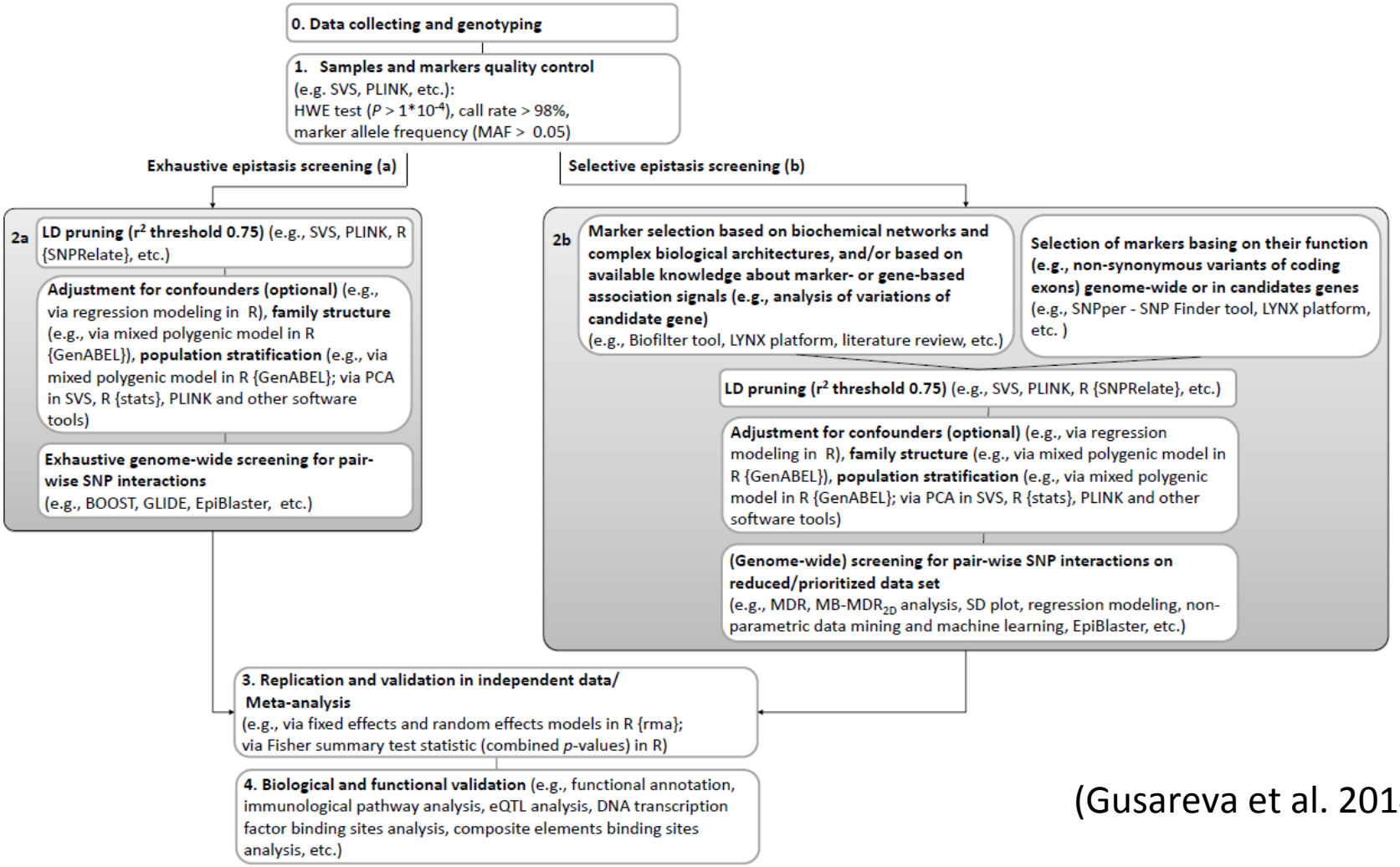



(slide: E Gusareva)

# Data Integration: a solution?!

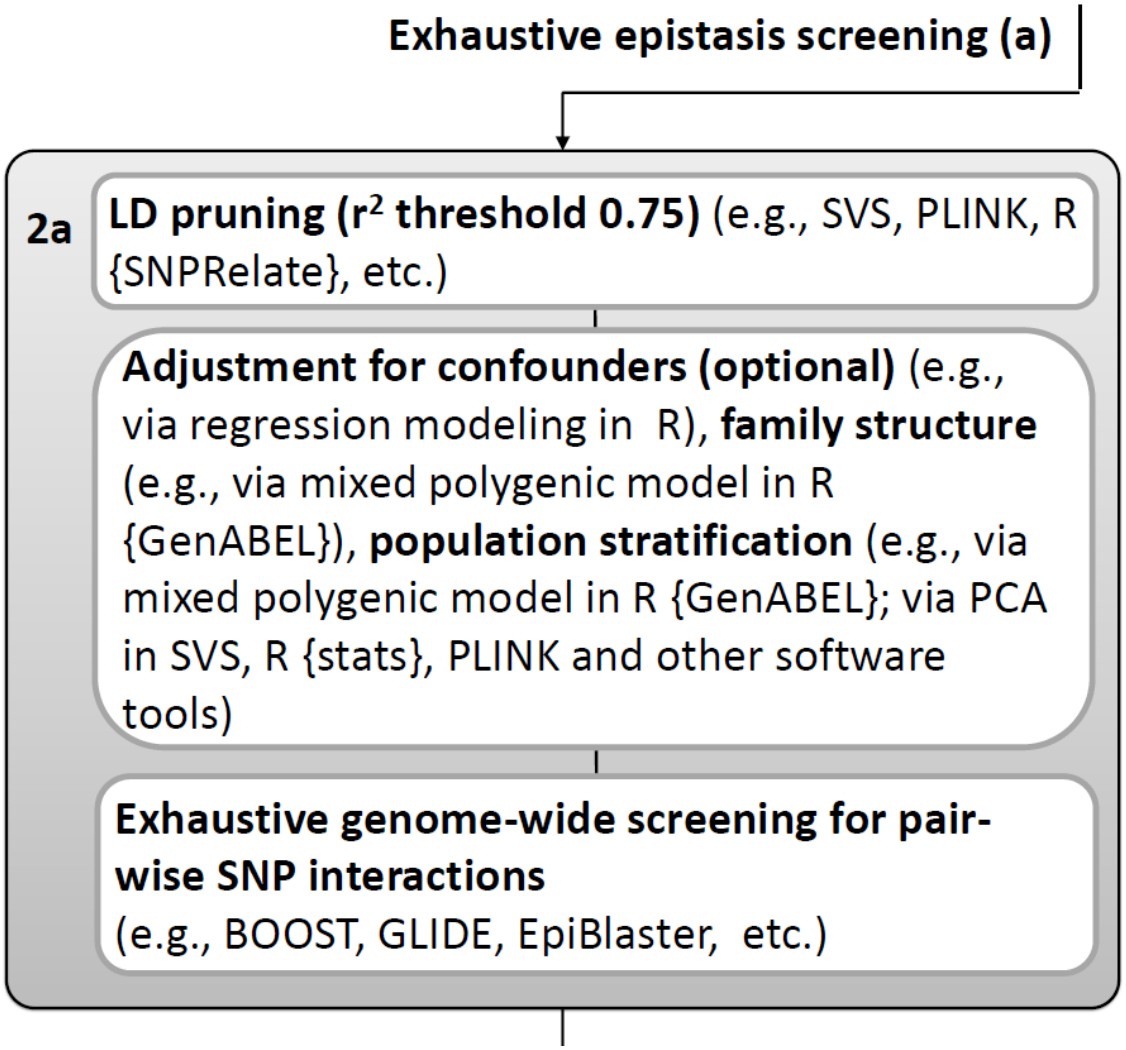
| Where? | How? | Comments |
|---|---|---|
| Data preparation / Quality control | Impute using different data resources | Filling in the gaps or inducing LD-driven interactions? |
| Variable selection | Use a priori knowledge about networks and genetical / biological interactions (e.g., Biofilter) | Feature selection (dimensionality reduction) or loosing information? |
| Modeling | "Integrative" analysis | Obtaining a multi-dimensional perspective or combining/merging data in a single analysis? |
| Interpretation (validation) | Use a posteriori knowledge (e.g., Gene Ontology Analysis, Biofilter – Bush et al. 2009) | Targeting known interactions or ruling out possibly relevant unknown interactions? |

# Integration is only part of the solution

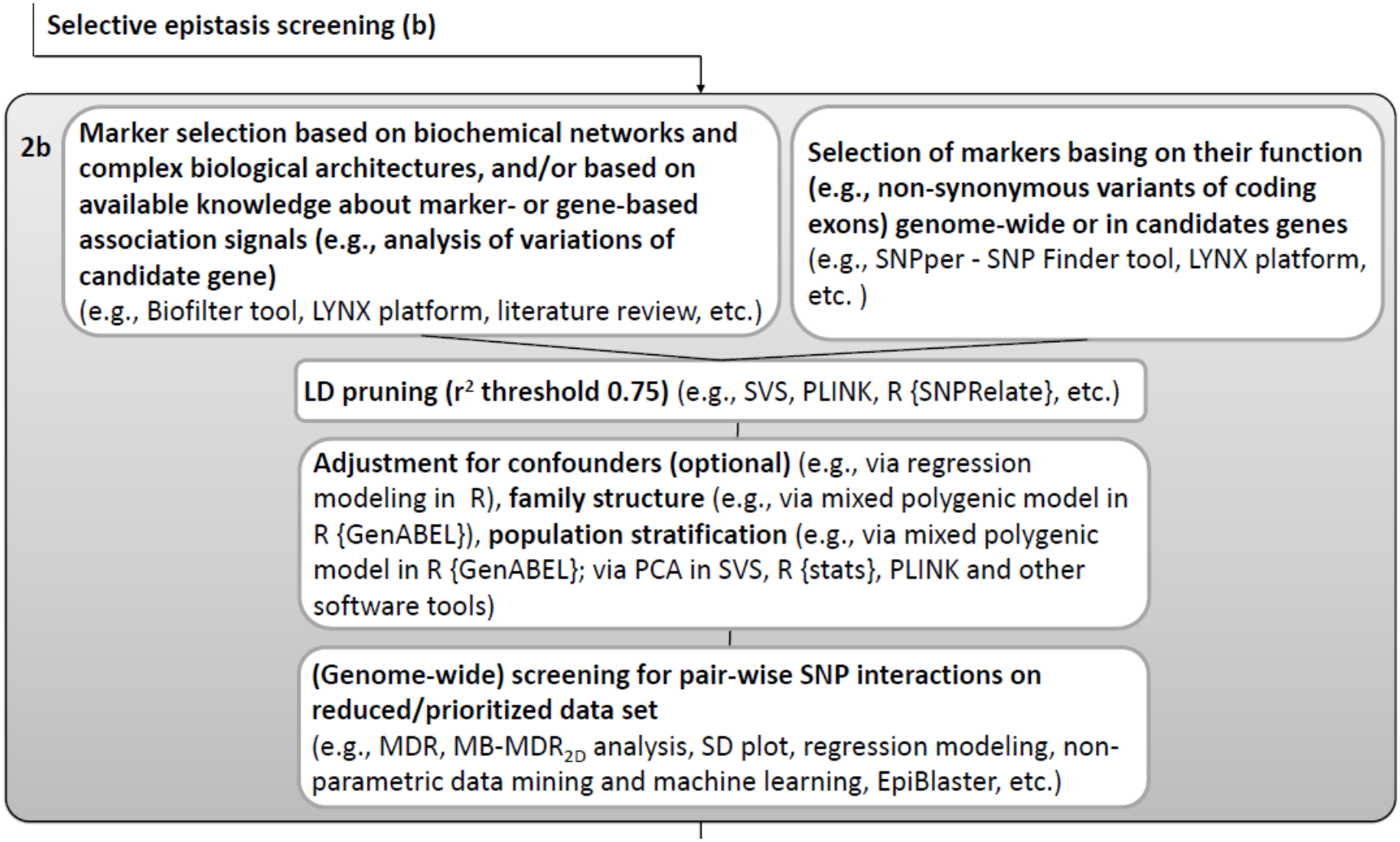



(Gusareva et al. 2014)

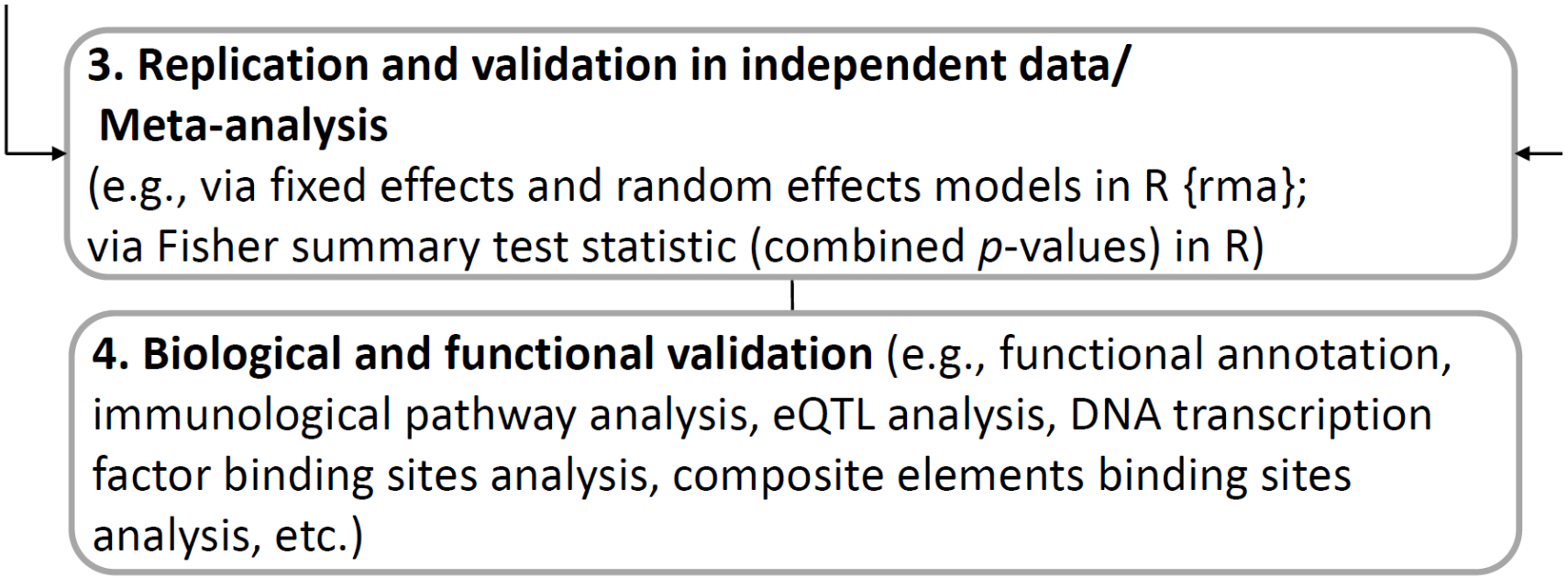
# Integration is only part of the solution



# Integration is only part of the solution



# Integration is only part of the solution




(Gusareva et al. 2014)

# General advice

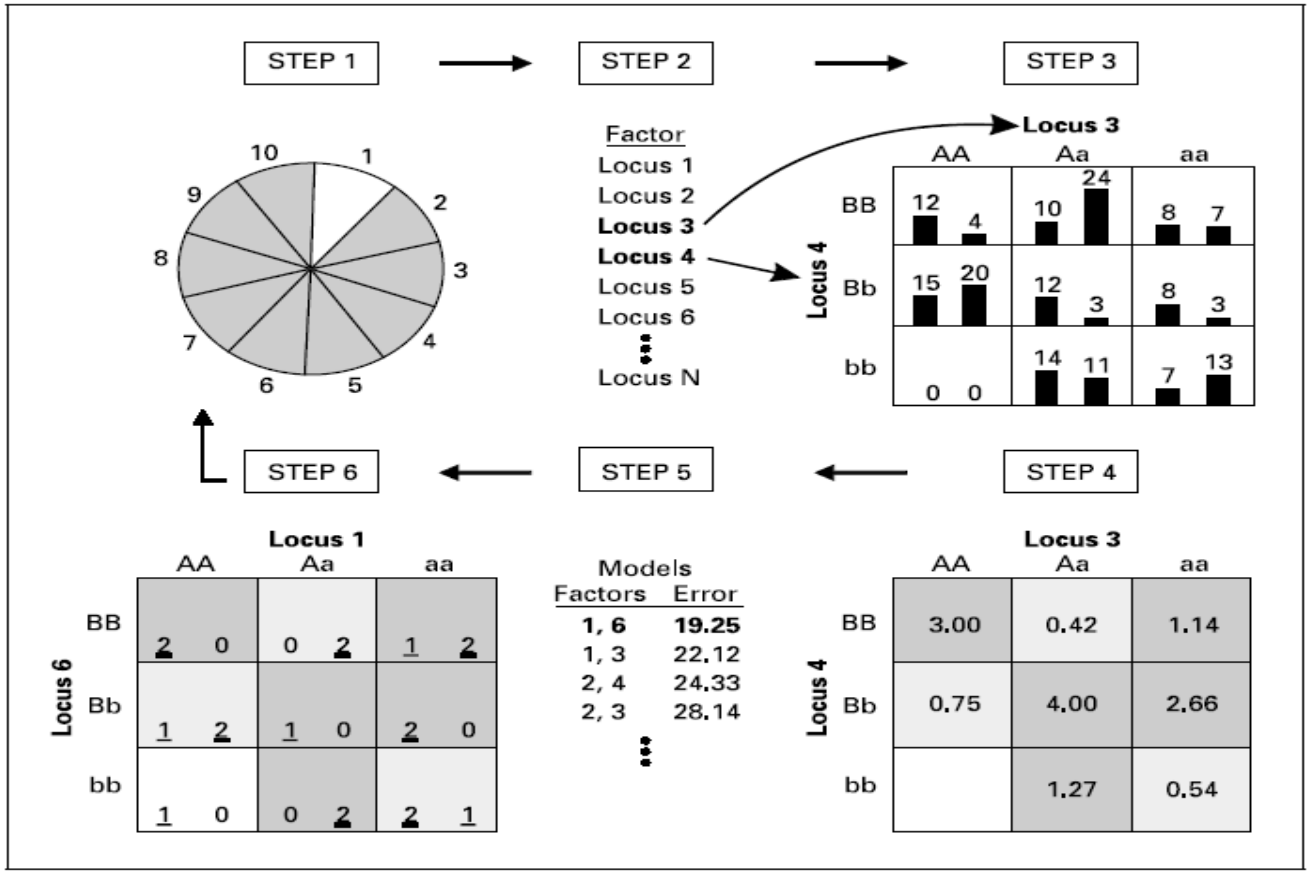
- The best advice towards success is to adopt different viewpoints to approach the biological problem (BIO3 examples on Alzheimer's)
- Plug and play … but not carelessly!


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

# An alternative viewpoint

# MB-MDR: not just another flavor of MDR

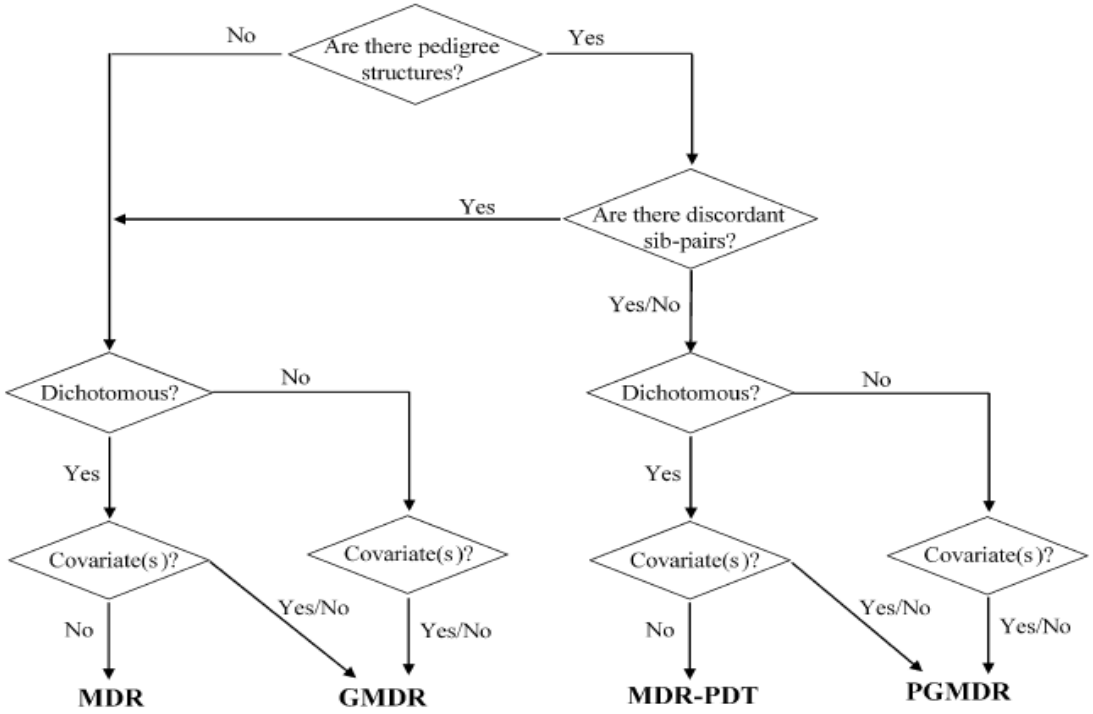
- Start: Multifactor Dimensionality Reduction by MD Ritchie et al. (2001)



# Historical notes about MB-MDR

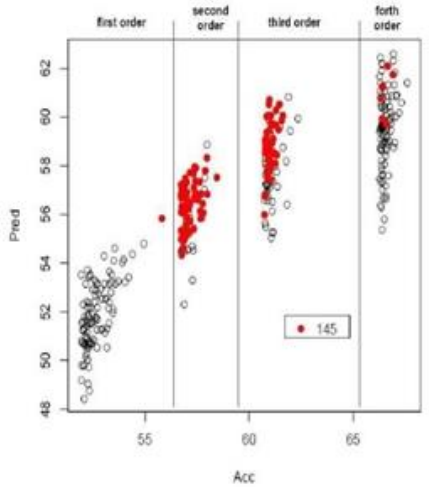
- Follow-up: Model-Based MDR by Calle et al. (2007)

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



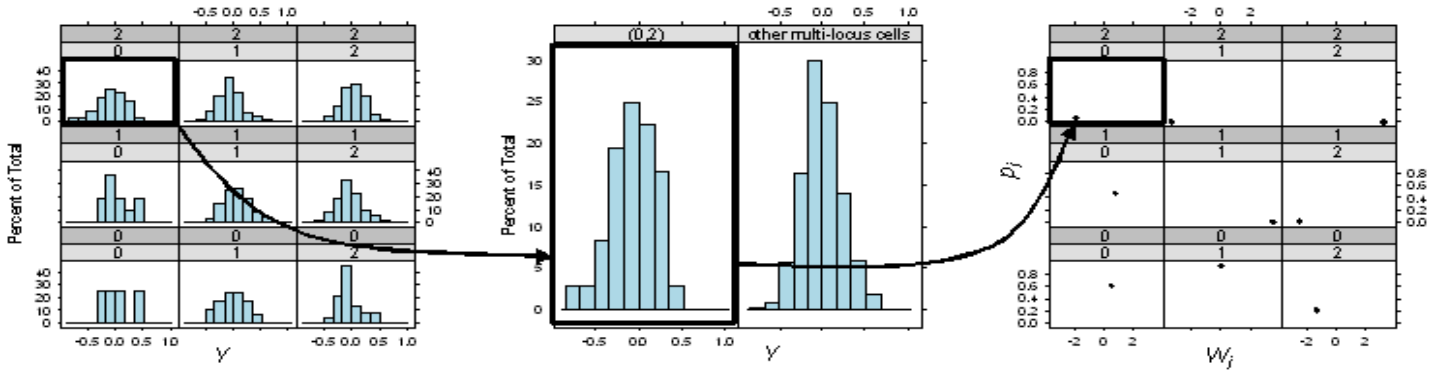
## Historical notes about MB-MDR

- Model-Based MDR by Calle et al. (2008)
  - Computation time is invested in optimal **association tests** to prioritize multilocus genotype combinations (e.g., high, low, no evidence) and in **statistically valid permutation-based methods** to assess joint statistical significance.
  - At the same time, a "quantification" of "interaction" signals can be obtained above and beyond **lower order effects**



# Historical notes about MB-MDR

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




- Stable score tests, one multilocus p-value and permutation-based strategy (Cattaert et al. 2010), rather than Wald tests, and MAF dependent empirical reference distributions (Calle et al. 2008)

# Historical notes about MB-MDR

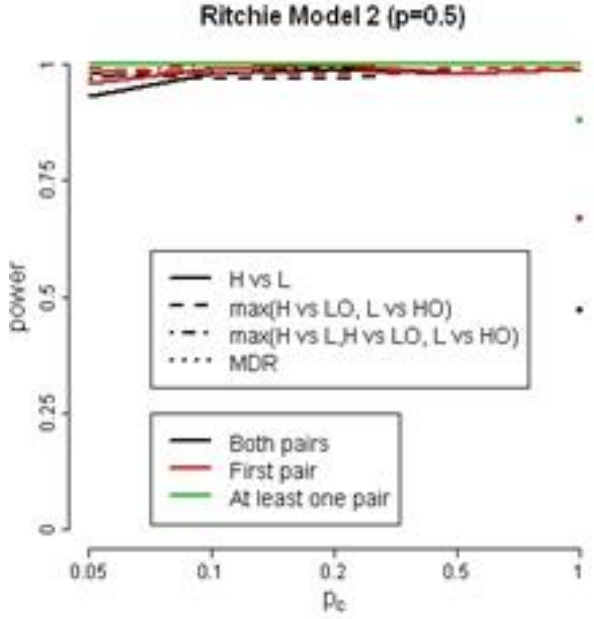
- Model-Based MDR by Cattaert et al (2011) – genetic heterogeneity

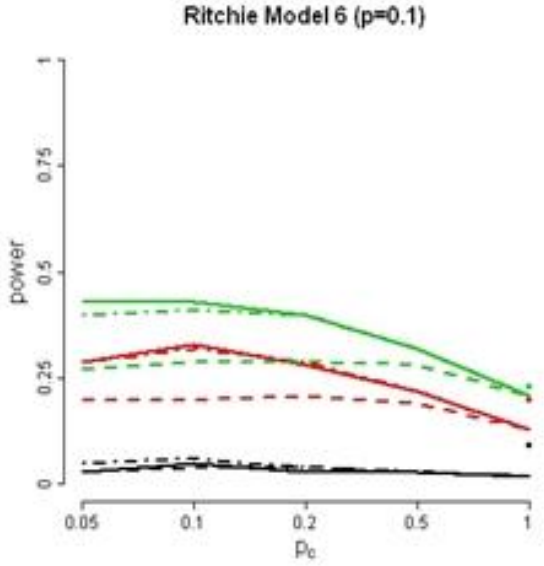
Model 2, $p = 0.5$

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

Model 6, $p = 0.1$

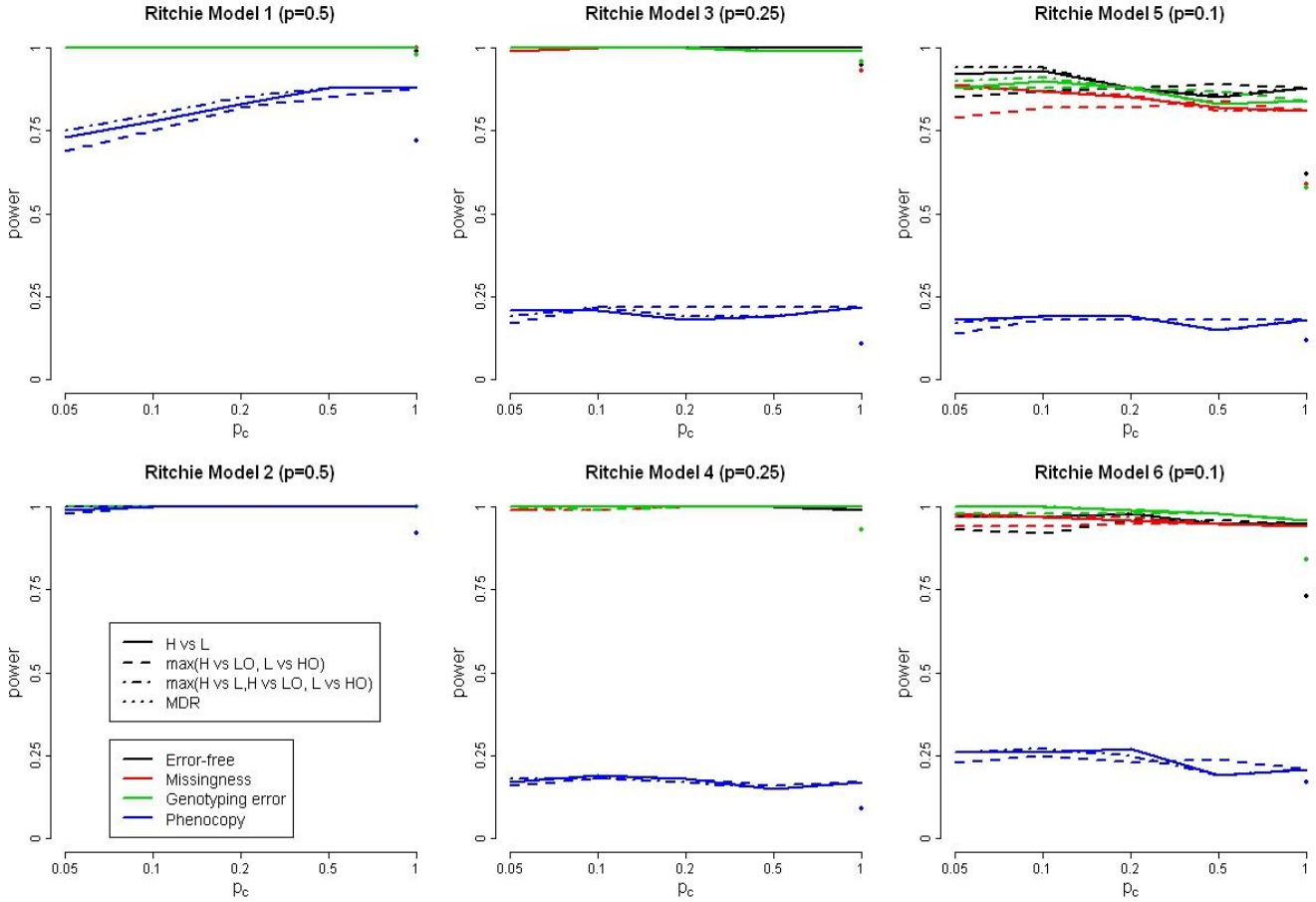
| | BB | Bb | Bb |
|---|---|---|---|
| AA | 0.09 | 0.001 | 0.02 |
| Aa | 0.08 | 0.07 | 0.005 |
| aa | 0.003 | 0.007 | 0.02 |



# Historical notes about MB-MDR

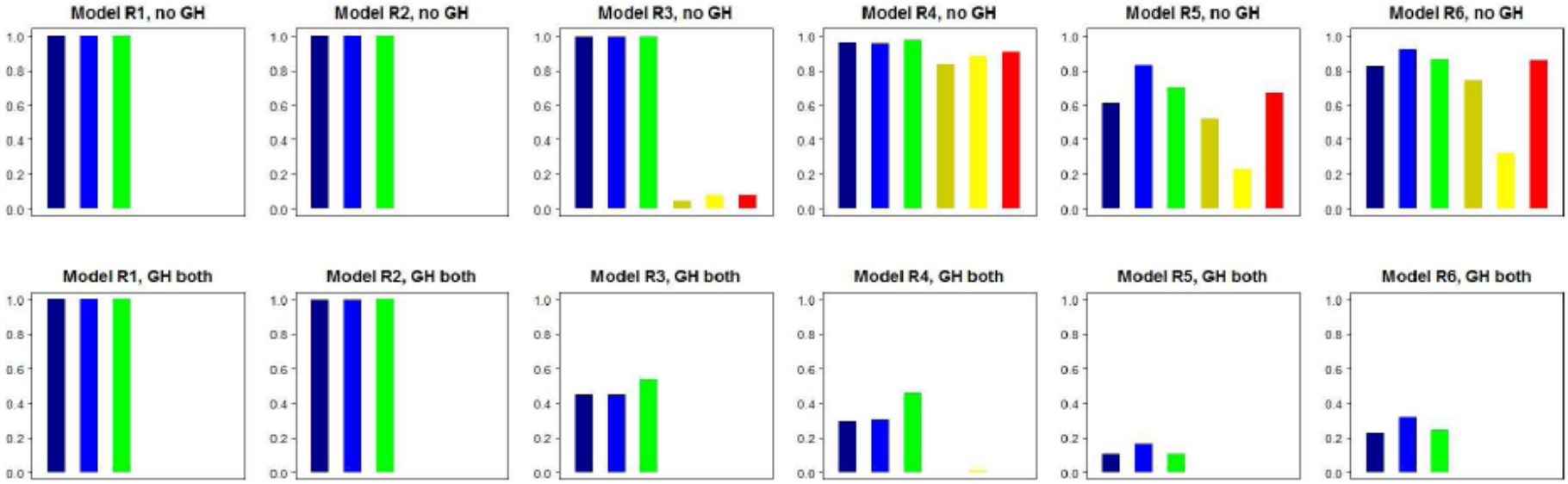
- Model-Based MDR by Cattaert et al (2011) – maximimal power



# Historical notes about MB-MDR

- Power performance

(example: pure epistasis scenario's; unpublished - 2010)




BOOST (dark blue)

EpiCruncher optimal options (light blue)

MB-MDR (green)

PLINK epistasis (dark yellow)

PLINK fast epistasis (light yellow)

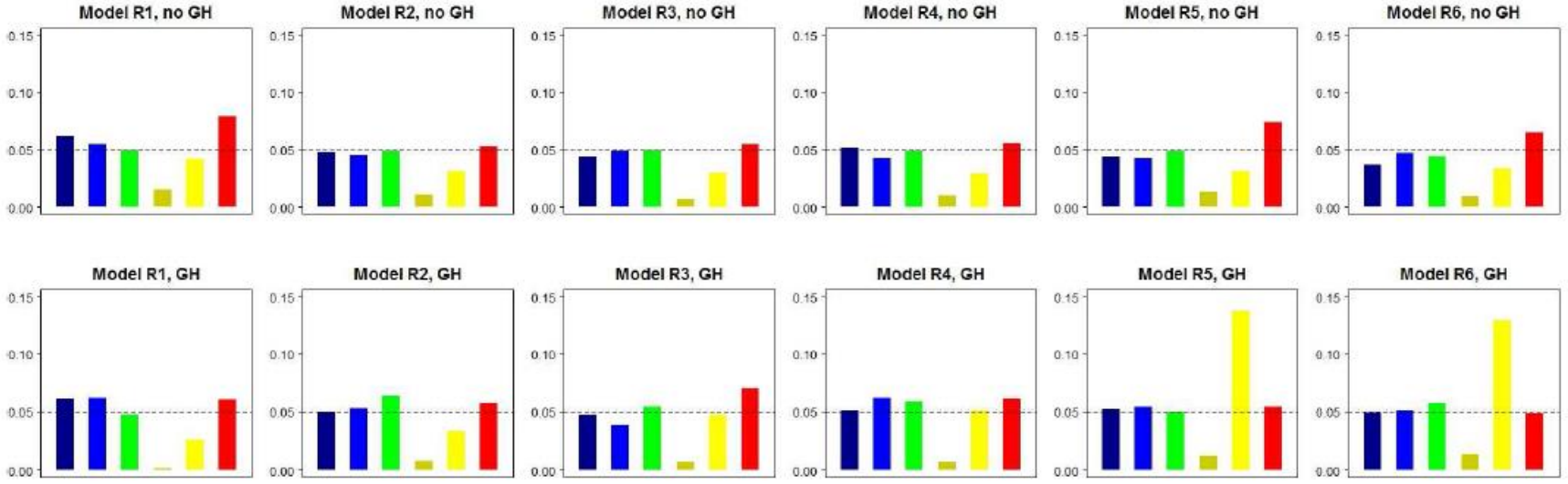
EPIBLASTER (red)

# Historical notes about MB-MDR

- False positives

  (example: pure epistasis scenario's;

  unpublished - 2010)




BOOST (dark blue)

EpiCruncher optimal options (light blue)

MB-MDR (green)

PLINK epistasis (dark yellow)

PLINK fast epistasis (light yellow)

EPIBLASTER (red)

# Nine methodological papers (2010-2013)

- **Calle** ML, Urrea V, Van Steen K (2010) mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits. Bioinformatics Applications Note 26 (17): 2198-2199 **[first MB-MDR software tool]**
- **Cattaert** T, Urrea V, Naj AC, De Lobel L, De Wit V, Fu M, Mahachie John JM, Shen H, Calle ML, Ritchie MD, Edwards T, Van Steen K. (2010) FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals, PLoS One 5 (4). **[first implementation of MB-MDR in C++, with improved features on multiple testing correction and improved association tests + recommendations on handling family-based designs]**
- **Cattaert** T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, Urrea V, Ritchie MD, Van Steen K (2010) Model-Based Multifactor Dimensionality Reduction for detecting epistasis in case-control data in the presence of noise (*invited paper*). Ann Hum Genet. 2011 Jan;75(1):78-89 **[detailed study of C++ MB-MDR performance with binary traits]**
- **Mahachie John** JM, Cattaert T, De Lobel L, Van Lishout F, Empain A, Van Steen K (2011) Comparison of genetic association strategies in the presence of rare alleles. BMC Proceedings, 5(Suppl 9):S32 **[first explorations on C++ MB-MDR applied to rare variants]**

- **Mahachie John** JM, Cattaert T, Van Lishout F, Van Steen K (2011) Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data. European Journal of Human Genetics 19, 696-703. **[detailed study of C++ MB-MDR performance with quantitative traits]**
- **Van Steen** K (2011) Travelling the world of gene-gene interactions *(invited paper).* Brief Bioinform 2012, Jan; 13(1):1-19. **[positioning of MB-MDR in general epistasis context]**
- **Mahachie John** JM , Cattaert T , Van Lishout F , Gusareva ES , Van Steen K (2012) Lower-Order Effects Adjustment in Quantitative Traits Model-Based Multifactor Dimensionality Reduction. PLoS ONE 7(1): e29594. doi:10.1371/journal.pone.0029594 **[recommendations on lower-order effects adjustments]**
- **Mahachie John** JM, Van Lishout F, Gusareva ES, Van Steen K (2012) A Robustness Study of Parametric and Non-parametric Tests in Model-Based Multifactor Dimensionality Reduction for Epistasis Detection. BioData Min. 2013 Apr 25;6(1):9**[recommendations on quantitative trait analysis]**
- **Van Lishout** F, Mahachie John JM, Gusareva ES, Urrea V, Cleynen I, Theâtre E, Charloteaux B, Calle ML, Wehenkel L, Van Steen K (2012) An efficient algorithm to perform multiple testing in epistasis screening. BMC Bioinformatics. 2013 Apr 24;14:138 **[C++ MB-MDR made faster!]**

# The importance of speed

- **Situation in 2012:**

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

Parallel workflow was tested on a cluster composed of 10 blades, containing each four Quad-Core AMD Opteron(tm) Processor 2352 2.1 GHz. Sequential executions were performed on a single core of this cluster. Results prefixed by "≈": extrapolated.

- **Situation in 2014:**

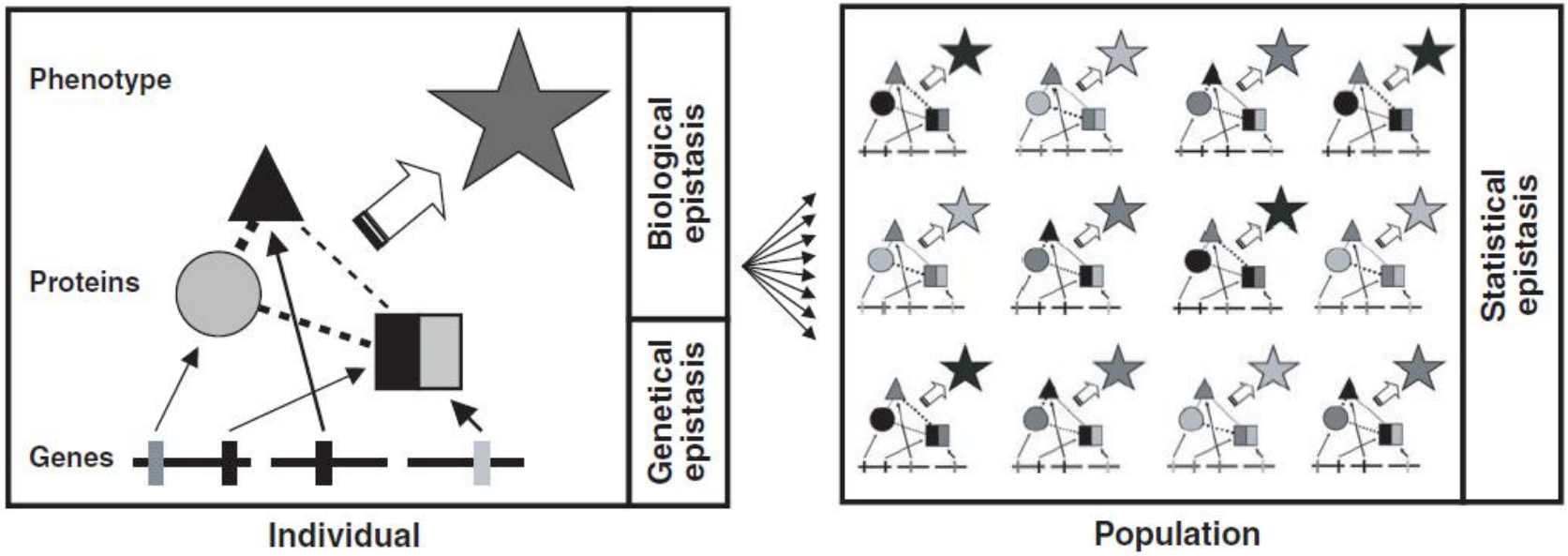
“gamma MaxT”: 10-fold increase

# Up-scaling Interaction Analyses

# “Big Data”

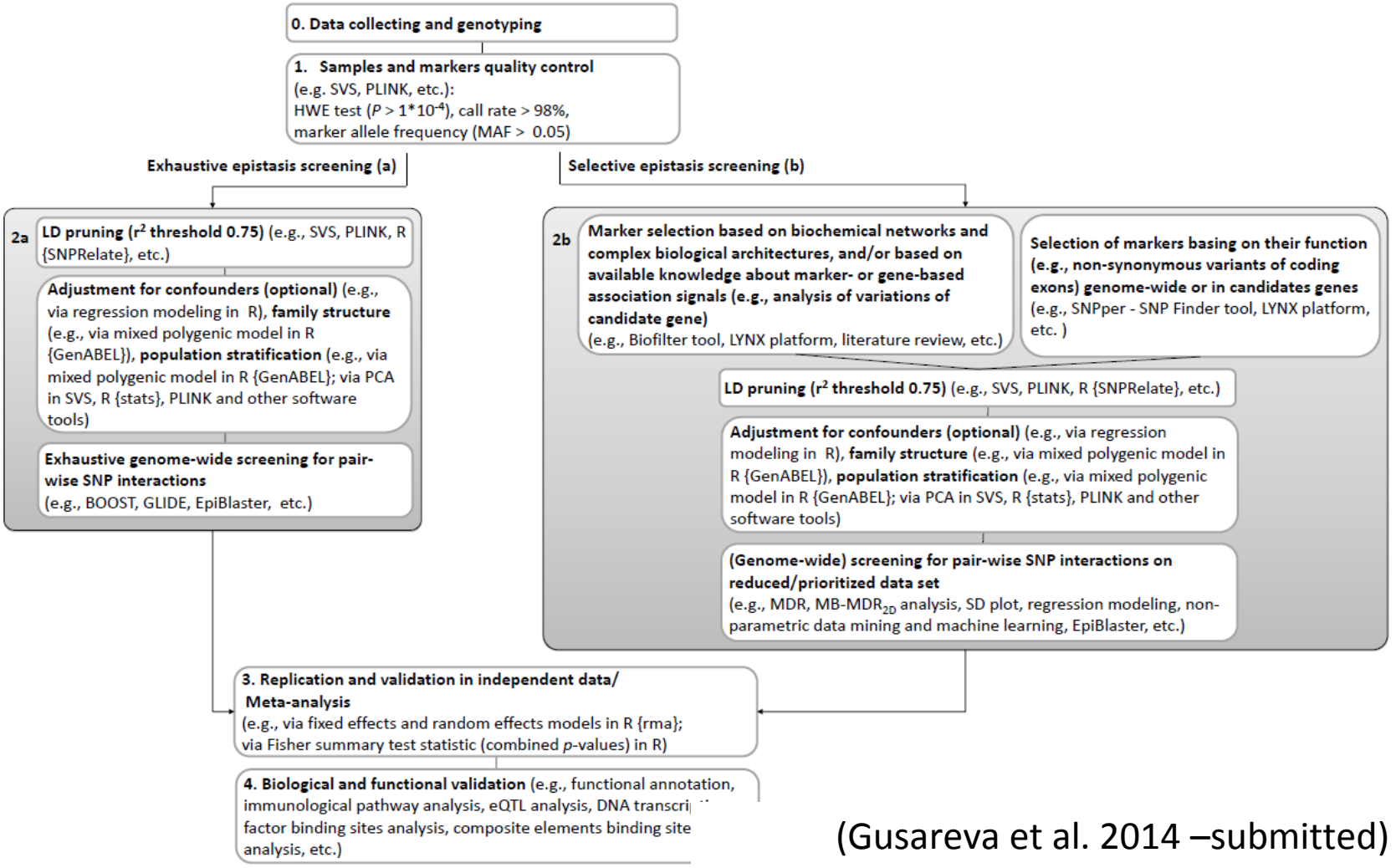
# Bridging the gap between statistics and biology

**Epistasis** … 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)

# **Towards a protocol for GWAI studies** (Gusareva et al. 2012-2013-2014)




(Gusareva et al. 2014 –submitted)

## No replication without a consensus -- about the methodology

- Multiple testing handling – speed (François Van Lishout) √
- Multi-stage designs incl marker selection (Kirill Bessonov) √
- Meta-analysis (Elena Gusareva) √
- LD between markers and long-distance between-marker associations (Jestinah Mahachie)
- Population stratification assessments by –omics (Kridsadakorn Chaichoompu) √
- The importance of epistasis and non-linear relationships in population genetics (Ramouna Fouladi) √
- Within- (Silvia Pineda) and between-gene architectures (K Bessonov) √
- Missing data handling (Kristel Van Steen)

## **No replication without a consensus** -- about the data

- No holy grail but some methods have more desirable properties than others:
  - "Algorithms for detecting epistatic interactions should be evaluated using simulated data, for reasons of both scalability and interpretation"
  - "The creation of realistic structure in simulated data is problematic, due to the complex nature and architecture of epistasis in humans, both of which are largely unknown"

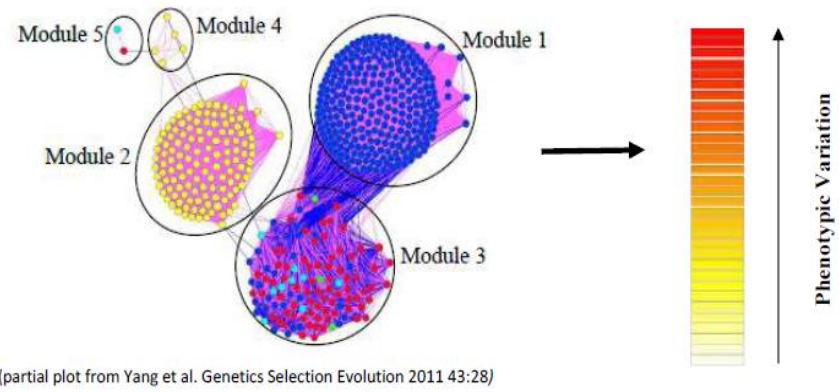
(Goudey et al. 2013)

- There is a need for good realistic reference data! (Develop an ensemble methods that combines best of several methodological worlds)

# The Rare Variant's Perspective

# Integromics as the umbrella framework

- **Thematic objectives:**
  - Develop network-based multi-omics data analysis pipeline using machine learning techniques
  - Identify population substructure using a systems biology integrated view
  - Identify pathways as disease outcome drivers via assessments of biomarker robustness and differential network expressions



(partial plot from Yang et al. Genetics Selection Evolution 2011 43:28)

# Integrated statistical epistasis networks

- *Genomic* MB-MDR naturally leads to *integrated* (statistical) interaction networks

  (nodes = regions of interest, to which features from different omics data types can be mapped; edges = defined by MB-MDR test results)

Pac Symp Biocomput. 2013:397-408.

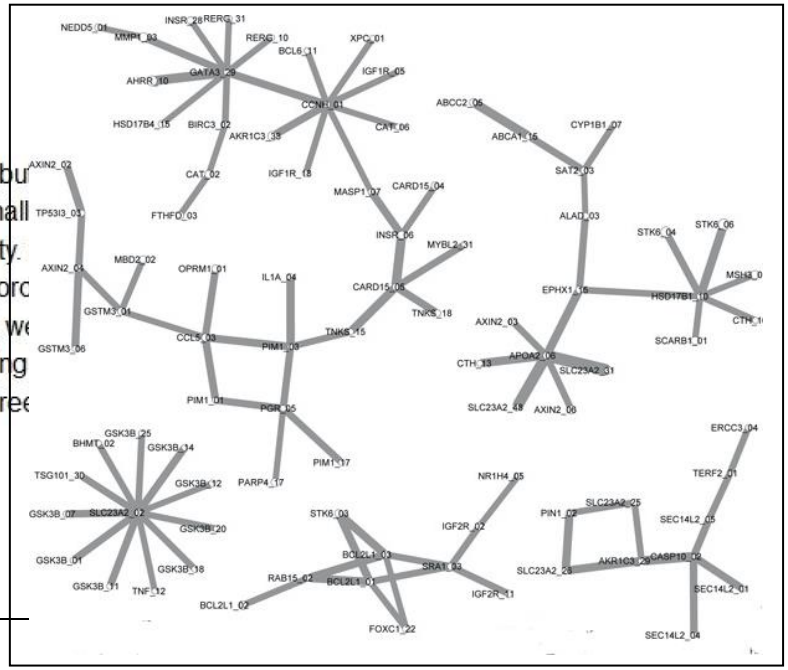
**Statistical epistasis networks reduce the computational complexity of searching three-locus genetic models.**

Hu T, Andrew AS, Karagas MR, Moore JH.

**Author information**

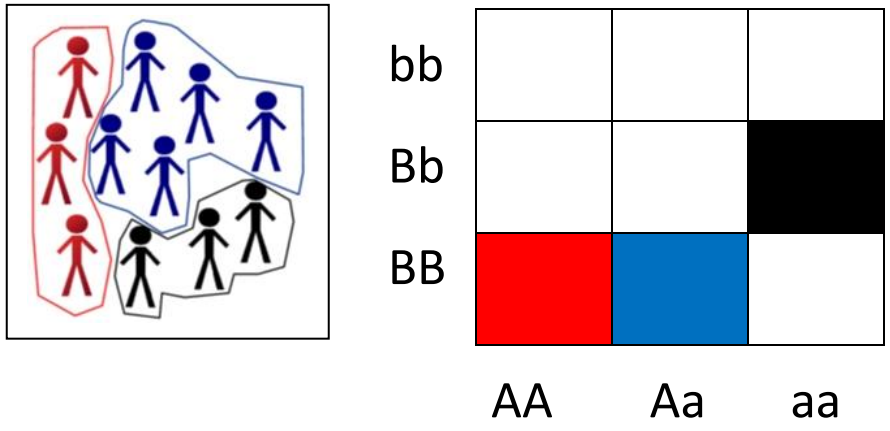
**Abstract**

The rapid development of sequencing technologies makes thousands to millions of genetic attribu
traits. Searching this enormous high-dimensional data space imposes a great computational chall
network-based approach to supervise the search for three-locus models of disease susceptibility.
strong pairwise epistatic interactions and provide a global interaction map to search for higher-orc
together in the networks. Applying this approach to a population-based bladder cancer dataset, we
variations in DNA repair and immune regulation pathways, which holds great potential for studying
validations. We demonstrate that our SEN-supervised search is able to find a small subset of three
substantially reduced computational cost.



# From MB-MDR to *genomic* MB-MDR

- Individuals may be similar wrt 2-locus genotypes: AABB (red), AaBB(blue), aaBb (black), …
- Individuals may be similar wrt features projectable to a gene




- Components of a gene: SNPs, rare variants, epigenetic markers, RNA expression → inter-related features

## *Genomic* MB-MDR 1D

- MB-MDR 1D has poor performance due to excellent power with "additive" coding and the non-parametric nature of MB-MDR (dosage coding)
- It is worthwhile to re-assess the power of genomic MB-MDR 1D due to the complexity of the "mode of inheritance" related to a gene as a whole (using different sources of information for that gene)
- When mapping common and rare variants to a gene or genomic region of interest, genomic MB-MDR leads to a novel association testing framework for sequence data

# Rare variant association testing

- Increase statistical power by aggregating single rare variants (RVs) into meaningful groups (e.g., defined by genes, pathways, or functionality – "Regions of Interest")
  - Collapse based on summary statistics → collapsing tests (e.g., li and Leal 2008 - CMC)
  - Use collapsed constructs in a regression framework (e.g., Lasso – Zhou et al. 2010 )
- Look for similarities between individuals based on their sequence data
  - Use multi-marker test by combining single-variant stats (Wu et al. 2011 - SKAT: ideas from kernel theory and regression)

# In search for desirable properties for a novel approach

- Adjust for confounders √
- Inherent flexible to a variety of trait types √
- Non-parametric in that no assumptions are made about the direction of individual SNP effects √
- Accommodate genotype uncertainty
- Incorporate biological information √
- Robust to LD to control false positive results (although relatively limited LD between rare variants; likely to have occurred fairly recently in the ancestry of the population)

# An integrated framework based on MB-MDR

## Step 1: Descriptor filtering

- At the end of this step, only descriptors that have "acceptable" representation and correlation with other descriptors are kept in the data

(don't throw away rare variants)

## Step 2: Choice of clustering approach

- Flexible integration of heterogeneous data (scaling)
- Flexible integration of data interdependencies
- Allowing low-variance descriptors

(Scalability, significant number of clusters from large-scale data)

## Step 3: Application of "classic" MB-MDR

GENOMIC

MB-MDR

# Prototype development - clustering

- Similarity (Liu et al. 2011 – inverse prob weighted clustering for RV/LFV/CV assoc. analysis)

| Individual 2 | Individual 1 | | |
|---|---|---|---|
| | ***aa*** | ***aA*** | ***AA*** |
| *aa* | $\frac{2}{p_a^2}$ | $\frac{1}{p_a^2}-\frac{1}{2p_a(1-p_a)}$ | $-\frac{1}{p_a(1-p_a)}$ |
| *aA* | $\frac{1}{p_a^2}-\frac{1}{2p_a(1-p_a)}$ | $\frac{1}{2}\left\{\left[\frac{1}{p_a^2}+\frac{1}{(1-p_a)^2}\right]-\frac{1}{p_a(1-p_a)}\right\}$ | $\frac{1}{(1-p_a)^2}-\frac{1}{p_a(1-p_a)}$ |
| *AA* | $-\frac{1}{p_a(1-p_a)}$ | $\frac{1}{(1-p_a)^2}-\frac{1}{p_a(1-p_a)}$ | $\frac{2}{(1-p_a)^2}$ |

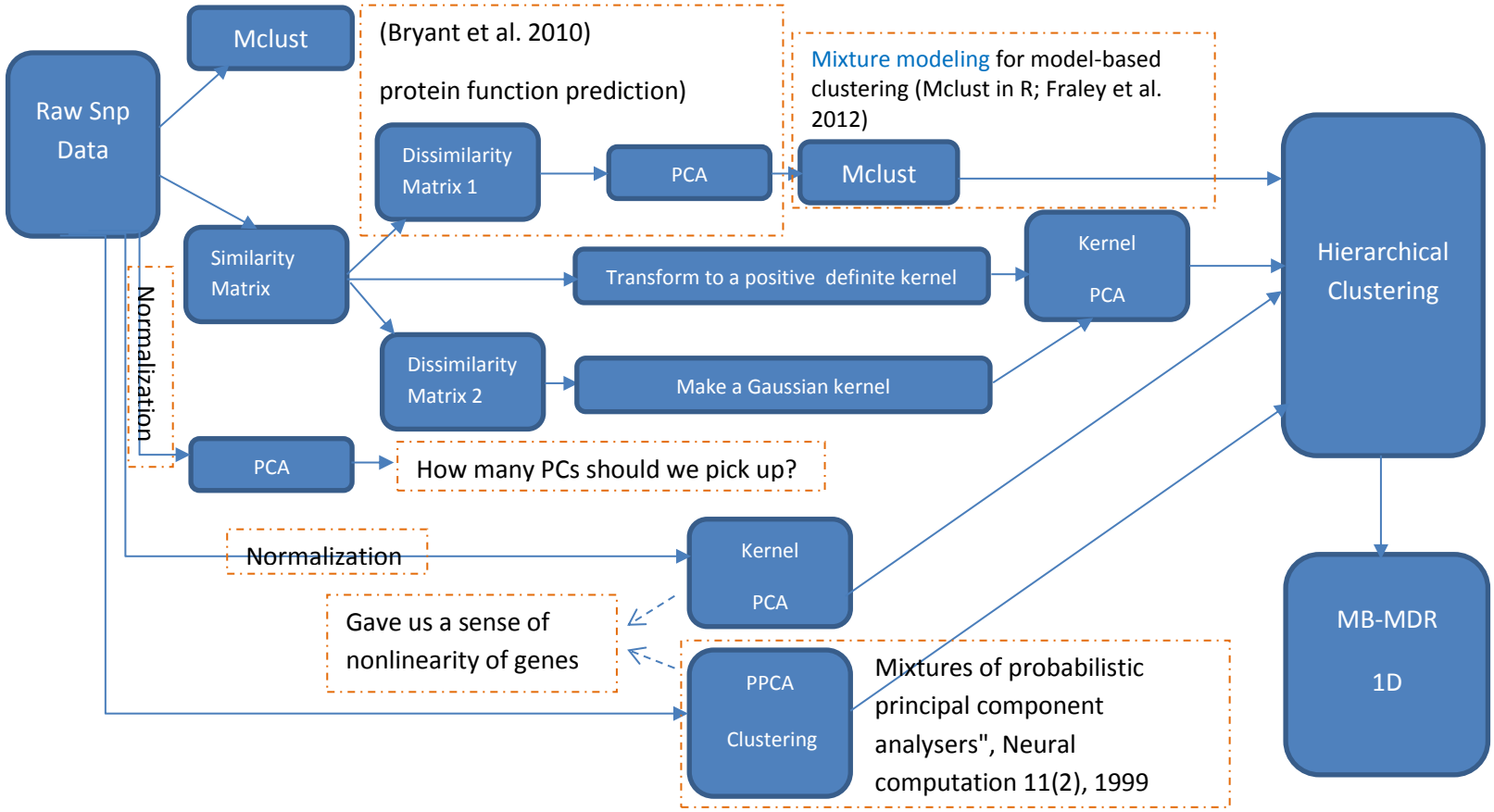
$p_a$ is the population frequency of minor allele $a$.

- Distance between individuals *i* and *j*:

$$d(i,j) = e^{-\beta\ \mathrm{sim}(i,j)}, \qquad \mathrm{sim}(i,j) = \sum_{k\ \in gene} \mathrm{sim}(i,j;k)$$

- PCA on distance features (Bryant et al. 2010 – protein function prediction)
- Mixture modeling for model-based clustering (Mclust in R; Fraley et al. 2012)

# Advanced development – clustering (steps 1 + 2)

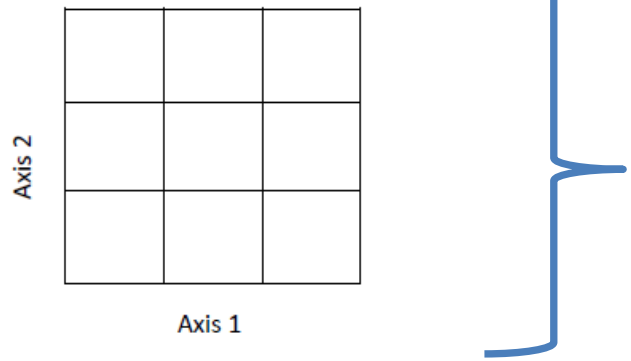


# Prototype development – genomic MB-MDR

– Importance of interpretability

> "If we identify a bird's species from its bodily shape, that predicts many other attributes: its coloration, its song, when it mates, whether and where it migrates, what it eats, its genome, etc. Bird species, then, is a good cluster" (http://www.stat.cmu.edu/~cshalizi/)

– Self-criticism: rethink the default options of MB-MDR (different contexts!)




1. SNP-based MB-MDR 2D : √
2. SNP-based MB-MDR 1D : ?

↓

2. Omics-based MB-MDR 1D
1. Omics-based MB-MDR 2D

# GAW17 mini-exome data

- Includes 200 replicates comprising 697 unrelated individuals (Kazma and Bailey 2011)
- The data provided by the GAW17 include a subset of genes grouped according to pathways that had sequence data available in the 1000 Genomes project
- All simulated singular SNP effects are assumed to be additive on the quantitative trait scale, such that each copy of the minor allele increases or decreases the mean trait value by an equal amount → any epistastic effect detected is a false positive
- Several traits were simulated

# GAW17 mini-exome data

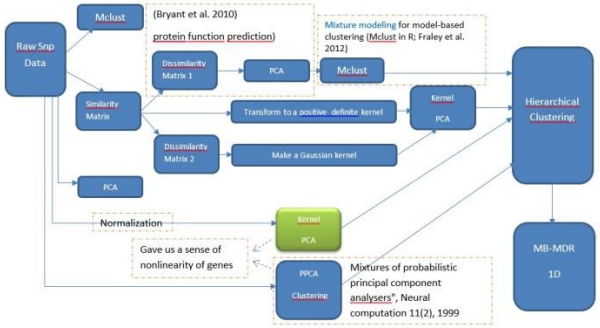
- Our analysis involved the quantitative trait Q1, which was simulated as a normally distributed phenotype.
- The SNPs with the largest effect size were located in the FLT1 gene ($h^2$ = 0.152, 0.083, 0.037) and the KDR gene ($h^2$ = 0.031, 0.031, 0.025) (Bailey-Wilson et al. 2011)
  - Restricted attention to the available single nucleotide polymorphisms (SNPs) on chromosome 4 (944 in total) in 81 genes.
- Note: of the 944 markers, 199 have MAF > %1 and 745 have MAF < %1 (79%)

# Genomic MB-MDR results

Selection probability (200 replicates) for MB-MDR 1D for 81 genes on chromosome 4

## 1D (gene-based association)

| KDR | All other genes |
|---|---|
| 166 times out of 200 | 87 times out of 200x80 (87/80 ~1.08 out of 200 per gene) |



# Genomic MB-MDR results

Selection probability (200 replicates) for MB-MDR 1D for 81 genes on chromosome 4

## 1D (gene-based association)

| KDR | All other genes |
|---|---|
| 166 times out of 200 | 87 times out of 200x80<br>(87/80 ~1 out of 200 per gene) |

## SKAT

| KDR | All other genes |
|---|---|
| 130 times out of 200 | 396 times out of 200x80<br>(396/80 ~5 out of 200 per gene) |

# Integration to enhance

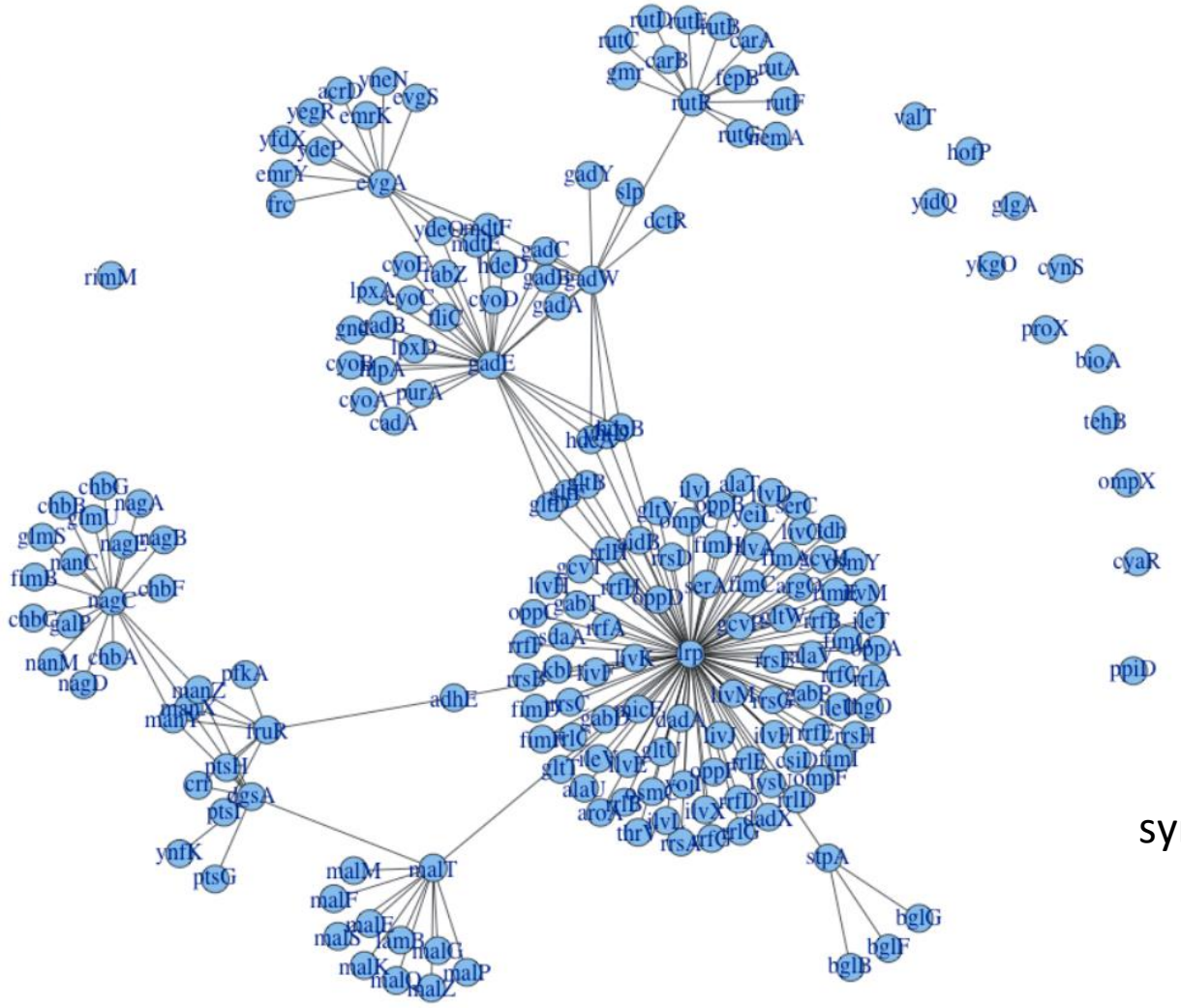
# Biological Network Construction

## Integrated statistical epistasis networks

- Statistical epistasis networks reduce the computational complexity of searching for higher-order interactions / multi-locus models
- They provide a natural framework to associate network properties or network "modules" to variations in the "outcome" (built in the edge def.)
- MB-MDR outcomes:
  - Disease related traits:
    - Censored
    - Quantitative (progression measurements)
    - Binary (disease status)
    - Multivariate (system groups)
  - Population types

# Integrated networks with machine learning

200 nodes
212 edges

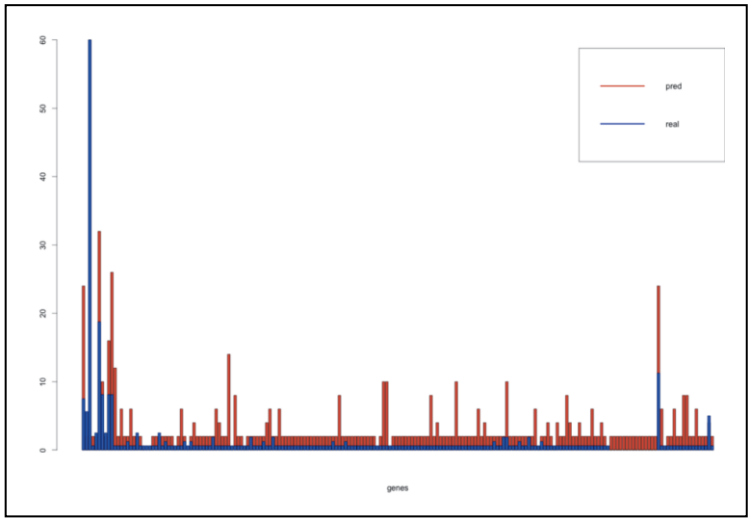


(GeneNetWeaver synthetic **gold standard** network based on transcription factor network (TFN) of E.coli)

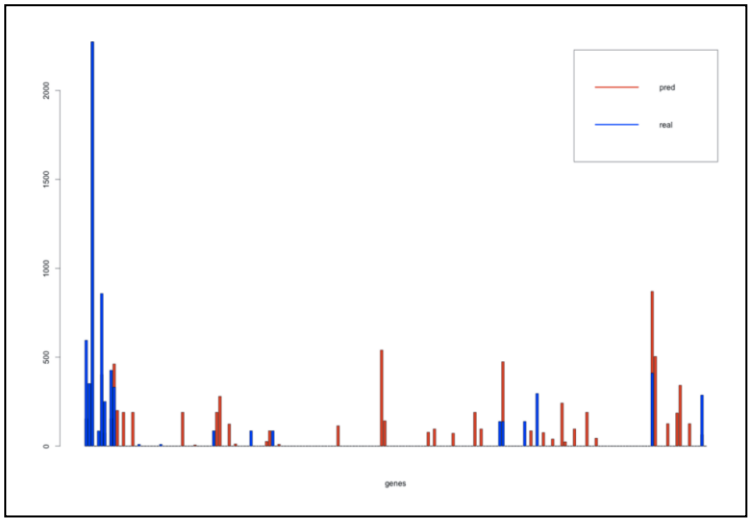
**“Regression2Net”** (Francesco Gadaleta)

- Uses penalized regression to identify interesting variables (but is flexible to accommodate other variable selection methods)
- Defines edges when specific stability criteria are met

Color legend: Predicted network ~ Gold Standard




Degree correlation = 0.86

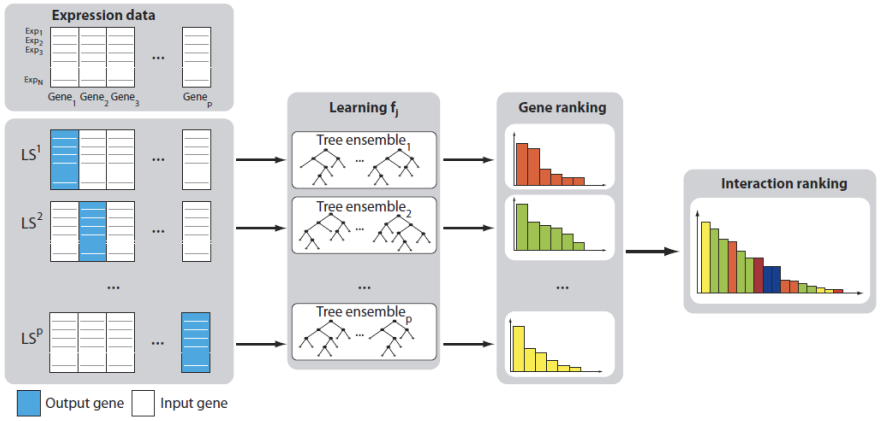



Betweenness correlation = 0.83

**Foresting in Integromics Inference** (Kirill Bessonov)

- Builds on the GENIE suite of Vân Anh et al. but uses "conditional inference trees/forests" (CIFs) instead of "random forests" with key performance differences

ctree uses a significance test procedure in order to select variables instead of selecting the variable that maximizes an information measure (e.g. Gini)




- Allows flexible integration of multiple features associated to a genomic region of interest
- Performance: Regression2Net > Conditional Inference Tree > GENIE3

# In conclusion

# Food for thought – ED GE

- Despite some critics, not all resources have been exploited to enhance Epistasis Detection (ED):
  - Reference data + Comparison of methods and creation of ensemble methods (cfr. ensemble clustering)
  - Assessment of the contribution of epistasis to human complex disease traits (including gene expression traits) [heritability, variation, prediction]
  - Assessment of epistasis influence on "evolution" (what characterizes population strata)
  - Awareness papers and lobbying (cfr. epistasis as conference topic)
  - Network construction involving E and dynamics ("adapt" statistical epistasis methods)

## Food for thought – BIO

- Improving GWAs power (T Park) / GWAIs power (K VanSteen)
  - Meta-analysis
  - Analysis of multiple SNPs
    - Regularized Regression (e.g., Elastic-Net, penalized regression - Ayers and Cordell 2013)
    - Interaction analysis (e.g., MB-MDR gammaMaxT)
    - Gene Set Analysis
  - Multivariate analysis
- Improving identification of disease mechanisms
  - NGS: apply genomic MB-MDR
  - Integromics: Fuse SNP-SNP epistasis networks with RNA networks

# Acknowledgement

## Systems and Modeling Unit, Montefiore Institute, University of Liège, Belgium

Machine Learning for Personalized Medicine

LA LIBERTÉ DE CHERCHER

EU PANCREAS

## Systems Biology and Chemical Biology Thematic Research Unit, GIGA-R, Liège,


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