Final Project: Exercise

Joshua Burkhart

December 14, 2015

Final Project: Exercise

Loading the demo data

```
library(DOQTL)
setwd("/Users/joshuaburkhart/SoftwareProjects/ComputationalGenetics/FinalProject")
load("DOQTL_demo.Rdata")
ls()
```

```
[1] "founder.probs" "pheno"
```

```
head(pheno)
```

```
   Sample.ID Sex Study Dose prop.bm.MN.RET
15        15   M     1  100        0.01150
16        16   M     1  100        0.01060
24        24   M     1  100        0.03430
35        35   M     1  100        0.02365
37        37   M     1  100        0.00620
44        44   M     1  100        0.01550
```

QTL Mapping

```
K = kinship.probs(founder.probs)
addcovar = matrix(pheno$Study)
rownames(addcovar) = rownames(pheno)
colnames(addcovar) = "Study"
setwd("/Users/joshuaburkhart/SoftwareProjects/ComputationalGenetics/FinalProject")
load("muga_snps.Rdata")
```

Q1. How does the biological question translate to a statistical hypothesis? Include the biological question, hypothesis and any assumptions in the write-up.

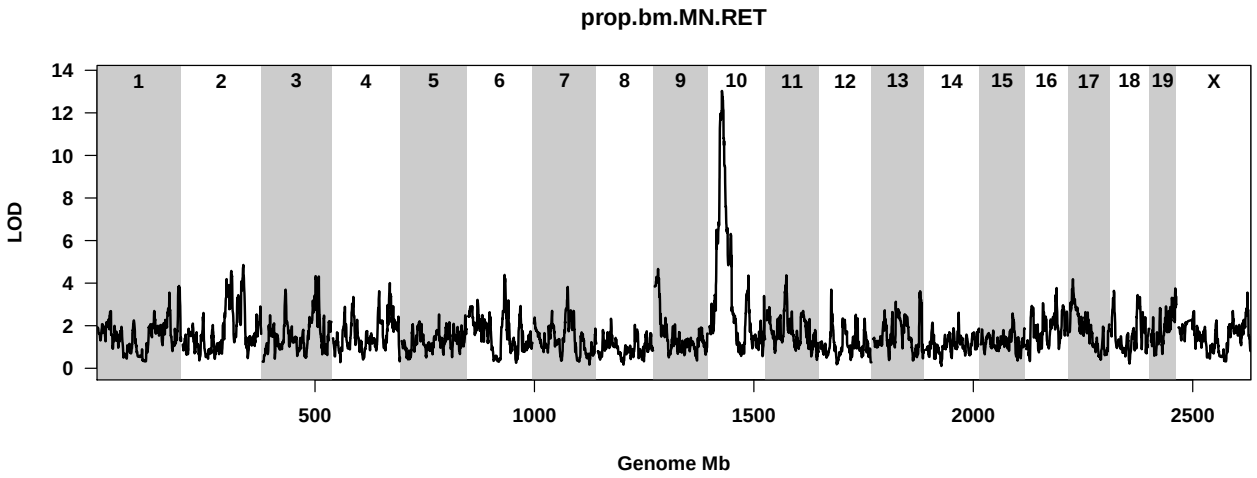
> The biological question asks: 'Variations in what region(s) of the genome is/are correlated with phenotype as measured by the number of micro-nucleated reticulocytes following benzene treatment?'
>
> The null hypothesis is that all regions in the genome vary independently from the variation seen in the phenotype.

The alternate hypothesis is that some of the variation in some of the regions in the genome are correlated with the phenotype.

```
qtl = scanone(pheno = pheno, pheno.col = "prop.bm.MN.RET", probs = founder.probs,
              K = K, addcovar = addcovar, snps = muga_snps)
```

```
[1] "prop.bm.MN.RET"
```

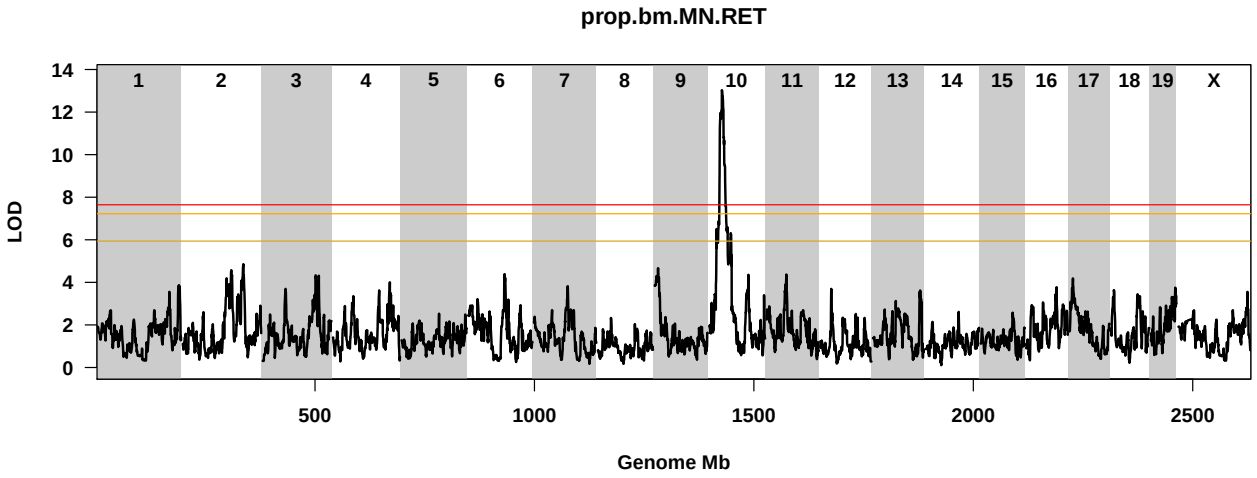
```
qtl.plot(qtl, main = "prop.bm.MN.RET")
```



Q2: Do you see evidence of a peak? If so, where? Save the plot for any results. Provide the figure and an appropriate figure-legend in your write-up.

A peak with a LOD score > 12 is present on chromosome 10.

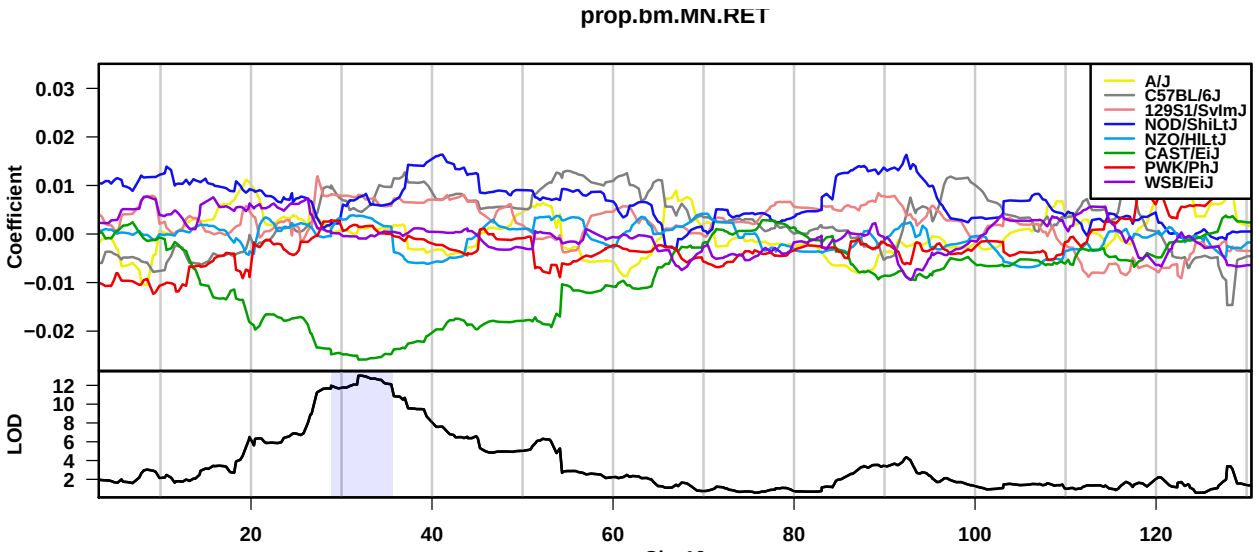
```
perms = scanone.perm(pheno = pheno, pheno.col = "prop.bm.MN.RET", probs = founder.probs,
    K = K, addcovar = addcovar, snps = muga_snps, nperm = 100)
thr = quantile(perms, c(0.95, 0.9, 0.37))
qtl.plot(qtl, sig.thr = thr, sig.col = c("red", "orange", "goldenrod"), main = "prop.bm.MN.RET")
```



Q3. Review and save the QTL plot with thresholds. Provide your assessment regarding statistical significance. Provide the figure and an appropriate figure-legend in your write-up.

The peak appears highly statistically significant and
shows typical QTL-characteristic jags on ramp supporting a biological underpinning for this result.

```
coef.plot(qtl, chr = 10, main = "prop.bm.MN.RET")
```



Q4. Do you see any strain effects? For example, do you see that any DO mice containing a particular strain allele have ou lower or higher levels of micro-nucleated reticulocytes? What does this mean? Provide the figure and an appropriate figure-legend in your write-up.

The CAST/EiJ strain allele appears to have much lower levels (<-.02) of micro-nucleation than the others. This may indicate our QTL applies to genotypes with CAST/EiJ only and not the others.

```
interval = bayesint(qtl$lod$A, chr = 10)
interval
```

```
           SNP_ID Chr Mb_NCBI38       cM          perc.var
1            <NA>  10  28.84078 11.78353              <NA>
4255 UNC100156403  10  32.24779 17.88470 0.342699171600896
3            <NA>  10  35.66660 11.37498              <NA>
                  lrs      lod
1                  10 28.84078
4255 60.0047281272647 13.02986
3                  10 35.66659
```

Q5. Provide the interval results in your write-up. How large is the support interval? Where is the proximal and distal end of the peak? What is the maximum LOD score and its location?

```
support_interval_Mb = 35.66659 - 28.84078
support_interval_cM = 11.78353 - 11.37498
support_interval_Mb
```

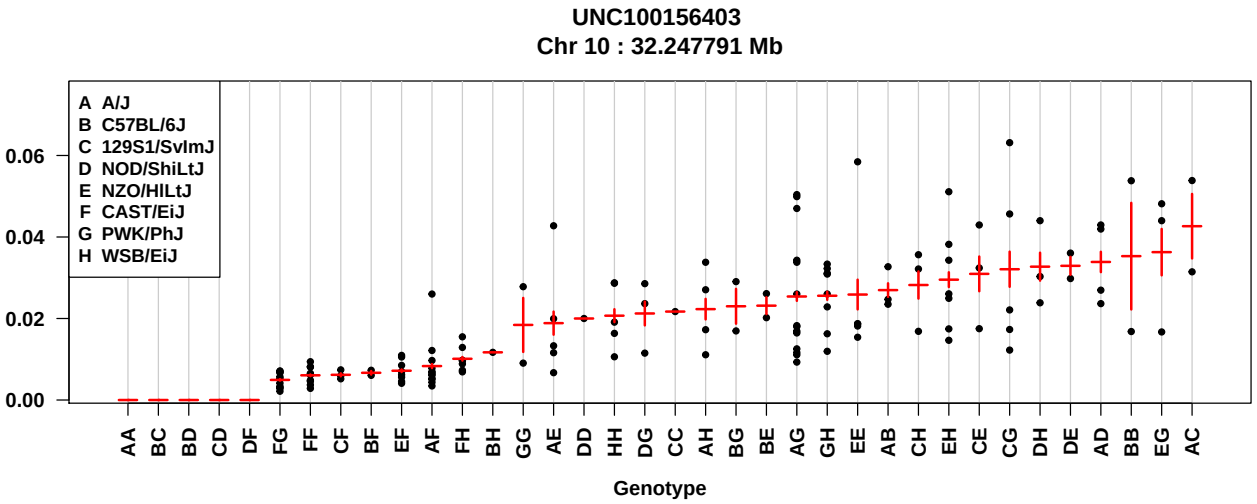
```
[1] 6.82581
```

```
support_interval_cM
```

```
[1] 0.40855
```

The support interval is 6.82581 Mb (0.40855 cM). The peak is centered at 32.24779 Mb with a maximum lod score of 13.02986. 28.84078 Mb is the proximal end of the peak and 35.66660 Mb is the distal end.

```
effect.plot(pheno = pheno, col = "prop.bm.MN.RET", founder.probs = founder.probs,
    snp.id = interval[2, 1], snps = muga_snps)
```



Q6. Save the plot and include it and the appropriate figure legend in your write-up. This plot shows the phenotype value on the Y-axis plotted against the 36 DO genotypes on the X-axis. Do all genotypes occur in this set of 143 samples? Do you see any patterns or trends?

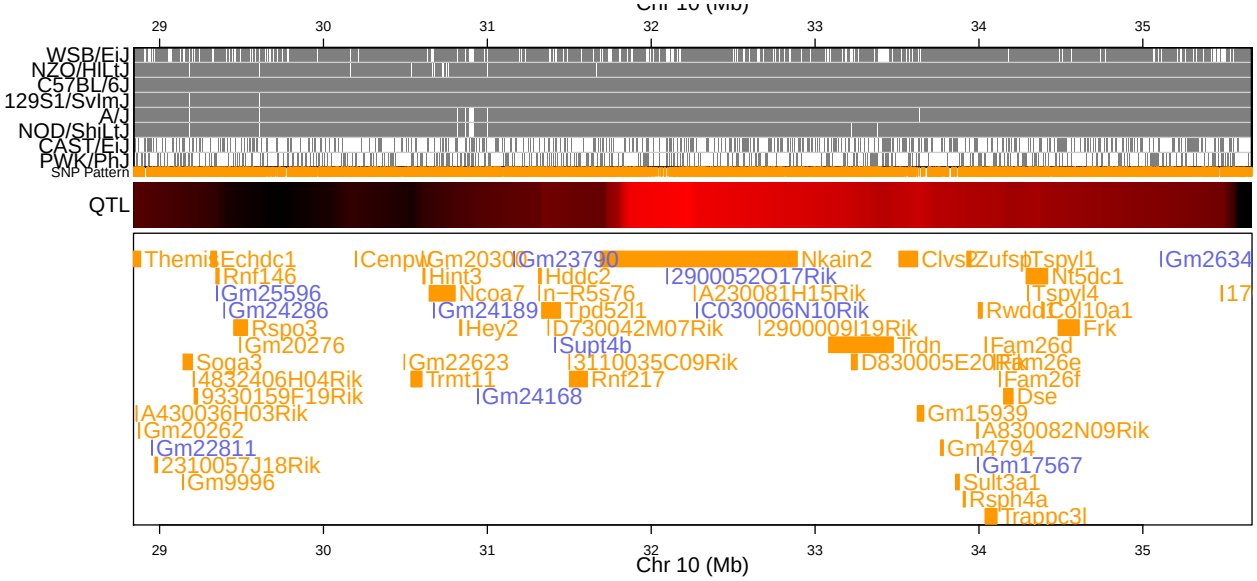
All 36 (8 choose 2 + 8) genotypes are plotted though some don't show values. Perhaps these genotypes are more difficult to cheive (e.g. sterile/degenerate/fatal/etc.). A stepwise pattern appears to model the data roughly. The first five genotypes show no phenotype value, the next eight show low phenotypic values, the next nine show higher phenotypic values, etc.

Searching for Candidate Genes

```
setwd("/Users/joshuaburkhart/SoftwareProjects/ComputationalGenetics/FinalProject")
cand.snps = variant.plot(var.file = "cc.snps.NCBI38.txt.gz", mgi.file = "MGI.20130703.sorted.txt.gz",
    chr = 10, start = interval[1, 3], end = interval[3, 3], pattern = "CAST/EiJ",
    qtl = qtl$lod$A[, c(2, 3, 7)])
```

```
[1] "Checking arguments..."
[1] "Retrieving Variants..."
[1] "105622 Variants in region."
[1] "Getting SNPs that match the pattern..."
[1] "27162 variants fit the pattern."
[1] "Categorizing pattern variants..."
```

```
[1] "53 genes have variants that fit the pattern between the 3'UTR and the 5'UTR (including introns)."
[1] "Getting genes in region..."
[1] "60 genes in region."
[1] "Drawing plot..."
```



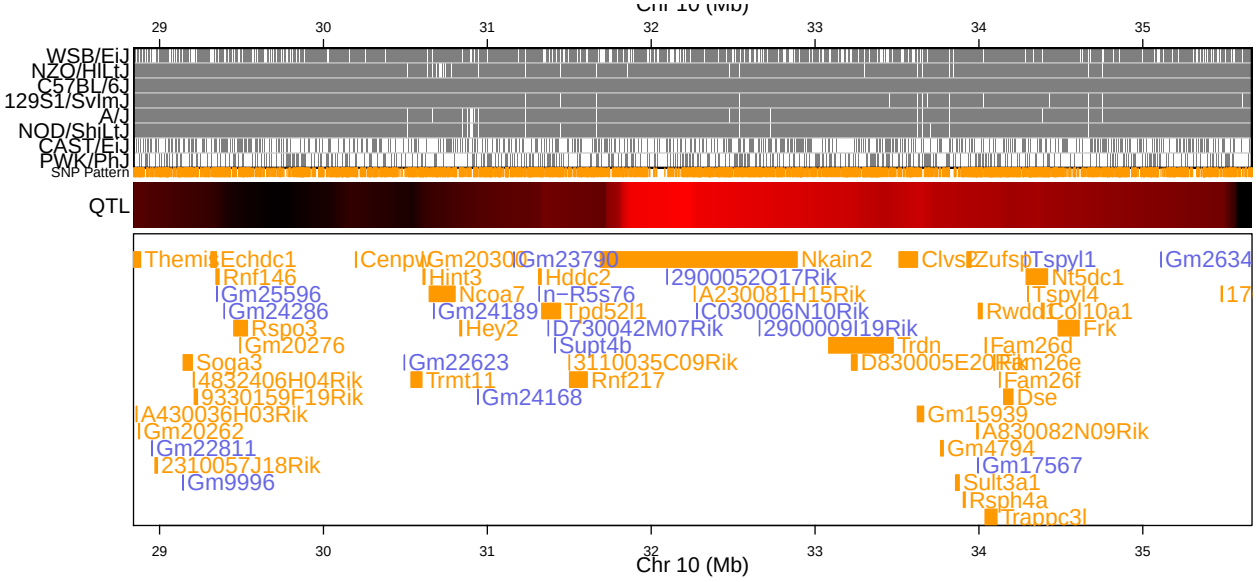
In this plot, the Sanger SNPs are plotted in the top panel with the reference allele in gray and the alternate allele in white. The SNPs that match the requested pattern are plotted as short orange bars beneath this. The LOD score is plotted as a heat map below this with black scores low and red scores high. Finally, the bottom panel shows the genes in the interval, with those containing SNPs that match the pattern plotted in orange.

Q7. How many genes (or non-coding RNAs) are in the interval? Do you see any evidence of private alleles? In some cases, this plot will narrow the support interval significantly. Include this plot and the appropriate figure legend in your write-up.

> 60 genes are reported in the region, 53 reported to fit the pattern.
>
> The peak appears to center over 4 genes: 2900052O17Rik, C0300006N10Rik, A230081H15Rik,and Nkain2. The latter two contain SNPs that match our supplied pattern (CAST/EiJ). It's possible a strain-specific allele of one of these genes is present.

```
cand.snps = variant.plot(var.file = "cc.indels.NCBI38.txt.gz", mgi.file = "MGI.20130703.sorted.txt.gz",
```

```
[1] "Checking arguments..."
[1] "Retrieving Variants..."
[1] "15604 Variants in region."
[1] "Getting SNPs that match the pattern..."
[1] "3943 variants fit the pattern."
[1] "Categorizing pattern variants..."
[1] "44 genes have variants that fit the pattern between the 3'UTR and the 5'UTR (including introns)."
[1] "Getting genes in region..."
[1] "60 genes in region."
[1] "Drawing plot..."
```

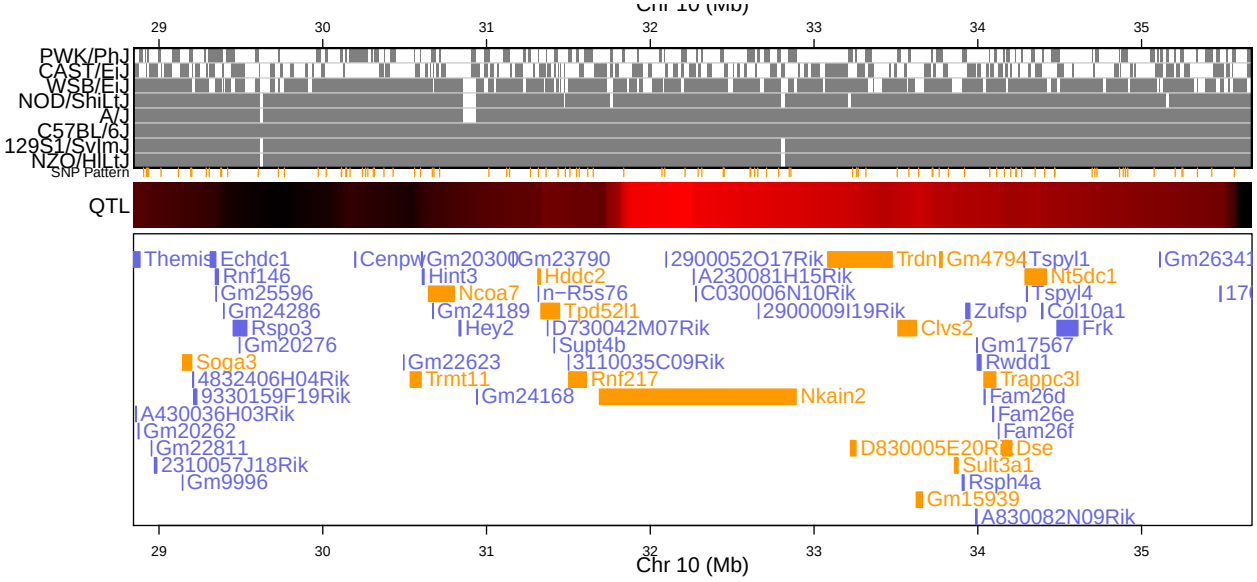


Q8. Can you narrow the candidate gene list from this based on the indels? Why or why not? Include this plot and the appropriate figure legend in your write-up.

This plot provides similar information to the SNPs plot above, indicating the peak is centered over 2900052O17Rik, C0300006N10Rik, A230081H15Rik,and Nkain2. Though perhaps validating our previous SNP assessment, it doesn't indicate we can further narrow our candidate list.

```
cand.snps = variant.plot(var.file = "cc.sv.NCBI38.txt.gz", mgi.file = "MGI.20130703.sorted.txt.gz",
    chr = 10, start = interval[1, 3], end = interval[3, 3], type = "sv", pattern = "CAST/EiJ",
    qtl = qtl$lod$A[, c(2, 3, 7)])
```

```
[1] "Checking arguments..."
[1] "Retrieving Variants..."
[1] "451 Variants in region."
[1] "Getting SNPs that match the pattern..."
[1] "105 variants fit the pattern."
[1] "Categorizing pattern variants..."
[1] "15 genes have variants that fit the pattern between the 3'UTR and the 5'UTR (including introns)."
[1] "Getting genes in region..."
[1] "60 genes in region."
[1] "Drawing plot..."
```



Q9: What other strategies could you use to narrow this region? Hint: think public data types and other types of data and studies. Provide the next steps that should be taken and any limitations of this study in your write-up.

Literature Review Previous RNA-Seq experiments? Previous CNV experiments? Putative Protein Product simulation New RNA-Seq experiment Knockout/Rescue experiments