Structural Genomics

Marco Ramoni
Children’s Hospital Informatics Program
Division of Health Sciences and Technology
Harvard Medical School and
Massachusetts Institute of Technology

October 2, 2009
HST 950
central dogma of molecular biology
On February 12, 2001 the Human Genome Project announces the completion of a first draft of the human genome.

Among the items on the agenda of the announcement, a statement figures prominently:

*A SNP map promises to revolutionize both mapping diseases and tracing human history.*

SNP are Single Nucleotide Polymorphisms, subtle variations of the human genome across individuals.

You can take this sentence as the announcement of a new era for population genetics.
One of the most striking statistics of the HGP is that 99.9% of the genome is expected to be identical.

In reality it is 99% identical.

The remaining 1% is mostly made by single nucleotide polymorphisms (SNPs).

A SNP is a the change of single base occurring in at least 5% of the population.
**terminology**

**Allele:** A sequence of DNA bases.

**Locus:** Physical location of an allele on a chromosome.

**Linkage:** Proximity of two alleles on a chromosome.

**Phenotype:** An outward, observable character (trait).

**Genotype:** The internally coded, inheritable information.

**Penetrance:** No. with phenotype / No. with allele.

**Distance (Physical):** Number of bases between two loci.

**Distance (Recombination):** Probability that two alleles will be transmitted together in the next generation.

**cMorgan:** Distance measure between two loci means that they have 1% chances of being separated (1 Mb).
genetic variations data

Courtesy of A. Riva
SNP microarrays: technology able to genotype 1 million SNPs in a single individual.

Advantage: no need to identify candidate genes and candidate SNPs.

Challenge: the dimensionality of the database – 1 million by number of subjects.

Common solution: filter one variation at the time.

Drawback: paradox of one variation at the time.
Task: Find basis (genotype) of diseases (phenotype).
Marker: Flag genomic regions in linkage disequilibrium.
Problem: *Real* genotype is not observable.
Strategy: Use marker as genotype proxy.
Condition: Linkage disequilibrium.
Dependency: Observable measure of dependency between marker and phenotype.
Linkage equilibrium: Loci Aa and Bb are in equilibrium if transmission probabilities $p_A$ and $p_B$ are independent.

$$p_{AB} = p_A p_B.$$ 

Haplotype: A combination of allele loci: $p_{AB}$, $p_{Ab}$, $p_{aB}$, $p_{ab}$.

Linkage disequilibrium: Loci linked in transmission as.

$$r^2 = \frac{(p_{AB} - p_A * p_B)}{(p_A * p_B * p_a * p_b)}$$

a measure of dependency between the two loci.

Markers: Linkage disequilibrium is the key of markers.
- LD (r²) distances can be used to identify haplotypes.
- Haplotypes are groups of SNPs transmitted in “blocks”.
- These blocks can be characterized by a subset of their SNPs (tags).
- Since they are the result of an underlying evolutionary process, they can be used to reconstruct ancestral DNA.
Dely et al. report an high-resolution analysis of the haplotype structure of 500Kbs on 5q31.

The resulting picture portraits the stretch separated in 11 blocks separated by recombination points.

Haplotype patterns travel together (blocks in LD) and therefore the authors infer 4 ancestral haplotypes.
Haplotypes: Not all combinations appear, we need fewer SNPs.  
Goal: Smallest set of SNPs deriving all the other SNPs.  
htSNPs: Tagging SNPs are called haplotype tagging SNPs.  
Problem: Intractable task (for 136 bases any relativistic machine would take more than the age of the universe).
stroke in sickle cell anemia

Anemia: SCA is caused by a single variant on β-globin.

Problem: phenotype ranges from asymptomatic to early childhood death.

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

Importance: a predictor of stroke would focus therapy.

Hypothesis: other genes modulate this risk of stroke.

Sebastiani et al, Nat Genet, 2005
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.
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.
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.
Rationale: Find the optimal coverage for an 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 5Kb.
Strategy: 1) Identify blocks of LD and 2) Choose the SNPs that represent blocks.
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.
Method: Parametric model building.

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

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

Sample: Large pedigree or multiple pedigrees.

Caveats: Multiple comparison, hard for complex traits.

$$LR = \frac{p(Data | M_1)}{p(Data | M_0)}$$
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.