A Computational Analysis of Cardiac Aging using mRNA and Protein Expression Data

Daniel Alfonsetti, Massachusetts Institute of Technology
The Jackson Laboratory
Summer Student Program 2018

Daniel Alfonsetti _____
Student

Susan McClatchy _____
Mentor

August 8, 2018

Abstract

The aging cell has been associated with dysfunctions in protein homeostasis even when disease is not present (de Magalhães). The aging cardiac proteome, like other aging cellular proteomes, displays increased protein oxidation and damage, increased ubiquitination, and dysfunction of the autophagy and ubiquitin quality control systems (Rabinovitch). The goal of this project was to make a pipeline that would further elucidate how genetic variation affects the aging heart using transcriptome and proteome wide heart expression data. We hypothesized that proteins forming the quality control systems in the heart would become less expressed in age. We found several genomic hotspots associated with different processes and changes in mRNA and protein expression levels. We suspect lincRNAs, which lie under many of the hotspots, to be the mediators of many of these associations. Of particular interest was a genomic location on chromosome 3 at 147.5 Mbp that was associated with the decline in expression of structural heart proteins with age. We also found that correlations between mRNA and their respective proteins decreases with age, but the way in which the decrease occurs is sex-dependent.

Introduction

In the elderly population, cardiovascular (CV) disease mortality increases exponentially with age (Rabinovitch) and is the leading cause of global death (AHA). However, it has been shown that those who have healthier cardiovascular systems also tend to have descendants with protection from CV-related deaths, suggesting a genetic

component to cardiac aging (Perls & Terry). For these reasons, genetics research into the processes of cardiac aging is warranted and has the potential to help increase human life expectancy in a very direct way.

To this end, we analyzed cross sectional data on 189 diversity outbred (DO) mice that were grouped into sex and times of sacrifice (6, 12, and 18 months). For each mouse, we analyzed founder strain haplotype probability data across the genome as well as expression levels of 21169 mRNA transcripts and 4193 proteins from the heart tissue. 4122 of these proteins had corresponding mRNA that were part of the 21169 measured transcripts. Protein abundance was measured by the Gygi lab at Harvard Medical School using mass spectrometry. Mouse genomes were sequenced using MUGA (the Mouse Universal Genotyping Array).

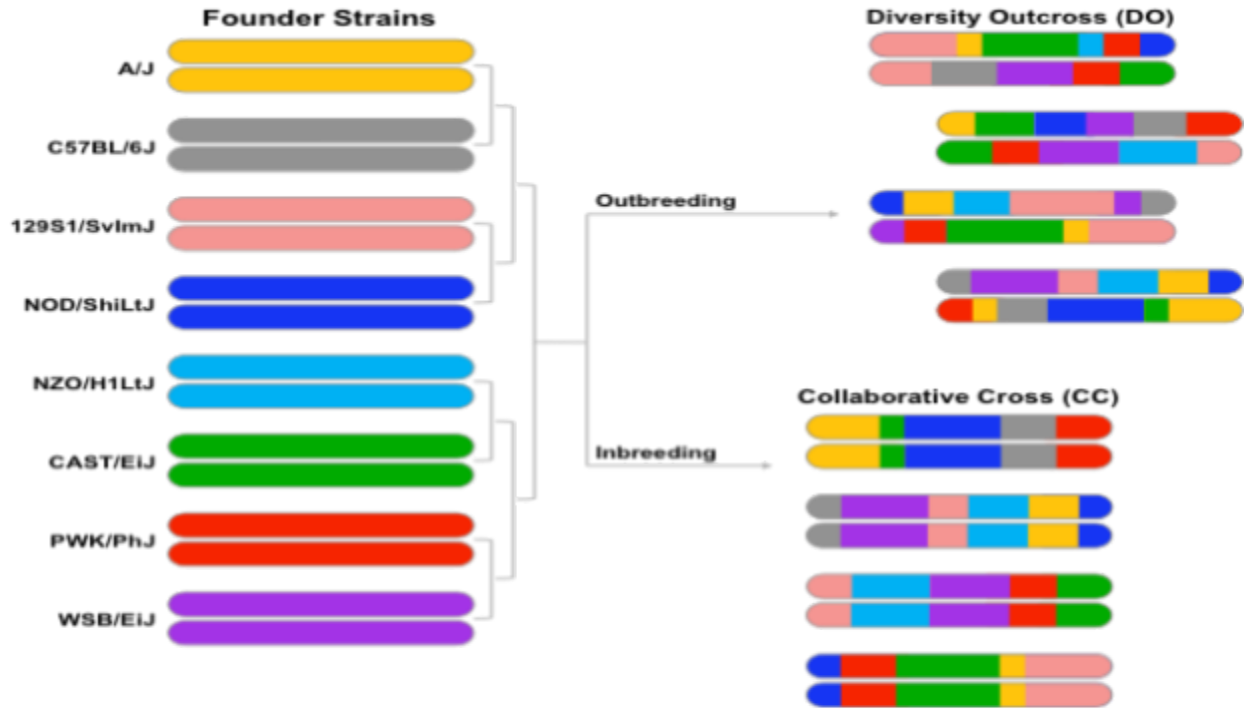

Figure 1. The diversity outbred mice (top right) are a heterogenous stock of mice that model human population genetics better than inbred strains do. They are useful when performing phenotype to genotype association mapping.

The diversity outbred mice are an advanced intercross population that are derived from eight founder strains: A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HlLtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ. Unlike inbred strains which are genetic clones of each other, the diversity outbred mice are heterogenous and have many more observable recombination events. At generation 19, the diversity outbred mice had about 500 visible recombination events while the inbred lines showed about 200. While some statistical power is lost due to more genetic variance, the increased number of recombination events in the DO stock increases the number of genes and proteins that are differentially expressed and allows for finer phenotype-to-genotype mapping

resolution. For these reasons, DO mice are better suited for genome-wide mapping studies (and are thus what we used for this study) while inbred strains are better suited for getting replicates of the same genotype when measuring a noisy trait (Gatti Mammalian Genetics Talk, 2018).

Our main research question was: how does mRNA and protein expression in the heart differ between ages and sexes? Different statistical analyses were performed on the data to answer this question. We will describe the process of each analysis and their results in the following sections and then conclude with a discussion of the findings in aggregate.

Dataset-Wide Differential Expression and Gene Set Enrichment Analysis

This analysis began with measuring the differential expression of each protein and transcript with respect to age and sex. We used t-tests to look for sex differences and ANOVAs for age differences. The p-values for these tests were provided to the functions of the Allez enrichment analysis library in R (Newton), which returned lists of enriched GO and KEGG categories. Using jaccard matrices to find the number of genes that overlapped in each enriched category, we further grouped our results. We have summarized the enriched categories in the table below.

		Differentially Expressed on...	
		Age	**Sex**
Expression Type	**mRNA**	• Leukocyte related categories • Endopeptidase activity	• RNA-splicing histone modification transcriptional activator activity

		• Blood coagulation complement activation	• Acetylation
	Protein	• Organelle subcompartment • golgi apparatus • Lysosome • Vesicle mediated transport	• RNA processing and splicing • Regulation of gene expression • Chromosome organization

mRNA-Protein Correlations

One question we wanted to ask with this data was whether or not the correlations between mRNA and the proteins they code for change with age. Since we had 4122 matching mRNA-protein pairs in our data set, we ran correlations for each of them using their expression levels as inputs. We found that these correlations decreased with age, but this was especially true for females (figure 2). One possible explanation for this decrease is simply that entropy is increasing with age in the cellular systems and the mRNA are not being translated as well as they were previously.

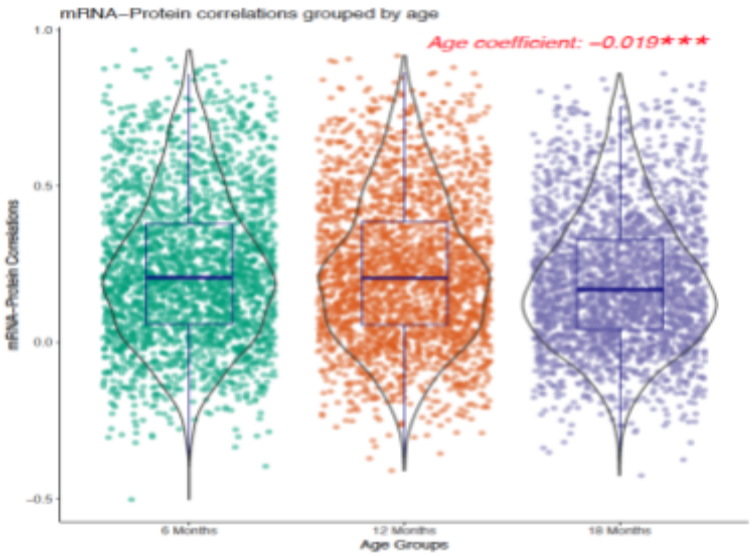

Figure 2. mRNA-protein correlations have a significant decrease with age ($p < 0.001$), but the way they decrease is dependent on sex. Males see an increase in correlation at 12 months and then a steep fall in correlations from 12 to 18 months. Females have a steep decrease from 6 to 12 months.

	Mean Correlation (Males)	Standard Deviation (Males)	Standard Error (Males)	Mean Correlation (Females)	Standard Deviation (Females)	Standard Error (Females)
6 Months	0.20985	0.2699	0.00544	0.25352	0.26682	0.00538
12 Months	0.24604	0.26435	0.00533	0.20478	0.26962	0.00544
18 Months	0.19198	0.24217	0.00488	0.19364	0.25695	0.00518

mRNA-Age and Protein-Age Correlations

Next, we wanted to identify how well mRNA and protein expression levels were correlated with age after regressing out sex. To do this, we took each protein that we also had corresponding mRNA expression data for and then calculated the protein-age correlation and the mRNA-age correlation while controlling for sex. We then plotted the resulting Pearson correlation coefficients for each pair and identified which pairs had significant correlations in both axes. Grouping by quadrant, we performed gene set enrichment analysis on the significant pairs (figure 3).

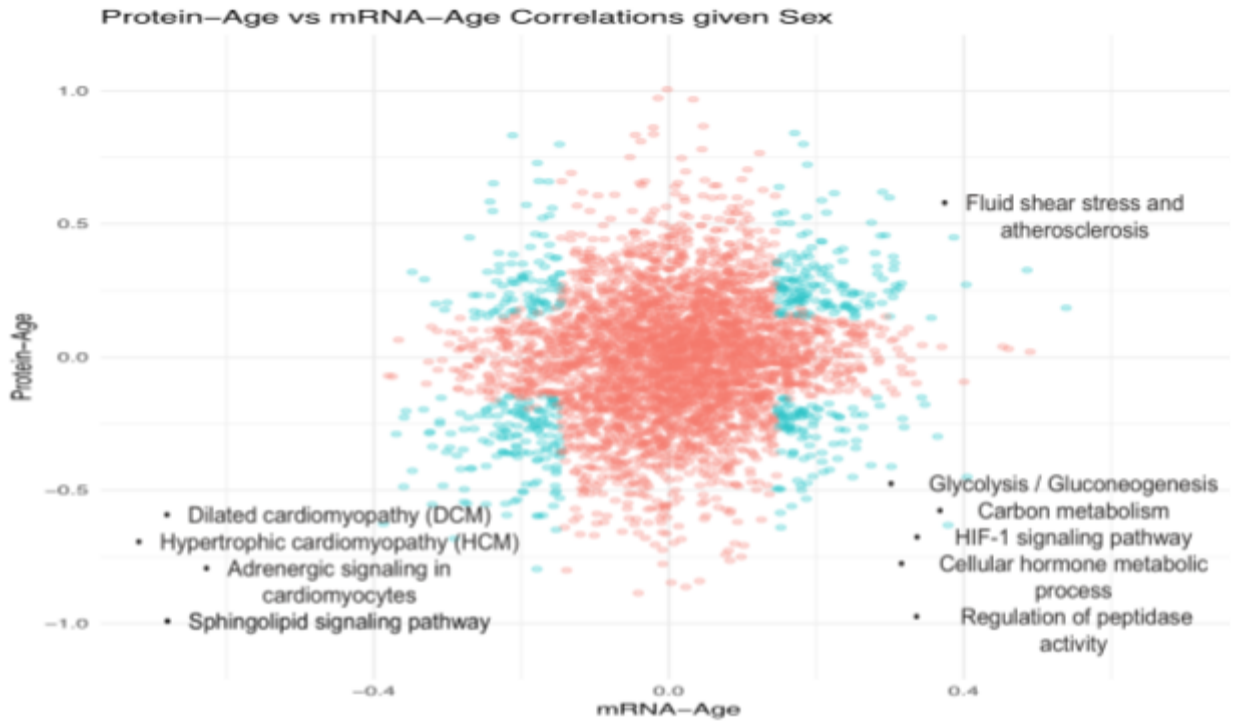

Figure 3. Each dot corresponds to a gene with both mRNA and protein expression data. The x-axis represents the mRNA-Age pearson correlation coefficient and the y-axis represents the protein-Age coefficient. Blue dots represent genes whose mRNA-age and protein-Age correlations are both significant ($p < 0.05$). Genes in each quadrant that had significance on both axes were grouped together and gene set enrichment was performed on them using the clusterProfiler R library (Yu). The labels in each quadrant are the GO and KEGG categories that were enriched for those genes. No categories came up as enriched for the upper left quadrant (low mRNA-Age correlation, high protein-Age correlation).

Quantitative Trait Locus (QTL) Mapping

Quantitative Trait Locus (QTL) mapping was the bulk of this research and had the most significant results. A quantitative trait is a trait that takes on continuous values and is affected by many genes. QTL mapping is a method of associating genomic location(s) to a quantitative trait of interest. A quantitative trait locus is the genomic location affecting the quantitative trait.

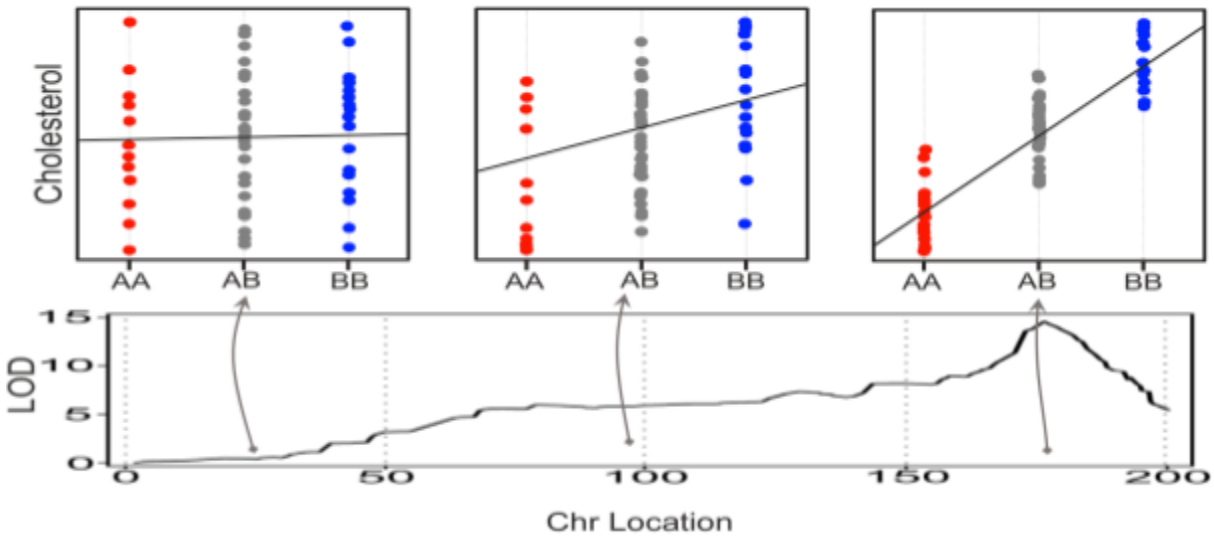

Figure 4. A pedagogical diagram showing a standard QTL scan (bottom) and how association between a SNP and a phenotype is quantified (top). The better the genotype at a particularly SNP is at predicting the phenotype, the more significant the SNP. Image credit: Dan Gatti.

Roughly speaking, QTL mapping works by regressing the phenotype of interest on genotype (or haplotype background, such as in our study) at each SNP (figure 4). SNPs with the most significant correlations therefore serve as markers indicating that the phenotype is associated with that area of the genome and that there is probably a gene (or genes) near that SNP that are mediating that SNP's genotype-phenotype correlation. As previously alluded to, the diversity outbred mice, from which our data came from, are particularly informative for this type of genome-wide analysis as their genetic diversity and increased observable recombination frequency allows for finer mapping resolution. QTLs run on protein expression phenotypes are called pQTLs while QTLs run on mRNA expression phenotypes are called eQTLs. If a gene's expression is controlled by a QTL that is not near that gene's genetic location, that QTL is said to be acting in 'trans', while QTL that are near the gene are said to be cis-QTL. In this study,

trans acting QTL were said to be QTL with a maximum peak that is at least 4 Mbp away from the associated gene's location (or on another chromosome altogether).

Different linear models can be used when performing the regression at each SNP. For example, instead of simply asking whether or not the genotype at a SNP affects protein expression, we can also ask if the effect of that genotype varies with age. A variable such as age in this example is called an interactive covariate. Interactive covariates differ from additive covariates in that additive covariates allow the phenotypic means of the levels of the covariate to differ, thus chiefly serving to increase the ability to detect QTL. They are used when the covariate is thought to strongly affect the phenotype. Interactive covariates, on the other hand, allow the effect of the QTL on the phenotype to vary with the levels of the interactive covariate (figure 5). When an interactive scan with additive covariates is paired with a purely additive scan, the effect of the genotype-covariate interaction can be identified by comparing the two results (Broman). For a more rigorous treatment of QTL analysis, see Karl Broman's book, "A Guide to QTL Mapping with R/qtl."

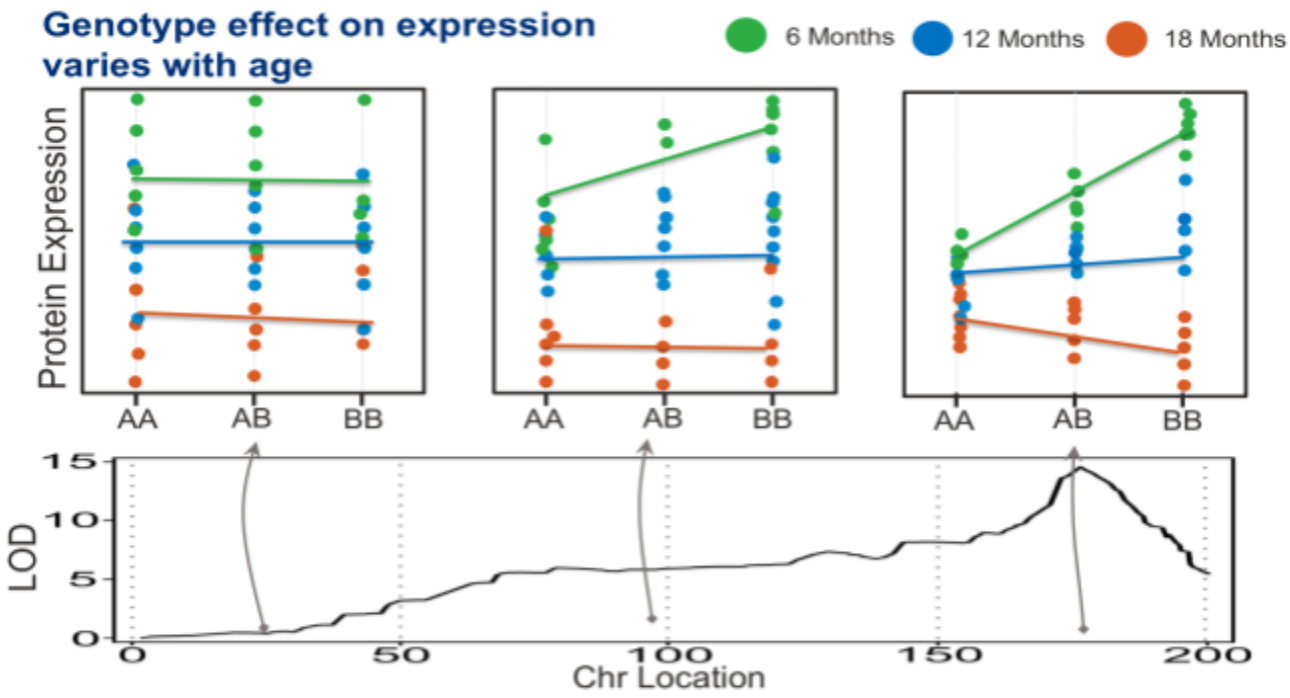

Figure 5. A pedagogical diagram exemplifying an age-interactive QTL scan for protein expression levels. The bottom half of the figure shows a standard QTL diagram with the x-axis representing genomic location and the y-axis representing how significantly the location is associated with the age-QTL interaction effect. You can see that protein expression is decreasing with age just by looking at the green-blue-orange stratifications, but you can also see that some locations in the genome give more details about how this change is occurring than others. For example, on the left, the genotype gives you no new predictive power about how protein expression changes with age. In the middle plot, we get some information since we see that if you have the BB genotype, you will probably have more protein when you are young as compared to your counterparts, but will then be on equal grounds with them as you age. On the right, knowing the genotype gives a lot of information since we see that if you have the BB genotype at that genomic location, you will start off with more protein than any of your peers, but when you are older, you will have the least. An AA genotype at this location, however, will have much more steady protein expression levels throughout their lifetime.

Since we were most interested in how genetic variation affects the hearts' transcriptome and proteome differently between age groups and sexes, we performed age interactive and sex interactive eQTL and pQTL analysis for all expressed proteins and mRNAs using the methods described above. Our additive covariates for each were sex and age.[1] We recorded any interactive QTL peak that had a LOD score above 6. In total, we had four interactive QTL peak tables: age-QTL interaction for protein expression, age-QTL interaction for mRNA expression, sex-QTL interaction for protein expression, and sex-QTL interaction for mRNA expression. We also recorded the peaks

[1] We originally used age, sex, batch, tag, and generation as additive covariates in each of the scans, but given that the data set only contains 189 mice, we suspected that we were overfitting to our data.

for the full and purely additive models, generating 6 additional QTL peak summary tables. However, the sets of interactive scans will be the focus for the remaining amount of this paper.

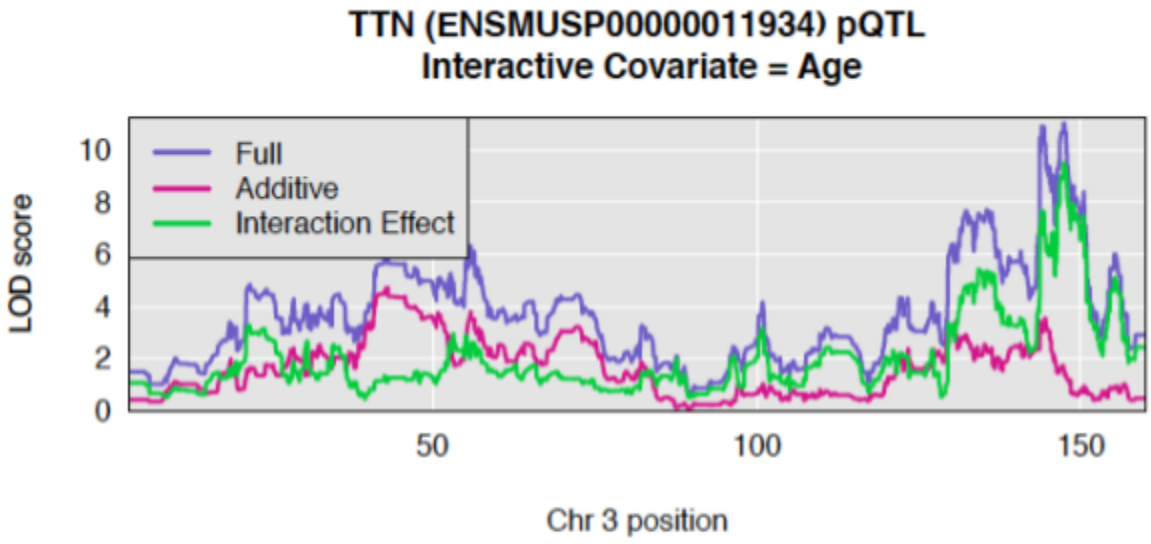

Figure 6. Example age-interactive scan for the TTN protein expression phenotype. The plot shows the large age-QTL interaction at chromosome 3 at 148 Mbp. The difference between the full scan (pheno ~ addcovars + QTL + QTL:intcovar) and the purely additive scan (pheno ~ addcovars + QTL) quantifies the significance of the interaction effect (QTL:intcovar). Interactive effect peaks such as the one displayed here were recorded. Locations in the genome that had many overlapping peaks (hotspot regions) from various expression phenotypes were then identified for each scan type set (protein-age, protein-sex, mrna-age, and mrna-sex).

For each of the four sets of interactive scans, we wanted to find ‘QTL hotspots’ - places in the genome that controlled how many different expression phenotypes changed with age (figure 7). To do this, we created a QTL density histogram using a sliding window paradigm where we would count the number of QTLs (with a somewhat arbitrary peak threshold of 7.2) in a specified interval length (i.e. window size) and then slide that interval length down the genome in increments. For this analysis, we used a window size of 4 Mbp and a sliding increment of 1 Mbp.

One might wonder why we didn't use permutation testing to find the 95th percentile (or some other percentile) of a distribution of genome-wide maximum LOD scores under the null hypothesis of no QTL (as is often done in QTL mapping) and then use it to filter out genes before creating the QTL histogram. The reason is because permutation testing is not particularly useful when searching for hotspots. The intuition behind this is because the permutation threshold will yield a threshold for any one phenotype to be considered significantly associated with a genomic location. However, investigation is also warranted if many phenotypes are seen to be associated with the same location in the genome, even if those latter associations are a bit more weak than what the permutation testing threshold would count as significant. Thus, using permutation testing before making the QTL histogram could potentially hide genomic locations that are drivers for entire sets of phenotypes (in our case, entire sets of transcript and protein expressions).

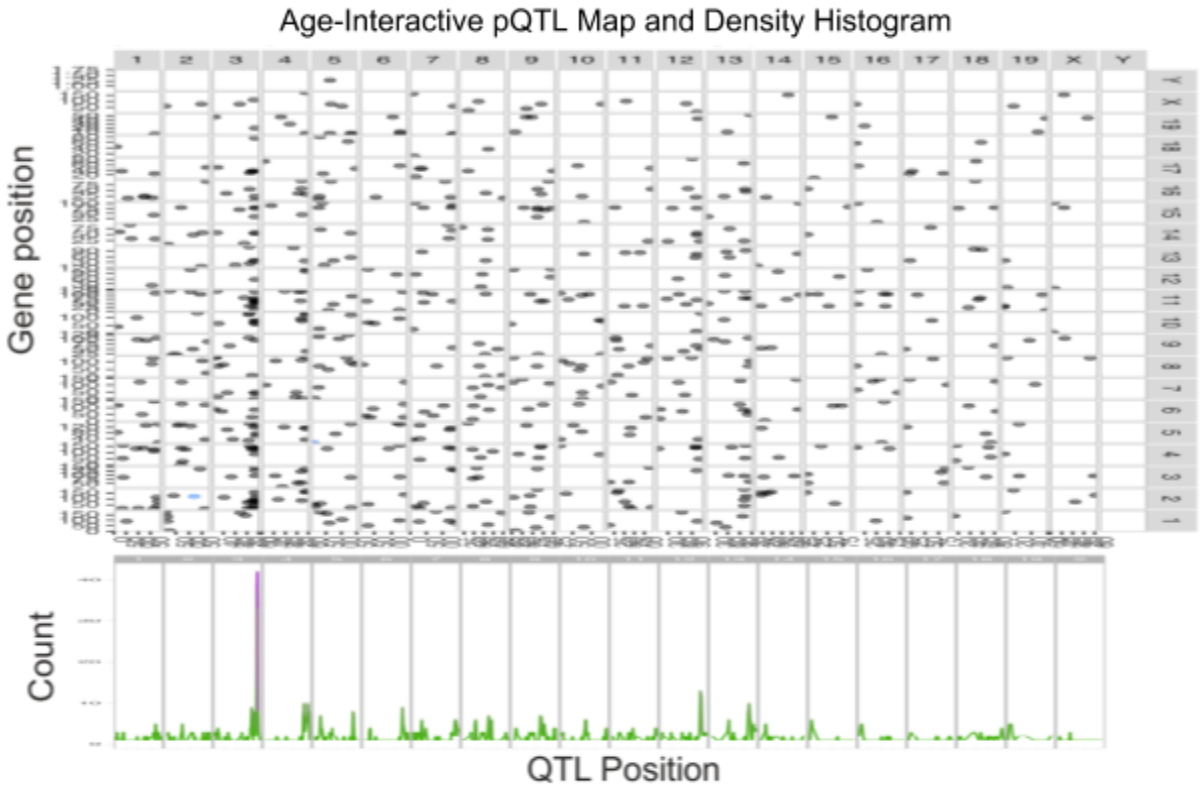

Figure 7. An example of a QTL map and QTL density histogram for age-interactive pQTL scans. QTL maps are used to identify regions in the genome that are affecting the change in expression of many genes. We can see that 42 genes share an age-interactive pQTL on chromosome 3 at around 148 Mbp, warranting further exploration.

The QTL density histograms identified several ‘trans bands’, i.e. hotspot regions in the genome where many phenotypes have trans-QTL that map there. For each hotspot we performed the following pipeline of analysis:

1. Identify enriched functional categories for the genes that have QTL in the hotspot (the ‘outcome’ genes)
2. For each outcome gene…
    a. Find allele effects on that chromosome for its expression (either mRNA or protein expression, depending on the QTL type)
    b. Perform SNP association mapping for its expression

c. Impute founder strain genotypes onto the DO genomes and group the mice by major and minor alleles at the most significant SNP for the gene's expression to see how the genotypes affect expression levels differently between the levels of the interactive covariate (either age or sex).
d. Perform mediation analysis, testing each gene within 10 Mbp of the hotspot peak location as mediators of the interaction effect. See "Supplemental Information - Mediation Analysis for Interactive QTL Mapping" for details. Our ability to detect mediators was hampered in many cases since we did not have expression data for many of the genes under the hotspots.

In total, we identified 11 QTL hotspots. This paper will proceed by sequentially summarizing the results from four hotspots that we thought were particularly interesting.

Age interactive pQTL Chromosome 3, ~ 148 Mbp Hotspot

There were 42 outcome genes comprising this hotspot. These genes were enriched for cardiac muscle GO and KEGG categories. Several of these genes had multiple protein isoforms with QTL in this region. Additionally, many of these genes (Ttn, Myh3, Myh6, Mybpc3, and Actg2) were associated with sarcomere structural proteins (figure 8). Sarcomeres are the basic unit of muscle contraction and are repeating structures in myofibrils which form striated muscle cells. The majority of these proteins showed decreased expression with age, with the most pronounced decrease occurring

for the individuals that were homozygous for the minor allele. Many of these proteins also showed a decrease in their corresponding mRNA expression as well.

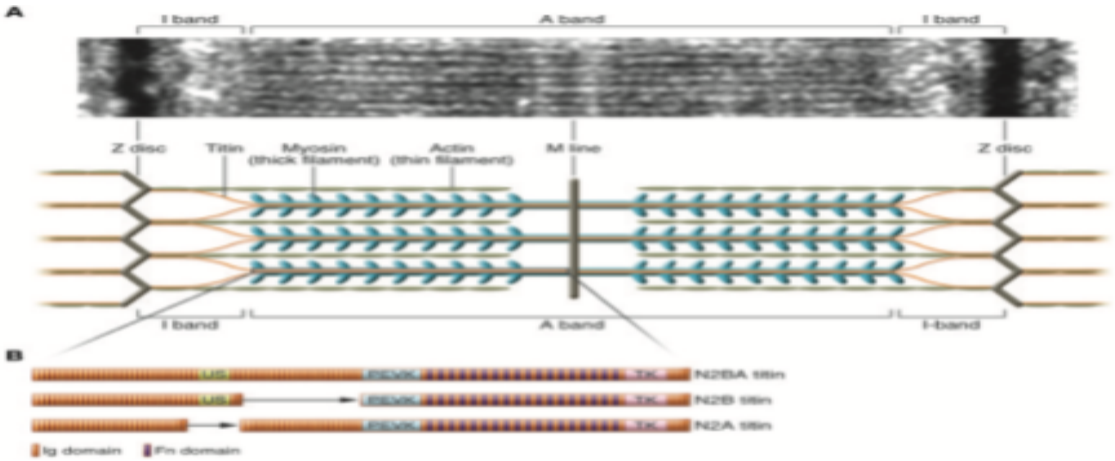

Figure 8. A diagram of a sarcomere. (Diagram A) Sarcomeres are the fundamental unit of muscle contraction found in sequence along the length of the myofibrils of striated muscle cells. They have thick and thin filaments which slide past each other. The thick filaments and thin filaments are made up of myosin and actin, respectively. Titin, Myosin, and Actin were all genes that had QTL in the age-interactive pQTL hotspot on chromosome 3 near 148 Mbp, suggesting that the way these proteins change with age is at least partly determined by genetic variation at this location in the genome. (Diagram B) Various isoforms of the TTN protein. Image credits: McNally

SNP association mapping at the maximum LOD peak for each outcome gene was then performed. For many of these structural protein encoding genes, we found that the most significantly associated SNPs lied directly on top of lincRNA (long intergenic non coding RNA) encoding regions (figure 9). Little research has been done on lincRNA, but evidence suggests that lincRNAs are often associated with gene expression regulation (Ulitsky). This would seem to be congruent with the results found here.

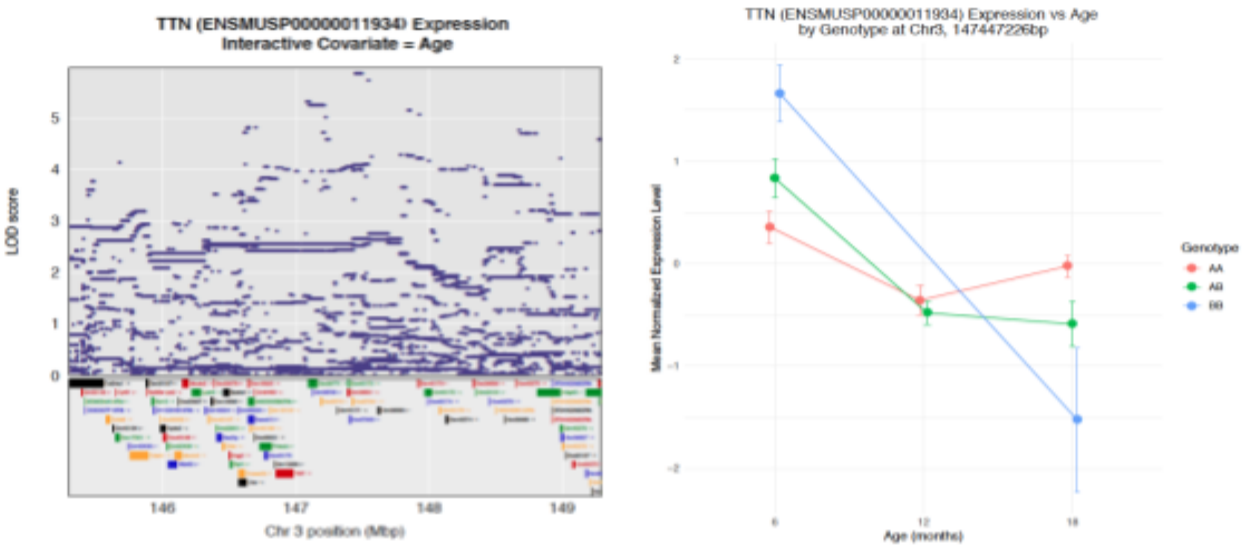

Figure 9. (Left) SNP association shows that the most significant SNP for TTN expression lies directly above lincRNA genes. (Right) TTN expression decreases with age, but the rate and magnitude of decrease is dependent on the genotype at this hotspot. Many of the other outcome proteins in this hotspot have similar expression-by genotype patterns, including ACTG2, MYH3, MYH6 and MYBPC3, all of which are associated with sarcomeres.

Expression levels of these proteins' corresponding transcripts were also stratified by genotype and their change with age was analyzed. Except for the Ttn gene, the expression levels for the transcripts all decreased with age on average (figure 10), but the pattern of decrease by genotype for these transcripts did not seem to be very well match the patterns of decline seen in their protein expression. This suggests that the change in protein expression with age is affected by other things besides a change in their transcript expression levels.

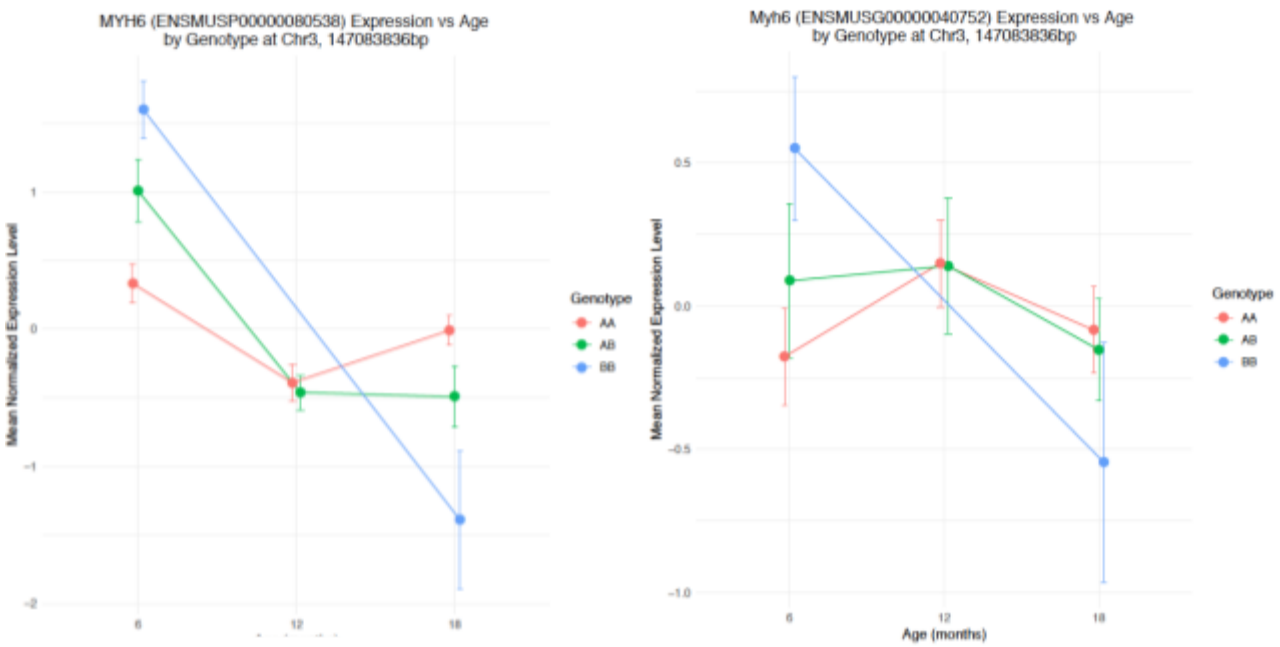

Figure 10. (Left) MYH6 protein expression displays a similar pattern to that of TTN. (Right) Furthermore, Myh6's mRNA expression decreases with age and roughly corresponds to its protein expression. The mRNA expression patterns for the other structural protein encoding genes in this hotspot (Myh3, Actg2, and Mybpc3) do not match their protein patterns as well, suggesting that the change in heart protein expression with age is controlled by more than just changing mRNA expression.

Mediation analysis failed to find any strong protein or mRNA mediators that lied directly underneath QTL peaks. However, 1110002E22RIK and SYDE2 proteins both mediated many of the QTL-age interaction effects, with the top interaction effect LOD drops ranging from 4 to 6 points. The 1110002E22Rik gene is located at 138 Mbp while the Syde2 gene lies closer to the hotspot peak at 146 Mbp.

Sex Interactive pQTL Chromosome 11, ~35 Mpb Hotspot

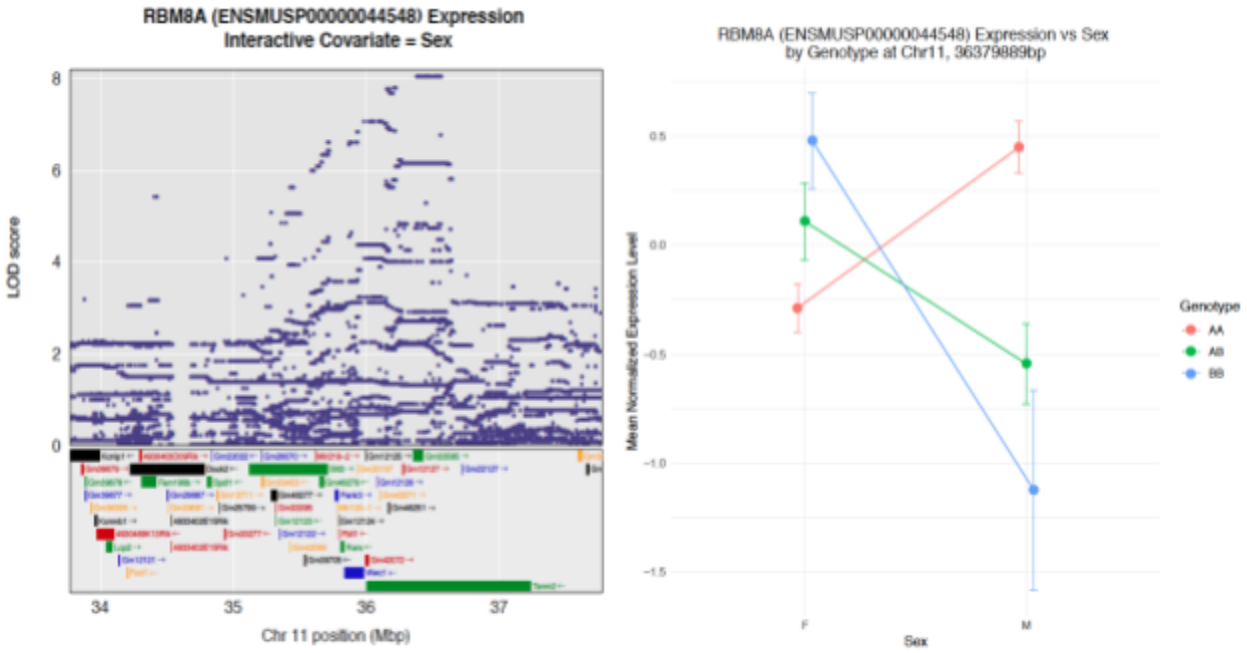

Figure 11. (Left) SNP association mapping shows the most significant SNPs lie directly above lincRNAs. Nearby at 35.8 Mbp is the Rars gene, which came up as a strong mediator of the sex-pQTL interaction for RBM8A and HDGF. (Right) Stratification by genotype at the most significant SNP shows that the AA allele results in males having higher RBM8A protein expression than females, but having an AB or BB allele switches the relationship. The HDGF expression pattern is nearly identical.

There were 18 genes that had pQTLs in this hotspot. RBM8A had the highest LOD score of any protein in this sex-interactive pQTL hotspot, followed by HDGF. The sex difference between HDGF and RBM8A expression was similar when grouped by genotype, with the AA allele giving males greater expression than females while the AB and BB alleles did the opposite (Figure 11). The mRNA expression pattern for Hdgf showed males having higher expression at all genotypes while the pattern for Rbm8a mRNA expression showed males having decreased expression at all genotypes. Mediation analysis showed that both the RBM8A and HDGF sex interactive pQTLs are strongly mediated by both Rars protein (interaction effect LOD drop of 8.9209) and

Rars mRNA expression (drop of 8.8682). Rars lies directly underneath the hotspot at 35.821444 Mbp.

Other outcome genes in this hotspot included the structural proteins found in the age interactive pQTL such as TTN and MYBPC3. Enriched GO categories for the outcome genes included cardiac-tissue morphogenesis, cardiac muscle contraction, negative regulation of gene expression, and negative regulation of translation.

Age Interactive eQTL Chromosome 11, ~112 Mpb Hotspot

We found an age-interactive eQTL hotspot on chromosome 11 containing 30 QTLs. Gene set enrichment analysis for genes in this hotspot identified multiple categories relating to the negative regulation of cell movement. The genes in these categories were Klf4, Ptprr, Rgcc, and Plxnb3. The outcome gene with the largest LOD score was Nhlrc3, with a LOD of 9.71. We noticed that the NZO allele effect for the expression of Nhlrc3 at 18 months was extremely pronounced in a downwards direction.

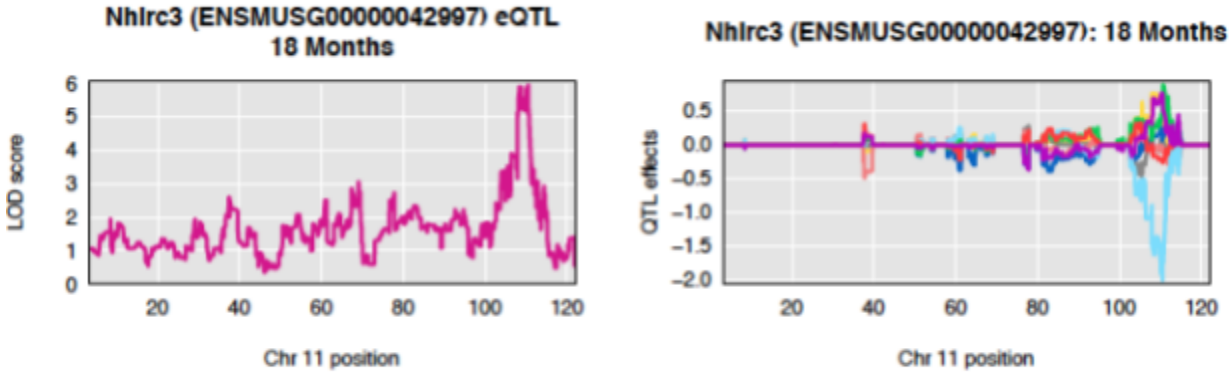

Figure 11. The NZO allele has strongly downregulated Nhlrc3 expression at 18 months.

SNP association mapping for Nhlrc3 expression showed us that the most significant SNP in the hotspot region lies above lincRNA and Abca (ATP-binding

cassette family) encoding genes. Stratification of expression levels by genotype showed that the AB genotype had a large decrease in expression with age while the AA genotype had a large increase in expression with age.

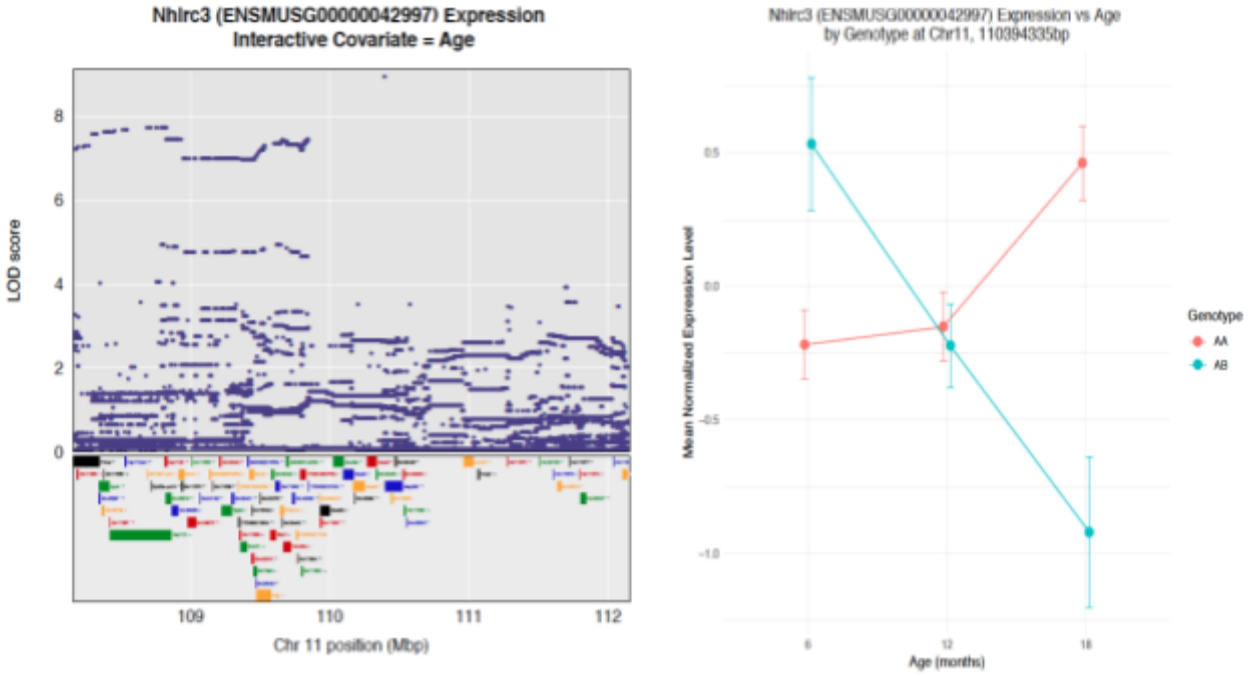

Figure 12. (Left) SNP association mapping for Nhlrc3 shows that the most significant SNP lies directly above lincRNA and Abca (ATP-binding cassette family) encoding genes. (Right) Stratification by genotype at the most significant SNP shows a fairly dramatic difference in how Nhlrc3 expression changes with age between the AA and AB genotypes.

Sex Interactive eQTL - Chromosome 10, ~116 Mbp Hotspot

We found 53 genes with eQTL in this hotspot. The two genes with the most significant LOD scores were Ryr2 (with a LOD score of 11.04) and Apoe (with a LOD score of 9.76). No categories came up as enriched for the genes in this hotspot. We noticed that the A/J allele had a strong genotype effect for the expression of Apoe (figure 12).

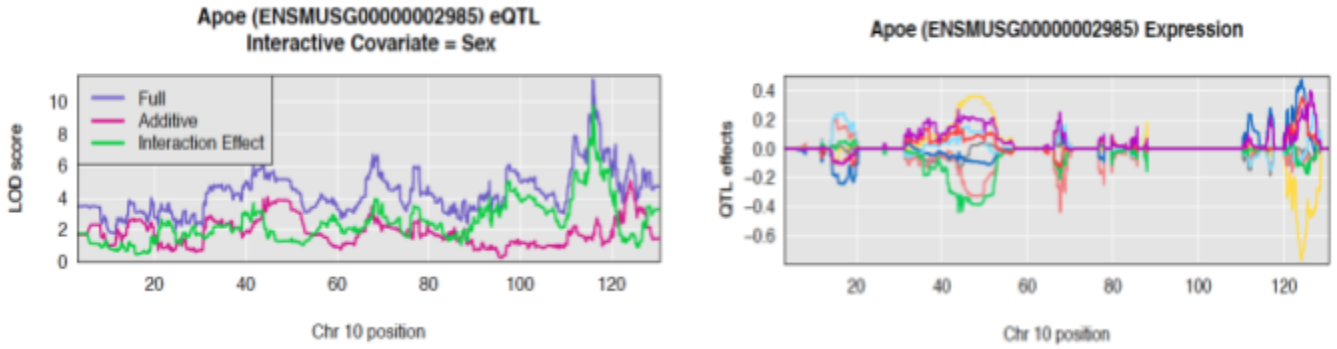

Figure 12. (Left) Apoe shows a strong sex dependent eQTL on chromosome 10 at 115.9 Mbp. (Right) The A/J allele seems to be associated with a downregulation of Apoe mRNA expression.

SNP association mapping for Apoe expression showed a sharp peak at around 115.8 Mbp (figure 13). The top candidate mediator for Apoe expression was TBC1D15 protein expression, which dropped the interaction effect LOD score by 4.6877. The Tbcd1d15 gene is located slightly to the left of the peak, at 115.224685 Mbp. Apoe expression differences between the sexes depended on genotype. The AA genotype was associated with more expression in males, while the AB and BB genotypes were associated with the reverse pattern. The expression by sex and genotype pattern for Ryr2 was similar to that of Apoe, with the A allele increasing relative expression in males and the B allele decreasing relative expression in males.

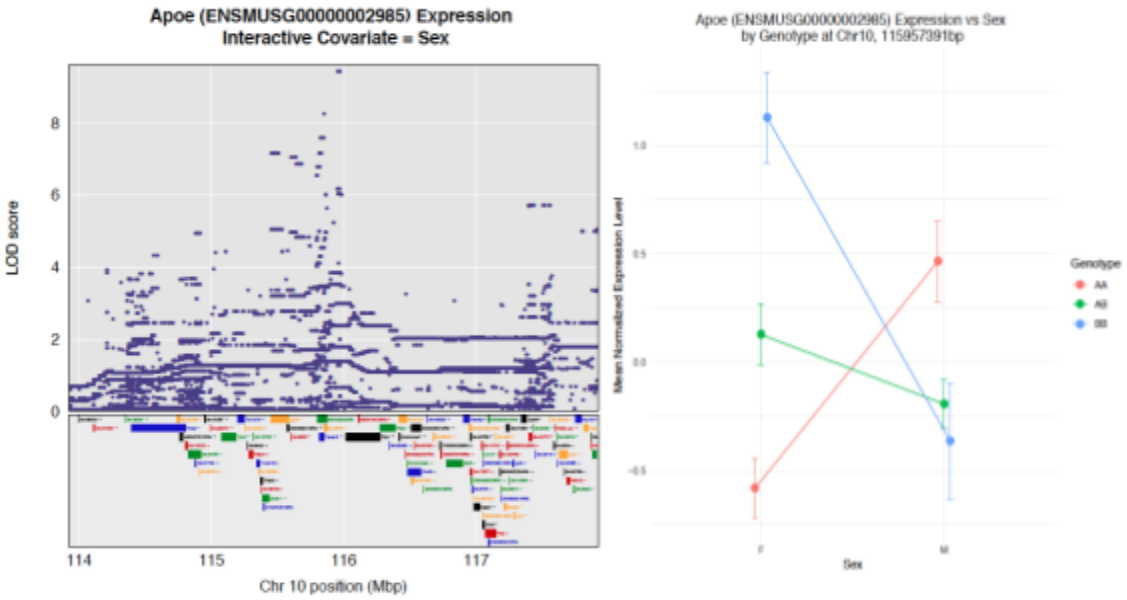

Figure 13. (Left) SNP association mapping at the Apoe sex-interactive eQTL shows a large and sharp peak with significant SNPs. (Right) Sex-dependent genotype effects at the SNP most significantly associated with the interaction effect shows that the AA genotype results in males having the highest expression levels and females having the lowest, while the BB genotype is associated with the opposite effect. The heterozygotes show intermediate levels of expression in both males and females.

Conclusion

We have determined that lincRNAs may influence gene and protein expression in cardiac muscle in an age-dependent manner. However, the transcript expression patterns by age and genotype do not always follow the same patterns that the protein expression patterns do. Thus, more work is needed to workout the relationship that lincRNAs have with structural heart proteins and aging. Future steps may include knocking out lincRNAs that lie underneath many of the QTLs from the age-interactive hotspot on chromosome 11 to test if they affect how proteins change with age in the heart.

Acknowledgements

This Summer Student Fellowship was supported by Educational Gifts to The Jackson Laboratory. I would also like to thank the people in the Churchill Lab who helped me throughout this project: Dan Gatti, Ph.D., Gary Churchill, Ph.D *(Principal Investigator),* Susan McClatchy *(Mentor),* and Yuka Takemon.

References

1. de Magalhães JP. From cells to ageing: a review of models and mechanisms of cellular senescence and their impact on human ageing. Exp Cell Res. 2004 Oct 15;300(1):1-10. Review. PubMed PMID: 15383309.
2. Quarles EK, Dai DF, Tocchi A, Basisty N, Gitari L, Rabinovitch PS. Quality control systems in cardiac aging.Ageing Res Rev. 2015 Sep;23(Pt A):101-15. doi: 10.1016/j.arr.2015.02.003. Epub 2015 Feb 19. Review. PubMed PMID: 25702865; PubMed Central PMCID: PMC4686341.
3. "Heart Disease and Stroke Statistics 2018 At-a-Glance." *American Heart Association*, https://healthmetrics.heart.org/wp-content/uploads/2018/02/At-A-Glance-Heart-Disease-and-Stroke-Statistics-2018.pdf. Accessed 15 June 2018.
4. Perls T, Terry D. Understanding the determinants of exceptional longevity. Ann Intern Med. 2003 Sep 2;139(5 Pt 2):445-9. Review. PubMed PMID: 12965974.
5. Newton MA, Wang Z. Multiset Statistics for Gene Set Analysis. Annu Rev Stat Appl. 2015 Apr;2:95-111. PubMed PMID: 25914887; PubMed Central PMCID: PMC4405258.
6. Broman KW, Sen S. A Guide to QTL Mapping with R/qtl. Dordrecht: Springer; 2009.
7. McNally EM, Golbus JR, Puckelwartz MJ. Genetic mutations and mechanisms in dilated cardiomyopathy. J Clin Invest. 2013 Jan;123(1):19-26. doi: 10.1172/JCI62862. Epub 2013 Jan 2. Review. PubMed PMID: 23281406; PubMed Central PMCID: PMC3533274.

8. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. Cell. 2013 Jul 3;154(1):26-46. doi: 10.1016/j.cell.2013.06.020. Review. PubMed PMID: 23827673; PubMed Central PMCID: PMC3924787.
9. Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, Raghupathy N, Svenson KL, Churchill GA, Gygi SP. Defining the consequences of genetic variation on a proteome-wide scale. Nature. 2016 Jun 23;534(7608):500-5. doi: 10.1038/nature18270. Epub 2016 Jun 15. PubMed PMID: 27309819; PubMed Central PMCID: PMC5292866.
10. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012 May;16(5):284-7. doi: 10.1089/omi.2011.0118. Epub 2012 Mar 28. PubMed PMID: 22455463; PubMed Central PMCID: PMC3339379.
11. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics. 2005 May 1;21(9):1943-9. Epub 2005 Jan 12. PubMed PMID: 15647293.
12. Kaushik S, Cuervo AM. Proteostasis and aging. Nat Med. 2015 Dec;21(12):1406-15. doi: 10.1038/nm.4001. Review. PubMed PMID: 26646497.

Extended Data and Information

Mediation Analysis for Interactive QTL Mapping

In the context of QTL mapping, mediation analysis is commonly used to identify the causal genes underneath a QTL peak that is driving the QTL-phenotype association. When using only additive QTL models, each gene under the peak for which expression data is recorded is added to the model as an additive covariate. (Here, we define ‘under’ to be 10Mbp or less, such as in a proteome wide analysis paper by Chick and Munger.) If the LOD score drops below the significance threshold determined for the original QTL scan when the candidate gene expression is added to the model, that gene is said to be a driver gene and to mediate the QTL-phenotype association. Either transcript or protein expression of candidate genes could be used.

For interactive QTL models, mediation analysis can be used to identify genes that are driving the interaction effect. Recall that the difference between the full model and the purely additive model is what quantifies the interaction effect. Therefore, candidate mediator genes (genes that lie under the interactive QTL effect peak) should be tested to see if they reduce this difference. To do this, one simply adds the expression of the candidate gene to both the full and additive models as another additive covariate and then checks if the difference between these two updated models is significantly smaller than the difference between the two models without the candidate gene expression as an additive covariate. Quantifying how much the difference should decrease by to be considered significant is outside the scope of this paper, although permutation testing may be one approach.

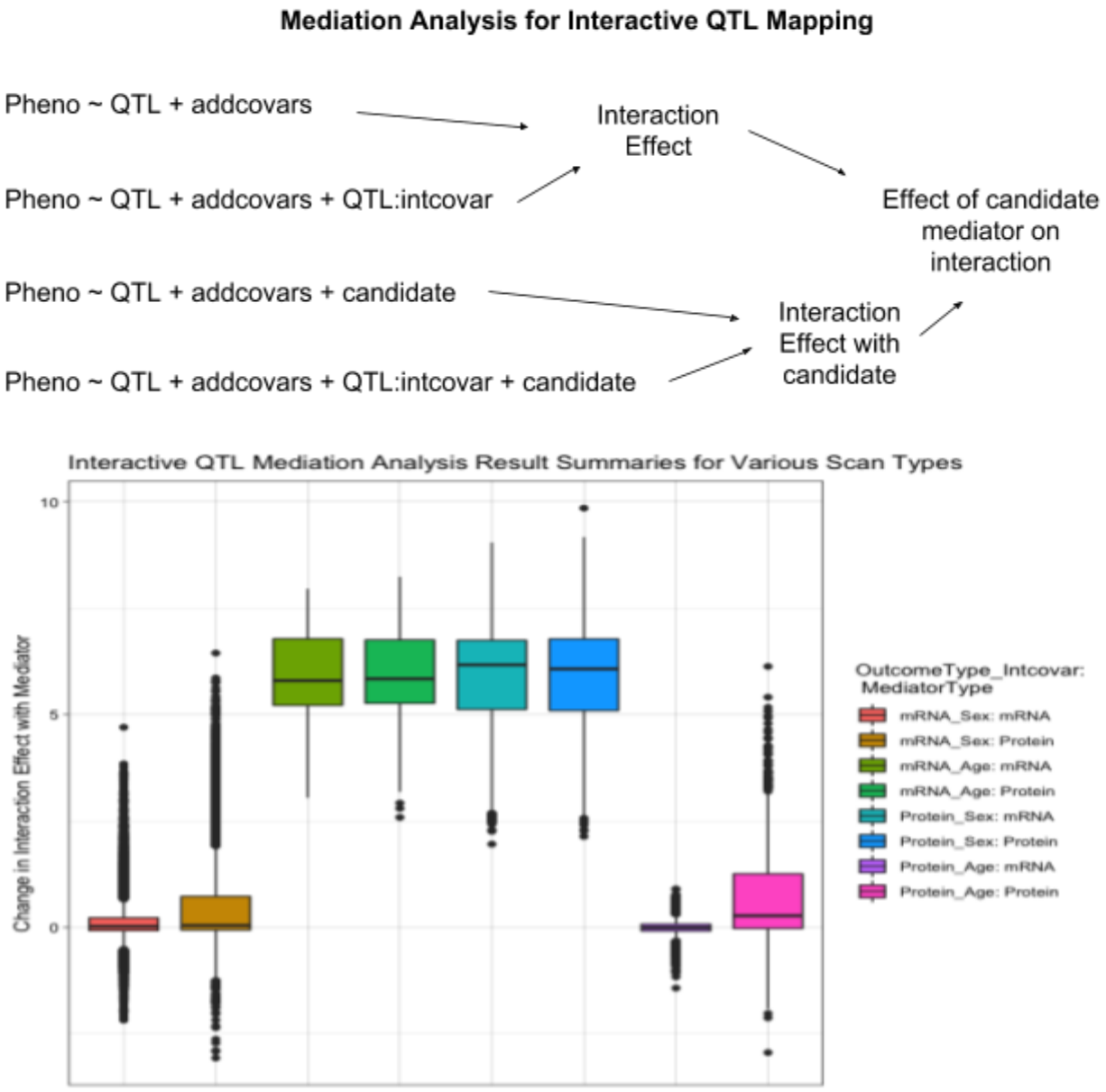

Figure 14. For each outcome gene in each interactive QTL hotspot, we checked all the genes within 10 Mbp of the outcome's QTL to see if they mediated the interaction effect through either their protein or mRNA expression. This graph shows the summary of the mediation results. The y axes is the difference between the LOD score quantifying the interaction effect without the mediator and the LOD score quantifying the interaction effect with the mediator. This graph seems to be showing that eQTL-Age interactions and pQTL-sex are well explained by the expression of nearby genes and proteins, but eQTL-Sex interactions and pQTL-Age interactions are not as much. (Note: We also generated a similar plot where the y-axis was change in interaction effect ratioed on the mean of the interaction effect without the mediator for a given group. Nearly the same pattern appeared.)