STAT 536: Statistics for Population Genetics

Overview of Methods in Statistical Genetics

Karin S. Dorman

Department of Statistics
Iowa State University

December 9, 2008

Kingsmore2008

- **Risk Analysis.** essentially a test of heritability, determines whether risk of disease is increased in related individuals
- **Segregation Analysis.** identify the nature of inheritance (dominant, recessive, partial dominant, penetrance, etc.) of a qualitative trait
- **Linkage-Based Studies.** excellent for identifying causal genes for Mendelian diseases (1300 such genes identified in humans), not so successful for identifying causal genes of complex phenotypes
- **Gene Association Studies.**
  - **case-control**
  - **family-based association studies**
  - **linkage disequilibrium mapping**
  - **genome-wide association studies**

Balmain2003 - Cancer

- **Rare, High-Penetrance Alleles.** multiple cases of disease in families, identified using genetic linkage and positional cloning
  - RB1: retinobastoma gene
  - p53: germline inherited predisposing gene
  - BRCA1, BRCA2: involved in DNA repair
- **Familial risk.** twins, repeat occurrence in multiple tissues of same host, familial studies show that genes play a large role
- **Polygenic model.**
  - Common variant-common disease model: common variants that arose just once predispose disease and can be identified through association studies, either by targetted search for candidates genes or genome-wide scans (the current motivation for the SNP database); the problem is lack of power to detect small effects
  - rare variant-recent origin model posits that disease is caused mainly by many, recently arisen new mutations, with some alleles arising several independent times; association studies will fail because there are multiple backgrounds on which these alleles exist

Pharoah2002

Question whether using genetic polymorphism information can predict risk of disease better than traditional risk factors.

Study relatives of affected inviduals in Anglian Breast Cancer Study.

Two models found to fit the data well:

- polygenic model
- single common recessive allele (frequency 0.24), but since no such genes have been found, not likely and did not fit multi-case families well

Botstein2003

- 1980: first proposal of genome-wide linkage analysis
- There has been an explosion in the number of Mendelian genes (1200) identified by positional cloning
  - hemochromatosis, nail patella syndrome, lactose interolance, chronic granulomatous disease, X-linked muscular distrophy, cystic fibrosis, Fanconi anemia, ataxia telangiectasia, neurofibromatosis I, retinoblastoma, breast cancer, polyposis colorectal cancer, Huntington disease,
- Linkage mapping: positional cloning starts here
  - families where disease segregates are typed for DNA polymorphisms (SNPs increasingly useful today)
  - confused by diagnosis ambiguities (misdiagnosis, heterogeneity, complex inheritance, phenocopies)
  - BRCA1 was identified by focusing on early onset cases to remove phenocopies, but this used improved linkage signal to focus the analysis, and such practices can lead to dubious outcome; it is better to id potential correlates and confirm on a *separate* dataset

Botstein2003 (cont.)

- homozygosity mapping: suitable for rare recessive diseases where affected individuals with known levels of inbreeding are examined for IBD regions (guaranteed same mutation, same background)
- linkage disequilibrium: getting closer by IBDness from common ancestor
  - look for association between allele in case/control
  - easy for recessives because no ambiguity in which allele is linked
  - familial information can be used to infer phase in other cases
  - deviations from HWE in cases can indicate linked alleles when inheritance is linked to diseas gene that is not recessive or multiplicative
  - ancestral haplotype reconstruction
- Association or LD mapping may work better for weak phenotypes
- map-based LD

Frequency Estimation

- **Data.** Typed markers in random samples of individuals or pedigrees.
- **Objective.** Estimate allele frequencies.
- **Discussion.** Allele frequency estimates are often needed for downstream linkage or association analysis. You know how to estimate allele frequencies from random samples of individuals, but more sophisticated methods are needed for random samples of pedigrees.

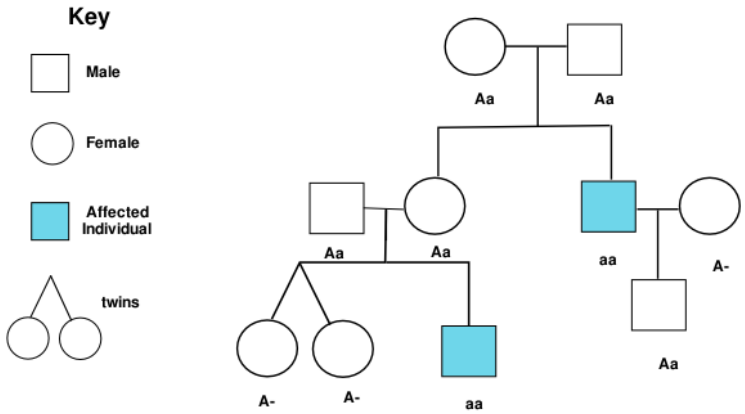
Equilibrium Testing

- **Data.** Marker data on randomly sampled individuals (not pedigrees).
- **Objective.** Test whether HWE or LE apply to particular markers in the population in question.
- **Discussion.** Disequilibrium can signify typing errors, linkage, selection, population structure, etc.

Segregation Analysis

- **Data.** Marker data and trait information.
- **Objective.** Estimate transmission parameters and penetrance parameters.
- **Discussion.** Useful as a prelude to linkage analysis in order to identify likely genetic models (e.g. recessive, dominant, etc.). Essentially, the goal is to predict trait based on genes (and possible other covariates, such as age, sex, etc.).

Mapping Markers

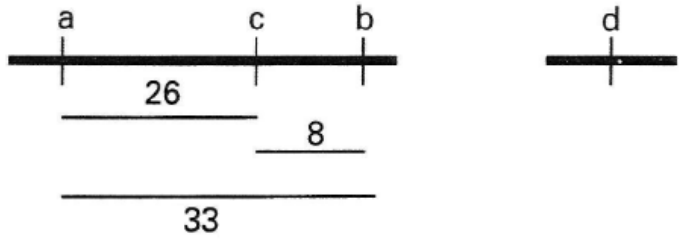


- **Data.** Typed markers in pedigrees.
- **Objective.** Order and estimate distances between markers based on observed or inferred recombination events.
- **Discussion.** Mapped markers are needed for mapping genes. With complete genome sequences in hand, it would seem that maps are readily available. Indeed, the order of loci may be known, but the recombination distance (related to probability of recombination in a generation) between loci may not be known. In addition, fully compiled genomic sequences are not available for all organisms. Use of poorly inferred or resolved maps can lead to confusing results. For example, adding more markers to a region of interest can lead to less resolution if the original map is wrong.

Mapping Markers

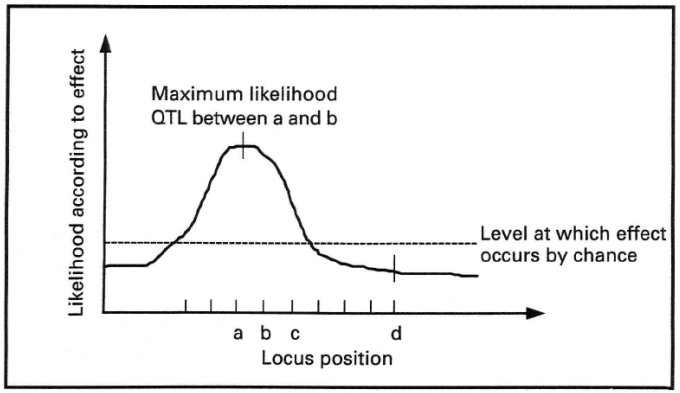
Recombination data

a-b = 33% (33 map units)
a-c = 26
c-b = 8
a-d = 50



- **Data.** Typed markers in pedigrees.
- **Objective.** Order and estimate distances between markers based on observed or inferred recombination events.
- **Discussion.** Mapped markers are needed for mapping genes. With complete genome sequences in hand, it would seem that maps are readily available. Indeed, the order of loci may be known, but the recombination distance (related to probability of recombination in a generation) between loci may not be known. In addition, fully compiled genomic sequences are not available for all organisms. Use of poorly inferred or resolved maps can lead to confusing results. For example, adding more markers to a region of interest can lead to less resolution if the original map is wrong.

Gene Mapping



- **Data.** Typed markers *plus* trait data in pedigrees.
- **Objective.** Estimate location of trait locus on an existing marker map based on observed or inferred recombination events.
- **Discussion.** Assuming there is a single trait locus with a particular, known pattern of inheritance, find its location by sliding it along the known marker map and identifying the position that maximizes the joint likelihood of observed marker and trait data. A complication is that a single locus may not explain the trait. One can assume a mixture model, where a fraction α of pedigrees have a linked trait locus, and the rest $1 - \alpha$ have an unlinked trait determinant, but this is a modest generalization. In general, this method is only good for mapping Mendelian loci.

Haplotyping

- **Data.** Marker data on pedigrees
- **Objective.** Estimate haplotypes and recombination events in order to identify markers shared by affected members of a pedigree.
- **Discussion.** You performed a simple example of this procedure using the EM algorithm to infer the haplotypes (*AB* or *Ab*) of double heterozygote *AaBb* individuals.

Allele Sharing

- **Data.** Marker data *plus* affected status on pedigrees.
- **Objective.** Identify excess sharing of marker alleles among affecteds in pedigrees in order to identify regions possibly associated with a 0/1 trait.
- **Discussion.** Similar to *gene mapping* except that it does not assume a particular genetic model. In particular, it can handle more complex traits, where multiple loci with arbitrary modes of inheritance control a trait. Also related to *association mapping* except that it operates on pedigrees.

Mistyping

- **Data.** Marker data on pedigrees.
- **Objective.** Identify marker assignments that are likely errors.
- **Discussion.** Markers are typically Mendelian traits, so evidence that they are not satisfying Mendelian inheritance (e.g. AC offspring of a $AA \times AA$ mating pair) could indicate data problems (mistyping or misinformation about pedigrees). Furthermore, given information about recombination probabilities, some haplotypes are very unlikely (e.g. $ABC/abc \times abc/abc$ with offspring AbC/abc requires two recombination events and my be unlikely if r_{AB} and r_{BC} are both small). Very small error rates (e.g. 1%) can lead to invalid gene mapping or association results, so this type of data verification is recommended before pedigree analysis.

Pedigree Selection

- **Data.** Marker data on related individuals.
- **Objective.** Identify the relationship (pedigree) between the individuals.
- **Discussion.** Applications include paternity testing, determination of twin zygosity, and body identification.

Kinship Calculations

- **Data.** Pedigree information.
- **Objective.** Identify the probability of sharing IBD alleles given known relationship.
- **Discussion.** Implications for studies on trait heritability.

Case/Control Analysis

- **Data.** Marker data on random sample of individuals who are also classified as case/control using some phenotype information.
- **Objective.** Identify markers that are *associated* with the trait.
- **Discussion.** The method works if the causative gene was introduced into the population on very few haplotypes (1 or few introductions) and there has been insufficient time for linkage disequilibrium to disappear.

QTL Association

- **Data.** Quantitative trait(s) measured on randomly sampled individuals or pedigrees.
- **Objective.** Identify markers that are associated with a quantitative trait.
- **Discussion.**

Transmission/Disequilibrium Test (TDT)

- **Data.** Marker data and affected status in parent/offspring combinations.
- **Objective.** Identify which markers tend to get transmitted with the affect status, thus identifying loci associated with the disease.
- **Discussion.** Proposed as an alternative method to association mapping that does not suffer from the problems of population stratification. The downside is that it relies on observing recombination events that occur in a single generation. It cannot localize putative disease genes with as much precision as association studies based on population data.

QTL Mapping

- **Data.** Marker data and quantitative trait measured on pedigrees.
- **Objective.** Map QTL (quantitative trait loci) that are important for a quantitative trait
- **Discussion.** Suppose you know that a major Mendelian gene underlies a quantitative trait, but you don't know where it is located. Then, this option places that locus at various positions along a known map and finds the position that maximizes the likelihood of the observed trait data.

Polygenic Traits

- **Data.** Quantitative trait measured on known pedigrees.
- **Objective.** Estimate heritability or variance components (additive, dominance, environmental, possible interactions)
- **Discussion.** Generalizes your study on estimating heritability to general pedigrees. Every pair of individuals share some fraction of alleles IBD, and the extent to which they share genes also determines the extent to which they share similar traits.