RegulomeDB and HaploReg Exercises

Exercise #1
rs2816316 has been associated with Celiac Disease in the European population by two studies (Hunt, …, van Heel (2008) *Nature Genetics* and Dubois, …, van Heel (2010) *Nature Genetics*). rs2816316 lies thousands of base pairs upstream of protein coding gene RGS1 in an intergenic region of the genome.

You decide to further investigate this SNP using RegulomeDB and HaploReg.

1. What score does RegulomeDB assign to rs2816316? Is this SNP likely to affect transcription factor binding?

2. Using HaploReg, determine if there are any SNPs in high LD with rs2816316. Are any of these SNPs more likely to be causal?

3. Using RegulomeDB, determine the scores for each of the SNPs in LD with rs2816316 that you think may be causal. Is there a SNP that is likely to affect transcription factor binding? Which SNP(s) would you further investigate?

Exercise #2
You are interested in studying genetics variants associated with Amyotrophic lateral sclerosis (ALS), which causes muscle atrophy due to the degeneration of motor neurons. Eleven studies have reported 66 SNPs associated with ALS. Since little is known about the disease, you decided to investigate these genetic variants.

1. Using HaploReg, determine if there are enrichments for enhancers in any ENCODE cell types for these ALS SNPs. Are there enrichments in DNase regions?

2. Perform the same analysis using Roadmap epigenomes. Are disease relevant tissue and cell types enriched?

Exercise #3
rs6774494 has been associated with Nasopharyngeal carcinoma in the Chinese population (Bei, …, Zeng (2010) *Nature Genetics*).

1. What score does RegulomeDB assign to rs6774494? Is this SNP likely to affect transcription factor binding?
2. Using HaploReg, determine if there are any SNPs in high LD with rs2816316. Are any of these SNPs more likely to be causal?

3. How would your results change if you used the default settings for HaploReg (i.e. European LD?)
SOLUTIONS

Exercise #1

rs2816316 has been associated with Celiac Disease in the European population by two studies (Hunt, …, van Heel (2008) *Nature Genetics* and Dubois, …, van Heel (2010) *Nature Genetics*). rs2816316 lies thousands of base pairs up stream of protein coding gene RGS1 in an intergenic region of the genome.

You decide to further investigate this SNP using RegulomeDB and HaploReg.

1. What score does RegulomeDB assign to rs2816316? Is this SNP likely to affect transcription factor binding?

   RegulomeDB assigns rs2816316 a score of 5, which means that there is minimal binding evidence. This SNP is not likely to affect TF binding.

2. Using HaploReg, determine if there are any SNPs in high LD with rs2816316. Are any of these SNPs more likely to be causal?

   There are 25 SNPs in LD ($r^2 > 0.8$) with rs2816316. There are three SNPs that overlap TF binding sites: rs2816305, rs2984920 and rs7535818. These SNPs also overlap DHSs, promoter marks, and enhancer marks for several cell lines.
rs2984920 is a strong candidate as it overlaps regulatory marks in the most cell lines. It also disrupts a PU.1 motif (Log odds drop from 14.5 to 2.9) and overlaps a PU.1 binding site. It is also in the promoter of RGS1.

rs7535818 primarily overlaps POL2 binding sites which suggests it would not affect regulation but is instead in an actively transcribed region. It is also in the promoter/first intron of RGS1.
rs2816305 also overlaps regulatory regions and TFs. It overlaps some motifs but not those corresponding to TFs with overlapping binding sites. However, it is important to remember that not all TFs are surveyed by ENCODE.

### Proteins bound by ChIP (ENCODE)

<table>
<thead>
<tr>
<th>Cell ID</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM12878</td>
<td>BCL11A</td>
</tr>
<tr>
<td>GM12878</td>
<td>OCT2</td>
</tr>
<tr>
<td>GM12878</td>
<td>PBX3</td>
</tr>
<tr>
<td>GM12878</td>
<td>POL24H8</td>
</tr>
<tr>
<td>GM12878</td>
<td>POU2F2</td>
</tr>
<tr>
<td>GM12878</td>
<td>TBP</td>
</tr>
</tbody>
</table>

### Regulatory motifs altered

<table>
<thead>
<tr>
<th>PWM</th>
<th>Strand</th>
<th>Ref</th>
<th>Alt</th>
<th>Match on:</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP-1_3</td>
<td>+</td>
<td>13.2</td>
<td>1.4</td>
<td>RCTGRTHVDCYD8</td>
</tr>
<tr>
<td>NF-Y_known2</td>
<td>-</td>
<td>-5.7</td>
<td>6.3</td>
<td>CTGATTGCTG</td>
</tr>
<tr>
<td>Pax-8_1</td>
<td>+</td>
<td>6.2</td>
<td>10.4</td>
<td>DSHEBRWPHMTGADNHV</td>
</tr>
</tbody>
</table>

3. Using RegulomeDB, determine the scores for each of the SNPs in LD with rs2816316 that you think may be casual. Is there a SNP that is likely to affect transcription factor binding? Which SNP(s) would you further investigate?

- rs2816305 = 1d
- rs2984920 = 2a
- rs7535818 = 3a
It would be worthwhile to further investigate both rs2984920 and rs2816305. rs2816305 has the lowest RegulomeDB score since it was reported to be an eQTL for RGS1. It does not overlap a motif corresponding to a bound TF but is in a regulatory region. rs2984920 lies in the promoter of RGS1 and overlaps motifs for several bound TFs including PU.1 and NFkB (discovered by RegulomeDB). rs2984920 and rs2816305 are also in LD, so the eQTL signal from rs2816305 could be due to rs2984920. Both SNPs would be worth investigating further to determine the casual variant.

Exercise #2
You are interested in studying genetics variants associated with Amyotrophic lateral sclerosis (ALS), which causes muscle atrophy due to degeneration of motor neurons. Eleven studies have reported 66 SNPs associated with ALS. Since little is known about the disease, you decided to investigate these genetic variants.

1. Using HaploReg, determine if there are enrichments for enhancers in any ENCODE cell types for these ALS SNPs. Are there enrichments in DNase regions?
   - HepG2 – Strong Enhancers
   - HMEC – Strong Enhancers
   - GM12878 – All Enhancers & Strong Enhancers
   - DNase: HFF-Myc, HA-sp, Th2, and GM18507

2. Perform the same analysis using Roadmap epigenomes. Are these disease relevant tissue and cell types enriched?
   - Colon: All Enhancers
   - Penis Foreskin: Strong Enhancers
   - Brain Substantia Nigra: All Enhancers
   - Brain Inferior Temporal lobe: All Enhancers
   - Brain Cingulate Gyrus: All Enhancers
Exercise #3
r6774494 has been associated with Nasopharyngeal carcinoma in the Chinese population (Bei, ..., Zeng (2010) Nature Genetics)

1. What score does RegulomeDB assign to r6774494? Is this SNP likely to affect transcription factor binding?
   There is no data for r6774494 in RegulomeDB so no score is assigned. This SNP is unlikely to affect TF binding.

2. Using HaploReg, determine if there are there any SNPs in high LD with r2816316. Are any of these SNPs more likely to be causal?
   There are ten SNPs in LD with r2816316. r9869781 and r13322424 overlap enhancer marks, promoter marks, DNase peaks and TF binding sites and would be worth investigating further.
3. How would your results change if you used the default settings for HaploReg (i.e. European LD?)

If you use European LD, there are only two SNPs in LD with rs6774494 and neither are as likely to affect TF binding as rs9869781 and rs1332242.