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?

• **Methodological challenges: a toy example**
  – Why GWAI$s$?

• **Towards a novel integrated analysis framework**
  – Based on MB-MDR
  – The need to deal with ...
  – Link with integrative analyses

• **In conclusion**
Bio³: Biostatistics – Biomedicine - Bioinformatics
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

- 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.

![Diagram showing DNA methylation, Phenotype, Gene expression relationships]
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\textsuperscript{1,2*} and Nuria Malats\textsuperscript{3}, on behalf of the COST Action BM1204 participants\textsuperscript{4}.

\textsuperscript{1} Systems and Modeling Unit, Montefiore Institute, University of Liège, Liège, Belgium
\textsuperscript{2} Bioinformatics and Modeling, GIGA-R, University of Liege, Avenue de l’Hôpital 1, Liège, Belgium
\textsuperscript{3} Genetic & Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
\textsuperscript{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

Step 1: Understand the biological (statistical) problem
Step 2: Know your data
Step 3: Preprocess cleanse, prepare for step 4
Step 4: Integrate select analytic tool, validate, replicate, report
Step 5: Interpret reconcile statistics with biology, visualize

Data Integration Stage
- Early: Data fusion, similar data types, increase power, weights
- Intermediate: Dissimilar data types, alternative data representations
- Final: Post-analysis, combine analytic results

(Hamid et al. 2009)
Systems information by integration (Joyce and Palsson 2006)

<table>
<thead>
<tr>
<th>Genomics</th>
<th>Transcriptomics</th>
<th>Proteomics</th>
<th>Metabolomics</th>
<th>Protein–DNA interactions</th>
<th>Protein–protein interactions</th>
<th>Fluxomics</th>
<th>Phenomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomics (sequence annotation)</td>
<td>• ORF validation</td>
<td>• SNP effect on protein activity or abundance</td>
<td>• Enzyme annotation</td>
<td>• Binding-site identification</td>
<td>• Functional annotation</td>
<td>• Functional annotation</td>
<td>• Functional annotation (71, 101)</td>
</tr>
<tr>
<td>Transcriptomics (microarray, SAGE)</td>
<td>• Protein: transcript correlation</td>
<td>• Enzyme annotation (100)</td>
<td>• Gene-regulatory networks</td>
<td>• Functional annotation (9)</td>
<td>• Protein complex identification</td>
<td>• Functional annotation</td>
<td>• Functional annotation (62)</td>
</tr>
<tr>
<td>Proteomics (abundance, post-translational modification)</td>
<td>• Enzyme annotation (99)</td>
<td>• Regulatory complex identification</td>
<td>• Differential complex formation</td>
<td>• Enzyme capacity</td>
<td>• Functional annotation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolomics (metabolite abundance)</td>
<td>• Metabolic-transcriptional response</td>
<td>• Metabolic pathway bottlenecks</td>
<td>• Metabolic flexibility</td>
<td>• Metabolic engineering (109)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein–DNA interactions (ChIP-chip)</td>
<td>• Signalling cascades (99, 102)</td>
<td>• Protein–protein interactions (yeast 2H, coAP–MS)</td>
<td>• Dynamic network responses (94)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Pathway identification activity (89)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluxomics (isotopic tracing)</td>
<td></td>
<td></td>
<td></td>
<td>• Metabolic engineering</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step 1**

- Formulating the biological (statistical) problem
• 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

- 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

- 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)

Step 5
• Interpretation
Integrative analytics

Top down versus bottom up

- Start with candidate genes, pathways and build up
- Start with x-ome and filter down
Integrative analytics

Crude division:

Kernels

Networks

Components

<table>
<thead>
<tr>
<th>Overview</th>
<th>Classification</th>
<th>Discrimination</th>
<th>Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trends</td>
<td>Pattern Recognition</td>
<td>Discriminating between groups</td>
<td>Comparing blocks of omics data</td>
</tr>
<tr>
<td>Outliers</td>
<td>Diagnostics</td>
<td>Biomarker candidates</td>
<td>Metab vs Proteomic vs Genomic</td>
</tr>
<tr>
<td>Quality Control</td>
<td>Healthy/Diseased</td>
<td>Comparing studies or instrumentation</td>
<td>Correlation spectroscopy (STOCSY)</td>
</tr>
<tr>
<td>Biological Diversity</td>
<td>Toxicity mechanisms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient Monitoring</td>
<td>Disease progression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td>SIMCA</td>
<td>PLS-DA</td>
<td>O2-PLS</td>
</tr>
</tbody>
</table>
Finding the most appropriate method for your research question

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

(http://www.metabolomics.se)
Taking baby steps: starting from GWAs

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

Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma
Methodological challenges
- a toy example -
Methodological aspects: scaling up from GWAs to GWAIs

4 meta-GWAIs

2 meta-GWA

3 GWAIIs

1 GWAs

Complexity / Burden

Storage and computational efficiency

IT and Biological complexity

IT and Parsimony, Model assumptions

Statistical modeling
Methodological aspects: scaling up from GWAs to GWAIs

(Kilpatrick 2009)
Towards a novel integrated framework
“genomic MB-MDR”
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 develop-

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)

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

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

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

Isabelle Cleynen¹*, Jestinah M. Ma hac hie John²,³, Liesbet Henckaerts⁴, Wouter Van Moerkercke¹, Paul Rutgeerts¹, Kristel Van Steen²,³, Severine Vermeire¹

¹ Department of Gastroenterology, KU Leuven, Leuven, Belgium, ² Systems and Medicine, UCL, Brussels, Belgium, ³ Bioinformatics and Modeling, GIGA-R, University of Liège, Liège, Belgium, ⁴ Biotechnology, Ghent University, Ghent, Belgium.
Detecting structure in patients: subphenotyping

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

Bärbel Maus¹,², Camille Jung³,⁴,⁵, Jestinah M. Mahachie John¹,², Jean-Pierre Hugot³,⁴,⁶, Emmanuelle Génin⁷,⁸, Kristel Van Steen¹,²

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

(Bootstrap p-value; AIC: 9 groups; BIC: 3 groups)
Detecting structure in patients: subphenotyping

- 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

Genome-wide association interaction analysis for Alzheimer disease

Elena S. Gusareva a,b,* Minerva M. Carrasquillo c, Céline Bellenguez d,e,f, Samuel Colon f, Neill R. Graff-Radford i, Ronald C. Petersen j, Dennis W. Jestinah M. Mahachie John a,b, Kyrylo Bessonov a,b, Christine Van Broeckhoven l, the GERAD1 Consortium l, Denise Harold k, Julie Williams k, Philippe Amouyel l, 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 GWAI (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

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene name</th>
<th>Function</th>
<th>Location</th>
<th>Epistatic SNPs</th>
<th>Main effect for AlzD</th>
<th>Population (N cases/N controls)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>INS</td>
<td>Insulin</td>
<td>Glucose metabolism</td>
<td>11p15.5</td>
<td>rs689</td>
<td>no</td>
<td>Germans (104/123)</td>
<td>Brune et al., 2003</td>
</tr>
<tr>
<td>PPARA</td>
<td>Peroxisome proliferator-activated receptor alpha</td>
<td>Glucose and lipid metabolism</td>
<td>22q13.31</td>
<td>rs1800206</td>
<td>yes</td>
<td>Northern Europeans (336/2426)</td>
<td>Kölsch et al., 2012</td>
</tr>
<tr>
<td>IL1A</td>
<td>Interleukin 1 alpha</td>
<td>Inflammatory cytokine</td>
<td>2q13</td>
<td>rs3783550</td>
<td>no</td>
<td>Northern Europeans (336/2426)</td>
<td>Heun et al., 2012</td>
</tr>
<tr>
<td>PPARA</td>
<td>Peroxisome proliferator-activated receptor alpha</td>
<td>Glucose and lipid metabolism</td>
<td>22q13.31</td>
<td>rs1800206</td>
<td>yes</td>
<td>Northern Europeans (336/2426)</td>
<td>Heun et al., 2012</td>
</tr>
<tr>
<td>IL1B</td>
<td>Interleukin 1 beta</td>
<td>Inflammatory cytokine</td>
<td>2q13</td>
<td>rs16944</td>
<td>no</td>
<td>Northern Europeans (336/2426)</td>
<td>Heun et al., 2012</td>
</tr>
<tr>
<td>PPARA</td>
<td>Peroxisome proliferator-activated receptor alpha</td>
<td>Glucose and lipid metabolism</td>
<td>22q13.31</td>
<td>rs1800206</td>
<td>yes</td>
<td>Northern Europeans (336/2426)</td>
<td>Heun et al., 2012</td>
</tr>
<tr>
<td>IL10</td>
<td>Interleukin 10</td>
<td>Inflammatory cytokine</td>
<td>1q32.1</td>
<td>rs1800896</td>
<td>yes</td>
<td>Northern Europeans (336/2426)</td>
<td>Heun et al., 2012</td>
</tr>
<tr>
<td>PPARA</td>
<td>Peroxisome proliferator-activated receptor alpha</td>
<td>Glucose and lipid metabolism</td>
<td>22q13.31</td>
<td>rs4253766</td>
<td>no</td>
<td>Northern Europeans (336/2426)</td>
<td>Combarros et al., 2010</td>
</tr>
<tr>
<td>IL1A</td>
<td>Interleukin 1 alpha</td>
<td>Inflammatory cytokine</td>
<td>2q13</td>
<td>rs1800587</td>
<td>no</td>
<td>Northern Europeans (336/2426)</td>
<td>Combarros et al., 2010</td>
</tr>
<tr>
<td>DBH</td>
<td>b-Hydroxylase</td>
<td>Onverts dopamine to norepinephrine in the synaptic vesicles of postganglionic sympathetic neurons</td>
<td>9q34.2</td>
<td>rs1611115</td>
<td>yes</td>
<td>Northern Europeans (336/2426)</td>
<td>Robson et al., 2004</td>
</tr>
<tr>
<td>TF</td>
<td>Transferrin</td>
<td>Iron metabolism</td>
<td>3q22.1</td>
<td>rs1049296</td>
<td>no</td>
<td>UK (191/269)</td>
<td>Kauwe et al., 2010</td>
</tr>
<tr>
<td>HFE</td>
<td>Hemochromatosis</td>
<td>Iron metabolism</td>
<td>6p22.2</td>
<td>rs1800562</td>
<td>yes</td>
<td>Caucasians USA (1166/1404)</td>
<td>Lehmann et al., 2012</td>
</tr>
<tr>
<td>TF</td>
<td>Transferrin</td>
<td>Iron metabolism</td>
<td>3q22.1</td>
<td>rs1130459</td>
<td>no</td>
<td>North Europeans (336/2426)</td>
<td>Lehmann et al., 2012</td>
</tr>
<tr>
<td>HFE</td>
<td>Hemochromatosis</td>
<td>Iron metabolism</td>
<td>6p22.2</td>
<td>rs1799945</td>
<td>yes</td>
<td>North Europeans (336/2426)</td>
<td>Lehmann et al., 2012</td>
</tr>
<tr>
<td>MTHFR</td>
<td>Methylenetetrahydrofolate reductase</td>
<td>Homocysteine metabolism useful for normal brain functioning</td>
<td>1p36.22</td>
<td>rs1801131</td>
<td>yes</td>
<td>Indians (80/120)</td>
<td>Mansoori et al., 2012</td>
</tr>
<tr>
<td>IL6</td>
<td>Interleukin 6</td>
<td>Pro-inflammatory cytokine</td>
<td>7p15.3</td>
<td>rs1800795</td>
<td>no</td>
<td>North Spain (232/191)</td>
<td>Infante et al., 2004</td>
</tr>
<tr>
<td>IL10</td>
<td>Interleukin 10</td>
<td>Limit inflammation in the brain</td>
<td>1q32.1</td>
<td>rs1800871</td>
<td>yes</td>
<td>North Europeans (336/2426)</td>
<td>Combarros et al., 2009</td>
</tr>
<tr>
<td>IL6</td>
<td>Interleukin 6</td>
<td>Pro-inflammatory cytokine</td>
<td>7p15.3</td>
<td>rs2069837</td>
<td>yes</td>
<td>North Europeans (336/2426)</td>
<td>Combarros et al., 2009</td>
</tr>
<tr>
<td>ABCA1</td>
<td>ATP-binding cassette transporter A1</td>
<td>Intracellular cholesterol transport and maintance of cell cholesterol balance</td>
<td>9q31.1</td>
<td>rs2422493</td>
<td>no</td>
<td>Spanish (631/731)</td>
<td>Rodríguez-Rodríguez et al., 2010</td>
</tr>
<tr>
<td>NPC1</td>
<td>Niemann-Pick C1</td>
<td></td>
<td>18q11.2</td>
<td>rs18050810</td>
<td>no</td>
<td>Spanish (631/731)</td>
<td>Rodríguez-Rodríguez et al., 2010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs4800488</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs2236707</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs2510344</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Chromosome</td>
<td>SNP</td>
<td>Association</td>
<td>Population Size</td>
<td>Study Reference</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------</td>
<td>--------------</td>
<td>--------------</td>
<td>-----------------</td>
<td>----------------------------------</td>
<td></td>
</tr>
<tr>
<td>LRP1</td>
<td>Low density lipoprotein receptor-related protein 1</td>
<td>12q13.3</td>
<td>rs1799986</td>
<td>no</td>
<td>Spanish (246/237)</td>
<td>Vázquez-Higuera et al., 2009</td>
<td></td>
</tr>
<tr>
<td>MAPT</td>
<td>Microtubule-associated protein tau</td>
<td>17q21.33</td>
<td>rs2471738</td>
<td>no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSK3B</td>
<td>Glycogen synthase kinase-3 beta</td>
<td>3q13.33</td>
<td>rs334558</td>
<td>no</td>
<td>Spanish (246/237)</td>
<td>Vázquez-Higuera et al., 2009</td>
<td></td>
</tr>
<tr>
<td>CDK5R1</td>
<td>Cyclindependent kinase 5</td>
<td>17q11.2</td>
<td>rs735555</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NR1H2</td>
<td>Liver X receptor beta</td>
<td>19q13.33</td>
<td>rs1052533</td>
<td>no</td>
<td>Spanish (414/442)</td>
<td>Infante et al., 2010</td>
<td></td>
</tr>
<tr>
<td>HMOX1</td>
<td>Heme oxygenase-1</td>
<td>22q12.3</td>
<td>rs2071746</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

"+" - 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, ...)

Université de Liège
Unable to replicate is a bad thing?

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

Casey S. Greene¹, Nadia M. Penrod¹, Scott M. Williams², Jason H. Moore¹,²,³,⁴,⁵,⁶*

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
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)

<table>
<thead>
<tr>
<th></th>
<th>BB</th>
<th>Bb</th>
<th>bb</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aa</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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)

Bryant et al. 2010:
protein function prediction

(Bryant et al. 2010)

Mixture modeling for model-based clustering (Mclust in R; Fraley et al. 2012)

Gave us a sense of nonlinearity of genes

Mixtures of probabilistic principal component analysers", Neural computation 11(2), 1999

How many PCs should we pick up?

(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 GWAs - Bio3’s research lines

Research line 1
GxG and GxE interactions
- region of interest
- SNPs
- cis/trans-eQTL epistasis
- replication
- meta-analysis
- ancestry-dependence

Research line 2
INTEGROMICS
- statistical interaction networks
- network perturbation
- network differentiation
- network fusion
- omics integration

Research line 3
P-STRUCT
- epigenetic markers
- omics integration
- patient heterogeneity
- population structure
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)

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 attributes. Searching this enormous high-dimensional data space imposes a great computational challenge. Our network-based approach can supervise the search for three-locus models of disease susceptibility by identifying strong pairwise epistatic interactions and provide a global interaction map to search for higher-order interactions together in the networks. Applying this approach to a population-based bladder cancer dataset, we observe new 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 these 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 GWAI 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

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


- **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]


Other references

URLs:

- 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