Methodological Aspects in Integromics

Kristel Van Steen, PhD[2] (*)

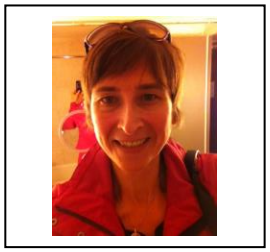
kristel.vansteen@ulg.ac.be

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

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

Outline

- **Integromics**
  - Definition and motivation
  - Building blocks / Bottom up versus top down?
- **Methodological challenges: a toy example**
  - Why GWAIs?
- **Towards a novel integrated analysis framework**
  - Based on MB-MDR
  - The need to deal with …
  - Link with integrative analyses
- **In conclusion**



BIO

X X

?

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

GIGA

Groupe Interdisciplinarie 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 González[2] and Sébastien Déjean[3]

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

- Joint analysis
- Challenging statistics
  - Regularized
  - Generalized

What's in a name?

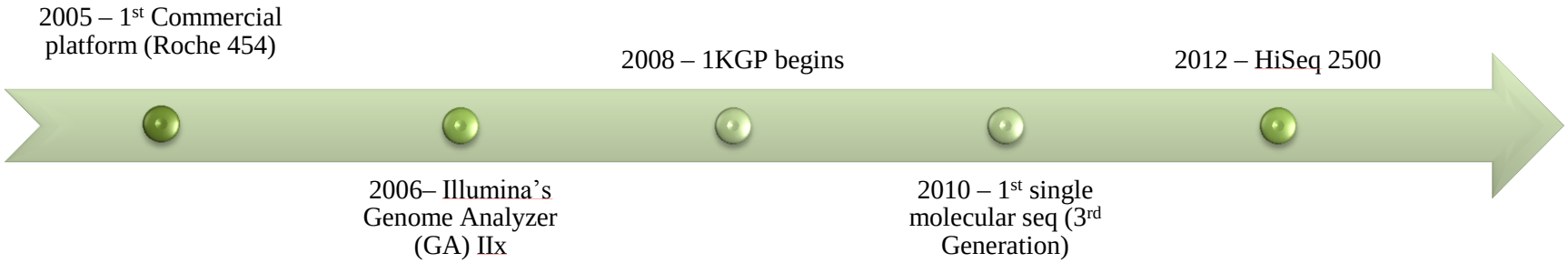
- **Data fusion** refers to fusing records on the same entity into a single file, and involves putting measures in place to detect and remove erroneous or conflicting data (Wang et al., 2014).
- Some definitions for "data fusion" use "data integration" in their definition. Although some data integration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



(Picture: Brooke L. Fridley - IGES 2014)

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.

DNA methylation ↔ Phenotype

DNA methylation ↔ Gene expression

Gene expression ↔ Phenotype

CANDIDATE LOCI

Is there room for data integration?

- Reasons for limited nr of initiatives in “truly integrating”?
  - 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

Data integration: Motivation and Opportunity

Perspectives on Data Integration in Human Complex Disease Analysis

Kristel Van Steen[1,2*] and Nuria Malats[3], on behalf of the COST Action BM1204 participants[4].

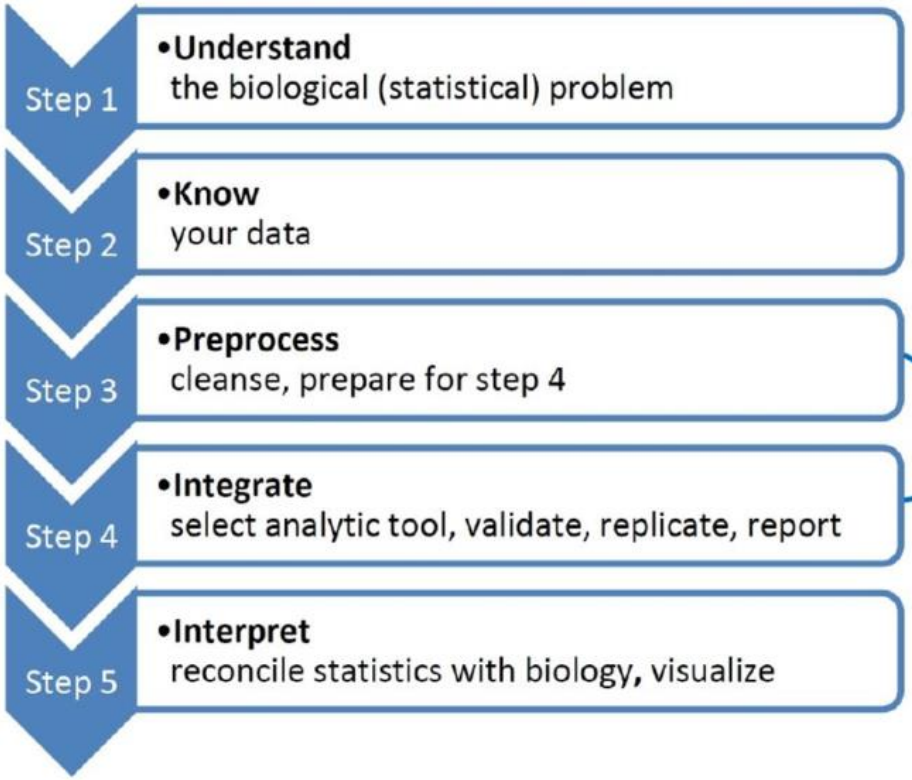
[1] Systems and Modeling Unit, Montefiore Institute, University of Liége, Liége, Belgium
[2] Bioinformatics and Modeling, GIGA-R, University of Liege, Avenue de l'Hôpital 1, Liége, Belgium
[3] Genetic & Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[4] http://www.cost.eu/domains_actions/bmbs/Actions/BM1204

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

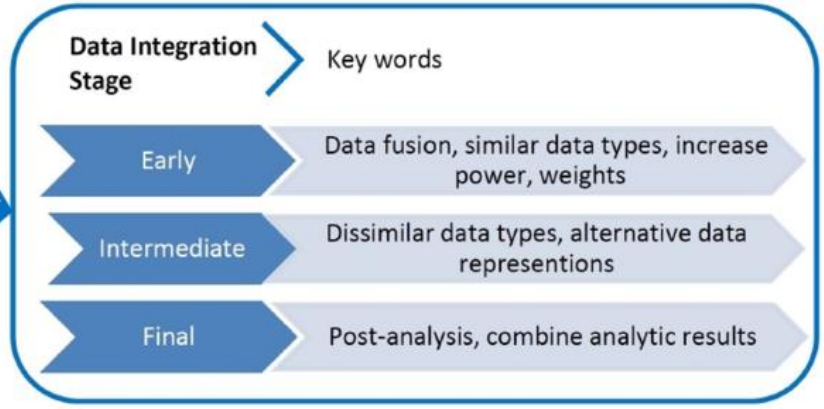
So we have the **motive,** and the **opportunity** …

(Boston Globe)

Building blocks of a "data integration" pipeline



(Hamid et al. 2009)



Systems information by integration (Joyce and Palsson 2006)

Genomics	Transcriptomics	Proteomics	Metabolomics	Protein–DNA interactions	Protein–protein interactions	Fluxomics	Phenomics
Genomics (sequence annotation)	• ORF validation • Regulatory element identification[74]	• SNP effect on protein activity or abundance	• Enzyme annotation	• Binding-site identification[75]	• Functional annotation[79]	• Functional annotation	• Functional annotation[71,103] • Biomarkers[125]
	Transcriptomics (microarray, SAGE)	• Protein: transcript correlation[20]	• Enzyme annotation[109]	• Gene-regulatory networks[76]	• Functional annotation[89] • Protein complex identification[82]		• Functional annotation[102]
		Proteomics (abundance, post-translational modification)	• Enzyme annotation[99]	• Regulatory complex identification	• Differential complex formation	• Enzyme capacity	• Functional annotation
			Metabolomics (metabolite abundance)	• Metabolic-transcriptional response		• Metabolic pathway bottlenecks	• Metabolic flexibility • Metabolic engineering[109]
				Protein–DNA interactions (ChIP–chip)	• Signalling cascades[89,102]		• Dynamic network responses[84]
					Protein–protein interactions (yeast 2H, coAP–MS)		• Pathway identification activity[89]
						Fluxomics (isotopic tracing)	• Metabolic engineering
							Phenomics (phenotype arrays, RNAi screens, synthetic lethals)

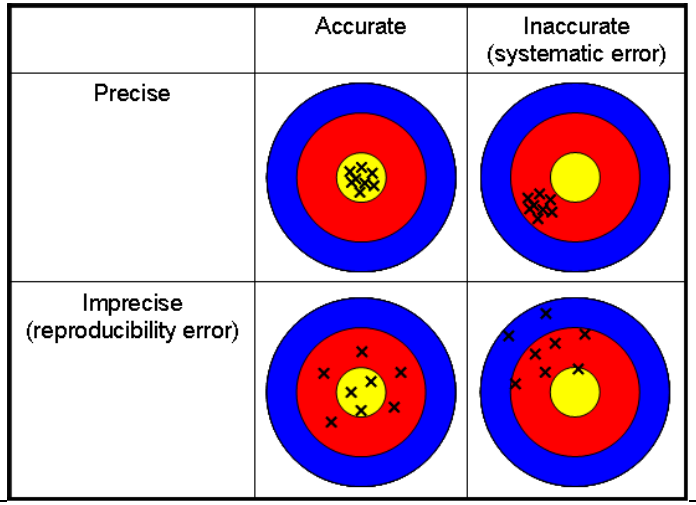
Step 1

- Formulating the biological (statistical) problem

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)
- EDA / Weighting: quality + information



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

Building blocks of a "data integration" pipeline

Step 3

- Data preprocessing

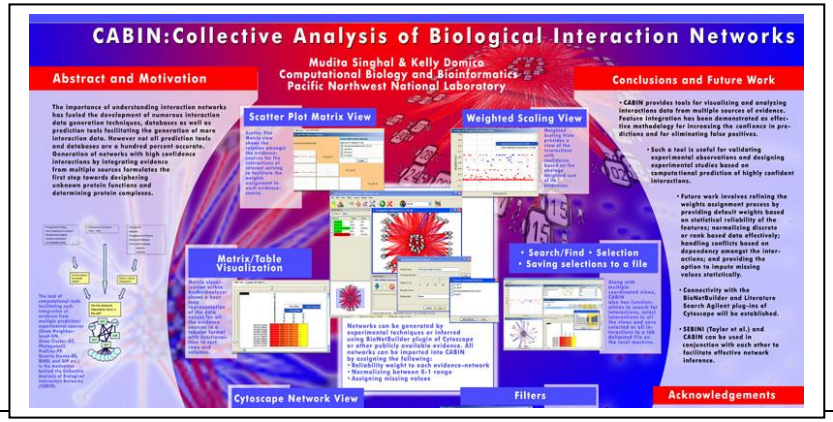
- Approaches for preprocessing vary depending on the type and nature of data:
  - e.g., arrays: background correction, normalization, quality assessment, which may differ from one platform to another
- Data (pre)processing can be done **at any step of the data integration** process:
  - e.g., at the **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 and providing a "functional explanation":
  - (Experimental) validation helps in the "understanding", but becomes cumbersome in integromics settings
- There is a huge challenge in visualizing the steps of and the results from an integrated analysis: **visual analytics**



Building blocks of a "data integration" pipeline

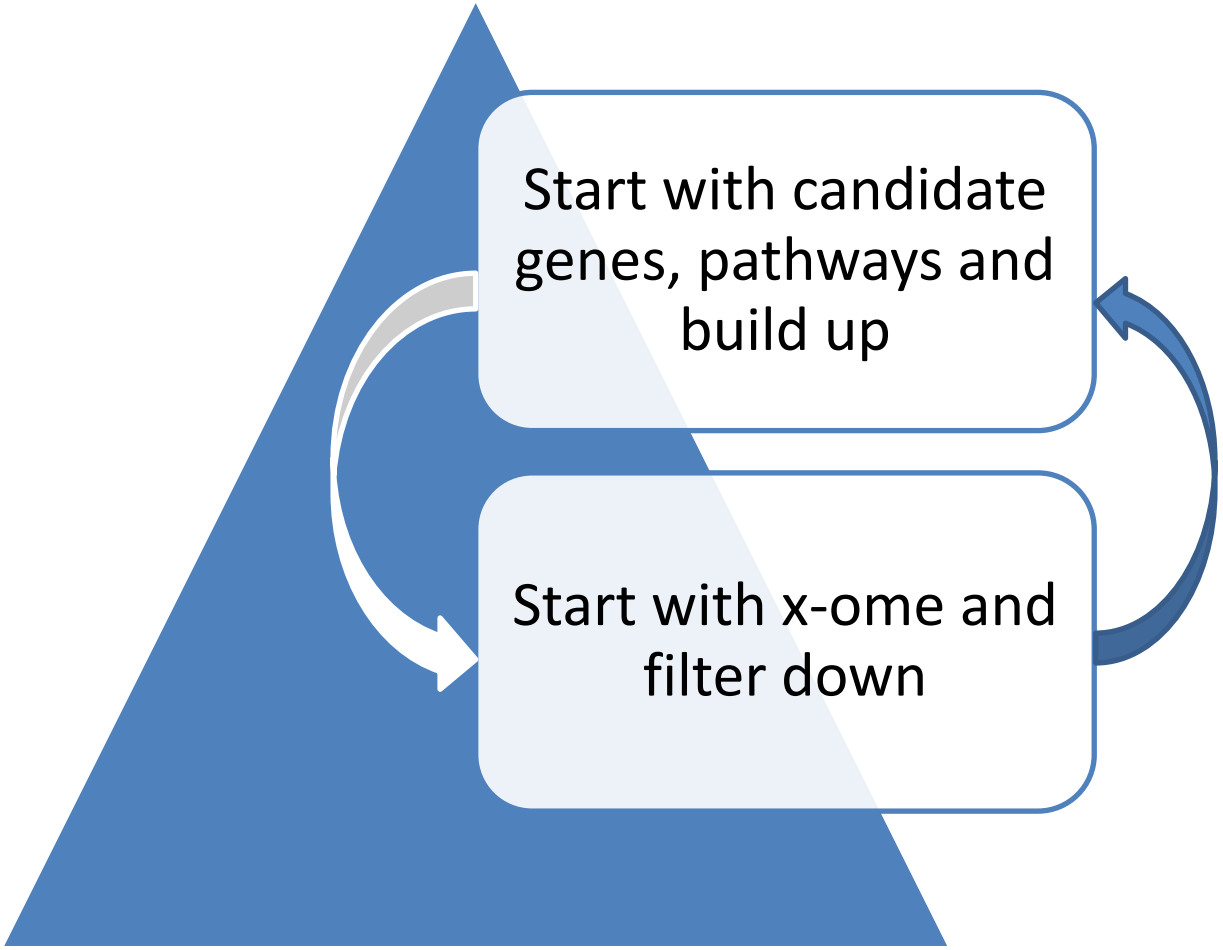
Step 5

- Interpretation

- Post-linking to several external biological data bases. Beware of "black-box" data base linking …. There is a need to:
  - allow for uncertainty involved in the data source entries
  - acknowledge the complementary characteristics of each of the available data sources
  - assess and incorporate "optimal" scoring systems to accumulate evidence from these data bases
  - allow for different assignment strategies (e.g., from genetic variants to genes)

Integrative analytics

Top down versus bottom up

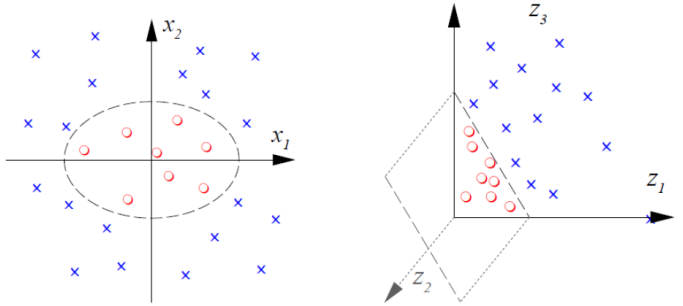


Integrative analytics

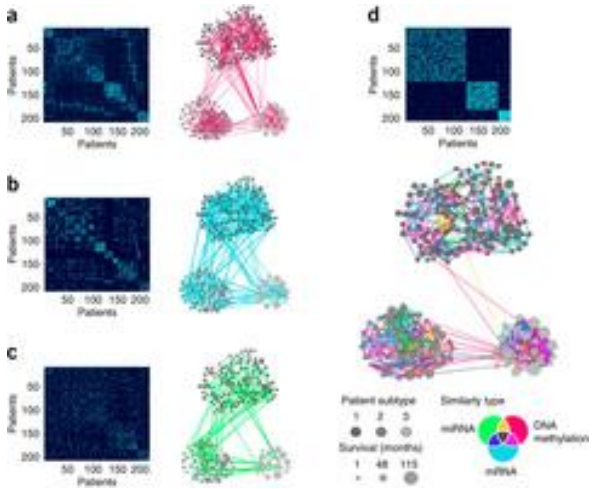
Crude division:

Kernels

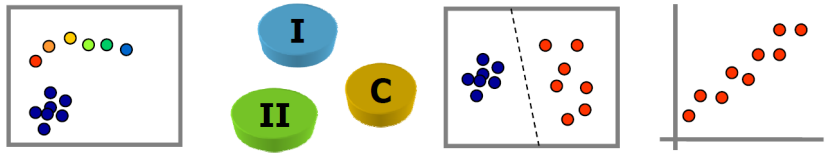
$$\Phi : \mathbf{R}^2 \rightarrow \mathbf{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$



Networks

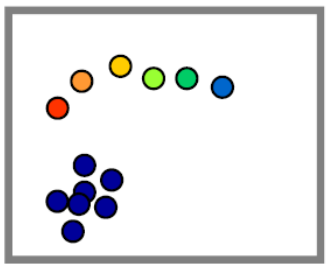


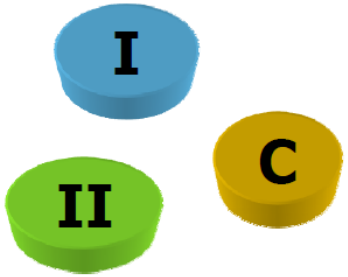
Components

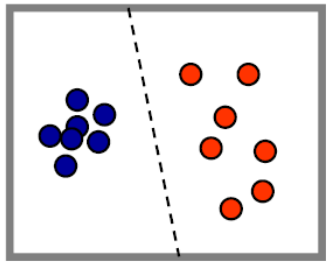


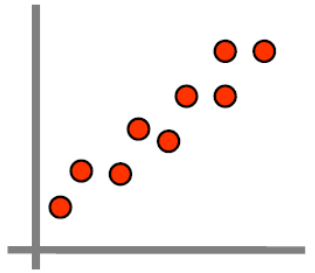
Overview	Classification	Discrimination	Regression
Trends Outliers Quality Control Biological Diversity Patient Monitoring	Pattern Recognition Diagnostics Healthy/Diseased Toxicity mechanisms Disease progression	Discriminating between groups Biomarker candidates Comparing studies or instrumentation	Comparing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

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)

Taking baby steps: starting from GWAs

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,
[1]Departn
Medical
[2]Departn
and Wor

OPEN ACCESS Freely available online
PLOS ONE

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

Li-Chung Chuang[1,2], Chung-Feng Kao[1], Wei-Liang Shih[1], 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 ARTIC
published: 31 May
doi: 10.3389/fgene.2013.0

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

Lin Li[1†], Michael Kabesch[2], Emmanuelle Bouzigon[3,4], Florence Demenais[3,4], Martin Farrall[5], Miriam F. Moffatt[6], Xihong Lin[1] 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

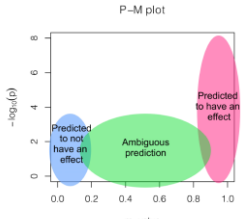
Université de Liège

Methodological challenges

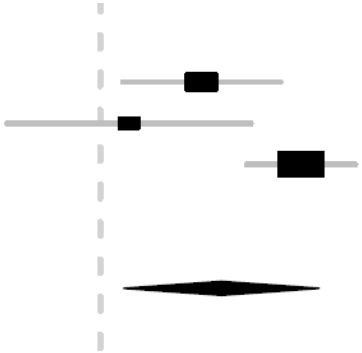
- a toy example -

Methodological aspects: scaling up from GWAs to GWAIs

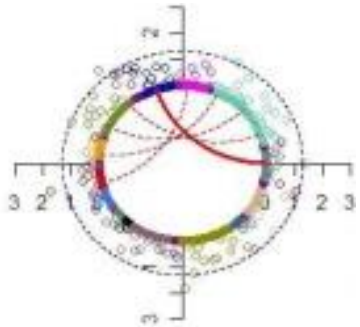
4 meta-GWAIs



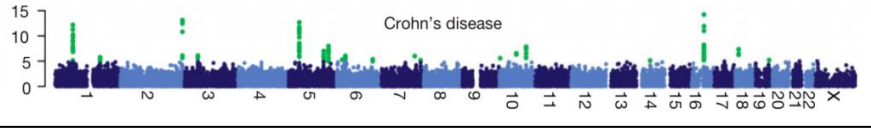
2 meta-GWA



3 GWAIs



1 GWAs



Statistical modeling

IT

4

1

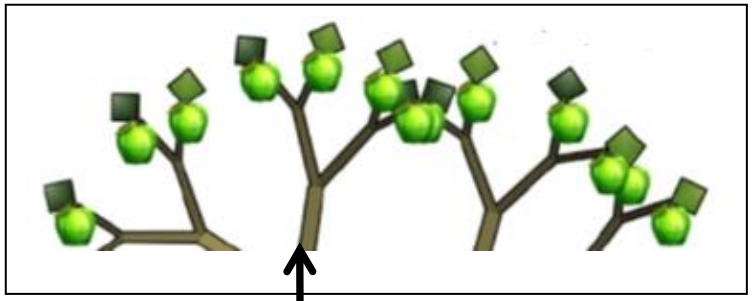
Complexity / Burden

Storage and computational efficiency

IT and Biological complexity

~~Parsimony, Model assumptions~~

Methodological aspects: scaling up from GWAs to GWAIs

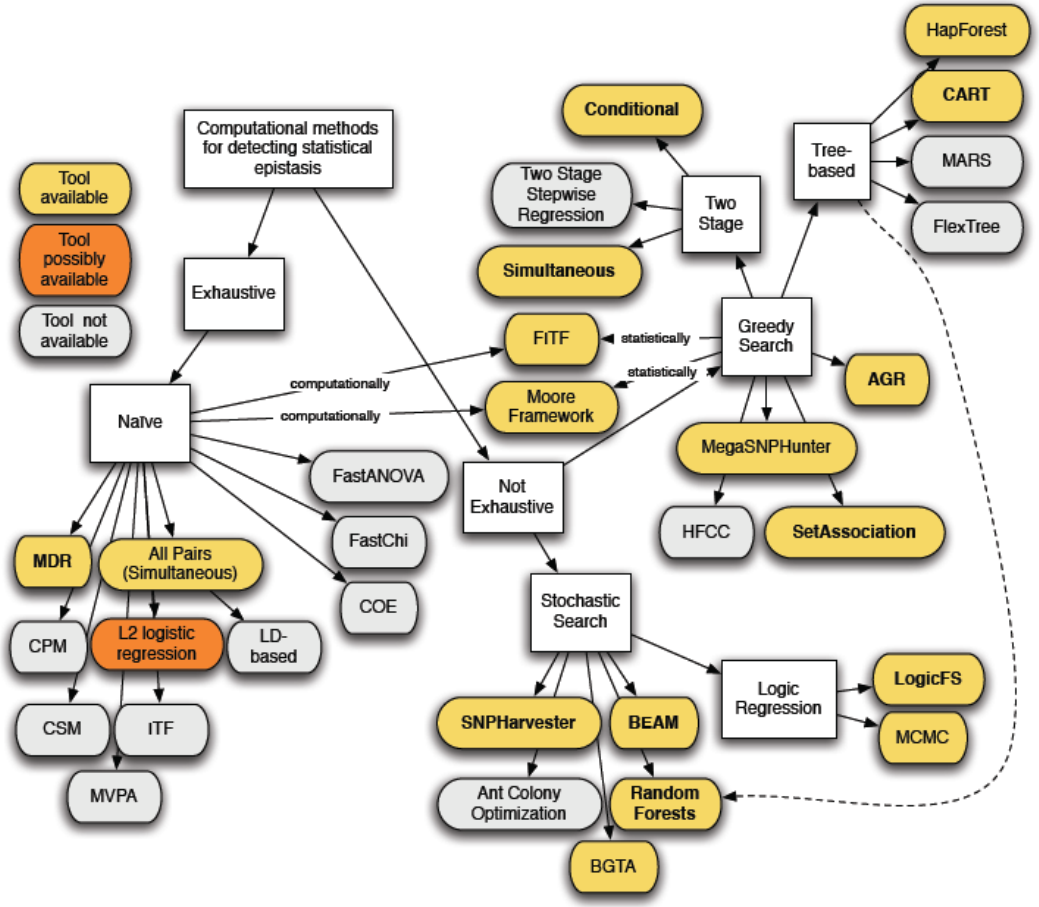


4

3

IT and Biological complexity

~~Parsimony, Model assumptions~~



(Kilpatrick 2009)

Towards a novel integrated framework "genomic MB-MDR"

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

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

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	$\approx$ 270 days	25 hours 12 min	$\approx$ 290 days	$\approx$ 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

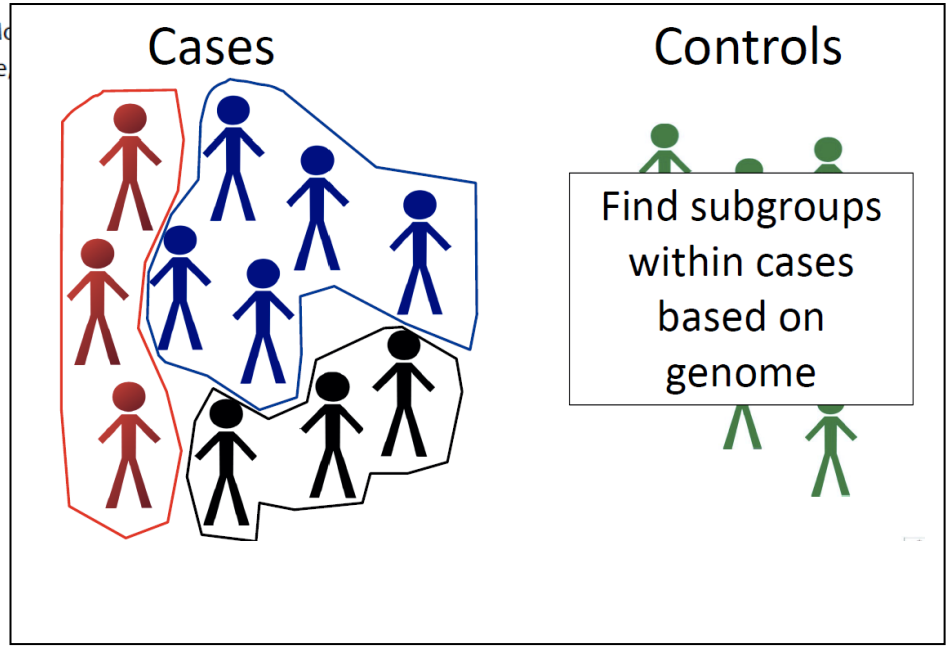
OPEN ACCESS Freely available online

PLoS one

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

Isabelle Cleynen[1]*, Jestinah M. Mahachie John[2,3], Liesbet Henckaerts[4], Wouter Van Moerkercke[1], Paul Rutgeerts[1], Kristel Van Steen[2,3], Severine Vermeire[1]

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



Detecting structure in patients: subphenotyping

OPEN ACCESS Freely available online PLOS ONE

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)

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

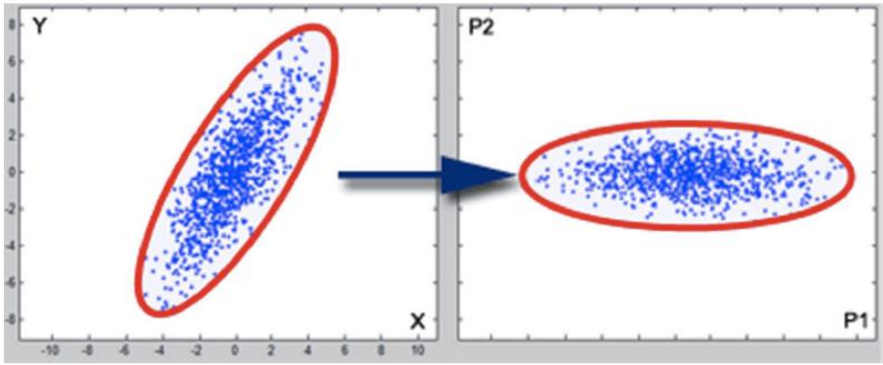
Detecting structure in patients: subphenotyping

		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					

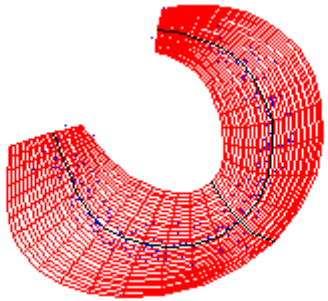
- Adjusted Rand Index between latent class analysis (LCA), PAM clustering and hierarchical clustering using Ward linkage and squared Eucl. distance (using population unadjusted and adjusted SNP data)
- Clusters ~ Clinical features: focus on populations with a similar genetic background

Detecting structure in patients / populations

- Orthogonal linear transformation of the data



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



Meta-analysis

Meta-GWAIs

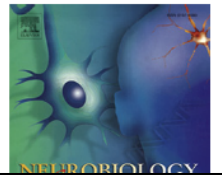
ARTICLE IN PRESS

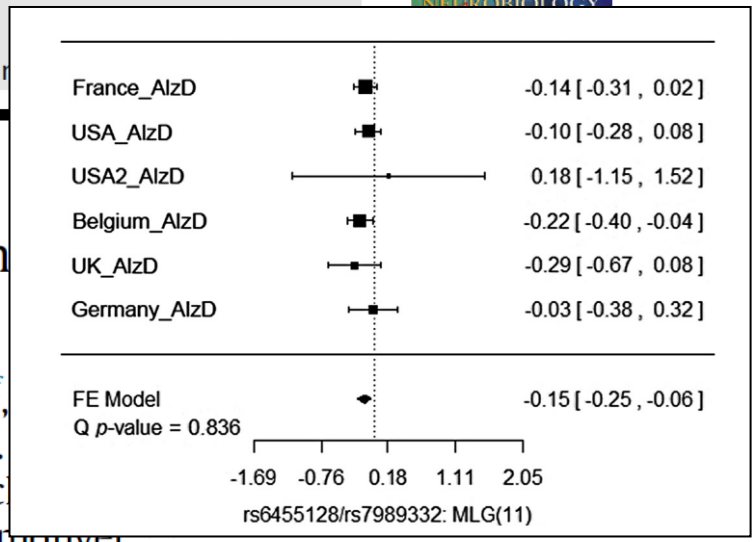
Neurobiology of Aging xxx (2014) 1–8

Contents lists available at ScienceDirect

Neurobiology of Aging

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

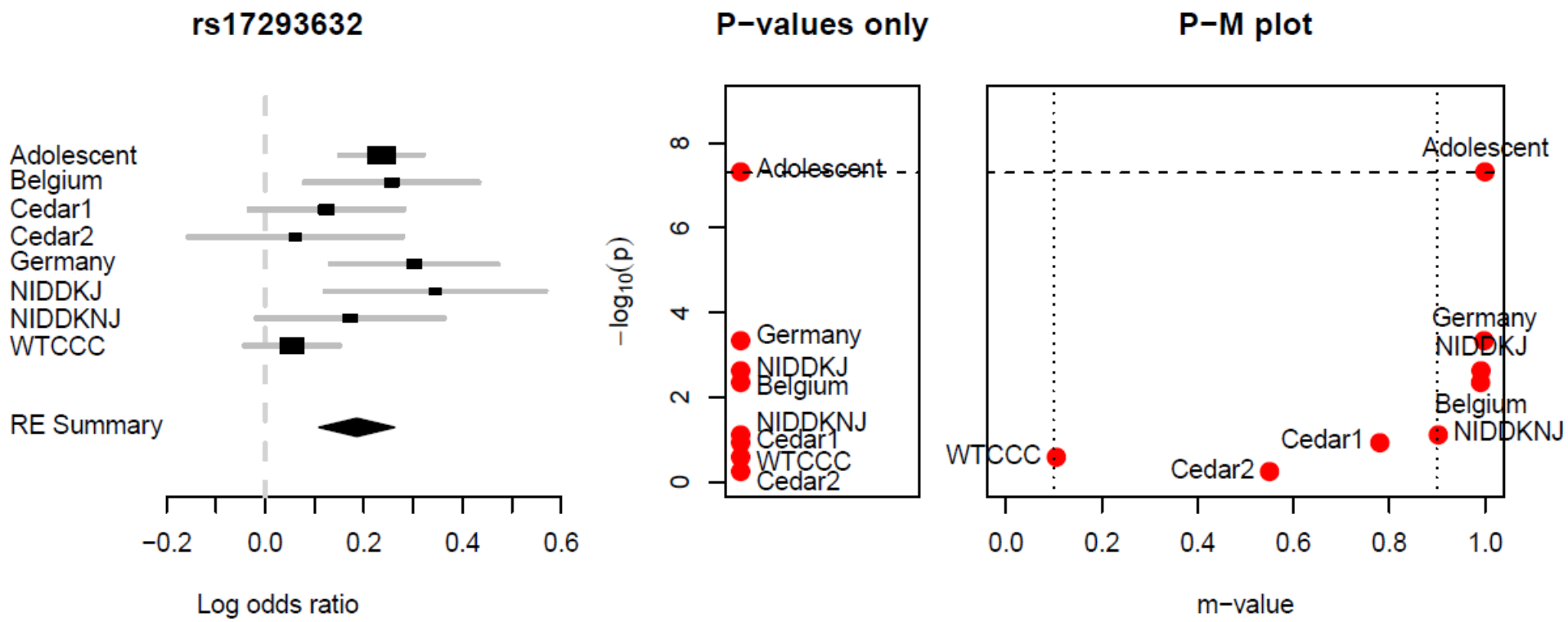




Genome-wide association interaction analysis for Alzh disease☆

Elena S. Gusareva [a,b,*], Minerva M. Carrasquillo [c], Céline Bellenguez [d,e,f], Samuel Colon [c], Neill R. Graff-Radford [i], Ronald C. Petersen [j], Dennis W. Jestinah M. Mahachie John [a,b], Kyrylo Bessonov [a,b], Christine Van Broec the GERAD1 Consortium[1], Denise Harold [k], Julie Williams [k], Philippe Amouyel Kristel Sleegers [g,h], Nilüfer Ertekin-Taner [c,i], Jean-Charles Lambert [d,e,f], Kristel Van Steen [a,b]

Dealing with increased heterogeneity



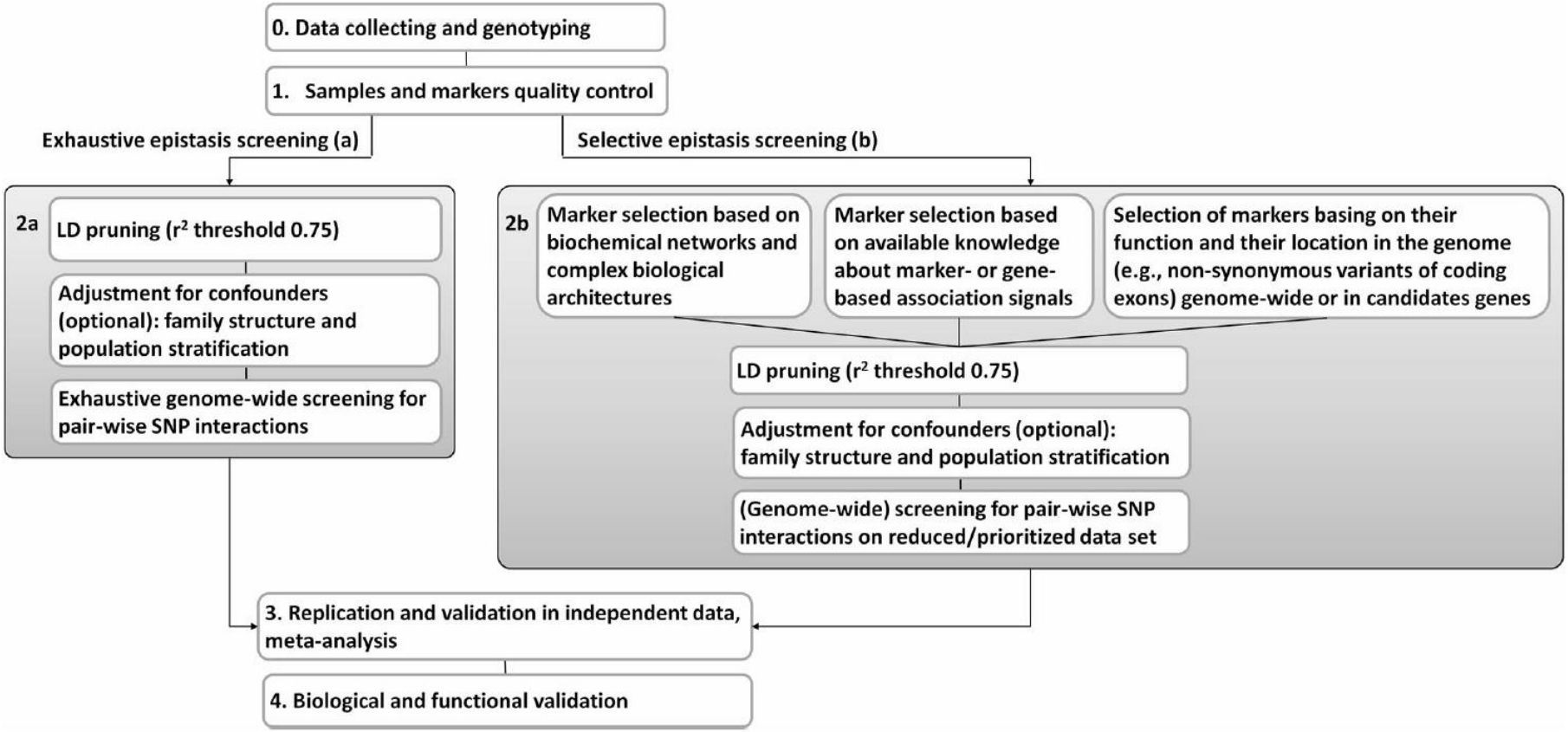
(Han and Eskin 2012)

Meta-GWAI studies

- Given the availability of a comprehensive meta-analysis toolbox, it may be surprising that hardly any meta-GWAIs have been published as the core topic of the publication.
- This may in part be explained by the absence of strict guidelines or best practices for epistasis analysis, and the fact that new epistasis screening approaches arise every day.
- Additional complicating factors include:
  - Traditional meta-analysis methods in genetic association studies usually assume a specific genetic model of action to summarize the effect of genetic markers on a phenotype.
  - GWA imputation strategies ensure that different data sets are made comparable, but most be revised in the context of GWAIs.

Interpretation

Statistical versus biological epistasis

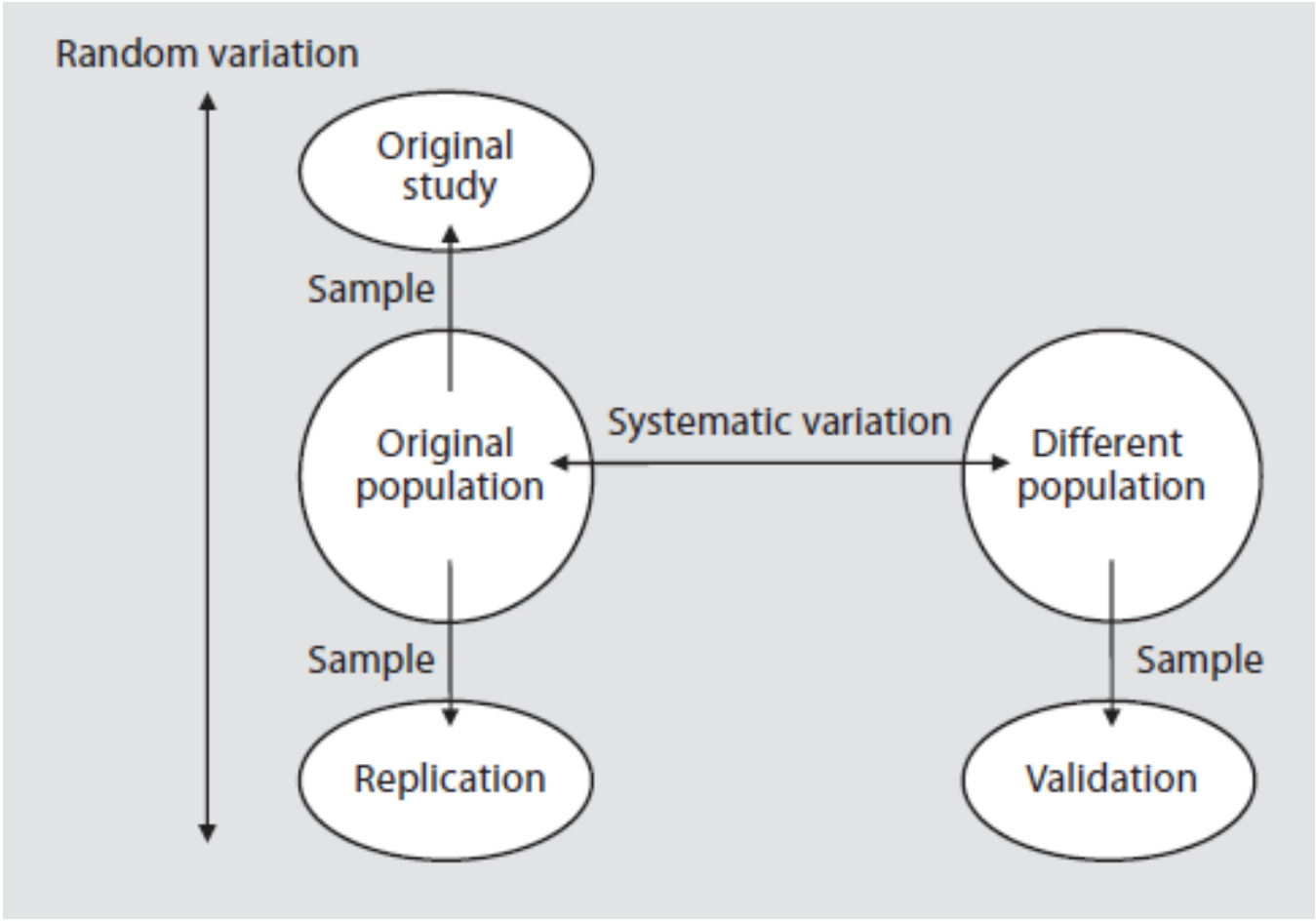


- Protocol for GWAIs (analytic blocks are highlighted)

(Gusareva et al. 2014)

Replication and validation

Difference between “replication” and “validation”



(Igl et al. 2009)

Replication using tagSNPs (often no functional consequence)

- "Due to variation in allele frequency and underlying linkage disequilibrium patterns (influenced by imputation …) between two datasets, it is highly unlikely that the same combination of tagSNPs would be associated in the same statistical interaction 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)

Available "knowledge" about epistasis: Alzheimer's disease

Gene	Gene name	Function	Location	Epistatic SNPs	Main effect for AlzD	Population (N cases/N controls)	Reference
INS	Insulin	Glucose metabolism	11p15.5	rs689	no	Germans (104/123)	Brune et al., 2003
PPARA	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs1800206	yes	Northern Europeans (336/2426)	Kölsch et al., 2012
IL1A	Interleukin 1 alfa	Inflammatory cytokine	2q13	rs3783550	no	Northern Europeans (336/2426)	Heun et al., 2012
PPARA	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs1800206	yes		
IL1B	Interleukin 1 beta	Inflammatory cytokine	2q13	rs16944	no	Northern Europeans (336/2426)	Heun et al., 2012
PPARA	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs1800206	yes		
IL10	Interleukin 10	Inflammatory cytokine	1q32.1	rs1800896	yes	Northern Europeans (336/2426)	Heun et al., 2012
PPARA	Peroxisome proliferator-activated receptor alpha	Glucose and lipid metabolism	22q13.31	rs4253766	no		
IL1A	Interleukin 1 alfa	Inflammatory cytokine	2q13	rs1800587	no	Northern Europeans (336/2426)	Combarros et al., 2010
DBH	b-Hydroxylase	Onverts dopamine to norepinephrine in the synaptic vesicles of postganglionic sympathetic neurons	9q34.2	rs1611115	yes		
TF	Transferrin	Iron metabolism	3q22.1	rs1049296	no	UK (191/269)	Robson et al., 2004
HFE	Hemochromatosis		6p22.2	rs1800562	yes	Caucasians USA (1166/1404)	Kauwe et al., 2010
						North Europeans (336/2426)	Lehmann et al., 2012
TF	Transferrin	Iron metabolism	3q22.1	rs1130459	no	North Europeans (336/2426)	Lehmann et al., 2012
HFE	Hemochromatosis		6p22.2	rs1799945	yes		
MTHFR	Methylenetetrahydrofolate reductase	Homocysteine metabolism useful for normal brain functioning	1p36.22	rs1801131	yes	Indians (80/120)	Mansoori et al., 2012
IL6	Interleukin 6	Pro-inflammatory cytokine	7p15.3	rs1800795	no		
IL10	Interleukin 10	Limit inflammation in the brain	1q32.1	rs1800871	yes	North Spains (232/191) ,	Infante et al., 2004
IL6	Interleukin 6	Pro-inflammatory cytokine	7p15.3	rs2069837	yes	North Europeans (336/2426)	Combarros et al., 2009
ABCA1	ATP-binding cassette transporter A1	Intracellular cholesterol transport and maintance of cell cholesterol balance	9q31.1	rs2422493	no	Spanish (631/731)	Rodríguez-Rodríguez et al., 2010
NPC1	Niemann-Pick C1		18q11.2	rs18050810 rs4800488 rs2236707 rs2510344	no		

LRP1	low density lipoprotein receptor-related protein 1	Neuronal uptake of cholesterol	12q13.3	rs1799986	no	Spanish (246/237)	Vázquez-Higuera et al., 2009
MAPT	Microtubule-associated protein tau		17q21.33	rs2471738	no		
GSK3B	Glycogen synthase kinase-3 beta	Abnormal hyperphosphorylation of tau, neuronal uptake of cholesterol	3q13.33	rs334558	no	Spanish (246/237)	Vázquez-Higuera et al., 2009
CDK5R1	Cyclindependent kinase 5		17q11.2	rs735555			
NR1H2	Liver X receptor beta	Cholesterol metabolism	19q13.33	rs1052533 rs1405655	no	Spanish (414/442)	Infante et al., 2010
HMOX1	Heme oxygenase-1		22q12.3	rs2071746			

Different levels

- Genetic marker
- Broader locus
- Gene
- Window including either one of the previous
- Pathway

Candidate gene pairs: Alzheimer's disease (Elena Gusareva)

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

MTHFR	IL10	IL1A	IL1B	TF	HFE	IL6	ABCA1	DBH	INS	LRP1	CDK5R1	MAPT	NPC1	NR1H2	HMOX1	PPARA	
	+	ns	+	+	+	+	+	+	+	+	ns	+	+	+	ns	+	MTHFR
		+	+	+	ns	ns	+	+	ns	+	ns	+	ns	ns	+	+	IL10
			ns	+	+	+	+	ns	+	ns	ns	+	ns	ns	ns	+	IL1A
				+	ns	ns	+	ns	ns	+	ns	+	+	ns	ns	ns	IL1B
					+	+	+	+	ns	+	ns	+	+	+	+	+	TF
						+	+	ns	+	+	ns	+	+	+	ns	+	HFE
							+	+	ns	ns	ns	+	+	+	+	+	IL6
								+	+	+	ns	+	+	+	+	+	ABCA1
									+	+	ns	+	+	ns	+	+	DBH
										ns	ns	+	ns	ns	+	+	INS
											ns	+	ns	ns	+	+	LRP1
												ns	ns	ns	ns	ns	CDK5R1
													+	ns	+	+	MAPT
														ns	ns	+	NPC1
															ns	ns	NR1H2
																+	HMOX1
																	PPARA

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

(the marginal *p*-value < 0.05 for the MB-MDR$_{2D}$ analysis)

Replication is highlighted by green; no replication is highlighted by red.

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)

No replication without a consensus -- about the methodology

- Multiple testing handling
- Multi-stage designs incl marker selection
- Meta-analysis
- LD between markers and long-distance between-marker associations

  Population stratification assessments by –omics
- The importance of epistasis and non-linear relationships in

  population genetics
- Within- and between-gene architectures
- Missing data handling (coarsening, …)

Unable to replicate is a bad thing?

OPEN ACCESS Freely available online

PLoS one

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

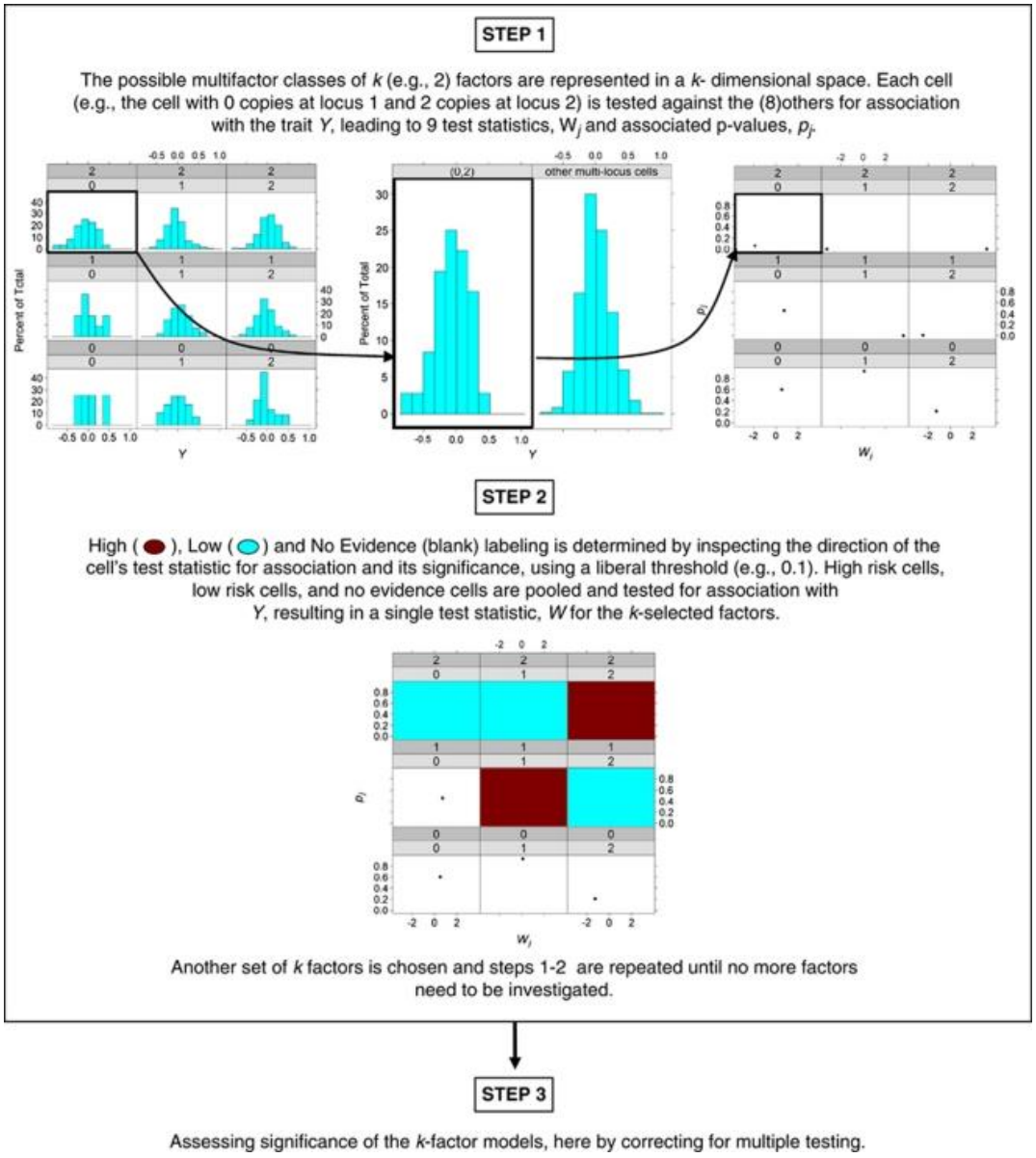
Casey S. Greene[1], Nadia M. Penrod[1], Scott M. Williams[2], Jason H. Moore[1,2,3,4,5,6]*

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

Combining it all: genomic MB-MDR



MB-MDR (Calle 2008**, BIO3)**

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

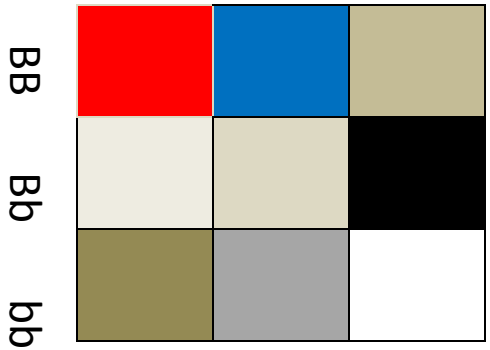
Step 2: Label and reduce dimensionality by pooling equally-labelled cells.

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

Gene-based or set-based testing

MB-MDR 2D

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



1 dimension = 1 genetic maker

(grouping based on 2-locus genotypes)

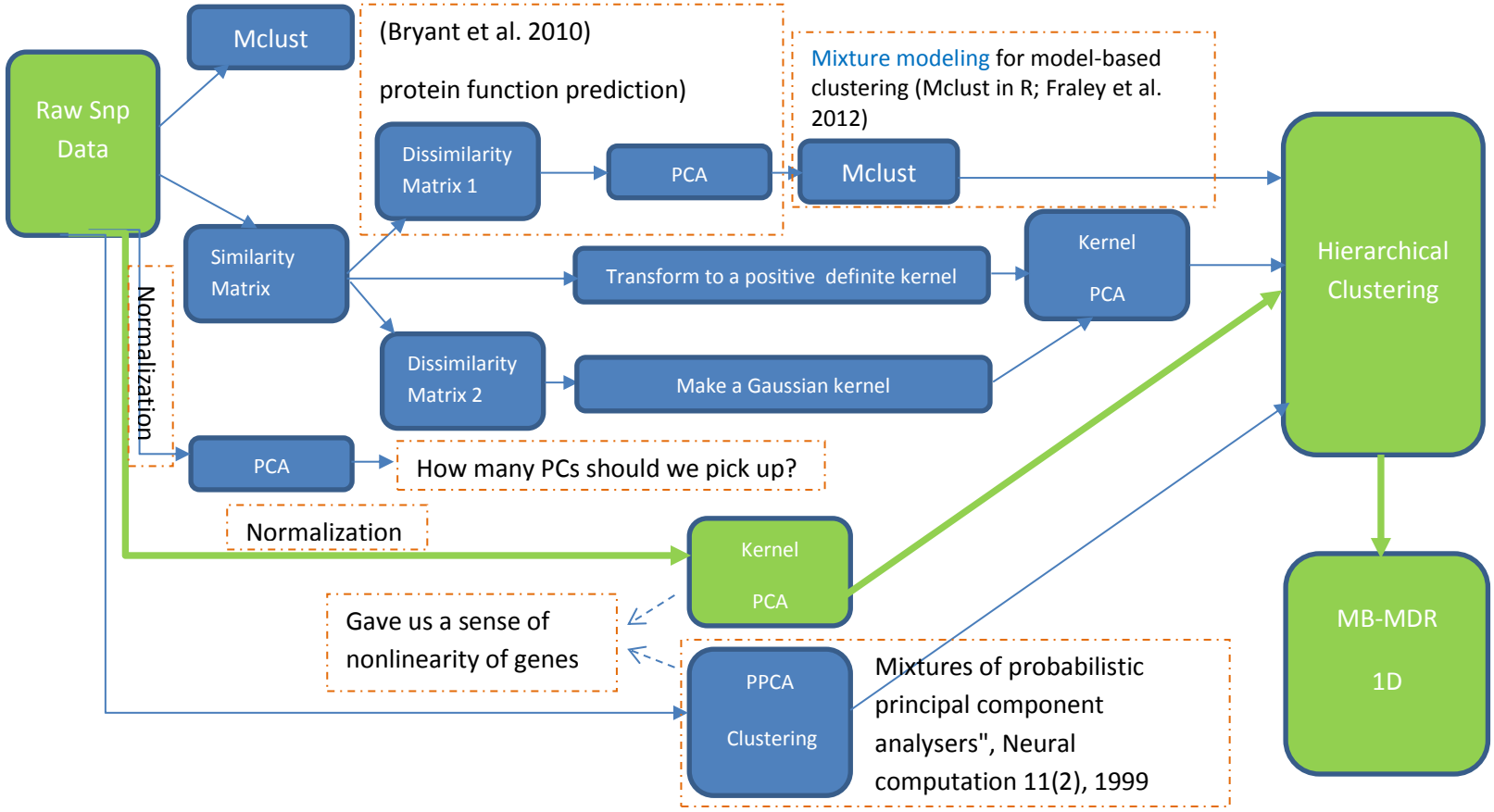
Genomic MB-MDR 1D

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

1 dimension = 1 ROI

(grouping on features mapped to the ROI)

Genomic MB-MDR: step 1 (descriptor filtering) **+ step 2** (clustering)



(adapted slide from Fouladi 2014)

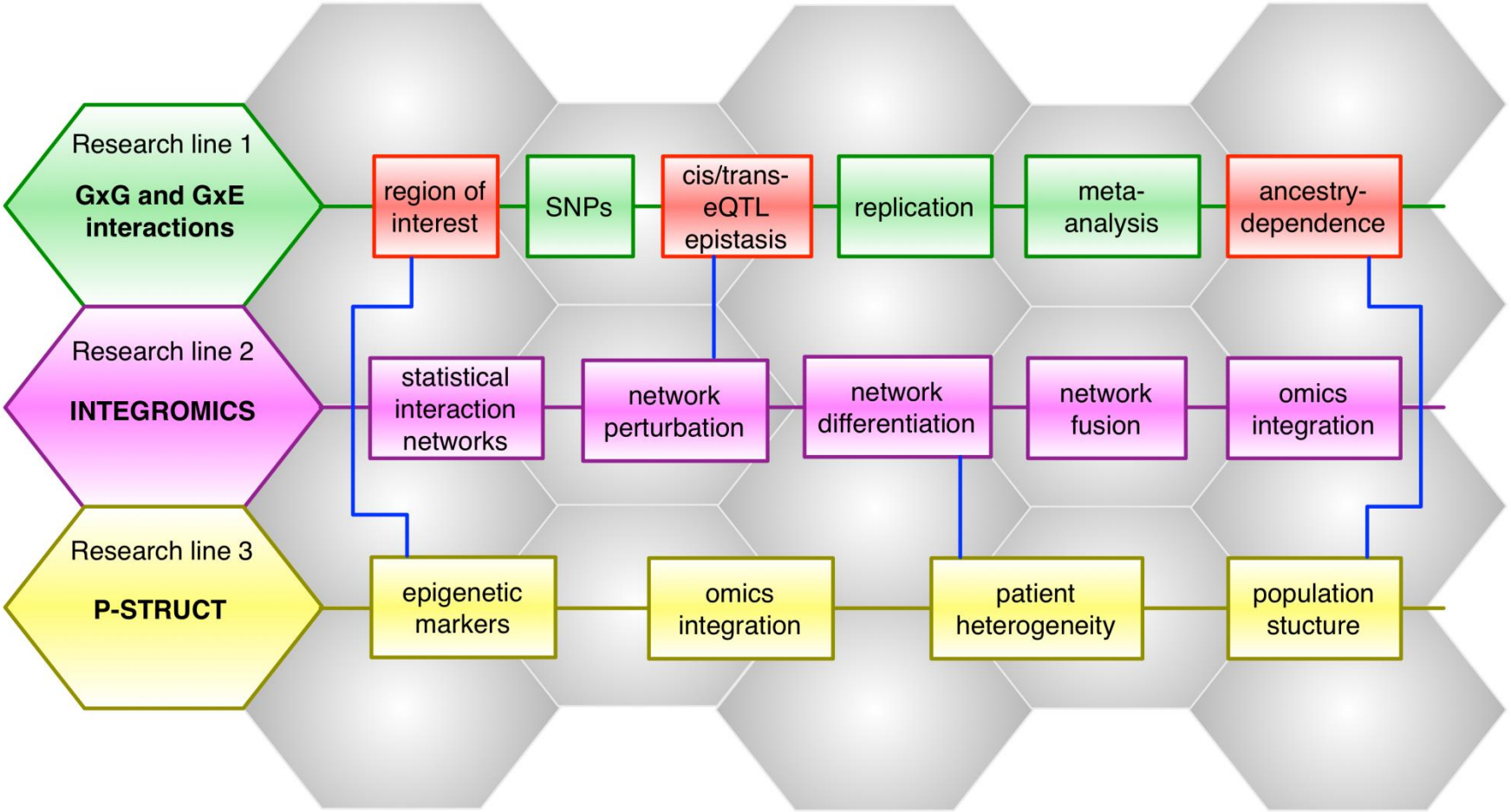
Genomic MB-MDR

- Interpretability depends on the quality of the clusters
  - Good clusters should generalize well. The clusters should continue to describe new observations of the same features
  - Good clusters should generalize to new features

  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

- Re-think the MB-MDR default options (different contexts!)

Starting from GWAIs - Bio3's research lines



Integration to enhance biological network construction

- 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.

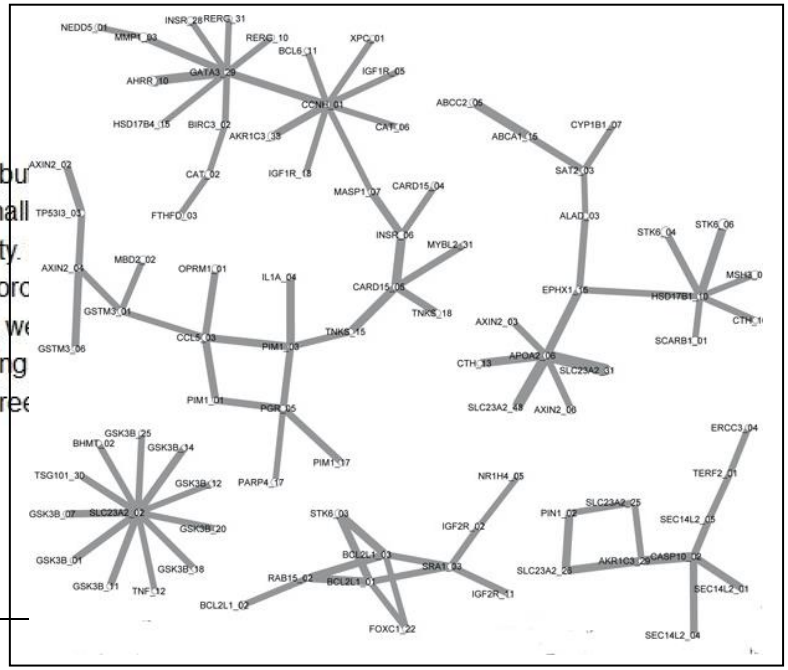
Statistical epistasis networks reduce the computational complexity of searching three-locus genetic models.

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

Author information

Abstract

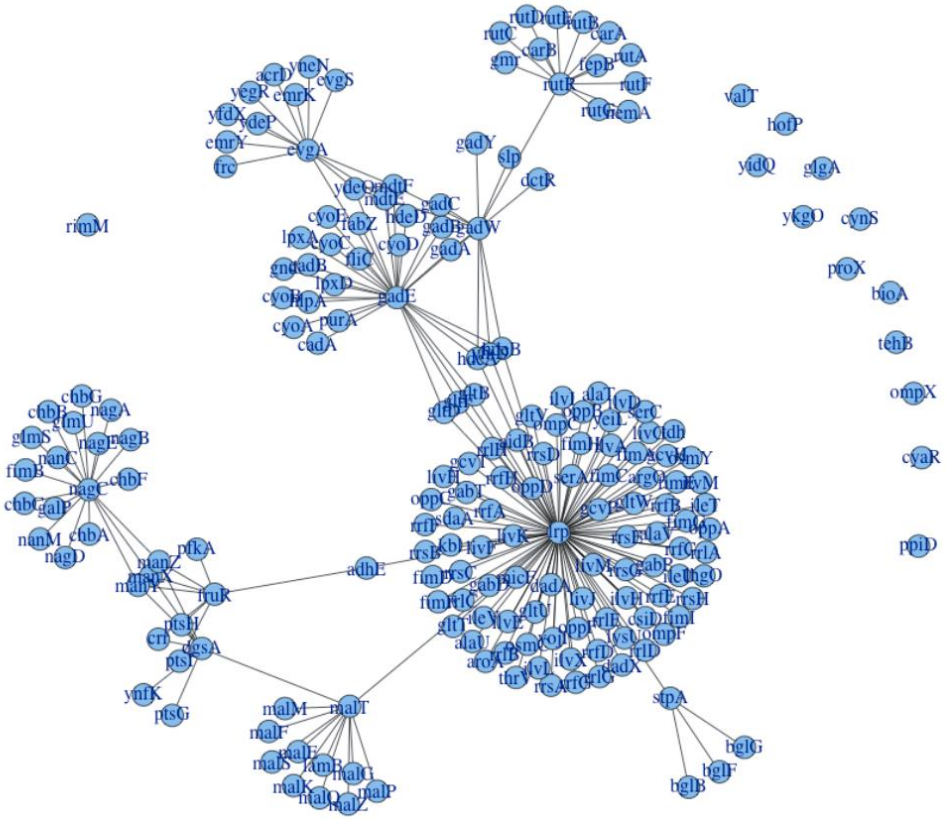
The rapid development of sequencing technologies makes thousands to millions of genetic attribu
traits. Searching this enormous high-dimensional data space imposes a great computational chall
network-based approach to supervise the search for three-locus models of disease susceptibility.
strong pairwise 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.



Integration to enhance biological network construction

- Regression-based frameworks

200 nodes
212 edges

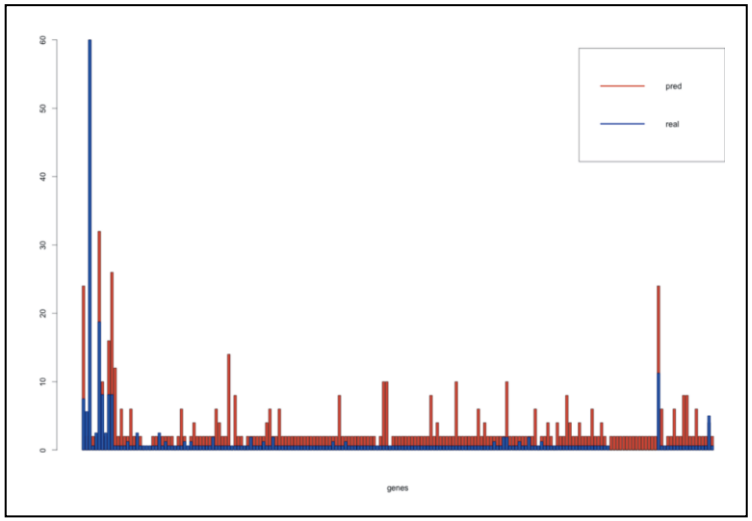


(GeneNetWeaver synthetic **gold standard** network based on transcription factor network (TFN) of E.coli)

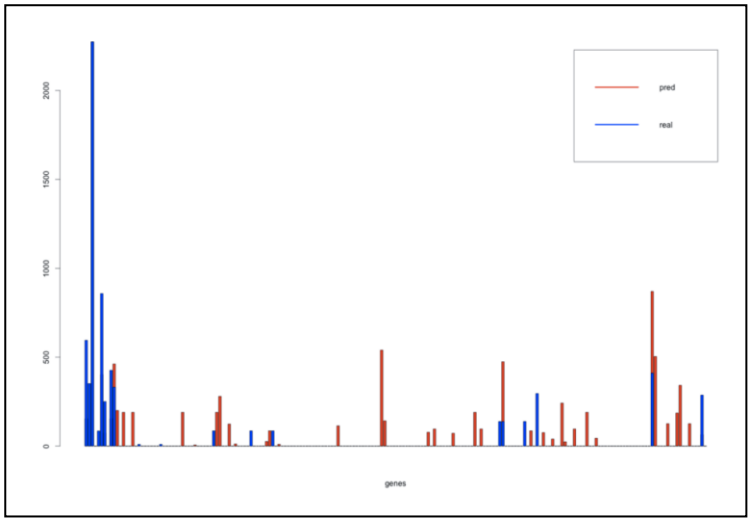
Regression2Net (Francesco Gadaleta)

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

Color legend: Predicted network ~ Gold Standard



Degree correlation = 0.86

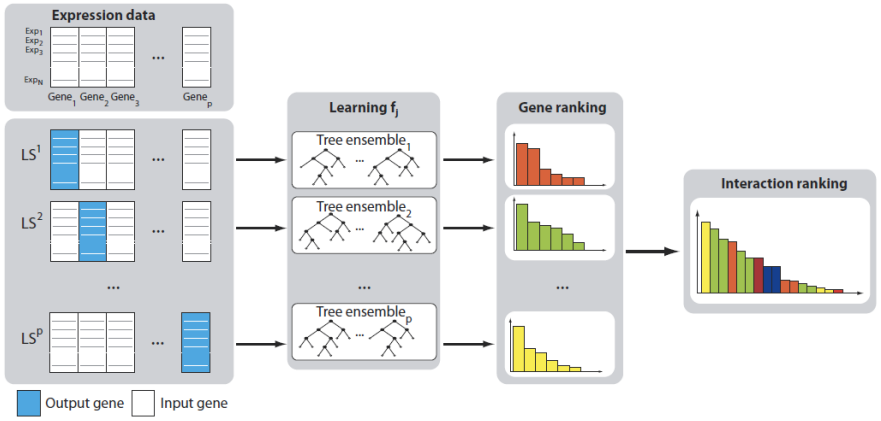


Betweenness correlation = 0.83

Foresting in Integromics Inference (Kirill Bessonov)

- Ensemble methods: e.g; GENIE suite of Vân Anh et al.
- Alternatively, use "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

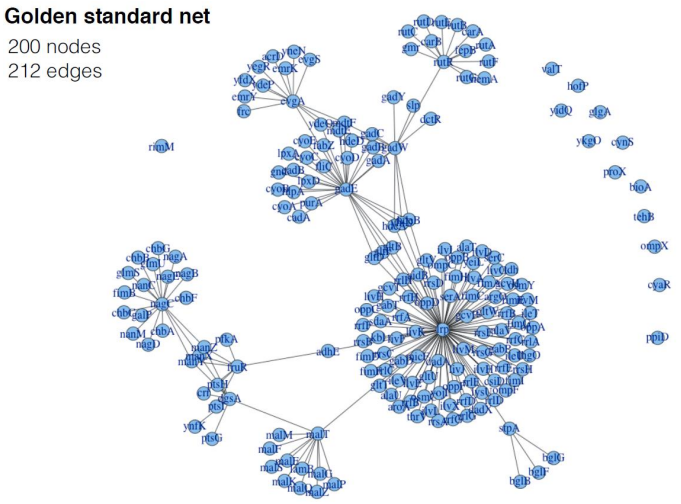
In conclusion

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.
- Toy example on GWAIs can be instrumental in understanding what matters in the context of a complex “integromics” world

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)



Integromics: Mission ..possible?

(Mission Impossible @ google)

Acknowledgement

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

Machine Learning for Personalized Medicine

fnrs LA LIBERTÉ DE CHERCHER

EU PANCREAS

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

Université de Liège

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, 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!]**

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
- High hanging fruit plot – Moore and Williams 2009