Matric eQTL analysis

Bioinformatics UiB

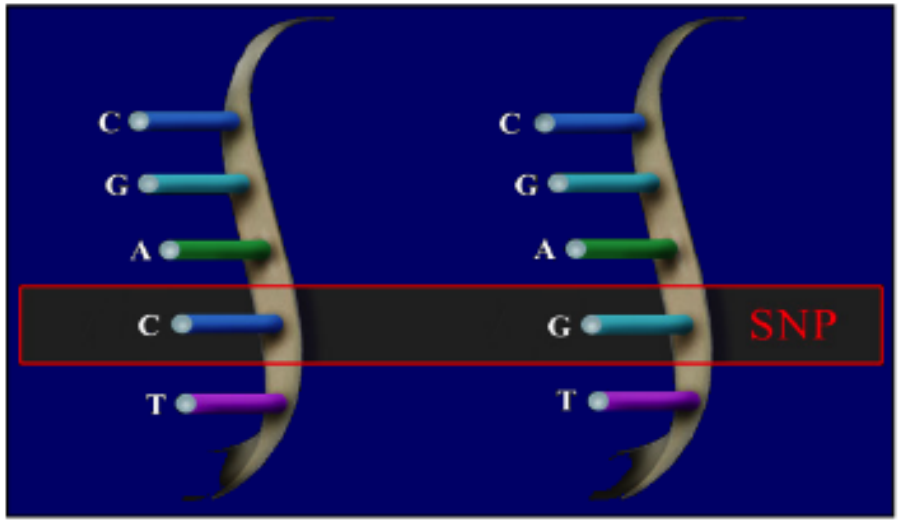
Mariyam Khan

UNIVERSITY OF BERGEN

Table of contents

Eqtl basics

- eQTLs expression quantitative trait loci.



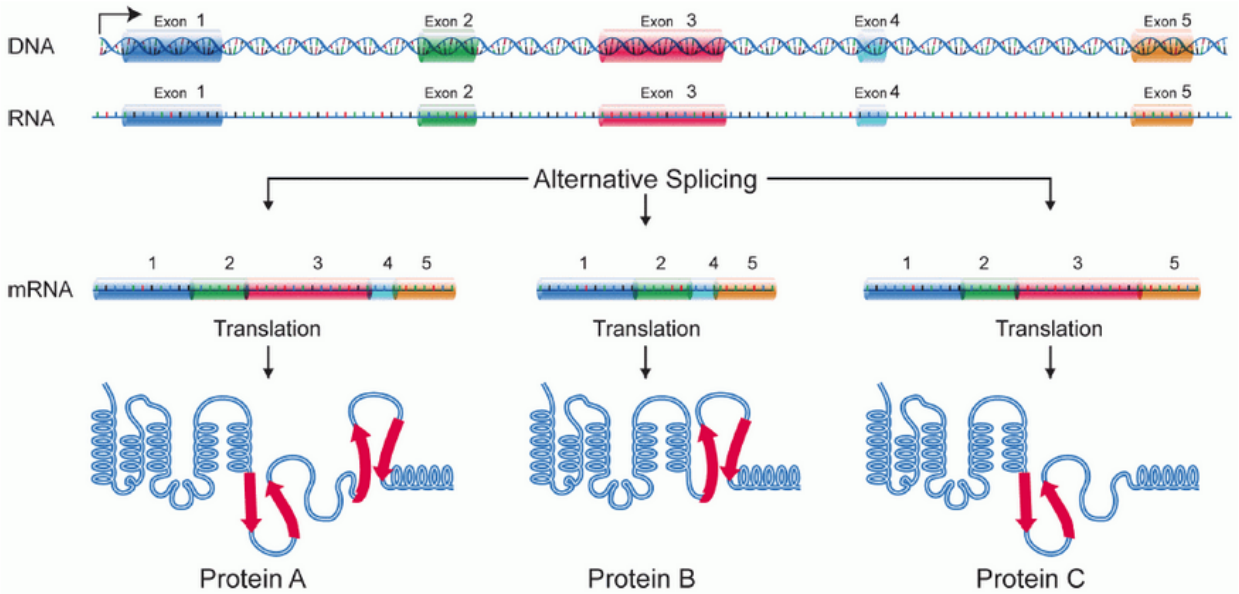
- Certain SNPs can either enhance or disrupt the expression of a certain gene.

Identify, Analyze, and Interpret eQTLs in the genome!

Eqtl mechanisms

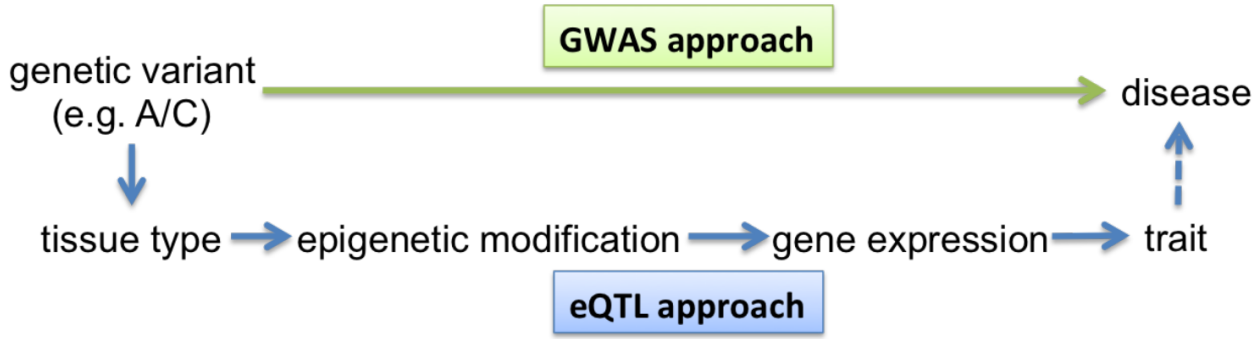
How eQTLs influence the expression of their associated genes?

- Altered transcription factor binding
- Histone modifications
- Alternative splicing of mRNA
- miRNA silencing



What is eqtl analysis?

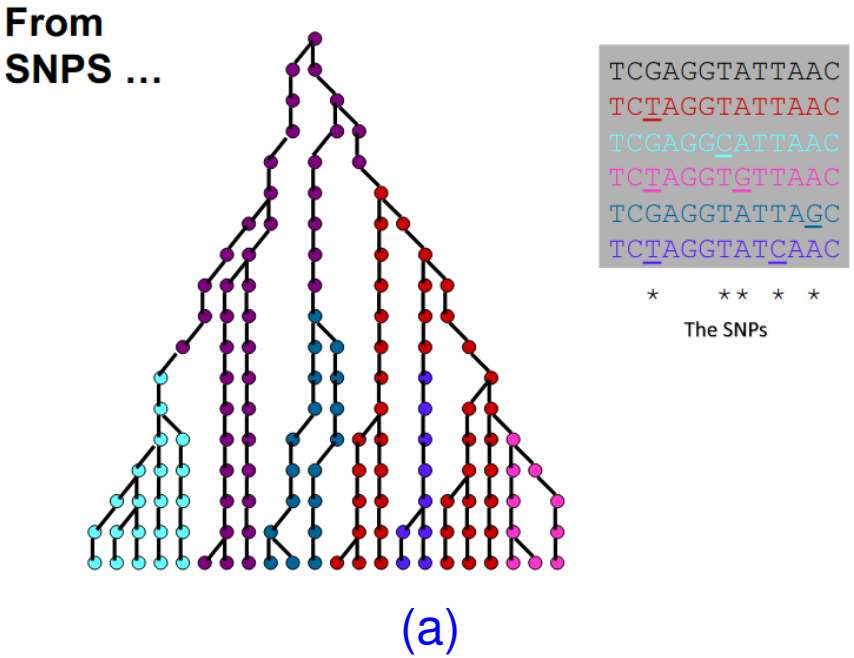
- Eqtl analysis links variations in gene expression levels to genotypes.

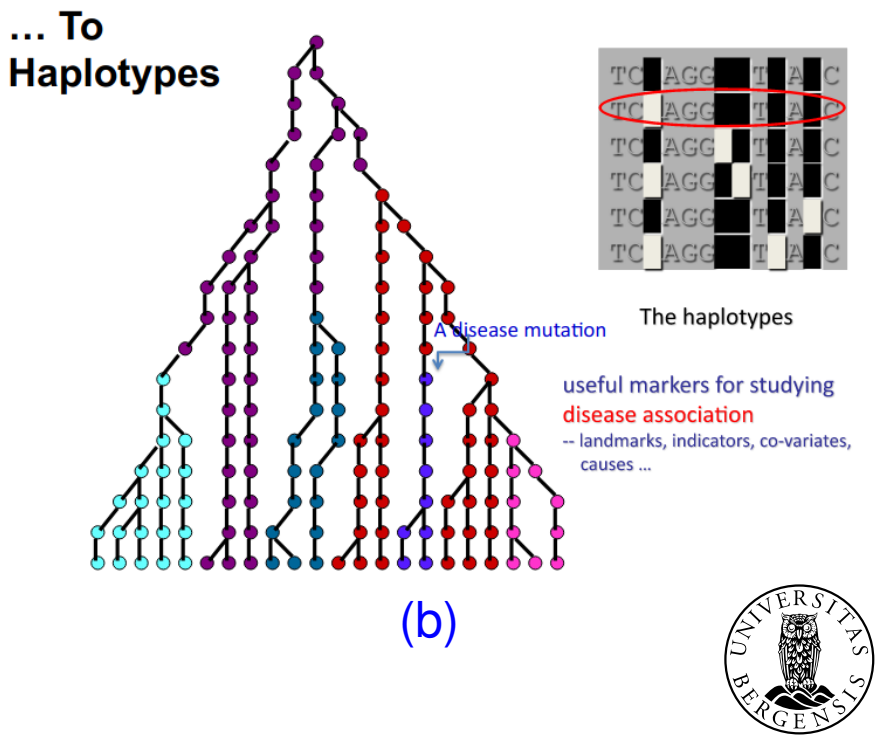


- **Nature of the phenotype being examined** : normalized gene expression levels.
- **Tissue specific** expression patterns of mRNA could vary greatly between tissue-types within the same individual.

Why eqtl analysis?

- Discover genetic factors causing certain diseases.
- Determine hotspots
- Construct causal networks
- Select genes for clinical trials.



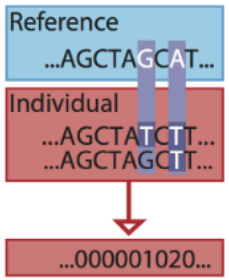


Structure of an eQTL Study

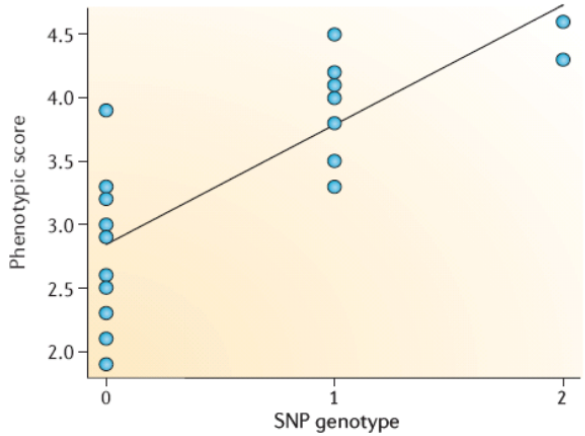
- Consider each gene's expression as a quantitative trait.
- Extract cells of the tissue of interest and their RNA.
- Measure expression of proteins by microarray or RNA-seq analysis.

geneid	Sam_01	Sam_02	...
Gene_01	4.91	4.63	...
Gene_02	13.78	13.14	...
...	...	...	...

- Regress expression levels of each gene on genotypes.



(c)



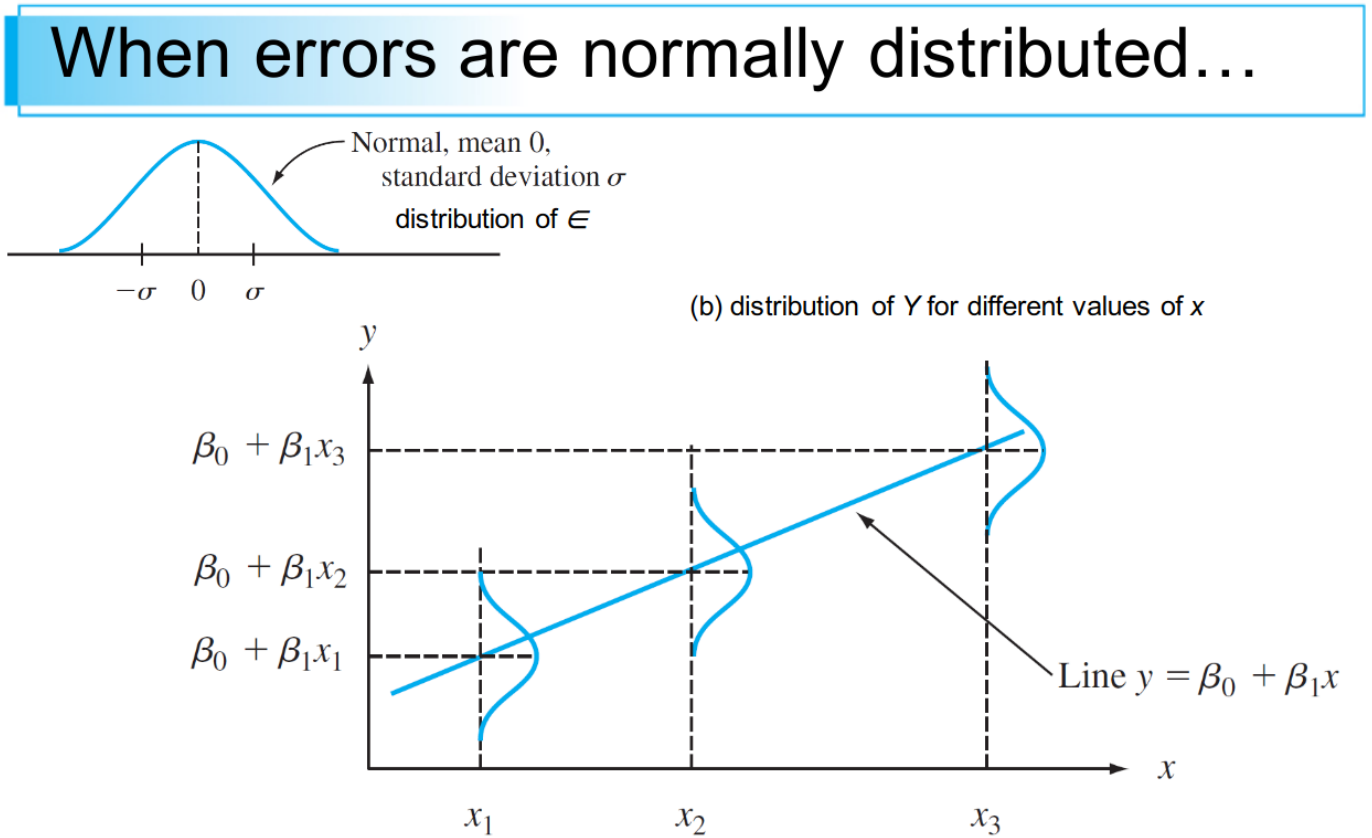
(d)

Simple linear regression

- For each gene-SNP pair, the association between gene expression y and genotype x is assumed to be linear:

$$y = \beta_0 + \beta_1 x + \epsilon, \ \textit{where}\ \epsilon \sim \mathcal{N}(0, \sigma^2). \quad (1)$$

Each SNP is encoded by 0,1 and 2 according to the frequency of the minor allele.



Cost function

We want to find the estimates for β_0 and β_1 such that we find the "best fit" for our data. Note, for each sample i, we have $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. This implies,

$$p(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(\epsilon_i)^2}{2\sigma^2}\right) \tag{2}$$

Since $E[\epsilon_i] = 0$,
$E[y_i|x_i] = \beta_0 + \beta_1 x_i$
And $Var(\epsilon_i) = \sigma^2$,
$Var(y_i|x_i) = \sigma^2$
This implies,

$$p(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right) \tag{3}$$

This is called the likelihood function

Estimates for β_0 and β_1

Obtain estimates for β_0 and β_1 by maximizing

$$p(y|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right) \tag{4}$$

for $i = 1, ..., n$ samples.

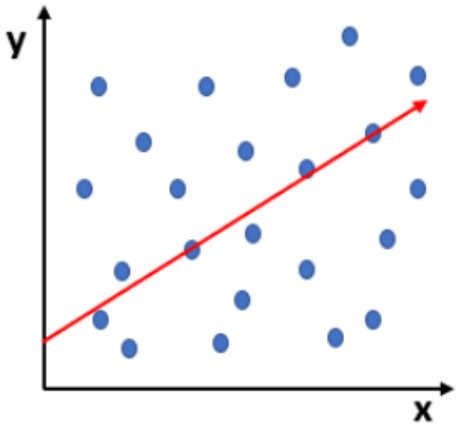
- Maximizing the log of this likelihood function is the same as minimizing the least-squares equation $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$
- Calculate $\frac{\partial p(y|x)}{\partial \beta_0}$ and $\frac{\partial p(y|x)}{\partial \beta_1}$ for $\hat{\beta}_0$ and $\hat{\beta}_1$
- We get for β_0,

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \tag{5}$$

- We get for β_1,

$$\hat{\beta}_1 = \frac{\sum_{i=1}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Why test if the fitted line is significant or not?



- Data in randomly scattered and doesn't seem to follow linear trend.
- Perform Hypothesis Testing to see how significant β is.
  - Set the Hypothesis
  - Set the Significance Level
  - Compute the test statistics
  - Make a decision

Hypothesis Test

- We start by saying β is not significant, i.e., there is no relationship between g and s, therefore slope $\beta = 0$.

$$\begin{aligned} \text{Null Hypothesis } (H_0) &: \beta = 0 \\ \text{Alternate Hypothesis } (H_A) &: \beta \neq 0 \end{aligned} \tag{2}$$

- A statistical hypothesis test evaluates the plausibility of H_0 in light of the data.
- Given the data I have observed, how plausible is that $\beta = 0$, accounting for the experimental noise?

To answer the above question, we need to compare the **observed data** to data that **could have been observed if H_0 were true.**

Samples from the null distribution!

Compute a Test statistic

A hypothesis test is typically specified in terms of a **test statistic**

Definition (Test Statistic)

A numerical summary of a data-set that reduces the data to one value that can be used to perform the hypothesis test.

Given H_0, H_1 and data $y = \{y_1, ..., y_n\}$:

- From the data, compute a relevant test statistic t(y). t(y) should be chosen so that it can differentiate between H_0 and H_1.

$$t(y) \text{ is probably } \begin{cases} \text{small under } H_0 \\ \text{large under } H_1 \end{cases} \tag{7}$$

- **Test whether a coin is fair**
  - Data : sequence of 20 heads and tails.
  - Statistic T : **number T out of the 20 flips that produced a head**

Obtain the Null distribution

Definition (Null Distribution)

The null distribution is the probability distribution of the test statistic when the null hypothesis is true.

- Obtaining a null distribution : A probability distribution over the possible outcomes of g(Y) under H_0.
- **Test whether a coin is fair**
  - Null Hypothesis is that the coin is fair.
  - Record 20 flips of this coin, **under the null hypothesis** that the coin is fair, T should have a binomial distribution with parameters 0.5 and 20.
  - A random variable X follows the binomial distribution with parameters $n \in \mathrm{N}$ and $p \in [0, 1]$, $X \sim \mathrm{B}(n, p)$. The probability of getting exactly k successes in n independent Bernoulli trials is given by

$$P(X) = \binom{n}{k} \cdot p^k q^{n-k}$$

Compute the p-value

- In a hypothesis test, a Type I error occurs when you reject a null hypothesis that is actually true. Statistical significance, α is probability of Type 1 error given null hypothesis is true.

Type I and II Errors

Decision	H_0 True	H_0 False
Do Not Reject H_0	Correct Decision $1 - \alpha$	Incorrect Decision Type II Error β
Rejct H_0	Incorrect Decision Type I Error α	Correct Decision $1 - \beta$

Actual Situation "Truth"

$$\alpha = P(Type\ I\ Error) \quad \beta = P(Type\ II\ Error)$$

Compute the p-value

- Consider an observed test-statistic t from unknown distribution T. Then the p-value p is probability of observing a test-statistic value at least as "extreme" as t if null hypothesis H_0 is true

$$p = Pr(|T| \geq |t||H_0) \tag{9}$$

  If the p-value is small $\Rightarrow$ evidence against H_0
  If the p-value is large $\Rightarrow$ not evidence against H_0
- In the example to test if a coin is fair or not:
  - Set the level of significance of 0.05
- Let us say we observed 14 heads i.e. T = 14
- We can find out that the probability of observing at least 14 is 0.015 $\geq$ 0.05

Hence we cannot reject the null hypothesis!

Example of Hypothesis test

Consider a simple hypothesis test:
$\{Y_1, ..., Y_n\} \sim i.i.d.\ P$, with mean μ and variance σ^2

$$H_0 : \mu = \mu_0 \tag{10}$$
$$H_1 : \mu \neq \mu_0$$

To test H_0, we need a test statistic and its distribution under H_0.

- $\bar{Y}$ **would make a good test statistic**
- Its distribution is approximately known:
  - $E(\bar{Y}) = \mu$
  - $V(\bar{Y}) = \frac{\sigma^2}{n}$
  - $\bar{Y}$ **is approximately normal**.

Example of Hypothesis Test

- Under H_0:

$$f(\mathbf{Y}) = \frac{\bar{Y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{11}$$

  is approximately standard normal and we write $f(\mathbf{Y}) \sim normal(0, 1)$. Is $f(\mathbf{Y})$ a statistic?
  - Problem: Don't usually know σ^2.
  - Solution: Approximate σ^2 with s^2.

- **One sample t-statistic:**

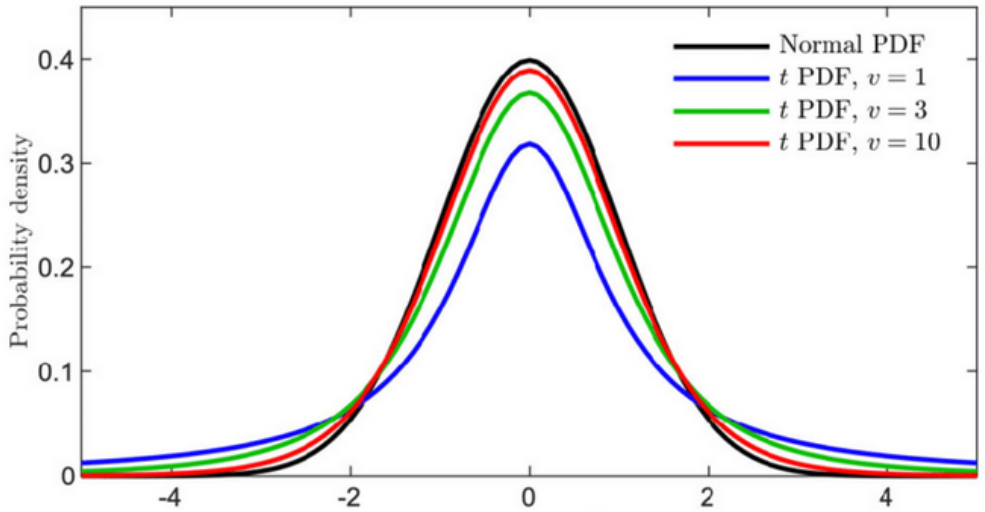
$$t(\mathbf{Y}) = \frac{\bar{Y} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{12}$$

Null distribution of the t-statistic

What is the null distribution of $t(\mathbf{Y})$?

$$t(\mathbf{Y}) = \frac{\bar{Y} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}, \text{ if } E(Y) = \mu_0 \tag{13}$$

This is a Student's t-distribution with n-1 degrees of freedom.



Basic Framework of Anova

- Want to study the effect of a qualitative variable (i.e SNP) on a quantitative variable (i.e gene expression levels).
- Characteristics that differentiate qualitative variables is called levels (three genotypes for a SNP)
- Motivating Anova:

Genotyping of a single SNP

- AA: 82, 83, 97
- AG: 83, 78, 68
- GG: 38, 59, 55

- Hypotheses of ANOVA:
  - H_0 : The mean expression levels of all 3 genotypes under consideration are equal
    H_1 : The mean expression levels are not all equal.

Rational of ANOVA

- Basic idea is to partition total variation of the data into two sources

    1. Variation within levels (groups)
    2. Variation between levels (groups)

- If H_0 is true the *standardized* variances are equal to one another

The Details

Our Data:

AA:	82, 83, 97	$\bar{x}_{1.} = (82 + 83 + 97)/3 = 87.3$
AG:	83, 78, 68	$\bar{x}_{2.} = (83 + 78 + 68)/3 = 76.3$
GG:	38, 59, 55	$\bar{x}_{3.} = (38 + 59 + 55)/3 = 50.6$

- Let X_{ij} denote the data from the i^{th} level and j^{th} observation

- Overall, or ***grand mean***, is:

$$\bar{x}_{..} = \sum_{i=1}^{K} \sum_{j=1}^{J} \frac{x_{ij}}{N}$$

$$\bar{x}_{..} = \frac{82 + 83 + 97 + 83 + 78 + 68 + 38 + 59 + 55}{9} = 71.4$$

Partitioning Total Variation

- Recall, variation is simply average squared deviations from the mean

$$SST = SST_G + SST_E$$

$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij} - \bar{x}_{..})^2 \qquad \sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.} - \bar{x}_{..})^2 \qquad \sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij} - \bar{x}_{i.})^2$$

Sum of squared deviations about the grand mean across all N observations

Sum of squared deviations for each group mean about the grand mean

Sum of squared deviations for all observations within each group from that group mean, summed across all groups

In Our Example

$$SST = SST_G + SST_E$$

$$\sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij} - \bar{x}_{..})^2 \qquad \sum_{i=1}^{K} n_i \bullet (\bar{x}_{i.} - \bar{x}_{..})^2 \qquad \sum_{i=1}^{K}\sum_{j=1}^{J}(x_{ij} - \bar{x}_{i.})^2$$

$$(82-71.4)^2 + (83-71.4)^2 + (97-71.4)^2 +$$
$$(83-71.4)^2 + (78-71.4)^2 + (68-71.4)^2 +$$
$$(38-71.4)^2 + (59-71.4)^2 + (55-71.4)^2 =$$

$$3 \bullet (87.3-71.4)^2 +$$
$$3 \bullet (76.3-71.4)^2 +$$
$$3 \bullet (50.6-71.4)^2 =$$

$$(82-87.3)^2 + (83-87.3)^2 + (97-87.3)^2 +$$
$$(83-76.3)^2 + (78-76.3)^2 + (68-76.3)^2 +$$
$$(38-50.6)^2 + (59-50.6)^2 + (55-50.6)^2 =$$

2630.2 2124.2 506

Calculating Mean Squares

- To make the sum of squares comparable, we divide each one by their associated degrees of freedom
    - SST_G = k - 1 (3 - 1 = 2)
    - SST_E = N - k (9 - 3 = 6)
    - SST_T = N - 1 (9 - 1 = 8)

- MST_G = 2124.2 / 2 = 1062.1

- MST_E = 506 / 6 = 84.3

Almost There… Calculating F Statistic

- The test statistic is the ratio of group and error mean squares

$$F = \frac{MST_G}{MST_E} = \frac{1062.2}{84.3} = 12.59$$

- If H_0 is true MST_G and MST_E are equal

- Critical value for rejection region is $F_{\alpha, k-1, N-k}$

- If we define $\alpha = 0.05$, then $F_{0.05, 2, 6} = 5.14$

Multiple testing problem

Multiple testing refers to any instance that involves the simultaneous testing of more than one hypothesis.

Why Multiple Testing Matters

Genomics = Lots of Data = Lots of Hypothesis Tests

A typical microarray experiment might result in performing 10000 separate hypothesis tests. If we use a standard p-value cut-off of 0.05, we'd expect ***500*** genes to be deemed "significant" by chance.

Multiple testing problem

Why Multiple Testing Matters

- In general, if we perform m hypothesis tests, what is the probability of at least 1 false positive?

$$P(\text{Making an error}) = \alpha$$

$$P(\text{Not making an error}) = 1 - \alpha$$

$$P(\text{Not making an error in m tests}) = (1 - \alpha)^m$$

$$P(\text{Making at least 1 error in m tests}) = 1 - (1 - \alpha)^m$$

20 hypotheses to test with a significance level of 0.05, a 64% chance of observing at least one error!

Ways to control Type 1 error

Testing of multiple hypotheses, if you have 100 hypotheses, you reject 5 true nulls by chance!

- One solution is **Bonferroni correction:**
  - If we have a 100 hypotheses each with a significance level of $\alpha = 0.05$, choose new significance level,

$$\alpha^* = \frac{0.05}{100} = \frac{\text{Original}\,\alpha}{\text{no. of hypotheses}} \tag{14}$$

- This ensures that the Family wise error rate (FWER) $\leq \alpha$
- $FWER = P(\textbf{making at least one type I error}) = P(V \geq 1)$
- V is the number of false positives (Type I error) (also called "false discoveries").

False Discovery rate

	Null hypothesis is true (H_0)	Alternative hypothesis is true (H_A)	Total
Test is declared significant	V	S	R
Test is declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

- m is the total number hypotheses tested
- m_0 is the number of true null hypotheses
- V is the number of false positives (Type I error) (also called "false discoveries")
- S is the number of true positives (also called "true discoveries")
- T is the number of false negatives (Type II error)
- U is the number of true negatives
- $R = V + S$ is the number of rejected null hypotheses

False Discovery rate

R is an observable random variable, and S, T, U, and V are unobservable random variables.

- False discovery rate (FDR) is the expected proportion of tests which are incorrectly called significant out of all the tests which are called significant.
- **FDR = E[Q]**

$$Q = \begin{cases} \frac{V}{R} \; R \geq 0 \\ 0 \; R = 0 \end{cases} \tag{15}$$

- The Benjamini-Hochberg (BH) method is a procedure which controls the false discovery rate so that $FDR \leq \alpha$

Benjamini and Hochberg procedure

Suppose we have computed the p-values for M hypothesis tests:

$$H_{0j} \; vs \; H_{1j} \; j = \{1, ..., m\} \tag{16}$$

The Benjamini-Hochberg method can be performed as follows

- Decide on an FDR level q^*
- Order the p-values

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)} \tag{17}$$

- Let k be the largest i such that

$$p_{(i)} \leq \frac{i}{m} q^* \tag{18}$$

  (i.e. start with $p_{(m)}$ and go backwards)
- Reject hypotheses 1...k, i.e $1 - k$ discoveries
- Adjusted p-values will be the minimal $\frac{m}{i} p$ upto that point.

Example

B&H FDR Example

Controlling the FDR at $\delta = 0.05$

Rank (j)	P-value	(j/m)× δ	Reject H_0 ?
1	0.0008	0.005	1
2	0.009	0.010	1
3	0.165	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

Computation for Eqtl analysis

- Computationally expensive test for association of billions of transcript-SNP pairs.
  - millions of SNPs
  - tens and thousands of gene expression transcripts

Over ten million tests!

Matrix eQTL:

- New software for efficient eqtl analysis.
- 2-3 orders of magnitude faster than existing popular tools while finding the same eqtls.

What Matrix eQTL can do

- Test for association b/w each SNP and each transcript by modelling the effect of genotype as additive linear or categorical.
- Can test for significance of genotype-covariate interaction.
- Covariates such as gender, population structure, clinical variables.
- Perform a separate test for each SNP-gene pair and correct for multiple comparisons by calculating FDR.
- Supports separate p-value thresholds and FDR calculation for cis and trans eqtls.

What we will learn

- **How this performance is achieved:**
  - prepossessing
  - expressing the most computationally intensive part of the algorithm in terms of large matrix operations.
- How it supports linear/ANOVA models + models with correlated and heteroskedastic errors.
- How the issue of multiple testing is adressed by calculatiing FDR.
- How to use the software + Exercises.

Simple Linear Regression in Matrix eQTL

For each gene-SNP pair, the association between gene expression g and genotype s is assumed to be linear:

$$g = \alpha + \beta s + \epsilon, \ \textit{where}\ \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{19}$$

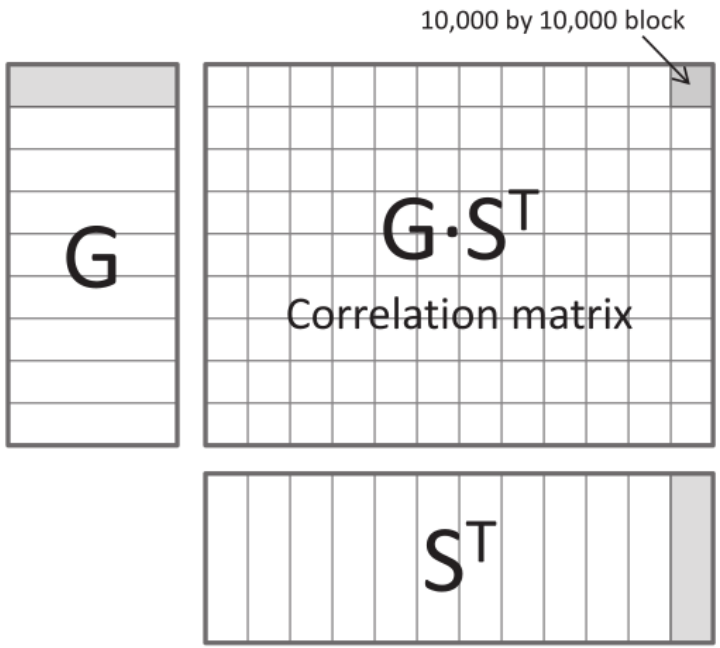
- Test statistic of choice : $|r| = cor(g, s)$
- Standardize the genotype and gene expression variables, g and s.
- Calculation of the sample correlation

$$r_{gs} = cor(s, g) = \frac{\sum (s_i - \tilde{s})(g_i - \tilde{g})}{\sqrt{\sum (s_i - \tilde{s})^2 \sum (g_i - \tilde{g})^2}} = \sum s_i g_i = \langle s, g \rangle \tag{20}$$

where $\langle s, g \rangle$ denotes the inner product between vectors s and g.

Simple Linear Regression in Matrix eQTL

- Let G be the gene expression matrix
  - Each row contains measurements for a single gene across samples.
- Let S be the genotype matrix
  - Each row contains measurements for a single SNP across samples.
- Matrix of all gene- SNP correlations can be calculated in one large matrix multiplication GS^T



Algorithm for Simple Linear Regression

Full GS^T would require hundreds of gigabytes of RAM. Slice the data matrices in blocks of up to 10000 variables and perform the analysis separately for each pair of blocks.

- Split input matrices into blocks of up to 10 000 variables
- Standardize variables of both gene expression and genotype matrices.
- For each pair of blocks:
  - Calculate the corresponding block of the correlation matrix in one matrix multiplication.
  - Find correlations which have absolute value that exceeds a predefined threshold.
  - For the selected correlations, calculate and report the corresponding test statistic, p-value, FDR and other variables of interest.

Model with covariates

Include covariates in an eQTL model to account for such effects as population stratification, gender, age, white blood count and other clinical variables.

$$g = \alpha + \gamma x + \beta s + \epsilon \tag{21}$$

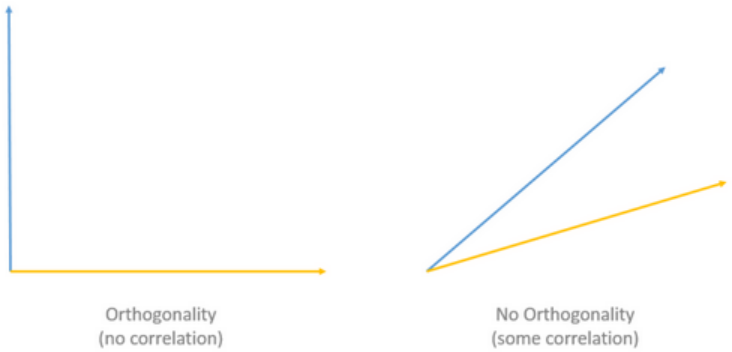
For fast computations, the previous model can be reduced to testing of the simple linear regression model by orthogonalizing g and s with respect to x. The algorithm for the analysis is then:

- Center variables g, x, and s to remove constant α from the model.
- Orthogonalize g and s with respect to x:

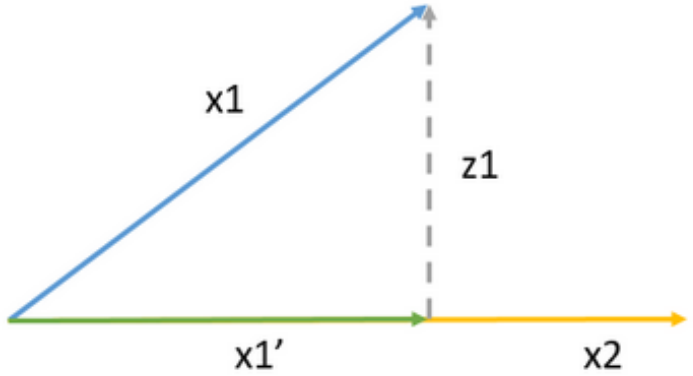
$$\tilde{g} = g - \langle g, x \rangle x \ , \ \tilde{s} = s - \langle s, x \rangle x. \tag{22}$$

- Perform the analysis for the simple linear regression $\tilde{g} = \beta \tilde{s} + e$ using one less degree of freedom for the test statistic to account for the removed covariate.

Orthogonalization



(e)



(f)

- x1 is orthogonally projected onto x2, thereby production green vector x1'.
- Subtract x1' from the blue vector x1, here is the residual vector z1
- z1 is the orthogonalized vector of x1 with respect to x2.

ANOVA model

Treat each genotype variable as categorical and modeling its effect on gene expression.

$$g = \alpha + \beta_1 s_1 + \beta_2 s_2 + \epsilon \tag{23}$$

where $s_1 = I(s = 1)$ and $s_2 = I(s = 2)$ are dummy variables constructed for the SNP's.

ANOVA algorithm

- Center variables g, s_1 and s_2 to remove the constant α from the model.
- Orthogonalize s_2 with respect to s_1 for every marker

$$\tilde{s}_2 = s_2 - \langle s_2, s_1 \rangle s_1 \tag{24}$$

- Standardize s_1 and $\tilde{s}_2$
- Calculate test statistic:

$$R^2 = \langle g, s \rangle^2 - \langle g, \tilde{s}_2 \rangle^2 \tag{25}$$

  via large matrix operations.
- The threshold for R^2 and p-values can be derived from the formula for F-test

$$F = \frac{(n-k-1)R^2}{k(1-R^2)} \tag{26}$$

  where $k = 2$ is the number of regressors (s_1 and s_2).

Heteroskedastic and/or correlated errors

Errors can be heteroskedastic if the quality of the measurements differs across samples. The errors may also be correlated if the samples are taken from related individuals.

$$g = \alpha + \beta s + u, \ \textit{where } U \sim \mathcal{N}(0, \sigma^2 K). \tag{27}$$

K is a known positive-definite covariance matrix

- For this model, transform the input variables to make the errors independent and identically distributed:

$$\tilde{g} = K^{-\frac{1}{2}} g, \ \tilde{s} = K^{-\frac{1}{2}} s, \ \tilde{q} = K^{-\frac{1}{2}} I_n \tag{28}$$

  where I_n is a vector of length n with unit elements.
- The new model equation is homoskedastic, has independent errors, but does not include a constant.

$$\tilde{g} = \alpha \tilde{q} + \beta \tilde{s} + e, \ \textit{where } e \sim \mathcal{N}(0, \sigma^2). \tag{29}$$

Test using the algorithm for the linear model with covariates.