BTRY 7210: Topics in Quantitative Genomics and Genetics

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

March 12, 2015
Lecture 4: What we can infer about eQTL and by leveraging eQTL?
Areas where we can use genome-wide data analysis to learn about eQTL

- Identifying causal genotype candidates
- Leveraging eQTL to identify new regulatory relationships
- Learning about the conditional dependencies of eQTL
- Inferring the possible impacts of eQTL on complex phenotypes
Reminder: an eQTL

Genotype-phenotype association

\[ A_1 \rightarrow A_2 \Rightarrow \Delta \bar{Y} \]
Can we identify the causal genotype / polymorphism?

- Given the right study conditions and available data, yes (sometimes)!
Can we identify the causal genotype / polymorphism?

- Given the right study conditions and available data, yes (sometimes)!

```
rs2927608 G>A  rs10233171 T>C  rs1059307 G>T
NRSF binding motif  SP1 binding motif  Nkx2-5 binding motif

25.1kb CNV (esv3640682-3)

rs9897034 C>T
YY1 binding motif
```

```
ERAP2  LRRC37A2  NT5C3B  SNHG5  WBSCR27
```

```
Can we identify the causal genotype / polymorphism?

- The only case of “perfect” experimental validation of an eQTL (to date: Lee et al. 2014 Science; 343:119):

**E**

<table>
<thead>
<tr>
<th>Gene</th>
<th>rs Number</th>
<th>Variant</th>
<th>Promoter Region</th>
<th>Luciferase Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLFN5</td>
<td>rs11080327</td>
<td>A/G</td>
<td>Without IFNβ</td>
<td>+1.5</td>
</tr>
<tr>
<td>SLFN5</td>
<td>rs11080327</td>
<td>A/G</td>
<td>With IFNβ</td>
<td>+2.0</td>
</tr>
<tr>
<td>CLEC4F</td>
<td>rs35856355</td>
<td>C/A</td>
<td>Without IFNβ</td>
<td>+1.2</td>
</tr>
<tr>
<td>CLEC4F</td>
<td>rs35856355</td>
<td>C/A</td>
<td>With IFNβ</td>
<td>+2.5</td>
</tr>
<tr>
<td>ARL5B</td>
<td>rs2130531</td>
<td>G/A</td>
<td>Without IFNβ</td>
<td>+1.8</td>
</tr>
<tr>
<td>ARL5B</td>
<td>rs2130531</td>
<td>G/A</td>
<td>With IFNβ</td>
<td>+2.2</td>
</tr>
</tbody>
</table>

**F**

- **SLFN5** rs11080327
  - Wildtype HEK293
  - SLFN5 A>G HEK293
  - Scatter plot showing a significant difference in luciferase activity between wildtype and SLFN5 A>G HEK293 cell lines.
eQTL network analysis using probabilistic graphical models

- We assume a graphical model $G = (V, E)$ describes the joint distribution where each random variable is represented as a vertex and each edge represents a conditional relationship:

$$Pr(Y) = Pr(Y_1, Y_2, \ldots, Y_k)$$

$$V = (Y) = Y_1, Y_2, \ldots, Y_k$$

$$E(Y_i, Y_j) \implies Pr(Y_i|Y_j) \neq Pr(Y_i)$$
Causal network graphical modeling

• We could discover network relationships with directed models, which would imply the direction of causality. However, there is an equivalence class problem:
By leveraging an eQTL, we can reduce some of the equivalencies.
Some examples...

<table>
<thead>
<tr>
<th>cis-gene</th>
<th>trans-gene</th>
<th>e_uQTL</th>
<th>relationship present</th>
<th>relationship not present</th>
<th>+ positive correlation</th>
<th>- negative correlation</th>
</tr>
</thead>
</table>

- KRTAP21–2
- MRPS18B
- SLC41A2
- SELK
- CLEC5A
- PAPPA
- P2RY2
- SNRPC
- MAPK8IP1
- DCAKD
- LINC00937
- PRSS53
- LOC100130264
- CDYL2

- Adipose (GTEx)
- Adipose (MuTHER)
- B Cells (Geuvadis)
- B Cells (MuTHER)
- Blood (GTEx)
- Blood (DGN)
- Breast (GTEx)
- Breast (TCGA Ctrl)
- Lung (GTEx)
- Lung (TCGA Ctrl)
- Skin (GTEx)
- Skin (MuTHER)

- ERAP2
- LRRC37A2

- relationship present
- relationship not present
- positive correlation
- negative correlation
Conceptually: what are the impacts of eQTL we are measuring?

- In most human genome-wide eQTL studies, expression is measured in a “tissue” (very broadly defined, e.g. blood, skin...) sampled under uncontrolled conditions (i.e. in vivo)

- This means each sample has an unknown mixture of cell populations, has unknown factors that apply to it (e.g. exposure of individuals in the study to X, etc.)

- Many eQTL cannot be replicated, such that many may be dependent (=only exist) given unknown conditional dependencies

- Are such results useful? When are they useful?
If we find that eQTL replicate for a known condition...

### e.g. all conditions

### e.g. for a specific tissue

**Dataset-wide Bonferroni**
- Dataset-wide Bonferroni signif.
- Gene not measured

**Locally Bonferroni signif.**
- + positive correlation
- - negative correlation
- relationship present
- relationship not present

**cis-gene**
- cis-gene

**trans-gene**
- trans-gene

**eQTL**
- eQTL

**Adipose (GTeX)**
- Adipose (MuTHER)
- B Cells (Geuvadis)
- B Cells (MuTHER)
- Blood (GTeX)
- Blood (DGN)
- Breast (GTeX)
- Breast (TCGA Ctrl)
- Lung (GTeX)
- Lung (TCGA Ctrl)
- Skin (GTeX)
- Skin (MuTHER)
- Nerve (GTeX)
- Muscle (GTeX)
- Thyroid (GTeX)

**KRTAP21**
- MRPS18B
- SLC41A2
- SELK
- CLEC5A
- PAPPA
- P2RY2
- SNRPC
- MAPK8IP1
- DCARK
- LINC00937
- PRSS53
- LOC101928188
- FGD3
- OR2A1
- UNC93B1
- ARPP21

**ERAP2**

**LRRC37A2**

**WBSCR27**
- cis-gene
- trans-gene
- eQTL
- relationship present
- relationship not present
- + positive correlation
- - negative correlation
What are known conditional dependencies useful for?

- If we are sure that eQTL (ideally we know the causal alleles) are responsible for a change in gene expression, we at least know this is functional genetic variation, i.e. it at LEAST impacts gene expression under certain these certain conditions.

- Such eQTL may also more effects on many other phentoypes, e.g. many (many) studies are finding that eQTL and GWAS hits co-locate in the genome, suggesting a common genetic basis.

- If an eQTL is specific to a tissue (or under a specific condition) this might help draw a connection to a specific disease.
That’s it for today!