BTRY 7210: Topics in Quantitative Genomics and Genetics

Jason Mezey
Biological Statistics and Computational Biology (BSCB)
Department of Genetic Medicine
jgm45@cornell.edu

February 12, 2015
Lecture 3: Biological false positive control in eQTL analysis
Reminder: what is an eQTL?

A_1 \rightarrow A_2 \Rightarrow \triangle \bar{Y}

rs27290 genotype

ERAP2 expression

A/A  A/G  G/G

depends on the reference; microsatellites or Short Sequence Repeats in cases where there are different numbers of short repeated motifs in one genome compared to another; Copy Number Variation in cases where one individual has a repeat of a long stretch of DNA compared to another individual or has less 'copies' of a repeated long stretch; transposable elements well-characterized segments of DNA that can excise and move to another location; sometimes these elements code for the proteins that cause them to move; chromosomal rearrangements and aberrations in any case where a large portion of a chromosome is moved to another location or cases where there is an 'extra' chromosome. Note that in all of these cases, describing the mutation depends on finding the 'same place' in the genome, something that we will take for granted in this course. Also, note that for our purpose, we don't care about the actual description of the genetic difference, only that there is a difference and we will code all such differences using the same system.

The central goal of the methodology we will learn in this course is to identify causal mutations that have an effect on a phenotype, in any aspect of an organism we can measure. We can define causal as follows:

causal mutation \equiv \text{a position in the genome where an experimental manipulation of the DNA produces an effect on the phenotype on average.}

There are a couple of strange components to this definition. When we say experimental manipulation, we mean a case where we physically change the state of the DNA at a position in the genome, i.e., produce a mutation. This is actually possible in some organisms (mice, flies, etc.), but we often will not actually do this manipulation, but rather find places where we assume such a manipulation will produce an effect. When we say an 'effect' on the phenotype, we ideally mean that if we were to take an individual, produce an identical clone grown under the same environmental conditions, where the only difference is a single experimental mutation, the phenotype would be different between the two individuals. In general, it is impossible to keep all conditions exactly the same, so we consider a case where, if we were to produce an experimental manipulation keeping most genetic and environmental factors the same, the result would be a change in the phenotype on average, i.e., that is to say, if we were to average over several manipulations, some of the manipulations would result in a change of the phenotype.
Typical outcome of a genome-wide eQTL analysis I

-\log_{10}(p\text{-value})

Chromosome

ERAP2

eQTL \ (p < 10^{-30})

rs27290 genotype

A/A

rs1908530 genotype

T/T

no eQTL \ (n.s.)

A/G

G/G

T/C

C/C
Typical outcome of a genome-wide eQTL analysis II
Considering cis- vs trans- eQTLs
Considering cis- vs trans- eQTL II
eQTL false positives

- In a genome-wide eQTL analysis we are more concerned about errors in identifying a position as having an eQTL when there is not one, as opposed to “missing” a location that has an eQTL (why?)

- The first of these errors is a “biological false positive” (the second is a “biological false negative”)

- There are both statistical and experimental (and combined) reasons why we can get a biological false positive:
  - Type I error (statistical)
  - Genotyping / Expression measurement errors (experimental)
  - Cases of disequilibrium when there is no linkage (experimental)
  - Unaccounted for covariates (both)
Type I error (statistical)

<table>
<thead>
<tr>
<th>Cannot reject $H_0$</th>
<th>$H_0$ is true</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$1 - \alpha$, (correct)</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>$\alpha$, type I error</td>
</tr>
</tbody>
</table>

- Note that to avoid Type I errors when performing the many tests involved in an eQTL analysis, we use multiple test corrections, generally Bonferroni or Benjamini-Hochberg.

- What are the trade-offs of the more conservative (Bonferroni) vs less conservative (Benjamini-Hochberg) controls of Type I error?
Genotyping error can produce false positives, where genotypes with lower minor allele frequencies (MAF) are more susceptible (why?)

Do these impact both cis- and trans- hits? How might we detect that a genotyping error has occurred from the patterns of linkage disequilibrium (LD)?
• Expression measurement error can produce false positives, where both gene expression array and RNA-Seq measurements are susceptible for different reasons (why?)
A possible problem with arrays is probes that are polymorphic in the population (=do not perfectly match the DNA of an individual).

A possible problem with RNA-Seq is pseudogenes that lead to incorrect mapping of sequencing reads (=incorrect quantification of mRNA at a gene for an individual).
Disequilibrium but not LD (experimental)

- A rule of genetics is that genotypes that tend to be physically close to each other in the genome have a high correlations (=linkage disequilibrium or LD) and the further away genotypes are from one another, the lower their correlations (in general):

- While generally true, this is not always the case (!!)

- Some issues to watch out for: (a) X and Y chromosomes, (b) low MAF markers in high marker coverage experiments, (c) inbreeding designs
Unaccounted for covariates (statistical and experimental)

• If the marker is not correlated with a causal polymorphism but the factor is correlated with BOTH the phenotype and the marker such that a test of the marker using our framework will produce a false positive (!!):

\[ Cov(Y, X_z) \neq 0. \]
\[ Cov(X_a, X_z) \neq 0 \]
\[ H_0 : \beta_a = 0 \cap \beta_d = 0 \]
\[ H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \]
\[ Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \]

• This can also lead to reduced power...
Unaccounted for covariates (statistical and experimental)

- Therefore, if we have a factor that is correlated with our phenotype and we do not handle it in some manner in our analysis, we risk producing false positives AND/OR reduce the power of our tests!

- The good news is that, assuming we have measured the factor (i.e. it is part of our eQTL dataset) then we can incorporate the factor in our model as a covariate:

  \[ Y = \beta_\mu + X_a\beta_a + X_d\beta_d + X_z\beta_z + \epsilon \]

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

- The effect of this is that we will estimate the covariate model parameter and this will account for the correlation of the factor with phenotype (such that we can test for our marker correlation without false positives / lower power!)

- In general, we should include covariates that might have an impact, how do we assess this? Do we include every possible covariate (why might this be a bad idea)? What if we haven't measured the covariate causing the problem(!)?
Visual diagnostics for assessing statistical model fit

- Remember that in our eQTL analysis, we are not after the “right” statistical model but rather a model that we can fit to genotype-expression pairs that produces as few biological false positives as possible while still allowing us to discover eQTL.

- Any alternative statistical modeling strategy (linear, mixed, non-parametric, etc.) and statistical model (which covariates are included, etc.) is potentially acceptable.

- However, you should always check genome-wide diagnostics of model fit before you believe the eQTL you identify (you should also visually check the expression by genotype plots, the local LD of the eQTL hit, etc.).

- By far, the two most valuable we have found are Quantile-Quantile plots (QQ) and Heatmap plots (where the latter are not in general usage).
Quantile-Quantile (QQ) plots

- For an expression variable impacted by an eQTL (i.e. you have found markers at a position significantly associated with the expression variable, you need to plot the p-values for all genotypes in your study when tested versus this expression variable using the following approach:
  
  - If you performed $N$ tests, take the -log (base 10) of each of the p-values and put them in rank order from smallest to largest
  
  - Create a vector of $N$ values evenly spaces from 1 to $1 / N$ (how do we do this?), take the -log of each of these values and rank them from smallest to largest
  
  - Take the pair of the smallest of values of each of these lists and plot a point on an x-y plot with the observed -log p-value on the y-axis and the spaced -log value on the x-axis
  
  - Repeat for the next smallest pair, for the next, etc. until you have plotted all $N$ pairs in order
  
  - This is a QQ plot and YOU SHOULD ONLY CONSIDER YOUR eQTL association to be correct if they look like the following:
Heatmap plots

- A heatmap plot is constructed by calculating p-values for every expression-genotype pair and plotting genotypes in the columns and expression variables on the rows, where colors indicate significance (while-yellow=not significant, red=significant).

- If you see horizontal stripes or vertical stripes THIS INDICATES A PROBLEM (!!).

- The horizontal stripes indicate an unaccounted for “population structure” covariate (i.e. your sample could be divided into multiple populations that differ in allele frequencies at many markers where these populations also differ in the mean value of an expressed gene = every allele that differs in frequency will be statistically associated with the gene!!) - how can we account for such population structure covariates?

- The vertical stripes indicate an unaccounted for “hidden factor” covariate (i.e. a covariate that you have not measured that impacts gene expression levels that is also correlated with markers) - how can we account for such hidden factor covariates?

- In general, YOU SHOULD NOT CONSIDER an eQTL in a horizontal or vertical stripe to be a true eQTL (!!).

- Why don’t people use heatmaps? What is a practical alternative?
Are false positives a problem in eQTL analyses?

All eQTL  
 cis-eQTL only  
 trans-eQTL only

Not really... while trans-eQTL overlap is worse than cis-eQTL overlap, Bonferroni eQTL have more overlap in both cases.
That’s it for today!