Human Variations

Marco F Ramoni, PhD
September 15th, 2005

Biomedical Computing
6.872 / HST 950
Outline

Properties of the Genome

Basics
- 80s revolution and HGP;
- Genetic polymorphisms;
- Evolution and selection;

Genetic diseases
- Tracking genetic diseases;
- Traits and complex traits;

Genomic diseases
- Blocks of heredity;
- Tracking blocks.

The Genetic Study of the Future

Candidates identification
- Find the genes;
- Find the SNPs;

Study design
- Case/control studies;
- Pedigree studies;
- Trios, sibs and TDT;

Study analysis
- Single gene association;
- Multivariate association;
- Validation.
The context: Sickle cell anemia is a monogenic disorder due to a mutation on the β-globin (HBB) at 11p15.5.

The problem: SCA phenotype ranges from asymptomatic to early childhood death.

The phenotype: SCA subjects have an increased risk of stroke (6-8%) before 18 yrs.

The hypothesis: Other genes modulate this risk of stroke.

HBB Sequence in Normal Adult Hemoglobin (Hb A):

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTG</td>
<td>Leu</td>
</tr>
<tr>
<td>ACT</td>
<td>Thr</td>
</tr>
<tr>
<td>CCT</td>
<td>Pro</td>
</tr>
<tr>
<td>GAG</td>
<td>Glu</td>
</tr>
<tr>
<td>GAG</td>
<td>Glu</td>
</tr>
<tr>
<td>AAG</td>
<td>Lys</td>
</tr>
<tr>
<td>TCT</td>
<td>Ser</td>
</tr>
</tbody>
</table>

HBB Sequence in Mutant Adult Hemoglobin (Hb S):

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTG</td>
<td>Leu</td>
</tr>
<tr>
<td>ACT</td>
<td>Thr</td>
</tr>
<tr>
<td>CCT</td>
<td>Pro</td>
</tr>
<tr>
<td>GTG</td>
<td>Val</td>
</tr>
<tr>
<td>GAG</td>
<td>Glu</td>
</tr>
<tr>
<td>AAG</td>
<td>Lys</td>
</tr>
<tr>
<td>TCT</td>
<td>Ser</td>
</tr>
</tbody>
</table>

Figure by MIT OCW.
Finding Candidate Genes

Rationale: Bar a genome-wide scan you need likely culprits.

Start: OMIM (NCBI/NIH)

Extend:
✓ Literature;
✓ Regions;
✓ Microsatellites;
✓ Mechanisms of actions (pathways);

Refinement: Cast a large net and run a wide scan on a subset of patients.

See the OMIM, Online Mendelian Inheritance in Man.
Finding The Right SNPs

Option 1. Finding the causative SNP:
   Rationale: Find the cause, select if there is a functional role.
   Drawback: What is functional? Exons, promoter, splicing, etc.

Option 2. Finding related SNPs:
   Rationale: Chose SNPs that represent the gene through LD.
   Drawback: Tough to get the causative root.

Figure removed due to copyright reasons.
Hunting Causative SNPs

Strategy: Select the SNPs on the basis of their role.

Options: Non synonymous, in exons, in promoter, in other regulatory region.

Source: dbSNP (NCBI/NIH).

Faster: Portal SNPPER.

Bonus: Primer design.

Example: Select all the SNPs in CST3 located on exons.

Filtering: From 146 to 26.

Problem: Uncovered regions.

Courtesy of Dr. Alberto A. Riva. Used with permission.
Fishing Across Genes

Rationale: Find the optimal coverage for the entire gene.

Problem: We need to know how SNPs are transmitted together in the population.

Source: HapMap.org

Hapmap: Genotype of 30 trios in 4 populations every 5k bases.

Strategy: 1) Identify blocks of LD and 2) Choose the SNPs that represent these blocks.

Figure removed due to copyright reasons.
Genome Wide Scan

- Technologies for genotyping:
- By SNP (individual primer);
- By Sample/Locus;
- Genome-wide: GeneChip® Mapping 100K Set (soon 500k) using a technology similar to expression arrays.
- 500k means 1 SNP every 6, close to the resolution of the HapMap.

Figure removed due to copyright reasons.
Study Design

Classification by sampling strategy:

Association: Unrelated subjects with/out phenotype.
Case/Control: Two sets of subjects, with and without.
Cohort: Natural emergent phenotype from study.
Pedigrees: Traditional studies focused on heredity.
Large pedigree: One family across generations.
Triads: Sets of nuclear families (parents/child).
Sib-pairs: Sets of pair of siblings.

Classification by experimental strategy:
Double sided: Case/control studies.
Single sided: e.g. trios of affected children.
Analysis Methods

- Study designs and analysis methods interact.
- We review five main analysis types:
  - **Association studies**: Case/control association.
  - **Linkage analysis**: Traditional analysis of pedigrees.
  - **Allele-sharing**: Find patterns better than random.
  - **TDT**: transmission disequilibrium test.
- Typically, these collections are hypothesis driven.
- The challenge is to collect data so that the resulting analysis will have enough power.
Association Studies

Method: Parametric method of association.
Strategy: Traditional case vs control approach.
Test: Various tests of association.
Sample: Split group of affected/unaffected individuals.
Caveats: Risk of stratifications (admixtures) - case/control split by populations.
Advantages: Easily extended to complex traits and ideal for exploratory studies.
Linkage Analysis

Method: Parametric model building.

Strategy: Compare a model with dependency between phenotype and allele against independence model.

Test: Likelihood ratio - or lod score $\log(LR)$.

\[
LR = \frac{p(\text{Data} \mid M_1)}{p(\text{Data} \mid M_0)}
\]

Sample: Large pedigree or multiple pedigrees.

Caveats: Multiple comparison, hard for complex traits.
Allele Sharing

**Method:** Non parametric method to assess linkage.

**Test:** An allele is transmitted in affected individuals more than it would be expected by chance.

**Sample:** It uses affected relatives in a pedigree, counts how many times a region is identical-by-descent (IBD) from a common ancestor, and compares this with expected value at random.

**Caveats:** Weak test, large samples required.
Transmission Disequilibrium Test

Method: Track alleles from parents to affected children.
Strategy: Transmitted=case / non transmitted=controls.
Test: Transmission disequilibrium test (TDT).
Sample: Triads of affected child and parents.
Caveats: Test is not efficient and is prone to false negatives.
Advantages: Powerful test and stratification not an issue.

6.872/HST 950
Stroke Study Design

Design: Nation-wide cohort study of over 4000 African American in 26 centers.

Subjects: 1392 SCA subjects with at least one complication from SCA (92 with stroke, 6.2%).

Genes: 80 candidate genes involved in vaso-regulation, inflammation, cell adhesion, coagulation, hemostasis, proliferation, oxidative biology and other functions.

SNPs: Coverage selected with bias to function (256).

Risk factors: $\alpha$-thalassemia, history, age, gender.

Filtering: Missing data and Hardy-Weinberg on unaffected reduces the set to 108 SNPs on 80 genes.
Single Gene Association

Method: One SNP at the time.
Analysis: Test statistics (like we had an hypothesis).
Style: Observational by pseudo hypothesis-driven.
Results: A list of SNP/genes.
Validation: Replication.

Table removed due to copyright reasons.
Please see:

Spurious Association/Confounding

- Association of shoe size (S) and literacy (L) in kids.
- If I act on S, I will not change L: If you buy bigger shoes, will your kids learn more words?
- No: age (A) make S and L conditionally independent.
Missed Associations

Gene environment interaction:

No association between genotype and phenotype

Association appears conditional on an environmental factor
Bayesian Networks

Definition: Direct acyclic graph (DAG) encoding conditional independence/dependence.

Qualitative:
- **Node**: stochastic variables (SNPs, phenotypes, etc).
- **Arcs**: Directed stochastic dependencies between parents and children.

Quantitative:
- **CPT**: Conditional probability tables (distributions) that shape the dependency.

<table>
<thead>
<tr>
<th>G</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>True</td>
</tr>
<tr>
<td>Aa</td>
<td>True</td>
</tr>
<tr>
<td>aa</td>
<td>True</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>G</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>False</td>
</tr>
<tr>
<td>Aa</td>
<td>False</td>
</tr>
<tr>
<td>aa</td>
<td>False</td>
</tr>
</tbody>
</table>

6.872/HST 950
Learning Networks

Processes: Data are generated by processes.

Probability: The set of all models is a stochastic variable $M$ with a probability distribution $p(M)$.

Selection: Find the most probable model given the data.

$$p(M | \Delta) = \frac{p(\Delta, M)}{p(\Delta)} = \frac{p(\Delta | M) p(M)}{p(\Delta)}$$

Estimation: Probabilities can be seen as relative frequencies:

$$p(x_i | \pi_i) = \frac{n(x_i | \pi_i)}{\sum_j n(x_j | \pi_i)}$$

$$p(x_j | \pi_i) = \frac{a_{ij} + n(x_j | \pi_i)}{\sum_j a_{ij} + n(x_j | \pi_i)}$$
Figure removed due to copyright reasons.


Prognostic Modeling

Prediction: The method used for the predictive validation can be used to compute the risk of stroke given a patient’s genotypes.

Prognosis: We can build tables of risks for patients and predict the occurrence of stroke in 5 years.

Extension: How about this risk scheme as a model of stroke in the general population?

<table>
<thead>
<tr>
<th>Risk</th>
<th>ANXA2.6</th>
<th>BMP6.10</th>
<th>BMP6.12</th>
<th>SELP.14</th>
<th>TGFB3.10</th>
<th>ERG.2</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rs26910500</td>
<td>rs267196</td>
<td>rs408505</td>
<td>rs3917733</td>
<td>rs284875</td>
<td>rs989554</td>
<td></td>
</tr>
<tr>
<td>0.007 (0;0.03)</td>
<td>AG</td>
<td>TT</td>
<td>TT</td>
<td>CT</td>
<td>CT</td>
<td>AG</td>
<td>1</td>
</tr>
<tr>
<td>0.06 (0;0.38)</td>
<td>AG</td>
<td>TT</td>
<td>TT</td>
<td>CT</td>
<td>CC</td>
<td>AG</td>
<td>4</td>
</tr>
<tr>
<td>0.185 (0.09;0.30)</td>
<td>AA</td>
<td>TT</td>
<td>CT</td>
<td>CC</td>
<td>CC</td>
<td>AA</td>
<td>50</td>
</tr>
<tr>
<td>0.727 (0.61;0.83)</td>
<td>AA</td>
<td>TT</td>
<td>CC</td>
<td>CC</td>
<td>CC</td>
<td>AA</td>
<td>64</td>
</tr>
<tr>
<td>0.868 (0.70;0.97)</td>
<td>GG</td>
<td>TT</td>
<td>CC</td>
<td>CC</td>
<td>CC</td>
<td>AA</td>
<td>21</td>
</tr>
<tr>
<td>0.968 (0.79;1)</td>
<td>GG</td>
<td>TT</td>
<td>CC</td>
<td>CT</td>
<td>CC</td>
<td>AA</td>
<td>8</td>
</tr>
</tbody>
</table>
Predictive Validation

Cross Validation: 98.8%.

Validation: Stroke prediction of 114 subjects in different population (not the cohort).

Accuracy: 98.2%: TPR=100%; TNR=98.1% (2 errors).

Logistic regression: Identify regressors at p-value < 0.05.

Model: 5 (SELP/BMP6) & HbF.

Accuracy: 88% accurate: TPR: 0.57% (3 errors); TNR: 0.9% (10 errors).

Figure removed due to copyright reasons.
Why we do not find the causes for complex traits?
Because we look at one gene at the time.
Genes work together (need more than one gene to get the phenotype) but also in a redundant way (phenotype through alternative paths).
Long distance disequilibrium, reveals more complex structures in the population.
Prediction is necessary.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Symbol</th>
<th>Position</th>
<th>Single Gene Accuracy</th>
<th>Cont</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADCY9</td>
<td></td>
<td>16p13.3</td>
<td>71.93%</td>
<td>2%</td>
</tr>
<tr>
<td>ANXA2</td>
<td></td>
<td>15q22.2</td>
<td>43.86%</td>
<td>2%</td>
</tr>
<tr>
<td>BMP6</td>
<td></td>
<td>6p24.3</td>
<td>83.33%</td>
<td>5%</td>
</tr>
<tr>
<td>CSF2</td>
<td></td>
<td>5q23.3</td>
<td>50.88%</td>
<td>1%</td>
</tr>
<tr>
<td>ECE1</td>
<td></td>
<td>1p36.12</td>
<td>13.15%</td>
<td>0.2%</td>
</tr>
<tr>
<td>ERG</td>
<td></td>
<td>21q22.2</td>
<td>42.98%</td>
<td>1%</td>
</tr>
<tr>
<td>MET</td>
<td></td>
<td>7q31.2</td>
<td>23.68%</td>
<td>1%</td>
</tr>
<tr>
<td>SCYA</td>
<td></td>
<td>17q11.2</td>
<td>55.14%</td>
<td>1%</td>
</tr>
<tr>
<td>SELP</td>
<td></td>
<td>1q24.2</td>
<td>80.70%</td>
<td>7%</td>
</tr>
<tr>
<td>TEK</td>
<td></td>
<td>9p21.2</td>
<td>8%</td>
<td>1%</td>
</tr>
<tr>
<td>TGFBR3</td>
<td></td>
<td>1p22.1</td>
<td>50.88%</td>
<td>2%</td>
</tr>
<tr>
<td>HbF.P</td>
<td></td>
<td></td>
<td>72.81%</td>
<td>1%</td>
</tr>
</tbody>
</table>
Take Home Messages

There are two types of science: physics and stamp collecting.  
Ernest Rutherford

Revolution: The –omic scale changes the way of biomedical sciences, makes it predictive/quantitative.

Discovery: The genome is too complex for simple hypothesis, hypotheses have to be discovered.

Proof: The burden of proof has to be based on prediction, as we expect from good science.

Potential: The potential of this changes goes beyond the still fantastic power to understand and heal.