Harvard-MIT Division of Health Sciences and Technology
HST.508: Quantitative Genomics, Fall 2005
Instructors: Leonid Mirny, Robert Berwick, Alvin Kho, Isaac Kohane

Children's Hospital Informatics Program
Harvard Medical School

# Human Variations
## Genes, Genotypes and Generations

Marco F. Ramoni
Children's Hospital Informatics Program at
Harvard-MIT Division of Health Sciences and Technology
Harvard Partners Center for Genetics and Genomics
Harvard Medical School

# 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

## 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;

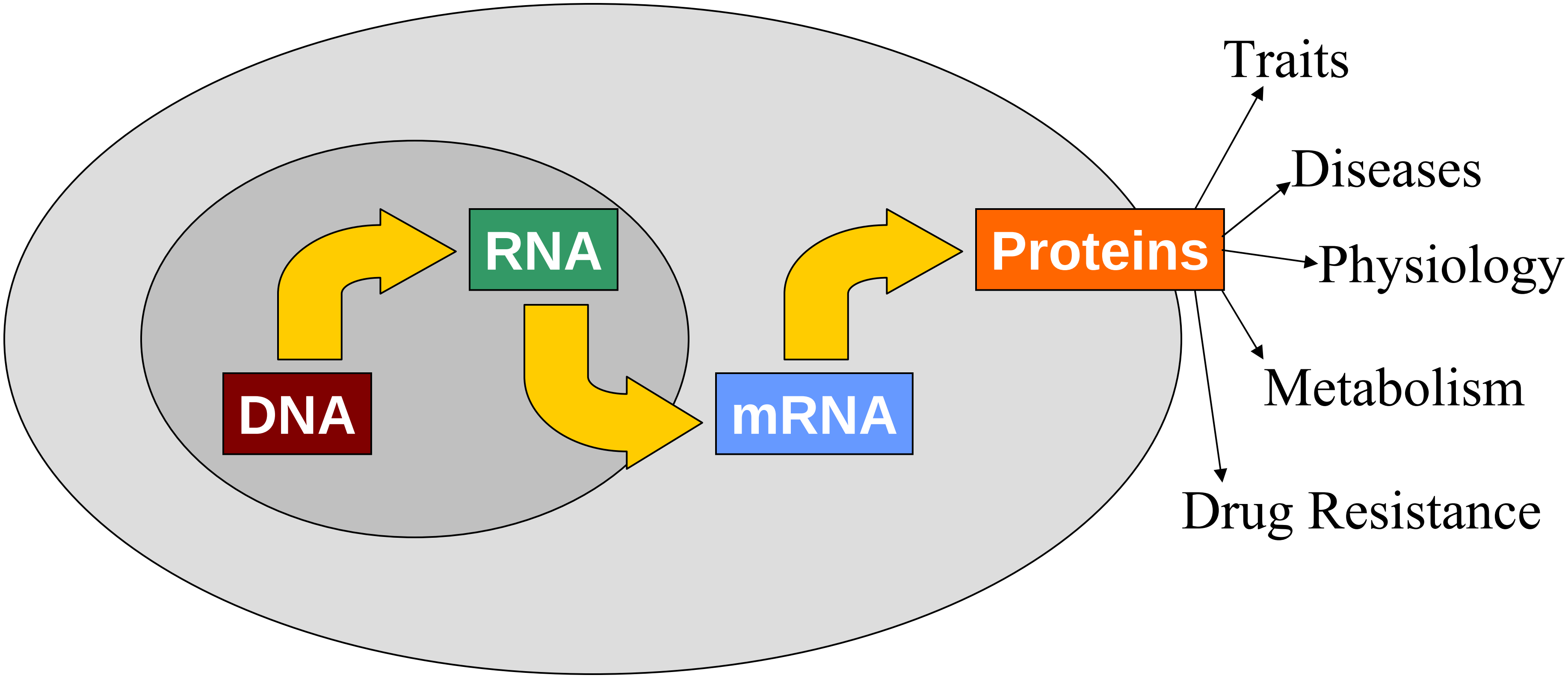
Study design

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

Study analysis

- Single gene association;
- Multivariate association;
- Validation.

# 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.

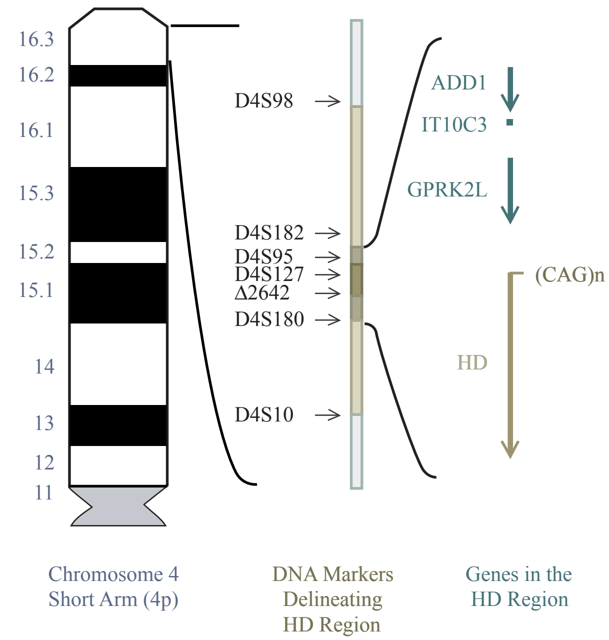



Figure by MIT OCW.

# 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.

Phenotype: An outward, observable character (trait).

Genotype: The internally coded, inheritable information.

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

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

Cosegregation: Alleles (or traits) transmitted together.

# Distances

Physical distance: 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.

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

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

Physical maps: in base-pairs. (Human autosomal map: 3000Mb).

Linkage maps: in centiMorgan (Male 2851cM, Female: 4296cM).

Physical/Linkage: A genetic distance of 1 cM is roughly equal to a physical distance of 1 million base pairs (1Mb).

# 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

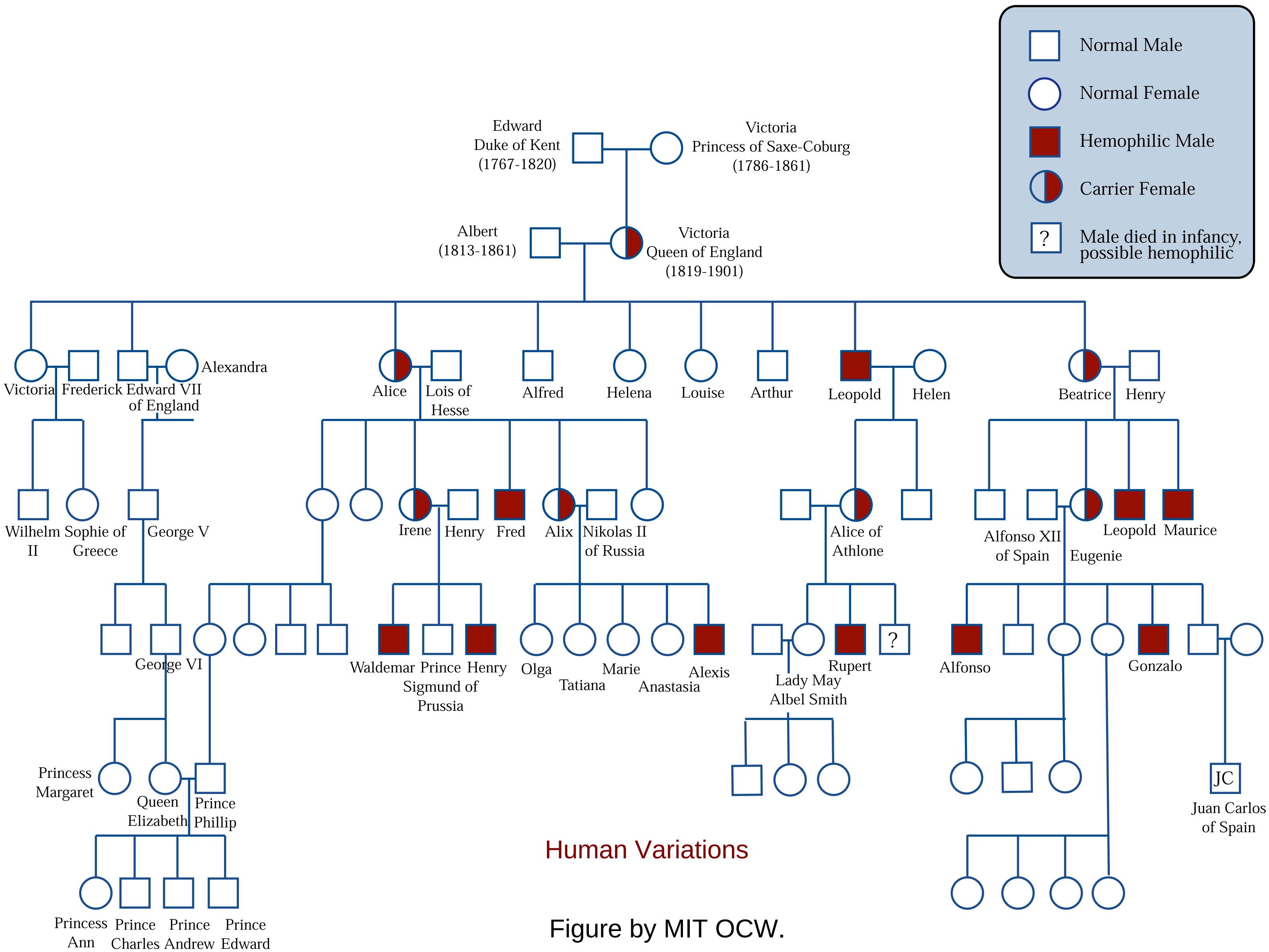





Figure by MIT OCW.

# Single Nucleotide Polymorphisms

- Variations of a single base between individuals:

  **... ATGCGATCGATACTCGATAACTCCCGA ...**

  **... ATGCGATCGATACGCGATAACTCCCGA ...**

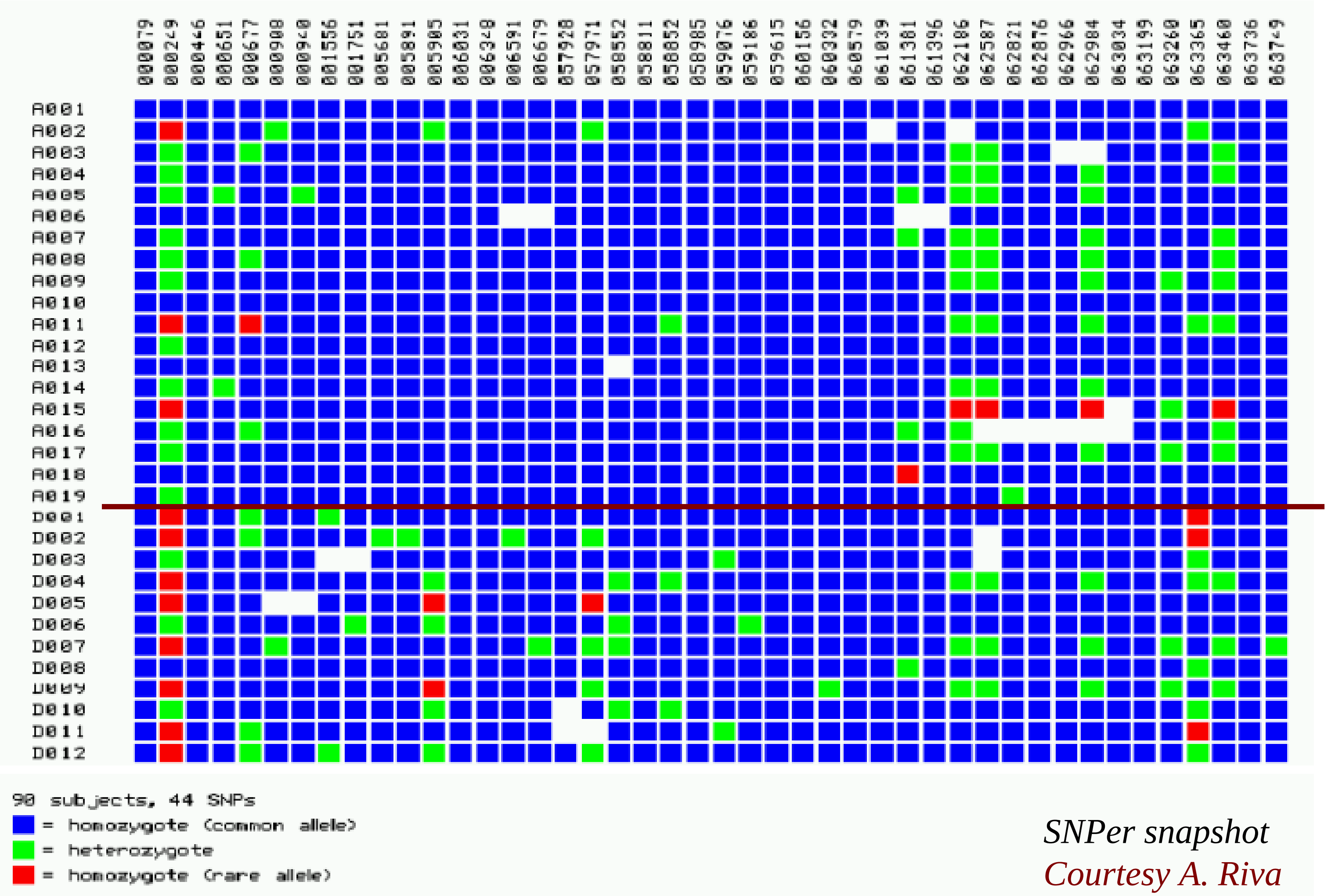
- A SNP must occur in at least 1% of the population.
- SNPs are the most common type of variations.
- Differently to microsatellites or RTLPs, SNPs may occur in coding regions:

  cSNP: SNP occurring in a coding region.

  rSNP: SNP occurring in a regulatory region.

  sSNP: Coding SNP with no change on amino acid.

# 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 ($2pq$) = $2(0.98 \times 0.02) = 0.04$= 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:

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

Selection: Effect on the $1^{st}$ 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 |

# Not Always

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 $\pi_A$ and $\pi_B$ 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.

$$r^2 = \frac{(\pi_{AB} - \pi_A \pi_B)^2}{\pi_A \pi_B \pi_a \pi_b}$$

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).

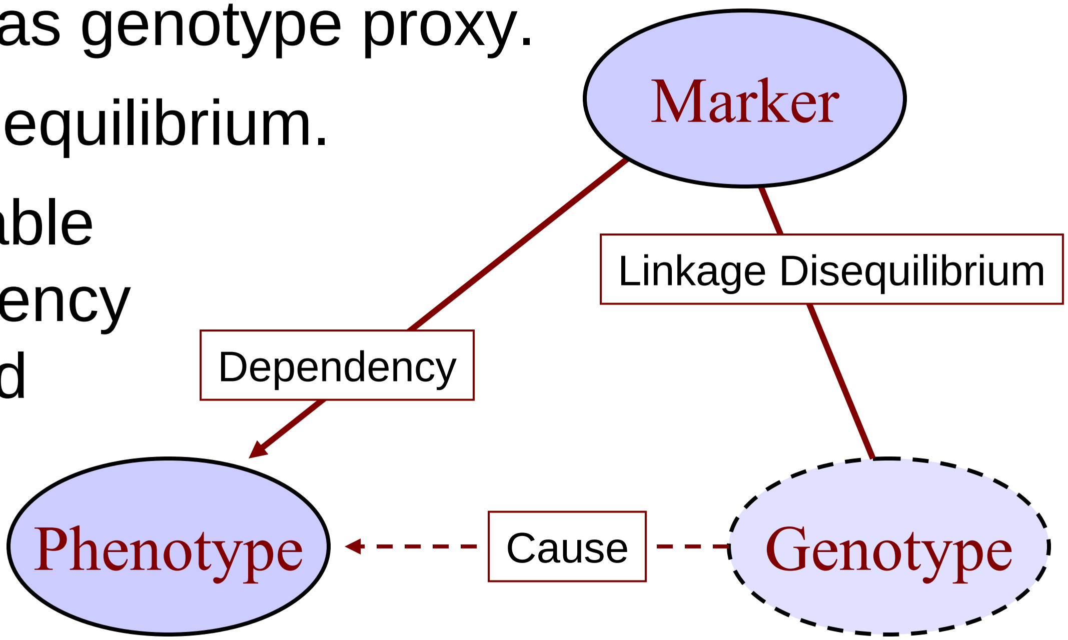
Marker: Flag genomic regions in linkage disequilibrium.

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.

# 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.

Figure removed due to copyright reasons.

# 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.

Figure removed due to copyright reasons.

# Haplotype Tagging

Haplotypes: As not all combinations appear, we need fewer SNPs.

Goal: Smallest set of SNPs deriving all the other SNPs.

htSNPs: These 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).

Figure removed due to copyright reasons.

# The Genomic Study of the Future

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.

Figure removed due to copyright reasons.

# 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.

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.

# Association Studies

Method: Parametric method of association.

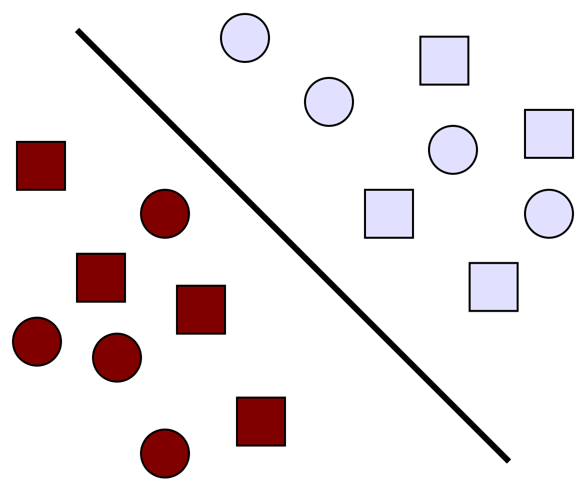
Strategy: Traditional case vs control approach.

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.



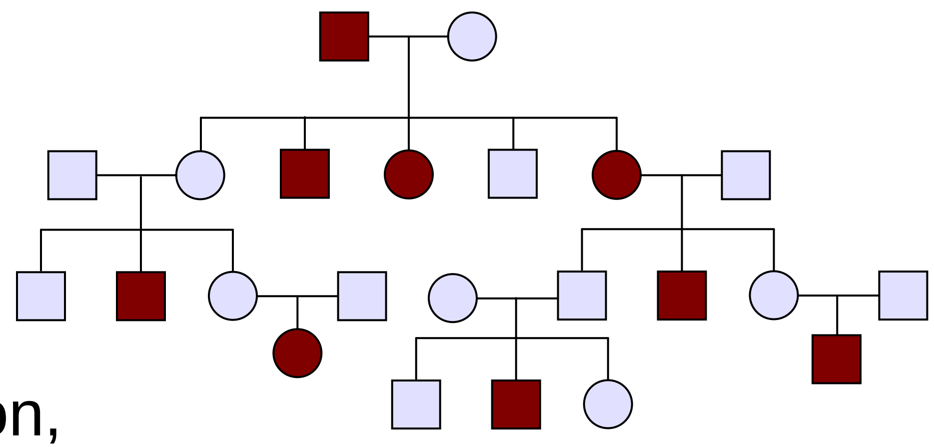
# Linkage Analysis

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).

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



Sample: Large pedigree or multiple pedigrees.

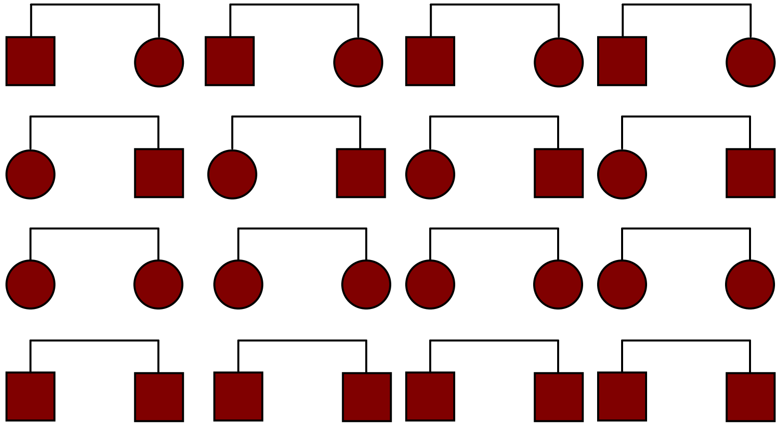
Caveats: Multiple comparison, hard for complex traits.

# 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 with expected value at random.



Caveats: Weak test, large samples required.

# Transmission Disequilibrium Test

Method: Track alleles from parents to affected children.

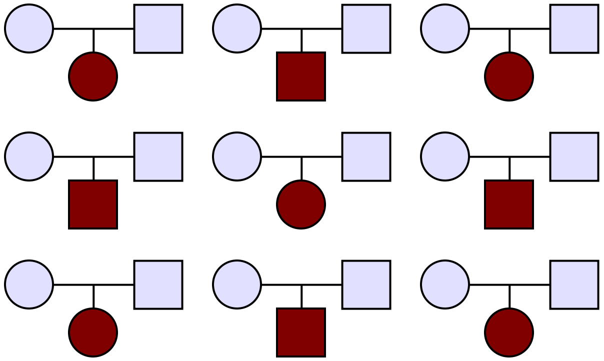
Strategy: Transmitted=case / non transmitted=controls.

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.



# 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.

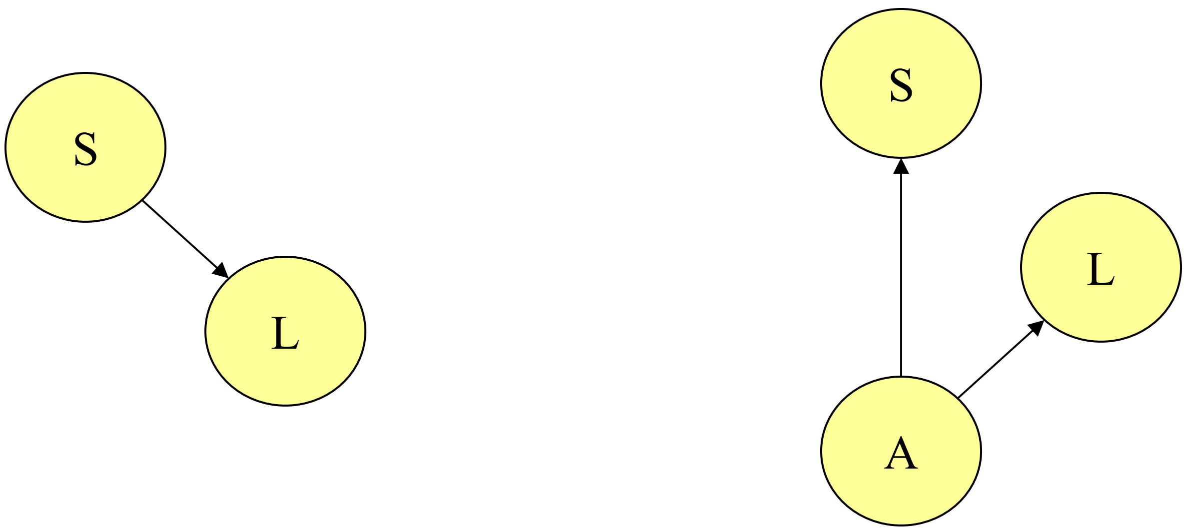
Validation: Replication.

Table removed due to copyright reasons.

Please see table 2 in Hoppe, et al. "Gene interactions and stroke risk in children with sickle cell anemia." *Blood* 103 (Mar 2004): 2391-2396.

# Spurious Association/Confounding

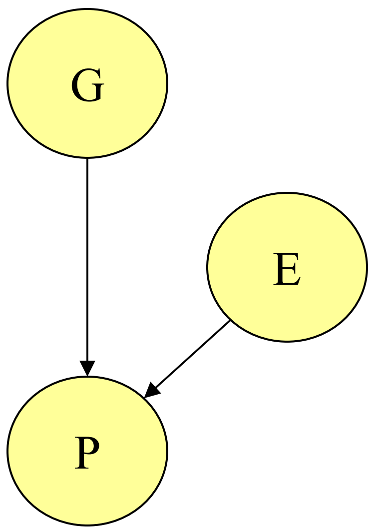
- Association of shoe size (S) and literacy (L) in kids.
- If I act on S, I will not change L: If you buy bigger shoes, will your kids learn more words?
- No: age (A) make S and L conditionally independent.



# Missed Associations

Gene environment interaction:


No association between genotype and phenotype




Association appears conditional on an environmental factor

# Bayesian Networks

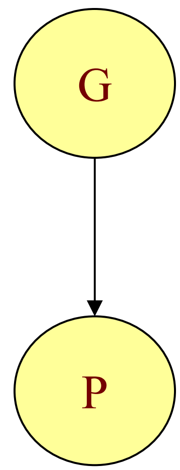
Definition: Direct acyclic graph (DAG) encoding conditional independence/dependence.

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.




| G | | |
|---|---|---|
| *AA* | *Aa* | *aa* |
| 0.3 | 0.6 | 0.1 |

| G | P | |
|---|---|---|
| | *True* | *False* |
| AA | 0.3 | 0.7 |
| Aa | 0.5 | 0.5 |
| aa | 0.9 | 0.1 |

# Learning Networks

Processes: Data are generated by processes.

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

Selection: Find the most probable model given the data.

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

Estimation: Probabilities can be seen as relative frequencies:

$$p(x_j \mid \pi_i) = \frac{n(x_j \mid \pi_i)}{\sum_j n(x_j \mid \pi_i)} \qquad p(x_j \mid \pi_i) = \frac{a_{ij} + n(x_j \mid \pi_i)}{\sum_j a_{ij} + n(x_j \mid \pi_i)}$$

# Network

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?

| Risk | ANXA2.6 *hCV26910500* | BMP6.10 *rs267196* | BMP6.12 *rs408505* | SELP.14 *rs3917733* | TGFBR3.10 *rs284875* | ERG.2 *rs989554* | N |
|---|---|---|---|---|---|---|---|
| 0.007 (0;0.03) | AG | TT | TT | CT | CT | AG | 1 |
| 0.06 (0;0.38) | AG | TT | TT | CT | CC | AG | 4 |
| 0.185 (0.09;0.30) | AA | TT | CT | CC | CC | AA | 50 |
| 0.727 (0.61;0.83) | AA | TT | CC | CC | CC | AA | 64 |
| 0.868 (0.70;0.97) | GG | TT | CC | CC | CC | AA | 21 |
| 0.968 (0.79;1) | GG | TT | CC | CT | CC | AA | 8 |

# Predictive Validation

**Cross Validation**: 98.8%.

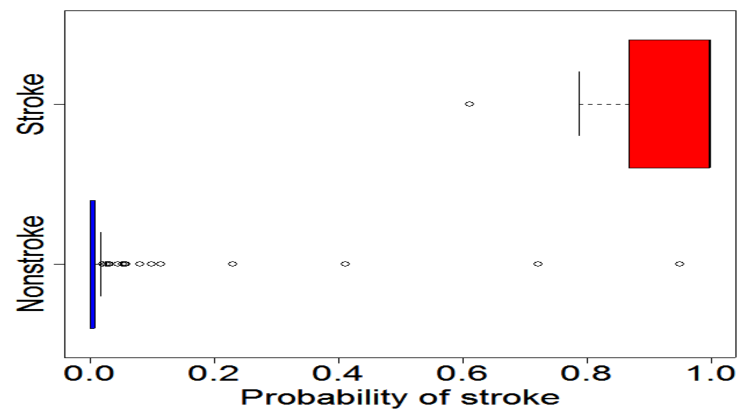
**Validation**: Stroke prediction of 114 subjects in different population (not the cohort).

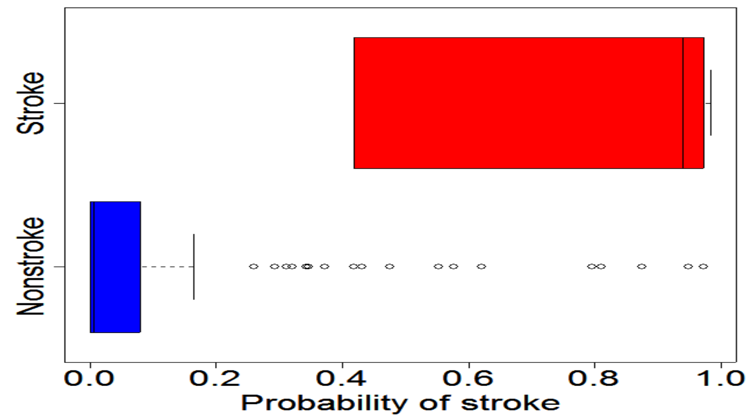
**Accuracy**: 98.2%: TPR=100%; TNR=98.1% (2 errors).

**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).



# A Holistic System

- 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.

| Gene Symbol | Position | Single Gene *Accuracy* | *Cont* |
|---|---|---|---|
| ADCY9 | 16p13.3 | 71.93% | 2% |
| ANXA2 | 15q22.2 | 43.86% | 2% |
| BMP6 | 6p24.3 | 83.33% | 5% |
| CSF2 | 5q23.3 | 50.88% | 1% |
| ECE1 | 1p36.12 | 13.15% | 0.2% |
| ERG | 21q22.2 | 42.98% | 1% |
| MET | 7q31.2 | 23.68% | 1% |
| SCYA | 17q11.2 | 55.14% | 1% |
| SELP | 1q24.2 | 80.70% | 7% |
| TEK | 9p21.2 | 8% | 1% |
| TGFBR3 | 1p22.1 | 50.88% | 2% |
| HbF.P | | 72.81% | 1% |

# Human Variation Omnibus

Definition: The Human Variation Omnibus (HVO) is a open repository of genotype studies.

Ancestors: Gene Expression Onmibus.

Aims:

- Collection/Distribution: Collect and distribute data related to publications.
- Transparence: Facilitate reproducibility.
- Reusability: Re-use data for search, comparison and candidate SNP/genes identification.
- Integration: Integration of multiple data sources to obtain a overall perspective on the problem.

# The Architecture

Figure removed due to copyright reasons.

# Collection and Storage

Submission: Data are submitted as a single study file.

Challenge: Make submission easy but get as much information as possible.

Portability: Across subject areas.

- Phenotype: MeSH.
- Genotype: dbSNP and Celera.
- Exposures: Standardized (gender, race, etc).

Enforcement: Today, microarray experiment data are published (submitted) at paper submission time through editorial policies (Nature, Science, PNAS).

Figure removed due to copyright reasons.

Figure removed due to copyright reasons.

Human Variations

# Retrieval and Exploration

Retrieval: The general aim is distribution.

- By Study: download as files.
- By Phenotype: Useful for single variant validation.
- By Genotype: Useful for candidate genes analysis.

Exploration: Novel analytical tools.

- Single Variation Associations: Across phenotypes with different statistical methods.
- Genomic Properties : Linkage disequilibrium, haplotype analysis, haplotype tagging.
- Virtual Operations: Candidate genes, sample size simulations, etc.

Figure removed due to copyright reasons.

Figure removed due to copyright reasons.