ABSTRACT

GILLESPIE, KEVIN. Systems Genetics of Lung Response in the Collaborative Cross: A Comparative Analysis. (Under the direction of David Aylor.)

We compared Expression Quantitative Trait Loci (eQTL) from two studies of the same incipient lines from a recombinant inbred line (RIL) population known as the Collaborative Cross (CC). In these previous experiments mice were exposed to either House Dust Mite (HDM) allergen or Influenza A (IAV) PR8. After a period of four days, bronchoalveolar lavage (BAL) RNA was obtained and assayed for gene expression. We sought to perform eQTL mapping, and then compared the results in order to investigate the extent to which eQTL colocalized to the same intervals, or different intervals between studies. We discovered collections of genes all under the control of a single genomic region, termed trans-bands, that were unique to either experiment. These regions represent areas of the genome where regulators of many downstream responses exist. We use functional annotation to ascertain what types of genes are downstream of these trans-bands. We found 2,274 genes, of the 9,612 genes that were common to both studies, were associated with eQTL intervals common to both studies. The eQTL interval upstream of a shared gene could either represent a single eQTL for both experiments or multiple eQTL experiencing context dependent effects within a single region. We used the estimated allele effects for the 8 founder strains of the CC throughout the eQTL interval to further exclude colocalized regions where the compared effects did not agree between studies. If these patterns don't match we make the case for multiple loci. Where each locus then may influence expression of the downstream gene. By doing this, we demonstrate that while a significant number of eQTL detected in both studies overlap in their constructed intervals, the actual number of shared eQTL positions are likely much fewer between the two studies based on the underlying allele effects at these intervals. These observations are important to the assessment of the true extent to which eQTL are shared between experiments. Assessing the shared eQTL also has implications on which common variants influence multiple common diseases. We also present the first full comparison of two experiments conducted in the same CC RIL mice.
Systems Genetics of Lung Response in the Collaborative Cross: A Comparative Analysis

by
Kevin Gillespie

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Functional Genomics

Raleigh, North Carolina
2016

APPROVED BY:

_________________________  _______________________
Dahlia Nielsen               Zhao-Bang Zeng

_________________________
David Aylor
Chair of Advisory Committee
DEDICATION

To my family and those friends that have lifted me up.
BIOGRAPHY

The author was born in Macon, GA and received his B.S. in genetics at the University of Georgia in 2011.
ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. David Aylor, for his help and guidance, as well as my committee, Dr. Dahlia Nielsen and Dr. Zhao-Bang Zeng, for their valuable input. I'd also like to thank Natalie Robles for continually being available in moments of weakness, and pressing me to persevere.
TABLE OF CONTENTS

LIST OF TABLES .................................................................................. vii

LIST OF FIGURES .............................................................................. viii

Chapter 1  INTRODUCTION .................................................................. 1
  1.1 Comparing eQTL studies ............................................................... 1
  1.2 Collaborative Cross ....................................................................... 2
  1.3 Attractive features of the population ............................................. 3
  1.4 AAD Previous Work ...................................................................... 3
  1.5 FLU Previous Work ...................................................................... 4
  1.6 Approach ..................................................................................... 5
  1.7 Specific Goals .............................................................................. 6

Chapter 2  MATERIALS & METHODS .................................................... 8
  2.1 Expression Arrays ......................................................................... 8
  2.2 Genotyping Arrays ....................................................................... 9
  2.3 eQTL Mapping ............................................................................. 9
  2.4 Allele Effect Patterns ................................................................... 10
  2.5 Gene Set Analysis ........................................................................ 10

Chapter 3  RESULTS ........................................................................... 12
  3.1 Expressed Genes .......................................................................... 12
  3.2 AAD eQTL .................................................................................. 13
  3.3 FLU eQTL .................................................................................. 13
  3.4 Shared eQTL ................................................................................ 17
  3.5 Allele Effects Comparisons ............................................................ 18

Chapter 4  DISCUSSION ..................................................................... 21
  4.1 Functional enrichment within gene sets ......................................... 21
  4.2 Trans-eQTL Bands ...................................................................... 22
  4.3 Shared eQTL after AEP comparison .............................................. 23
  4.4 Extent of eQTL sharing between these studies ............................... 24

BIBLIOGRAPHY ............................................................................... 26

APPENDICES ..................................................................................... 29
  Appendix A  QTL Mapping ................................................................. 30
    A.1 Hematology ............................................................................. 30
    A.2 Peak conditioning ................................................................... 30
    A.3 Conditioning for Sample Batches .............................................. 31
    A.4 Removing Outliers .................................................................. 31
    A.5 Manhattan plots ..................................................................... 31
    A.6 Results ................................................................................... 31
<table>
<thead>
<tr>
<th>Appendix B</th>
<th>Insufficient samples lead to overfit data</th>
<th>33</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.1</td>
<td>Uncharacteristic AEP</td>
<td>33</td>
</tr>
<tr>
<td>B.2</td>
<td>Pseudo counts</td>
<td>35</td>
</tr>
<tr>
<td>Appendix C</td>
<td>Differential expression</td>
<td>36</td>
</tr>
<tr>
<td>C.1</td>
<td>Response Groups</td>
<td>36</td>
</tr>
<tr>
<td>C.2</td>
<td>Differentially expressed genes</td>
<td>36</td>
</tr>
<tr>
<td>C.3</td>
<td>Results</td>
<td>37</td>
</tr>
<tr>
<td>Appendix D</td>
<td>Conditional correlation analysis</td>
<td>38</td>
</tr>
<tr>
<td>D.1</td>
<td>FLU trans-eQTL</td>
<td>38</td>
</tr>
<tr>
<td>D.2</td>
<td>Conditioning on Mx1 locus</td>
<td>40</td>
</tr>
<tr>
<td>D.3</td>
<td>Results</td>
<td>40</td>
</tr>
<tr>
<td>Appendix E</td>
<td>Modulated modularity clustering</td>
<td>41</td>
</tr>
<tr>
<td>E.1</td>
<td>Modulated Modularity Clustering</td>
<td>41</td>
</tr>
<tr>
<td>E.2</td>
<td>AAD Example</td>
<td>42</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 3.1  Functional Annotation Mx1  ......................................................... 15
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>Analysis Process</td>
<td>7</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Global plot for eQTL in both AAD and FLU studies</td>
<td>14</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Partitioning of genes from individual experiments through analysis</td>
<td>16</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Example of matching AEP</td>
<td>18</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Example of inverted AEP</td>
<td>19</td>
</tr>
<tr>
<td>Figure A.1</td>
<td>Manhattan plot of log(RBC Diameter Width)</td>
<td>32</td>
</tr>
<tr>
<td>Figure B.1</td>
<td>Addition of simulated PWK mice</td>
<td>34</td>
</tr>
<tr>
<td>Figure D.1</td>
<td>Conditional Correlation scan of gene downstream of Mx1</td>
<td>39</td>
</tr>
<tr>
<td>Figure E.1</td>
<td>MMC of AAD differential expression</td>
<td>42</td>
</tr>
<tr>
<td>Figure E.2</td>
<td>MMC of AAD eQTL</td>
<td>43</td>
</tr>
</tbody>
</table>
We utilized identical methods to perform expression Quantitative Trait Locus (eQTL) mapping in two studies of mice from the Collaborative Cross (CC) that were exposed to either Influenza A (IAV) or House Dust Mite (HDM). The eQTL detected in these two studies were compared, and their intervals and the associated gene were used to measure the number of eQTL that were unique to an experimental condition or shared between them. We performed functional annotation on the downstream genes associated with eQTL specific to one of three major groups, the Associated Airway Disease experiment (AAD), the Influenza experiment (FLU), and genes downstream of eQTL that had the same interval in both experiments.

1.1 Comparing eQTL studies

Expression QTL mapping uses gene expression as a quantitative trait and attempts to correlate variation in expression with variation in genotype. [1, 2] The results are expression quantitative trait loci, which are simply regions of the genome containing variants that influence the expression levels of a gene. Many eQTL mapping studies have uncovered relationships between regions of the genome and genes involved in a number of biological processes such as plant yield, insect behavior, plant morphology, and complex diseases. [3–5] Here, we are interested in exploring the extent to
1.2. COLLABORATIVE CROSS

which eQTL are shared between lung responses to different experimental conditions. Attempts to compare multiple sets of eQTL results with one another are more rare. Studies have attempted to detect regulatory regions that are shared between immune responses triggered from different environmental insults, but have been limited to comparisons between cell types. [6–9] Studies have also shown that 50% of the cis-eQTL between two different tissues can be shared. [10] Studies have also attempted to reconcile shared associations from multiple diseases as well. In the area of Irritable Bowel Diseases (IBD) work analyzing shared regulatory modules of Crohn's disease and ulcerative colitis have uncovered new modules of shared susceptibility. [11] The Genotype-Tissue Expression (GTEx) project was started through the National Institute of Health (NIH) common fund, and is an example of a resource — currently in development — that attempts to catalog these relationships between gene expression variation and genetic variation in humans. The project uses tissues gathered from deceased donors to obtain gene expression and genotype information, which is used to perform eQTL analysis in each tissue type collected. With this information, and the medical records of the donors, this resource allows us to assess many of the common variants associated with changes in gene expression that could influence susceptibility to many complex human diseases. [12] These types comparisons are interesting when considering complex diseases with similar etiologies. The assumption is that if the pathology of diseases are similar, the regulatory architecture may be as well. These studies illustrate the existence of common variants within a single cell type between environmental insults, and that there are shared modules of susceptibility between disease pathologies. There is, however, less information as to the extent to which regulatory variants are shared between disease pathologies. Here we describe one such comparison of eQTL detected between mice of incipient lines of a RIL panel, the Collaborative Cross (CC).

1.2 Collaborative Cross

We selected two specific studies in order to minimize the ways that design, subject and treatment may confound the results. The two studies analyzed here are in mice of the incipient lines to the CC RIL population. These CC mice are not fully inbred, being on average 6.7 generations into the inbreeding process they display an average homozygosity of 84.1% across their genomes. [3, 13, 14] The advantages of this population are similar to those found in the Collaborative Cross.

The CC itself is a population of mice generated from crossing eight founder strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HILtJ, and three wild derived strains CAST/EiJ, PWK/PhJ, and WSB/EiJ) crossed so that the resulting progeny equally represent all 8 of these founders. [3, 13–15] The panel was designed with a number of key factors in mind. It is reproducible, because in a RIL population individuals from generation to generation are genetically identical. [3, 15] This makes it
replicative as well, and allows the same genotype to be tested against multiple experimental conditions multiple times. [3, 15] This is of particular interest to our specific goal of making comparisons between eQTL experiments. It is diverse because, as mentioned above, it uses lines from all three major subspecies of mice, but they are also equal in that the lines are all made of roughly the same number of genetic regions from each founder strain. [3, 15] These qualities make this panel an ideal population to investigate complex traits with low heritability, such as disease susceptibility, and were an important factor in selecting the two studies to be described below.

1.3 Attractive features of the population

The two studies, the Associated Airway Disease (AAD) arm and the Influenza A (FLU) arm were selected from the four arms of an initial set of studies in the CC, because they overcome some of the limitations of comparative analysis. [16] The mice in each study were genetically similar and used the same tissue, Bronchoalveolar Lavage (BAL), for RNA sampling. [3, 17–20] While the analysis methods of these studies differed, here, they will be re-analyzed in an identical fashion to minimize technical noise. [16] For these reasons a comparative analysis between these two sets of data should be effective in allowing us to address the extent of shared eQTL between these experimental lung insults. It will also allow us to better understand the suites of genes that define AAD and IAV immune responses.

1.4 AAD Previous Work

Both datasets utilized in this comparative analysis were previously analyzed independently of one another. In the AAD study the investigators sought to characterize allergic responses in pre-CC mice post exposure to house dust mite allergen. To this end they performed a number of analyses profiling: QTL, Quantitative Trait Transcripts (QTTs) and eQTL. They found the strongest phenotypic responses when assessing eosinophil count, an important cell type indicator of allergic response, and \( il5 \) levels. [19] They performed QTL mapping and focused the rest of the analysis on a QTL detected for eosinophil count, as an indicator of AAD. To analyze QTTs highly correlated with eosinophil count they performed linear regression using the expression level of a transcript as an independent variable. In this way, they uncovered 2,852 genes that were associated with eosinophil count. Many of these genes were well known cytokines associated with AAD. Finally, results from the eQTL mapping analysis discovered some 6,457 eQTL associated with about 4,556 genes. Most of these eQTL, 73.7\% (4,759 eQTL), were very near the gene they were associated with. They also illustrated an intuitive relationship between the proximity of an eQTL and its associated LOD score.
This same pattern has been observed in all of the analysis here as well. This study also observed two regions on chromosome 2 where there were clusters of eQTL that were very far away from the gene they influenced. These regions where expression patterns for many genes are all associated with the same genomic locus are referred to as trans-bands. Within the trans-band centered at 147.6Mb were genes enriched for genes found in the QTT analysis and therefore important to eosinophil count. Additionally, they found that two of the founder strains (A/J and C57BL/6J) were nearly identical at alleles across the trans-band region. [19]

Finally, this analysis sought to characterize any eQTL within a previously defined QTL on chromosome 11 for eosinophil count. They used conditional correlation analysis to assess which eQTL might be candidates responsible for the eosinophil QTL. A candidate eQTL should reduce the LOD score of the target QTL when variation at the eQTL locus is controlled for. The gene that contributed the greatest LOD drop was Tlcd2, which is negatively correlated with eosinophil count. Therefore, it was considered a candidate negative regulator of eosinophil recruitment. [19] A second study of the AAD experiment a conditional correlation test of probes within a previously observed QTL for CXCL1 on Chr 7 (Dpc1) concluded that a variant in the gene, Zfp30, within this region is the eQTL underlying Dpc1. [20]

1.5 FLU Previous Work

The FLU studies sought to characterize the pre-CC mice responses to exposure to Influenza A PR8. Similarly to the AAD studies they profiled a number of phenotypes and gene expression. They performed differential expression analysis between subsets of the most extreme responding mice in the experiment. These mice were either in a group with high day 4 weight loss and immunohistochemical staining or had low scores for these measures. They chose the top up regulated and down regulated genes to perform eQTL analysis. From these 2,671 transcripts they discovered 21 that had eQTL associated. The group went on to experimentally verify expression changes driven by 17 of these eQTL via qPCR. [17] This small set largely agreed in expression levels with the observed allele effects.

A second set of studies utilized weighted gene coexpression network analysis (WGCNA) with the top 6,000 most variable and interconnected transcripts from the gene expression array. These transcripts were reduced to twelve module eigengenes, which were enriched for ontologies such as cellular signaling and immune response. [18] They performed QTL scans with a number of collected phenotypes, such as day 4 weight loss, log titer, etc., as well as the eigengene modules. They found that most of the phenotypes and the three of the modules overlapped with the HrI1 locus that contains the Mxl gene, a GTPase that is associated with inhibiting Influenza viral replication.
Additionally, they characterized the eight founder genetic backgrounds at *Mx1* gene. The found a specific allele variant, that caused an amino acid substitution (Gly83Arg), corresponding to the CAST/EiJ founder strain that does not confer the viral replication inhibition yet still protects against weight loss. They found that A/J, 129S1/SvImJ and NOD/ShiLtJ all had the same 2kb deletion as the C57BL/6J founder. This haplotype renders *Mx1* non-functional. WSB/EiJ also had a non-functional gene, but this was due to a nonsense mutation within exon 10.

The group then characterized CC variation at another significant QTL, referred to as *HrI2*, located on chromosome 7. They selected 99 mice that had a non-functional *Mx1* gene, based on inferred founder state, and performed WGCNA on the 60 mice RNA extracts from this group that produced usable microarray expression data. This analysis produced 11 modules, where 8 modules were enriched for T-cell and signaling processes as well as inflammatory response. They performed this same process for two other host responses to IAV *HrI* QTL. One QTL contributed control to the variation in pulmonary edema, termed *HrI3*, and was located on chromosome 1. The last QTL detected, *HrI4*, contributed to variation in airway neutrophils on chromosome 15. None of the modules produced by WGCNA were found to associate with these QTL and could not be used to for GO term enrichment.

In order to further reduce the likely interval around significant QTL peaks the investigators set out to correlate QTL allele effect patterns by separating the founders into two groups based on largest distance between groups. They then excluded SNPs where the members of the causative SNP do not agree with the predicted pattern. In this way they reduced the interval for both *HrI3* and *HrI4* dramatically, which reduced the number of causal genes within the region. They used the same process for *HrI2*, but in this example the groups were separated into three groups an increased resistance group (A/J), a susceptibility group (129S1/SvImJ) and all the other founders were considered intermediate. In this case they looked for SNPs that discriminate both A/J and 129S1/SvImJ from the rest of the founders. This allowed them to identify genes that had associated SNPs within the QTL interval for either founder or that were shared between them. [18] This process of excluding markers that don’t agree with the allele effect patterns estimated in the mapping process is a very important feature of the CC that allows us to assess the support of a shared eQTL in the current analysis.

### 1.6 Approach

To make this comparison possible we performed eQTL mapping on both sets of pre-existing data in the exact same fashion. This reduces the amount of technical noise between the two comparisons, which is referred to as heterogeneity. [16] There are many sources of heterogeneity that will be
1.7. SPECIFIC GOALS

important considerations before the analysis is performed. The analysis scheme described below is depicted in Figure 1.1. Briefly, the data from both studies will be curated, to attempt to minimize sources of heterogeneity and prepare them for our specific analysis, eQTL mapping will then be performed and the results will be obtained from each study. We separate the results by their estimated intervals, the likely region where an eQTL effect exists, to separate those that overlap in both studies or are discrete to a single study. Each of the sets of genes, that have eQTL associated, is used as a separate gene set for functional annotation analysis for GO terms and pathway enrichment. This gives us an idea of what types of genes are being influenced by the eQTL detected in either or both studies. The eQTL that have overlapping intervals will have their allele effect patterns compared as well. One of the advantages of the CC population is that QTL mapping is done in a way that allows us to observe the allelic effects of the eight founder strains at each significant eQTL association. These allele effects can then be grouped, based on their effect on the trait, into high or low groups. When these groups and the correlation between the individual effects match it suggests that the two eQTL are potentially a single eQTL, while non-matching patterns suggest the opposite is true, and there are likely multiple eQTL within the overlapping region. Using this as a second filter we can make a better estimate of the true extent of shared eQTL between the two lung responses.

1.7 Specific Goals

This study will be the first full eQTL mapping of the FLU study arm, and by re-analyzing these similar experiments in a similar way we can assess which eQTL detected in either study are discrete or shared. By using the allele effect patterns we can give more or less support to a shared eQTL between studies. With these results we gain more information on what regulatory regions are unique to these insults and which are common to both.
The beginning of this analysis is to perform eQTL analysis. The results are then added to a database and then assessed for genes, eQTL, and allele effects that are similar in both independent studies. The goal of these comparisons is to measure the extents to which eQTL are shared between two similar experiments with differing complex disease foci.

**Figure 1.1 Analysis Process**

The beginning of this analysis is to perform eQTL analysis. The results are then added to a database and then assessed for genes, eQTL, and allele effects that are similar in both independent studies. The goal of these comparisons is to measure the extents to which eQTL are shared between two similar experiments with differing complex disease foci.
2.1 Expression Arrays

The expression array data was graciously provided from the principle investigators of the two previous studies of the pre-CC RIL population. The AAD study assayed 138 male mice BAL RNA extracts on the Illumina WG 6v2.0 expression array. [19, 20] The FLU study assayed 99 male mice BAL RNA extracts on the Agilent MWG 4x44k expression array. [17, 18] Both arrays were pre-processed using the array platform specific software (Genome Studio or AgilentPreProcess). The results were then fed into the limma R package for quality filtering and background normalization. While the arrays contained more than 40k probes each, many of them did not reach expression levels above background in the respective experiments. The results were two matrices of 24,013 and 17,593 log2 normalized probe profiles for 13,437 and 11,529 entrez genes respectively in the ADD and FLU studies. [17–20] The result was a substantial number of probes that were unable to be assayed and so a large number of genes were removed from downstream analyses. More than 70% of these genes were shared between arrays. These overlapping genes constitute the initial population of genes for the rest of the downstream comparisons.
2.2 Genotyping Arrays

The principle investigators also graciously provided the genotype information in the two expression studies. All the mice were assayed on either the A or B version of the Mouse Diversity Array. Both of these arrays contain over 180k SNPs.\cite{3, 14, 17, 19} The AAD study was assayed on both, so the missing markers were imputed between the two arrays.\cite{21} This allele information from the genotyping microarrays was then processed, previously, using the HAPPY R package.\cite{22}

Briefly, the genotype allele information from each study was provided along with a Hidden Markov Model (HMM), containing the probabilities that an allele originated from one of the eight founders, to a forward-backward algorithm that generates a set of 8 emission state probabilities at each marker for each mouse.\cite{22} Regions where the emission state probabilities are essentially constant for all the mice in a study were reduced to representative intervals.\cite{3} This process produced genotype information for 27,039 and 12,100 marker intervals in the AAD and FLU study respectively. All these steps were previously performed in the independent analyses, and the results were here processed into a 3D array that the DOQTL software can interpret.

2.3 eQTL Mapping

To perform eQTL analysis on gene expression data, the expression array data was quantile normalized. A kinship matrix was created by calculating the relatedness of each mouse's genotype to the other mice in the experiment. A covariate matrix was created that included the sex of each mouse. It can hold additional information if it is available and useful, which is described in. (Appendix 4)

The DOQTL R package was used to perform the eQTL scans.\cite{23, 24} A nominal scan was performed to generate nominal LOD score for each marker interval for each probe in the expression array. 5,000 additional permutation scans were performed to generate a significance threshold for the discrimination of significant LOD score associations in the nominal scan. This also facilitated the calculation of adjusted p-values for those significant peaks that are greater than the permuted threshold.

For each significant peak in the nominal scan a 95% Bayesian Credibility Interval (CI) was constructed. This interval represents the region most likely to contain the position contributing control for the quantitative trait. This interval, an associated LOD score, an adjusted p-value and the gene the eQTL was associated with are used as a summary. The top 500 eQTL, judged by having the smallest adjusted p-value and greatest LOD score, were used as gene sets for functional analysis.

The eQTL results were all added to a database in order to assess which genes that were common to both studies also had eQTL that had overlapping credibility intervals. These shared eQTL were
subdivided into those that were very near their associated gene, termed cis-eQTL, and those that were far away from their associated gene, termed trans-eQTL. Both sets of associated genes were analyzed for functional significance.

### 2.4 Allele Effect Patterns

Allele effect patterns (AEP) reflect the estimated effect on expression of having the allele from the corresponding haplotype. We generated plots of these effects across a specific eQTL interval when the estimated intervals overlapped in the two studies. These AEP were compared and separated based on a set of criterion described below. The correlation of the eight allele effects between experiments was used to discriminate those comparisons that were inverted (values near -1) and comparisons that likely match (values near 1). An inverted comparison implies that the same founder strain has opposing effects in each study, while a matching effect implies the two effects synergistic. Those comparisons whose correlation is near 0 are termed ambiguous. These effects are too nebulous for us to obtain quality groupings and so they are censored from this study. Next, silhouette widths were assess for different k-means clustering of the effects ($k = [2-7]$). The $k$ that minimized the interquartile distance and maximized the mean of the silhouette widths was selected as the appropriate cluster number. Silhouette widths measure the cohesion of the cluster compared to other clusters. A high value for each silhouette implies appropriate clustering, whereas many widths that are very low implies an inappropriate $k$ value. The effects are separated this way into two or three groups based on kmeans clustering with ($k = 2$ or $3$). Any more than three groups become difficult to score and were left out of this analysis. This step creates two or three groups of effects in either high and low, or high, intermediate and low groups respectively. If there is sufficient evidence for more than 3 clusters the comparison is considered ambiguous. If the clusters do not agree in the comparison then it is considered a non-match.

The use of allele effects provides a way of excluding overlapping intervals where the effects between studies do not match. It has been illustrated in the previous work in the pre-CC that we can reduce a single interval be excluding those positions that do not agree with the estimated effects. [3, 18] Here we apply the same rational to refute the idea that an eQTL is the same in both studies if the corresponding AEP do not match between them.

### 2.5 Gene Set Analysis

The results of eQTL analysis are positions within the genome that influence the expression of some gene expression target. We can gain more insight into the biological characteristics of an exposure
by making lists of these genes, within an experiment, that are associated with eQTL and performing gene set analysis. All gene set enrichments seeks to find biological terms, GO terms and pathways in this analysis that are over-represented in the provided list of genes. This indicates that the provided list is biased for genes corresponding to that term or pathway. We were carried out all enrichment using DAVID as well as Consensus Path Database (CPDB) from the Max Planck Institute for Molecular Genetics. DAVID was employed for functional annotation clustering and CPDB was utilized for gene set overrepresentation analysis, when results for annotation clustering failed to produce any strong enrichment. Enrichment was generally expected to be greater than 3 or 4 for DAVID enrichment to be considered of decent quality. These enrichment values or greater are generally associated with high FDR corrected P-values as well. [25–27] Annotation acceptance was stringent with the main threshold for significance set at a q-value of at least less than 0.05. Using this value corrects for multiple testing and still indicates that the ontology or pathway is significantly associated. We used functional annotation methods on genes associated with eQTL that were unique to an exposure or shared. We also performed gene set enrichment on subsets of these groups, such as those genes associated with trans-bands and those genes that had matching AEP. These results give us more information about which biological pathways and processes are important between exposures, and which ones are likely influencing both outcomes.
3.1 Expressed Genes

We compared the probes from each array that were above background to assess what genes were assayed on both platforms. The arrays in the AAD and FLU studies had 71% and 83% of their genes in common. Of the 13,437 and 11,529 genes that had levels of expression above the background level (in the AAD and FLU studies) the overlap came out to 9,612 genes. These 9,612 genes serve as the initial population for all subsequent comparisons. We analyzed differential expression between high response and low response subsets of the mice classified by phenotype. [17, 19] We found that these analyses yield very different numbers of differentially expressed transcripts, with the FLU study having more than 100 times the number of differentially expressed transcripts than the AAD study. Both sets were enriched for GO terms and pathways that correspond to their respective exposure. (Appendix C).
3.2 AAD eQTL

The AAD study produced 6,971 eQTL for 5,031 genes (5,031/9,612; 52.34%), shown in blue in figure 3.1. We found, just as in the previous study, that the majority of eQTL in this analysis were cis-eQTL (> 90%). In total there were 6,317 cis-eQTL associated with 4,717 genes, and 651 trans-eQTL associated with 610 genes. For the 500 genes with the largest adjusted p-values, the GO terms leukocyte and lymphocyte mediated immunity were significantly over-represented (q-value = 0.00289, 0.00544). There was also enrichment for antigen processing and presentation (q-value = 0.0489).

Additionally, trans-eQTL bands were observed on chromosomes 2, 9, and 15, blue vertical clusters in figure 3.1. Each of these bands contained more than 50 trans-eQTL under the control of a single region. The trans-eQTL band at chromosome 2 had been previously described to be associated with eosinophil QTT. [19] It contained 102 trans-eQTL and was enriched for GO terms related to cell death (q-value = 0.0226) and DNA binding (q-value = 0.00478). The trans-bands on chromosome 9 and 15 were also analyzed for functional annotation enrichment. Chromosome 9 contained 64 trans-eQTL and was significantly enriched for chemokine signaling and showed over-representation for GO terms like interleukin15 production as well as CD4 receptor binding (all q-value < 0.01). The 56 trans-eQTL within the chromosome 15 trans-eQTL band contained no significant enrichments after FDR.

3.3 FLU eQTL

eQTL analysis on the FLU study resulted in 6,649 eQTL for 5,074 genes (5,075/9,612; 52.78%), shown in red in figure 3.1. Whereas in the AAD study the majority of eQTL were near the gene of influence, the FLU study contained many more eQTL that were distant from the associated gene. In total there were 3,979 cis-eQTL associated with 3,337 genes, and 2,662 trans-eQTL associated with 2,222 genes. The increase in trans-eQTL is a result of a very strong trans-band on chromosome 16. For the 500 genes with the largest adjusted p-values the GO term enrichments are for metabolic processes often associated with immune response (q-value < 0.05).
Figure 3.1 Global plot for eQTL in both AAD and FLU studies

This image illustrates the global positions of each eQTL-gene pair from both the AAD (blue) and FLU (red) analyses. The strong diagonal band represents the cis-eQTL from both experiments. The three distinct blue and red vertical bands represent the trans-bands that are discrete to an insult. This illustrates how these bands could be used to classify a response.
Table 3.1 Functional Annotation Mx1

This table illustrates pathways overrepresented from the top 300 genes within the chromosome 16 trans band in the FLU data. There is strong representation for interleukin signaling, Interferon signaling and Jak-STAT signaling, all pathways important for immune and inflammatory responses.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Set</th>
<th>Candidates Contained</th>
<th>P-value</th>
<th>Q-value</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regulation of IFNG signaling</td>
<td>11</td>
<td>3(27.3%)</td>
<td>0.000417</td>
<td>0.0348</td>
<td>Reactome</td>
</tr>
<tr>
<td>IL-3 Signaling Pathway</td>
<td>100</td>
<td>7(7.0%)</td>
<td>0.000514</td>
<td>0.0348</td>
<td>Wikipathways</td>
</tr>
<tr>
<td>Transcriptional misregulation in cancer</td>
<td>180</td>
<td>9(5.0%)</td>
<td>0.000956</td>
<td>0.0348</td>
<td>KEGG</td>
</tr>
<tr>
<td>Cytokine-cytokine receptor interaction</td>
<td>266</td>
<td>11(4.3%)</td>
<td>0.00098</td>
<td>0.0348</td>
<td>KEGG</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>83</td>
<td>6(7.2%)</td>
<td>0.0011</td>
<td>0.0348</td>
<td>Wikipathways</td>
</tr>
<tr>
<td>Interferon gamma signaling</td>
<td>15</td>
<td>3(20.0%)</td>
<td>0.0011</td>
<td>0.0348</td>
<td>Reactome</td>
</tr>
<tr>
<td>Jak-STAT signaling pathway</td>
<td>155</td>
<td>8(5.2%)</td>
<td>0.00149</td>
<td>0.0348</td>
<td>KEGG</td>
</tr>
<tr>
<td>Cytokine Signaling in Immune System</td>
<td>193</td>
<td>9(4.7%)</td>
<td>0.00156</td>
<td>0.0348</td>
<td>Reactome</td>
</tr>
<tr>
<td>IL-4 signaling Pathway</td>
<td>61</td>
<td>5(8.2%)</td>
<td>0.00164</td>
<td>0.0348</td>
<td>Wikipathways</td>
</tr>
<tr>
<td>Interferon Signaling</td>
<td>61</td>
<td>5(8.2%)</td>
<td>0.00164</td>
<td>0.0348</td>
<td>Reactome</td>
</tr>
</tbody>
</table>
3.3. FLU EQTL

The noticeable clustering of trans-eQTL around chromosome 16, depicted in red in figure 3.1, correspond to a well-known master regulator of influenza infection, *Mx1*. A mid-sized GTPase, which is involved in protecting against viral replication. This position alone contained 2,018 trans-eQTL that coordinate the expression of 1,722 downstream genes. 33.94% of all genes had an eQTL located at *Mx1*. Gene set analysis on this trans-band confirms it's strong association with immune response processes (Depicted in table 3.1). Additional trans-bands on chromosome 2 and 7, shown in red vertical clusters in figure 3.1, were also assessed for annotations. Chromosome 2 contained 42 trans-eQTL and was overrepresented for transport genes (q-value < 0.01), and chromosome 7 contained 114 trans-eQTL and was overrepresented for antigen processing and ubiquitination (q-value < 0.05).

**Figure 3.2** Partitioning of genes from individual experiments through analysis

Illustrations of comparisons after individual eQTL scans are completed. The emphasis is on those sets of genes that are discrete to one study or the other, or are shared between them. The next assessment is for the individual eQTL observed. Finally, for those eQTL that are shared we observe the number who has matching, non-matching, or inverted allele effects (ambiguous AEP excluded).
3.4 Shared eQTL

The results of both independent eQTL analyses were compared and eQTL that were observed in both studies were assessed. There were 6,920 genes (71.99%) that had eQTL in at least one of the studies. Only 1,846 and 1,889 genes had eQTL in only the AAD or FLU study respectively. The genes within these two groups were mostly from the trans-bands (AAD = 30.6% and FLU = 55.6%) observed during the independent mapping of the eQTL. There were 984 eQTL, 41.13% of the eQTL only in the FLU study, that originate from the \( Mx1 \) locus. There were significantly more genes (3,185) that had at least one eQTL in both studies than would be expected in a population of 9,612 genes by chance (p-value = 2.09e-49).

When we compared the eQTL intervals of the shared genes, we found that only 2,281 genes also shared an interval. This step removed 904 genes, and indicates that there are fewer shared eQTL regions than there are shared genes between studies. Of these genes, 2,274 are genes with cis-eQTL and 7 are genes for trans-eQTL. There were significantly more cis-eQTL observed with matching intervals than would be expected given the number of cis-eQTL observed in either study (p-value = 2.21e-32). This significantly large number of cis-eQTL could be the result of the distinct trans-bands observed in each experiment. It was expected that there would be more cis-eQTL since many genes in both studies will have nearby eQTL. The distinct trans-bands contributed to this overrepresentation of cis-eQTL overlaps as well. With a third of the eQTL in the FLU study being associated with the \( Mx1 \) trans-band it is especially unsurprising that there are fewer shared trans-eQTL. These genes correspond to 2,974 and 2,525 eQTL respectively for AAD and FLU studies, where discrepancies are found where two eQTL were discovered for the same gene. The enrichment for GO terms associated with this subset of genes were for peptide antigen binding (q-value < 0.05) and the lysosome KEGG pathway (q-value < 0.05).

The other 903 genes that were shared between studies, but did not share the same eQTL interval, were compared. These genes that were not colocализed to the same eQTL position were in most cases due to a trans-eQTL in one experiment associated with a cis-eQTL for the same gene in the separate study (>98%). 643 genes had trans-eQTL for \( Mx1 \) in the FLU study matching with a cis-eQTL in the FLU study. One example of these is the \( Fcer1g \) gene, which codes for the high-affinity IgE receptor. This gene had a cis-eQTL near its location on chromosome 1 in the AAD study, but in the FLU Study, it had a trans-eQTL near the position of the well-known IAV resistance gene \( Mx1 \), which could indicate its upstream influence on the expression of \( Fcer1g \).

Overall, the group of genes with \( Mx1 \) trans-eQTL in the FLU study and cis-eQTL in the AAD study were enriched for GO terms associated with cellular proliferation involved in lung morphogenesis (q-value = 0.000436), regulation cell communication and signal transduction (q-value = 0.00117), and adaptive immune response (q-value = 0.0167).
3.5 Allele Effects Comparisons

Allele effect pattern comparisons were made to assess whether eQTL observed in both studies are supported at the allelic level. AEP that were not ambiguous were considered either matching, not matching, or inverted. The gross majority of these comparisons resulted in non-match classifier as shown in figure 5. Only about 27.2% (193/708) of the comparisons had matching AEP. This indicates our previous report of significant overlaps in eQTL only extend to their constructed intervals and overrepresent the real number of shared variants between insults. Those AEP that don’t match suggest that the region that is shared contain multiple positions contributing control to the gene in different conditions. Matching AEP could indicate support of the same locus being involved in both studies. While inverted and non-matching effects could indicate the presence of multiple variants from the same founder strain.

Figure 3.3 Example of matching AEP

This figure illustrates and example of shared eQTL with matching allele effects in both studies. These types of results indicate that, likely, the same regulatory element is responsible for both associations.

An example of a set of matching AEP comparison is the Interleukin17 receptor d, *Il17rd*, depicted in figure 3.3, which is an important receptor that is involved with autoimmune diseases and...
3.5. ALLELE EFFECTS COMPARISONS

inflammation. CAST/EiJ and PWK/PhJ founder strains are associated with reduced levels of *Il17rd* expression. While the other strains group together and appear to influence higher expression of the *Il17rd* gene. This information implies that this gene is important in both insults, and that this same eQTL interval exerts control on the expression of the gene in both insults. The fact that in both experiments the same founder strain drives this association indicates that they may both in fact be driven by the same causal variant.

The Bottomly et al. paper concludes that there were three poorly characterized genes within the flu dataset that were important to resistance to IAV infection. [17] One of these uncharacterized genes, *BC022687*, serves as a second example of matching AEP. The same eQTL interval associated with this gene was observed in the AAD study as well. [19] This same eQTL interval, associated with *BC022687*, is detected in both datasets of the current analysis as well.

Figure 3.4 Example of inverted AEP

This figure illustrates and example of shared eQTL with inverted allele effects in both studies. These types of results indicate that, likely, different regulatory elements are responsible for both associations, but that the same founder strain harbors the effects.

Inverted allele effects were also observed in the comparisons. This occurs when the causal strain for one eQTL has opposing effects in the two studies. In particular the Janus kinase, *Jak1*, illustrates
one such inversion shown here in figure 3.4. The PWK/PhJ allele is associated with lower expression in the AAD study and higher expression in the FLU study.
4.1 Functional enrichment within gene sets

We assessed a number of gene sets that were uniquely controlled by one of the two experimental conditions. Three main groups of genes were analyzed for functional enrichment. These groups include differentially expressed genes (Appendix-C), trans-bands, and genes downstream of eQTL intervals that are found in only one of the two studies. The enrichments produced from these groups had members that often indicated immune response and in some cases were unique to either AAD or IAV. The example of the chromosome 16 trans-band for Mx1 clearly indicates an influenza infection response being mounted in these animals. Functional enrichment at AAD study trans-bands, such as the one on chromosome 9 indicate cell communication GO terms such as IL-15 production.

In the instance of selecting the top cis-eQTL, the fact that they are distributed across the genome makes gene set analysis difficult. There are significant ontologies and pathways observed, but often the q-values are much larger than with an analysis using differential expression. This issue is not resolved by using modulated modularity clustering (MMC) [28]. The results are many small modules containing only a few correlated genes. This issue is addressed in Appendix 5, briefly the results of clustering are a series of modules that contain one or two genes. These sets are not enough for proper gene-set analysis, and also seem to indicate that many of the most significant eQTL are related to genes with unique expression patterns. These expression patterns tend to be distinct from one another. This makes sense, because instead of selecting genes that all pass some filtering cut-off
you are actually performing a filtering for gene expression that is uniquely varying with a genomic region, which will results in distinct patterns difficult to cluster.

Because of this difficulty we compared these datasets to ascertain genes with eQTL discrete to an insult before gene set analysis. These results, where common eQTL between studies are removed, showed better enrichment during gene set analysis. This makes sense, by removing eQTL common to multiple insults we expect to distill those eQTL that are specific to an insult. There were some high quality enrichments to note. The FLU eQTL, for instance, were enriched for Influenza A response pathways at a q-value \(< 0.05\), which supports the notion that there is an over-representation of genes involved with host response to flu within the set of flu specific eQTL. Additionally, AAD eQTL were enriched for leukocyte and lymphocyte mediated immunity (q-value \(< 0.05\)).

There were some interesting enrichments observed within the two studies as well that support some emerging new systems related to these diseases. Cilia are important first defenses against respiratory insults. The secretion of ligands that bind to the surface of virions and allergens prevent these from interacting with the cells themselves and immobilize the particle in the mucosal film of the alveolar epithelium. [29] These cilia also play an important role in clearing this mucous from the lungs and thus clearing some of the particles contained within. [29] There were enrichments within the AAD study specific eQTL that implicate cilial formation and morphology as significant ontologies. Enrichment for collagen pathways reinforces the AAD pathology of collagenitis. Concurrent eQTL in both studies for the gene \(BC022687\) link cilia morphology to effects in both AAD and FLU studies. The FLU study eQTL appeared to show enrichment for the insulin pathway as well as type II diabetes mellitus. These associations agree with the literature that indicates more severe influenza infections and higher susceptibility in mice that have characteristics of type II diabetes. [30] Both studies showed consistent enrichment for metabolic processes. These processes have been implicated previously in differences in response to insults. [31] The immune system and all the cell types that play a role in its function are tightly regulated, and so it is unsurprising that subtle perturbations can contribute to substantial differences in response.

### 4.2 Trans-eQTL Bands

In this study trans-bands could be employed to classify an AAD or IAV response. These bands, by their definition, are clusters of associations of many gene expression traits with a single genomic locus. The biological hypothesis behind these trans bands is that they are regions associated with large-scale regulation of the expression of the genome. The example presented here are the 3 trans bands characterized on chromosomes 2, 9, and 15 in the AAD eQTL scans. Genes in these trans-bands were clearly associated with signaling and classic immune response signaling pathways.

Likewise in assessing the FLU trans-eQTL bands there were similar enrichments for transport genes and antigen processing in chromosomes 2 and 7 respectively. There was strong enrichment
at the well-defined master regulator gene, *Mx1*, which lies on chromosome 16 and has been well characterized previously. [32] This trans-band alone contributed a great deal of the eQTL in the FLU study (30.9%) and almost all of the trans-eQTL (75.8%). This helps to explain why a significant number of genes were observed that had eQTL in both experiments, before we delved into comparing the intervals and AEP. 903 of the genes that had eQTL in both studies did not have eQTL at the same interval. More than 70% of these genes with different intervals in the different experiments were a result of trans-eQTL for *Mx1* in FLU that corresponded to a cis-eQTL in the AAD study, and even more fall into this group when we consider trans-eQTL for AAD that matched an *Mx1* localized eQTL for the same gene (78.7%). The above example of the IgE receptor, *Fcer1g*, illustrates this with the AAD eQTL on chromosome 1 and the FLU eQTL on chromosome 16 near the IAV resistance gene *Mx1*. This association is supported by observations of increased IgE levels in virus-infected individuals. [33, 34]

Not only were the signals within these trans-bands enriched for treatment specific ontologies. They also served as unique identifiers to either study. The only near overlap were the two trans-bands on chromosome 2 in both studies, but the trans-band in the AAD study is more distal on the chromosome than in the FLU study indicating that their intervals do not overlap. The existence of these treatment specific trans-bands alone allows us to discriminate between exposures. This is an important factor biologically in that these distinct large subsets of genes are under the regulatory control of distinct loci dependent on environmental conditions.

### 4.3 Shared eQTL after AEP comparison

In this analysis we detected more than 2000 genes which had eQTL in both experiments that colocalized to the same genomic region. These genes could be considered core genes influencing both lung immune responses. However, because we do not have BAL RNA extracts from control mice we can not be sure that all of these genes are relevant to lung immune response. It is possible that some of these shared eQTL would actually be shared in a control group as well. These would represents variation in these gene expression not explained by either treatment. We would expect these eQTL to be mostly cis-eQTL for the control case, and some of those cis-eQTL would likely have an overlap with those shared between the two test exposures. An analysis of the eQTL in a control group of mice would give us some information about which of our shared eQTL are independent of treatment. It would be interesting to also compare the allele effects between this control group and the two experimental conditions. This would give us additional information to assess whether the two observed experimental eQTL were the same variant or different variants.

Another of our major interests in comparing eQTL studies is to assess if any eQTL, detected that are common to multiple environments, are in fact the same eQTL. We, thus, assessed the shared eQTL intervals between environmental conditions to address the question of the extent to which
common variants are shared between two complex diseases. Previous studies have shown these overlaps between treatments in dendritic cells. [6, 8] Here we present a comparison between BAL recovered from related mice to address the extent of sharing between AAD and IAV pathologies.

For those eQTL intervals that overlapped between studies it was important to observe the allele effect patterns. It is well established that cis-regulatory regions cluster within the upstream promoters of genes. [35] Therefore, it is not enough to simply find overlapping eQTL intervals. The AEP provide us with a way of giving support or opposition to the presence of a single variant between studies. The example above of Il17rd illustrates a potentially true variant associated with response in two different insults, because the eQTL intervals overlap and the AEPs match between insults. Previous literature has implicated deficiencies in its expression to increased induction of NF-κB, Il-6 and keratinocyte chemoattractant (KC). [36] These results and the biological implication seem to imply that this same eQTL is important in both experimental conditions. This could imply that these mice have less neutrophil recruitment due to the lower expression of this required receptor. Neutrophil recruitment is an important and balanced process for combating foreign invaders. Too much recruitment can result in increased inflammation and damage to the host while too few neutrophils may not adequately protect and clear invaders. [36–38]

The matching AEP for the uncharacterized BC022687 gene is also an interesting result that implies a single eQTL in both insults. The gene is also annotated as Cspp1, and is associated with spindle fiber formation. Other studies have shown that deficiencies in this gene contribute to poor ciliogenesis and impaired sonic hedgehog, SHH, signaling. [39] This could potentially help explain the purpose of this shared eQTL, as signaling between cells is important for both innate and adaptive immune responses. [37, 38]

### 4.4 Extent of eQTL sharing between these studies

In this study we show that a majority of shared eQTL intervals, more than 90% in these studies, in fact have non-matching AEP. This was because the patterns of allele effects were not the same between the two overlapping intervals. In the case of a single variant effecting both traits we expect them to have similar allele effects. Therefore, it is more likely in these non-matching cases that there are multiple causal variant beneath the estimated eQTL region, and that they are discrete to an environmental condition, which is consistent with the biology. It is well established that there are a great many regulatory regions upstream of any gene within the genome. Not all of these regulatory motifs are active at all times. Therefore, it is reasonable to assume that individual variants may effect the recognition or affinity of some trans-acting factor binds to the regulatory motif and that these effects are only detectable when the specific regulatory cascade is triggered in an exposure. The inverted example of the Jak1 gene illustrates this fact. Jak1 is associated with induction of the Th2 immune responses and is upstream of the Th2-dominant type of asthma. It has been shown
that lowered expression reduced viral titers and inhibited virion formation. [40] In the context of these two results it is interesting that there is this apparent inversion. This could indicate that different positions, but the same founder allele contribute to the observed associations. There is less evidence in support of the same locus variant causing both up and down regulation dependent on the treatment. Model cases can be imagined of more complex interactions at the locus, such as a regulatory motif that has context dependent effects on gene expression. However, the more likely reason is that while the same founder strain is driving effects in both conditions, the observed expression levels, in either study, are the result of different regulatory loci within the detected interval. This interpretation is further supported by the observation that CAST/EiJ and PWK/PhJ and WSB/EiJ founder strains alleles are driving expression differences for many of the detected eQTL. All three strains are derived from wild caught mice, while the other five strains originate from inbred laboratory strains [3, 13]. The fact that the majority of matching allele effects were for the wild-derived founder strain may further over estimate the extent of shared eQTL, because this may imply that some matching AEP may obscure multiple eQTL.

With this in mind, the minority of overlapping eQTL that were driven by the same founders in both studies have the most evidence of being the same regulatory element. They represent common variants that underlie multiple complex diseases, making them of increased interest when investigating common gene modules underlying multiple similar disease states. It appears that while many genes may have pleiotropic effects dependent on context, the regulatory scheme that influence the interactions of these genes appear to be more context dependent. This refutes the idea that there are large numbers of common variants influencing multiple complex diseases. These results seem to indicate that the underlying regulatory make-up of even pathologically similar diseases are likely very different and made up of many small regulatory perturbations distinct to the insult, even when there is substantial overlap between the gene sets.


APPENDICES
A.1 Hematology

QTL analysis is centered around the analysis of variance for a phenotypic trait at markers scattered throughout the genome of individuals in an experiment, and because of this fact we must assess whether our traits conform to a roughly Normal distribution in order for the tests to reflect real results. To this end we plotted histograms and performed Shapiro-Wilk tests of each phenotype as a qualitative and quantitative approach to assessing normality. After analysis we selected proper transforms between (raw (untransformed), log, and square root) for each phenotype based on maximizing the score of W in the Shapiro-Wilks test between the above transformation scalars.

A.2 Peak conditioning

Previous research in incipient lines of the Collaborative Cross from which the Diversity Outbred population originates had observed a large QTL peak for MCV (Mean Red Blood Cell Volume) on Chr7. [21] This same peak was observed consistently in this data set and so in order to search for QTL of smaller effect within phenotypes where this spike was observed we used the allelic status of individuals at a SNP within this interval to use as an additive covariate and suppress this peak.
We successfully suppressed the peak in this interval, but the results yielded no new QTL. In fact a previously observed QTL peak for RDW vanished when conditioning for the MCV peak thus leading us to the conclusion that this peak is likely just an artifact of a second locus that is potentially in LD with this peak for MCV.

A.3 Conditioning for Sample Batches

A second concern was that of batch effects, since not all samples were collected or analyzed on the same day, we wanted to perform scans were we included an additive covariate for samples within the same batch. The resulting scans were only subtly different than the pervious scans of the transformed data. After this we sequentially dropped batches from the analysis (approximating a Drop-D permutation). This further confirmed that the peak for RDW was an artifact as the peak could be lost depending on the batches that were actually used in the scan.

A.4 Removing Outliers

We performed analyses similar to above with phenotype files where outliers had been removed and this again yielded no new QTL and did not significantly change the results of scans except for again suppressing the peak for RDW on Chr1 again further indicating that it was in fact an artifact.

A.5 Manhattan plots

After creating the scans we decided to also use manhattan plots to visualize the scans. An example of this type of plot is depicted in figure A.1 These plots had the added advantage of allowing us to highlight individual SNPs of particularly high association with a peak along the genome. This made it easier to select the most likely SNPs to control for in subsequent conditional scans and to ascertain the genotypic state of individuals in order to do post-hoc tests of allele effects.

A.6 Results

Previous results had pointed to several new QTL that could be involved with the traits measured; however, upon data transformation these peaks disappeared. It seems that these peaks were associated non-normal nature of the raw data for these traits as only peaks observed in transformed (to more closely approximate normality) data vanished. This resulted in two peaks the classically
A.6. RESULTS

APPENDIX A. QTL MAPPING

Figure A.1 Manhattan plot of log(RBC Diameter Width)

This plot depicts a manhattan plot generated from a scan of RBC diameter width. There is one clear QTL on chromosome 7, near the same region of a mean cell volume (MCV) QTL, and another suggestive peak on chromosome 1. The advantage of this type of plot is that we can highlight individual SNPs of particular interest at the position of QTL.

observed peak at Chr7 and a single peak at Chr1 for the Log of RDW (red blood cell distribution width).
APPENDIX

B

INSUFFICIENT SAMPLES LEAD TO OVERFIT DATA

B.1 Uncharacteristic AEP

An issue that we encountered when performing some QTL analysis on a set of CC mice that were provided from collaborator was the issue of too few samples. The study attempted to explore QTL associated with a number of phenotypic parameters. The issue was noticed when observing AEP for significant peaks in the nominal scan. The allele effect corresponding to the PWK/PhJ and WSB/EiJ founders had uncharacteristically sharp spikes greatly exceeding the range of commonly observed allele effects. This is illustrated in figure B.1.A To address this strange issue we went back to look at the input data.

To this end the data was plotted for the phenotypic value of the mice dependent on the most likely haplotype construction of the genotype at the significant peak associated marker. The results indicated that, while we had 75 CC mice at certain regions of the genome some of the founder haplotypes were not represented. This in effect creates a model that is overfit for the available data.
B.1. UNCHARACTERISTIC AEP APPENDIX B. INSUFFICIENT SAMPLES LEAD TO OVERFIT DATA

Figure B.1 Addition of simulated PWK mice

A) Illustrates the abnormal AEP for a detected QTL. The large estimates for the WSB/EiJ and PWK/PhJ are a result of having no mice with the PWK/PhJ haplotype within the region of the QTL. (B) To address this we simulated inbred PWK/PhJ individuals and added them to the data by taking a mean of the phenotype being tested. We rescanned and generate the above AEP plot. The scan no longer has the large estimates of the two strains after this simulated addition. This is a strong indicator that the absence of the haplotype at the position of the QTL is the cause of the poor estimation of the effects.
B.2 Pseudo counts

To verify that this lack of representation of a founder contributes to an overfit mode and thus results in inaccurate estimates of the beta coefficient associated with the last of the eight founder states I introduced simulated mice. These mice were created to be entirely PWK/PhJ inbred mice and have an average phenotypic value, so they are not the mice driving QTL detection, and re-scan the genomes. With the addition of a few of these simulated mice, where there were previously no PWK/PhJ haplotypes successful controlled the wild estimations of the WSB/EiJ coefficient (Figure B.1.B). While it is odd that the coefficient corresponding to the WSB/EiJ founder strain was the coefficient to be characteristic a careful consideration of linear regression explains the phenomena. Generally, each beta coefficient tries to minimize the sum of square errors (SSE) for a sample. However, since there is no data for the estimation of the effect of the PWK/PhJ allele. The variance of the estimates are large leading to rapid large changes to the corresponding $\beta$-coefficient. Because of this and the fact that WSB/EiJ is the effect when we remove all the other effects, the estimates of its $\beta$-coefficient varies wildly as well.
C.1 Response Groups

Differential expression was assessed in each study. In each study mice were separated into high and low response group. In the AAD study low response mice had log10(eosinophil count) > 5.5 and log(il5 + 1) > 5.0. The low response group of mice had log10(eosinophil count) < 4 and log(il5 + 1) < 1.5. The FLU mice were separated similarly to the methods of Bottomly et al. [17] Briefly, the high response group had immunohistochemical (IHC) staining scores of more than 4 and a day 4-weight loss of greater than 15

These groups were used as contrasts in the limma R package to perform differential expression analysis. [41] Those probes in either array that had greater than 1.5 Fold Change were chosen as differentially expressed and the top 300 in each study were assessed for gene set overrepresentation. Additionally, the overlap of differentially expressed genes was made into a gene set.

C.2 Differentially expressed genes

Differential expression analysis in the AAD study found 253 probes differentially expressed (FC > 1.5). However, there was not very strong evidence that many genes were significantly differentially
expressed based on the associated p-values. This could be a result of the complex architecture underlying the AAD responses. Regardless, within this set there is significant enrichment for annotation corresponding to inflammatory immune response. The FLU expression data were also analyzed for differential expression, and more than 2,667 genes were found to be differentially expressed (FC > 1.5). The top 300 of these differentially expressed genes showed strong enrichment for IAV response GO terms and pathways.

C.3 Results

Assessment of differential expression was limited in regard to the AAD study. There were few genes that had large fold change and even fewer that had associated adjusted p-values less than 0.05. This could be a result of the more subtle nature of AAD phenotypes to measure or that there is no murine equivalent to human asthma. This gene set was overrepresented for genes involved in interleukin receptor interactions and inflammatory response (q-value < 0.01). These annotations coincide with classical pathways and phenotypes of asthma. The most highly differentially expressed FLU genes were enriched for processes such as signalling and chemotaxis (q-value < 0.01) and a number of important chemokine and cytokine signaling pathways such as Cytokine-cytokine, Jak-STAT, and TLR pathways (q-value < 0.01). These annotations seem to confirm that differential expression for T-cell recruitment is occurring. A large number of the genes that had large fold change also had large adjusted p-values as well indicating that they are also significantly differentially expressed. Taken with the immune system signaling annotations the data indicate a differential response to the respiratory insult.
One of the issues observed with the first pass analysis of the Flu dataset was that the split between cis- and trans- eQTL was almost 1:1. This is atypical of eQTL analysis, which generally shows a large proportion of cis-eQTL. However, historically eQTL studies of external perturbations, like viral infection, have produced higher proportions of trans- eQTL. Most of the excess of trans-eQTL in this study (2,054) were found on Chromosome 16 close to a well known Inuenza resistance master regulator $Mx1$. This region was previously observed in Bottomly et al. [17] and the founder allele effects at this locus were grouped by Ferris et al. [18] Sometimes referred to as $HrI1$ (Host resistance to Inuenza 1) $Mx1$ is thought to be the main gene driving this locus signal. [18] A second locus $HrI2$ was also previously observed on Chromosome 7 that showed trans regulatory activity. [18] It is common in QTL studies to perform conditional scans where a locus with a very strong and expected signal is used as a covariate for additional scans. These scans may detect new signals that would be lost without the covariate.
Figure D.1 Conditional Correlation scan of gene downstream of Mx1

This figure depicts a single scan (black) that was performed and yielded a strong peak above the significance threshold localized to the $Mx1$ trans-band. We performed conditional scans (red) where we controlled variation at the $Mx1$ gene, based on AEP, and the scan was not significant for $Mx1$, but did have an interesting new significant peak at chromosome 4.
D.2 Conditioning on *Mx1* locus

There were 5,277 detected eQTL in this analysis. Of the total eQTL, 2868 (54.35%) were found to reside on the chromosome where the gene they correlate with is found. These eQTL are termed cis-eQTL and the low percentage was unexpected. To understand this a scatter plot of the total collection of eQTL was performed. In this plot we see a few regions where there are a large number of trans-eQTL. This may help to explain the high number of trans-eQTL, since we know *Mx1* is strongly tied to resistance to Influenza A, which is located in the trans-band on chromosome 16. In fact, there were 1782 trans-eQTL on chromosome 16, which is 75.67% of the total number of trans-eQTL.

Conditional Scan There were 3,908 detected eQTL in this analysis. Of the total eQTL, 3,100 (79.33%) were cis-eQTL. To further compare these results to the previous ones we created another scatterplot of cis- versus trans- eQTL. The main differences between the original scan and this one were a significant reduction in trans- eQTL on both *HrI1* and *HrI2*, which are on chromosomes 16 and 4 respectively. There was also a noticeable increase in trans- eQTL on chromosome 4.

D.3 Results

These results show that when we control for the known master regulator of in response, *Mx1*, we observe a reduction in trans-eQTL peaks near its position as well as a number of trans-eQTL in a band on chromosome 7 known as *HrI2*. There was also an increase in trans-eQTL within a narrow interval of chromosome 4, this could be an interesting new locus of interest only visible after performing conditional scans. I have not looked more deeply into this locus, but a cursory look into other regions associated with *Mx1* yielded a region of pseudogenized DNA on chromosome 4 in humans.
E.1 Modulated Modularity Clustering

For differentially expressed groups and the top significant eQTL associated genes I performed MMC on the expression data to find clusters of probes that were expressed similarly to one another, but differently from other clusters of probes. MMC results are summarized by probe with the module the expression trait belongs to, a within module correlation, and a total correlation. Additional information summarizes the quality of the overall clustering with a number of clusters and a value for the estimated tuning parameter for the optimum clustering selected by the program.
E.2 AAD Example

For both datasets using MMC with the top 300 genes that had greater than 1.5 FC, as described in the differential expression analysis appendix, yielded good quality clustering. The modules were large, but within them were genes that were all had highly similar annotation after gene set analysis. Some of the enrichments in some modules contained genes that we would expect to see associated with the diseases. Some modules in the AAD study had distinct members of immune response pathways, however few of the modules had significant numbers of members for significant enrichment in gene set analysis.

The eQTL results were much more difficult to interpret. Using the expression patterns of the top ranked 500 genes with eQTL did not generate the expected types of MMC module clusters. Many of these were small and often contained only one or two genes. This is partly a product of the nature of eQTL score and distance relationships. Most of the tops ranked eQTL were for cis-eQTL and this resulted in collecting an unbiased sample of the genes that had strongest signatures. The results are
Figure E.2 MMC of AAD eQTL.

This heatmap was generated from gene expression that had the smallest adjusted p-value in the nominal AAD eQTL scan. 300 genes were clustered and the results are the above correlated modules. Most notably, these modules are much smaller than those in Figure 8, which means fewer genes are part of each module. The fact that these genes have unique expression patterns that vary uniquely with variation at the eQTL locus supports the observation of many smaller modules in this clustering attempt.
a bunch of genes with unique expression patterns and thus clustered poorly, even when the total set is enriched for immune system functions. Many of these small groups of genes have associations with metabolic pathways. This could again be indication that metabolic processes play a large role in the variable responses to allergen insult. A hypothesis that has some support in the literature. [31]