Part II. Genetics:
  2. ANOVA, Multi-marker model

Lecture 8 – Jan 29, 2015
CSE 427 Computational Biology
Instructor: Su-In Lee
TAs: Scott Lundberg, Safiye Celik
TTh 12:00-1:20 @ MGH 238

Outline
- Statistical methods for mapping QTL
  - Analysis of variance
  - Statistical significance of the LOD score
  - Multi-marker model
  - L1/L2-regularized linear regression
  - Cross validation

The simplest method: ANOVA
- “Analysis of variance”: assumes the presence of single QTL
- For each marker: Split mice into groups according to their genotypes at each marker.
- Compute the F-statistic (perform the t-test)
- Repeat for each typed marker

The simplest method: ANOVA
- t-test/F-statistic will tell us if there is sufficient evidence to believe that measurements from one condition (i.e. genotype) is significantly different from another.
- LOD score ("Logarithm of the odds favoring linkage")
  \[ \text{LOD score} = \log_{10} \text{likelihood ratio, comparing single-QTL model to the "no QTL anywhere" model.} \]
LOD score

- “Logarithm of odds favoring linkage”
- For each marker $k$,

$$\text{LOD}_k = \log_{10} \left( \frac{P(D | \text{marker}_k \text{ is the QTL})}{P(D | \text{no QTL})} \right)$$

Let’s say we are given trait values obtained from $n$ mouse individuals – $y_1, \ldots, y_n$ 

We divide the $n$ mouse individuals based on genotype values (AA/AB) on marker $k$

Let’s say that marker $k$ is the QTL
Let's say that there is no QTL and the trait values are normally distributed.

\[ \text{LOD score} = \log_{10} \left( \frac{p(D | \text{marker}_k \text{ is the QTL})}{p(D | \text{no QTL})} \right) \]

\[ = \log_{10} P(y_1, \ldots, y_n | \hat{\mu}_y, \hat{\sigma}_y) - \log_{10} P(y_1, \ldots, y_n | \hat{\mu}_y, \hat{\sigma}_y) - \sum_i \log_{10} P(y_i | \hat{\mu}_y, \hat{\sigma}_y) \]

Outline

- Basic concepts
  - Meiosis, genetic recombination
  - Allele, allele frequencies, genotype frequencies
  - Genotyping

- Statistical methods for mapping QTL
  - What is QTL?
  - Experimental animals
  - Analysis of variance
  - Statistical significance of the LOD score
Statistical significance

- Large LOD score → evidence for QTL
- Question: How large is large?
- Answer: Consider distribution of LOD score if there were no QTL.

Null distribution of the LOD scores at a particular marker $k$ (solid curve)

Only ~3% of chance that the genomic position gets LOD score ≥ 1.

How can we get the null distribution?

- We need a distribution of LOD score at a particular marker $k$.
- This means that we need many LOD scores to form a null distribution.
- How can we obtain many LOD scores at a particular marker $k$?
- We can do permutation tests.

Permutation tests: LOD score

- We divide the $n$ mouse individuals based on genotype values (AA/AB) on marker $k$.
- Permute the trait values.
- Repeat many times to obtain many LOD scores.

Null distribution of the LOD scores at a particular marker $k$ (solid curve)
Statistical significance

- **Question:** How large is large?
- **Answer:** Consider distribution of LOD score if there were no QTL.
- **Answer:** Consider distribution of maximum LOD score if there were no QTL anywhere in the genome-wide search.

![Null distribution of the LOD scores at a particular marker and LOD score from a genome scan.](image)

LOD thresholds

- LOD threshold = 95th percentile of the distribution of genome-wide max LOD, when there are no QTL anywhere.

More on LOD thresholds

- **Methods for obtaining thresholds**
  - Analytical calculations (assuming dense map of markers) [Lander & Botstein, 1989]
  - Computer simulations
  - Permutation/randomized test [Churchill & Doerge, 1994]

- **Appropriate threshold depends on:**
  - Size of genome
  - Number of typed markers
  - Pattern of missing data
  - Stringency of significance threshold
  - Type of cross (e.g. F2 intercross vs backcross)

ANOVA at marker loci

- **Advantages**
  - Simple.
  - Easily incorporate covariates (e.g. environmental factors, sex, etc).
  - Easily extended to more complex models.

- **Disadvantages**
  - Must exclude individuals with missing genotype data.
  - Imperfect information about QTL location.
  - Suffers in low density scans.
  - Only considers one QTL at a time (assumes the presence of a single QTL).
Outline

- Statistical methods for mapping QTL
  - Analysis of variance
  - Statistical significance of the LOD score
  - Multi-marker model
  - L1/L2-regularized linear regression
  - Cross validation

Modeling multiple QTLs

- Advantages
  - Reduce the residual variation and obtain greater power to detect additional QTLs.
  - Identification of (epistatic) interactions between QTLs requires the joint modeling of multiple QTLs.

Multiple marker model

- Let \( Y = \) phenotype, \( X = \) genotype data.

- Imagine a small number of QTL with genotypes \( X_1 \ldots X_p \)
  - \( 2^p \) or \( 3^p \) distinct genotypes for backcross and intercross, respectively

- We assume that \( Y | X \sim N \left[ \mu(X_1 \ldots X_p), \sigma^2 \right] \)

- Linearity \( \mu(X_1 \ldots X_p) = \mu_0 + \sum_j \nu_j X_j \)

Why Linear Regression?

- Suppose we want to model the outcome variable \( Y \) in terms of three predictors, \( X_1, X_2, X_3 \)
  \( Y = f(X_1, X_2, X_3) \)

- Typically will not have enough data to try and directly estimate \( f \)

- Therefore, we usually have to assume that it has some restricted form, such as linear
  \( Y = X_1 + X_2 + X_3 \)