Diseases as complex phenomena

- Complex phenotypes and diseases are the result of a high number of genomic and epi-genomic variations, directly affecting:
  - Genes in which variations appear (i.e. target genes)
  - Regulators of target genes (transcription factors, chromatin structural proteins, receptors, etc…)
  - Genomic regulatory elements (promoters, enhancers, insulators, silencers, etc.)
- The effects on gene expression, physical interaction and regulation are also modified by non-genetic factors, such as differentiation (tissue-specific expression), and environmental stimuli.
Diseases as complex phenomena

- TISSUE-SPECIFIC GENE REGULATION
- ENVIRONMENT
- GENETIC VARIATION
- CHROMATIN ACCESSIBILITY
- REGULATORY ELEMENTS: Promoters, Enhancers, Insulators, Silencers...
- GENE EXPRESSION
- EPIGENOME: CpG islands, DNAme, Histone methylation, Histone variants, ...
- DISEASE

GWAS as big data

- **GWAS** consider a very large number of *genome-wide variants* to test their association with a particular trait.

- **Frontotemporal dementia** case-study
  - **EXAMINED INDIVIDUALS**
    - 532 cases
    - 926 controls
    - **1458 individuals**, overall
  - **EXAMINED MARKERS** (for each individual)
    - ~ 0.7M directly genotyped SNPs (*)
    - More than 1.8M imputed SNPs (*)
    - Overall, more than **2.54M markers** (*)

(*) Single Nucleotide Polymorphisms.
High stringency in GWAS

The biological problem

- GWAS are capable of testing the association between each single marker and the trait of interest.

- However, the observed variability in diseased phenotypes is often due to complex interactions among genetic markers having small and synergic effects.

- We also need to functionally characterize millions of potentially causal markers, to understand not only their association with the trait of interest, but also the biological mechanism at the base of the perturbation processes causing the disease.
From SNPs to molecular pathways

- Mapping millions of SNPs onto genes. Possible ways:
  - Overlap/Proximity
  - Linkage disequilibrium (LD)

- Computing LD and handling genotype data for more than 2.5M SNPs and ~1500 individuals. Solutions:
  - Parallel programming (R parallel)
  - Data indexing (SQLite)

Analysis environment

- RefSeq
- SNP
- Data analysis
- Storage, Indexing, Pre-Processing
- Case control data
- Association data, Genotypes
Analysis environment

Case-control data

RefSeq

Genomic region annotation

TSS

Data retrieval from HapMap, SNP mapping

Functional annotation

Network analysis SEM analysis

Data/Genome stats

Analysis workflow (I)

SNP association test

SNP-to-gene map

Gene association

Case-control study

Genotypes

Gene vector space

PCA1 (1st principal component)
Gene association

- **Input data:**
  - Marker coordinates
  - Marker association P-values (likelihood ratio test from GWAS)
  - Genotypes (directly assayed and imputed)

- A marker X is considered in linkage with a gene Y if the maximum LD with a Y-overlapping marker, measured as R-squared, is above a user-defined threshold (default: 0.1).

- Gene-mapping is entirely based on marker-marker co-segregation during genetic recombination.

- A gene is considered to significantly discriminate cases from controls when its gene-wise P-value (from GATES method) is less than 0.05

Advantages and assumptions of LD

- Intergenic markers are assigned to genes based on a specific **biological criterion**, instead of an arbitrary proximity threshold.

- **One marker may influence more than one gene.** The higher the R-squared value, the stronger the influence.

- Also non-significantly associated markers are considered when calculating gene-wise association. LD methods map markers to genes on the base of the LD alone. This will recover potentially associated variants, having **small and synergic effects** on the observed trait, although they are not singularly relevant to the trait itself.
SNP to genes mapping

GWAS are characterized by a large number of loci tested for association.

Problems:
- **Reduced statistical power.** This problem is enhanced by the high stringency (typically $P < 5 \times 10^{-8}$) required to control FP associations.
- The **LD structure of a population** may generate surrogate SNPs being in strong LD with a true disease-causing variant.
- **Surrogate SNPs may differ** across different populations, having different LD structure, leading to non-replicable results.

SNP to genes mapping

Mapping SNPs to genes …
- Reduces the amount of multiple tests
- Include variants that are singularly not significantly associated
- Offers a replicable set of loci to be tested across populations
- Is prone to functional annotation and enrichment analysis
- Offers an easier way to perform pathway and network analysis

The GATES method (extended Simes procedure) is used to assign a gene-wise P-value from its SNP P-values, and does need permutation or simulation to evaluate gene-based empirical significance (Li et al, 2011)
**Analysis workflow (II)**

Gene list: NBEAL2, ELPA, ELP4, TM4CS, SMAP4, G1, G2, G3

SEM multi-group analysis (global testing)

Significantly perturbed pathways

Map gene list to metabolic pathways

Network model from each pathway

**Structural Equation Modeling (SEM)**

- The SEM approach is capable of **detecting**, **estimating**, and **testing** the causal relationships among variables (genes), given the underlying conceptual model.
- The network obtained from the metabolic pathways involving the discriminant genes, is converted to a SEM model (a system of linear equations) and then tested using a **SEM multi-group analysis**.
- This analysis allows the user not only to test the differences between cases and controls, but also the **missing relationships** that may further explain this difference.
Structural Equation Modeling (SEM)

- SEM are mixed graphs $G = (V, E)$, where
  - The directed relationship $k \rightarrow j$ (parent-child) is represented by the structural equation system $j = 1, \ldots, q$: endogenous genes:
    \[
    y_j = \beta_{jC} C + \sum_{k \in pa(j)} \beta_{jk} y_k + u_j
    \]
  - The bi-directed relationship $k \leftrightarrow j$ (spouse-spouse) is represented by the covariance structure:
    \[
    \text{cov}(u_j; u_k) = \begin{cases} 
    \Psi_{jk} & \text{if } j \in sp(k) \\
    0 & \text{otherwise}
    \end{cases}
    \]

**GWAS-SEM**

- $C = \text{case & control}$ indicator; $y_j = j$th gene; $pa(j) = \text{parental set of } j$; $sp(j) = \text{spouse set of } j$; $\beta_{jk} = \text{regression coefficients}$; $\Psi_{jk} = \text{variance-covariances}$; $u_j = \text{unobserved genes of } j \text{ unexplained by the model.}$

<table>
<thead>
<tr>
<th>SAMPLE01</th>
<th>SAMPLE02</th>
<th>SAMPLE03</th>
<th>SAMPLE04</th>
<th>SAMPLE05</th>
<th>SAMPLE06</th>
<th>SAMPLE07</th>
<th>SAMPLE08</th>
<th>SAMPLE09</th>
<th>SAMPLE10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>0.04</td>
<td>0.09</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.05</td>
<td>0.02</td>
<td>0.05</td>
<td>0.03</td>
<td>0.02</td>
<td>0.09</td>
</tr>
<tr>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.05</td>
<td>0.02</td>
<td>0.05</td>
<td>0.03</td>
<td>0.02</td>
<td>0.09</td>
</tr>
<tr>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.05</td>
<td>0.02</td>
<td>0.05</td>
<td>0.03</td>
<td>0.02</td>
<td>0.09</td>
</tr>
</tbody>
</table>

$C = 1$
GWAS-SEM constrained

\[ y_j = \beta_{jc}C + \sum_{k \in \text{pa}(j)} \beta_{jk}y_k + u_j \]

\( \beta_{jk} = \{-1, 0, 1\} \) represents the KEGG connection type:

\( c(\text{"activation"}, \text{"compound"}, \text{"binding/association"}, \text{"expression"}, \text{"inhibition"}, \text{"activation}\text{-}\text{phosphorylation"}, \text{"phosphorylation"}, \text{"indirect"}, \text{"inhibition}\text{-}\text{phosphorylation"}, \text{"dephosphorylation}\text{-}\text{inhibition"}, \text{"dissociation"}, \text{"dephosphorylation"}, \text{"activation}\text{-}\text{dephosphorylation"}, \text{"state"}, \text{"activation}\text{-}\text{indirect"}, \text{"inhibition}\text{-}\text{ubiquination"}, \text{"ubiquination"}, \text{"expression}\text{-}\text{indirect"}, \text{"indirect}\text{-}\text{inhibition"}, \text{"repression"}, \text{"binding}\text{-}\text{association}\text{-}\text{phosphorylation"}, \text{"dissociation}\text{-}\text{phosphorylation"}, \text{"indirect}\text{-}\text{phosphorylation"}) \)

\( c(1, 0, 0, 1, -1, 1, 0, -1, -1, 0, 0, 1, 0, 1, -1, 0, 1, -1, 0, 0, 0) \)

\( \beta_{jc} = \{0, \beta_{hc}\} \) where \( h = \text{GATES significantly discriminant genes} \)

Gene covariance matrix, \( \Sigma(\theta) \)

- SEM parameters are estimated so that the implied (expected) gene covariance matrix:

\[ \Sigma(\theta) = \begin{bmatrix} \sigma_{11}(\theta) & \sigma_{12}(\theta) & \sigma_{13}(\theta) & \sigma_{14}(\theta) \\ \sigma_{21}(\theta) & \sigma_{22}(\theta) & \sigma_{23}(\theta) & \sigma_{24}(\theta) \\ \sigma_{31}(\theta) & \sigma_{32}(\theta) & \sigma_{33}(\theta) & \sigma_{34}(\theta) \\ \sigma_{41}(\theta) & \sigma_{42}(\theta) & \sigma_{43}(\theta) & \sigma_{44}(\theta) \end{bmatrix} \]

where \( \theta = (\beta, \Psi) \)

is close to the observed gene covariance matrix, \( S \).

- pathway P-values are obtained from bootstrap LRTs given the null hypothesis: \( H_0: \Sigma(\theta_0) = \Sigma(\theta_1) \)
### Pathway testing

**Network testing**

**Single pathway testing**

**SEM multi-group analysis**

**Node testing (between groups)**
Test whether a gene is implied in differentiating cases from controls

**Edge testing (between groups)**
Test whether group membership has effects on gene-gene relationships

**Edge testing (within groups)**
Test whether the relationships defined by model hold also in the data

### Table: Pathway testing results

<table>
<thead>
<tr>
<th>Name</th>
<th>pSize</th>
<th>mSize</th>
<th>p-value</th>
<th>pFDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Endocrine and other factor-regulated calcium reabsorption</td>
<td>3</td>
<td>1</td>
<td>0.009268</td>
<td>0.05262</td>
</tr>
<tr>
<td>2. Alzheimer’s disease</td>
<td>3</td>
<td>1</td>
<td>0.026586</td>
<td>0.08262</td>
</tr>
<tr>
<td>3. Dopaminergic synapse</td>
<td>9</td>
<td>1</td>
<td>0.061451</td>
<td>0.12762</td>
</tr>
<tr>
<td>4. Long-term potentiation</td>
<td>3</td>
<td>1</td>
<td>0.067656</td>
<td>0.13048</td>
</tr>
<tr>
<td>5. Serotonergic synapse</td>
<td>8</td>
<td>1</td>
<td>0.083253</td>
<td>0.149856</td>
</tr>
<tr>
<td>6. Cholinergic synapse</td>
<td>7</td>
<td>1</td>
<td>0.092734</td>
<td>0.156489</td>
</tr>
<tr>
<td>7. GnRH signaling pathway</td>
<td>6</td>
<td>1</td>
<td>0.103512</td>
<td>0.161542</td>
</tr>
<tr>
<td>8. Calcium signaling pathway</td>
<td>9</td>
<td>1</td>
<td>0.110081</td>
<td>0.161542</td>
</tr>
<tr>
<td>9. Gastric acid secretion</td>
<td>6</td>
<td>1</td>
<td>0.113678</td>
<td>0.161542</td>
</tr>
<tr>
<td>10. Glutamatergic synapse</td>
<td>8</td>
<td>1</td>
<td>0.127202</td>
<td>0.171723</td>
</tr>
<tr>
<td>11. Estrogen signaling pathway</td>
<td>8</td>
<td>1</td>
<td>0.139011</td>
<td>0.178729</td>
</tr>
<tr>
<td>12. Vascular smooth muscle contraction</td>
<td>10</td>
<td>1</td>
<td>0.149579</td>
<td>0.183574</td>
</tr>
<tr>
<td>13. Inflammatory mediator regulation of TRP channels</td>
<td>8</td>
<td>1</td>
<td>0.169341</td>
<td>0.190509</td>
</tr>
<tr>
<td>14. Melanogenesis</td>
<td>10</td>
<td>1</td>
<td>0.178209</td>
<td>0.192465</td>
</tr>
<tr>
<td>15. Adrenergic signaling in cardiomyocytes</td>
<td>14</td>
<td>1</td>
<td>0.185928</td>
<td>0.193079</td>
</tr>
<tr>
<td>16. Rap1 signaling pathway</td>
<td>14</td>
<td>1</td>
<td>0.284347</td>
<td>0.284347</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Endocrine and other factor-regulated calcium reabsorption

As an example, we show partial results obtained by screening perturbed genes on chromosome three.

References


