A QTL Mapping Study

Mark Klick

December 19, 2015

1 Introduction

Historically, identifying the part of the genome responsible for modulating the variation in a complex trait has not been very successful.[1] However, the development of high-throughput genotyping arrays has greatly improved our ability to identify a putative list of genes that significantly contribute to phenotypic variation. The biological question concerning the number, location, and behavior of possible genes that influence a quantitative trait (QT) translates easily into a statistical question. The key to identifying a quantitative trait locus (QTL), is using statistical methods to describe the correlation between the complex trait and genotyping array marker data. Typically, the null hypothesis assumes that for all the markers within the array, QT values are independent of the genotype.[1] The Bayesian statistical approach used in this study, which will be explained further in the methods section, assumes no QTLs exist as the null hypothesis.[2]

Our study measures the level of a QT, micronucleated reticulocytes (MN), in response to prolonged benzene exposure in 141 mice. The MN level is an indicator of how much DNA damage has occurred in the bone marrow. The population of mice are a result of inbred line crosses of eight founder strains. Our goal is to map the genes that influence the level of DNA damage in the bone marrow using computational QTL mapping methods.[2]

2 Methods

Several methods for mapping QTL are available such as single marker mapping or interval mapping.[1] We employ an interval mapping method developed by Churchill and Sen that specifically addresses previous issues with QTL mapping methods e.g. covariates, non-normal trait distributions, epistatic QTL and the issues of multiple simultaneous searches.[2] The method allows for previous issues with QTL studies to be addressed by separating the QTL mapping problem into two parts that are statistically independent: the genotypic relationship part and the linkage part.[2] The genetic model describes the relationship between the QTL and the QT. Churchill and Sen, assume that the proportion of explained phenotypic variance by a marker makes it more likely to be close to a

QTL.[2] Thus they impute psuedomarkers and regress each imputed genotype against the phenotype to achieve an average LOD score over the pseudomarkers. The location of the QTL relies on the log10 of the posterior distribution of the QTL locations, given us the Bayesian credible interval.[2] More details regarding their approach can be found in their manuscript, the gestalt of their method, is that it provides us with the location of an accurate genomic interval that contains the most likely causative markers of the MN phenotype.[2]

The R package DOQTL serves as the host for the required pieces of the analysis. DOQTL provides functions to locate QTLs and calculate Bayesian support intervals. To identify QTL peaks, the function scanone takes into account relevant information like kinship and the markers within the MUGA genotyping array that was used to genotype the mice. There is a debate between using permutation or bootstrapping for re-sampling of the data to assess the statistical significance of identified peaks.[1] We use permutation to assess the statistical significance of these peaks because permutation retains the summary information of the QT, whereas bootstrapping does not.[1] However, the family structure of mice must be carefully considered when applying permutation to avoid type I errors.[3] In our analysis, we perform one thousand permutations of scanones results with p-values of 0.5,0.1,and 0.63.

A QTL peak that passes this significance threshold is then assessed for allelic contribution of the founder strains. The chromosome that contains the significant peak may have a different level of the MN phenotype for each combo of genotypes consisting of founder strain alleles. A support interval for that chromosome is obtained using the Bayesian Credible Interval method explained earlier.[2] DOQTL provides a function bayesint that calculates this support interval. Given the support interval, we can again look at the distribution of the MN phenotype across all the genotypes at the peak. This interval represents the region containing causative SNPs, so any genotype that has a high or low QT value is of interest to us. One allele may have a significant effect on the QT value and therefore should be investigated further.

Given an identified allele that has a significant effect on the QT, we search within our support interval to check for SNPs,indels, and structural variations for that allele. By doing this, we hope to narrow the list of candidate genes within the previously determined support interval. All of the annotation was obtained from the Jackson Lab ftp server. The resulting QTL mapping analysis of the MN phenotype using our method is described below.

3 Results

A statistically significant QTL peak (fig.1) that passed our 1000 permutation test threshold was found on chromosome 10. A closer look at the peak (fig.2) reveals that the CAST/EiJ allele seems to have a strain effect due to its significantly lower model coefficient. An allele with a significantly higher model coefficient, implies that there is more variation in the QT attributed to that allele. A low model coefficient value means that there is less variation seen in the QT value for that allele.

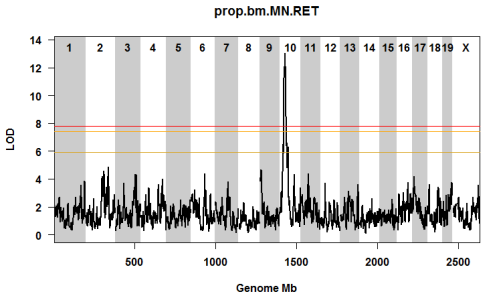

Figure 1: The QTL peak positions are shown with chromosome position on the x-axis and LOD score on the y-axis. Significance thresholds of 0.95, 0.9 , and 0.37 resulting from permutation tests are shown as horizontal lines.

The computed Bayesian support interval for the significant QTL peak on chromosome 10 can be seen in (fig.2). Its important to keep in mind that within this support interval resides the most likely causative SNPs found at that locus. This interval is essential to narrowing in on the most likely causative genes within this area. The support interval spanned approximately 6.8mb. The proximal and distal ends of the peak are located at 28.84078mb and 35.66660mb respectively. The maximum LOD score at this locus is 35.6 and occurs at the distal end of the peak.

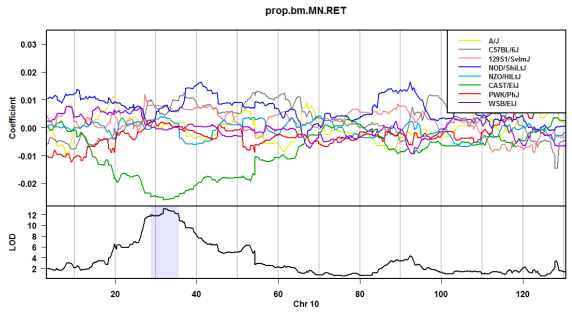

Figure 2: Zoomed in significant max QTL peak on chromosome 10 is shown. The top panel show the founder allele effects along the chromosome with the model coefficient on the y-axis. The computed Bayesian support interval is shown in blue. A strain effect of CAST/EiJ can be observed.

Taking into account what we learned from (fig.2), i.e. the CAST/EiJ allele had the largest strain effect, we now want to look at the MN values associated with each possible combination of alleles. We can see from (fig.3) that there are several genotypes that do not occur in this set of 141 samples. We can also see that each genotype containing the CASt/EiJ allele has roughly the same MN value distribution. Based on what we saw in (fig.2) and (fig.3),

we believe is the CAST/EiJ strain is most different from the others because there is a consistent low MN value. A consistent low MN value for all genotypes containing the CAST/EiJ, suggests that at this QTL peak there may be SNPs or genes that influence/regulate a resistance to DNA damage in the bone marrow caused by benzene. Therefore we searched for candidate markers and genes within the support interval specific to the CAST/EiJ strain.

Searching the support interval on chromosome 10 in finer detail requires the SNP file and the gene location file, both files contain the information necessary to identify candidate SNPs specific to CAST/EiJ. We can see that in (fig.4) there are 53 genes and noncoding RNA that correspond to CAST/EiJ found within this interval. There seems to be convincing evidence of private alleles specific to CAST/EiJ seen by the amount of white alternate alleles that exist within the support interval. CAST/EiJ shows a high level of genetic differentiation at this QTL peak which suggests the CAST/EiJ SNPs with the highest LOD score have a very high chance of modulating the previously observed resistance to DNA damage.

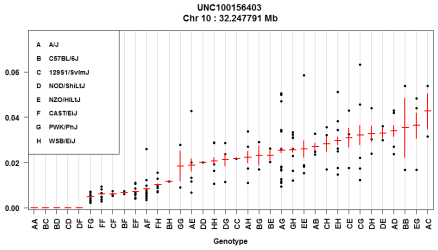

Figure 3: The 36 possible genotypes are shown on the x-axis and the phenotype value is on the y-axis. All genotypes containing the CAST/EiJ allele have a very consistent low phenotypic value.

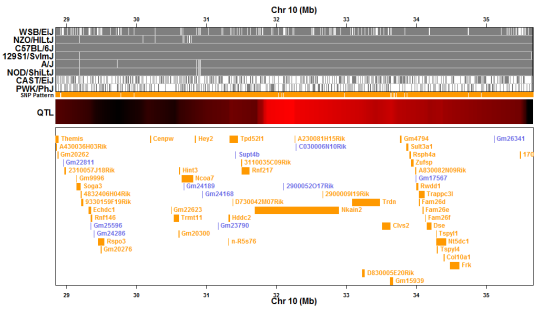

Figure 4: Sanger SNPs are plotted in the top panel with the reference allele in gray and the alternate allele in white. The SNPs for CAST/EiJ are shown as short orange bars beneath this. A QTL LOD score heatmap showing red values as high LOD scores and black values for a low LOD score. The bottom panel shows the genes in the support interval that contain CAST/EiJ SNPs. A list of 53 genes is obtained from this plot.

In hopes of narrowing the candidate gene list the indels and structural variants at this QTL pertaining to CAST/EiJ were looked at. Based on the indels located within the QTL, the gene list was narrowed from 53 to 44 genes in the interval. This technique narrowed the candidate gene list by about ten genes. Based on the structural variations located within the QT, the gene list was narrowed to 15 genes. However, there are often few annotations for structural variants, so this method can artificially narrow the candidate gene list.

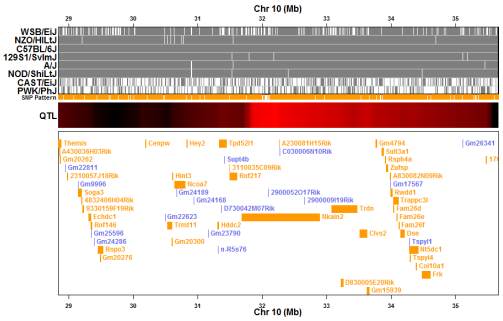

Figure 5: The bottom panel shows the genes in the support interval that contain CAST/EiJ indels. A list of 44 genes is obtained from this plot.

4 Discussion

We now have a list of 53 candidate causal genes that may influence a resistance to DNA damage via benzene exposure. Using annotations regarding indels associated with the CAST/EiJ strain, the candidate gene list was narrowed to 44 genes. This study aimed at mapping the QTL of the MN QT and we accomplished our goal. A manageable list of genes that potentially influence the level of DNA damage in the bone marrow has been acquired. This list of candidate genes is crucial for the next steps of this experiment.

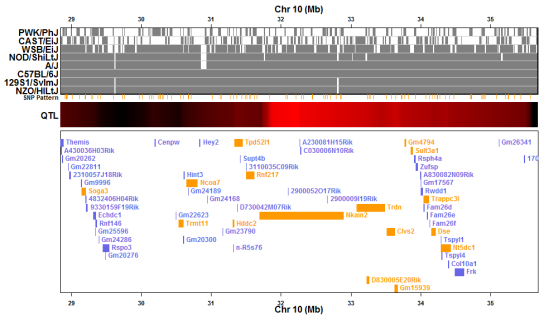

Figure 6: The bottom panel shows the genes in the support interval that contain CAST/EiJ structural variants (SV). A list of 15 genes is obtained from this plot but due to lack of known SVs this can artificially shorten the list of genes.

A few possible next steps for this study could be to perform an in-depth literature review of the list of genes to elucidate how these genes may be interacting. We could do a regulatory gene network analysis to try and elucidate the underlying mechanisms of the genes and how they relate to the low MN phenotype i.e. DNA damage resistance. Finally, we could use this list of genes for potential targets for gene knockout experiment. This can be done by knocking these genes out and observing the changes in the MN level.

Some limitations in our study are similar to the limitations inherent to QTL mapping studies described in [2]. First, the number of samples will affect the

total number of detectable QTL. If our study contained more than 141 samples then our Bayesian support interval might have been more refined. Another possible limitation in our study is the fact that the localization of QTL is directly influenced by the number of crossover events between alleles in the cross. [2] Despite these limitations, we feel like we have uncovered a putative list of causal genes that influence the amount of DNA damage that occurs when a mouse is exposed to the toxin benzene.

Wed like to thank Shannon McWeeny for being a great instructor and providing us with the experimental outline, data, and the DOQTL R package.

References

[1] R. W. Doerge. *Mapping and analysis of quantitative trait loci in experimental populations* Nature Reviews Genetics vol.3 (JANUARY 2002).

[2] S. Sen. G. A. Churchill. *A statistical framework for quantitative trait mapping* Genetics 159: 371387 (September 2001).

[3] G. A. Churchill. R. W. Doerge. *Naive application of permutation leads to type I inflated error rates* Genetics 178: 609-610 (January 2008).