GWAS

Shyam Visweswaran, MD, PhD
10 July 2014

Outline

- Genomic (common) diseases
- Study of genome
Genetic vs. genomic medicine

- Genetic medicine: is based on understanding the impact of single genes on disease
- Genomic medicine: is based on understanding the impact of the entire genome and environmental factors on disease

All diseases have a genetic component

Not just “genetic” diseases
10 leading causes of death

1. Heart disease (28.5%)
2. Cancer (22.8%)
3. Cerebrovascular diseases (6.7%)
4. Chronic lower respiratory disease (5.1%)
5. Injury (4.4%)
6. Diabetes (3.0%)
7. Pneumonia/Influenza (2.7%)
8. Alzheimer disease (2.4%)
9. Kidney disease (1.7%)
10. Septicemia (1.4%)

Genomic (common) diseases

- Typically, due to variations at multiple genes
- Relatively common in the population; of large public health importance
- Of great importance to affected individuals and families
- Care provided by all physicians
- **High prevalence and low severity**
Single Nucleotide Polymorphisms (SNPs)

- Commonest genomic variation
- Estimated 5-10 million SNPs in the human genome
- Very few of SNPs directly cause disease
- Non-disease SNPs may serve as markers to identify and map other genes that do cause disease

Single Nucleotide Polymorphisms (SNPs)

- SNP vs. mutation
  - Both are single nucleotide changes
  - Alleles: A = wild-type, a = mutant type
  - Genotypes: AA, Aa, aa
  - SNP = allele a occurs > 1% of population
  - Mutation = allele a occurs < 1% of population
- Mutation causes “genetic” disease
- SNP related “genomic” or “common” disease
Classification of SNPs

SNPs may occur at any position: intronic, exonic, promoter region, splice region, etc.

Hypotheses for genomic (common) diseases

- **Common Disease-Common Variant hypothesis**
  - common diseases are caused by several common variants of genes
  - each variant increases risk of disease modestly
  - an affected individual has several variants (= several SNPs)

- **Common Disease-Rare Variant hypothesis**
  - multiple rare variants underlie susceptibility to common diseases
  - an affected individual has one high risk variant (= one SNP)
Outline

- Genomic (common) diseases
- Study of genome

HapMap project

- How to study genomic diseases?
  - Sequence the entire DNA for cases and controls. 3 billion nucleotide pairs - too expensive!
  - Sequence only nucleotide pairs that vary i.e., SNPs. Estimated 10 million SNPs
- Map SNPs on the human genome
HapMap project

- HapMap project
  - 270 people from Africa, China, Japan and US
  - 3.1 million SNPs mapped so far
- SNP chips
  - 0.5 – 1 million SNPs can be identified by a single chip the size of a thumbnail
  - SNPs are chosen based on the HapMap data
  - Typically SNPs are distributed uniformly across the genome
  - Affymetrix, Illumina

GWAS

- GWAS = genome-wide association study
- Collect cases of a disease (say 1000) and controls (another 1000)
- Run a SNP chip on each to generate 0.5 – 1.0 million SNP values
- Find SNPs or combinations of SNPs that discriminate cases from control
- Two challenges in analysis of GWAS data
  1. Large number of statistical tests
  2. Epistasis
Challenge 1: Large number of statistical tests

- Simplest GWAS analysis: are there SNPs that singly increase risk of disease?
- Construct counts table for each SNP
- Do chi-squared test and compute p-value

<table>
<thead>
<tr>
<th>SNP</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseased</td>
<td>12</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Healthy</td>
<td>4</td>
<td>24</td>
<td>7</td>
</tr>
</tbody>
</table>

Large number of false positives

- 1 SNP = 1 chi-squared test
  - At p-value <= 0.05, there is 5% chance of a false positive
- 1 million SNPs = 1 million chi-squared tests
  - At p-value <= 0.05, 50,000 SNPs will be declared positive by chance alone!
- Simplest solution: Bonferroni correction
  - Use p-value <= 0.05 / 1 million = 5^-8
  - Hardly any SNP will be declared significant!
GWAS p-values

Prioritization of SNPs

- Typical solution: use a conservative p-value and validate selected SNPs in future studies
- Idea: use knowledge along with data to come up with a better list of SNPs to be studied
Knowledge

- Curated knowledge:
  - disease-specific biochemical pathways
  - functional properties of SNPs (coding vs. non-coding)
  - comparative genomics
  - prior evidence of genetic linkage
- Inferred knowledge:
  - Text-mining of medical literature
  - Data-mining of clinical records of patients

Knowledge + Data

- Systematic method for combining information across multiple sources of knowledge
- Use Bayes theorem to combine prior knowledge with evidence from data
- E.g., a SNP that is known to be in a coding region might be given a higher prior probability than one in an intergenic region

\[
P(\text{SNP} \rightarrow \text{disease} | \text{Data}) \propto P(\text{Data} | \text{SNP} \rightarrow \text{disease})P(\text{SNP} \rightarrow \text{disease})
\]

Data

Knowledge
SNP- and Gene-Based Biological Function to Set Prior Weights

- Genotyped SNPs
- Genotyped SNPs and SNPs in LD
- Genes

Prior weights:
- Regulatory impact: 2
- Protein coding change: 2
- RNA stability: 1
- Pathway: 1
- Disease: 2
- Literature: 0.5

Challenge 2: Epistasis

- Gene-gene interactions and gene-environment interactions are thought to underlie common diseases
- **Epistasis** refers to gene-gene interaction when the action of one gene is modified by one or several other genes
- Biologically, epistasis is likely mediated by physical interactions among biomolecules produced by the genes e.g., protein-protein interactions
Multifactor Dimensionality Reduction (MDR)

- Widely used method to detect linear and non-linear interactions among SNPs associated with disease
- Combinatorial method that searches over all possible combinations of SNPs
- Computationally expensive

Number of models

<table>
<thead>
<tr>
<th>#SNPs</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>200</td>
<td>19900</td>
<td>1313400</td>
<td>64684950</td>
<td>66018450</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>499500</td>
<td>166167000</td>
<td>41417124750</td>
<td>41583792250</td>
</tr>
</tbody>
</table>

MDR Algorithm
MDR

- MDR performs 10-fold cross-validation for each combination of SNPs to estimate the accuracy for each model.

A 2-SNP model for 170 cases

<table>
<thead>
<tr>
<th>SNP A</th>
<th>AA</th>
<th>AA</th>
<th>AA</th>
<th>Aa</th>
<th>Aa</th>
<th>aa</th>
<th>aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP B</td>
<td>BB</td>
<td>Bb</td>
<td>bb</td>
<td>BB</td>
<td>Bb</td>
<td>BB</td>
<td>Bb</td>
<td>Bb</td>
</tr>
<tr>
<td>Diseased</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>15</td>
<td>12</td>
<td>8</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Healthy</td>
<td>4</td>
<td>24</td>
<td>7</td>
<td>20</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>11</td>
</tr>
</tbody>
</table>