Analysis of Quantitative Trait Loci (QTL)

Janet Sinsheimer
Human Genetics, Biomathematics, Biostatistics
janet@mednet.ucla.edu
5357C Gonda
5-8002
References:

Chapters 4 and 15, Stratchen and Reed (2004). Human Molecular Genetics 3.


Chapter 8, Mathematical and Statistical Methods for Genetic Analysis by K. Lange (2002) Springer

What is a Quantitative Trait?

A quantitative trait has numerical values that can be ordered highest to lowest. Examples: height, weight, cholesterol level, reading scores.

There are discrete (count) values where the values differ by a fixed amount and continuous values where the difference in two values can be arbitrarily small. Most methods for quantitative traits assume that the data are continuous rather than counts.
Statistical Measures in Quantitative Genetics

The mean is a way to estimate the center of the distribution:

$$\bar{X} = \frac{\sum X_i}{N}$$

The variance is a way to estimate the spread of the distribution:

$$V_X = \frac{\sum (X_i - \bar{X})^2}{N - 1}$$

The standard deviation is another estimate of the spread:

$$SD = \sqrt{V_X}$$

The mean is a way to estimate the center of the distribution.

The variance is a way to estimate the spread of the distribution.

The standard deviation is another estimate of the spread.
Correlation between X and Y

[Graphs showing correlation patterns]
Correlation is an Important Concept in Quantitative Genetics

\[ Cov(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1} \]

The covariance is a way to measure the tendency for two variables to have the same location in their distributions.

\[ r = \text{cor} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}} \]

The correlation is a standardized (measurement scale free) covariance.

Perfect + correlation: \( r = 1 \)

No correlation: \( r = 0 \)
Why use Quantitative Traits?
(1) **More power.** Fewer subjects may need to be examined (phenotyped) if one uses the quantitative trait rather than dichotomizing it to create qualitative trait.

Individuals $w$ and $x$ have similar trait values, yet $w$ is grouped with $z$ and $x$ is grouped with $y$. Note that even among affecteds, knowing the trait value is useful ($v$ and $z$ are more similar than $v$ and $w$).
(2) The genotype to phenotype relationship may be more direct.

Affection with a disease could be the culmination of many underlying events involving gene products, environmental factors and gene-environment interactions. The underlying events may differ among people, resulting in heterogeneity.
When are Quantitative Traits Problematic?
(1) The quantitative trait doesn’t meet the assumptions of the proposed statistical method. For example, many methods assume the quantitative traits are unimodal but not all quantitative traits are unimodal.

(2) The values of the quantitative trait might be very unreliable.

(3) There are no good intermediate quantitative phenotypes for a particular disease. The quantitative traits available aren’t telling the whole story.
Basic Premise: The Quantitative Trait Values vary with Genotype

A gene that influences a quantitative trait’s value is called a quantitative trait locus (QTL).
The Architecture of the Genotypic Variance

The genotypic variance of the trait can be due to a single major QTL, a few QTLs with large effects (oligogenic) or many QTLs with small effects (polygenic), or combinations of all three.

The individual genes may act on the variance in an additive manner or there may be interaction, or epistatic, relationships between the genes.
How can a few genes and their alleles which are discrete lead to something that looks essentially normally distributed?
Quantitative Traits as the Result of Some Number of Bi-allelic Genes

Suppose 3 genes control the trait value. Also suppose that in our population, each gene has two equally frequent alleles.

Label the alleles from gene 1 as A and a, the alleles from gene 2 as B and b, and the alleles from gene 3 as C and c.

Let A, B, C add one to the phenotype value and let a, b, c subtract one from the phenotype value.

So an individual with the A/a B/b C/c genotype has phenotype value = 0

What if the genotype is A/A B/b c/c? or A/A B/B C/C?

Now “genotype” a sample of the population.
Even with only 3 genes, the discrete phenotype is well approximated by a normal distribution. The normal density is bell-like, symmetric, has mean \( \bar{X} \) and has inflection points (points where the second derivative is zero) at \( \bar{X} - SD \) and \( \bar{X} + SD \). As the number of genes increases, the approximation improves. This scenario is the idea behind a polygenic model.
Population Based Studies of Quantitative Traits
Analysis Conducted with genotypes

Trait values and genotypes: Let $Y$ denote the trait value. If $Y$ depends on the alleles at locus $T$, then we expect that knowing the alleles at $T$ would help us predict the value of $Y$. 

![Genotype frequency graph](chart.png)
But we don’t Actually Know the Position of the Trait Locus

As the distance between the marker locus and the trait locus decreases the marker alleles becomes a good predictor of the trait value.
How can We Test for Association between a SNP and a Quantitative Trait?

Collect phenotype and genotype data from a number of unrelated individuals.

Just like a case-control study but there are no cases and no controls – just individuals with a range of phenotype values.
Hypothetical Data Example

- Vitamin D receptor gene and women’s height.
- Measure 128 women between the ages of 30-40.
- Genotype a SNP in the Vitamin D receptor.
### Results

<table>
<thead>
<tr>
<th>Genotype</th>
<th>number</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1</td>
<td>53</td>
<td>63.01</td>
<td>3.71</td>
<td>55.76</td>
<td>71.25</td>
</tr>
<tr>
<td>1/2</td>
<td>54</td>
<td>64.52</td>
<td>2.86</td>
<td>58.18</td>
<td>71.82</td>
</tr>
<tr>
<td>2/2</td>
<td>21</td>
<td>66.39</td>
<td>2.95</td>
<td>60.66</td>
<td>70.48</td>
</tr>
</tbody>
</table>

Do the means change with the number of 2 alleles present in the genotype?
Trend (Correlation) Test

• Test whether $r$ (the correlation) is significantly different from zero.

• For our example: the calculated value of $r$ is 0.3452. It is significantly different from zero, $p$-value = 0.0001

• This trend test is the test used by Cho et al. (2009) in their association analysis
More Generally: QTL Association using Linear Regression

- Example: height and vitamin D receptor genotypes A/A, A/B, B/B. Let S be the number of B alleles in person i’s genotype.
- Linear regression allows us to include covariates and expand our sample to include males and females effects and other covariates such as age.

\[ \text{height}_i = \beta_0 + \beta_{\text{age}} \text{age}_i + \beta_{\text{cal} C}_i + \beta_{\text{male}} \text{male}_i + \beta_{\text{snp}} S_i + e_i \]

where

\[ e_i \sim N(0, \sigma^2) \]

- How can we code for an interaction between age and SNP?
Graph of Additive Acting Alleles

![Graph of Additive Acting Alleles](image-url)
What if Allele Effects aren’t Additive?
Dominance Effects

![Graph showing the relationship between average trait values and the number of risk alleles, highlighting the concepts of Additive Effect and Dominance effect.](image)
Cho et al. 2009 GWAS

- 8 Quantitative traits
- ~300K markers
  - Guard against false positives by using an arbitrary stage 1 threshold (P value <10^{-5} with P value <0.01 in both Ansung and Ansan)
  - Replication of SNPs passing stage 1 in another sample

- Study Sample:
  ~4000 individuals from Ansung (rural)
  ~4000 individuals from Ansan (urban)
Korean Genome Epidemiology Study (KoGES)

**Ansung Cohort**

**Ansan Cohort**

Genotyping w/ Affymetrix 5.0

Data Filtering

**Stage 1: KARE GWAS (QT Analyses)**

Identification of SNPs passed the stage 1 threshold
(GWA $p<1 \times 10^{-5}$, MAF $>5\%$, cohort $p<0.01$)

**Stage 2: Replication study from Health2 cohort**

Combined analysis of GWAS and replication study
Cho et al. Take Care to Remove Data that Could Lead to Spurious Results

- Samples with low call rates (<96%), n = 401
- Samples with evidence of contamination, n = 11
- Samples with sex inconsistencies, n = 41
- Samples with cryptic relatedness, n = 608
- Samples with serious illnesses, n = 101
Population Stratification can Cause Spurious Results

• If the sample consists of two or more populations with different allele frequencies and different trait value distributions, one could see evidence of an association.

• Because they have samples from two regions of Korea, Cho et al. check for evidence of population stratification.
Simple Illustration of Population Stratification for a Qualitative trait (cases/controls).

- Disease risk is not dependent on SNP allele.
  - Within a population allele frequencies for cases and controls are the same.
- Odds of disease differs pop 1 versus pop 2. OR = 9
  - 1000 cases (800 pop 1, 200 pop 2), 1000 controls (200 pop1, 800 pop 2)
- Allele frequencies differ between cases and controls
  - Prob(allele 1 | pop 1) = .7, Prob(allele 1 | pop 2) = .3
  - Assume HWE holds in both populations.
- Allele 1 frequency in cases = .8*.7+.2*.3 = .62
- Allele 1 frequencies in controls = .2*.7+.8*.3 = .38
- Would lead to a highly significant association result.
Graphical Display – Combined Populations

Pr (case)

genotype

0  0.5  1  1.5  2

0  0.2  0.4  0.6  0.8  1

disease  Fitted values
Graphical Display for Each Population

Pr (case)

Pop 1

Pop 2
<table>
<thead>
<tr>
<th>Trait</th>
<th>RS ID</th>
<th>Class</th>
<th>Locus</th>
<th>Nearby gene(s)</th>
<th>Minor allele MAF</th>
<th>Effect size 80% power</th>
<th>GWAS (p &lt; 8.84E-13)</th>
<th>Replication (p &lt; 7.86E-13)</th>
<th>Combined (p &lt; 1.67E-13)</th>
<th>Meta-analysis</th>
<th>Published</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMN</td>
<td>rs17176527</td>
<td>Unknown</td>
<td>6q24.1</td>
<td>LOC729076</td>
<td>A 0.25 ± 0.031 0.055 -0.310</td>
<td>2.127</td>
<td>1.2E-06 3.9E-01 3.7E-04</td>
<td>-0.142 ± 0.040 2.63E-06</td>
<td>5.7E-02</td>
<td>Frigyl et al.</td>
<td></td>
</tr>
<tr>
<td>WHR</td>
<td>rs2074356</td>
<td>Intron</td>
<td>16q12.2</td>
<td>FTI</td>
<td>A 0.13 ± 0.035 0.07 0.200</td>
<td>2.390</td>
<td>1.7E-06 2.2E-02 1.5E-07</td>
<td>0.235 ± 0.045 5.7E-02</td>
<td>1.5E-02</td>
<td>Friedman et al.</td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>rs17098410</td>
<td>Unknown</td>
<td>13q21.33</td>
<td>TXL5</td>
<td>T 0.14 ± 0.006 0.001 0.005</td>
<td>1.852</td>
<td>6.1E-06 5.3E-01 4.4E-03</td>
<td>0.003 ± 0.001 12.59E-06</td>
<td>1.9E-04</td>
<td>Gudbjartsson et al.</td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>rs17038182</td>
<td>Unknown</td>
<td>1q21</td>
<td>ZBTB58</td>
<td>G 0.21 ± 0.542 0.996 1.010</td>
<td>1.613</td>
<td>3.2E-08 2.5E-02 3.3E-08</td>
<td>0.401 ± 0.072 4.25E-07</td>
<td>3.9E-02</td>
<td>Gudbjartsson et al.</td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>rs15631373</td>
<td>Intron</td>
<td>3q23</td>
<td>ZBTB58</td>
<td>A 0.72 ± 0.492 0.991 0.760</td>
<td>2.363</td>
<td>5.6E-06 1.9E-05 5.6E-12</td>
<td>0.461 ± 0.067 2.4E-06</td>
<td>3.2E-01</td>
<td>Weidner et al.</td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>rs12273123</td>
<td>Intron</td>
<td>8q22.1</td>
<td>FLJ61</td>
<td>G 0.07 ± 0.781 0.16 -0.165</td>
<td>2.358</td>
<td>1.1E-06 2.0E-04 1.0E-09</td>
<td>0.710 ± 0.116 4.0E-02</td>
<td>3.2E-01</td>
<td>Gudbjartsson et al.</td>
<td></td>
</tr>
<tr>
<td>rs6081030</td>
<td>Intron</td>
<td>9q22.32</td>
<td>FPP2</td>
<td>BCL3, TBBX</td>
<td>G 0.15 ± 0.529 0.113 -0.330</td>
<td>2.127</td>
<td>2.7E-06 5.1E-01 9.9E-05</td>
<td>-0.316 ± 0.081 7.8E-01</td>
<td>5.4E-03</td>
<td>Gudbjartsson et al.</td>
<td></td>
</tr>
<tr>
<td>rs12079795</td>
<td>Intron</td>
<td>17q23.2</td>
<td>BCL3, TBBX</td>
<td>A 0.33 ± 0.399 0.085 0.315</td>
<td>1.712</td>
<td>2.5E-06 - - -</td>
<td>- -</td>
<td>-</td>
<td>Weidner et al.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs3916756</td>
<td>Intron</td>
<td>2q21.3</td>
<td>TFEM1</td>
<td>C 0.29 ± 0.445 0.096 0.310</td>
<td>1.712</td>
<td>2.5E-06 - - -</td>
<td>- -</td>
<td>-</td>
<td>Weidner et al.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs14463438</td>
<td>Intron</td>
<td>2q22.3</td>
<td>LRBP1</td>
<td>A 0.35 ± 0.370 0.082 0.315</td>
<td>1.712</td>
<td>2.5E-06 - - -</td>
<td>- -</td>
<td>-</td>
<td>Weidner et al.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>rs12249754</td>
<td>Unknown</td>
<td>12q21.33</td>
<td>ATP2B1</td>
<td>A 0.37 ± 0.310 0.266 -1.260</td>
<td>2.127</td>
<td>1.3E-06 7.0E-05 1.6E-09</td>
<td>0.661 ± 0.109 6.0E-01</td>
<td>4.1E-01</td>
<td>Gudbjartsson et al.</td>
<td></td>
</tr>
<tr>
<td>DBP</td>
<td>rs17249545</td>
<td>Unknown</td>
<td>12q21.33</td>
<td>ATP2B1</td>
<td>A 0.37 ± 0.310 0.266 -1.260</td>
<td>2.127</td>
<td>1.3E-06 7.0E-05 1.6E-09</td>
<td>0.661 ± 0.109 6.0E-01</td>
<td>4.1E-01</td>
<td>Gudbjartsson et al.</td>
<td></td>
</tr>
<tr>
<td>Pulse rate</td>
<td>rs12731740</td>
<td>Unknown</td>
<td>1q32.2</td>
<td>CA94</td>
<td>T 0.10 ± 0.993 0.195 1.030</td>
<td>1.712</td>
<td>2.5E-06 - - -</td>
<td>1.0E+05 1.8E+03 2.7E+00</td>
<td>1.0E-01</td>
<td>Gudbjartsson et al.</td>
<td></td>
</tr>
<tr>
<td>rs12200093</td>
<td>Unknown</td>
<td>6q22.31</td>
<td>LOC564602</td>
<td>A 0.49 ± 0.573 0.118 0.620</td>
<td>2.127</td>
<td>1.3E-06 7.0E-05 1.6E-09</td>
<td>0.661 ± 0.109 6.0E-01</td>
<td>4.1E-01</td>
<td>Gudbjartsson et al.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs11556715</td>
<td>Intron</td>
<td>2q12.1</td>
<td>CTSS</td>
<td>A 0.24 ± 0.630 0.141 0.855</td>
<td>2.127</td>
<td>1.3E-06 7.0E-05 1.6E-09</td>
<td>0.661 ± 0.109 6.0E-01</td>
<td>4.1E-01</td>
<td>Gudbjartsson et al.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7776725</td>
<td>Intron</td>
<td>7q31.31</td>
<td>FAM3C</td>
<td>C 0.13 ± 0.222 0.033 0.212</td>
<td>1.609</td>
<td>1.0E-11 1.9E-10</td>
<td>- -</td>
<td>-</td>
<td>Gudbjartsson et al.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6056677</td>
<td>Intron</td>
<td>13q14.11</td>
<td>FAM3C</td>
<td>T 0.45 ± 0.105 0.022 -0.110</td>
<td>2.127</td>
<td>1.3E-06 7.0E-05 1.6E-09</td>
<td>0.661 ± 0.109 6.0E-01</td>
<td>4.1E-01</td>
<td>Gudbjartsson et al.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7776725</td>
<td>Intron</td>
<td>7q31.31</td>
<td>FAM3C</td>
<td>T 0.13 ± 0.155 0.032 0.212</td>
<td>1.609</td>
<td>1.0E-11 1.9E-10</td>
<td>- -</td>
<td>-</td>
<td>Gudbjartsson et al.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1721400</td>
<td>Intron</td>
<td>7q14.1</td>
<td>TXNDC3, SFRP4, EPDR1</td>
<td>T 0.17 ± 0.149 0.028 -0.136</td>
<td>2.987</td>
<td>1.4E-07 6.0E-04</td>
<td>- -</td>
<td>-</td>
<td>Gudbjartsson et al.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Minor allele MAFs were estimated from the positive strand based on NCBI build 36. For the current study, the significance threshold is based on 1% per polygenic trait. The p-values were estimated from meta-analysis between KARE and replication studies. The p-values were adjusted for multiple testing using the Bonferroni correction. The p-values were then compared to the threshold of 5.7E-05.
What if we also have Family Data?

- Why can’t the linear regression model be applied to family data as is?
- What could go wrong with this model when we have related individuals?
QTL Association - with a Variance Components Model
Appropriate Study Design

• Families, unrelated individuals or a mixture can be used, best results with extended pedigrees.
• Quantitative trait outcomes that can be transformed to normality.
• Random ascertainment of families or ascertainment through explicitly defined probands are best.
Mean Part of the Model is the Same as Before

\[ \hat{y}_i = \beta_0 + \beta_{\text{age}} \text{age}_i + \beta_{\text{cal}} C_i + \beta_{\text{male}} \text{male}_i + \beta_{\text{snp}} \text{SNP}_i \]

(measured genotype analysis)

But now our model lets relatives share more genes in common on average than unrelated individuals. That is, the covariance among two relatives is not assumed to be zero.
Familial Correlation handled through Polygenic Inheritance

• Polygenic inheritance assumes that the trait is determined by many genes with small additive effects

• The distribution of the trait values in a population follows a normal distribution, whose location and scale depend on the mean and variance of the trait.

• The bivariate distributions of relative pairs are 2-dimensional normal distributions – ellipsoidal level curves whose shape depends on the heritability of the trait.
Example: Sibling Correlation for Height

Correlation is not caused by a single gene, caused by many genes and common environment.
Variance Components

Commonly made simplifying assumptions:
No shared environment effects
All genes are acting independently of each other
No gene – environment interactions

\[ \sigma^2_P \text{, is the sum of the genetic variance, } \sigma^2_G, \text{ and the environmental variance, } \sigma^2_E. \]
\[ \sigma^2_P = \sigma^2_E + \sigma^2_G \]
The Degree of Correlation between Two Relatives Depends on just how Closely Related they are

• An important measure of family relationship is the theoretical kinship coefficient.
• It is the probability that two genes, at a randomly chosen locus, one chosen randomly from individual i and one from j are identical by descent.
• The kinship coefficient, $\Phi_{ij}$, does not depend on the observed genotype data.
• The theoretical kinship coefficient is not enough to uniquely defend relationships. $\Phi_{ij} = 1/4$ for siblings and for parent-child.
• Let $\Delta_{ij} =$ the probability that i and j share two genes IBD.
Covariance among Relatives

- \( \text{cov}(Y_i, Y_j) = 2 \Phi_{ij} \sigma^2_A + \Delta_{ij} \sigma^2_D \)
- The more closely related \( i \) and \( j \), the greater the similarity between their trait values is expected to be.
- In our example, we will assume the dominance variance equals zero, then the covariance among relatives \( i \) and \( j \) is: \( \text{cov}(Y_i, Y_j) = 2 \Phi_{ij} \sigma^2_A \)
Correlation between X and Y

Could relatives trait values be this highly correlated?
Population Stratification can Cause Spurious Association in this Model

- Need to allow for different allelic effects in different populations.
Data Analysis

• 50 two generation nuclear families 236 individuals

• Use a variance component model: Trait levels as outcome, and sex, bmi, and age as covariates.

• Account for variation among relatives with an additive polygenic variance term.

• Compare results with and without a SNP as a covariate.
<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter values</th>
<th>$\beta_0$</th>
<th>$\beta_{female}$</th>
<th>$\beta_{bmi}$</th>
<th>$\beta_{age}$</th>
<th>$\sigma_A^2$</th>
<th>$\sigma_E^2$</th>
<th>$\beta_{snp}$</th>
<th>Log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>-SNP $H_0$</td>
<td></td>
<td>3.799</td>
<td>.1392</td>
<td>.0092</td>
<td>.003</td>
<td>.0245</td>
<td>.0286</td>
<td>.000</td>
<td>199.63</td>
</tr>
<tr>
<td>+SNP $H_A$</td>
<td></td>
<td>3.868</td>
<td>.1252</td>
<td>.0083</td>
<td>.002</td>
<td>.0272</td>
<td>.0255</td>
<td>-.037</td>
<td>203.72</td>
</tr>
</tbody>
</table>

\[ H_o : \beta_{snp} = 0 \]

\[ H_A : \beta_{snp} \neq 0 \]

\[ \hat{y}_i = \beta_0 + \beta_{female} Female_i + \beta_{bmi} bmi_i + \beta_{age} age_i + \beta_{snp} snp_i \]

\[ \text{var}(y_i) = \sigma_A^2 + \sigma_G^2 \]

\[ \text{cov}(y_i, y_j) = 2\Phi_{ij}\sigma_A^2 \]
Compare Models using Likelihood Ratio Tests

• Models must be nested to use the LRT. For example the no SNP model is a special case of SNP model
• The null hypothesis is the special case model
• Alternative hypothesis is the general case model
• \( LRT = 2 \times (LL_{\text{gen}} - LL_{\text{spec}}) \sim \chi^2 \) distribution, \( df = m_{\text{gen}} - m_{\text{spec}} \)
• P-value = probability of getting an LRT as big or bigger when the null hypothesis is true.
• If the p-value is less than 0.05 (for example) then reject the null hypothesis (in favor of the alternative hypothesis)
• If the p-value is greater than 0.05 then fail to reject the null hypothesis
• For our example: \( LRT = 2 \times (203.72 - 199.63) = 8.18 \)
• P-value = 0.0042
Family Based Association Methods for Quantitative Traits

- Methods that test whether the apparent transmission of an allele from parent to offspring depends on the offspring’s phenotype value
  - Examples: FBAT; Mendel’s Gamete Competition
- Methods that treat the genotype at the marker as a covariate in a regression analysis, “measured genotype analysis”.
  - Examples: Mendel’s QTL Association and Association given Linkage; QTDT (Abecassis)