Population Genetics in the Genomic Era

Marco F. Ramoni
Children's Hospital Informatics Program and
Harvard Partners Center for Genetics and Genomics
Harvard Medical School

HST 950

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

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.

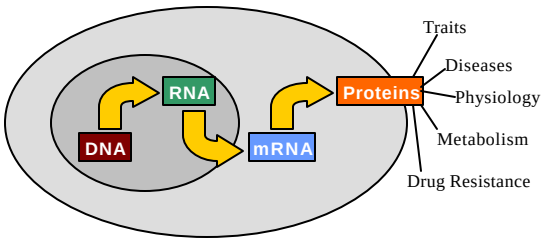
Early History: Genetic maps of plants and insects.

Outcast: Ernst Mayr called it " Beans bag genetics" .

Reasons: No markers to identify coding regions.

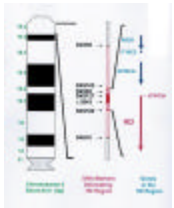
Markers: Botstein (1977) showed that naturally occurring DNA already contains markers identifying regions of the genome: polymorphisms.

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.

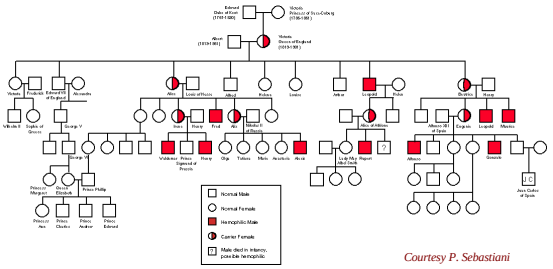
Correspondence between maps:
Male cM ~ 1.05 Mb; Female cM ~ 0.88Mb.

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 x X=carrier; xx=dies; XX=non carrier.
- A male (YX) can be affected or not affected: (xY= affected; XY=not affected).

Hemophilia: A Royal Disease



Courtesy P. Sebastiani

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.

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.

HST 512/513

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.

Reading SNP Maps

SNPer snapshot
Courtesy A. Riva

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 × 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.

AA	Aa	aa
36%	48%	16%

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

$$dp = \frac{spq^2}{1-sq^2} = \frac{(0.2)(0.6)(0.4)^2}{1-(0.2)(0.4)^2} = \frac{0.019}{0.968} = 0.02$$

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

AA	Aa	aa
39.7%	46.6%	13.7%
+3.7%	-1.4%	-2.3%

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.

	MM	MN	NN
Observed	363	634	282
Expected	361.54	636.93	280.53

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

Does it work?

Race and Sanger (1975) 1279 subjects' blood group.
p = p(M) = (2 x 363) + 634 / (2 x 1279) = 0.53167.

	MM	MN	NN
Observed	363	634	282
Expected	361.54	636.93	280.53

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.

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

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

Linkage disequilibrium: Loci linked in transmission as.

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

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).

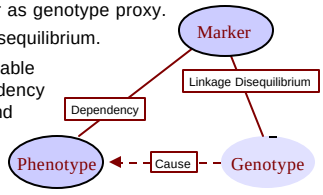
Marker: Flag genomic regions in linkage disequilibrium.

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.

Age	40	55	80
Carrier	37%	66%	85%
Non Carrier	0.4%	0.3%	8%

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.

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

marco_ramoni@harvard.edu