Population Genetics in the Post-Genomic Era

Marco F. Ramoni

Children's Hospital Informatics Program

Harvard Medical School

HST 950 (2003)

Harvard-MIT Division of Health Sciences and Technology
HST.950J: Medical Computing

Peter Szolovits, PhD
Isaac Kohane, MD, PhD
Lucila Ohno-Machado, MD, PhD

Introduction

•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.

Outline

- Background
  - 80s revolution and HGP
- Genetic Polymorphisms
  - Their nature
  - Types of polymorphisms
- Foundations
  - Terminology
  - Hardy Weinberg Law
  - Types of inheritance
- Complex Traits
  - Definition
  - Factors of Complexity
- Study and Experiment Design
  - Case Control Studies
  - Pedigree Studies
- Analysis Methods
  - Association Studies
  - Linkage Studies
  - Allele-sharing Studies
  - QTL Mapping
- The New Ways
  - Haplotypes
  - HapMap
  - htSNPs

Background

Intuition: We can find the genetic bases of observable characters (like diseases) without knowing how the actual coding works.

Origins: Sturtevant (1913) finds traits-causing genes.

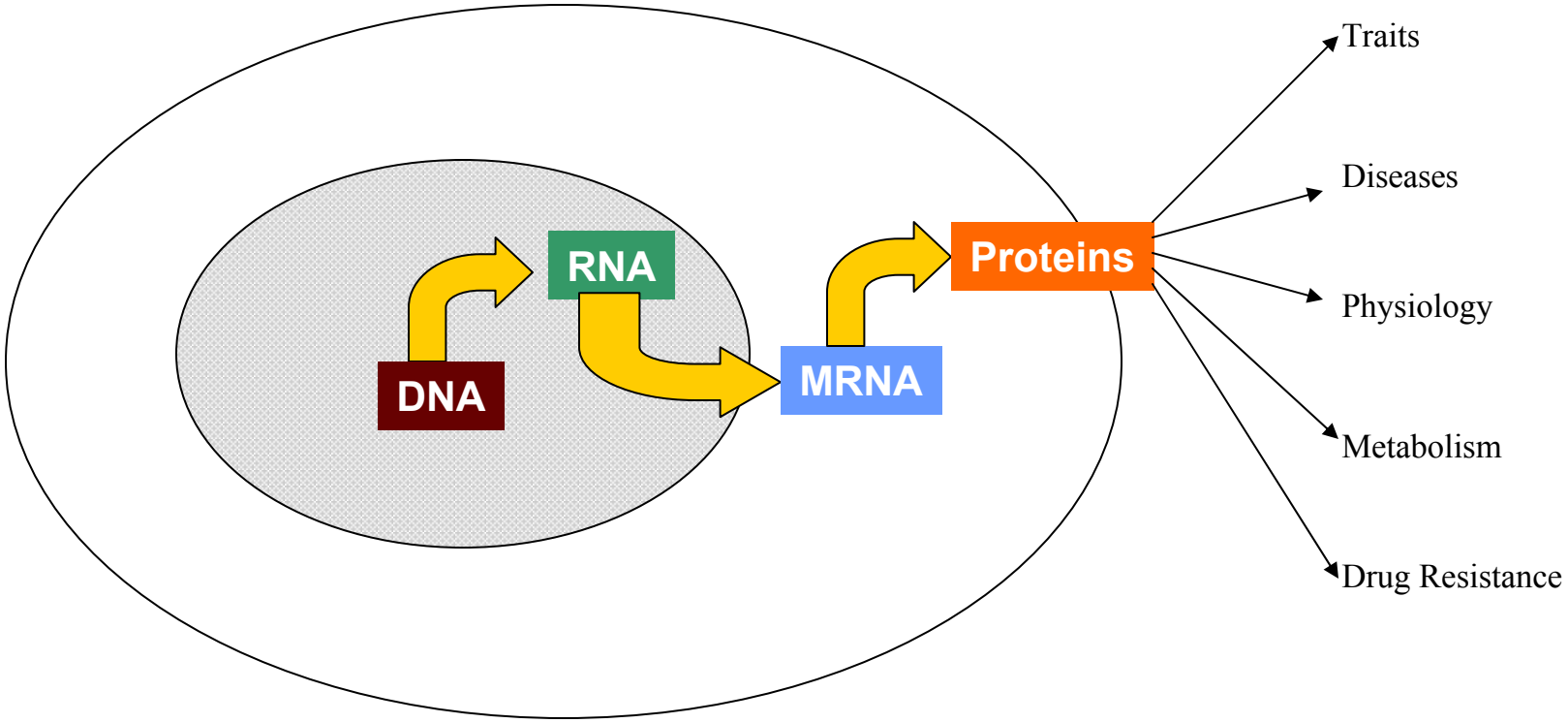
Early History: Genetic maps of plants and insects.

Outcast: Ernst Mayr called it ¡°Beans bag genetics¡±.

Reasons: No markers to identify coding regions.

Markers: Botstein (1977) showed that naturally occurring DNA already contains markers identifying regions of the genome: polymorphisms.

Central Dogma of Molecular Biology



The 80s Revolution and the HGP

•The intuition that polymorphisms could be used as markers sparkled the revolution.

•Mendelian (single gene) diseases:

Autosomal dominant (Huntington).

Autosomal recessive (C Fibrosis).

X-linked dominant (Rett).

X-linked recessive (Lesch-Nyhan).

•Today, over 400 single-gene diseases have been identified.

•This is the promise of the HGP.

Terminology

Allele: A sequence of DNA bases.

Locus: Physical location of an allele on a chromosome.

Linkage: Proximity of two alleles on a chromosome.

Marker: An allele of known position on a chromosome.

Distance: Number of base-pairs between two alleles.

centiMorgan: Probabilistic distance of two alleles.

Phenotype: An outward, observable character (trait).

Genotype: The internally coded, inheritable information.

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

Distances

•Physical distances between alleles are base-pairs.

But the recombination frequency is not constant.

•Segregation (Mendel's first law): Allele pairs separate during gamete formation and randomly reform pairs.

•A useful measure of distance is based on the probability of recombination: the Morgan.

•A distance of 1 centiMorgan (cM) between two loci means that they have 1% chances of being separated by recombination.

•A genetic distance of 1 cM is roughly equal to a physical distance of 1 million base pairs (1Mb).

More Terminology

Physical maps: Maps in base-pairs.

Human autosomal physical map: 3000Mb (bases).

Linkage maps: Maps in centiMorgan.

Human Male Map Length: 2851cM.

Human Female Map Length: 4296cM.

Correspondence between maps:

Male cM ~ 1.05 Mb; Female cM ~ 0.88Mb.

Cosegregation: Alleles (or traits) transmitted together.

Hemophilia, a Sex Linked Recessive

•Hemophilia is a X-linked recessive disease, that is fatal for women.

•X-linked means that the allele (DNA code which carries the disease) is on the X-chromosome.

•A woman (XX) can be carrier or non-carrier: if x=allele with disease, then xX=carrier; xx=dies; XX=non carrier.

•A male (YX) can be affected or not affected: (xY= affected; XY=not affected).

Hemophilia: A Royal Disease

Genetic Markers

One of the most celebrated findings of the human genome project is that humans share most DNA.

Still, there are subtle variations:

Simple Sequence Repeats (SSR): Stretches of 1 to

6 nucleotide repeated in tandem.

Microsatellite**:** Short tandem repeat (e.g. GATA) varying in number between individuals.

Single nucleotide polymorphism (SNP): Single base variation with at least 1% incidence in population.

Single Nucleotide Polymorphisms

Variations of a single base between individuals:

```
... ATGCGATCGATACTCGATAACTCCCGA ...
... ATGCGATCGATACGCGATAACTCCCGA ...
```

A SNP must occur in at least 1% of the population.

SNPs are the most common type of variations.

Differently to microsatellites or RTLPs, SNPs may occur in coding regions:

cSNP: SNP occurring in a coding region.

rSNP: SNP occurring in a regulatory region.

sSNP: Coding SNP with no change on amino acid.

Evolutionary Pressure

•Kreitman (1983) sequenced the first 11 alleles from nature: alcohol dehydrogenase locus in Drosophila.

•11 coding regions / 14 sites have alternative bases.

•13 variations are silent: ie do not change amino acid.

•With a random base change, we have 75% chances of changing the amino acid (i.e. creating a cSNP).

•Why this disparity?

•Drosophilae and larvae are found in fermenting fruits.

•Alcohol dehydrogenase is important in detoxification.

•A radical change in protein is a killer.

Hardy-Weinberg Law

Hardy-Weinberg Law (1908): Dictates the proportion of major (p), minor alleles (q) in equilibrium.

$$p^2 + 2pq + q^2 = 1.$$

Equilibrium: Hermaphroditic population gets equilibrium in one generation, a sexual population in two.

Example: How many Caucasian carriers of C. fibrosis?

Affected Caucasians (q^2) = 1/2,500.

Affected Alleles (q)= 1/50 = 0.02.

Non Affected Alleles (p) = (1 - 0.02) = 0.98.

Heterozygous (2pq) = 2(0.98 ¡Á 0.02) = 0.04= 1/25.

Assumptions

Random mating: Mating independent of allele.

- Inbreeding: Mating within pedigree;
- Associative mating: Selective of alleles (humans).

Infinite population: Sensible with 6 billions people.

Drift: Allele distributions depend on individuals offspring.

- Locality: Individuals mate locally;
- Small populations: Variations vanish or reach 100%.

Mutations contrast drift by introducing variations.

Heresy: This picture of evolution as equilibrium between drift and mutation does not include selection!

Natural Selection

Example: p=0.6 and q=0.4.

Fitness (w): AA=Aa=1, aa=0.8. Selection: s = 1-w = 0.2:

$$\delta p = \frac{Spq^2}{1-sq^2} = \frac{(0.2)(0.6)(0.4)^2}{1-(0.2)(0.4)^2} = \frac{0.019}{0.968} = 0.02$$

Selection: Effect on the 1st generation is A=0.62 a=0.38.

Rate: The rate decreases. Variations do not go away.

Does it work?

Race and Sanger (1975) 1279 subjects¡¯ blood group.

p = p(M) = (2 x 363) + 634 / (2 x 1279) = 0.53167.

Caveat: Beta-hemoglobin sickle-cell in West Africa:

	AA	*AS*	*SS*
Observed	25,374	5,482	64
Expected	25,561.98	5,106.03	254.98

Reason: Heterozygous selective advantage: Malaria.

Linkage Equilibrium/Disequilibrium

Linkage equilibrium: Loci Aa and Bb are in equilibrium if transmission probabilities π_A and π_B are independent.

$$\pi_{AB} = \pi_A \pi_B.$$

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

Linkage disequilibrium: Loci linked in transmission as.

$$y^2 = \frac{(\Pi_{AB} - \Pi_A \Pi_B)^2}{\Pi_A \Pi_B \Pi_a \Pi_b}$$

- a measure of dependency between the two loci.

Markers: Linkage disequilibrium is the key of markers.

Phenotype and Genotype

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.

Complex Traits

Problem: Traits don¡¯t always follow single-gene models.

Complex Trait: Phenotype/genotype interaction.

Multiple cause: Multiple genes create phenotype.

Multiple effect: Gene causes more than a phenotype.

Caveat: Some Mendelian traits are complex indeed.

Sickle cell anemia: A classic Mendelian recessive.

Pattern: Identical alleles at beta-globulin locus.

Complexity: Patients show different clinical courses, from early mortality to unrecognizable conditions.

Source: X-linked locus and early hemoglobin gene.

Reasons for Complex Traits

Incomplete Penetrance: Some individuals with genotype do not manifest trait. Breast cancer / BRCA1 locus.

Genetic Heterogeneity: Mutation of more than one gene can cause the trait. Difficult in non experiment setting.

- Retinitis pigmentosa: from any of 14 mutations.

Polygenic cause: Require more than one gene.

- Hirshsprung disease: needs mutation 13c and 21c.

Study Design

- Classification by sample strategy:
  - 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.
  - Case/control: Unrelated subjects with/out phenotype.
- Classification by experimental strategy:
  - Double sided: Case/control studies.
  - Single sided: e.g triads of affected children.

Analysis Methods

•Study designs and analysis methods interact.

•We review five main analysis types:

Linkage analysis: Traditional analysis of pedigrees.

Allele-sharing: Find patterns better than random.

Association studies: Case/control association.

TDT: transmission disequilibrium test.

Experimental crosses: Crosses in controlled setting.

•Typically, these collections are hypothesis driven.

•The challenge is to collect data so that the resulting analysis will have enough power.

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).

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

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 withexpected value at random.

Caveats: Weak test,large samples required.

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.

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.

Quantitative Trait Locus Maps

Method: Complex trait variable in intensity (alcoholism).

Complex Trait: Different since all about phenotype.

Strategy: Controlled mating to induce recombination.

Sample: Animal populations.

Test: Pedigree-based transmission tests.

Caveats: In humans, it is hard to have the controlled conditions of an animal setting.

Advantages: Highly controlled setting and, as in the case of hypertension in rats, animal models can stimulate and direct research in humans.

Genotyping Cohort Studies

•Traditionally, data collection is phenotype driven.

•During the past 3 years, an increasing number of large cohort studies started subjects genotyping.

- •Nurses Health Study.
- •Framingham Hearth Study.

•This process will provide genotypes for a vast array of phenotypes recorded in these cohort studies.

•This situation calls for a brand new generation of analytical tools able to provide high throughput *phenotyping* of whole-genome genotypes.

Feasibility: Time and Cost

Base: Number of SNPs per individual: 3,000,000

Costs: How much for a genome-wide SNP scan?

Cost of 1 SNP: 0.30-0.45$ (soon 0.10-0.20$)

Cost of a 10kb SNP map/individual: 90,000 (30,000)

Cost of a 1000 individuals study: 90,000k (30,000k)

Cost of 1000 complete maps: 900,000k (300,000k)

Time: How long does it take?

1 high throughput sequencer: 50,000 SNPs/day

Effort 1000 10kb SNP maps: ~700 days/man

Effort 1000 complete SNP maps: ~7000 days/man

Haplotypes

•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.

The Importance of Haplotypes

•Haplotypes make a SNP map of the human genome redundant: as some SNPs will be transmitted together, we only need a subset of SNPs to tag the entire region.

•NHGRI launched in October the HapMap project: *a description of the set of haplotype blocks and the SNPs that tag them. The HapMap will be valuable because it will reduce the number of SNPs required to examine the entire genome for association with a phenotype from all 10 million common SNPs to perhaps 200,000 to 300,000 htSNPs*.¡±

Identifying Haplotypes

•Dely et at. report an high-resolution analysis of the haplotype structure of a stretch of chromosome 5q31
•500Kbs long.

•There are 103 SNPs in the stretch.

•The SNPs were selected if the minor allele frequency was higher than 5%.

•Sample were 129 trios (nuclear families) of European descent with children affected by Crohn disease.

•Therefore, they had 258 transmitted and 258 non- transmitted chromosomes.

Haplotype Blocks

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.

htSNPs Identification

•Johnson et al. propose a transmission-based method to identify Haplotype Tag SNPs (htSNPs), the necessary SNPs to identify an haplotype.

•The method is claimed to capture the *majority* (80%) of the haplotype diversity observed within a region.

•They genotyped polymorphisms in INS H19 SDF1 TCF8 and GAS2 in 418 UK families with at least 2 siblings with diagnosed type 1 diabetes.

•They constructed haplotypes at CASP8 CASP10 and CFLAR of 598 Finnish families with type 1 diabetes.

Characterizing Phenotypes

Simple phenotypes: Mendelian diseases usually have also the advantage of simple (binary) phenotypes.

Complex diseases: Twenty years of AI in medicine show that often diseases do not obey these patterns.

Dissecting phenotypes: It is critical as dissecting gene expression patterns.

Animal models: QTL strategy may be an answer.

Opportunity: Better clinical definition of disease states.

Clinical data: With dropping costs of sequencing, good clinical data about patients are the real wealth.

Take Home Messages

The Past:

Hypothesis-driven phenotype-focused data;

Interesting discoveries with simple models.

The Present:

HGP changes the perspectives of genetic studies;

SNPs are a critical tool to break the code;

HGP technology makes sequencing a commodity.

The Future:

Exploratory genotyping will streamline discoveries;

Phenotypes will be the real goodies of the future;

The challenge is to handle complex traits.

A Must Read List

Human genome mapping:

Genomics Issue, *Nature*, February 2001

Genomics Issue, *Science*, February 2001

Visions of polymorphisms:

ES Lander and NJ Schork, *Science*, **265**, 1994

DG Wang *et al., Science,* **280**, 1998

ES Lander, *Nat Gen*, **21**, 1999

MJ Daly *et al., Nat Gen*, **29**, 2001

Visions of population genetics:

LL Cavalli-Sforza, *Genes, People, Languages*, 2001

Supported by NSF (ECS-0120309) and NIH/NHLBI (HL-99-024)