Biostatistics 666
Statistical Models in Human Genetics

Instructor

Gonçalo Abecasis

Course Logistics

Grading

Office Hours

Class Notes

Course Objective

- Provide an understanding of statistical models used in gene mapping studies

- Survey commonly used algorithms and procedures in genetic analysis

Assessment

- 10-12 Weekly Assignments
  - About 60% of the final mark

- 2 Half Term Assessments
  - About 40% of the final mark

Office Hours

- Please cross out times for which you are unavailable in the sheet going around

- Room M4132
  School of Public Health II

Class Website

- PDF versions of notes and problem sets

  www.sph.umich.edu/csg/abecasis/class/

- Please let me know about any mistakes!

Course Contents

Brief Overview

Genetic Mapping

"Compares the inheritance pattern
of a trait with the inheritance pattern
of chromosomal regions"

Positional Cloning

"Allows one to find where a gene is,
without knowing what it is."

Some of the Topics Covered

- Maximum Likelihood

- Modeling Genes in Populations

- Modeling Genes in Pedigrees

Modeling
Genes in Populations

- Hardy Weinberg Equilibrium

- Linkage Disequilibrium

- The Coalescent

- Methods for Haplotyping

Modeling
Genes within Pedigrees

- Elston-Stewart algorithm

- Lander-Green algorithm

- Genetic linkage tests

- Checking Genetic Data for Errors

- Family Based Association Tests

Let's Get Started!

The Basics

Today – Primer In Genetics

- How information is stored in DNA

- How DNA is inherited

- Types of DNA variation

- Common designs for Genetic studies

DNA – Information Store

- Encodes the information required for cells and organisms to function and produce new cells and organisms.

- DNA variation is responsible for many individual differences, some of which are medically important.

Human Genome

- Multiple chromosomes
  - 22 autosomes
    - Present in 2 copies per individual
    - One maternally and one paternally inherited copy
  - 1 pair of sex chromosomes
    - Females have two X chromosomes
    - Males have one X chromosome and one Y chromosome

- Total of ~3 x 10^9 bases (each A, C, T or G)

Inheritance of DNA

- Through recombination, a new "DNA string" is formed by combining two parental DNA strings

- Thus, each chromosome we carry is a mosaic of the two chromosomes carried by our parents

- Only a small number of changeovers between the two parental chromosomes
  - On average ~1 per Morgan (~10^8 bases)

- Copying of DNA sequences is imperfect and, for typical sequences, the error rate is about 1 per 10^8 bases copied

Human Variation

- Every chromosome is unique …

- … but when two chromosomes are compared most of their sequence is identical

- About 1 per 1,000 bases differs between pairs of chromosomes in the population
  - In the same individual
  - In the same geographic location
  - Across the world

DNA Sequences That Vary...

- Genes (protein coding sequences, which total <2% of all DNA)
    - ~30,000-35,000 in humans

- Pseudogenes
    - Ancient genes, inactivated through mutation

- Promoters and Enhancers
    - Sequences which control gene expression

- Repeat DNA
    - Useful for tracking DNA through families or populations

- Packaging sequences, "spacer" DNA, etc.

Important Vocabulary ...

- Locus
- Polymorphism
- Allele
- Mutation
- Linkage
- Genetic Marker
- Genotype

- Phenotype
  - Mendelian Traits
  - Complex Traits
- Chromosomal landmarks
  - Centromeres
  - Telomeres
- Gene
- RNA
- Protein

Data for a Genetic Study

- ## Pedigree
  - Set of individuals of known relationship

- ## Observed marker genotypes
  - SNPs, VNTRs, microsatellites

- ## Phenotype data for individuals

Genetic Markers

- Genetic variants that can be measured conveniently

- Typically, we characterize them by:
  - Number of Alleles
  - Frequency of Each Allele

  - These are summarized by the heterozygosity

- The most commonly used genetic markers are microsatellites and SNPs

Phenotypes

- Measured characters of individuals

- Mendelian Phenotypes
    - Completely determined by genes
    - e.g. Cystic Fibrosis, Retinoblastoma

- Complex Phenotypes
    - Controlled by multiple genes and environmental factors
    - eg. Diabetes, Inflammatory Bowel Disease

Ultimate Aim of Gene-Mapping Experiments

- Localize and identify variants that control interesting traits
  - Susceptibility to human disease
  - Phenotypic variation in the population

- The difficulty…
  - Testing several million variants is impractical…

3 Common Questions

- Are there genes influencing this trait?
  - Epidemiological studies

- Where are those genes?
  - Linkage analysis

- What are those genes?
  - Association analysis
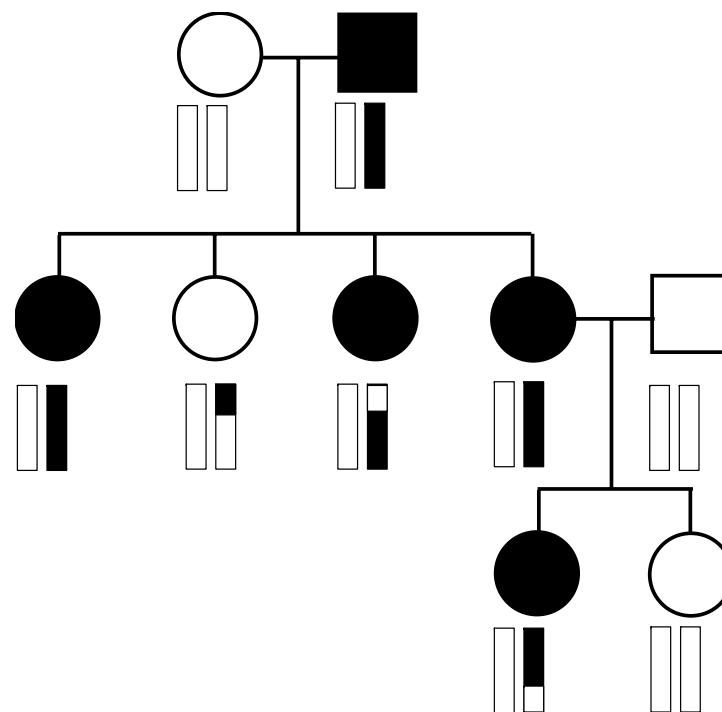
Is a trait genetic?

- Examine distribution of trait in the population and among relatives

- E.g. Inflammatory Bowel Disease (Crohn's)
  - General population
    - 1-3 cases per 1,000 individuals
  - Twins of affected individuals
    - 44% of monozygotic twins also have Crohn's
    - 3.8% of dizygotic twins also have Crohn's

Where are those genes?

- Find genetic markers that co-segregate with disease

- E.g. D16S3136 co-segregates with Crohn's

What are those genes?

- Identify genetic variants that are associated with disease…

- E.g. Mutations which disrupt NOD2 are much more common in Crohn's patients

	Crohn's	Controls
Arg702Trp:	11%	4%
Gly908Arg:	4%	2%
Leu1007fs	8%	4%

Common Designs for Genetic Studies
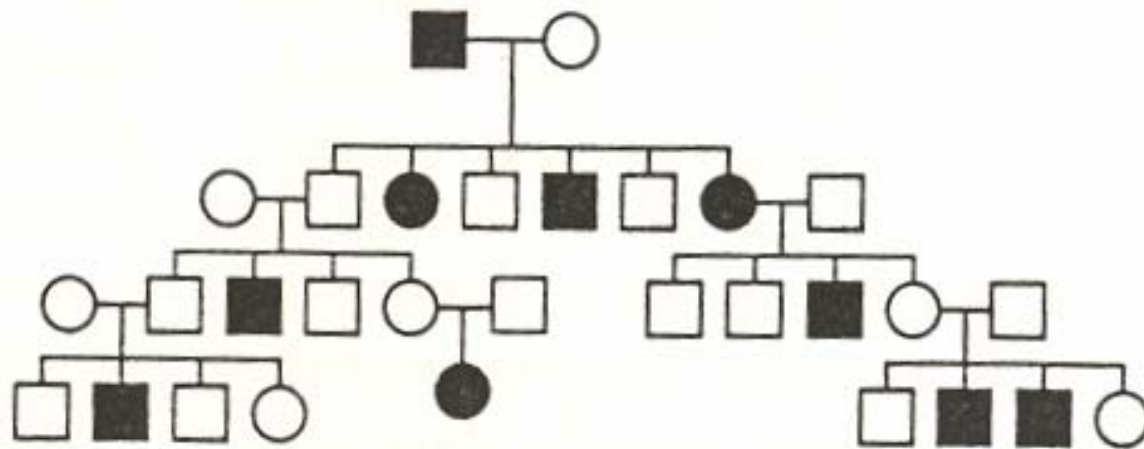
- Parametric Linkage analysis

- Allele-sharing methods

- Association analysis
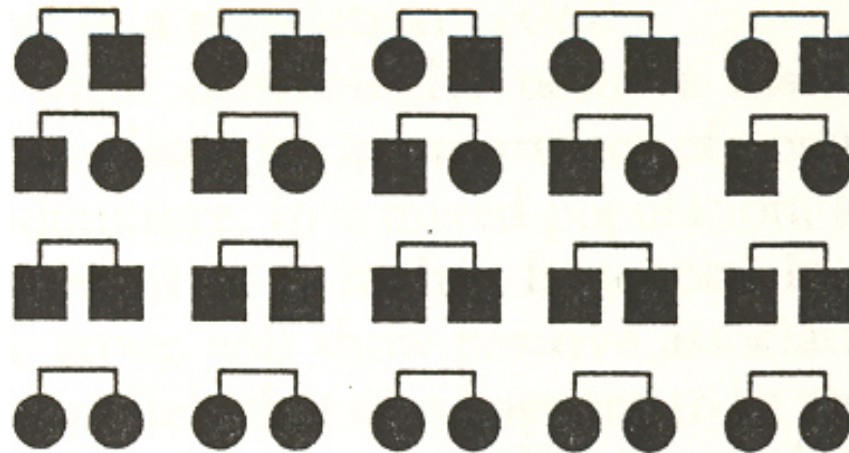
Parametric Linkage Analysis

- Evaluate a specific model and location
    - Allele frequencies at disease loci
    - Probability of disease for each genotype
- Potentially very powerful
- Vulnerable to heterogeneity, model misspecification
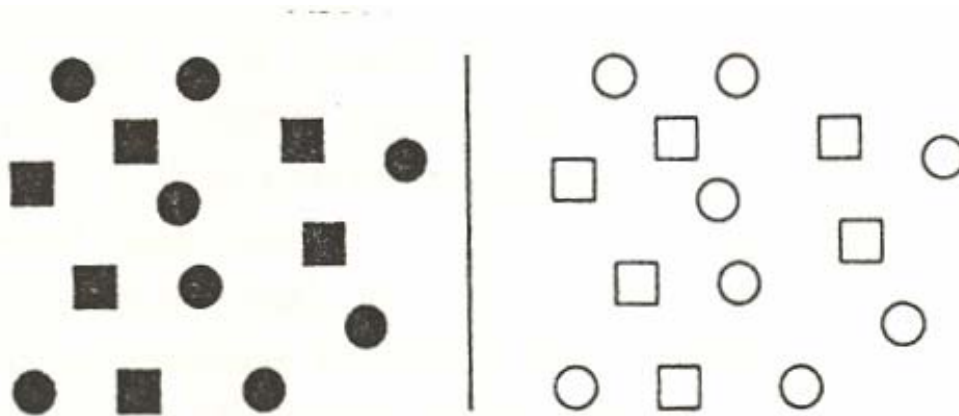
Allele Sharing Analysis

- Reject null hypothesis that sharing is random at a particular region

- Less powerful, but more robust

- Quantitative trait extensions exist
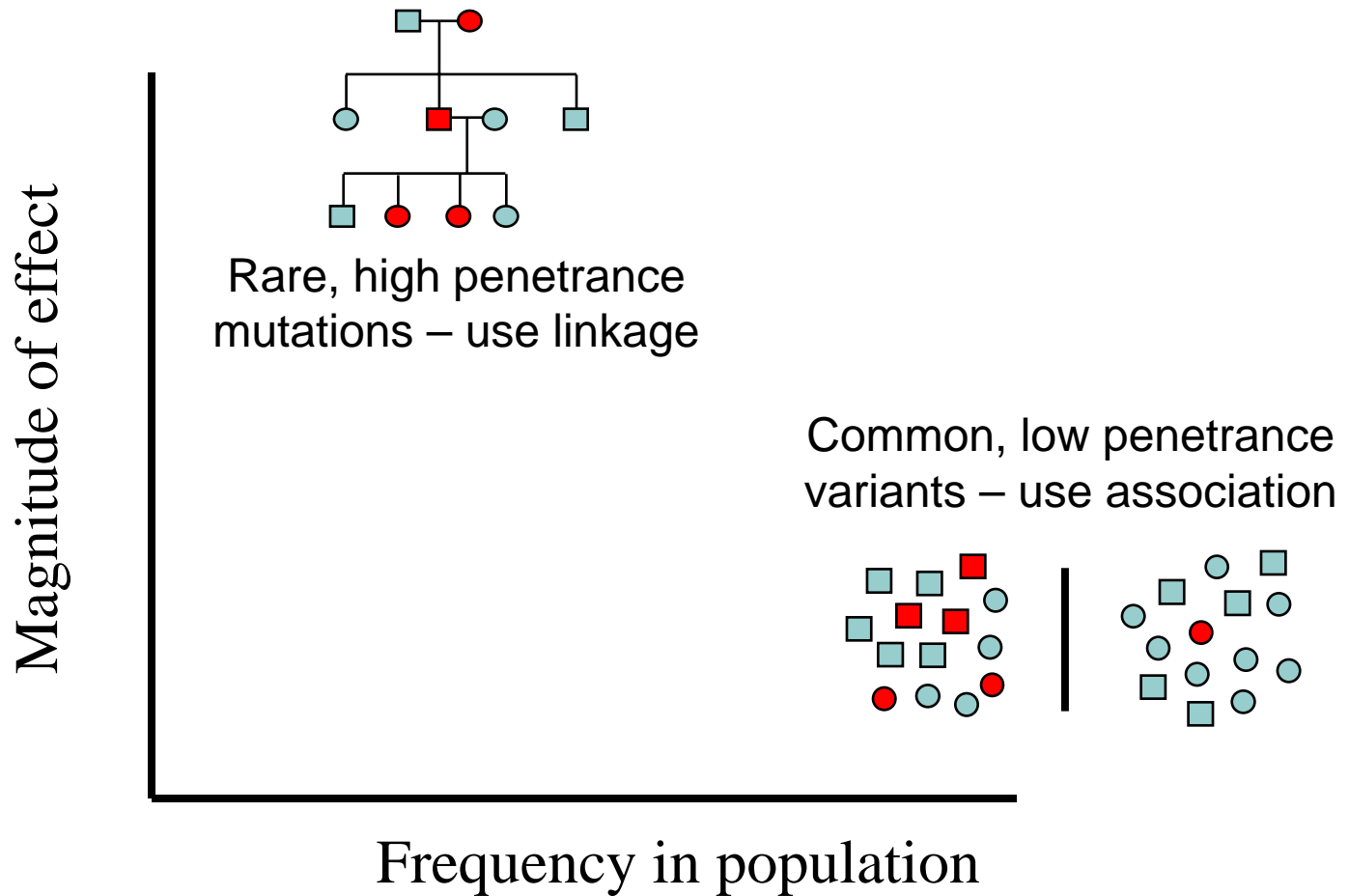
Association Analysis

- Simplest case compares frequency of allele among cases and controls
- Genome-wide search requires hundreds of thousands of markers
- Typically, focuses on candidate genes

Which Design to Choose?
The Right Choice Depends on the Alleles We Seek...

Rare, high penetrance mutations – use linkage

Common, low penetrance variants – use association

Magnitude of effect

Frequency in population

Genetic Linkage Studies

- Identify variants with relatively large contributions to disease risk

- Require only a coarse measurement of genetic variation
  - 400 – 800 microsatellites can extract most of the linkage information in typical pedigrees
  - Until recently, the only option for conducting whole genome studies

- High-throughput SNP genotyping has already sped up and facilitated these studies
  - Data analysis methods must select subset of independent SNPs or model disequilibrium between markers

Genetic Association Studies

- Identify genetic variants with relatively small individual contributions to disease risk

- Require detailed measurement of genetic variation
  - > 10,000,000 catalogued genetic variants, so …
  - Until recently, limited to candidate genes or regions
    - A hit-and-miss approach…

- SNP resources and decreasing assay costs now make it possible to examine 100,000s of markers

Recommended Reading

- An introduction to important issues in genetics:

  - Lander and Schork (1994)
    Science **265:**2037-48

- A good reference on molecular genetics:

  - Human Molecular Genetics
    Tom Strachan and Andrew Read

Reading for Next Lecture

- Will be discussing Hardy-Weinberg equilibrium
  - A basic feature of genotypes in human populations

- Wigginton, Cutler, Abecasis (2005)
  A note on exact tests of Hardy-Weinberg equilibrium.
  Am J Hum Genet **76**:887-93

- This paper describes an efficient method for testing Hardy-Weinberg equilibrium and includes many important historical references