From Statistical to Biological Interactions via Omics Integration
Outline

1. Thesis overview
2. Concepts
3. Genome-genome interactions
4. Trans-eQTL epistasis protocol
5. Gene expression networks
6. General conclusions
7. Future directions
Thesis Overview
Contributions 1 of 2

**Statistical epistasis networks**

- **Genetic Epistasis**

**Software development (CH 3)**
- Key: Epishell
- Ref: Grange and Bessonov(*) et al. 2016 Finding the tree for the forest: which epistasis analysis method to choose? – circulating among co-authors
  - (*): equal contribution

**SNP x SNP analysis protocol (CH 3)**
- Key: Robustness checks
- Ref: Bessonov et al. 2015 A cautionary note on the impact of protocol changes for genome-wide association SNP-SNP interaction studies: an example on ankylosing spondylitis. Hum Genet 134: 761-773
Contributions 2 of 2

**Biological Epistasis**

- **Trans-eQTL epistasis protocol (CH 4)**
  - Key: Multiple testing

- **Gene-expression networks (CH 5)**
  - Key: Conditional Inference Forests (CIFs)

- **Two-omics networks (CH 6)**
  - Key: Penalized regression
  - Ref: Gadaleta and Bessonov (*) et al. 2015
  - Integration of Gene Expression and Methylation to unravel biological networks in glioblastoma patients. (submitted to Genetic Epidemiology special issue (*) : equal contribution)

Gene regulatory networks
Concepts
Omics data

Environmental (E)

Omics (O)
- Genomics
- Epigenomics
- Transcriptomics
- Proteomics
- Metabolomics

Phenotypic (P)
- Phenomics
- Cancer
- Diabetes

- Smoking
- Age
Genomics: Single Nucleotide Polymorphisms

- Locus – physical location in the genome
- SNP - genetic marker
- Phenotype - observable trait
## Interactions

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene - gene (GxG)</td>
<td>mRNA</td>
</tr>
<tr>
<td>gene-gene (SNPxSNP)</td>
<td>SNPs</td>
</tr>
<tr>
<td>protein-protein</td>
<td>protein</td>
</tr>
<tr>
<td>gene-environment (GxE)</td>
<td>SNPs / environment</td>
</tr>
</tbody>
</table>

### Data omics layers
- **Genomics**, **Transcriptomics**, **Phenomics**

### Proteins
- P1
- P2
- P3

### mRNA, SNPs, DNA
- G1
- G2
- G3
Epistasis

Biological

One gene or allele **masking** the phenotypic expression of the other genes or alleles in the interaction.

~ not necessarily symmetric

Statistical

Departure from a specific **linear model** describing the relationship between predictive factors (here assumed to be alleles at different genetic loci)

~ symmetric in regression framework
Biological epistasis

- **se**\(^+\) red eyes (dominant)
- **se**\(^-\) brown eyes (recessive)
- **eyD** no eyes (dominant)
Statistical epistasis

Phenotype (color)

AA

Aa

Locus $i$

aa

Locus $j$

No epistasis

Epistasis
Omics integration

- Single-omics (transcriptomics / transcriptomics)
- Multi-omics (transcriptomics / metabolomics)
Challenges

• Data
  - Storage
  - Accessibility
  - Standardization

• Analysis
  - “Curse of dimensionality
    Large $p$, small $n$ problem
  - Systems view in omics integration
Nowadays ...
Genome-genome interactions

The impact of protocol changes for genome-wide association SNP x SNP interaction
Context: genome - phenome

- Phenotypic data layer (case / control)
- Genotypic data layer (SNP)
Context: genome – phenome interactions

- Which pair of markers affects phenotype?
  - Predictors - SNPs
  - Trait - phenotype

- Linkage disequilibrium (LD)
  - Association between alleles
• Genome-wide association interaction studies (GWAI)

• Goal
  ➢ Gene – gene interactions
  ➢ Assumes large number of individuals

• Linear regression model

\[ Y_{\text{trait}} = \beta_0 + \beta_1 X_{\text{locus } i} + \beta_2 X_{\text{locus } j} + \beta_3 X_{\text{locus } i} \times X_{\text{locus } j} + \varepsilon_i \]
Strategy: GWAI protocol

Data

Quality Control

Screening for loci

Interpretation

Replication

Biological validation

“Curse of dimensionality”
Large $p$, small $n$ problem
Multiple-testing
Strategy: GWAI protocol

Data

Quality Control

Screening for loci

Interpretation

Replication

Biological validation

Targeted Quality Control (QC) protocol
Strategy: GWAI protocol

Data

Quality Control

Screening for loci

Interpretation

Replication

Biological validation
Strategy: GWAI protocol

- Data
- Quality Control
- Screening for loci
- Interpretation
- Replication
- Biological validation
Strategy: GWAI protocol

- Data
- Quality Control
- Screening for loci
- Interpretation
- Replication
- Biological validation

Replication
Validation in a different population
Strategy: GWAI protocol

Data

- Quality Control
- Screening for loci
- Interpretation
- Replication
- Biological validation
**Strategy: GWAI protocol**

0. Data collection and genotyping

1. Samples and markers quality control

   - Exhaustive epistasis screening (a)
   - Selective epistasis screening (b)

2a.
   - LD pruning ($r^2 > 0.75$)
   - Adjustment for confounders
   - Screening for pair-wise SNP interactions

2b.
   - Biochemical networks
   - Available knowledge
   - Function and location

   - LD pruning ($r^2 > 0.75$)
   - Adjustment for confounders
   - Screening for pair-wise SNP interactions

3. Replication and validation in independent data

4. Biological and functional validation
Problem

• No standard GWAI protocol exists
  ➢ Choice of parameters
    ▪ Dataset
    ▪ Encoding
      ▪ Additive
      ▪ Co-dominant
    ▪ LD pruning

• Impact on the final epistasis findings
Application: Ankylosing spondylitis data

• Cases*
  ➢ 2005 ankylosing spondylitis (AS)

• Controls*
  ➢ 3000 British 1958 Birth Cohort (BC)
  ➢ 3000 National Blood Donors (NBS)

• Source
  ➢ Wellcome Trust Case Control Consortium (WTCCC2)

* European ancestry
Application: 8+2 GWAI protocols

- Exhaustive marker screening
  - Yes: Pre-selection of markers via Biofilter 2.0 (Implication idx = 3) (44,018)
  - No: LD pruning at $r^2 > 0.75$
    - Yes: MB-MDR MAXT
    - No: MB-MDR gammaMAXT
      - #9 co-dominant (487,780)
      - #10 co-dominant (321,565)

- LD pruning at $r^2 > 0.75$
  - Yes: #4 BOOST pre-selected LD pruned (30,426)
  - No: #3 BOOST pre-selected (44,018)
    - #7 additive (44,018)
    - #5 co-dominant (44,018)
Application: results overlap

(significant SNP pairs)
Application: percent overlap

MB-MDR

Exhaustive

Boost

Biofilter

LD up → #9 (6589)
LD down → #10 (2340)
LD up → #5, #7 (77, 84975)
LD down → #6, #8 (48, 57032)

LD up → #1 (226)
LD down → #2 (2165)
LD up → #3 (129)
LD down → #4 (84)
Application: distance

• Sort results
  ➢ Highest to lowest significance

• 207 common SNPs
  ➢ All protocols
  ➢ Significant and Non-significant

• Get rank values

• Calculate Euclidian distance

\[ D(1,2) = \sqrt{(5-125)^2 + (5-500)^2 + (120-500)^2} \]

\[ D(1,2) = 675.28 \]
Application: clustering of protocols

- 207 ranks
  - Common SNPs pairs
  - Not all significant
- Marker selection
  - Protocols
    - #1-#2
    - #3-#8
- Encoding
  - Protocols #5-#8
- LD pruning
  - Protocols
    - #3 and #4
    - #5 and #6
<table>
<thead>
<tr>
<th>GO ID</th>
<th>GO Term Description</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0007411</td>
<td>axon guidance</td>
<td>7.9E-77</td>
</tr>
<tr>
<td>GO:0030168</td>
<td>platelet activation</td>
<td>3.9E-58</td>
</tr>
<tr>
<td>GO:0055085</td>
<td>transmembrane transport</td>
<td>3.0E-50</td>
</tr>
<tr>
<td>GO:0007268</td>
<td>synaptic transmission</td>
<td>2.0E-36</td>
</tr>
</tbody>
</table>

* Fisher’s method (combined topGO p-values from 10 protocols)
Conclusions

• 10 GWAI protocols
  ➢ Dramatic changes
  ➢ Key factors
    ▪ Input markers
    ▪ Tool selection
    ▪ Encoding of lower order effects (additive / co-dominant)
    ▪ LD pruning

• Impact strength

Higher  Lower
Dataset  >  Encoding  >  LD pruning
Trans-eQTL epistasis protocol

Integrative network-based analysis of cis and trans regulatory effects in asthma
Context: genome - transcriptome

- Trait – expression (microarrays / RNAseq)
- Predictors – genotypic data (SNP arrays)
Context: problem

- Identification of genome – transcriptome interactions
- Avoid statistical artifacts
- Build epistatic statistical model (network)
Context: *Cis* eQTL

• Expression quantitative trait loci (eQTL)

• Marker in the TG ‘neighborhood’
• Distant marker affecting a TG
• TF – transcription factor
• Interaction
  - Between *trans* and *cis* loci
  - *Trans* locus modifies effect of *cis* locus on the TG
    - SNP_{trans} \times SNP_{cis} \rightarrow [TG]
Context: physical trans/cis loci mapping

- **ORF** – open reading frame
  - Codes a gene product (introns + exons)
### Strategy: *trans-eQTL epistasis protocol*

<table>
<thead>
<tr>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Childhood Asthma management program (CAMP)</td>
</tr>
<tr>
<td>• 177 asthmatics (smokers / non-smokers)</td>
</tr>
<tr>
<td>• Expression - microarrays</td>
</tr>
<tr>
<td>• Genotypic - SNP arrays</td>
</tr>
</tbody>
</table>
Strategy: *trans-eQTL epistasis protocol*

- Cis eQTLs
  - Generalized Least Squares (GLS)
  - 1763 cis eQTLs
Strategy: *trans-eQTL epistasis protocol*

- **Data**
- **Cis eQTLs**
- **Epistatic eQTLs**

- For 1763 *cis* eQTLs
- Find epistatic signals (*trans/cis* eQTLs)
  - MB-MDR
  - Step-wise permutation
    - Step 1 - $10^3$ permutations
    - Step 2 - $10^7$ permutations
- MB-MDR
Strategy: *trans-eQTL epistasis protocol*

- Build statistical epistatic network
- *Trans x cis, cis x cis, trans x trans* interactions
Strategy: *trans-eQTL epistasis protocol*

- **Data**
- **Cis eQTLs**
- **Episatic eQTL**
- **Network**
- **Validation**
  - Disease etiology
  - Previous knowledge
Simulations: null data

- FWER
  - within each trans/cis eQTL run
  - Mean 0.056
  - Median 0.04
Applications: statistical epistatic network

- 1459 nodes
- **red**: high
- **orange**: average
Application: mapping to pathways

1364 epistatic trans/cis eQTL p-value < 0.05

trans gene
pathway
trans
208

cis gene
pathway
cis
171

mRNA
Applications: significant genes overlap

Pathways overlap

Transcription Regulation (R-HSA-212436)
Cell adhesion pathways (R-HSA-418990)

Adaptive immunity (R-HSA-212436)
Signaling
Conclusions

• Impact of genetic component on expression
  ➢ Higher order interactions
    ▪ \textit{trans/cis} epistatic effects

• Global interaction map
  ▪ Epistatic network

• Disease-relevant results
Gene expression networks
Practical aspects of gene regulatory network inference
(CIFs)
- **Trait** – target gene (TG)
- **Predictors** – transcription factor (TF)
Did you see?
Did you see me...on the second row?
While I was upregulating.
Context: transcriptional networks

- **Regulators**
  - Transcription factors
- **Targets**
  - Target genes
- **Expression data**
- **Directed edges**
Infer a transcriptional network

- Correlation structure of genes
- Scaling (>1000 genes)
- $mtry$ parameter
  - performance impact
• Diseases
  ➢ Share genes
  ➢ Classify
  ➢ Etiology
Strategy: network inference via trees
**Strategy: Conditional Inference Forest**

- Select randomly $m$ variables ($mtry$) $X=\{x_1, \ldots, x_m\}$

- For each $x_i$ in $X$ test "global" null hypothesis $H_0$
  \[ H_0 = \cap_{j=1}^{m} H^j_0 \text{ and } H^j_0 : D(\mathbf{Y}|X_j) = D(\mathbf{Y}) \]

- Select one covariate $x_j$ with largest $c_{max}$

  \[ c_{max}(\mathbf{t}, \mu, \Sigma) = \max_{k=1,\ldots,pq} \left| \frac{(t - \mu)_k}{\sqrt{(\Sigma)_{kk}}} \right| = \left| \frac{t - \mu}{\sqrt{\Sigma}} \right| \]

- Assign $x_j$ to a node

- Split $x_j$

  - Maximize split test statistic $c_{split}$

    \[ c_{max}(t^A_{j\star}, \mu^A_{j\star}, \Sigma^A_{j\star}) = \max_k \left| \frac{(t^A - \mu)_k}{\sqrt{(\Sigma)_{kk}}} \right| \]
Advantages

• Threshold available
  ➢ useful in the absence of a gold standard

• Global test of independence*
  ➢ Avoids bias in variable selection  [5]
    ▪ 2-stages: 1) node variable selection and 2) splitting
  ➢ Accommodates different measurement scales

• Handles correlated variables
  ➢ Conditional permutation scheme (CIF\textsubscript{cond})

Strategy: Why not CIFs?

Disadvantages

• Computation time

  ➢ In the presence of multicollinearity (f.i. gene co-expression) the conditional variable importance measure is advocated

• Selects the features with the best “linear” association to the outcome

  ➢ Tends to miss non-linear associations

    ▪ Proposed solution

      ▪ Generalized additive models (GAM)
Strategy: CIF variants

- **CIT**
  - Single conditional inference tree

- **CIF**
  - original CIF
  - classical permutation scheme

- **CIF_{cond}**
  - original CIF
  - conditional permutation scheme

- **CIF_{mean}**
  - CIF without permutation
  - Averaging of node $p$-values or test-statistics

- **RF** – Random Forest
Strategy: CIF$_{\text{mean}}$

- No permutations are required
- Multiple-test control at each node
  - Bonferroni (samples)
  - Monte-Carlo
- Variable importance for each $x_j$
  - Average over n trees where $x_j$ is present

\[
\frac{\sum_t p_{xjt}}{n(X_j^t)}
\]

Number of trees containing $x_j$
Results: DREAM data

- Dialogue for Reverse Engineering Assessments and Methods
- Gold Standard available
- Predict
  - Gene regulatory network
- Data
  - Expression
## Results: DREAM Data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GS available</th>
<th>Real-life</th>
<th>Nr of genes</th>
<th>Nr of TFs</th>
<th>Nr of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>DREAM2 (E.coli)</td>
<td>Y</td>
<td>Y</td>
<td>3456</td>
<td>320</td>
<td>300</td>
</tr>
<tr>
<td>DREAM4 network 1</td>
<td>Y</td>
<td>N</td>
<td>100</td>
<td>100**</td>
<td>100</td>
</tr>
<tr>
<td>DREAM4 network 2</td>
<td>Y</td>
<td>N</td>
<td>100</td>
<td>100**</td>
<td>100</td>
</tr>
<tr>
<td>DREAM4 network 3</td>
<td>Y</td>
<td>N</td>
<td>100</td>
<td>100**</td>
<td>100</td>
</tr>
<tr>
<td>DREAM4 network 4</td>
<td>Y</td>
<td>N</td>
<td>100</td>
<td>100**</td>
<td>100</td>
</tr>
<tr>
<td>DREAM4 network 5</td>
<td>Y</td>
<td>N</td>
<td>100</td>
<td>100**</td>
<td>100</td>
</tr>
<tr>
<td>DREAM5 network 1</td>
<td>Y</td>
<td>N</td>
<td>1643</td>
<td>195</td>
<td>805</td>
</tr>
<tr>
<td>DREAM5 network 2 (E.coli)</td>
<td>Y</td>
<td>Y</td>
<td>4511</td>
<td>334</td>
<td>805</td>
</tr>
<tr>
<td>DREAM5 network 3 (S.cerevisiae)</td>
<td>Y</td>
<td>Y</td>
<td>5950</td>
<td>333</td>
<td>536</td>
</tr>
</tbody>
</table>
Results: measures of evaluation

• AUROC
  ➢ Area Under Receiver Operating Characteristic
  ➢ TPR / FPR

• AUPR
  ➢ Area Under Precision Recall
  ➢ Precision / Recall

• DREAM 4/5 score
  ➢ 1/2 * (ROC score + PR score)
  ➢ 25,000 of random networks (re-sampling)
Results: DREAM 4 gold standards

100 nodes each
Results: DREAM4

- Single tree (CIT)
  - Poor performance
- $\text{CIF}_{\text{cond}}$ performance slightly better than RF
Results: DREAM 5

- $\text{CIF}_{\text{mean}}$ comparable to RF and GENIE3 performance
- Little gain from Monte-Carlo MT
- Bonferroni is not to be recommended
Results: DREAM5 - \textit{mtry}

- \textit{mtry} parameter
  - Significant performance impact
  - Here \( k/3 \) is the top performer
• *mtry* parameter
  - Significant performance impact
  - here $k=5$ is the top performer
Conclusions

• CIFs provide comparable performance to RF

• CIFs are scalable
  ➢ Multi-thread runs
  ➢ $\text{CIF}_{\text{mean}}$ (12 min/100 genes/100 samples /1CPU)

• CIFs imply statistically sound variable selection
  ➢ Significance-based threshold selection
  ➢ No gold standard needed ($\text{CIF}_{\text{mean}}$)
General conclusions
Conclusions

- Small protocol changes in epistasis screening can have a major impact on replication and validation follow-up studies.
- Using prior information helps in obtaining more robust results, yet limits the detection of novel (not previously reported) gene-gene interactions.
- Sometimes pragmatic approaches to feature selection need to be taken in very small \( n \) datasets.
- Even in the absence of multicollinearity or highly correlated features, CIF\textsubscript{cond} showed comparable results with RF (DREAM4 score).
Future directions
Future perspectives

Genome-Genome Interactions

1. Optimal LD pruning threshold definition
   - Determine the lower bounds for LD pruning (now \( r^2 > 0.75 \))

2. Epistatic hits aggregation over protocols
   - Optimally combine complementary epistatic evidences from different epistasis detection routes
**Perspectives**

**Trans-eQTL epistasis protocol**

1. Apply the protocol to sufficiently large datasets (large $n$)
   - Carry out a thorough evaluation of false positives (FWER)
   - Assess the impact of the step-wise procedure (MB-MDR) on false positives

**Gene expression networks**

1. Increase computational efficiency (speed) in the conditional variable importance computations
Papers (14 published)

Statistical genetics / genetic epidemiology related to my PhD thesis


In preparation / submitted

3. Schleich F., Bessonov K, Van Steen K (2015). Exhaled volatile organic compounds are able to discriminate between neutrophilic and eosinophilic asthma. (submitted) - Patent #203-17

Data mining/molecular dynamics related


Biology related

3. Kyrylo Bessonov and Dr. George Harauz. “In-silico study of the myelin basic protein C-terminal α-helical peptide in DMPC and mixed DMPC/DMPE lipid bilayers.” Studies by Undergraduate Researchers at Guelph 2010; 4(1).
5. Mumdooh A.M Ahmed, Miguel De Avila, Eugenia Polverini, Kyrylo Bessonov, Vladimir V. Bamm, George Harauz, “Solution NMR structure and molecular dynamics simulations of murine 18.5-kDa myelin basic protein segment (S72-S107) in association with dodecylphosphocholine micelles”. Biochemistry
Acknowledgements

• Supervisor
  ➢ Prof. Dr. Dr. Kristel Van Steen

• BIO3 lab

• Committee members
  ➢ Pierre Geurts  ▪ Vincent Bours  ▪ Benno Schwikowski
  ➢ Patrick Meyer  ▪ Monika Stoll

• Funding
  ➢ FNRS, COST BM1204, ITN MLPM
  ➢ University of Liege
References


