

Methodological Aspects in Integromics

Kristel Van Steen, PhD² (*)

kristel.vansteen@ulg.ac.be

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

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

Outline

- **Integromics**

- Definition and motivation
- Building blocks / Bottom up versus top down?

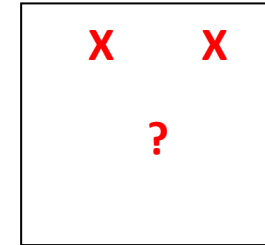
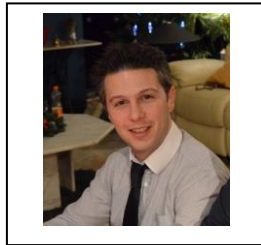
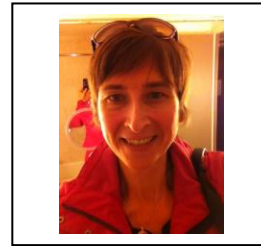
- **A novel integrated analysis framework based on dim. reduction**

- How does it work?
- Issues
- Simulation results

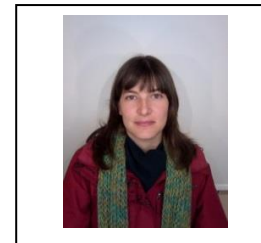
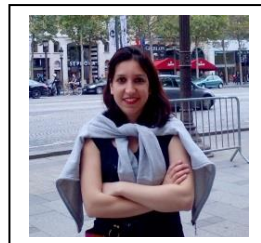
- **Bonuses of the novel framework “genomic MB-MDR”**

- Gene-based testing
- From SNP x SNP to Gene x Gene Interactions
- Integrated networks

- **In Conclusion**



Bio³: **Bi**ostatistics – **Bi**omedicine - **Bi**oinformatics





Groupe Interdisciplinaire de
Génoprotéomique Appliquée



Systems Biology and Chemical Biology

- Laboratory of molecular engineering and genetic engineering
- Laboratory of histology and mammalian cell culture
- Laboratory of mass spectrometry
- **Research unit of systems and modelling**
 - Algorithms and stochastic methods
 - Computational systems biology
 - Bioinformatics – Statistical Genetics

Integromics

Data integration: Definition

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 21 2009, pages 2855–2856
doi:10.1093/bioinformatics/btp515

Systems biology

integrOmics: an R package to unravel relationships between two omics datasets

Kim-Anh Lê Cao^{1,*}, Ignacio González² and Sébastien Déjean³

¹Institute for Molecular Biosciences and ARC Centre of Excellence in Bioinformatics, The University of Queensland, Brisbane QLD 4072, Australia, ²Plateforme Biopuces, Genopôle Toulouse Midi-Pyrénées, Institut National des Sciences Appliquées, F-31077 and ³Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse et CNRS, F-31062, France

- Joint analysis
- Challenging statistics – Regularized - Generalized
- Integrating different types of variables

What's in a name?

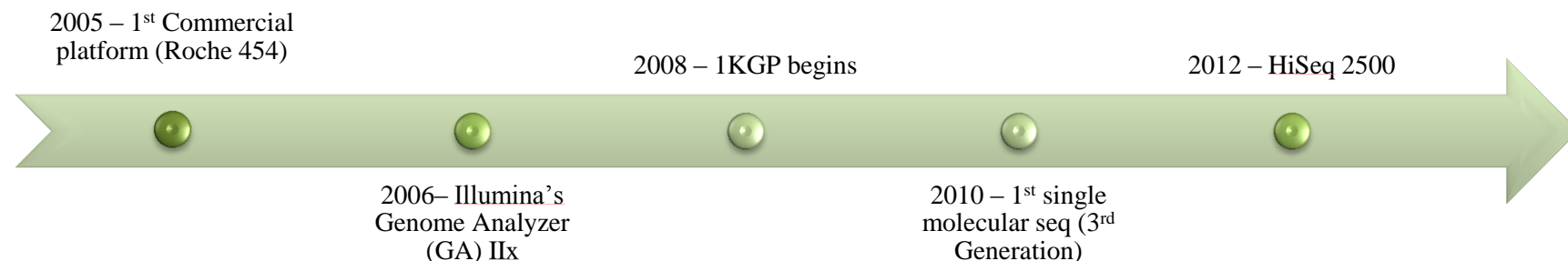
- **Data fusion** refers to fusing records on the same entity into a single file, and involves putting measures in place to detect and remove erroneous or conflicting data (Wang et al., 2014).
- Some definitions for “data fusion” use “data integration” in their definition. Although some data integration 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.

Multidisciplinary, interdisciplinary, transdisciplinary research

- An **omics multidisciplinary approach** divides the initial problem in data-specific sub-problems
 - disperse pieces of information are combined or integrated in a limited way /later stage in the study
- **Interdisciplinary efforts** adopt discipline-specific perspectives in a joint effort to solve a common problem
- A **trans-disciplinary approach** involves an active synergy between disciplines, to create a solution to the problem that otherwise could not have been found.
 - requires cross-talk between disciplines and a unified language that is accessible to all parties involved (Fawcett, 2013; Woods, 2007)

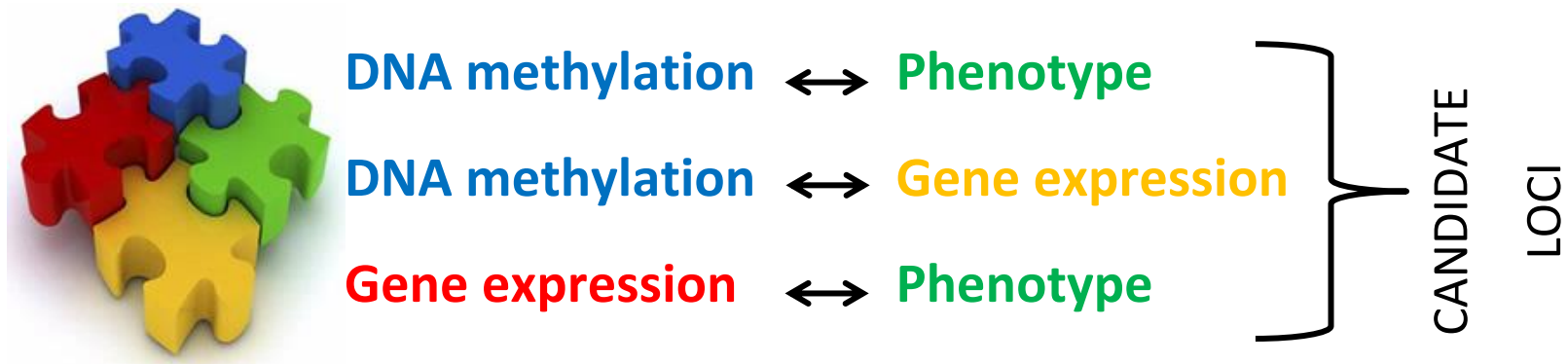
Data integration: Motivation and Opportunity

- The identification of causal or predictive variants/genes/mechanisms for disease-associated traits is characterized by “complex” networks of molecular phenotypes.
- Present technology and computer power allow building and processing large collections of these data types → Next Generation Sequencing.



Is there room for data integration?

- Observation 1: the super-rapid data generation is counterweighted by a slow-pace for data integration methods development.
- Observation 2: Most currently available integrative analytic tools pertain to pairing omics data and focus on between-data source relationships, making strong assumptions about within-data source architectures.



Is there room for data integration?

- A limited number of initiatives exist aiming to find the most optimal ways to analyze multiple, possibly related, omics data bases, while fully acknowledging the specific characteristics of each data type.

Is there room for data integration?

- Reasons?
 - There is an advantage in out-of-the-box thinking
 - Integrative methodologies have been developed in different sciences (e.g., computer science, engineering)
 - It is essential to thoroughly understand underlying assumptions of integrative methods in order to draw sound conclusions
 - Helps in minimizing the gap between bio and theoretical model

Is there room for data integration?

Perspectives on Data Integration in Human Complex Disease Analysis

Kristel Van Steen^{1,2*} and Nuria Malats³, on behalf of the COST Action BM1204 participants⁴.

¹ *Systems and Modeling Unit, Montefiore Institute, University of Liège, Liège, Belgium*

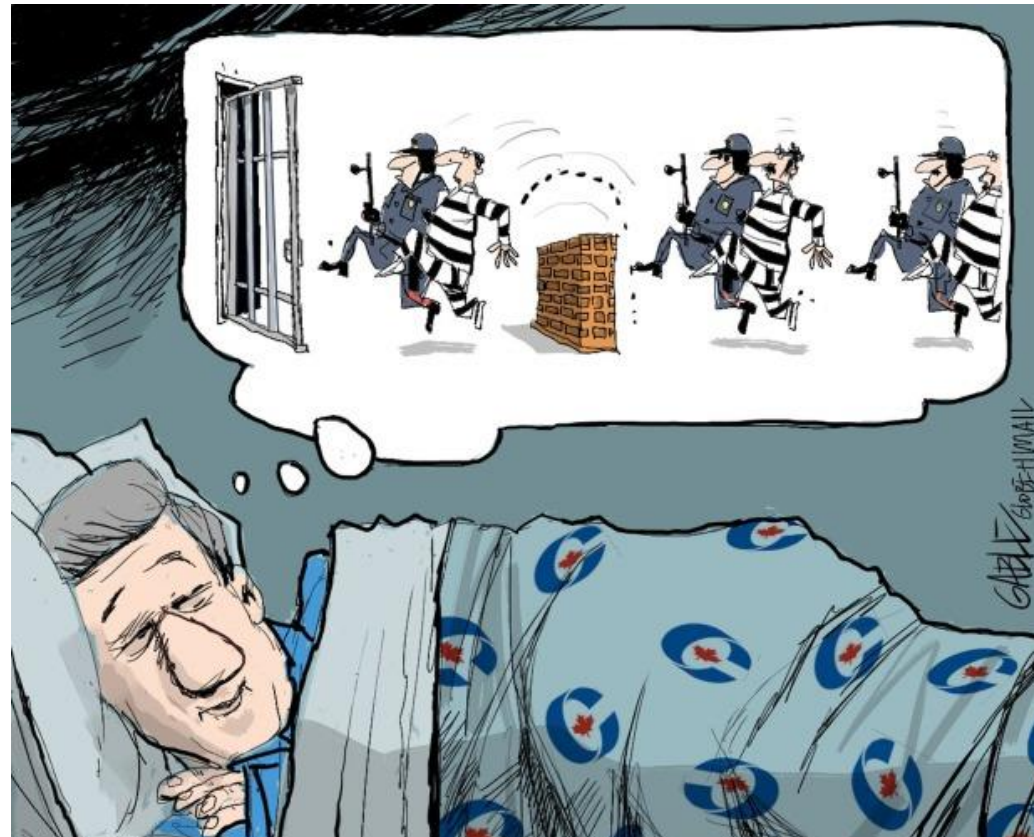
² *Bioinformatics and Modeling, GIGA-R, University of Liege, Avenue de l'Hôpital 1, Liège, Belgium*

³ *Genetic & Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain*

⁴ http://www.cost.eu/domains_actions/bmbs/Actions/BM1204

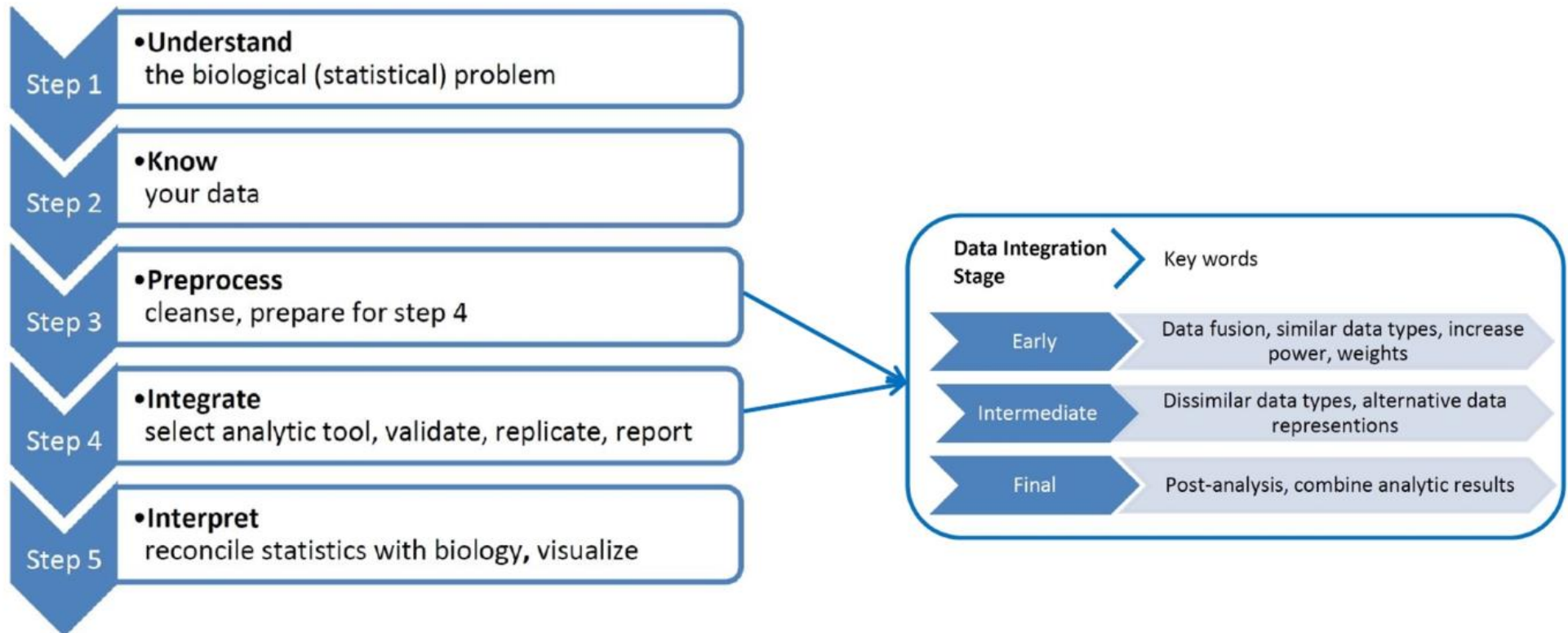
(Book chapter in “Big Data Analytics in Bioinformatics and Healthcare”, 2014 - accepted)

So we have the **motive**, and the **opportunity** ...



(Boston Globe)

Building blocks of a “data integration” pipeline



Building blocks of a “data integration” pipeline

Step 1

- Formulating the biological (statistical) problem

- Traditional biological research questions are for the most part **hypothesis-driven**: performing experiments to answer specific biological hypotheses
- In modern genomics, it is increasingly accepted to generate data in a relatively **hypothesis-free** setting: different questions can be formulated on the pool of data; data are mined with a variety of computational and statistical tools

Systems information by integration (Joyce and Palsson 2006)

Genomics	Transcriptomics	Proteomics	Metabolomics	Protein-DNA interactions	Protein-protein interactions	Fluxomics	Phenomics
Genomics (sequence annotation)	<ul style="list-style-type: none"> ORF validation Regulatory element identification⁷⁴ 	<ul style="list-style-type: none"> SNP effect on protein activity or abundance 	<ul style="list-style-type: none"> Enzyme annotation 	<ul style="list-style-type: none"> Binding-site identification⁷⁵ 	<ul style="list-style-type: none"> Functional annotation⁷⁹ 	<ul style="list-style-type: none"> Functional annotation 	<ul style="list-style-type: none"> Functional annotation^{71,103} Biomarkers¹²⁵
	Transcriptomics (microarray, SAGE)	<ul style="list-style-type: none"> Protein: transcript correlation²⁰ 	<ul style="list-style-type: none"> Enzyme annotation¹⁰⁹ 	<ul style="list-style-type: none"> Gene-regulatory networks⁷⁶ 	<ul style="list-style-type: none"> Functional annotation⁸⁹ Protein complex identification⁸² 		<ul style="list-style-type: none"> Functional annotation¹⁰²
		Proteomics (abundance, post-translational modification)	<ul style="list-style-type: none"> Enzyme annotation⁹⁹ 	<ul style="list-style-type: none"> Regulatory complex identification 	<ul style="list-style-type: none"> Differential complex formation 	<ul style="list-style-type: none"> Enzyme capacity 	<ul style="list-style-type: none"> Functional annotation
			Metabolomics (metabolite abundance)	<ul style="list-style-type: none"> Metabolic-transcriptional response 		<ul style="list-style-type: none"> Metabolic pathway bottlenecks 	<ul style="list-style-type: none"> Metabolic flexibility Metabolic engineering¹⁰⁹
				Protein-DNA interactions (ChIP-chip)	<ul style="list-style-type: none"> Signalling cascades^{89,102} 		<ul style="list-style-type: none"> Dynamic network responses⁸⁴
					Protein-protein interactions (yeast 2H, coAP-MS)		<ul style="list-style-type: none"> Pathway identification activity⁸⁹
						Fluxomics (isotopic tracing)	<ul style="list-style-type: none"> Metabolic engineering
							Phenomics (phenotype arrays, RNAi screens, synthetic lethals)

(Joyce and Palsson 2006)

Building blocks of a “data integration” pipeline

Step 2

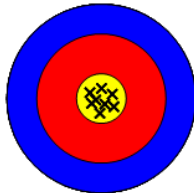
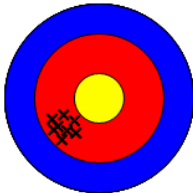
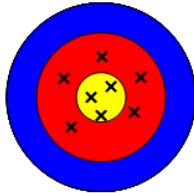
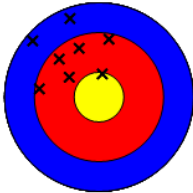
- Identifying the (characteristics of the) data types

- Current data integration methods fall into two different categories:
 - **similar data types** (across studies) or
 - **heterogeneous data types** (across studies as well as within studies).
- Heterogeneous: if two or more fundamentally different data sources are involved.

Step 2

- Identifying the (characteristics of the) data types

- Data characterization (in my opinion) refers to finding first evidences for
 - intrinsic properties (e.g., small sample sizes, standard formats)
 - layers of information; hierarchies; dimensionality
 - noise patterns (related to technology, platform, the lab; systematic and random errors)

	Accurate	Inaccurate (systematic error)
Precise		
Imprecise (reproducibility error)		

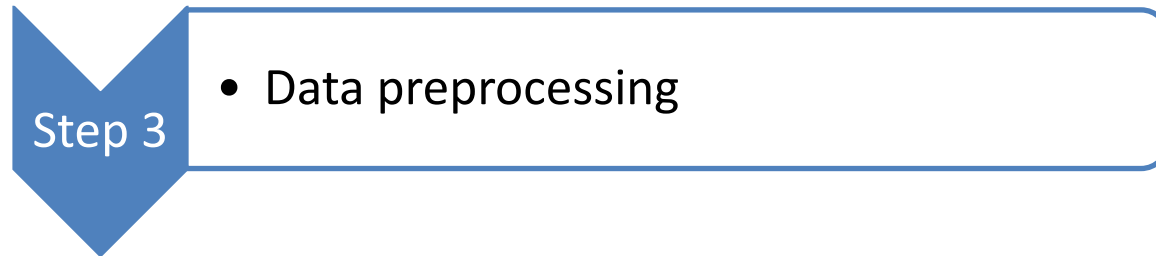
(<http://saturn.cis.rit.edu/>)

Step 2

- Identifying the (characteristics of the) data types

- Whether data are of similar or heterogeneous type, the issue of **quality** (each data sources is unavoidably subject to different levels of noise) and **informativity** is of great importance.
- Therefore, the concept of **weighting the data sources** with **quality and/or informativity scores** becomes an essential component of the framework?
- Step 2 to data integration is as important as a classical Exploratory Data Analysis (EDA) in statistical inference practice.

Building blocks of a “data integration” pipeline



- Approaches for preprocessing vary depending on the type and nature of data:
 - e.g., arrays: background correction, normalization, quality assessment, which may differ from one platform to another
- Data (pre)processing can be done **at any step of the data integration process**:
 - e.g., at the **initial stage**
 - e.g., **prior to statistical analysis** (related to model assumptions)

Building blocks of a “data integration” pipeline

Step 5

- Interpretation (after integrative analytics)

- Is about “understanding” the problem that was initially posed.
- Involves post-linking to several external biological data bases
- Interpretation often involves functional explanation (as part of functional genomics)
- There is a huge challenge in visualizing the steps of and the results from an integrated analysis: **visual analytics**
- (Experimental) validation helps in the “understanding”, but becomes cumbersome in integromics settings

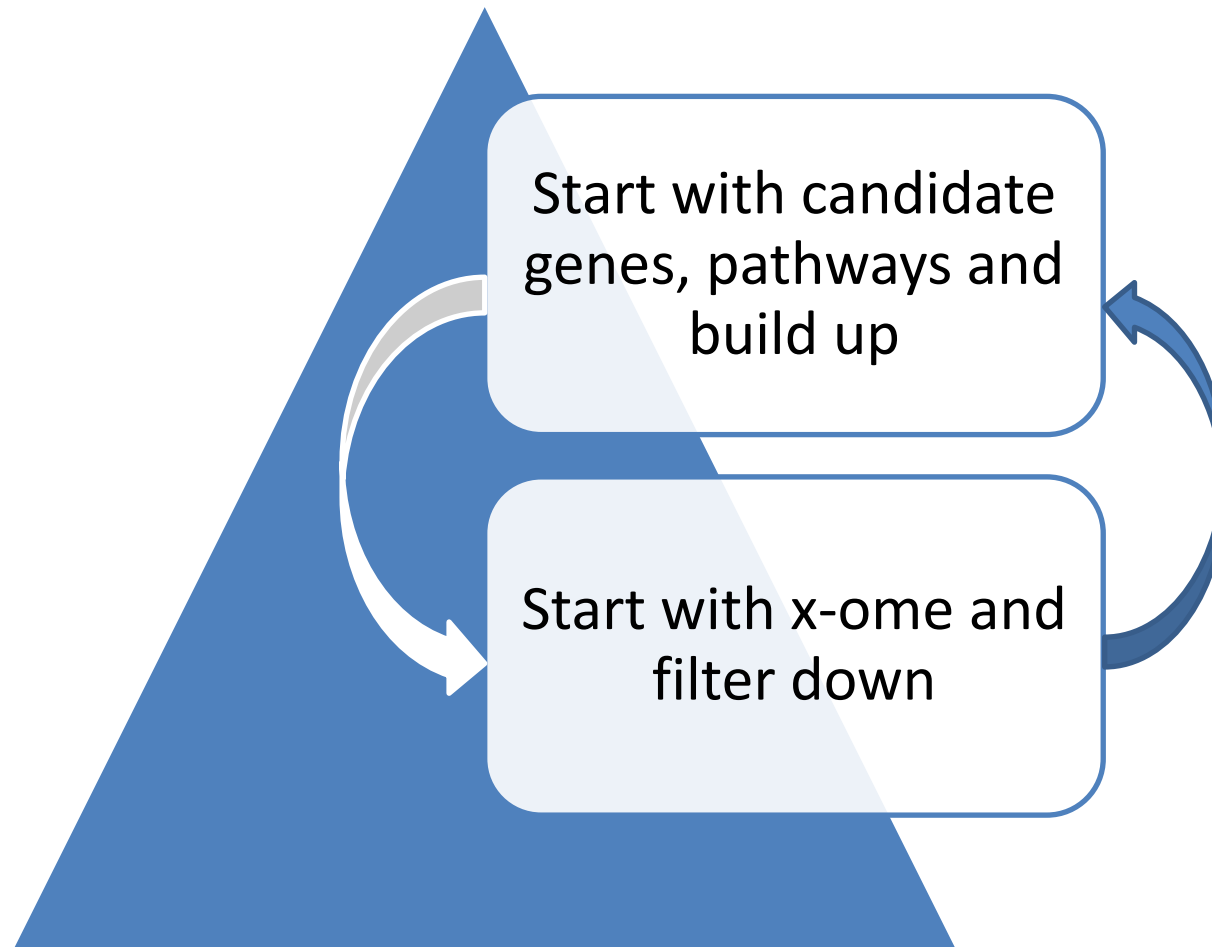
Step 5

- Interpretation

- Non-standard TODOs (?) when “integrating” evidences from biological data bases:
 - Assess and incorporate “optimal” scoring systems to accumulate evidence from these data bases
 - Allow for uncertainty involved in the data source entries
 - Acknowledge the complementary characteristics of each of the available data sources
 - Allow for different assignment strategies (e.g., from genetic variants to genes)

Integrative analytics

Top down versus bottom up

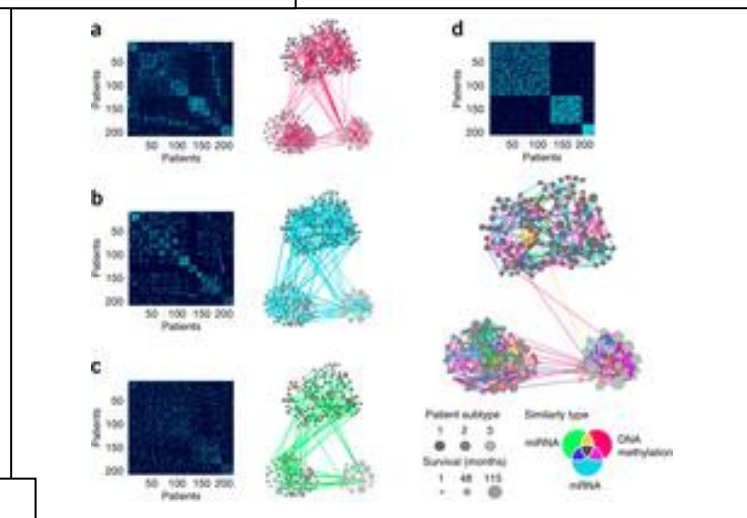
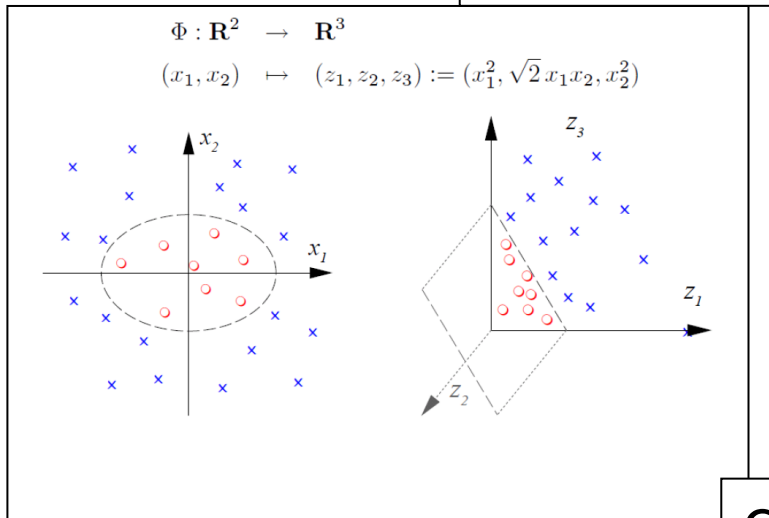


Integrative analytics

Crude division:

Kernels

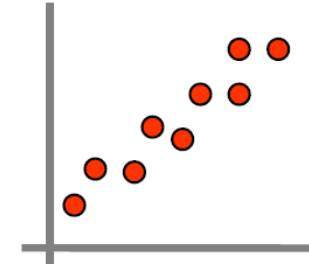
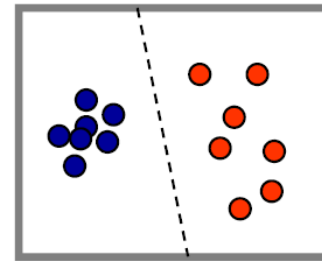
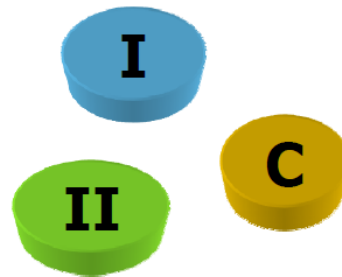
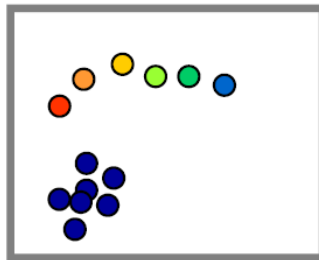
Networks



Components

Overview	Classification	Discrimination	Regression
<ul style="list-style-type: none"> Trends Outliers Quality Control Biological Diversity Patient Monitoring 	<ul style="list-style-type: none"> Pattern Recognition Diagnostics Healthy/Diseased Toxicity mechanisms Disease progression 	<ul style="list-style-type: none"> Discriminating between groups Biomarker candidates Comparing studies or instrumentation 	<ul style="list-style-type: none"> Comparing blocks of omics data Metab vs Proteomic vs Genomic Correlation spectroscopy (STOCSY)
PCA	SIMCA	PLS-DA OPLS-DA	O2-PLS

Finding the most appropriate method for your research question



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

(<http://www.metabolomics.se>)


Visual analytics

CABIN: Collective Analysis of Biological Interaction Networks

Mudita Singhal & Kelly Domica
Computational Biology and Bioinformatics
Pacific Northwest National Laboratory

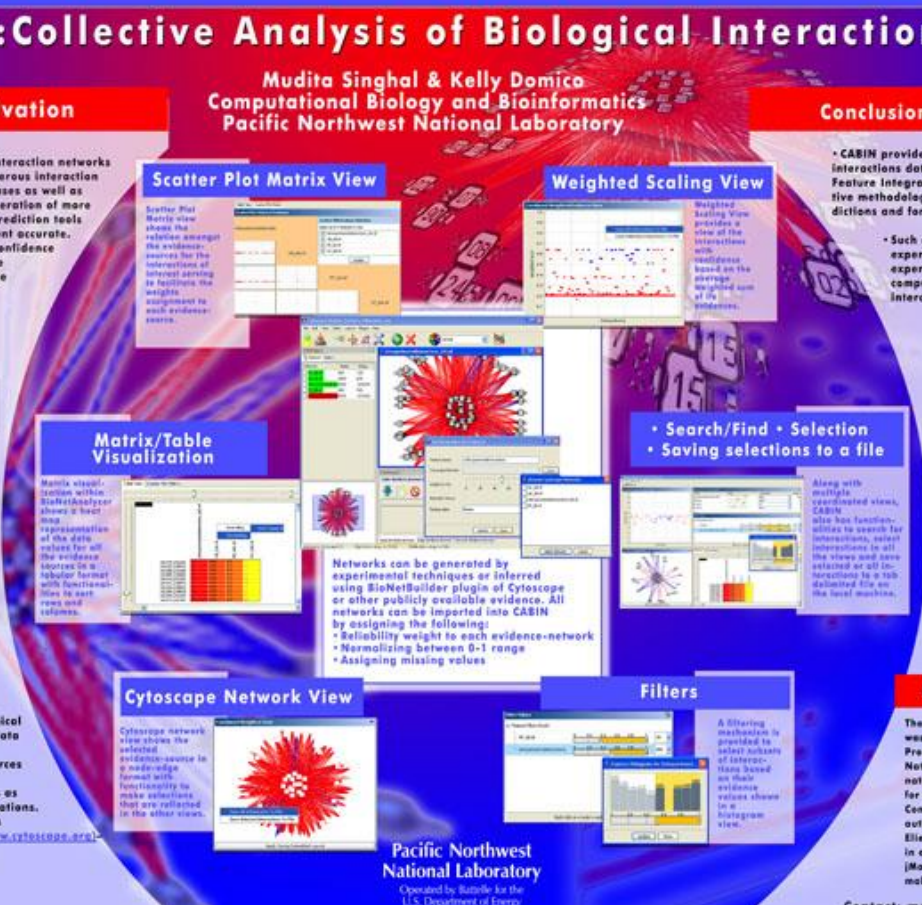
Abstract and Motivation

The importance of understanding interaction networks has fueled the development of numerous interaction data generation techniques, databases as well as prediction tools facilitating the generation of more interaction data. However not all prediction tools and databases are a hundred percent accurate. Generation of networks with high confidence interactions by integrating evidence from multiple sources formulates the first step towards deciphering unknown protein functions and determining protein complexes.



The tool of computational tools facilitating such integration of evidence from multiple prediction/experimental sources (Gene Neighbor-land-OK, Gene Cluster-OC, PathLinker, Profile-PP, BioGRID, BioGRID, SIBIN, and BIP etc.) is the motivation behind the Collective Analysis of Biological Interaction Networks (CABIN).

The CABIN: Collective Analysis of Biological Interaction Networks is an exploratory data analysis tool that enables fusion of interactions obtained from multiple sources of evidence, thereby increasing the confidence of computational predictions as well as validating experimental observations. CABIN has been written in JAVA™ and is available as a plugin for Cytoscape (www.cytoscape.org) an open source network visualization tool.



Scatter Plot Matrix View: Scatter Plot Matrix view shows the relation amongst the evidence sources for the interactions of interest. It facilitates the weights assignment to each evidence source.

Weighted Scaling View: Weighted Scaling View provides a view of the interaction network. Weights are based on the average weighted sum of its evidences.

Matrix/Table Visualization: Matrix visualization within ScatterAnalyzer shows a heat map representation of the data values for all the evidence sources in a tabular format with functionalities to sort rows and columns.

Cytoscape Network View: Cytoscape network view shows the selected evidence source in a node-edge format with functionality to make selections that are reflected in the other views.

Filters: A filtering mechanism is provided to select subsets of interactions based on their evidence values shown in a histogram view.

Search/Find/Selection: Saving selections to a file

Networks can be generated by experimental techniques or inferred using BioNetBuilder plugin of Cytoscape or other publicly available evidence. All networks can be imported into CABIN by assigning the following:
 • Reliability weight to each evidence-network
 • Normalizing between 0-1 range
 • Assigning missing values

Conclusions and Future Work

- CABIN provides tools for visualizing and analyzing interactions data from multiple sources of evidence. Feature integration has been demonstrated as effective methodology for increasing the confidence in predictions and for eliminating false positives.
- Such a tool is useful for validating experimental observations and designing experimental studies based on computational prediction of highly confident interactions.
- Future work involves refining the weights assignment process by providing default weights based on statistical reliability of the features; normalizing discrete or rank based data effectively; handling conflicts based on dependency amongst the interactions; and providing the option to impute missing values statistically.
- Connectivity with the BioNetBuilder and Literature Search Agilent plug-ins of Cytoscape will be established.
- SIBIN (Taylor et al.) and CABIN can be used in conjunction with each other to facilitate effective network inference.

Acknowledgements

The research described in this paper was conducted under the LDRD Program at the Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76OR21400. The authors acknowledge contributions from Elie Noullet (semio@pnl.gov) in obtaining the updated version of the jMatrixView open source library and making it available for public use.

Contact: mudita.singhal@pnl.gov PNL-02-0110

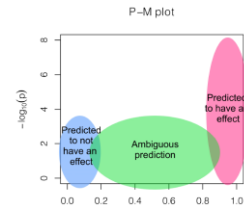
(<http://www.sysbio.org/capabilities/compbio/cabin.stm>)

Methodological challenges

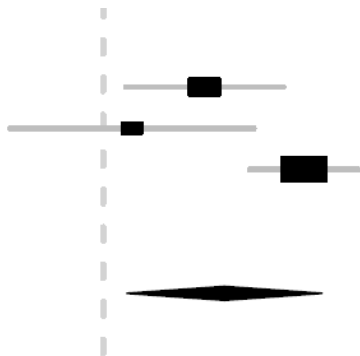
A toy example

Methodological aspects: scaling up from GWAs to GWAs

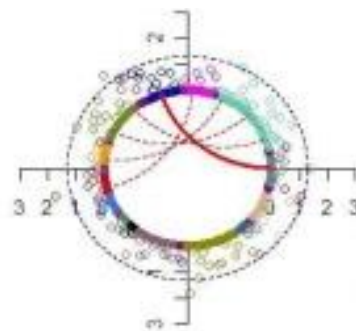
4 meta-GWAs



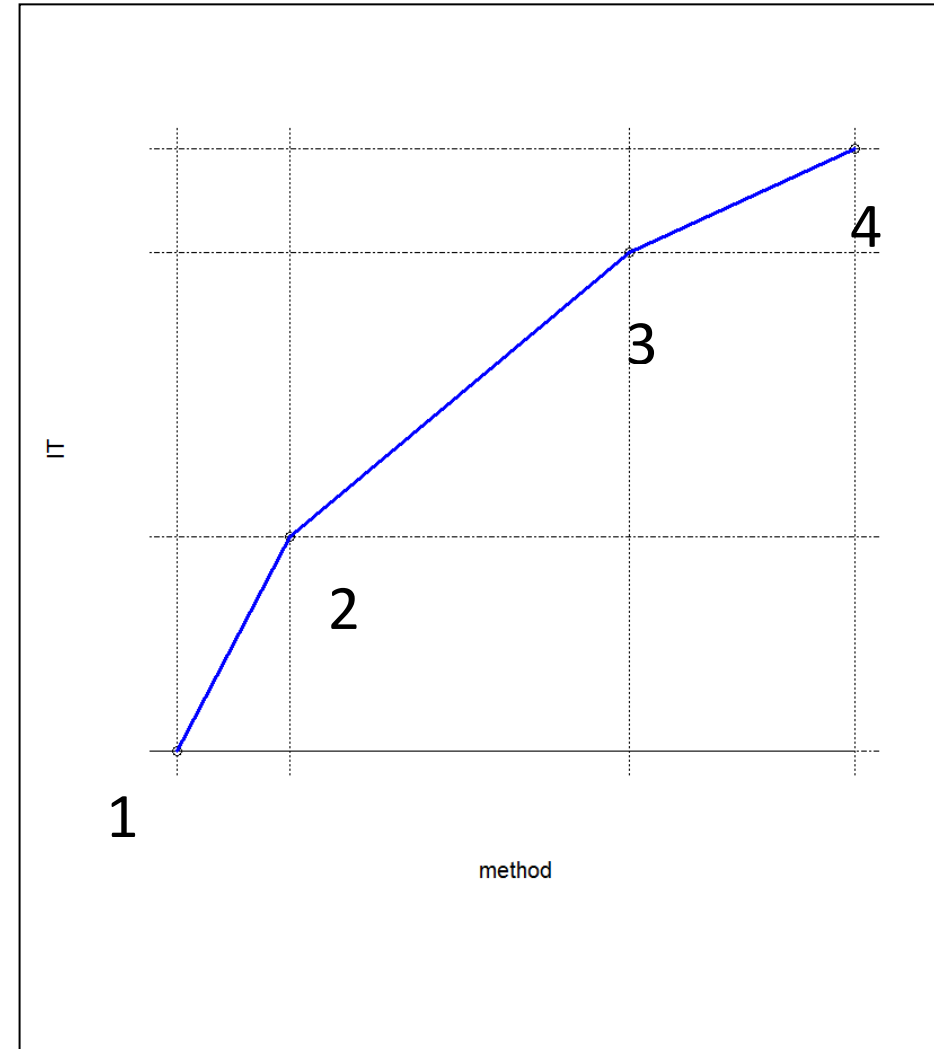
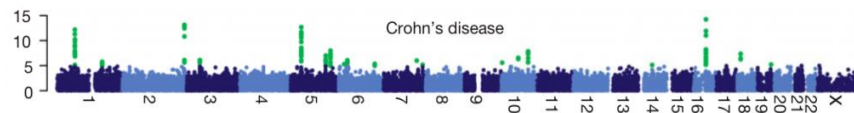
2 meta-GWA



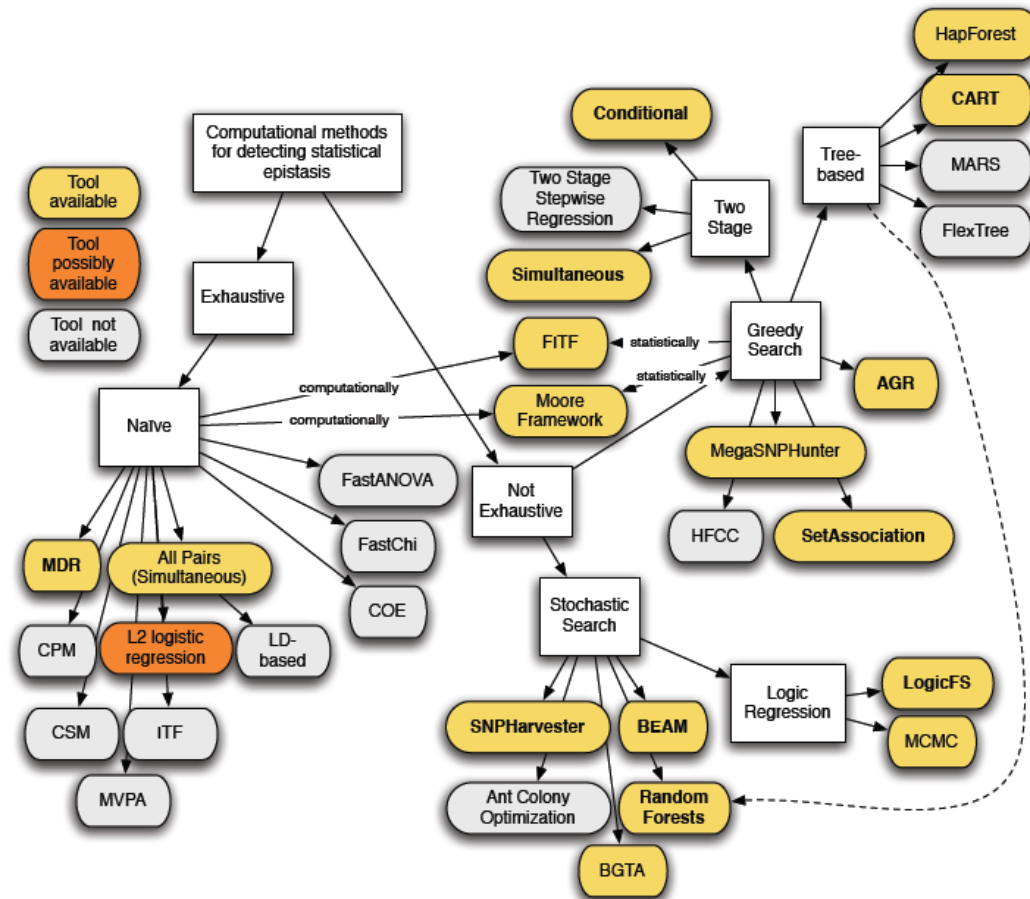
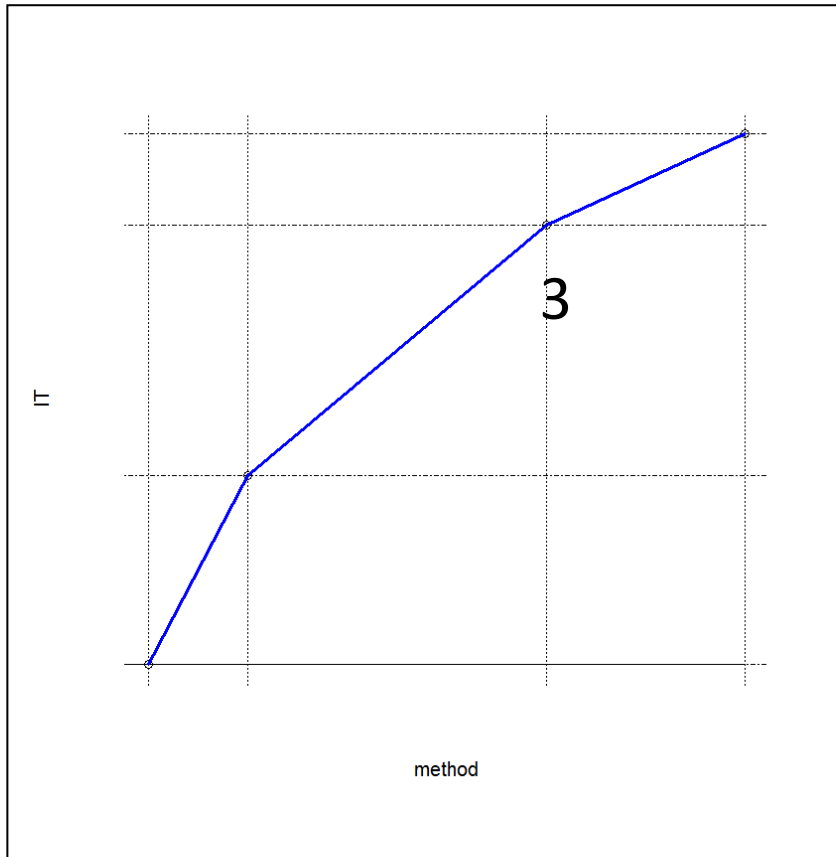
3 GWAs



1 GWAs

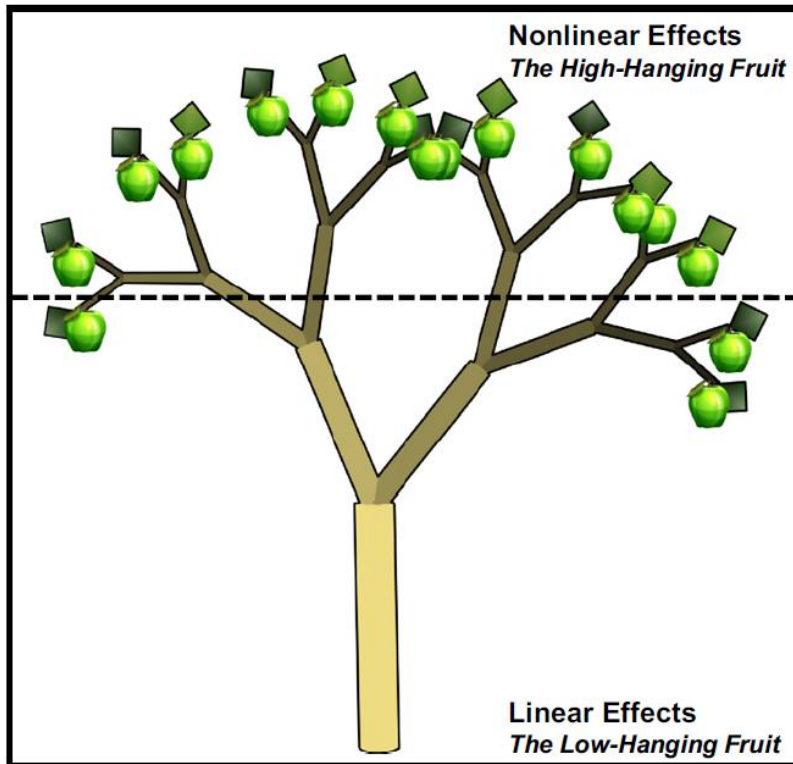


Methodological aspects: scaling up from GWAs to GWAs



(Kilpatrick 2009)


Methodological aspects: complexity



(Moore and Williams 2009)

- 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: nonlinear gene-gene interactions, gene-environment interaction, locus heterogeneity...

Methodological aspects: integration



NIH Public Access

Author Manuscript


Circ Cardiovasc Genet. Author manuscript; available in PMC 2012 October 1.

NIH-PA Author Manuscript

Published in final edited form as:
Circ Cardiovasc Genet. 2011 October 1; 4(5): 549–556. doi:10.1161/CIRCGENETICS.111.960393.

Protein Interaction-Based Genome-Wide Analysis of Incident Coronary Heart Disease

Majken Girman,
¹Department of Medical
²Department and Work

OPEN ACCESS Freely available online


Pathway Analysis Using Information from Allele-Specific Gene Methylation in Genome-Wide Association Studies for Bipolar Disorder

Li-Chung Chuang^{1,2}, Chung-Feng Kao¹, Wei-Liang Shih¹, Po-Hsiu Kuo^{1,3*}

¹ Department of Public Health & Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan, ² Department of Nursing, Cardinal Tien College of Healthcare & Management, I-Lan, Taiwan, ³ Research Center for Genes, Environment and Human Health, National Taiwan University,

ORIGINAL RESEARCH ARTICLE

published: 31 May 2013
doi: 10.3389/fgene.2013.00000

Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma

Lin Li^{1†}, Michael Kabesch², Emmanuelle Bouzigon^{3,4}, Florence Demenais^{3,4}, Martin Farrall⁵, Miriam F. Moffatt⁶, Xihong Lin¹ and Liming Liang^{1,7*}

¹ Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

² Department of Pediatric Pneumology and Allergy, KUNO University Children's Hospital Regensburg, Regensburg, Germany

³ INSERM, Genetic Variation and Human Diseases Unit, U946, Paris, France

⁴ Sorbonne Paris Cité, Institut Universitaire d'Hématologie, Université Paris Diderot, Paris, France

⁵ Wellcome Trust Centre for Human Genetics, Oxford, UK

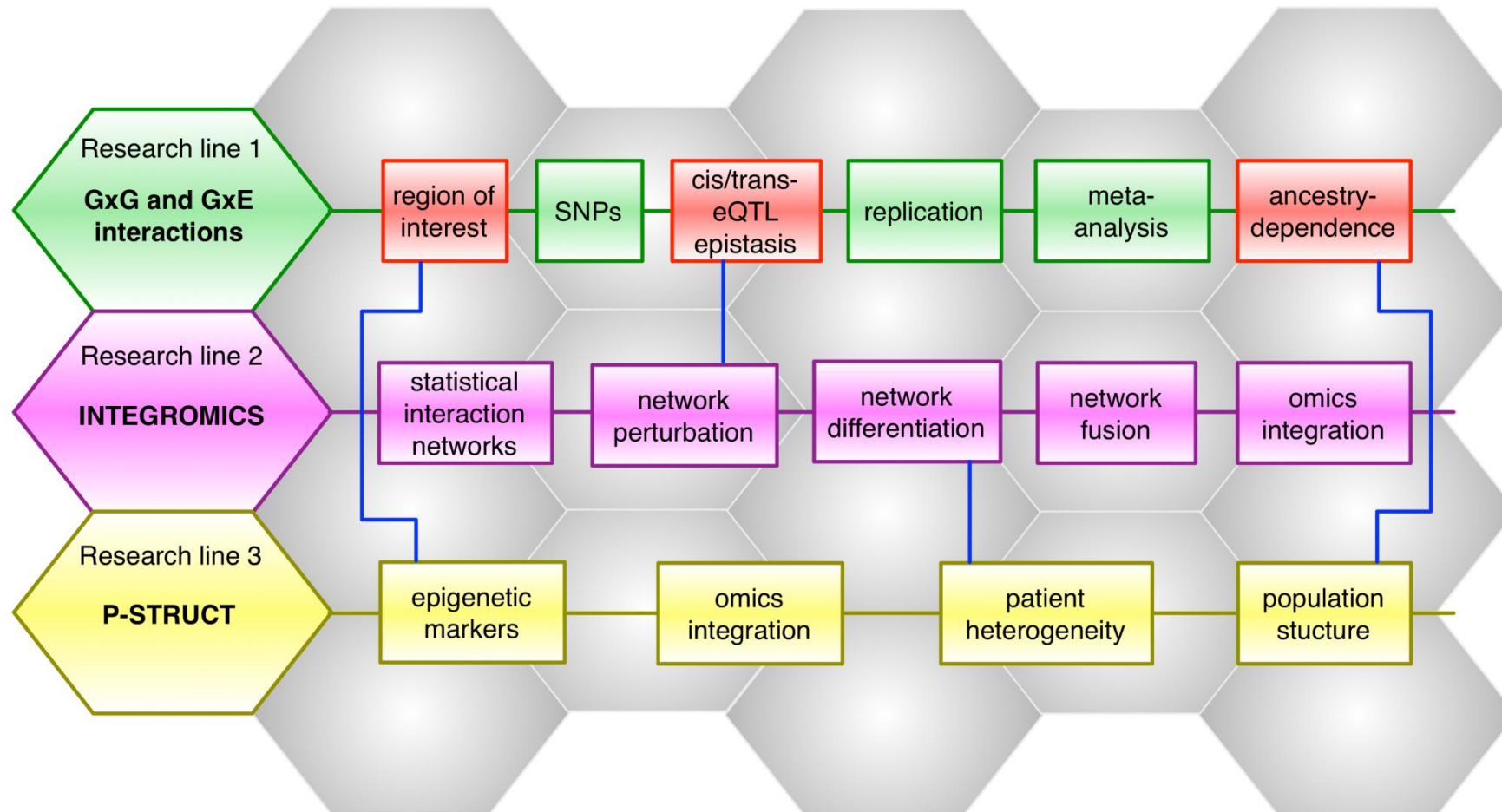
⁶ Molecular Genetics and Genomics Section, National Heart and Lung Institute, Imperial College London, London, UK

⁷ Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA

GWAs as a toy example

- Integrative analytics
 - Kernel-theory -----
 - Components-theory -----
 - Network theory -----
- GWAI analytics
 - Kernel-based methods to detect rare variant associations in the presence of interactions
 - PCA to capture population stratification
 - Statistical epistasis networks

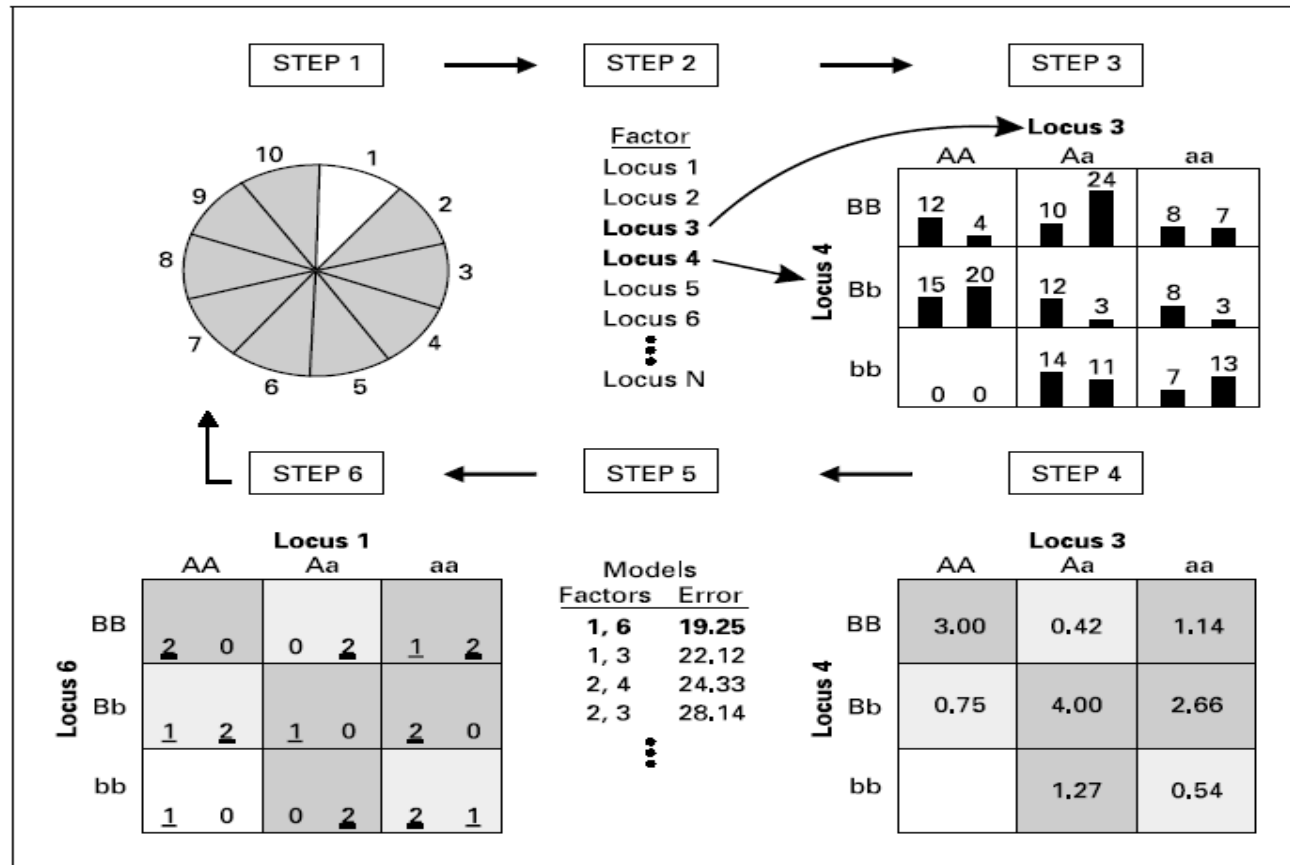
Bio3's research lines



Towards a novel integrated framework “(gen)omic MB-MDR”

Historical notes about MB-MDR

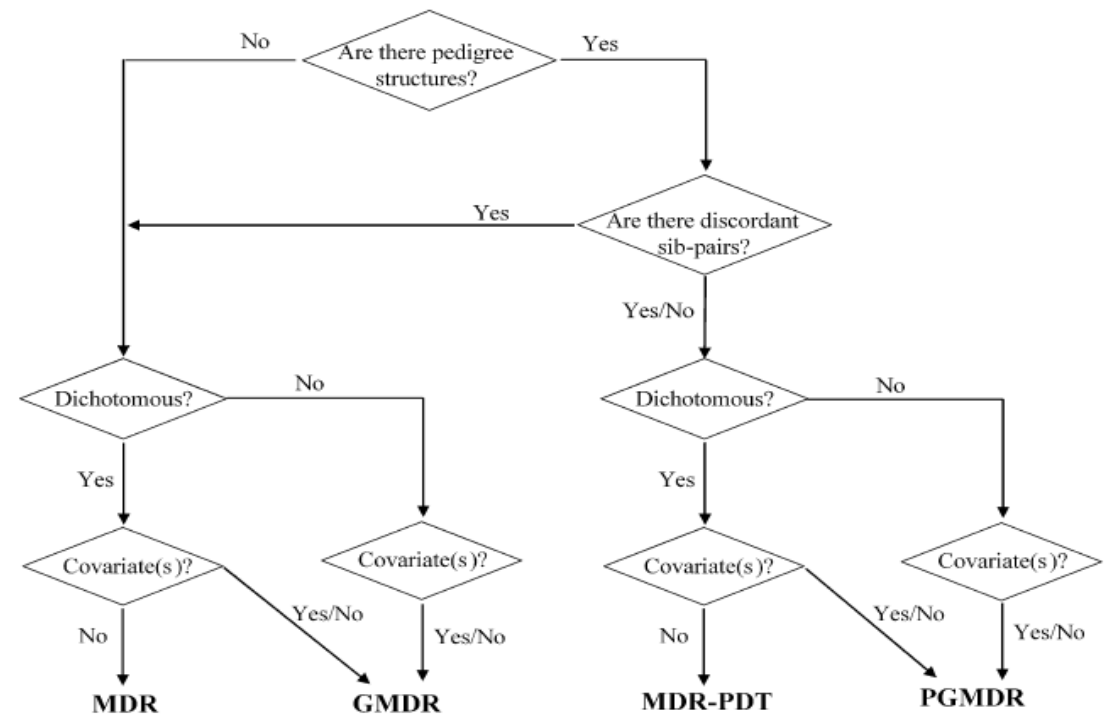
- Start: Multifactor Dimensionality Reduction by MD Ritchie et al. (2001)



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

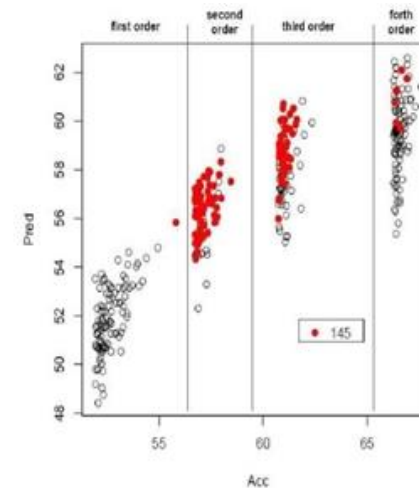


Historical notes about MB-MDR

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

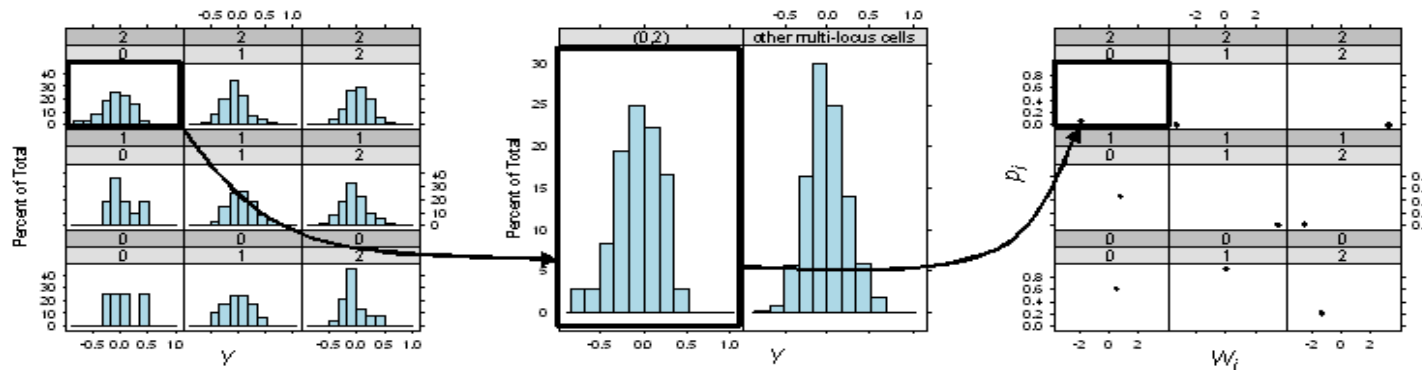
- Computation time is invested in optimal **association tests** to prioritize multilocus genotype combinations (e.g., high, low, no evidence) and in **statistically valid permutation-based methods** to assess joint statistical significance.

- At the same time, a “quantification” of “interaction” signals can be obtained above and beyond **lower order effects**

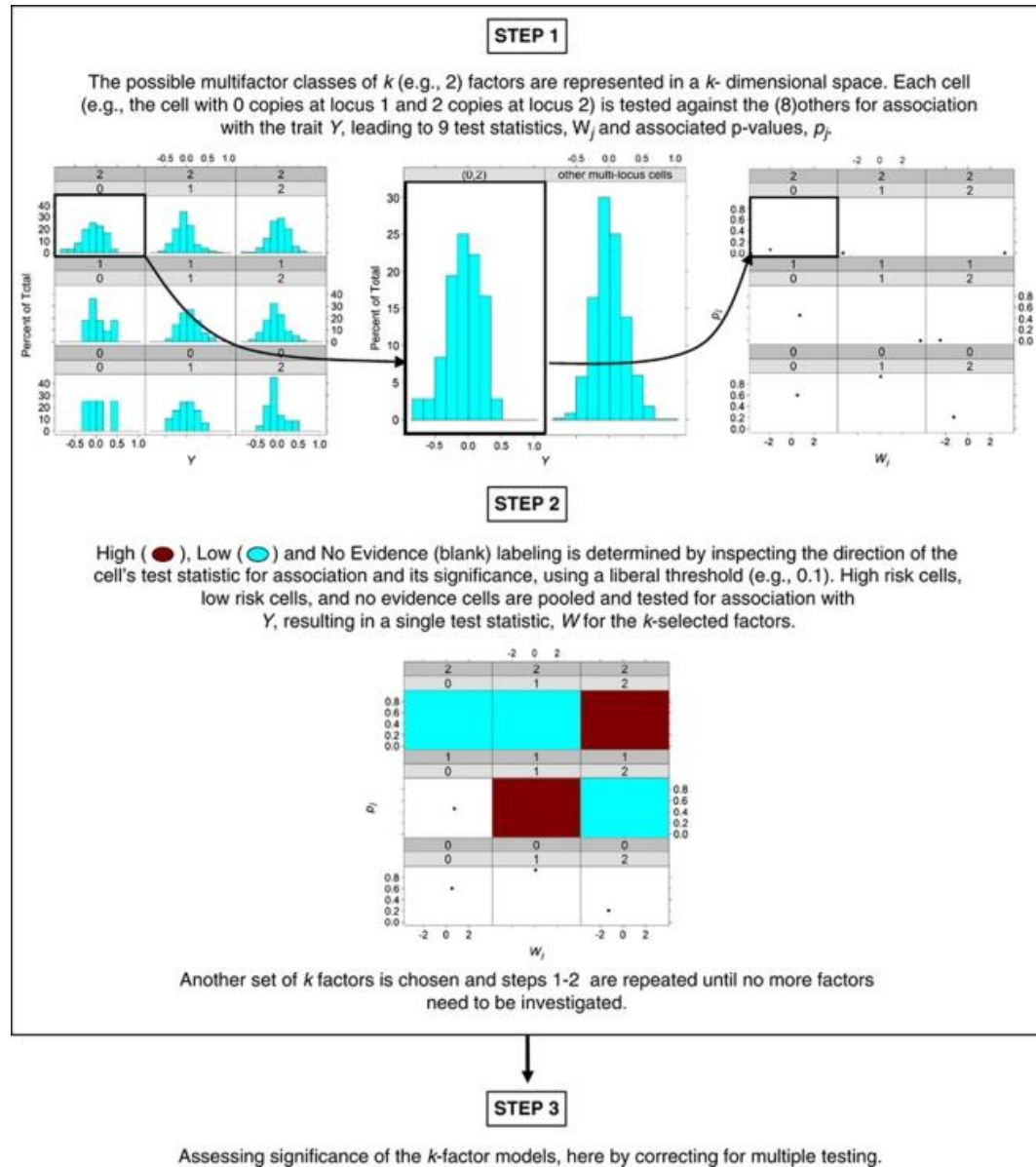


Historical notes about MB-MDR

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



- 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)



MB-MDR

Step 1: organization of data in multi-locus cells (here: 2D) and assessing relevance.

Step 2: Label and reduce dimensionality by pooling equally-labelled cells.

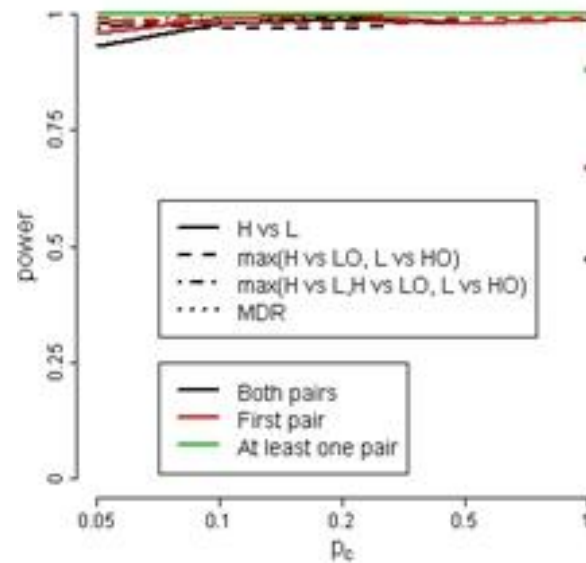
Step 3: Assess joint significance over all multi-locus models

Historical notes about MB-MDR

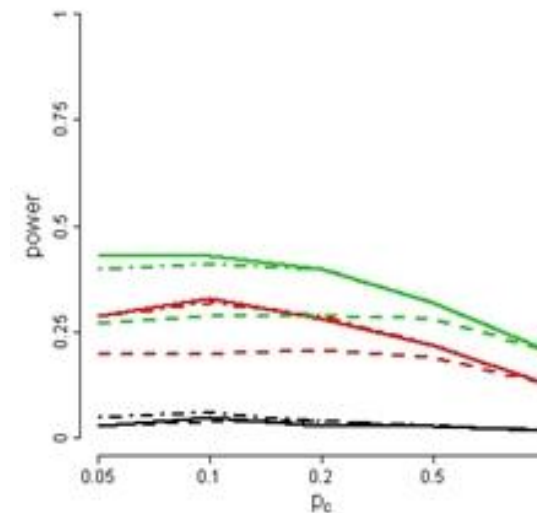
- Power: Model-Based MDR by Cattaert et al. (2011) – genetic heterogeneity

Model 2, $p = 0.5$

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

Ritchie Model 2 ($p=0.5$)Model 6, $p = 0.1$

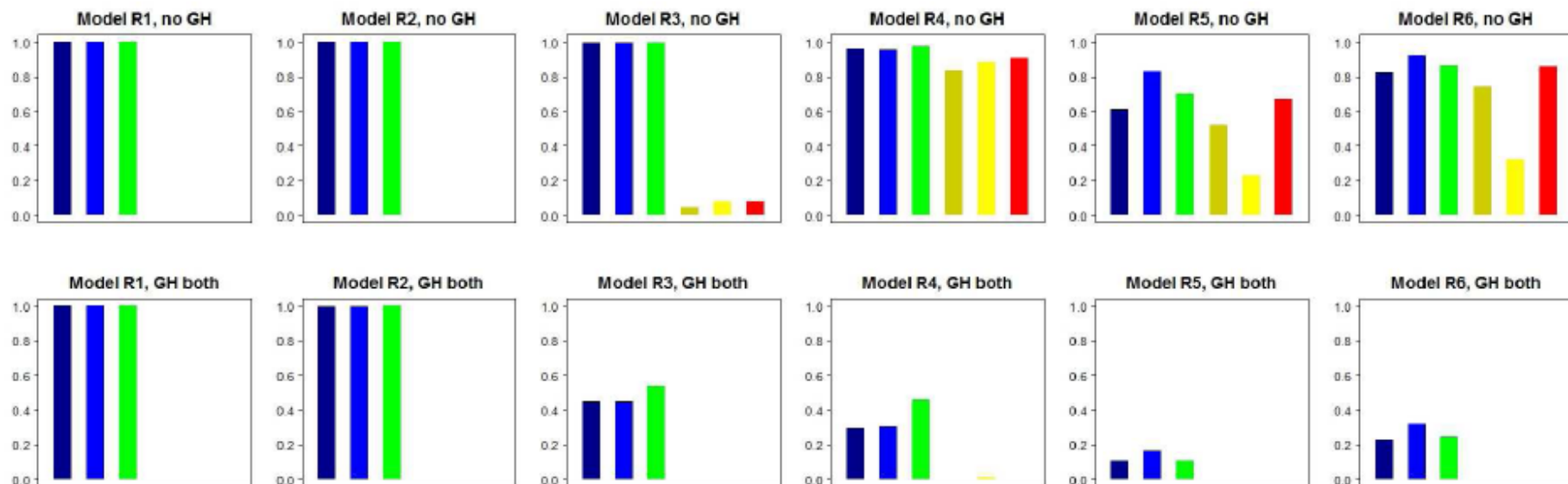
	BB	Bb	Bb
AA	0.09	0.001	0.02
Aa	0.08	0.07	0.005
aa	0.003	0.007	0.02

Ritchie Model 6 ($p=0.1$)

Historical notes about MB-MDR

- Power performance

(example: pure epistasis scenario's; unpublished – 2010-2014)



BOOST (dark blue)

EpiCruncher optimal options (light blue)

MB-MDR (green)

PLINK epistasis (dark yellow)

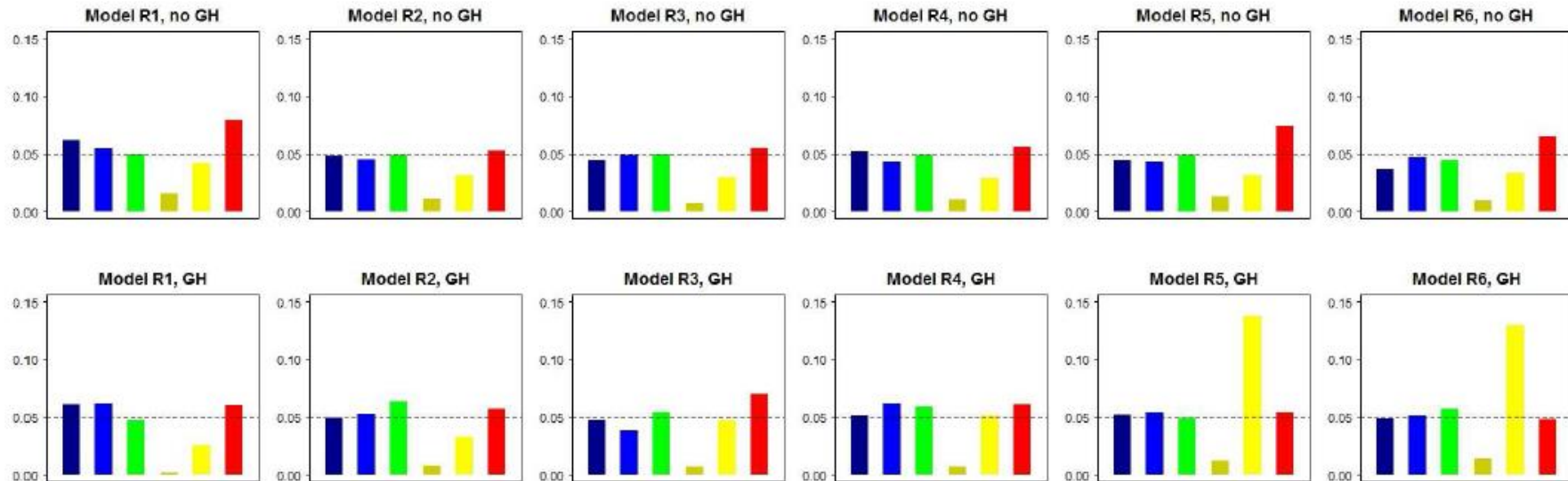
PLINK fast epistasis (light yellow)

EPIBLASTER (red)

Historical notes about MB-MDR

- FWER performance

(example: pure epistasis scenario's; unpublished – 2010-2014)



BOOST (dark blue)

EpiCruncher optimal options (light blue)

MB-MDR (green)

PLINK epistasis (dark yellow)

PLINK fast epistasis (light yellow)

EPIBLASTER (red)

Computational Efficiency

From GWAs to exomes: speed

- Situation in 2014 (Van Lishout et al. - manuscript in preparation)

SNPs	MBMDR-4.2.2 Binary trait sequential execution	MBMDR-4.2.2 Binary trait parallel workflow	MBMDR-4.2.2 Continuous trait sequential execution	MBMDR-4.2.2 Continuous trait parallel workflow
10^3	13 min 33 sec	20 sec	13 min 18 sec	18 sec
10^4	52 min 15 sec	1 min 05 sec	56 min 14 sec	53 sec
10^5	64 hours 35 min	22 min 15 sec	70 hours 3 min	20 min 28 sec
10^6	≈ 270 days	25 hours 12 min	≈ 290 days	≈ 24 hours

The parallel workflow was tested on a 256-core computer cluster (Intel L5420 2.5 GHz).
The sequential executions were performed on a single core of this cluster.

- Situation < 2013 (Van Lishout et al. 2013)

MB-MDR-3.0.2 binary trait sequential execution (input 10^5 SNPs): 1.5 years

MB-MDR-3.0.2 cnt trait sequential execution (input 10^5 SNPs): 3 years

Population and patient substructures

Detecting structure in patients: subphenotyping

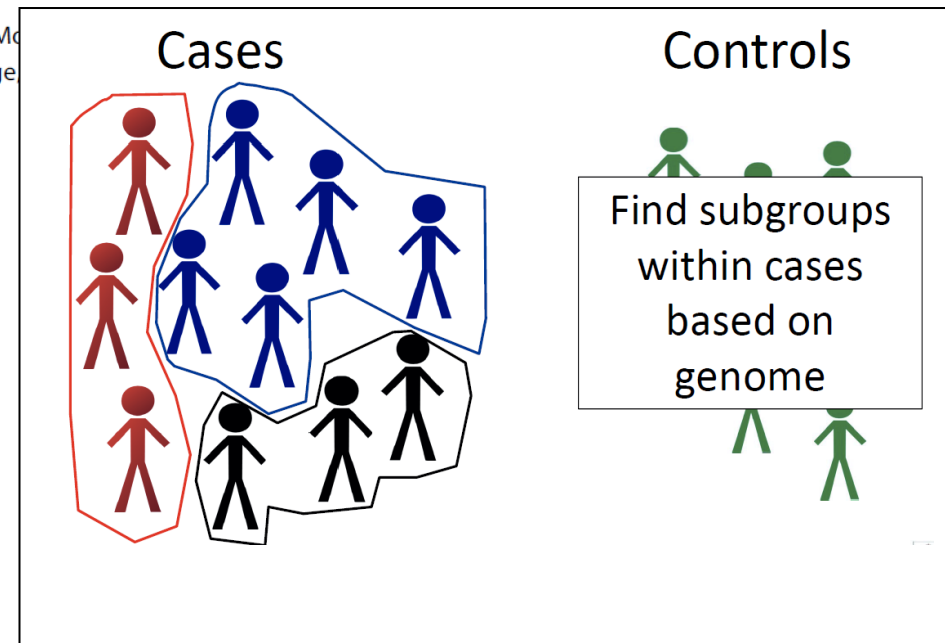
OPEN ACCESS Freely available online



Molecular Reclassification of Crohn's Disease by Cluster Analysis of Genetic Variants

Isabelle Cleynen^{1*}, Jestinah M. Mahachie John^{2,3}, Liesbet Henckaerts⁴, Wouter Van Moerkercke¹, Paul Rutgeerts¹, Kristel Van Steen^{2,3}, Severine Vermeire¹

¹ Department of Gastroenterology, KU Leuven, Leuven, Belgium, ² Systems and Modeling, University of Liège, Liège, Belgium, ³ Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium, ⁴ Department of Gastroenterology, KU Leuven, Leuven, Belgium



Detecting structure in patients: subphenotyping

OPEN ACCESS Freely available online



Molecular Reclassification of Crohn's Disease: A Cautionary Note on Population Stratification

Bärbel Maus^{1,2*}, Camille Jung^{3,4,5}, Jestinah M. Mahachie John^{1,2}, Jean-Pierre Hugot^{3,4,6}, Emmanuelle Génin^{7,8}, Kristel Van Steen^{1,2}

	H ₀ : 1 grp H _A : 2 grps	H ₀ : 2 grps H _A : 3 grps	...	H ₀ : 8 grps H _A : 9 grps	H ₀ : 9 grp H _A : 10 grps
-2LL Diff	897.5524	489.0997	...	140.6088	84.8221
p- value	<0.0001	<0.0001	...	<0.0001	0.4640

(Bootstrap p-value ; AIC : 9 groups ; BIC: 3 groups)

e, Liège
rt Deb
Génc

- **Latent class modeling**
applied to continuous pop-
adjusted SNP data requires
Gaussian distribution ...

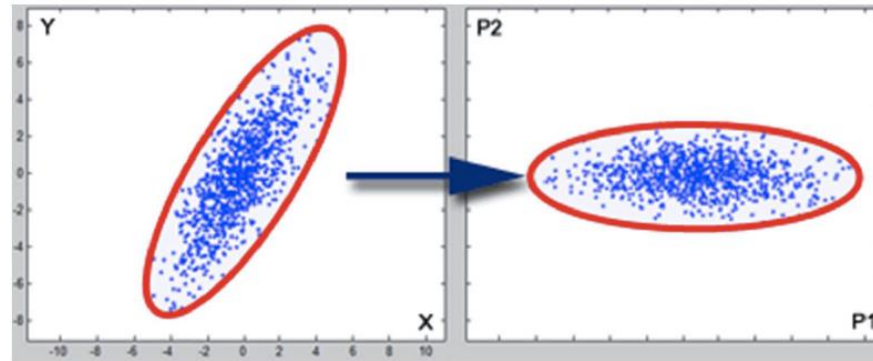
Detecting structure in patients: subphenotyping

		Unadjusted			Adjusted	
		LCA	PAM	HC	PAM	HC
Unadjusted	LCA		0.49	0.30	0.12	0.23
	PAM			0.23	0.20	0.13
	HC				0.04	0.54
Adjusted	PAM					0.04
	HC					

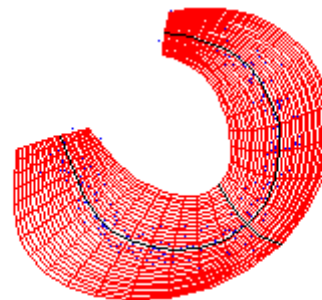
- Adjusted Rand Index between latent class analysis (LCA), PAM clustering and hierarchical clustering using Ward linkage and squared Eucl. distance (using population unadjusted and adjusted SNP data)
- Clusters ~ Clinical features: focus on populations with a similar genetic background

Detecting structure in patients / populations

- Orthogonal linear transformation of the data



- Non-linear PCA (e.g., based on an auto-associative neural networks)



Detecting structure in patients: sub-phenotyping

Hum Genet

DOI 10.1007/s00439-014-1480-y

REVIEW PAPER

Practical aspects of genome-wide association interaction analysis

Elena S. Gusareva · Kristel Van Steen

Received: 21 May 2014 / Accepted: 18 August 2014

© Springer-Verlag Berlin Heidelberg 2014

Abstract Large-scale epistasis studies can give new clues to system-level genetic mechanisms and a better understanding of the underlying biology of human complex disease traits. Though many novel methods have been proposed to carry out such studies, so far only a few of them have demonstrated replicable results. Here, we propose a minimal protocol for genome-wide association interaction (GWAi) analysis to identify gene–gene interactions from large-scale genomic data. The different steps of the devel-

Introduction

Genome-wide association (GWA) studies have been very successful in identifying predisposing genetic variants to a variety of complex traits (e.g., GWAS Diagram Browser for exploring GWA studies at <http://www.ebi.ac.uk/fgpt/gwas/> and the Catalog of Published Genome-Wide Association Studies at http://www.genome.gov/page.cfm?pageid=26525384&clearquery=1#result_table). Still, yet to identify

Meta-analysis

Meta-GWAs

ARTICLE IN PRESS

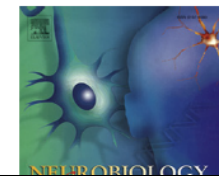
Neurobiology of Aging xxx (2014) 1–8



Contents lists available at [ScienceDirect](#)

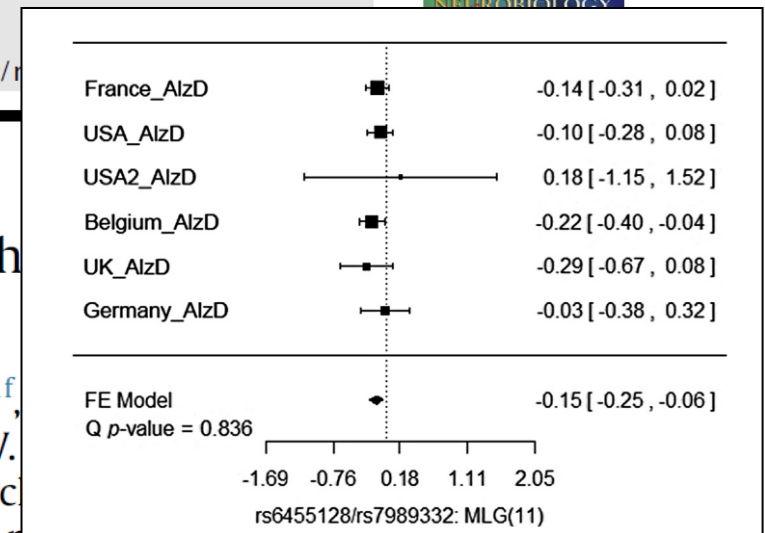
Neurobiology of Aging

journal homepage: www.elsevier.com/locate/na



Genome-wide association interaction analysis for Alzheimer disease[☆]

Elena S. Gusareva^{a,b,*}, Minerva M. Carrasquillo^c, Céline Bellenguez^{d,e,f}, Samuel Colon^c, Neill R. Graff-Radfordⁱ, Ronald C. Petersen^j, Dennis W. Jostina M. Mahachie John^{a,b}, Kyrylo Bessonov^{a,b}, Christine Van Broecklin¹, Denise Harold^k, Julie Williams^k, Philippe Amouyel¹, Kristel Slegers^{g,h}, Nilüfer Ertekin-Taner^{c,i}, Jean-Charles Lambert^{d,e,f}, Kristel Van Steen^{a,b}



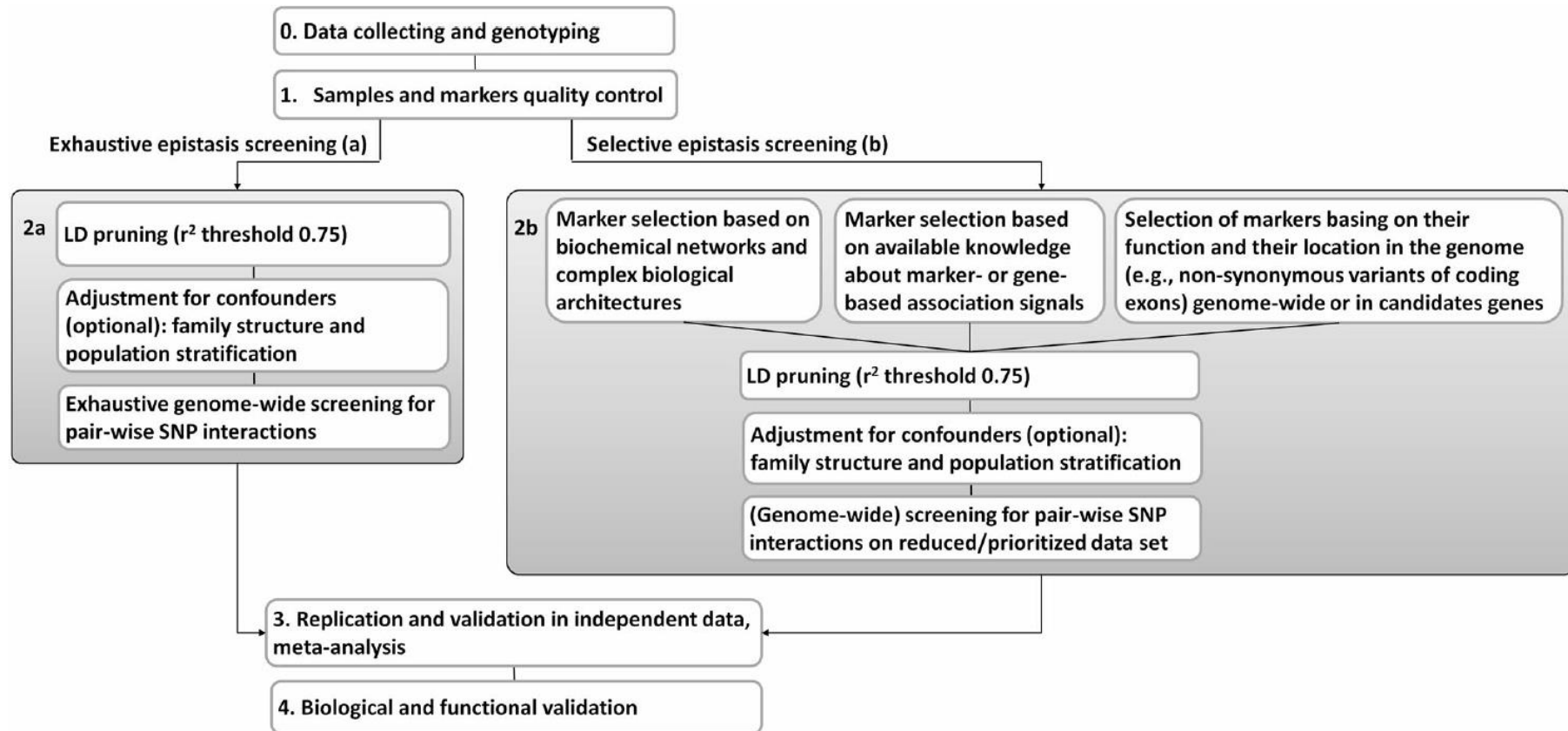
Imputation

Imputation: a blessing or a curse?

- When aligning two independent datasets by imputing missing markers, one SNP in a SNP-pair may be imputed in one dataset, whereas it is actually observed in another dataset.
 - So even when the same SNP pair is highlighted in a significant genetic interaction in this setting, can we really talk about **“replication”**?
- Imputation in GWAs is based on LD-blocks
 - Imputation can therefore induce increased LD between markers and hence **“redundant epistasis”** (should be dealt with appropriately)

Interpretation

Statistical versus biological epistasis

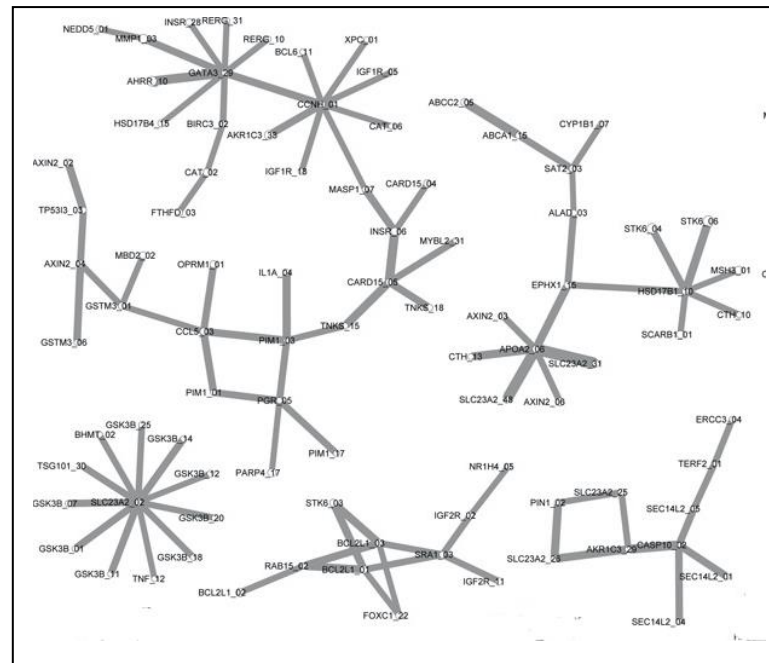


- Protocol for GWAs (analytic blocks are highlighted)

(Gusareva et al. 2014)

Statistical epistasis networks with MB-MDR

- Motivation:
Statistical epistasis networks can reduce the computational complexity of searching for higher (>2) order genetic models



(Hu et al. 2013)

Replication and validation

Unable to replicate is a bad thing?

OPEN ACCESS Freely available online



Failure to Replicate a Genetic Association May Provide Important Clues About Genetic Architecture

Casey S. Greene¹, Nadia M. Penrod¹, Scott M. Williams², Jason H. Moore^{1,2,3,4,5,6*}

¹ Department of Genetics, Dartmouth College, Lebanon, New Hampshire, United States of America, ² Vanderbilt University, Center for Human Genetics, Nashville, Tennessee, United States of America, ³ Department of Community and Family Medicine, Dartmouth Medical School, Lebanon, New Hampshire, United States of America, ⁴ Department of Computer Science, University of New Hampshire, Lebanon, New Hampshire, United States of America, ⁵ Department of Computer Science, University of Vermont, Burlington, Vermont, United States of America, ⁶ Translational Genomics Research Institute, Phoenix, Arizona, United States of America

Abstract

Replication has become the gold standard for assessing statistical results from genome-wide association studies. Unfortunately this replication requirement may cause real genetic effects to be missed. A real result can fail to replicate for numerous reasons including inadequate sample size or variability in phenotype definitions across independent samples. In genome-wide association studies the allele frequencies of polymorphisms may differ due to sampling error or population differences. We hypothesize that some statistically significant independent genetic effects may fail to replicate in an independent dataset when allele frequencies differ and the functional polymorphism interacts with one or more other functional polymorphisms. To test this hypothesis, we designed a simulation study in which case-control status was determined by two interacting polymorphisms with heritabilities ranging from 0.025 to 0.4 with replication sample sizes ranging from 400 to 1600 individuals. We show that the power to replicate the statistically significant independent main effect of one polymorphism can drop dramatically with a change of allele frequency of less than 0.1 at a second interacting polymorphism. We also show that differences in allele frequency can result in a reversal of allelic effects where a protective

Replication using tagSNPs (often no functional consequence)

- Genome-wide SNP genotyping platforms consist predominantly of tagSNPs from across the genome.
- When two or more tagSNPs are combined in a genetic interaction model, is it reasonable to assume that the same combination of tagSNPs interacts in an independent dataset?
- “Due to variation in allele frequency and underlying linkage disequilibrium patterns between two datasets, it is highly unlikely that the same combination of tagSNPs would be associated in the same statistical model.”
- “We would expect that the combination of underlying signals that those SNPs are tagging would replicate across datasets, rather than the tagSNPs themselves”
(Ritchie and Van Steen 2014 – under review)

No replication without an analytic consensus

- Multiple testing handling / speedy algorithms (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 (Kridsakorn Chaichoompu)
- Non-linear relationships in population genetics (Ramouna Fouladi)
- Within- (Silvia Pineda) and between-gene architectures (K Bessonov)
- Missing data handling (Kristel Van Steen)
- Gene-based or **set-based** testing (Elena Gusareva, Ramouna Fouladi)

Combining it all: genomic MB-MDR

Gene-based or set-based testing

MB-MDR

Individuals may be similar wrt 2-locus genotypes: AAbb (red)

BB			
Bb			
bb			
	AA	Aa	aa

1 dimension = 1 genetic maker
(grouping based on 2-locus genotypes)



Genomic MB-MDR

Individuals may be similar wrt “features” (common and rare variants, epigenetic markers)



1 dimension = 1 ROI
(grouping on features mapped to the ROI)

An integrated framework based on MB-MDR

Step 1: Descriptor filtering

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

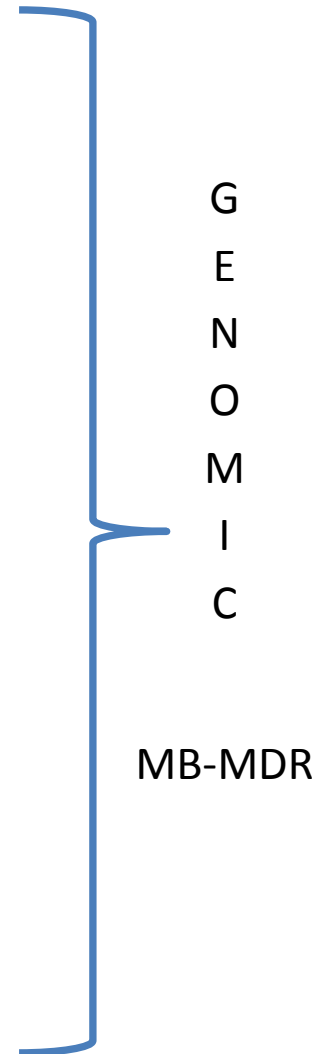
(don't throw away rare variants)

Step 2: Choice of clustering approach

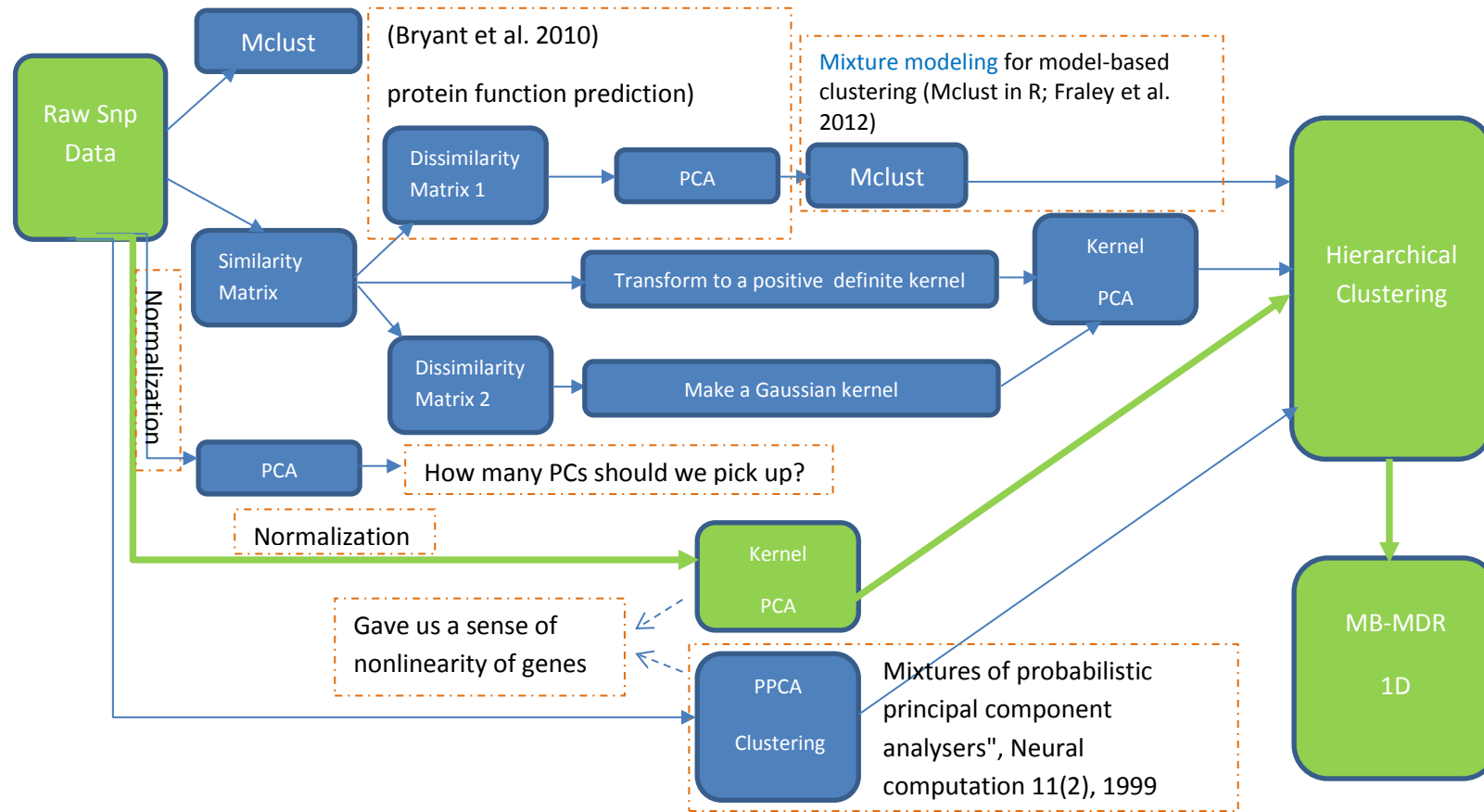
- Flexible integration of heterogeneous data (scaling)
- Flexible integration of data interdependencies
- Allowing low-variance descriptors

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

Step 3: Application of “classic” MB-MDR



Genomic MB-MDR step 1 + step 2



(adapted slide from Fouladi 2014)

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 out of 200 per gene)

SKAT-O

KDR	All other genes
130 times out of 200	396 times out of 200x80 (396/80 ~5 out of 200 per gene)

(Burden tests ☺: most variants are causal, with effects in same direction; SKAT ☺: large fraction of the variants are non-causal or the effects of causal variants are in different directions; SKAT-O: joins the best of both)

In conclusion

Genomic MB-MDR

- Genomic MB-MDR seems to be a flexible tool in different contexts
- Filtering: assigning “weights” to alleles → MAF in a control population, possible alterations in protein function, including measures produced by f.i. eXtasy (SIFT, Polyphen2,...) (Zuk et al. 2014)
- Interpretation
 - clusters (step 2), dimensionality reduced cells (step 3)

*“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.”*

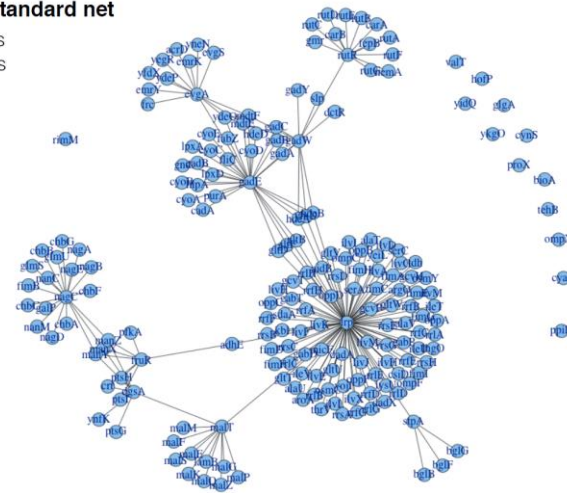
(<http://www.stat.cmu.edu/~cshalizi/>)

Genomic MB-MDR applied to ...

- Gene-based association analysis
(~GWIS - Huang et al 2011)
- Gene-gene statistical interactions
(~ GGG – Ma et al. 2013)
- Gene-gene statistical interaction networks
(~ correlation-based networks/differential network analysis, machine learning based or “forest”-based network construction)
- Integrating different types of omics data
(genetic + epigenetic variants)

Golden standard net

200 nodes
212 edges



Methodological aspects in integromics

- A series of challenges will need to be overcome:
 - protocol development for standardizing data generation and pre-processing or cleansing in integrative analysis contexts,
 - development of computationally efficient analytic tools to extract knowledge from dissimilar data types to answer particular research questions,
 - the establishment of validation and replication procedures, and tools to visualize results.

Mission ... possible



(Mission Impossible @ google)

Acknowledgement

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



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

Groupe Interdisciplinaire de Génomprotéomique Appliquée



References

- **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, Théâtre 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!]**

EXTRA SLIDES

From SNPs to genes with baby steps - rare variants setting

- Rare variants may help to explain some of the missing heritability of complex diseases; mixed messages towards their importance (when not considered with other structural variation in the genome)
- The low frequency of these rare variants raises issues about how best to analyze them (Bodmer and Bonilla 2009)
- In the general context of biostatistics, the low-frequency problem is known as the imbalance in the covariate distributions
- Several rare variants analysis methods have been evaluated during the GAW17 (Bailey-Wilson et al. 2011)
- All methods largely group into a few classes (Dering et al. 2011, Tachmazidou et al. 2012)

Rare variants setting

- Use aggregation of single rare variants (RVs) into meaningful groups or regions of interest (ROIs). Examples: genes, pathways, ...
 - Weighted burden tests (e.g., Liu and Leal 2010 - KBAC)
 - Use collapsed constructs in a regression framework (e.g., Lasso – Zhou et al. 2010)
- Use similarities between individuals based on their sequence data
 - Use multi-marker test while combining single-variant stats (Wu et al. 2011 - SKAT: ideas from kernel theory and regression)

Prototype development - clustering

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

Individual 2	Individual 1		
	<i>aa</i>	<i>aA</i>	<i>AA</i>
<i>aa</i>	$\frac{2}{p_a^2}$	$\frac{1}{p_a^2} - \frac{1}{2p_a(1-p_a)}$	$-\frac{1}{p_a(1-p_a)}$
<i>aA</i>	$\frac{1}{p_a^2} - \frac{1}{2p_a(1-p_a)}$	$\frac{1}{2} \left\{ \left[\frac{1}{p_a^2} + \frac{1}{(1-p_a)^2} \right] - \frac{1}{p_a(1-p_a)} \right\}$	$\frac{1}{(1-p_a)^2} - \frac{1}{p_a(1-p_a)}$
<i>AA</i>	$-\frac{1}{p_a(1-p_a)}$	$\frac{1}{(1-p_a)^2} - \frac{1}{p_a(1-p_a)}$	$\frac{2}{(1-p_a)^2}$

p_a is the population frequency of minor allele a .

- Distance between individuals i and j :

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

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

Extensive simulation study (work in progress) - methods

- KBAC (Liu and Leal 2010): each unique pattern of multi-locus genotypic configurations is tabulated, and the associated risk of disease for each configuration is modeled via a mixture distribution (estimated via non-parametr. kernel density)
- SKAT-O (Lee et al. 2012): optimized SKAT
- CLUSTER (Lin 2014): combines the association signals of variants that are more likely to be causal / incorporates the spatial information of variants (same direction or spatially close → optimal)
- IL-K (Ionita-Laza et al. 2012): scan-statistic approach that identifies clusters of rare disease variants / extension to Kulldorff scan statistic (deleterious effects); RBT (Ionita-Laza et al. 2011) max of 2 KBAC tests

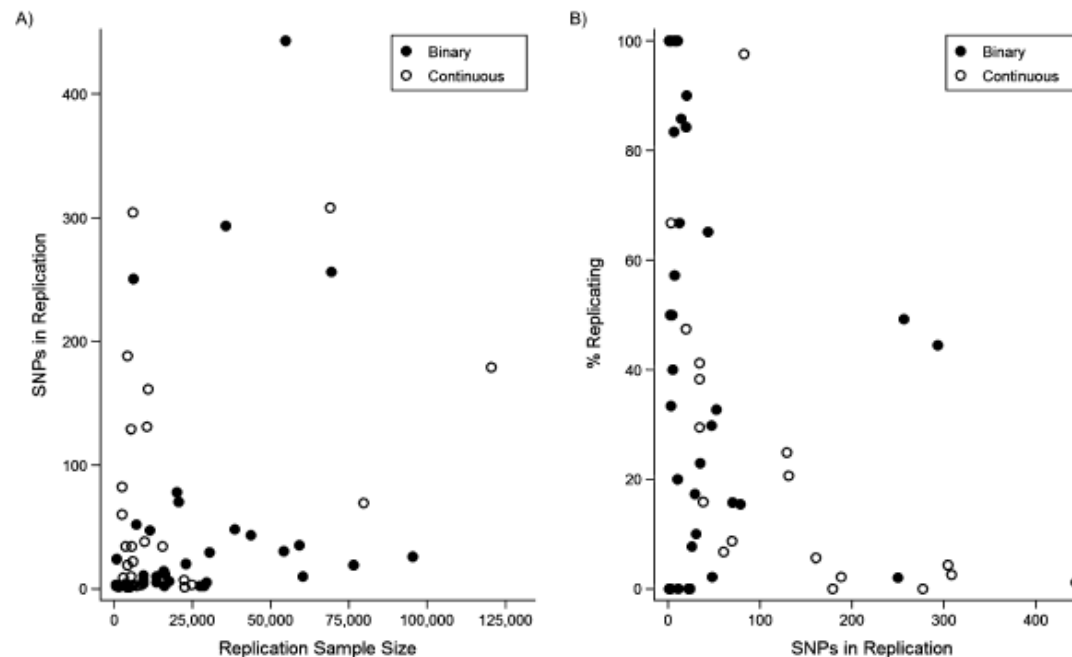
Extensive simulation study (work in progress) – real-life data

ExomeChip v1.1 data

- BBMRI-NL
 - UMCU (Utrecht, the Netherlands) – ~8000 individuals
 - UMCN (Nijmegen, the Netherlands) – ~1900 individuals
- BioMe Biobank data (Mount Sinai School of Medicine, USA)
 - ~10,000 individuals

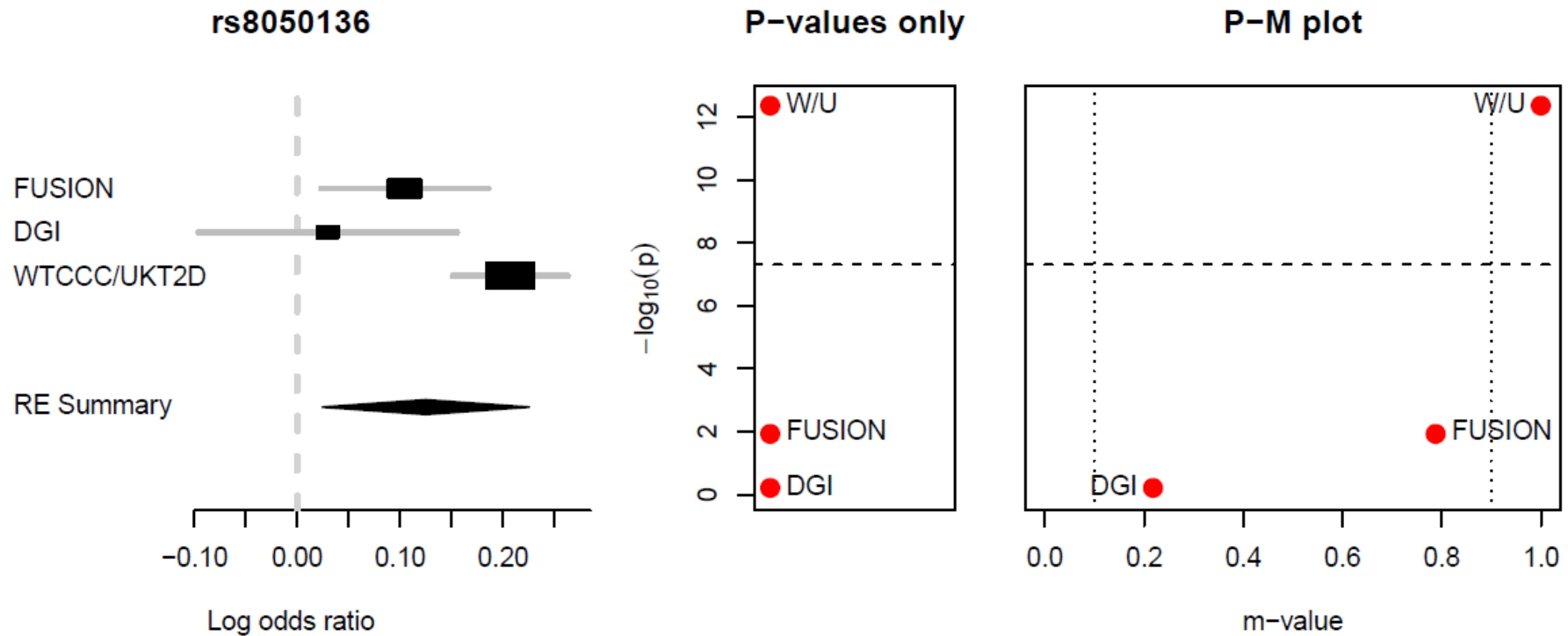
Imputation: a blessing or a curse? Part B: multiple testing?

Figure 3. A) Relation between the number of single nucleotide polymorphisms (SNPs) selected for replication and replication sample size. Unexpectedly, the number of SNPs selected for replication does not increase with the increase in the replication sample size. B) Relation between the proportion of replicated SNPs and the number of SNPs selected for replication. The proportion of SNPs successfully replicated increases with the decrease in the number of SNPs selected for replication.



(Gögele et al 2012)

Meta-analyses



(Han and Eskin 2012)

Other references

URLs:

- Kernel plot: http://www.ipam.ucla.edu/publications/ccstut/ccstut_9744.pdf
- Network plot: <http://www.nature.com/nmeth/journal/v11/n3/full/nmeth.2810.html>
- Components plot :
http://www.metabolomics.se/Courses/MVA/MVA%20in%20Omics_Handouts_Exercises_Solutions_Thu-Fri.pdf
- GWA related plots (levels of complexity): <http://genomesunzipped.org> – J Barrett