DEVELOPING COMPUTATIONAL TOOLS AND DATASETS TO INVESTIGATE THE GENOMIC LOCI ASSOCIATED WITH DISEASE

Nicole E. Kramer

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirement for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology in the School of Medicine.

Chapel Hill
2023

Approved by:

Douglas H. Phanstiel

Karen L. Mohlke

Terry S. Furey

David H. Gotz

Jason L. Stein

© 2023
Nicole E. Kramer
ALL RIGHTS RESERVED

ABSTRACT

Nicole E. Kramer: Developing computational tools and datasets to investigate the genomic loci associated with disease
(Under the direction of Douglas H. Phanstiel)

The majority of genetic variants associated with complex diseases are located in non-coding, regulatory regions of the genome. Understanding the genetic mechanisms of the progression of these diseases has been largely advanced by sequencing-based genomic techniques including RNA-seq, ChIP-seq, Hi-C, genome-wide association studies (GWAS), and Quantitative Trait Locus (QTL) mapping. However, the genetic underpinnings of disease have been difficult to interpret largely because (1) currently available visualization software lacks the ability to efficiently and programmatically integrate large volumes of complex multi-omic data and (2) there are few datasets in disease-relevant cell types in which genomic changes are tracked in response to disease-specific stimuli. In the first part of this work I describe plotgardener, a new R programmatic library for efficiently and reproducibly plotting publication-quality, multi-panel genomic figures. Plotgardener provides customizable genomic plotting and annotation functions that allows users to size and arrange plots in precisely-defined coordinate systems based upon user-defined units of measurement. I include example use cases with plotgardener, both with genomic data and ggplot2 objects, and also have extensively documented and freely available code for the package through Bioconductor and GitHub. I then go on to create and investigate the first response allelic imbalance (AI) and eQTL (reQTL) datasets using an ex vivo model of osteoarthritis (OA) whereby chondrocytes are stimulated with fibronectin fragment (FN-f), a known OA trigger. AI analysis revealed 55 unique genetic variants exhibiting AI at 58 positional genes only after FN-f treatment, with some of these genes exhibiting differential expression. reQTL mapping identified 384 eGenes specific to FN-f treated samples, and colocalization of identified reQTLs with GWAS of various OA phenotypes revealed one robust colocalization of a reQTL with multiple OA phenotypes. I also use plotgardener to visualize these datasets within the context of the genes and linkage disequilibrium (LD) structure of the region. Overall,

these studies have resulted in the creation of a broadly applicable genomic visualization tool and novel datasets to provide critical insights into the genetic basis of osteoarthritis.

For all the women in STEM – we got this.

ACKNOWLEDGMENTS

I have so many amazing people to thank for their endless support throughout and beyond my PhD. First and foremost to my family, particularly my parents, Chandell and Marc: I am so lucky to have such kind, giving, intelligent, hilarious, and encouraging people as my role models guiding me through life. You have always given me the freedom to pursue whatever I wanted and have been with me through every step of the way. Thank you for your advice, words of encouragement, doing my taxes, and attempting to describe my work to people. I love you guys so much. Jared, thanks for reminding me how crazy my work is, even if my paper wasn't actually 100 pages. Ita – you are such an inspiring matriarch and your love always fuels me. Te quiero mucho, mucho, mucho…Poppa – I'm sad you didn't live to see me defend and graduate but I know you've still been rooting for me the whole way. I'm pretty sure I get my love of science and math from you.

Thank you to my absolutely wonderful friends who have become my second family. Mary Kate, I can't believe you've stuck with me since Pre-K but here we are. Your spontaneous cards, flowers, and mini cupcakes have gotten me through some of the toughest parts of my PhD and make me feel like you're right here with me even though you're many miles away. To my echo chamber ladies, Marielle and Kate: you guys are my friend soulmates. I live for our Bachelor nights, Stitch N' Bitch sessions, Pedro Pascal obsessing, Bandido's adventures, Target runs, and just our random shenanigans. You guys are always there for me, make me cry tears of joy, and are some of the only reasons I made it through grad school without losing my mind. I still remember the first day I met you, Marielle, and how nervous I was that you wouldn't want to be my friend as much as I wanted to be yours. Thank you for becoming my sea wife and supporting me every day since. JP, thank you for being one of the kindest and most thoughtful people to ever grace my life. You always put everything in such great perspective and remind me why I got into science in the first place. Thank you for all the coffee, boba, chocolate, and hugs.

To the Phanstiel Lab, past and present, who are not only my labmates but also my friends – you guys are the smartest people I have ever met and I feel so lucky to work with you, learn from you, and

spend time with you. Katie, thank you for being such an inspiration and role model. Not only are you a brilliant scientist, but are such a shining light of joy. Sarah, you are one of the coolest people I know. Thank you for sharing your amazing ice cream flavors, your vast selection of games, and your endless knowledge of random facts. Jess, thank you for making me feel like one of the funniest people ever. I live for your laughter and absolutely love working with you. Zack, you are so absolutely hilarious and never fail to make me feel better about everything. Eric (and Amelia, you're married into the lab), I cannot imagine going through grad school with anyone else but you as my twin. I know we bicker like siblings, but I hope you know how much I look up to you and respect you. I have learned so much about coding from you, and it's always been such a relief to have a friend like you to confide in along the way. Here's to us, we did it, buddy.

Thank you so much to all my mentors at UNC. To Doug: thank you so much for taking me on as a student and training me to become a thoughtful computational scientist. Thank you for always encouraging me to go for the talk or award, giving me an extra boost when grad school was wearing me down, and giving me opportunities to gain new skills. I feel so lucky to have a PI like you who focuses on science but puts more focus on the people behind the science. To Karen: I have learned so much from you, not just about science but also about being a female scientific leader. Thank you for being my committee chair, allowing me into your lab meetings, and providing endless support and advice. To my committee, David, Jason, and Terry: thank you for being my biggest advocates and your guidance throughout the PhD journey. Thank you to Will Valdar and John Cornett for all their BCB support. The BCB community is so lucky to have such wonderful people leading us.

Lastly, thank you to some random things that have been with me throughout my PhD: Taylor Swift (for good tears and sad tears), all the dogs I pet, Gordon Ramsay (I would not have finished my dissertation without Hell's Kitchen), mini Squishmallows, yarn (knitting is the best stress reliever), and GoGo Squeez applesauce (the best grad school snack).

PREFACE

This work was supported in part by a grant from the National Institute of General Medical Sciences under award 5T32 GM067553.

Chapter 2 has been previously published with the following citation: Kramer, N.E., Davis, E.S., Wenger, C.D., Deoudes, E.M., Parker, S.M., Love, M. I., & Phanstiel, D. H. (2022) Plotgardener: cultivating precise multi-panel figures in R. Bioinformatics 38(7): 2042-2045. doi: 10.1093/bioinformatics/btac057. Other contributors for this work include Eric S. Davis (conceptualization, software, writing), Craig D. Wenger (software), Erika M. Deoudes (visualization), Sarah M. Parker (software), Michael I. Love (software, writing), and Douglas H. Phanstiel (conceptualization, writing, supervision, funding acquisition). Hyejung Won and Jason Stein provided helpful discussions and feedback during software development. Muhammad Saad Shamim and Neva Durand provided assistance with the strawr package. This work was supported by grants from the NIH (D.H.P, R35-GM128645; N.E.K. and E.S.D., T32- GM067553; M.I.L. R01-MH118349, R01-HG009937) and the NSF (S.M.P. GRFP DGE-1650116).

For Chapters 3 and 4, donor tissue was obtained through the Gift of Hope Organ and Tissue Donor Network (Elmhurt, IL). Chondrocyte extraction and isolation was performed by Philip Coryell in the lab of Richard F. Loeser. DNA and RNA extractions were performed by Eliza Thulson and Susan D'Costa. Genotyping was done at the Mammalian Genotyping Core at the University of North Carolina at Chapel Hill. RNA-seq library preparation and sequencing was done at the New York Genome Center. Jason L. Stein, Nil Aygün, and Michael I. Love provided assistance with allelic imbalance analysis. Karen L. Mohlke, Kevin W. Currin, Sarah M. Brotman, Jason L. Stein, Brandon D. Le, and Jordan M. Valone advised in design and implementation of standard eQTL and response eQTL models.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

aFC	Allelic fold change
AI	Allelic imbalance
API	Application programming interface
ASE	Allele-specific expression
bp	Base pairs
caQTL	Chromatin accessibility quantitative trait loci
ChIP	Chromatin immunoprecipitation
CRAN	Comprehensive R Archive Network
DNA	Deoxyribonucleic acid
eGene	eQTL gene
eQTL	Expression quantitative trait loci
FDR	False discovery rate
FN-f	Fibronectin fragments
GATK	Genome analysis toolkit
GB	Gigabytes
GO	Gene ontology
GTEx	Genome-tissue expression project
GWAS	Genome-wide association study
IGV	Integrative Genomics Viewer
KEGG	Kyoto encyclopedia of genes and genomes
LD	Linkage disequilibrium
Mb	Megabases
OA	Osteoarthritis
PBS	Phosphate buffered saline
PC	Principal component
PCA	Principal component analysis
PEER	Probabilistic estimation of expression residuals

QTL	Quantitative trait loci
reQTL	Response expression quantitative trait loci
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
TSS	Transcription start site

CHAPTER 1: INTRODUCTION

1.1 Gene regulation and disease

Despite the presence of numerous specialized cell types carrying out specific functions within the human body, the same sequence of DNA base pairs makes up the functional blueprint of every cell in an individual. The complexities of this system lie in its overall regulation – which genes are expressed and when, how much are they expressed, and how does their expression change in different environments? The key players of this regulation, including enhancers, promoters, repressors, insulators, and other transcription factor binding sites (Levine & Tjian, 2003), mostly reside in the vast non-coding genome (Alexander et al., 2010; ENCODE Project Consortium, 2012). Although genome-wide efforts have sought to understand the complicated interplay and dynamics of this system, many aspects of genetic regulation are still poorly understood.

Different individuals, particularly from different populations, have many single base pair variations in their genomes, termed single nucleotide polymorphisms (SNPs). It is estimated that there are at least 11 million SNPs in the human genome (1000 Genomes Project Consortium et al., 2015), which can in turn alter its regulation and function. Certain variations can result in single gene disorders like cystic fibrosis and sickle-cell anemia, whereby a mutation at a single gene results in the diseased phenotype (Spataro et al., 2017). However, most traits and diseases are complex and result from the interaction of several genomic loci housing susceptibility alleles and various environmental factors (Almouzni et al., 2014; Jackson et al., 2018). Furthermore, much of this disease-associated genetic variation resides in the non-coding genome and is thought to alter the complex transcriptional regulatory landscape and lead to the misregulation of gene expression (Gonzaga-Jauregui et al., 2012; Lee & Young, 2013; Zhang & Lupski, 2015). While recent studies have made significant progress to disentangle the genetics of numerous diseases, it still remains difficult to fully understand the mechanisms of causal genes and variants involved in disease susceptibility. Understanding disease-related genetic variation in different populations would provide novel avenues for preventative treatments and curing therapies.

Osteoarthritis (OA) is one such polygenic disease that requires detailed genetic investigation in order to identify targets for therapeutic intervention. OA is a progressive joint disease that affects over 250 million people worldwide and is one of the leading causes of disability and pain (Hunter & Bierma-Zeinstra, 2019; Loeser et al., 2012). There are currently no treatments other than symptom management or end-stage joint replacements, which leaves gene regulation mechanisms as a lucrative space for identifying novel disease-modifying treatments. Thus far, genome-wide association studies (GWAS) have been one of the primary methods in studying the effects of genetic variation on OA phenotypes. These GWAS involve mapping alleles at SNPs in controls and cases, which can encompass OA at any skeletal joint (primarily the knees and hips) both by joint stratification and combining case types. To date, a combination of numerous rigorous OA GWAS has resulted in 124 SNPs significantly associated with OA spanning 95 independent loci across the genome (arcOGEN Consortium et al., 2012; Boer et al., 2021; Casalone et al., 2018; Castaño Betancourt et al., 2012; Castaño-Betancourt et al., 2016; Day-Williams et al., 2011; den Hollander et al., 2017; Evangelou et al., 2013, 2014; Evans et al., 2015; Hackinger et al., 2017; Kerkhof et al., 2010; Liu et al., 2017; Miyamoto et al., 2007, 2008; Nakajima et al., 2010; Panoutsopoulou et al., 2017; Styrkarsdottir et al., 2014, 2017, 2018; Tachmazidou et al., 2019; Valdes et al., 2011; Yau et al., 2017; Zengini et al., 2018). These studies have identified broad genetic risk loci that provide the starting point for disentangling OA etiology but do not reveal the mechanistic impact of causal non-coding variants on OA gene regulation.

1.2 Challenges of studying disease-relevant genetic regulation

The increasing sample sizes and diversifying populations of GWAS has revealed more loci associated with traits and diseases, but it still remains challenging to translate these results into causal variants and their molecular mechanisms for a variety of reasons. The linked inheritance of nearby SNPs, termed linkage disequilibrium (LD), results in their correlation with each other and thus makes it difficult to distinguish the actual causal variant. In numerous diseases, it is also unclear which cell types these SNPs act in to drive disease versus which cell types are merely affected as a consequence of disease. Lastly, more than 90% of disease associated variants lie in the non-coding genome and indirectly regulate the expression of one or many genes (Cano-Gamez & Trynka, 2020). Thus, the natural next step for understanding the biological relevance of these results is to generate molecular datasets in a disease-

relevant system and integrate them with GWAS to prioritize variants. Several studies across various traits have used integrative, multi-omic approaches to identify genes impacted by distal, non-coding SNPs. For example, one study used epigenomics, comparative genomics, human genetics, genome editing, and sample perturbations to disentangle the mechanisms of the FTO locus and its association to obesity. The C risk allele at rs1421085 was found to disrupt a conserved motif for the ARID5B repressor, which led to an increase in IRX3 and IRX5 expression during adipocyte differentiation and in turn shifted adipocytes towards the energy-storing white adipocyte development (Claussnitzer et al., 2015). From this example it is clear that multiple layers of genomic and genetic data are required to understand the effects of non-coding variants on a phenotype.

Beyond the complications of interpreting GWAS results, studying gene regulation in OA in particular has been difficult due to the heterogeneity of the disease. In addition to genetic factors, OA is highly influenced by various environmental factors including age or degree of wear and tear of the joint. Furthermore, symptoms can be highly variable and involve single or multiple joints (Abramoff & Caldera, 2020). Thus, it is hard to distinguish genetic effects and their response to disease stimuli when interrogating the effects of genetic variation on gene expression and regulation. Possible solutions to studying the disease include in vivo with animal models, which lack full control in fine-tuning and probing genetic-specific effects, and in vitro models, which diminish the biological relevance of a study when cells are taken out of the disease context. Ex vivo models provide an opportunity to study OA through articular chondrocytes, which are the only cells found in cartilage and are associated with OA, isolated from healthy donor tissue and stimulated with an extracellular matrix component (fibronectin fragments, or FN-f) found in degrading cartilage and shown to initiate a positive feedback loop of matrix destruction (Forsyth et al., 2002; Homandberg, 1999; Homandberg et al., 1998; Pulai et al., 2005; Xie et al., 1992). This system allows for the control of joint location, OA severity, and environmental factors to increase power in identifying causal genetic variation and its effect on gene regulation throughout an OA-like response. The FN-f chondrocyte model of OA has been characterized as an effective model for recapitulating OA transcriptomics (Reed et al., 2021), so Chapters 3 and 4 leverage this fine-tuned system to study OA gene regulation mechanisms in a disease-relevant context.

1.3 Genomic datasets for investigating disease-associated genetic loci

Many genome-wide, sequencing-based assays have been used to study the complexities of non-coding genomic features and gain mechanistic information about these loci. Measures of chromatin accessibility, gene expression, 3D chromatin contact frequencies, and histone marks for transcription factor activity functionally annotate the genome, but do not usually connect genetic variation to these readouts. Quantitative trait loci (QTL) mapping is a powerful statistical method to understand associations between genetic variants and molecular markers. In particular, expression QTLs (eQTLs) link SNPs to gene regulatory mechanisms at specific genes, and integrating these datasets with GWAS can serve as a way to connect genetic effect on molecular traits with complex phenotypes, making this a useful method to identify the genes involved in traits and diseases. Numerous studies and large consortiums like the Genotype Tissue Expression project (GTEx) (GTEx Consortium, 2013) have mapped eQTLs in a vast array of tissues and in increasingly large sample sizes, but many of these identified eQTLs are shared across tissues and do not colocalize with disease-associated GWAS loci (GTEx Consortium, 2020). One hypothesis for this lack of colocalization is that disease variants must be studied in the correct cell type and genetic regulation in this cell type may only be detected after a disease-relevant stimulus (Umans et al., 2021). For example, one previous study identified eQTLs involved in the macrophage immune response that colocalized with disease risk loci only after stimulation with IFNγ and/or Salmonella (Alasoo et al., 2018).

Despite this finding, there are limited numbers of disease response-specific genomic datasets, particularly in the study of OA. Only one study has leveraged the chondrocyte FN-f model of OA to prioritize OA genetic variants through Hi-C DNA looping information (Thulson et al., 2022). However, OA integrative epigenomics and QTL studies have only focused on steady-state disease conditions to map methylation, protein, and expression QTLs, primarily in diseased bulk cartilage or chondrocytes extracted from diseased cartilage alone (Kreitmaier et al., 2022; Steinberg et al., 2017, 2021). While the identified QTLs have revealed compelling colocalizations with OA GWAS, there are still many loci that are yet to colocalize with OA-relevant QTLs that could potentially be resolved with stimulated cell-type-specific QTL datasets. In Chapter 4, I map the first set of OA response eQTLs (reQTLs) using the ex vivo OA model of FN-f perturbed chondrocytes to identify novel putative OA risk genes.

In addition to understanding the effects of genetic variation on total gene expression through eQTLs, we can use the relative expression of gene alleles to further investigate allele-specific effects of causal variants on gene expression. Allelic imbalance (AI), or allele-specific expression (ASE), refers to the unequal expression of alleles of a gene in diploid individuals, which is driven by genetic variants located in or near the gene. This analysis can be used as a complementary approach to eQTL mapping to characterize the genetic basis of gene expression. Compared to eQTL studies, AI can be investigated with smaller donor sample sizes with higher power because the two measured alleles are expressed within the same individual, controlling the genetic and environmental background (Almlöf et al., 2012). AI OA studies, like OA QTL studies, have also only been conducted in bulk and/or steady-state diseased OA tissues thus far (Bos et al., 2012; Coutinho de Almeida et al., 2022; den Hollander et al., 2019; Gee et al., 2014; Raine et al., 2013; Reynard et al., 2014; Southam et al., 2007). In Chapter 3, I generate a complementary AI dataset in the OA FN-f model system to capture stimulus-specific and chondrocyte-specific AI variants and their regulation of gene expression.

1.4 Genomic data visualization as a tool for data interpretation

The increasing need for larger and more complex genomic datasets to study genetic regulation in disease has created a dire need for computational tools that can efficiently parse, interpret, and visualize these data. Response eQTL and AI datasets, while informative on their own, are further strengthened by integration with other genomic data, including GWAS, chromatin accessibility, DNA looping, and gene annotations. Improved tools for handling the large-scale landscape of studying the regulatory genome increase the reproducibility and communication of scientific findings as well as contribute to the forming of novel hypotheses that can contribute to understanding the genetic basis of disease.

Numerous tools have already revolutionized our ability to interpret complex scientific data. Genomic browsers like the Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al., 2013), UCSC Genome Browser (Kent et al., 2002), and WashU Epigenome Browser (Li et al., 2019) specialize in preliminarily surveying through stacked linear genomic tracks for quickly exploring genomic regions of interest. The introduction of chromatin conformation assays like Hi-C has led to genomic browsers specialized in visualizing 3D data representations like Juicebox (Durand et al., 2016) and HiGlass (Kerpedjiev et al., 2018). Despite their contribution to the scientific process, these browsers are poorly

suited for integration with data analysis workflows and creating easily reproducible visualizations with code. Programmatic libraries have increased the reproducibility and automation of genomic figure generation, particularly through bioinformatic tools and infrastructure developed by the Bioconductor project (Gentleman et al., 2004) for the R programming language (R Core Team, n.d.). The Bioconductor project has established widely used data structures, packages for analysis and visualization, and various workflows that have been instrumental in analyzing genomic datasets. For example, almost all studies studying differential gene expression use edgeR (Robinson et al., 2010) or DESeq2 (Love et al., 2014), which are both a part of Bioconductor.

Although several visualization packages are available through Bioconductor or the Comprehensive R Archive Network (CRAN), these packages do not effectively integrate, align, and arrange multiple large-scale genomic data types into multi-panel, publication-quality figures. Genomic-specific visualization packages like rtracklayer (Lawrence et al., 2009) and Gviz (Hahne & Ivanek, 2016) are best suited for plotting single-paneled tracks of small regions of data, which can be difficult to scale to look at the increasing number of datasets and disease-associated loci. Packages that can make and arrange multi-paneled plots like cowplot (Wilke, 2020), egg (Auguie, 2019), gridExtra (Auguie, 2017), multipanelfigure (Graumann & Cotton, 2018), patchwork (Pedersen, 2020), and Sushi (Phanstiel et al., 2014) give users little control over the exact sizing and arrangement of figures by using relative positioning and limiting arrangements to rigid grid-style layouts. Figures generated with these libraries often require finishing with graphic design software, which decreases the reproducibility of using computational visualization methods.

In Chapter 2, I describe plotgardener, an R/Bioconductor genomic visualization package for building entirely programmatic, complex multi-panel genomic figures. Plotgardener uses an absolute, coordinate-based system for sizing and arranging plots and supports a wide range of high-throughput genomic data. I have written extensive documentation illustrating plotgardener's numerous use cases, and I have used it for generating figures for my own datasets in Chapters 3 and 4.

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Abramoff, B., & Caldera, F. E. (2020). Osteoarthritis: Pathology, Diagnosis, and Treatment Options. *The Medical Clinics of North America*, *104*(2), 293–311.

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., HIPSCI Consortium, Hale, C., Dougan, G., & Gaffney, D. J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature Genetics*, *50*(3), 424–431.

Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M., & Gerstein, M. B. (2010). Annotating non-coding regions of the genome. *Nature Reviews. Genetics*, *11*(8), 559–571.

Almlöf, J. C., Lundmark, P., Lundmark, A., Ge, B., Maouche, S., Göring, H. H. H., Liljedahl, U., Enström, C., Brocheton, J., Proust, C., Godefroy, T., Sambrook, J. G., Jolley, J., Crisp-Hihn, A., Foad, N., Lloyd-Jones, H., Stephens, J., Gwilliam, R., Rice, C. M., … Syvänen, A.-C. (2012). Powerful identification of cis-regulatory SNPs in human primary monocytes using allele-specific gene expression. *PloS One*, *7*(12), e52260.

Almouzni, G., Altucci, L., Amati, B., Ashley, N., Baulcombe, D., Beaujean, N., Bock, C., Bongcam-Rudloff, E., Bousquet, J., Braun, S., Bressac-de Paillerets, B., Bussemakers, M., Clarke, L., Conesa, A., Estivill, X., Fazeli, A., Grgurević, N., Gut, I., Heijmans, B. T., … Widschwendter, M. (2014). Relationship between genome and epigenome--challenges and requirements for future research. *BMC Genomics*, *15*(1), 487.

arcOGEN Consortium, arcOGEN Collaborators, Zeggini, E., Panoutsopoulou, K., Southam, L., Rayner, N. W., Day-Williams, A. G., Lopes, M. C., Boraska, V., Esko, T., Evangelou, E., Hoffman, A., Houwing-Duistermaat, J. J., Ingvarsson, T., Jonsdottir, I., Jonnson, H., Kerkhof, H. J., Kloppenburg, M., Bos, S. D., … Loughlin, J. (2012). Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study. *The Lancet*, *380*(9844), 815–823.

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra

Auguie, B. (2019). *egg: Extensions for "ggplot2": Custom Geom, Custom Themes, Plot Alignment, Labelled Panels, Symmetric Scales, and Fixed Panel Size*. https://CRAN.R-project.org/package=egg

Boer, C. G., Hatzikotoulas, K., Southam, L., Stefánsdóttir, L., Zhang, Y., de Almeida, R. C., Wu, T. T., Zheng, J., Hartley, A., Teder-Laving, M., Skogholt, A. H., Terao, C., Zengini, E., Alexiadis, G., Barysenka, A., Bjornsdottir, G., Gabrielsen, M. E., Gilly, A., Ingvarsson, T., … Zeggini, E. (2021). Deciphering osteoarthritis genetics across 826,690 individuals from 9 populations. *Cell*, *0*(0). https://doi.org/10.1016/j.cell.2021.07.038

Bos, S. D., Bovée, J. V. M. G., Duijnisveld, B. J., Raine, E. V. A., van Dalen, W. J., Ramos, Y. F. M., van der Breggen, R., Nelissen, R. G. H. H., Slagboom, P. E., Loughlin, J., & Meulenbelt, I. (2012). Increased type II deiodinase protein in OA-affected cartilage and allelic imbalance of OA risk polymorphism rs225014 at DIO2 in human OA joint tissues. *Annals of the Rheumatic Diseases*, *71*(7), 1254–1258.

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, *11*, 424.

Casalone, E., Tachmazidou, I., Zengini, E., Hatzikotoulas, K., Hackinger, S., Suveges, D., Steinberg, J., Rayner, N. W., arcOGEN Consortium, Wilkinson, J. M., Panoutsopoulou, K., & Zeggini, E. (2018). A novel variant in GLIS3 is associated with osteoarthritis. *Annals of the Rheumatic Diseases*, *77*(4), 620–623.

Castaño Betancourt, M. C., Cailotto, F., Kerkhof, H. J., Cornelis, F. M. F., Doherty, S. A., Hart, D. J., Hofman, A., Luyten, F. P., Maciewicz, R. A., Mangino, M., Metrustry, S., Muir, K., Peters, M. J., Rivadeneira, F., Wheeler, M., Zhang, W., Arden, N., Spector, T. D., Uitterlinden, A. G., … van Meurs, J. B. J. (2012). Genome-wide association and functional studies identify the DOT1L gene to be involved in cartilage thickness and hip osteoarthritis. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(21), 8218–8223.

Castaño-Betancourt, M. C., Evans, D. S., Ramos, Y. F. M., Boer, C. G., Metrustry, S., Liu, Y., den Hollander, W., van Rooij, J., Kraus, V. B., Yau, M. S., Mitchell, B. D., Muir, K., Hofman, A., Doherty, M., Doherty, S., Zhang, W., Kraaij, R., Rivadeneira, F., Barrett-Connor, E., … van Meurs, J. B. J. (2016). Novel Genetic Variants for Cartilage Thickness and Hip Osteoarthritis. *PLoS Genetics*, *12*(10), e1006260.

Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I. S., Beaudry, J. L., Puviindran, V., Abdennur, N. A., Liu, J., Svensson, P.-A., Hsu, Y.-H., Drucker, D. J., Mellgren, G., Hui, C.-C., Hauner, H., & Kellis, M. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England Journal of Medicine*, *373*(10), 895–907.

Coutinho de Almeida, R., Tuerlings, M., Ramos, Y., Den Hollander, W., Suchiman, E., Lakenberg, N., Nelissen, R. G. H. H., Mei, H., & Meulenbelt, I. (2022). Allelic expression imbalance in articular cartilage and subchondral bone refined genome-wide association signals in osteoarthritis. *Rheumatology* . https://doi.org/10.1093/rheumatology/keac498

Day-Williams, A. G., Southam, L., Panoutsopoulou, K., Rayner, N. W., Esko, T., Estrada, K., Helgadottir, H. T., Hofman, A., Ingvarsson, T., Jonsson, H., Keis, A., Kerkhof, H. J. M., Thorleifsson, G., Arden, N. K., Carr, A., Chapman, K., Deloukas, P., Loughlin, J., McCaskie, A., … Zeggini, E. (2011). A variant in MCF2L is associated with osteoarthritis. *American Journal of Human Genetics*, *89*(3), 446–450.

den Hollander, W., Boer, C. G., Hart, D. J., Yau, M. S., Ramos, Y. F. M., Metrustry, S., Broer, L., Deelen, J., Cupples, L. A., Rivadeneira, F., Kloppenburg, M., Peters, M., Spector, T. D., Hofman, A., Slagboom, P. E., Nelissen, R. G. H. H., Uitterlinden, A. G., Felson, D. T., Valdes, A. M., … van Meurs, J. J. B. (2017). Genome-wide association and functional studies identify a role for matrix Gla protein in osteoarthritis of the hand. *Annals of the Rheumatic Diseases*, *76*(12), 2046–2053.

den Hollander, W., Pulyakhina, I., Boer, C., Bomer, N., van der Breggen, R., Arindrarto, W., Couthino de Almeida, R., Lakenberg, N., Sentner, T., Laros, J. F. J., 't Hoen, P. A. C., Slagboom, E. P. E., Nelissen, R. G. H. H., van Meurs, J., Ramos, Y. F. M., & Meulenbelt, I. (2019). Annotating Transcriptional Effects of Genetic Variants in Disease-Relevant Tissue: Transcriptome-Wide Allelic Imbalance in Osteoarthritic Cartilage. *Arthritis & Rheumatology (Hoboken, N.J.)*, *71*(4), 561–570.

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems*, *3*(1), 99–101.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74.

Evangelou, E., Kerkhof, H. J., Styrkarsdottir, U., Ntzani, E. E., Bos, S. D., Esko, T., Evans, D. S., Metrustry, S., Panoutsopoulou, K., Ramos, Y. F. M., Thorleifsson, G., Tsilidis, K. K., arcOGEN Consortium, Arden, N., Aslam, N., Bellamy, N., Birrell, F., Blanco, F. J., Carr, A., … Valdes, A. M. (2014). A meta-analysis of genome-wide association studies identifies novel variants associated with osteoarthritis of the hip. *Annals of the Rheumatic Diseases*, *73*(12), 2130–2136.

Evangelou, E., Valdes, A. M., Castano-Betancourt, M. C., Doherty, M., Doherty, S., Esko, T., Ingvarsson, T., Ioannidis, J. P. A., Kloppenburg, M., Metspalu, A., Ntzani, E. E., Panoutsopoulou, K., Slagboom, P. E., Southam, L., Spector, T. D., Styrkarsdottir, U., Stefanson, K., Uitterlinden, A. G., Wheeler, M., … arcOGEN consortium, the TREAT-OA consortium. (2013). The DOT1L rs12982744 polymorphism is

associated with osteoarthritis of the hip with genome-wide statistical significance in males. *Annals of the Rheumatic Diseases*, *72*(7), 1264–1265.

Evans, D. S., Cailotto, F., Parimi, N., Valdes, A. M., Castaño-Betancourt, M. C., Liu, Y., Kaplan, R. C., Bidlingmaier, M., Vasan, R. S., Teumer, A., Tranah, G. J., Nevitt, M. C., Cummings, S. R., Orwoll, E. S., Barrett-Connor, E., Renner, J. B., Jordan, J. M., Doherty, M., Doherty, S. A., … Lane, N. E. (2015). Genome-wide association and functional studies identify a role for IGFBP3 in hip osteoarthritis. *Annals of the Rheumatic Diseases*, *74*(10), 1861–1867.

Forsyth, C. B., Pulai, J., & Loeser, R. F. (2002). Fibronectin fragments and blocking antibodies to alpha2beta1 and alpha5beta1 integrins stimulate mitogen-activated protein kinase signaling and increase collagenase 3 (matrix metalloproteinase 13) production by human articular chondrocytes. *Arthritis and Rheumatism*, *46*(9), 2368–2376.

Gee, F., Clubbs, C. F., Raine, E. V. A., Reynard, L. N., & Loughlin, J. (2014). Allelic expression analysis of the osteoarthritis susceptibility locus that maps to chromosome 3p21 reveals cis-acting eQTLs at GNL3 and SPCS1. *BMC Medical Genetics*, *15*, 53.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., … Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80.

Gonzaga-Jauregui, C., Lupski, J. R., & Gibbs, R. A. (2012). Human genome sequencing in health and disease. *Annual Review of Medicine*, *63*, 35–61.

Graumann, J., & Cotton, R. (2018). multipanelfigure: Simple Assembly of Multiple Plots and Images into a Compound Figure. In *Journal of Statistical Software, Code Snippets* (Vol. 84, Issue 3, pp. 1–10). https://doi.org/10.18637/jss.v084.c03

GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585.

GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318–1330.

Hackinger, S., Trajanoska, K., Styrkarsdottir, U., Zengini, E., Steinberg, J., Ritchie, G. R. S., Hatzikotoulas, K., Gilly, A., Evangelou, E., Kemp, J. P., arcOGEN Consortium, GEFOS Consortium, Evans, D., Ingvarsson, T., Jonsson, H., Thorsteinsdottir, U., Stefansson, K., McCaskie, A. W., Brooks, R. A., Wilkinson, J. M., … Zeggini, E. (2017). Evaluation of shared genetic aetiology between osteoarthritis and bone mineral density identifies SMAD3 as a novel osteoarthritis risk locus. *Human Molecular Genetics*, *26*(19), 3850–3858.

Hahne, F., & Ivanek, R. (2016). Visualizing Genomic Data Using Gviz and Bioconductor. In E. Mathé & S. Davis (Eds.), *Statistical Genomics: Methods and Protocols* (pp. 335–351). Springer New York.

Homandberg, G. A. (1999). Potential regulation of cartilage metabolism in osteoarthritis by fibronectin fragments. *Frontiers in Bioscience: A Journal and Virtual Library*, *4*, D713–D730.

Homandberg, G. A., Wen, C., & Hui, F. (1998). Cartilage damaging activities of fibronectin fragments derived from cartilage and synovial fluid. *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society*, *6*(4), 231–244.

Hunter, D. J., & Bierma-Zeinstra, S. (2019). Osteoarthritis. *The Lancet*, *393*(10182), 1745–1759.

Jackson, M., Marks, L., May, G. H. W., & Wilson, J. B. (2018). The genetic basis of disease. *Essays in Biochemistry*, *62*(5), 643–723.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006.

Kerkhof, H. J. M., Lories, R. J., Meulenbelt, I., Jonsdottir, I., Valdes, A. M., Arp, P., Ingvarsson, T., Jhamai, M., Jonsson, H., Stolk, L., Thorleifsson, G., Zhai, G., Zhang, F., Zhu, Y., van der Breggen, R., Carr, A., Doherty, M., Doherty, S., Felson, D. T., … van Meurs, J. B. J. (2010). A genome-wide association study identifies an osteoarthritis susceptibility locus on chromosome 7q22. *Arthritis and Rheumatism*, *62*(2), 499–510.

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J. M., Ouellette, S. B., Azhir, A., Kumar, N., Hwang, J., Lee, S., Alver, B. H., Pfister, H., Mirny, L. A., Park, P. J., & Gehlenborg, N. (2018). HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biology*, *19*(1), 125.

Kreitmaier, P., Suderman, M., Southam, L., Coutinho de Almeida, R., Hatzikotoulas, K., Meulenbelt, I., Steinberg, J., Relton, C. L., Wilkinson, J. M., & Zeggini, E. (2022). An epigenome-wide view of osteoarthritis in primary tissues. *American Journal of Human Genetics*, *109*(7), 1255–1271.

Lawrence, M., Gentleman, R., & Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* , *25*(14), 1841–1842.

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, *152*(6), 1237–1251.

Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, *424*(6945), 147–151.

Li, D., Hsu, S., Purushotham, D., Sears, R. L., & Wang, T. (2019). WashU Epigenome Browser update 2019. *Nucleic Acids Research*, *47*(W1), W158–W165.

Liu, Y., Yau, M. S., Yerges-Armstrong, L. M., Duggan, D. J., Renner, J. B., Hochberg, M. C., Mitchell, B. D., Jackson, R. D., & Jordan, J. M. (2017). Genetic Determinants of Radiographic Knee Osteoarthritis in African Americans. *The Journal of Rheumatology*, *44*(11), 1652–1658.

Loeser, R. F., Goldring, S. R., Scanzello, C. R., & Goldring, M. B. (2012). Osteoarthritis: a disease of the joint as an organ. *Arthritis and Rheumatism*, *64*(6), 1697–1707.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Miyamoto, Y., Mabuchi, A., Shi, D., Kubo, T., Takatori, Y., Saito, S., Fujioka, M., Sudo, A., Uchida, A., Yamamoto, S., Ozaki, K., Takigawa, M., Tanaka, T., Nakamura, Y., Jiang, Q., & Ikegawa, S. (2007). A functional polymorphism in the 5' UTR of GDF5 is associated with susceptibility to osteoarthritis. *Nature Genetics*, *39*(4), 529–533.

Miyamoto, Y., Shi, D., Nakajima, M., Ozaki, K., Sudo, A., Kotani, A., Uchida, A., Tanaka, T., Fukui, N., Tsunoda, T., Takahashi, A., Nakamura, Y., Jiang, Q., & Ikegawa, S. (2008). Common variants in DVWA on chromosome 3p24.3 are associated with susceptibility to knee osteoarthritis. *Nature Genetics*, *40*(8), 994–998.

Nakajima, M., Takahashi, A., Kou, I., Rodriguez-Fontenla, C., Gomez-Reino, J. J., Furuichi, T., Dai, J., Sudo, A., Uchida, A., Fukui, N., Kubo, M., Kamatani, N., Tsunoda, T., Malizos, K. N., Tsezou, A., Gonzalez, A., Nakamura, Y., & Ikegawa, S. (2010). New sequence variants in HLA class II/III region associated with susceptibility to knee osteoarthritis identified by genome-wide association study. *PloS One*, *5*(3), e9723.

Panoutsopoulou, K., Thiagarajah, S., Zengini, E., Day-Williams, A. G., Ramos, Y. F., Meessen, J. M., Huetink, K., Nelissen, R. G., Southam, L., Rayner, N. W., arcOGEN Consortium, Doherty, M., Meulenbelt,

I., Zeggini, E., & Wilkinson, J. M. (2017). Radiographic endophenotyping in hip osteoarthritis improves the precision of genetic association analysis. *Annals of the Rheumatic Diseases*, *76*(7), 1199–1206.

Pedersen, T. L. (2020). *patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork

Phanstiel, D. H., Boyle, A. P., Araya, C. L., & Snyder, M. P. (2014). Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* , *30*(19), 2808–2810.

Pulai, J. I., Chen, H., Im, H.-J., Kumar, S., Hanning, C., Hegde, P. S., & Loeser, R. F. (2005). NF-kappa B mediates the stimulation of cytokine and chemokine expression by human articular chondrocytes in response to fibronectin fragments. *Journal of Immunology* , *174*(9), 5781–5788.

Raine, E. V. A., Dodd, A. W., Reynard, L. N., & Loughlin, J. (2013). Allelic expression analysis of the osteoarthritis susceptibility gene COL11A1 in human joint tissues. *BMC Musculoskeletal Disorders*, *14*, 85.

R Core Team. (n.d.). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org

Reed, K. S. M., Ulici, V., Kim, C., Chubinskaya, S., Loeser, R. F., & Phanstiel, D. H. (2021). Transcriptional response of human articular chondrocytes treated with fibronectin fragments: an in vitro model of the osteoarthritis phenotype. *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society*, *29*(2), 235–247.

Reynard, L. N., Bui, C., Syddall, C. M., & Loughlin, J. (2014). CpG methylation regulates allelic expression of GDF5 by modulating binding of SP1 and SP3 repressor proteins to the osteoarthritis susceptibility SNP rs143383. *Human Genetics*, *133*(8), 1059–1073.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , *26*(1), 139–140.

Southam, L., Rodriguez-Lopez, J., Wilkins, J. M., Pombo-Suarez, M., Snelling, S., Gomez-Reino, J. J., Chapman, K., Gonzalez, A., & Loughlin, J. (2007). An SNP in the 5'-UTR of GDF5 is associated with osteoarthritis susceptibility in Europeans and with in vivo differences in allelic expression in articular cartilage. *Human Molecular Genetics*, *16*(18), 2226–2232.

Spataro, N., Rodríguez, J. A., Navarro, A., & Bosch, E. (2017). Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Human Molecular Genetics*, *26*(3), 489–500.

Steinberg, J., Ritchie, G. R. S., Roumeliotis, T. I., Jayasuriya, R. L., Clark, M. J., Brooks, R. A., Binch, A. L. A., Shah, K. M., Coyle, R., Pardo, M., Le Maitre, C. L., Ramos, Y. F. M., Nelissen, R. G. H. H., Meulenbelt, I., McCaskie, A. W., Choudhary, J. S., Wilkinson, J. M., & Zeggini, E. (2017). Integrative epigenomics, transcriptomics and proteomics of patient chondrocytes reveal genes and pathways involved in osteoarthritis. *Scientific Reports*, *7*(1), 8935.

Steinberg, J., Southam, L., Roumeliotis, T. I., Clark, M. J., Jayasuriya, R. L., Swift, D., Shah, K. M., Butterfield, N. C., Brooks, R. A., McCaskie, A. W., Bassett, J. H. D., Williams, G. R., Choudhary, J. S., Wilkinson, J. M., & Zeggini, E. (2021). A molecular quantitative trait locus map for osteoarthritis. *Nature Communications*, *12*(1), 1309.

Styrkarsdottir, U., Helgason, H., Sigurdsson, A., Norddahl, G. L., Agustsdottir, A. B., Reynard, L. N., Villalvilla, A., Halldorsson, G. H., Jonasdottir, A., Magnusdottir, A., Oddson, A., Sulem, G., Zink, F., Sveinbjornsson, G., Helgason, A., Johannsdottir, H. S., Helgadottir, A., Stefansson, H., Gretarsdottir, S., … Stefansson, K. (2017). Whole-genome sequencing identifies rare genotypes in COMP and CHADL associated with high risk of hip osteoarthritis. *Nature Genetics*, *49*(5), 801–805.

Styrkarsdottir, U., Lund, S. H., Thorleifsson, G., Zink, F., Stefansson, O. A., Sigurdsson, J. K., Juliusson, K., Bjarnadottir, K., Sigurbjornsdottir, S., Jonsson, S., Norland, K., Stefansdottir, L., Sigurdsson, A., Sveinbjornsson, G., Oddsson, A., Bjornsdottir, G., Gudmundsson, R. L., Halldorsson, G. H., Rafnar, T., … Stefansson, K. (2018). Meta-analysis of Icelandic and UK data sets identifies missense variants in SMO, IL11, COL11A1 and 13 more new loci associated with osteoarthritis. *Nature Genetics*, *50*(12), 1681–1687.

Styrkarsdottir, U., Thorleifsson, G., Helgadottir, H. T., Bomer, N., Metrustry, S., Bierma-Zeinstra, S., Strijbosch, A. M., Evangelou, E., Hart, D., Beekman, M., Jonasdottir, A., Sigurdsson, A., Eiriksson, F. F., Thorsteinsdottir, M., Frigge, M. L., Kong, A., Gudjonsson, S. A., Magnusson, O. T., Masson, G., … Stefansson. (2014). Severe osteoarthritis of the hand associates with common variants within the ALDH1A2 gene and with rare variants at 1p31. *Nature Genetics*, *46*(5), 498–502.

Tachmazidou, I., Hatzikotoulas, K., Southam, L., Esparza-Gordillo, J., Haberland, V., Zheng, J., Johnson, T., Koprulu, M., Zengini, E., Steinberg, J., Wilkinson, J. M., Bhatnagar, S., Hoffman, J. D., Buchan, N., Süveges, D., arcOGEN Consortium, Yerges-Armstrong, L., Smith, G. D., Gaunt, T. R., … Zeggini, E. (2019). Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nature Genetics*, *51*(2), 230–236.

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192.

Thulson, E., Davis, E. S., D'Costa, S., Coryell, P. R., Kramer, N. E., Mohlke, K. L., Loeser, R. F., Diekman, B. O., & Phanstiel, D. H. (2022). 3D chromatin structure in chondrocytes identifies putative osteoarthritis risk genes. *Genetics*, *222*(4). https://doi.org/10.1093/genetics/iyac141

Umans, B. D., Battle, A., & Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? *Trends in Genetics: TIG*, *37*(2), 109–124.

Valdes, A. M., Evangelou, E., Kerkhof, H. J. M., Tamm, A., Doherty, S. A., Kisand, K., Tamm, A., Kerna, I., Uitterlinden, A., Hofman, A., Rivadeneira, F., Cooper, C., Dennison, E. M., Zhang, W., Muir, K. R., Ioannidis, J. P. A., Wheeler, M., Maciewicz, R. A., van Meurs, J. B., … Doherty, M. (2011). The GDF5 rs143383 polymorphism is associated with osteoarthritis of the knee with genome-wide statistical significance. *Annals of the Rheumatic Diseases*, *70*(5), 873–875.

Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2."* https://CRAN.R-project.org/package=cowplot

Xie, D. L., Meyers, R., & Homandberg, G. A. (1992). Fibronectin fragments in osteoarthritic synovial fluid. *The Journal of Rheumatology*, *19*(9), 1448–1452.

Yau, M. S., Yerges-Armstrong, L. M., Liu, Y., Lewis, C. E., Duggan, D. J., Renner, J. B., Torner, J., Felson, D. T., McCulloch, C. E., Kwoh, C. K., Nevitt, M. C., Hochberg, M. C., Mitchell, B. D., Jordan, J. M., & Jackson, R. D. (2017). Genome-Wide Association Study of Radiographic Knee Osteoarthritis in North American Caucasians. *Arthritis & Rheumatology (Hoboken, N.J.)*, *69*(2), 343–351.

Zengini, E., Hatzikotoulas, K., Tachmazidou, I., Steinberg, J., Hartwig, F. P., Southam, L., Hackinger, S., Boer, C. G., Styrkarsdottir, U., Gilly, A., Suveges, D., Killian, B., Ingvarsson, T., Jonsson, H., Babis, G. C., McCaskie, A., Uitterlinden, A. G., van Meurs, J. B. J., Thorsteinsdottir, U., … Zeggini, E. (2018). Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis. *Nature Genetics*, *50*(4), 549–558.

Zhang, F., & Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human Molecular Genetics*, *24*(R1), R102–R110.

CHAPTER 2: PLOTGARDENER: CULTIVATING PRECISE MULTI-PANEL FIGURES IN R[1]

2.1 Introduction

The increasing size, complexity, and sheer volume of multi-omic data sets has created a dire need for tools to efficiently visualize, interpret, and communicate the underlying biological signals present in these data. Towards this end, genome browsers, including the UCSC Genome Browser and IGV, have revolutionized our ability to investigate genomic data in a rapid and intuitive fashion, using a stacked linear representation of a wide variety of data types and annotations (Abeel et al., 2012; Carver et al., 2009; Chelaru et al., 2014; Flicek et al., 2011; Freese et al., 2016; Kent et al., 2002; Thorvaldsdóttir et al., 2013; Zhou et al., 2011). Recently, more specialized browsers like Juicebox (Durand et al., 2016) and HiGlass (Kerpedjiev et al., 2018) have increased the ability to visualize non-linear data types, such as 3D chromatin contact frequency (Djekidel et al., 2017; Wang et al., 2018). Furthermore, an ever-increasing array of programmatic libraries and browser APIs now allow code-based, integrated data analysis and construction of browser tracks, which has improved reproducibility and automation (Durinck et al., 2009; Hahne & Ivanek, 2016; Lawrence et al., 2009; Wickham, 2016; Yin et al., 2012).

While these tools have been transformative for data exploration, they are largely based on single-panel figures and vertical stacking of genomic tracks and are often ill-suited for the generation of complex multi-panel figures that include both genomic and non-genomic plot types. Such complex figures are often critical for evaluating the underlying biology and are almost always used to present multi-omic data in publications. Thus, a tool specifically designed to programmatically create and arrange publication-quality multi-panel figures is critical to extend the rigor, reproducibility, and clarity of scientific data visualizations.

Currently existing R packages like patchwork (Pedersen, 2020), cowplot (Wilke, 2020), gridExtra (Auguie, 2017), egg (Auguie, 2019), multipanelfigure (Graumann & Cotton, 2018), and Sushi (Phanstiel et

[1] The work in this chapter has been previously published. The original citation is as follows: Kramer, N.E., Davis, E.S., Wenger, C.D., Deoudes, E.M., Parker, S.M., Love, M. I., & Phanstiel, D. H. (2022) Plotgardener: cultivating precise multi-panel figures in R. Bioinformatics 38(7): 2042-2045. doi: 10.1093/bioinformatics/btac057.

al., 2014) can be used to arrange multi-panel plots. However, these layout packages use relative positioning to place plots and are limited to standard grid-style layouts, giving users little control over precise sizing and arrangement. Furthermore, these packages arrange and align entire plot panels as opposed to internal plot elements like axes. Figures generated with these tools often need finishing in graphic design software such as Adobe Illustrator (Adobe Inc., 2019), Inkscape (Inkscape Project, 2020), PowerPoint (Microsoft Corporation, 2018), and Keynote (Apple Inc., n.d.). In addition to the cost of purchasing proprietary graphic design software and the steep learning curve often associated with their use, generating multi-panel figures with these software requires non programmatic, manual user interactions, a labor intensive process that decreases reproducibility.

Here we introduce plotgardener, an R package for absolute coordinate-based plot placement and sizing of complex multi-panel plots. This paradigm gives users precise control over size, placement, typefaces, font sizes, and virtually all plot aesthetics without the need for graphic design software. Plotgardener (1) supports a vast array of genomic data types, (2) allows precise placement and sizing of genomic and non-genomic figures, (3) is tightly integrated with the Bioconductor environment (Gentleman et al., 2004), and (4) is optimized for speed and user-experience. The code is open source, extensively documented, and freely available via GitHub and Bioconductor.

2.2 Philosophy

The defining feature of plotgardener that separates it from virtually all other genomic visualization tools is that it allows exact sizing and placement of plots using an absolute, coordinate-based plotting system (**Figure 2.1**). Each plot, axis, and annotation is placed independently according to user-specified positions and dimensions. Each plot or feature extends from edge to edge of the defined coordinates, allowing for precise control and perfect alignment of plots. Rulers and guidelines can be temporarily added for ease of plotting and then removed prior to file generation. Adding additional plots does not shift or resize existing ones, so figures can be built incrementally and adjusted without affecting other figure panels, allowing rapid and easy construction of publication-quality multi-panel figures.

2.3 Data Types

Plotgardener can display a vast array of genomic data types which can be provided as either external files or R data classes. Plotgardener has 45 functions for plotting and annotating diverse genomic data types, including genome sequences, gene/transcript annotations, chromosome ideograms,

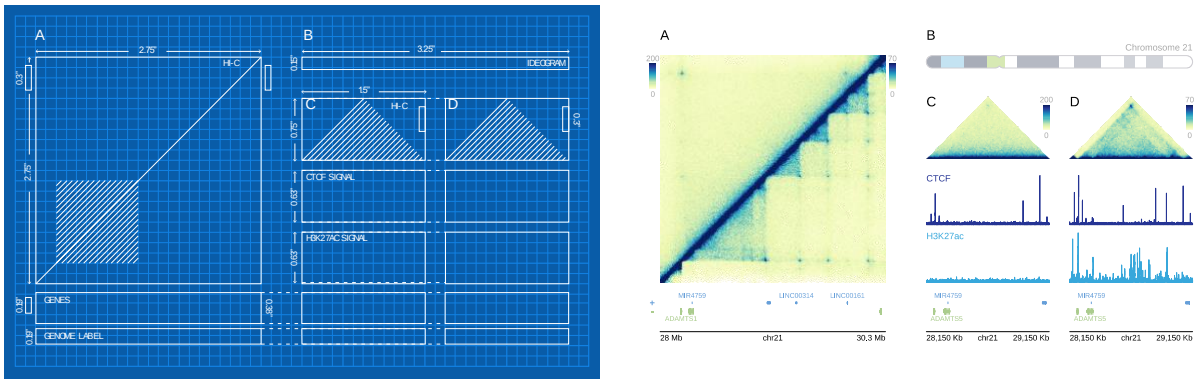

Figure 2.1. Plotgardener uses a coordinate-based plotting system to size and arrange plots. **(A)** Blueprint outline of a multi-omic figure to be created with specified dimensions and placements on a defined page. **(B)** Multi-panel, multi-omicc figure programmatically created with plotgardener using the sizing and placement coordinates from (A). The plotgardener functions used to create this figure include pageCreate, plotHicSquare, annoHeatmapLegend, plotGenes, annoGenomeLabel, plotIdeogram, plotHicTriangle, plotSignal, and plotText. Code to reproduce this plot is included in the plotgardener package.

signal tracks, GWAS Manhattan plots, genomic ranges (e.g. peaks, reads, contact domains, etc), paired ranges (e.g. paired-end reads, chromatin loops, structural rearrangements, QTLs, etc), and 3D chromatin contact frequencies. Plotgardener plotting functions automatically recognize and read compressed, indexed file types including ".bam", ".bigwig", and ".hic", allowing for rapid and memory-efficient reading and plotting of large genomic data. **Figure 2.S1** displays the runtime required to read and plot various types of genomic data. Even with file sizes exceeding 50 GBs, plotgardener can read and plot data in under a second. Multiple classes of R objects are supported, including "data.frame", "data.table", "tibble", "GRanges", and "GInteractions". Plotgardener automatically detects whether the input is a file path or an R object and handles them accordingly, providing a seamless and flexible experience for the user. Furthermore, plotgardener provides the additional reading functions readHic and readBigwig to allow users to work with their raw datasets within the R environment.

2.4 Plotting Workflow

Plotgardener functions can be grouped into four main categories of exported functions – page, reading, plotting, and annotation – which provide a framework for modular plots and annotations to be arranged together on a page. The first step to building a complex, multi-panel figure is to initialize a plotgardener page with pageCreate. The plotgardener page provides the coordinate system for users to size and place plots. pageCreate allows users to define the width, height, and units of the page so users can make a variety of figure sizes in the units of their preference, including inches, centimeters, and millimeters. To assist the placement of figure elements, gridlines can be set in the vertical and horizontal directions with the xgrid and ygrid parameters within pageCreate, and additional guides can be added with pageGuideVertical and pageGuideHorizontal. Once the page is defined, users are free to add plots with precise dimensions and placement coordinates in relation to the page. The page sets the origin of the figure at the top left corner of the page, and the just parameter in plotting functions provides additional flexibility by allowing users to change the placement reference point. The just parameter can be set using character strings or numeric values, as shown in **Figure 2.S2**.

Plotgardener is modular and separates plotting and annotating into two different categories of functions. Once a user creates a plot and places it on the page, the resulting plot object can be passed into various annotation functions through the plot parameter. This parameter allows the annotation function to inherit genomic region and plot location information from the main plot. Possible annotations include genome labels, heatmap legends, Hi-C pixel and domain highlights, axes, and genomic region highlights.

Once a user has plotted and arranged a multi-panel figure with the elements of their choice, they have the option to customize the aesthetics of the figure. Each plot and annotation can be customized for a variety of aesthetic parameters, including colors/palettes, line colors, line widths, font families, font sizes, font colors, and labeling options. In addition, any grid lines or guides used during plotting can be removed with pageGuideHide. After this step, the plotgardener figure is ready for export with any of the built in R graphics devices for saving plots.

2.5 Bioconductor Integration

Plotgardener is tightly integrated with the Bioconductor ecosystem (Gentleman et al., 2004), making it compatible with many existing workflows. It has 29 built-in genomes and associated annotations but can easily accommodate custom genomes and annotations using Bioconductor TxDb (Lawrence et al., 2013), OrgDb (Pagès et al., 2021), and Bsgenome (Pagès, 2021) packages and/or objects. Plotgardener leverages these annotation resources on behalf of the user to obtain and plot chromosome sizes, gene and transcript structures, and nucleotide sequences. By preconfiguring the genome builds and associated feature data, plotgardener allows users to focus their attention on layout and to quickly visualize their data rather than spending time and effort on curation and organization of sequences and genome annotations.

2.6 User Experience

Plotgardener includes a variety of user-friendly features to maximize ease of use for both novices and experienced R programmers. We describe just some of these features here. Parameters can be set within each function call or passed in a pgParams object for more efficient code. Genomic coordinates can be set either by supplying the chromosome, start, and end position or by providing a gene name (e.g. IL1B), reference genome name (e.g. “hg19”), and optional base pair window around the gene (e.g. 50,000 bp). Resolution of Hi-C contact matrices, signal tracks, and gene tracks are automatically determined based on the genomic range being plotted, but can be overwritten if desired. When genomic regions are too large to label all genes, plotGenes and plotTranscripts will choose which genes/transcripts to label based on frequency of appearance in publications. Similarly, annoGenomeLabel/plotGenomeLabel can detect appropriate resolutions to display nucleotides as colored boxes or colored letters. Users can provide their own priorities or select individual genes to highlight with text and colors. A colorby function allows users to flexibly color genomic features by quantitative and qualitative attributes. Plotgardener is open source, version controlled, and extensively documented via articles and vignettes (https://phanstiellab.github.io/plotgardener/).

2.7 ggplot and Beyond

In addition to its included functions for plotting and annotating genomic data, plotgardener allows for the absolute sizing and placement of non-genomic plots, shapes, and images within a plotgardener

page. Users can make multi-panel figures seamlessly by integrating and aligning plotgardener and non-plotgardener plots, including other Bioconductor grid Graphics-based visualizations like ComplexHeatmap (Gu et al., 2016) (**Figure 2.S3**). Users can also create coordinate-based layouts entirely composed of external plot types and objects. For example, plotgardener was used to arrange and add text annotations to the ggplot2 plot objects featured in **Figure 2.S4**. Plotgardener intuitively sizes, arranges, and overlays plots, text, and geometric objects to make complex figure arrangements beyond basic grid-style or relative layouts.

2.8 Future Directions

The plotgardener package is actively maintained via GitHub issues and undergoes regular build reports and unit tests to ensure consistency and robustness. We are actively developing the package with suggestions from the genomic plotting community to refine functions and add additional features, and other potential future additions include more plotting functions, templates for common arrangements, convenient functions for multiplotting, enhanced ggplot2 integration, and more.

In summary, plotgardener provides a new paradigm for generating complex publication-quality figures of both genomic and non-genomic data types, making it an invaluable tool for R users and data scientists from virtually any discipline.

2.9 Methods

Visualization Methods

Plotgardener is an open-source extension for R, building its visualization functions from primitive graphical functions in the grid package (R Core Team, 2021). Each plot and annotation is drawn within its own defined graphical region, or viewport, and then placed on a larger plotgardener page. These viewports give the power to specify the size and placement of plot containers and clip data to precise genomic and data axis measurements. To obtain large, reference genomic annotation data, plotgardener integrates and utilizes packages and data objects through Bioconductor.

Gene and transcript label publication frequency mining

Annotations for genes in PubMed articles were obtained from the PubTator text mining tool (Wei et al., 2013) and counted for each unique gene ID. Publication frequencies were matched via gene ID to Bioconductor transcript database (TxDb) gene IDs for the 29 built-in plotgardener genomes.

Evaluating runtimes of plotgardener plotting functions

To calculate plotgardener plotting runtimes, we used the R package microbenchmark (Mersmann, 2019). plotHicSquare, plotSignal, plotGenes, and plotRanges functions were timed for various genomic region sizes and resolutions. Each condition was timed on 20 random genomic regions generated by BedtoolsR (Patwardhan et al., 2019).

Data availability

Various publicly available datasets are included with a supplementary plotgardenerData package and were used to demonstrate the functionalities of plotgardener. Hi-C datasets from the GM12878 and IMR90 cell lines were downloaded from GEO (Barrett et al., 2013) under the accession code GSE63525. CTCF ChIP-seq signal files for the GM12878 and IMR90 cell lines were downloaded from the ENCODE portal (ENCODE Project Consortium, 2012) with accession codes ENCFF312KXX and ENCFF603PYX. H3K27ac ChIP-seq signal files for the GM12878 and IMR90 cell lines were downloaded from the NIH Roadmap Epigenomics Project (Bernstein et al., 2010) with reference epigenome identifiers E116 and E017. COVID-19 case data was downloaded from The COVID Tracking Project (https://covidtracking.com/). State population data and state COVID-19 vaccination data were downloaded from the Johns Hopkins Centers for Civic Impact COVID-19 GitHub repository (https://github.com/govex/COVID-19/).

2.10 Acknowledgments

We would like to thank Hyejung Won and Jason Stein for helpful discussions and feedback. We thank Muhammad Saad Shamim and Neva Durand for assistance with the strawr package.

2.11 Supplemental Figures

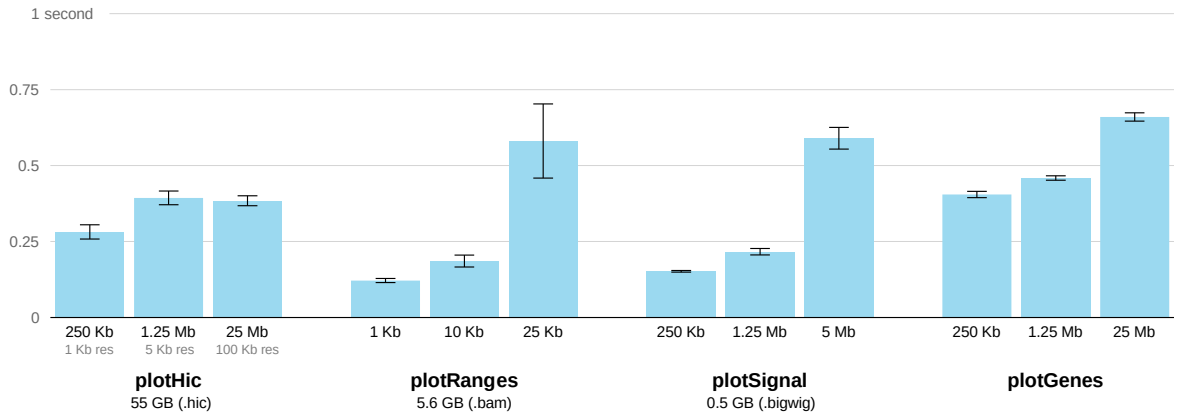

Figure 2.S1. Plotgardener function runtimes.
A bar plot depicting mean runtimes for reading and plotting genomic data across various sizes and resolutions using plotHicSquare, plotRanges, plotSignal, and plotGenes functions. Times were calculated for 20 randomly chosen gene regions for each bar. Error bars indicate standard error. File sizes for the input data are indicated below each set of bars for each function: plotHicSquare (55 GB .hic file), plotRanges (5.6 GB .bam file), plotSignal (0.5 GB .bigwig), plotGenes (NA, data stored as an internal object).

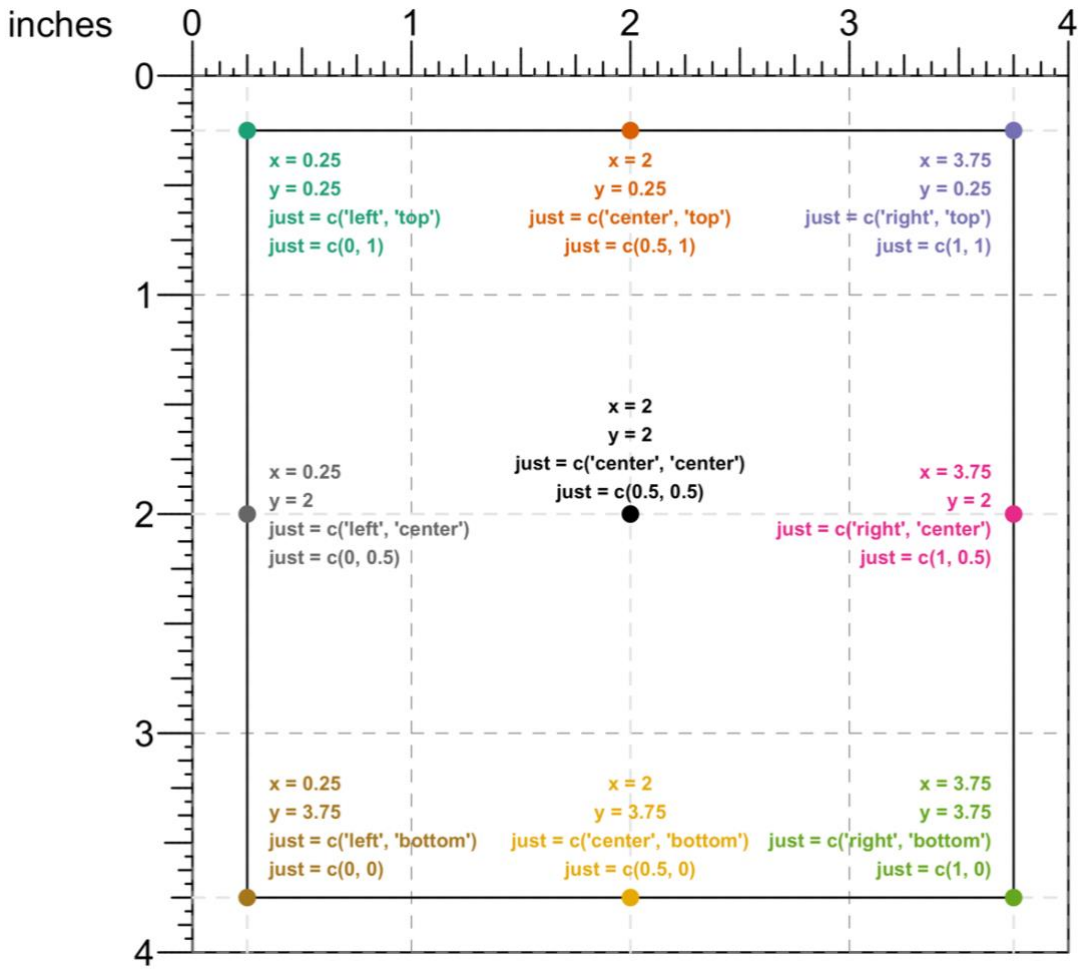

Figure 2.S2. Plotgardener plotting justifications.
Diagram illustrating the various character and numeric justification settings that can be specified with the just parameter of plotting functions. 9 different points along a plot's rectangular edges can be used as the reference point for plotting. Each point shows the the x-coordinate, y-coordinate, justification in characters, and justification in numbers that would be used for plotting the displayed box. More information can be found in the plotgardener vignette "The plotgardener page."

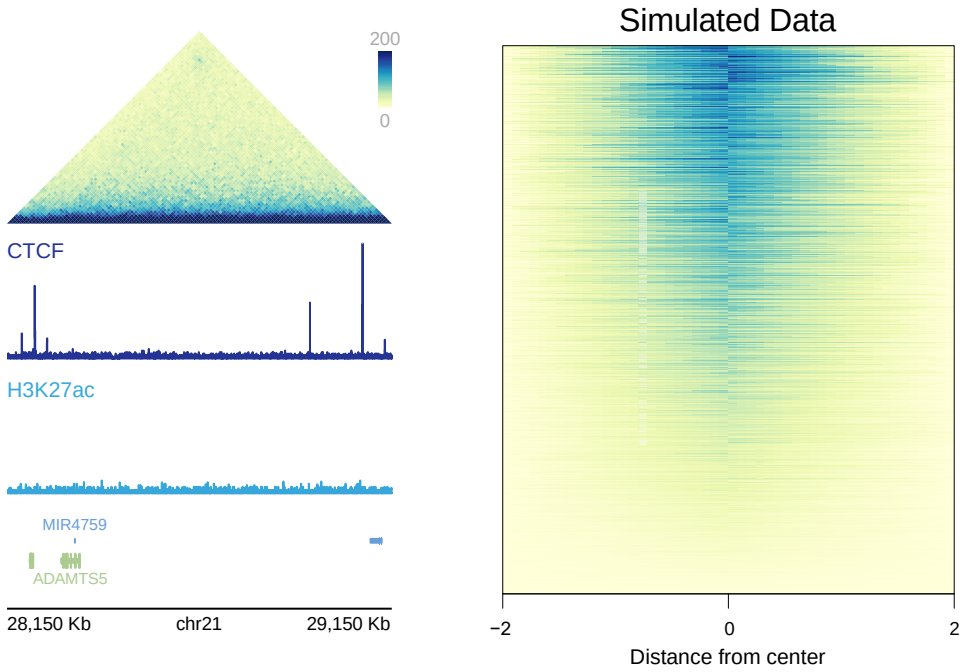

Figure 2.S3. Integration of plotgardener plot objects and Bioconductor ComplexHeatmap. A ComplexHeatmap of the density of simulated ChIP-seq data was incorporated into a plotgardener figure with plotgardener plotting and annotation objects. Left: Triangular Hi-C heatmap, heatmap legend, CTCF and H3K27ac signal tracks, gene track, and genome label plotted with plotgardener functions plotHicTriangle, annoHeatmapLegend, plotSignal, plotGenes, and annoGenomeLabel. Right: Density heatmap of simulated ChIP-seq data produced with ComplexHeatmap and incorporated into the figure using the plotgardener function plotGG. Code to reproduce this figure can be found in the plotgardener vignette “Incorporating ggplots and other grid-based Bioconductor visualizations.”

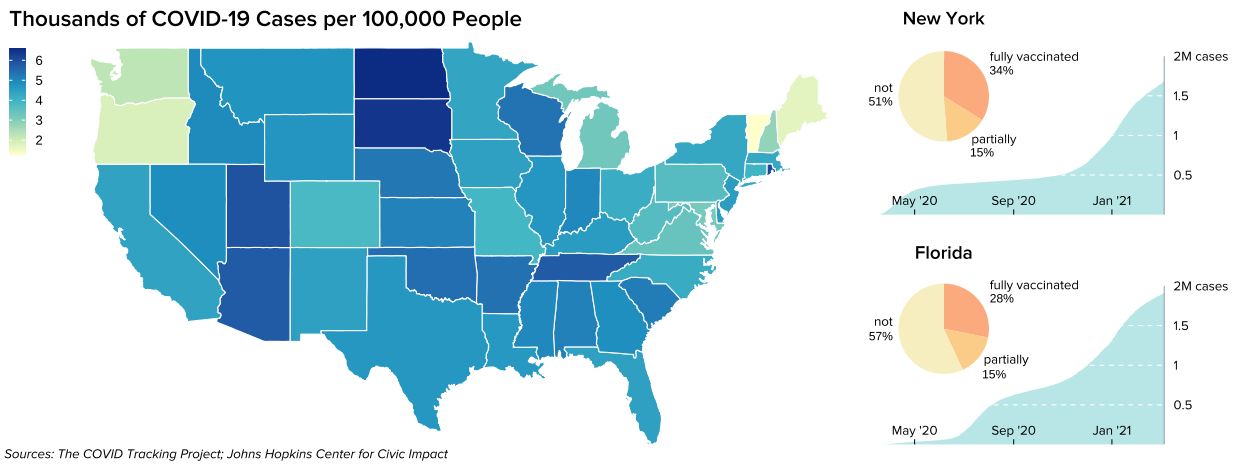

Figure 2.S4. Precise arrangement of ggplot2 objects with plotgardener.
Five ggplot2 objects and additional text elements were arranged into this multi-panel figure using plotgardener. Left: A map of the United States depicts COVID-19 cases per 100,000 people in each state. Right: Pie charts depict state vaccination percentages and line plots describe cumulative COVID-19 cases in New York and Florida. The plotgardener functions used to create this figure include pageCreate, plotGG, and plotText. Code to reproduce this plot is included in the plotgardener package.

REFERENCES

Abeel, T., Van Parys, T., Saeys, Y., Galagan, J., & Van de Peer, Y. (2012). GenomeView: a next-generation genome browser. *Nucleic Acids Research*, *40*(2), e12.

Adobe Inc. (2019). *Adobe Illustrator* (CC 2019 (23.0.3)) [Computer software].https://adobe.com/products/illustrator

Apple Inc. (n.d.). *Keynote* (Version 11.0.1) [MacOS]. https://www.apple.com/keynote/

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra

Auguie, B. (2019). *egg: Extensions for "ggplot2": Custom Geom, Custom Themes, Plot Alignment, Labelled Panels, Symmetric Scales, and Fixed Panel Size*. https://CRAN.R-project.org/package=egg

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research*, *41*(Database issue), D991–D995.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., & Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, *28*(10), 1045–1048.

Carver, T., Thomson, N., Bleasby, A., Berriman, M., & Parkhill, J. (2009). DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* , *25*(1), 119–120.

Chelaru, F., Smith, L., Goldstein, N., & Bravo, H. C. (2014). Epiviz: interactive visual analytics for functional genomics data. *Nature Methods*, *11*(9), 938–940.

Djekidel, M. N., Wang, M., Zhang, M. Q., & Gao, J. (2017). HiC-3DViewer: a new tool to visualize Hi-C data in 3D space. *Quantitative Biology*, *5*(2), 183–190.

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems*, *3*(1), 99–101.

Durinck, S., Bullard, J., Spellman, P. T., & Dudoit, S. (2009). GenomeGraphs: integrated genomic data visualization with R. *BMC Bioinformatics*, *10*, 2.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., … Searle, S. M. J. (2011). Ensembl 2011. *Nucleic Acids Research*, *39*(Database issue), D800–D806.

Freese, N. H., Norris, D. C., & Loraine, A. E. (2016). Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* , *32*(14), 2089–2095.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., … Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80.

Graumann, J., & Cotton, R. (2018). multipanelfigure: Simple Assembly of Multiple Plots and Images into a Compound Figure. In *Journal of Statistical Software, Code Snippets* (Vol. 84, Issue 3, pp. 1–10). https://doi.org/10.18637/jss.v084.c03

Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. “Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data.” *Bioinformatics* 32 (18): 2847–49.

Hahne, F., & Ivanek, R. (2016). Visualizing Genomic Data Using Gviz and Bioconductor. In E. Mathé & S. Davis (Eds.), *Statistical Genomics: Methods and Protocols* (pp. 335–351). Springer New York.

Inkscape Project. (2020). *Inkscape* (Version 0.92.5) [Computer software]. https://inkscape.org

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006.

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J. M., Ouellette, S. B., Azhir, A., Kumar, N., Hwang, J., Lee, S., Alver, B. H., Pfister, H., Mirny, L. A., Park, P. J., & Gehlenborg, N. (2018). HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biology*, *19*(1), 125.

Lawrence, M., Gentleman, R., & Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* , *25*(14), 1841–1842.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., & Carey, V. (2013). Software for Computing and Annotating Genomic Ranges. In *PLoS Computational Biology* (Vol. 9). https://doi.org/10.1371/journal.pcbi.1003118

Mersmann, O. (2019). *microbenchmark: Accurate Timing Functions*. https://CRAN.R-project.org/package=microbenchmark

Microsoft Corporation. (2018). *Microsoft PowerPoint* (2019 (16.0)) [Computer software]. https://office.microsoft.com/PowerPoint

Pagès, H. (2021). *BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs*. https://bioconductor.org/packages/BSgenome

Pagès, H., Carlson, M., Falcon, S., & Li, N. (2021). *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. https://bioconductor.org/packages/AnnotationDbi

Patwardhan, M. N., Wenger, C. D., Davis, E. S., & Phanstiel, D. H. (2019). Bedtoolsr: An R package for genomic data analysis and manipulation. *Journal of Open Source Software*, *4*(44). https://doi.org/10.21105/joss.01742

Pedersen, T. L. (2020). *patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork

Phanstiel, D. H., Boyle, A. P., Araya, C. L., & Snyder, M. P. (2014). Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* , *30*(19), 2808–2810.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192.

Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M. N. K., Li, Y., Hu, M., Hardison, R., Wang, T., & Yue, F. (2018). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biology*, *19*(1), 151.

Wei, C.-H., Kao, H.-Y., & Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research*, *41*(Web Server issue), W518–W522.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2."* https://CRAN.R-project.org/package=cowplot

Yin, T., Cook, D., & Lawrence, M. (2012). ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biology*, *13*(8), R77.

Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E. A., Koebbe, B. C., Nielsen, C., Hirst, M., Farnham, P., Kuhn, R. M., Zhu, J., Smirnov, I., Kent, W. J., Haussler, D., Madden, P. A. F., Costello, J. F., & Wang, T. (2011). The Human Epigenome Browser at Washington University. *Nature Methods*, *8*(12), 989–990.

CHAPTER 3: OSTEOARTHRITIS RELEVANT GENETIC VARIATION AFFECTS GENE EXPRESSION THROUGH ALLELIC IMBALANCE

3.1 Introduction

Osteoarthritis (OA) is an age-related degenerative disease of synovial joints which impacts over 500 million people worldwide and is one of the leading causes of disability and pain (Hunter and Bierma-Zeinstra 2019; Hunter et al. 2020). OA is a disease of the whole joint, but it is primarily characterized by the progressive degradation and loss of articular cartilage as well as the remodeling of joint tissues driven by a variety of inflammatory mediators (Loeser et al. 2012). Currently, there is no disease-modifying treatment for OA with only end-stage treatments available in the form of costly total joint replacements.

It is known that OA has a strong genetic component (Aubourg et al. 2022; MacGregor and Spector 1999), and many genome-wide and gene-targeted studies have aimed to characterize the genetic landscape of the disease and identify putative genes and mechanisms that contribute to disease progression. In particular, numerous studies have identified risk genes whose altered expression contributes to disease progression and recent genome-wide association studies (GWAS) have identified many SNPs that are associated with the disease (Tachmazidou et al. 2019; Boer et al. 2021). However, it still remains difficult to determine the causal SNPs and their molecular mechanisms because many lie in non-coding regions of the genome and most likely modulate OA pathology through genetic regulation.

One method to capture effects of genetic regulatory variants *in cis* is allele-specific expression (ASE) analysis or allelic imbalance (AI). Allele-specific expression refers to the unequal expression of alleles among heterozygous variants in diploids. Previous studies have identified AI events of OA risk and susceptibility genes, both with targeted analyses (Raine et al. 2013; Reynard et al. 2014; Southam et al. 2007) and transcriptome-wide surveys. Notable examples of genes with significant AI that contribute to OA association include *GNL3* and *SPCS1* (Gee et al. 2014), *DIO2* (Bos et al. 2012), *CRLF1* (den Hollander et al. 2019), and *MALAT1* (Coutinho de Almeida et al. 2022). While these examples were found

in diseased OA tissue, many genetic regulatory events occur in response to disease-specific stimuli and must be studied in the appropriate disease context (Umans et al. 2021).

One possible model for simulating the OA transcriptional landscape is fibronectin fragment-treated chondrocytes. Fibronectin is an extracellular protein present in cartilage (Loeser 2014), and fibronectin fragments (FN-f) have been shown to recapitulate many features of OA, including the production of matrix degradation enzymes and pro-inflammatory cytokines (Homandberg et al. 1998; D. L. Xie et al. 1992; Homandberg 1999; Collins et al. 2019). Chondrocytes are the only cell type found in cartilage and synthesize and maintain the cartilage matrix (Bhosale and Richardson 2008), making them the appropriate cell target for such a stimulus. A previous study has characterized the response of chondrocytes isolated from normal cartilage to acute FN-f treatment, confirming robust transcriptional changes that recapitulate the transcriptional landscape of OA (Reed et al. 2021).

In this study, we utilize the above-described model of OA with 79 paired samples of control and FN-f treated chondrocytes isolated from donor articular cartilage to characterize transcriptome-wide AI events in the context of a disease-relevant stimulus and identify sites of putative disease-driven genetic variation. We also overlap AI SNPs with genes expressed differentially between unstimulated and stimulated chondrocytes and intersected our data with OA GWAS SNPs to identify potentially novel OA susceptibility variants and genes affected by AI.

3.2 Results

3.2.1 FN-f treatment recapitulates previously published transcriptional response

To validate the utility of our FN-f treated chondrocytes in capturing OA-relevant transcriptional changes, we conducted differential gene expression analysis and compared gene expression changes in response to FN-f to a previous study that utilized the same system and compared it to genes implicated in OA. We observed 4201 significantly differential genes (FDR < 0.01, absolute log2FC > 2), with 1915 genes differentially upregulated and 2286 genes differentially downregulated in response to FN-f. When compared to Reed et al. (2021), our samples capture 93.1% of the significant differential genes identified in that study (**Figure 3.1A**). Our dataset contains notable upregulated and downregulated genes identified in Reed et al. with known implications in OA, including *CXCL2* (log2FC 8.88, FDR-adjusted p-value < 2.23e-308), *IL6* (log2FC 10.4, FDR-adjusted p-value < 2.23e-308), *MMP13* (log2FC 4.7, FDR-adjusted p-

value = 5.753e-246), and *GDF5* (log2FC -3.49j, FDR-adjusted p-value = 4.494e-270) (**Figure 3.1B, Figure S3.1**). Chemokines and pro-inflammatory

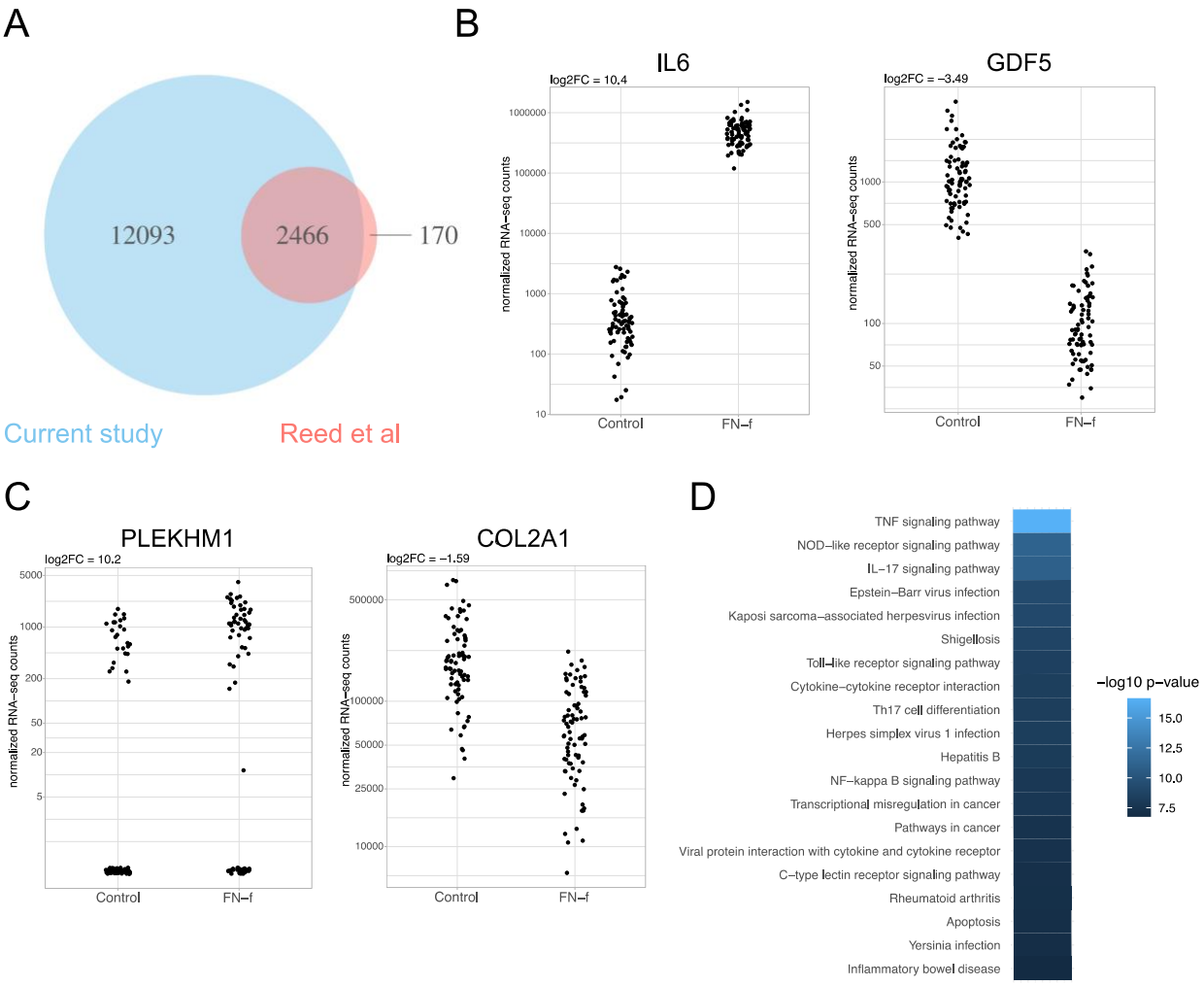

Figure 3.1. Differential expression analysis between FN-f treated and control chondrocytes. (A) Venn diagram depicting the overlap of identified differential genes and previously published differential chondrocyte expression from Reed et al. (2021). **(B)** Donor RNA-seq data for previously identified upregulated *IL6* and downregulated *GDF5* OA-relevant genes. **(C)** Donor RNA-seq data for genes with connections to OA but were not identified in Reed et al. (2021). **(D)** Top 20 KEGG pathways enriched in response to FN-f. Heatmap color represents -log10 p-value enrichment for labeled pathway.

cytokines like *CXCL2* and *IL6* are involved in OA pathogenesis, particularly in the disruption of the inflammatory processes affected during disease progression (Molnar et al. 2021). *MMP13* is a matrix metallopeptidase that degrades type II collagen and is thought to play a role in the progressive cartilage degradation associated with OA (Forsyth et al. 2002). *GDF5* is a growth/differentiation factor with a major role in cartilage and joint development and has been cited as a major susceptibility gene for OA

(Hatakeyama et al. 2004; Kania et al. 2020; Sun et al. 2021; Boer et al. 2021). Only 107 Reed et al. (2021) genes were not marked as significantly differential in our dataset, with 55 of these genes falling into the “up-early” and “down-early” temporal clusters of response which we did not capture with our treated samples. Our study also found genes that were not identified in Reed et al., but have connections to OA. Examples of these genes include *PLEKHM1*, which is related to autophagy and has been shown to play a role in osteoarthritis (Kao et al. 2022) and two genes highlighted as high confidence effector genes in the most recent OA GWAS: wnt family member *WNT10B* and collagen II gene *COL2A1* (Boer et al. 2021) (**Figure 3.1C, Figure S3.1**).

To connect our genes to likely phenotypic functions and pathways, we performed GO and KEGG pathway enrichment analyses. In line with the findings of Reed et al. (2021), differential regulated genes were strongly enriched for processes related to stimulus response and signaling, inflammatory response and cytokine production, and cell morphogenesis. Similarly, differentially regulated genes were enriched for TNF, IL-17, and NF-kappa B pathways (**Figure 3.1D**), which were highlighted by Reed et al. (2021) and have known implications in the inflammatory changes and cartilage destruction of OA (van den Bosch et al. 2020; Choi et al. 2019; Lu et al. 2006). Taken together, these results demonstrate a highly powerful and robust use of the chondrocyte FN-f model system to capture OA-relevant transcriptional changes.

3.2.2 Transcriptome-wide AI in FN-f model of OA

To understand OA-relevant *cis*-regulated gene expression changes, we robustly characterized allelic imbalance at heterozygous variants in control and FN-f treated samples. We quantified allele-specific RNA-seq counts at heterozygous genomic sites and filtered these variants based on read counts, DNA-RNA heterozygotes concordance, and the number of heterozygotes donors for the variant. After this stringent filtering to ensure robust quantification of allelic imbalance, we performed statistical tests for allelic imbalance for 25440 SNPs with DESeq2 (Love et al. 2014). Model fitting and testing resulted in 741 tested SNPs after removal of outliers. The distribution of mean alternative allele fractions in heterozygote donors in control and FN-f conditions among these tested variants showed fractions both above and below 0.5, with a range of values from 0.28 to 0.72 in control and 0.22 to 0.72 in FN-f (**Figure 3.2A**), confirming an appropriate dataset of heterozygous AI candidates. We observed a total of 179 significant

SNPs (FDR > 0.05; abs(log2FC) of reference and alternate alleles > log2(1.1)) marked by allelic imbalance, with 46 unique significant control SNPs, 55 unique significant SNPs in FN-f treated samples, and 78 significant SNPs shared by both conditions (**Figure 3.2B**). These SNPs correspond to 57, 58, and 74 unique positional genes, respectively. Mean alternative allele fractions among significant AI variants follow similar trends as the tested variants but accurately select against variants that show approximate even ratios of reference and alternate allele counts (**Figure 3.2C**).

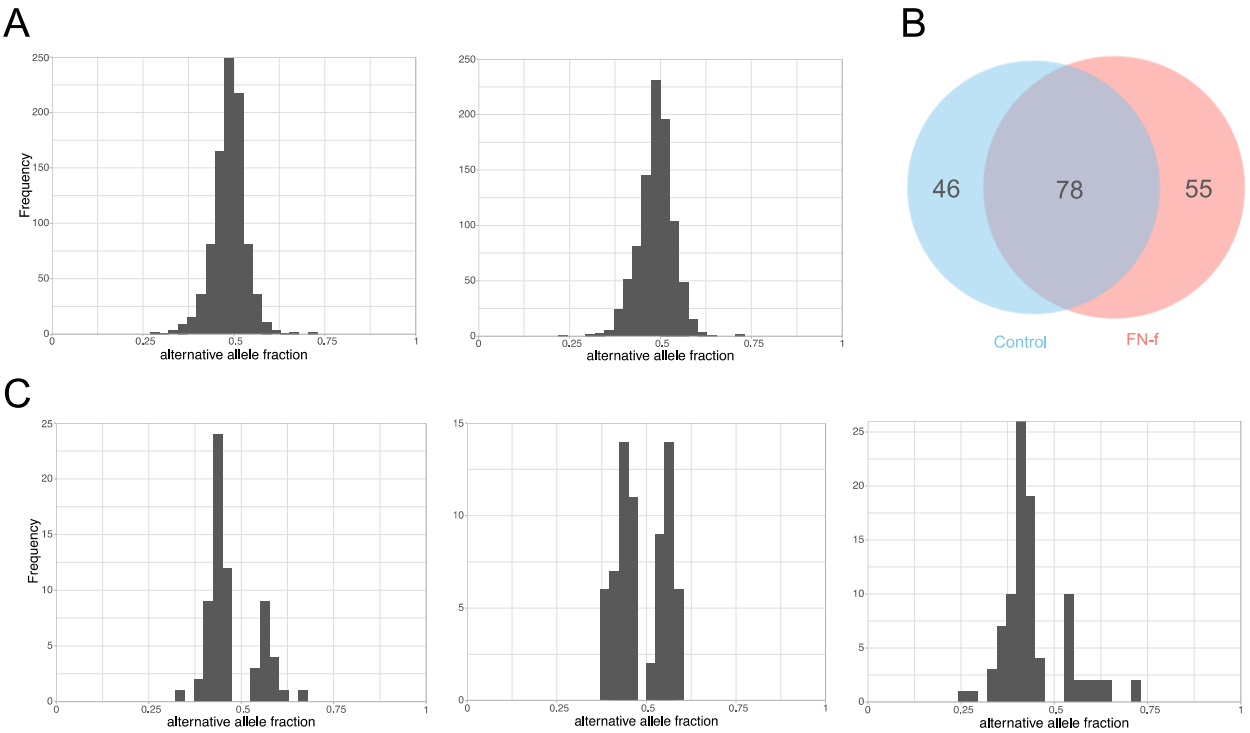

Figure 3.2. Allelic imbalance events in control and FN-f chondrocytes. (A) Distribution of mean alternative allele fraction results from DESeq2 in control (left) and FN-f (right) conditions. **(B)** Venn diagram showing significant AI variants (FDR > 0.05, abs(log2FC) > 1.1) found in control, FN-f, or both conditions. **(C)** Distribution of mean alternative allele fractions for significant AI variants identified with DESeq in control (left), FN-f (middle), and both (right) conditions.

A notable example of a site significantly associated with AI in heterozygous donors in both control and FN-f treated conditions was rs78820491, which is positionally located within the gene *NQO2*. This site consistently showed a higher expression of the reference A allele than the alternative C allele across all samples (**Figure 3.3A**) in both conditions (control FDR-adjusted p-value = 3.935e-3; FN-f FDR-adjusted p-value = 1.220e-4), possibly suggesting that the decreased alternative allele expression may be a consistent site of genetic variation in chondrocytes. One donor's alternative allelic fractions exhibited

opposite trends than the other four donors, with a slightly higher alternative allele fraction after FN-f treatment as opposed to a slightly lower alternative allele fraction after FN-f treatment. These differences may be due to other donor-specific genetic differences or upstream genetic mechanisms that result in allele-specific gene expression changes at the variant, though we do not know these mechanisms through this data.

Many SNPs were only tagged by significant AI in one condition, particularly in samples treated with FN-f, but these effects were subtle. For example, rs3177065 showed overall lower expression of the G alternative allele within the *FOSL2* gene and AI was only detected as statistically significant after FN-f treatment (control FDR-adjusted p-value = 0.1633; FN-f FDR-adjusted p-value = 0.0147) (**Figure 3.3B**). We also observed subtle allelic fraction changes with the opposite direction of effect. For example, rs12418317 showed overall higher expression of the A alternative allele within the *LIN7C* gene with statistically significant AI in FN-f-treated samples (control FDR-adjusted p-value = 0.405; FN-f FDR-adjusted p-value = 5.373e-3) (**Figure 3.3C**). Interestingly, *FOSL2* was identified as a downregulated gene involved in OA (J. Xie et al. 2021) and *LIN7C* was found to be associated with higher bone mineral density and hence higher risk of OA (Yerges-Armstrong et al. 2014). These results identify statistically significant AI in previously implicated OA risk genes with imbalanced alternative allele effect directions aligning with directions of previously identified gene upregulation and downregulation.

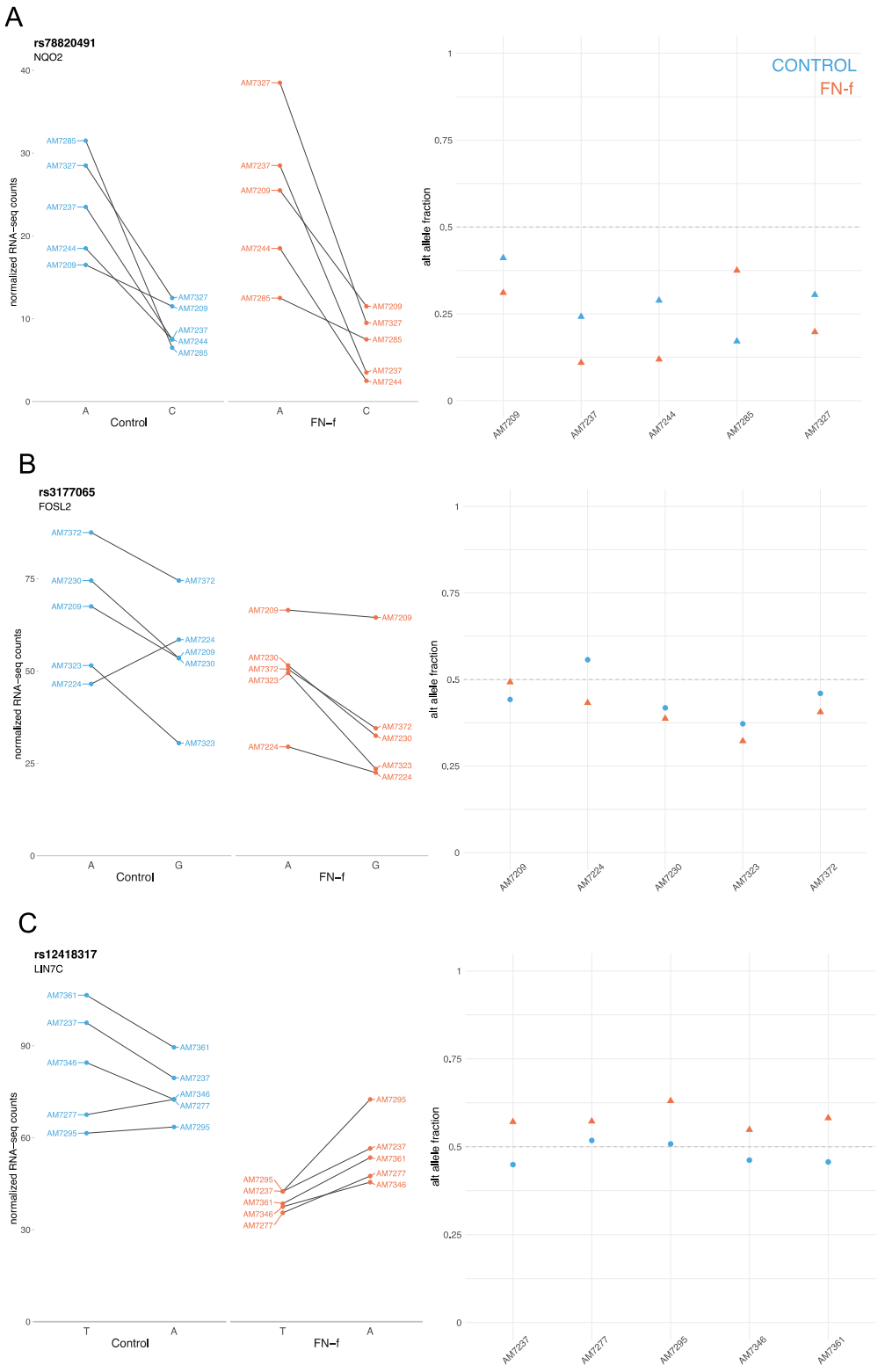

Figure 3.3. *NQO2* expression exhibits significant allelic imbalance in control and FN-f chondrocytes while *FOSL2* and *LIN7C* expression are associated with allelic imbalance after FN-f stimulation. (A) (Left) Normalized RNA-seq counts for reference A allele versus alternative C allele at rs78820491 in *NQO2*. (Right) Alternative allele fractions for heterozygous donors at rs78820491 in

NQO2. Control AI FDR-adjusted p-value = 3.935e-3; FN-f AI FDR-adjusted p-value = 1.220e-4. **(B)** (Left) Normalized RNA-seq counts for reference A versus alternative G allele at rs3177065 in *FOSL2*. (Right) Heterozygote alternate allele fractions at rs3177065 in *FOSL2.* Control AI FDR-adjusted p-value = 0.1633; FN-f AI FDR-adjusted p-value = 0.0147. **(C)** (Left) Normalized RNA-seq counts for reference T versus alternative A allele at rs12418317 in *LIN7C*. (Right) Heterozygote alternate allele fractions at rs12418317 in *LIN7C*. Blue points represent data in a given donor's control samples and orange points represent data in the given donor's FN-f treated samples. Control AI FDR-adjusted p-value = 0.405; FN-f AI FDR-adjusted p-value = 5.373e-3. Triangular points indicate an FDR-adjusted p-value < 0.05 in the corresponding condition. Alternative allele fractions represent the number of alternative allele counts over total read counts, with horizontal dashed lines depicting equal expression ratios of alternative and reference alleles.

3.2.3 AI genes intersect with genes differentially expressed between untreated and FN-f treated chondrocytes

We next intersected our AI candidates with genes identified from differential gene expression analysis to identify differential upregulated and downregulated OA-relevant genes tagged by genetic variation through AI. Of the 4201 significantly differential genes, 26 genes were tagged by 27 SNPs that exhibited significant AI, which includes 21 upregulated genes marked by 22 SNPs and 5 downregulated genes marked by 5 SNPs. Examples of differential genes that contained variants showing significant FN-f-associated AI in FN-f treated samples were *SH3PXD2B* and *GALNT8*. *SH3PXD2B* was consistently upregulated (log2FC 2.19), with the AI at rs17074773 trending towards higher expression of the alternative A allele (FN-f FDR-adjusted p-value = 3.5e-8) (**Figure 3.4A**). *SH3PXD2B* encodes a protein involved in extracellular matrix degradation and has been previously connected to OA as a gene linked to musculoskeletal phenotypes in humans and mice (Boer et al. 2021; Mortier et al. 2019). Conversely, AI at rs11063346 trended towards lower expression of the alternative A allele (FN-f FDR-adjusted p-value = 1.7e-4) with *GALNT8* expression being downregulated with FN-f treatment (log2FC -1.2) (**Figure 3.4B**). *GALNT8* encodes a protein linked to protein modification and protein glycosylation pathways, though it has not been previously implicated in OA.

We also assessed differential gene-based AI at these genes with ASEP (Fan et al. 2020) and found the significant upregulated genes *RGS5*, *SLC25A37*, and *IFIT3* to mark significantly differential AI ($p < 0.05$). Connecting these genes back to our SNP-based AI tests, these genes positionally map to the SNPs rs15049, rs11779396, and rs17119665, which were found to be significant in the FN-f condition (FN-f FDR-adjusted p-values 0.0104, 0.0367, and 0.00407, respectively). These variants had mean alternative allele fractions among heterozygous donors of 0.44, 0.44, and 0.53, respectively. For a full list

of significant AI variants after FN-f treatment that intersected with differential AI genes, please refer to **Table 3.1**. Please refer to **Table S3.1** to compare AI results from DESeq2 and ASEP methods.

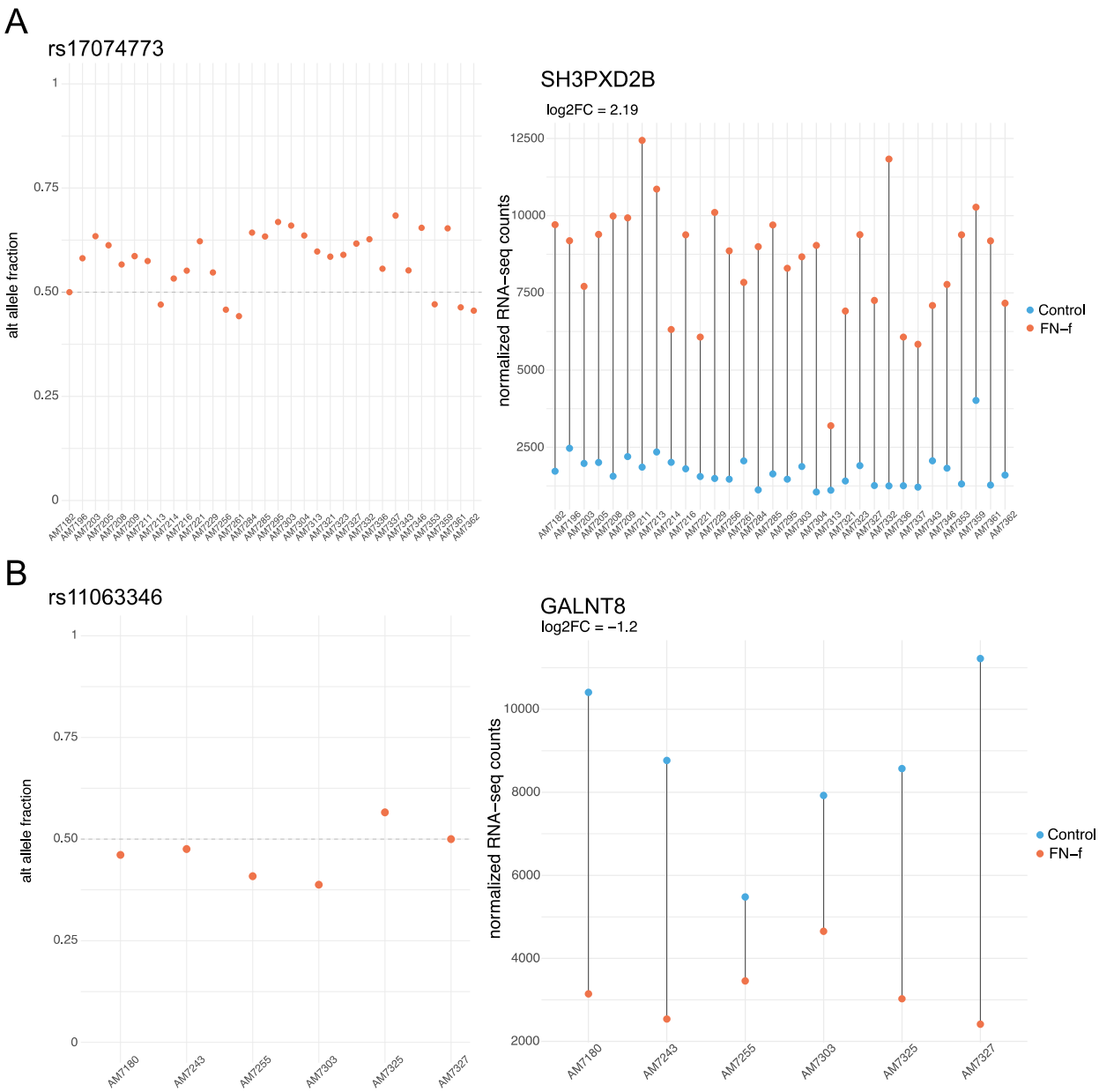

Figure 3.4. Differentially upregulated and downregulated genes intersect with significant FN-f allelic imbalanced SNPs. (A) (Left) Alternative allele fractions (A allele counts over total counts) of heterozygous donors at rs17074773 (AI FN-f FDR-adjusted p-value = 3.4997e-8) maps to upregulated expression of *SH3PXD2B* after FN-f treatment as compared to control (log2FC 2.19), represented by relative normalized RNA-seq counts for heterozyous donors (right). **(B)** (Left) Alternative allele fractions (A allele counts over total counts) of heterozygous donors as rs11063346 (AI FN-f FDR-adjusted p-value = 1.7112e-4) maps to downregulated expression of *GALNT8* after FN-f treatment as compared to control (log2FC -1.2), represented by relative normalized RNA-seq counts for heterozygous donors (right). In alternative allele count fraction plots, the horizontal dashed lines represent even ratios of reference and alternative alleles. In plots illustrating gene expression, blue points represent control donor samples and orange points represent corresponding FN-f-treated donor samples.

SNP	Positional Gene	Reference Allele	Alternative Allele	AI FDR-adjusted p-value	AI log2 Fold Change	Gene FDR-adjusted p-value	Gene log2 Fold Change
rs17074773	*SH3PXD2B*	G	A	3.50E-08	0.4488	< 2.23E-308	2.1889
rs2278225	*LHFPL2*	G	A	8.08E-08	-0.1630	2.92E-111	1.6160
rs11063346	*GALNT8*	G	A	1.71E-04	-0.2405	4.24E-47	-1.2020
rs1044276	*IL4I1*	T	C	9.72E-04	-0.2560	< 2.23E-308	3.7850
rs2077439	*MDM2*	T	G	9.72E-04	-0.4665	< 2.23E-308	1.4163
rs17119665	*IFIT3*	A	G	4.071E-03	0.1419	1.70E-24	1.1469
rs2292692	*ADAM12*	G	A	6.428E-03	0.2743	5.91E-55	1.1712
rs4752904	*PTPRJ*	G	C	7.253E-03	0.2032	< 2.23E-308	1.8815
rs15049	*RGS5*	T	G	1.040E-02	-0.5596	3.21E-54	1.9634
rs12364724	*TNKS1BP1*	G	C	1.177E-02	-0.6250	< 2.23E-308	1.1715
rs10493018	*PNRC2*	A	C	2.599E-02	-0.3697	< 2.23E-308	1.3048
rs11779396	*SLC25A37*	C	G	3.666E-02	-0.3819	1.13E-153	1.6889
rs3198487	*BTN3A1*	G	C	3.696E-02	-0.2984	< 2.23E-308	2.0313
rs61729681	*PARP12*	C	A	3.799E-02	0.3629	7.81E-127	1.7316

Table 3.1. Intersection of significant allelic imbalance SNPs with positional genes exhibiting significant differential expression.

3.2.4 AI variants overlap with genetic association signals in various OA phenotypes

To connect our AI results to OA risk loci reported in OA GWAS, we intersected our AI SNPs and their proxies with variants with reported OA genetic association signals. Using GWAS meta-analysis for OA phenotypes defined as “All OA”, “Finger OA”, “Hand OA”, “Hip OA”, “KneeHip OA”, “Knee OA”, “Spine OA”, “Total Hip Replacement (THR)”, “Thumb OA”, “Total Joint Replacement (TJR)”, and “Total Knee Replacement (TKR)”, we compared significant FN-f-specific AI SNPs to GWAS SNPs with a nominal genetic association (p-value < 0.05). As shown in **Table 3.2**, 34 unique SNPs were nominally associated in one or more of 10 OA phenotypes and were marked with significant AI after FN-f treatment in our dataset. As an example, rs11779396 was associated with 5 OA phenotypes either directly (All OA, Finger OA, Knee OA) or by proxy SNP in high LD (KneeHip OA and Spine OA). This SNP was significantly associated with FN-f AI and positionally marked *SLC25A37* with a direction of effect trending towards lower expression of the alternative G allele. Many of the identified AI SNPs with associations in the OA GWAS mapped to positional genes that also exhibited significant differential expression within our data, including *SH3PXD2B*, *LHFPL2*, *GALNT8*, *MDM2*, *IFIT3*, *PTPRJ*, *RGS5, PNRC2*, *SLC25A3*, *BTNA31*, and *PARP12*.

Allelic Imbalance					GWAS						
SNP	Positional Gene	Ref	Alt	FDR	SNP	R2	Beta	p	OA	EA	EAF
rs385543	*CFH*	A	G	3.01E-09	rs460376	1	0.364	0.0344	S	T	0.0072
rs17074773	*SH3PXD2B*	G	A	3.50E-08	rs2731722	0.83	-0.0151	0.0439	K	T	0.33
rs2278225	*LHFPL2*	G	A	8.08E-08	rs3822476	0.83	0.0759	0.00668	THR	T	0.955
rs11063346	*LOC105369614*	G	A	1.72E-04	rs11063346	NA	0.04	0.04747	THR	A	0.0803
rs11063346	*GALNT8*	G	A	1.72E-04	rs11063346	NA	0.04	0.04747	THR	A	0.0803
rs11063346	*KCNA6*	G	A	1.72E-04	rs11063346	NA	0.04	0.04747	THR	A	0.0803
rs1874340974	*TMEM120B*	A	G	2.51E-04	rs1874340974	NA	-0.0158	0.00841	All	A	0.8244
rs1874340974	*RHOF*	A	G	2.51E-04	rs1874340974	NA	-0.0158	0.00841	All	A	0.8244
rs1874340974	*TMEM120B*	A	G	2.51E-04	rs1874340974	NA	-0.0275	0.04034	S	A	0.828
rs1874340974	*RHOF*	A	G	2.51E-04	rs1874340974	NA	-0.0275	0.4034	S	A	0.828
rs1874340974	*TMEM120B*	A	G	2.51E-04	rs1874340974	NA	-0.0194	0.03966	K	A	0.827
rs1874340974	*RHOF*	A	G	2.51E-04	rs1874340974	NA	-0.0194	0.03966	K	A	0.827
rs1874340974	*TMEM120B*	A	G	2.51E-04	rs12303659	0.9	0.0237	0.02417	KH	A	0.176
rs1874340974	*RHOF*	A	G	2.51E-04	rs12303659	0.9	0.0237	0.02417	KH	A	0.176
rs3741588	*TMEM120B*	A	G	2.51E-04	rs3741588	NA	-0.0158	0.0084	All	A	0.824
rs3741588	*RHOF*	A	G	2.51E-04	rs3741588	NA	-0.0158	0.0084	All	A	0.824
rs3741588	*TMEM120B*	A	G	2.51E-04	rs3741588	NA	-0.0275	0.0403	S	A	0.828
rs3741588	*RHOF*	A	G	2.51E-04	rs3741588	NA	-0.0275	0.0403	S	A	0.828
rs3741588	*TMEM120B*	A	G	2.51E-04	rs3741588	NA	-0.0194	0.0397	K	A	0.827
rs3741588	*RHOF*	A	G	2.51E-04	rs3741588	NA	-0.0194	0.0397	K	A	0.827
rs3741588	*TMEM120B*	A	G	2.51E-04	rs7308348	0.84	-0.0167	0.0396	KH	T	0.825
rs3741588	*RHOF*	A	G	2.51E-04	rs7308348	0.84	-0.0167	0.0396	KH	T	0.825
rs7228940	*CEP192*	G	A	1.55E-03	rs143666215	0.82	-0.0306	0.0464	K	D	0.117
rs7228940	*CEP192*	G	A	1.55E-03	rs60945352	1	-0.0845	0.0464	THR	T	0.131
rs2077439	*MDM2*	T	G	3.99E-03	rs7296057	0.86	-0.605	0.0403	K	C	0.9996
rs17119665	*IFIT3*	A	G	4.07E-03	rs12250860	1	-0.1633	0.0363	F	A	0.0148
rs34644316	*ELK3*	C	T	4.26E-03	rs35891480	1	-0.0775	0.0132	K	A	0.0878
rs34644316	*ELK3*	C	T	4.26E-03	rs75936517	0.87	-0.0481	0.0359	Ha	T	0.0784
rs12460570	*ZNF253*	C	G	4.49E-03	rs3841054	1	0.0536	0.0488	F	D	0.1614
rs1221896942	*UBFD1*	A	G	4.78E-03	rs1221896942	0.94	0.0614	0.0164	K	A	0.962
rs541504918	*UBFD1*	A	G	4.78E-03	rs9937564	0.94	0.0614	0.0164	K	A	0.962
rs12418317	*LIN7C*	T	A	5.37E-03	rs16917051	0.83	-0.1672	0.0252	F	T	0.0163
rs3797851	*HAVCR2*	G	T	6.45E-03	rs3797851	NA	0.0271	0.0465	S	T	0.161
rs3797851	*MED7*	G	T	6.45E-03	rs3797851	NA	0.0271	0.0465	S	T	0.161
rs4752904	*PTPRJ*	G	C	7.25E-03	rs4752894	0.9	-0.0209	0.0288	H	A	0.605
rs4752904	*PTPRJ*	G	C	7.25E-03	rs4752894	0.9	-0.0247	0.0305	THR	A	0.604
rs7301926	*STX2*	T	C	9.92E-03	rs61346189	1	-0.55	0.0204	Ha	A	0.999
rs15049	*RGS5*	T	G	1.04E-02	rs15049	NA	-0.0396	0.00773	H	T	0.895
rs62222237	*N6AMT1*	A	G	1.04E-02	rs18702983	0.84	0.0481	0.0128	THR	A	0.121
rs62222237	*N6AMT1*	A	G	1.04E-02	rs187023983	0.84	0.0358	0.0415	H	A	0.121
rs1804094	*KIAA0040*	C	G	1.30E-02	rs1804094	NA	0.076	0.0121	Fr	C	0.897
rs1804094	*KIAA0040*	C	G	1.30E-02	rs1804094	NA	0.0556	0.0292	Ha	C	0.896
rs1076669	*ECE1*	G	A	1.38E-02	rs1076669	NA	0.0464	0.03423	Ha	A	0.0842
rs1076669	*ECE1*	G	A	1.38E-02	rs1076669	NA	0.0634	0.02873	Th	A	0.0895
rs12713193	*FAM228B*	C	T	2.09E-02	rs6721278	1	0.0204	4567	K	T	0.1396
rs41291167	*BIRC6*	T	C	2.41E-02	rs41291167	NA	0.0417	0.0473	THR	T	0.924
rs10493018	*PNRC2*	A	C	2.60E-02	rs10493018	NA	0.122	0.0176	Th	A	0.952
rs10493018	*PNRC2*	A	C	2.60E-02	rs10493018	NA	0.105	0.0202	Ha	A	0.953
rs10493018	*PNRC2*	A	C	2.60E-02	rs10493018	NA	0.12	0.0358	Fr	A	0.963
rs77410650	*CMAHP*	T	C	3.10E-02	rs77410650	NA	-0.0482	0.0428	TKR	T	0.93
rs77410650	*CMAHP*	T	C	3.10E-02	rs35073828	0.8	-0.0389	0.038	TJR	D	0.93
rs77410650	*CMAHP*	T	C	3.10E-02	rs916539	1	0.0633	0.04	Th	T	0.08
rs12367881	*ALG10*	T	C	3.13E-02	rs71447679	0.82	0.0272	0.02399	H	D	0.3896
rs11779396	*SLC25A37*	C	G	3.66E-02	rs11779396	NA	0.0222	0.0147	All	C	0.993
rs11779396	*SLC25A37*	C	G	3.66E-02	rs11779396	NA	-0.795	0.0219	F	C	0.932
rs11779396	*SLC25A37*	C	G	3.66E-02	rs11779396	NA	0.0375	0.00954	K	C	0.935
rs11779396	*SLC25A37*	C	G	3.66E-02	rs73671493	0.86	-0.0253	0.0401	KH	C	0.0658
rs11779396	*SLC25A37*	C	G	3.66E-02	rs11779370	1	0.0411	0.0475	S	C	0.9358
rs3198487	*BTN3A1*	G	C	3.70E-02	rs3198487	NA	-0.022	0.02679	K	C	0.155
rs3198487	*BTN3A1*	G	C	3.70E-02	rs3198487	NA	-0.0169	0.047	KH	C	0.153
rs3198487	*BTN3A1*	G	C	3.70E-02	rs111540572	0.97	0.044	0.0359	TKR	T	0.859
rs3198487	*BTN3A1*	G	C	3.70E-02	rs55775018	0.97	-0.0132	0.0462	All	T	0.146
rs61729681	*PARP12*	C	A	3.80E-02	rs61729681	NA	-0.0387	0.0273	K	A	0.0462
rs61729681	*PARP12*	C	A	3.80E-02	rs61729681	NA	-0.0335	0.0241	KH	A	0.0469
rs41296175	*LINC01140*	A	T	3.90E-02	rs41296175	NA	-0.0333	0.0332	All	A	0.978
rs41296175	*HS2ST1*	A	T	3.90E-02	rs41296175	NA	-0.0333	0.0332	All	A	0.978
rs2231250	*AUP1*	G	C	3.90E-02	rs2231250	NA	0.0192	0.00271	All	C	0.157
rs2231250	*AUP1*	G	C	3.90E-02	rs2231250	NA	0.0176	0.04008	KH	C	0.1558

rs2231250	*AUP1*	G	C	3.90E-02	rs2231250	NA	0.0379	0.027	TKR	C	0.149
rs2231250	*AUP1*	G	C	3.90E-02	rs2231250	NA	0.0217	0.03	K	C	0.159
rs10409531	*ZNF470-DT*	C	T	3.90E-02	rs111831807	0.84	-0.028	0.0387	S	D	0.686
rs10409531	*ZFP28*	C	T	3.90E-02	rs111831807	0.84	-0.028	0.0387	S	D	0.686
rs10409531	*ZNF470-DT*	C	T	3.90E-02	rs3065645	0.93	-0.0399	0.0467	Th	D	0.319
rs10409531	*ZFP28*	C	T	3.90E-02	rs3065645	0.93	-0.0399	0.0467	Th	D	0.319
rs7215868	*NLRP1*	T	C	4.41E-02	rs7215868	NA	-0.0354	0.0278	K	T	0.949
rs56301120	*EPHX1*	G	A	4.44E-02	rs56301120	NA	-0.0582	0.0076	Ha	A	0.086
rs56301120	*TMEM63A*	G	A	4.44E-02	rs56301120	NA	-0.0582	0.0076	Ha	A	0.086
rs56301120	*EPHX1*	G	A	4.44E-02	rs56301120	NA	-0.0606	0.0434	Th	A	0.0876
rs56301120	*TMEM63A*	G	A	4.44E-02	rs56301120	NA	-0.0606	0.0434	Th	A	0.0876

Table 3.2. Intersection of significant FN-f specific allelic imbalance SNPs with nominally significant OA GWAS SNPs. Ref = reference allele; Alt = alternative allele; FDR = FN-f FDR-adjusted p-value; R2 = R2 measure of linkage disequilibrium of proxy SNP with AI SNP; OA = OA GWAS phenotype; EA = effect allele; EAF = effect allele frequency; All = All OA; F = Finger OA; H = Hip OA; Ha = Hand OA; K = Knee OA; KH = KneeHip OA; S = Spine OA, Th = Thumb OA; THR = Total Hip Replacement; TJR = Total Joint Replacement; TKR = Total Knee Replacement.

3.3 Discussion

In this study, we used a dataset of matched donors with a robust transcriptomic model of OA to identify disease-relevant genetic variants acting through imbalanced expression of alleles. We not only confirmed the validity of the chondrocyte FN-f system in recapitulating the transcriptomic changes of OA progression and cartilage degradation, but also use the largest sample size of this system thus far (n = 79 unique donors) to create the first FN-f response-specific ASE dataset. The AI variants identified provide an initial set of candidate sites with which to investigate the effects of genetic variation on gene expression changes during disease progression. Among the genes these SNPs positionally marked, we confirmed previously known chondrocyte and OA-relevant genes (e.g. *COL1A2* (Snelgrove et al. 2005), which had AI in both conditions; *FOSL2* (J. Xie et al. 2021), *LIN7C* (Yerges-Armstrong et al. 2014), *MDM2* (Jiang et al. 2022), *RGS5* (Appleton et al. 2006), and *SLC25A37* (Rai et al. 2019), which were tagged by AI only after FN-f treatment) and found genes not previously connected to OA (e.g. *GALNT8*, *PARP12, LHFPL2*, and *IFIT3*) that could potentially serve as new candidates for investigation in relation to OA.

An interesting aspect of our dataset was the variants that marked significant AI both in control and FN-f treated conditions, which previous OA-related AI datasets cannot identify with diseased cartilage samples alone. *NQO2* was one such gene marked by statistically significant AI at rs78820491. *NQO2* encodes a member of the thioredoxin family of enzymes and harbors the rs78820491 A>C SNP where both sample conditions showed significantly decreased expression of the alternate C allele among all heterozygote donors. *NQO2* was not differentially expressed in our data, but its related quinone oxidoreductase *NQO1* was upregulated. Both NQO2 and NQO1 enzymes are factors related to the

regulation of the immune response, with the loss of NQO2 and NQO1 linked to lack of NF-κB activation with NQO1-null and NQO2-null mice having a predisposition to collagen-induced arthritis (Iskander et al. 2006). Based on our findings, we hypothesize that the upregulation of *NQO1* compensates for the lower expressed alternative C allele of *NQO2* in all chondrocytes, which may still leave a potential predisposition to OA but mitigating some inhibited immune function. This AI event may not serve as a suitable therapeutic target for OA since it exhibits AI for both conditions in our model and may be a site regulated by similar genetic mechanisms regardless of disease state.

Among the AI events only identified after FN-f treatment, we were able to find previously implicated OA risk genes and hypothesize possible mechanisms for conferring or tagging OA risk through AI. For example, decreased expression of *FOSL2* in a recent study of OA risk genes linked *FOSL2* to OA susceptibility (J. Xie et al. 2021), and part of this susceptibility may be conferred in heterozygotes through lower expression of the alternative G allele at rs3177065. Similarly, *LIN7C* is associated with higher bone mineral density which has been linked observationally to a higher risk of OA, particularly in the knee (Yerges-Armstrong et al. 2014). This study linked rs10835187 with association to OA to *LIN7C* as the nearest gene, and our dataset reveals AI at the SNP rs12418317 whereby heterozygotes show increased expression of the alternative A allele. This may confirm the relevance of *LIN7C* to OA through imbalanced expression of alleles mediated by genetic risk variants.

Notable AI SNPs also overlapped significantly differentially expressed genes, including rs17074773 in *SH3PXD2B* and rs11063346 in *GALNT8*. In the latest OA GWAS meta-analysis, *SH3PXD2B* was identified as a likely effector gene for OA linked to the lead SNP rs3884606 (Boer et al. 2021). *SH3PXD2B* encodes an adapter protein involved in cell adhesion with mutations in the gene being linked to Borrone dermato-cardio-skeletal syndrome. Borrone demato-cardio-skeletal-syndrome is characterized by traits like thickened joints (Wilson et al. 2014), giving it a plausible connection to OA. Not only was *SH3PXD2B* expression upregulated both in our study and a previous study utilizing the FN-f model of OA, but it was also marked by highly significant AI (FDR 3.5e-8) at rs17074773 G > A with higher expression of the alternative A allele. With these combined lines of evidence, we hypothesize that increasing expression of the alternative A allele confers higher expression of *SH3PXD2B* which may result in abnormal cell adhesion within the joints and lead to susceptibility for diseases like OA. *GALNT8*

has never been previously connected to OA, but was shown to be downregulated with FN-f treatment and marked by AI at rs11063346 G > A. *GALNT8* encodes one of a family of glycosylation enzymes whose altered activity has been implicated in cancer metastasis. Our data reveals that this gene could also be relevant in OA, where lower expression of the alternative A drives lower expression of the gene and may confer OA susceptibility through misregulated EGFR signaling, though future functional studies are required to verify this hypothesis.

By comparing AI SNPs with OA-associated SNPs from GWAS data, we were able to connect the genetic variation at our SNPs with 11 different OA phenotypes. Although we were not able to observe any overlap with the lead variants or proxies reported in Boer et al. (2021), we found 34 unique nominally significant SNPs that were tagged by significant AI. Numerous significant AI SNPs were found in multiple OA phenotypes, which could possibly suggest shared genetic mechanisms act at these sites for related OA phenotypes. We did not observe an enrichment of GWAS putative OA risk SNPs in our dataset, which may point to the subtlety of the allelic imbalance events we observed and might suggest that these SNPs may not be functional drivers of gene expression changes. It is also possible that the variants identified through this analysis are merely markers with which to measure allele-specific gene expression as a consequence of unknown disease mechanisms.

The dataset presented here is the first characterization of AI events in the FN-f chondrocyte model of OA. With such a large sample size of paired donors (n = 79), we are able to recapitulate OA-relevant transcriptomic changes and find differential genes and sites of AI not found in previous studies. However, all the AI events identified here are merely statistical associations within an OA model and any hypotheses must be validated. Nonetheless, these data serve as an initial examination in understanding the effects of *in cis* gene regulation at these sites of genetic variation. When comparing to previous OA AI studies, we did not necessarily identify any discrepancies from previous OA AI datasets, but we did not find previously tagged AI SNPs to be significant in our dataset. This may occur due to several reasons. First, FN-f treatment induces OA-relevant changes, but may still fail to capture some OA-relevant genes and mechanisms. This is to be expected using an approximate model of the disease and does not discount the validity of such a model, but attests to the benefits and drawbacks for both model and disease tissue systems to study OA genetics. It is also possible that previous studies reported false

positive AI hits that we corrected for using other statistical methods. We ensured stringent quality control and filtering of our data at multiple stages, and tested for AI using a Wald test with a beta binomial model or a generalized linear mixed-effects model accounting for donor-specific random effects. Using different statistical models and methods than previous studies may have yielded fewer false positives. However, it may also be worthwhile to explore different statistical models to test for AI with our dataset, like the binomial and negative binomial models. Furthermore, despite our large sample size of donors, we cannot ensure sufficient numbers of heterozygotes at each variant site and may not have included numerous variants in our statistical tests. Though previous AI studies required comparable ratios of heterozygotes for their AI tests (e.g. 2 heterozygotes for 47 donors (Hollander et al. 2019) and 5 heterozygotes for 85 or 74 donors (Aygün et al. 2021)), 5 heterozygotes may not be a sufficient cutoff to test for AI. However, a more stringent heterozygote cutoff would further limit the variants we are able to test. Lastly, our study may be more relevant for capturing early-stage OA susceptibility AI SNPs that previous studies could not identify using diseased cartilage.

In summary, we have produced a dataset identifying AI in response to an OA-relevant stimulus to characterize genetic variation that may contribute to disease risk or serve as indicators of disease risk. Combining evidence of differential gene regulation and condition-specific AI allowed us to present a dataset of variants that are potentially affected by disease-specific mechanisms and result in disease-relevant gene expression changes. These results and the analysis framework presented here will allow for further investigation into the effects and consequences of OA genetic mechanisms as we continue to disentangle the genetics of OA and search for novel therapeutic targets.

3.4 Materials and Methods

Sample collection and treatment

Primary articular chondrocytes were isolated by enzymatic digestion from human talar cartilage obtained from 79 tissue donors without a history of arthritis (see Table S3.2 for donor sexes, ages, and self-reported ancestries), through the Gift of Hope Organ and Tissue Donor Network (Elmhurt, IL) as previously described (Reed et al. 2021; Loeser et al. 2003). After serum starvation, cells were treated with either purified 42-kDa endotoxin-free recombinant FN-f (1 μM final concentration in PBS), prepared as previously described, or with PBS as a control (Wood et al. 2016). After 18 hours of treatment with FN-f or

PBS, RNA was isolated using the RNeasy kit from Qiagen. Samples were sent for library preparation and sequencing at the New York Genome Center.

RNA-sequencing data processing

RNA-seq libraries were sequenced to an average depth of approximately 80 million reads per sample at the New York Genome Center. FASTQ files that were from the same library but sequenced on multiple flow cells were merged. Low quality reads and adapters were trimmed with TrimGalore! v0.6.2 (Krueger). We performed quality control of each library with FastQC v0.11.8 (Andrews 2010). These trimmed reads were quantified for differential gene expression analysis with Salmon v1.4.0 (Patro et al. 2017) against the hg38 transcriptome. All programs were run with default settings.

Genotype processing

We performed genotyping using the Illumina Human Infinium Global Diversity Array platform and exported SNP genotypes into PLINK format with the Illumina software GenomeStudio. Quality control was performed with PLINK v1.9 (Purcell et al. 2007) to filter out SNPs with missing genotype rate > 10% (--geno 0.1), deviations from Hardy-Weinberg equilibrium at a p-value < 1 x 10e-6 (--hwe 10^-6), and minor allele frequency < 1% (--maf 0.01). Reported sample sexes were confirmed based on heterozygosity on the X chromosome. We estimated the population structure of our samples after combining it with data from the 1000 Genomes Project using EIGENSTRAT v7.2.1 (Price et al. 2006; Patterson et al. 2006). Data was imputed with Eagle2 (v2.4) phasing (Loh et al. 2016) against the TOPMed Imputation Server against the TOPMed reference panel (version R2 on GRC38) (Das et al. 2016). Following imputation, we retained SNPs with missing genotype rate < 10% (--geno 0.1), deviations from Hardy-Weinberg equilibrium at a p-value > 1 x 10e-6 (--hwe 10^-6), minor allele frequency > 1% (--maf 0.01), and sufficient imputation quality (R2 > 0.3). The final dataset contained 10419216 autosomal variants for 79 donor samples.

Sample quality control

We detected sample swaps or mixing between samples by evaluating the consistency of genotypes called from RNA-seq and genotyping array using VerifyBamID v1.1.3 (Jun et al. 2012). We detected genotyping sample swaps (n = 2) and corrected them. All RNA-seq libraries had [FREEMIX] >

0.04 and [CHIPMIX] < 0.04 and were thus kept for subsequent analyses. After quality control, we retained 79 unique control donors and 79 unique FN-f treated donors with genotyping and RNA-seq data.

Differential gene expression

To identify differential genes, transcript-level quantifications for each sample were summarized and converted to gene-level scaled transcripts in R with tximeta (Love et al. 2020). Differential analysis was conducted with DESeq2 (Love et al. 2014) using a design to adjust for donor variability and to calculate differences between treatment conditions (~ donor + condition). Differential genes were defined as those with p-values < 0.01 and absolute log2FC < 2.

Comparison of differential genes with data from Reed et al. (2021)

Sample gene-level summaries from Reed et al. (2021) were downloaded at GEO accession GSE150411. To account for the multiple clusters of temporal patterns found within this dataset, we isolated a significant differential dataset with genes of p-values < 0.01 and absolute log2FC < 1 for the 18-hour time point. Our compared dataset defined significant differential genes with these same thresholds.

GO term and KEGG pathway enrichment analysis

Significantly enriched (p-value < 0.05) Gene Ontology (GO) terms among significant differential genes were found with terms obtained through topGO (Alexa and Rahnenfuhrer 2021) and tested with a Fisher's exact test. Significant enriched (p-value < 0.05) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways among significant differential genes were identified using KEGGREST (Tenenbaum and Maintainer 2021) with a Wilcoxon rank-sum test.

Allele-specific expression analysis

Trimmed FASTQ files from RNA-seq data processing were aligned to the GENCODE.GRCh38.p13 reference genome with STAR aligner v2.7.10a (Dobin et al. 2013). To reduce reference mapping bias, we used the WASP algorithm (van de Geijn et al. 2014) implemented within STAR with the –waspOutputMode tag and filtered reads that did not pass WASP filtering. For each sample, we counted allele-specific reads that overlapped with non-duplicated, bi-allelic variants identified in our genotyping data with GATK ASEReadCounter v4.2.5.0 (McKenna et al. 2010). A sample was considered heterozygous for a variant when it had at least 10 total counts from both alleles and at least 2

counts from either allele. We only retained variants with at least 5 heterozygous donors. We further pruned our dataset by stringently correcting for potential genotyping and sequencing errors. We assessed concordance between genotypes called from DNA versus RNA and removed any variants that were discordant between these datasets.

We used two different methods to evaluate allelic imbalance: (1) We used DESeq2 (Love et al. 2014) with a design to account for donor variability and with the ability to detect significant differences in the reference allele versus alternative allele in each condition (~0 + condition:donor + condition:allele). The log2 fold change of alternative allele counts to reference allele counts was calculated with a Wald test and fitType = “mean”. Significant ASE sites were defined as variants with an FDR-adjusted p-value (Benjamini-Hochberg (Benjamini and Hochberg 1995)) below 0.05 and absolute value log2 fold change > log2(1.1). (2) We used ASEP (Fan et al. 2020) to detect AI on a gene-level while taking into account shared information amongst individuals. Haplotype phases for SNPs were obtained from processed genotyping data. Data was prepared for ASEP for the two-condition analysis. Genes were considered to exhibit significantly differential ASE with p-values < 0.05.

For both analyses, AI sites were mapped to positional genes based on their coordinates in the Bioconductor TxDb.Hsapiens.UCSC.hg38.knownGene with the IRanges findOverlaps function (Lawrence et al. 2013), omitting genes that were not single stranded.

Overlap with Boer et al. (2021) OA GWAS

Genome-wide association summary statistics for OA phenotypes identified in Boer et al. (2021) were obtained from the Musculoskeletal Knowledge Portal (Kiel et al. 2020). Positional variants were mapped to rsIDs against the dbSNP155.GRCh37.p13 reference. Nominal significant SNPs in GWAS datasets were defined with a p-value < 0.05. LD proxies for significant AI SNPs were identified using a within study panel and R2 values were calculated with the –ld function in PLINK v1.9 (Purcell et al. 2007). Datasets were compared on the basis of rsID.

3.5 Supplemental Figures and Tables

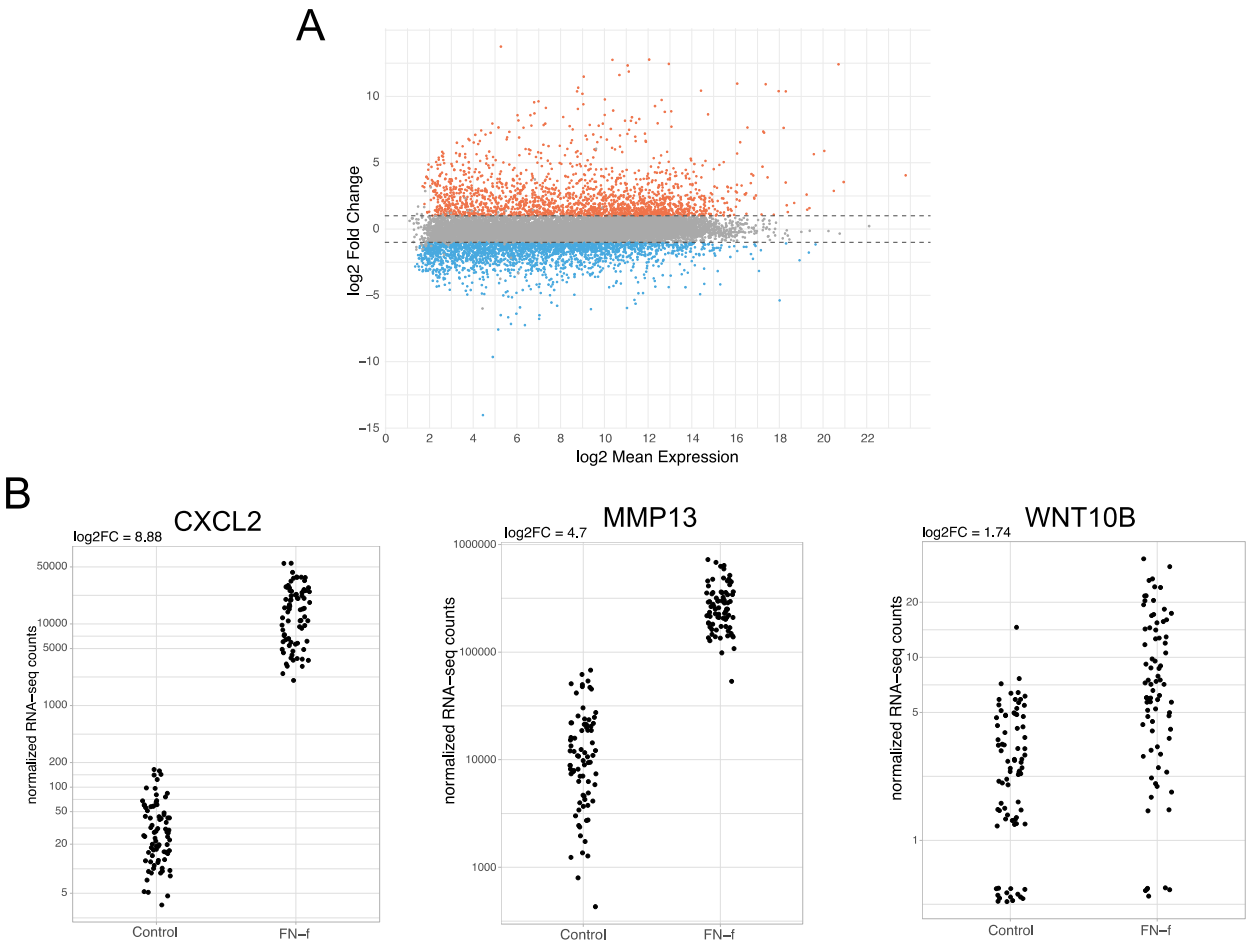

Figure 3.S1. Paired differential expression between control and FN-f treated chondrocytes. (A) MA plot depicting significant upregulated genes in orange (FDR > 0.05; abs(log2FC) > 2) and significant downregulated genes in blue (FDR > 0.05; abs(log2FC) < 2). **(B)** Normalized RNA-seq counts between control and FN-f treated samples for additional examples of genes previously implicated in OA: *CXCL2* (log2FC 8.88, FDR-adjusted p-value < 2.23e-308), *MMP13* (log2FC 4.7, FDR-adjusted p-value = 5.753e-246), and *WNT10B* (log2FC 1.74, FDR-adjusted p-value = 1.947e-16).

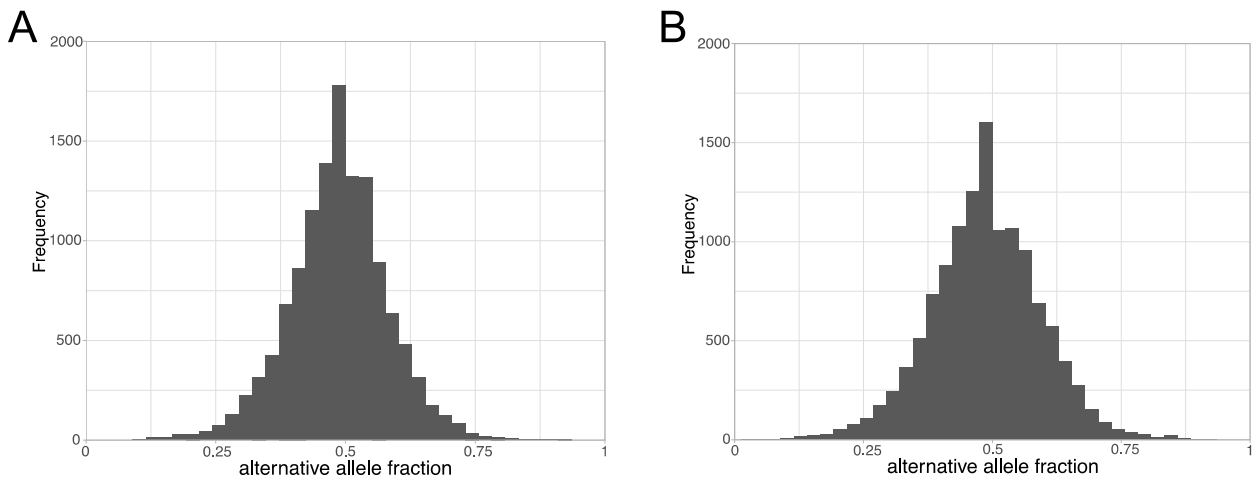

Figure 3.S2. Distributions of allelic imbalance events for all donors. (A) Alternative allele fractions for all control donor samples**. (B)** Alternative allele fractions for all donor samples treated with FN-f.

SNP	Positional Gene	Reference Allele	Alternative Allele	Mean Alternative Allele Fraction	DESeq FDR-adjusted p-value	ASEP p-value
rs385543	*CFH*	A	G	0.41	3.01E-09	0
rs5756130	*MYH9*	C	T	0.43	1.17E-04	0.0496
rs3741588	*RHOF*	A	G	0.39	2.51E-04	0.0385
rs1874340974	*RHOF*	A	G	0.39	2.51E-04	0.0385
rs1044276	*NUP62*	T	C	0.46	9.72E-04	0.00309
rs61364522	*SYNPO*	G	C	0.47	0.001127	0.00397
rs17119665	*IFIT3*	A	G	0.53	0.004071	0.00206
rs17119665	*LOC101926887*	A	G	0.53	0.004071	0.00212
rs17119665	*LIPA*	A	G	0.53	0.004071	0.00331
rs17119665	*LOC105378419*	A	G	0.53	0.004071	0.00347
rs34644316	*ELK3*	C	T	0.6	0.004261	0.022
rs720745	*CCAR2*	G	T	0.43	0.005158	0.00895
rs7301926	*STX2*	T	C	0.44	0.009925	0.0294
rs15049	*RGS5*	T	G	0.44	0.01040	0
rs62222237	*N6AMT1*	A	G	0.43	0.01045	0.00165
rs1804094	*KIAA0040*	C	G	0.42	0.01299	0
rs77410650	*CMAHP*	T	C	0.58	0.03102	1.00E-06
rs11779396	*SLC25A37*	C	G	0.44	0.03666	0.001
rs74984838	*PTGFRN*	A	G	0.56	0.04351	0.00454
rs7215868	*NLRP1*	T	C	0.56	0.04408	0.00855

Table 3.S1. Intersection of significant allelic imbalance SNPs from DESeq2 with positional genes exhibiting significant differential allelic imbalance with ASEP.

Donor ID	Sex	Age	Race
AM7180	M	39	C
AM7181	M	84	C
AM7182	F	65	C
AM7188	M	62	C
AM7189	M	37	C
AM7196	M	73	BL
AM7197	M	50	C
AM7203	M	59	Unknown
AM7204	M	49	BL
AM7205	M	66	C
AM7208	M	38	BL
AM7209	M	63	BL
AM7211	M	55	C
AM7213	M	64	C
AM7214	M	72	C
AM7215	M	66	C
AM7216	F	58	BL
AM7221	F	63	C
AM7223	M	71	C
AM7224	F	61	BL
AM7226	F	65	C
AM7228	M	63	BL
AM7229	F	62	C
AM7230	M	72	BL
AM7236	M	62	C
AM7237	F	58	BL
AM7241	M	65	C
AM7242	M	63	C
AM7243	M	67	C
AM7244	M	71	C
AM7255	M	50	C
AM7256	M	70	C
AM7260	M	71	C
AM7261	F	54	C
AM7266	M	63	C
AM7269	F	57	BL
AM7270	M	67	C
AM7272	M	68	C

AM7273	M	72	BL
AM7277	M	43	C
AM7278	M	51	HISP
AM7280	M	68	C
AM7283	M	57	C
AM7284	M	71	C
AM7285	M	77	HISP
AM7294	M	57	C
AM7295	M	76	Unknown
AM7302	M	70	HISP
AM7303	F	65	C
AM7304	F	54	BL
AM7312	M	61	C
AM7313	M	63	C
AM7318	M	74	C
AM7319	F	49	BL
AM7320	M	45	C
AM7321	M	76	C
AM7323	M	75	BL
AM7325	M	75	C
AM7327	M	55	C
AM7328	M	38	C
AM7329	M	50	C
AM7332	M	34	BL
AM7333	M	69	C
AM7334	M	68	HISP
AM7336	F	61	C
AM7337	M	77	Unknown
AM7343	M	68	C
AM7344	M	65	C
AM7345	M	50	ASIAN
AM7346	M	54	HISP
AM7352	M	72	HISP
AM7353	M	67	C
AM7354	M	50	C
AM7356	F	49	C
AM7359	F	58	C
AM7361	M	46	BL
AM7362	M	72	C

AM7365	M	39	C
AM7372	M	76	ARAB
	Totals	Mean age ± SD years	Totals
	15 Females 64 Males	61 ± 11	1 ARAB 1 ASIAN 3 Unknown 6 HISP 16 BL 52 C

Table 3.S2. Donor sample sexes, ages, and reported ancestry. M = Male; F = Female; ASIAN = Asian; ARAB = Arab; BL = Black; C = Caucasian; HISP = Hispanic.

REFERENCES

Alexa, Adrian, and Jorg Rahnenfuhrer. 2021. *topGO: Enrichment Analysis for Gene Ontology*.

Andrews, S. 2010. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Appleton, C. Thomas G., Claudine G. James, and Frank Beier. 2006. "Regulator of G-Protein Signaling (RGS) Proteins Differentially Control Chondrocyte Differentiation." *Journal of Cellular Physiology* 207 (3): 735–45.

Aubourg, G., S. J. Rice, P. Bruce-Wootton, and J. Loughlin. 2022. "Genetics of Osteoarthritis." *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 30 (5): 636–49.

Aygün, Nil, Angela L. Elwell, Dan Liang, Michael J. Lafferty, Kerry E. Cheek, Kenan P. Courtney, Jessica Mory, et al. 2021. "Brain-Trait-Associated Variants Impact Cell-Type-Specific Gene Regulation during Neurogenesis." *American Journal of Human Genetics* 108 (9): 1647–68.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society* 57 (1): 289–300.

Bhosale, Abhijit M., and James B. Richardson. 2008. "Articular Cartilage: Structure, Injuries and Review of Management." *British Medical Bulletin* 87 (August): 77–95.

Boer, Cindy G., Konstantinos Hatzikotoulas, Lorraine Southam, Lilja Stefánsdóttir, Yanfei Zhang, Rodrigo Coutinho de Almeida, Tian T. Wu, et al. 2021. "Deciphering Osteoarthritis Genetics across 826,690 Individuals from 9 Populations." *Cell* 0 (0). https://doi.org/10.1016/j.cell.2021.07.038.

Bosch, M. H. J. van den, P. L. E. M. van Lent, and P. M. van der Kraan. 2020. "Identifying Effector Molecules, Cells, and Cytokines of Innate Immunity in OA." *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 28 (5): 532–43.

Bos, Steffan D., Judith V. M. G. Bovée, Bouke J. Duijnisveld, Emma V. A. Raine, Wouter J. van Dalen, Yolande F. M. Ramos, Ruud van der Breggen, et al. 2012. "Increased Type II Deiodinase Protein in OA-Affected Cartilage and Allelic Imbalance of OA Risk Polymorphism rs225014 at DIO2 in Human OA Joint Tissues." *Annals of the Rheumatic Diseases* 71 (7): 1254–58.

Choi, Moon-Chang, Jiwon Jo, Jonggwan Park, Hee Kyoung Kang, and Yoonkyung Park. 2019. "NF-κB Signaling Pathways in Osteoarthritic Cartilage Destruction." *Cells* 8 (7). https://doi.org/10.3390/cells8070734.

Collins, J. A., L. Arbeeva, S. Chubinskaya, and R. F. Loeser. 2019. "Articular Chondrocytes Isolated from the Knee and Ankle Joints of Human Tissue Donors Demonstrate Similar Redox-Regulated MAP Kinase and Akt Signaling." *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 27 (4): 703–11.

Coutinho de Almeida, Rodrigo, Margo Tuerlings, Yolande Ramos, Wouter Den Hollander, Eka Suchiman, Nico Lakenberg, Rob G. H. H. Nelissen, Hailiang Mei, and Ingrid Meulenbelt. 2022. "Allelic Expression Imbalance in Articular Cartilage and Subchondral Bone Refined Genome-Wide Association Signals in Osteoarthritis." *Rheumatology* , August. https://doi.org/10.1093/rheumatology/keac498.

Das, Sayantan, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E. Locke, Alan Kwong, Scott I. Vrieze, et al. 2016. "Next-Generation Genotype Imputation Service and Methods." *Nature Genetics* 48 (10): 1284–87.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

Fan, Jiaxin, Jian Hu, Chenyi Xue, Hanrui Zhang, Katalin Susztak, Muredach P. Reilly, Rui Xiao, and Mingyao Li. 2020. “ASEP: Gene-Based Detection of Allele-Specific Expression across Individuals in a Population by RNA Sequencing.” *PLoS Genetics* 16 (5): e1008786.

Forsyth, Christopher B., Judit Pulai, and Richard F. Loeser. 2002. “Fibronectin Fragments and Blocking Antibodies to alpha2beta1 and alpha5beta1 Integrins Stimulate Mitogen-Activated Protein Kinase Signaling and Increase Collagenase 3 (matrix Metalloproteinase 13) Production by Human Articular Chondrocytes.” *Arthritis and Rheumatism* 46 (9): 2368–76.

Gee, Fiona, Clare F. Clubbs, Emma V. A. Raine, Louise N. Reynard, and John Loughlin. 2014. “Allelic Expression Analysis of the Osteoarthritis Susceptibility Locus That Maps to Chromosome 3p21 Reveals Cis-Acting eQTLs at GNL3 and SPCS1.” *BMC Medical Genetics* 15 (May): 53.

Geijn, Bryce van de, Graham McVicker, Yoav Gilad, and Jonathan K. Pritchard. 2014. “WASP: Allele-Specific Software for Robust Discovery of Molecular Quantitative Trait Loci.” *bioRxiv*. https://doi.org/10.1101/011221.

Hatakeyama, Yuji, Rocky S. Tuan, and Lillian Shum. 2004. “Distinct Functions of BMP4 and GDF5 in the Regulation of Chondrogenesis.” *Journal of Cellular Biochemistry* 91 (6): 1204–17.

Hollander, Wouter den, Irina Pulyakhina, Cindy Boer, Nils Bomer, Ruud van der Breggen, Wibowo Arindrarto, Rodrigo Couthino de Almeida, et al. 2019. “Annotating Transcriptional Effects of Genetic Variants in Disease-Relevant Tissue: Transcriptome-Wide Allelic Imbalance in Osteoarthritic Cartilage.” *Arthritis & Rheumatology (Hoboken, N.J.)* 71 (4): 561–70.

Homandberg, G. A. 1999. “Potential Regulation of Cartilage Metabolism in Osteoarthritis by Fibronectin Fragments.” *Frontiers in Bioscience: A Journal and Virtual Library* 4 (October): D713–30.

Homandberg, G. A., C. Wen, and F. Hui. 1998. “Cartilage Damaging Activities of Fibronectin Fragments Derived from Cartilage and Synovial Fluid.” *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 6 (4): 231–44.

Hunter, David J., and Sita Bierma-Zeinstra. 2019. “Osteoarthritis.” *The Lancet* 393 (10182): 1745–59.

Hunter, David J., Lyn March, and Mabel Chew. 2020. “Osteoarthritis in 2020 and beyond: A Lancet Commission.” *The Lancet* 396 (10264): 1711–12.

Iskander, Karim, Jessica Li, Shuhua Han, Biao Zheng, and Anil K. Jaiswal. 2006. “NQO1 and NQO2 Regulation of Humoral Immunity and Autoimmunity *.” *The Journal of Biological Chemistry* 281 (41): 30917–24.

Jiang, Jiahao, Shuaihua Feng, Zexiang Li, Yangqian Luo, Zhenyuan Wang, Mingyang Li, and Guanbao Wu. 2022. “The Expression of MDM2 Gene Promoted Chondrocyte Proliferation in Rats with Osteoarthritis via the Wnt/β-Catenin Pathway.” *Cellular and Molecular Biology* 67 (6): 236–41.

Jun, Goo, Matthew Flickinger, Kurt N. Hetrick, Jane M. Romm, Kimberly F. Doheny, Gonçalo R. Abecasis, Michael Boehnke, and Hyun Min Kang. 2012. “Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data.” *American Journal of Human Genetics* 91 (5): 839–48.

Kania, Karolina, Fabio Colella, Anna H. K. Riemen, Hui Wang, Kenneth A. Howard, Thomas Aigner, Francesco Dell’Accio, Terence D. Capellini, Anke J. Roelofs, and Cosimo De Bari. 2020. “Regulation of Gdf5 Expression in Joint Remodelling, Repair and Osteoarthritis.” *Scientific Reports* 10 (1): 157.

Kao, Wei-Chun, Jian-Chih Chen, Ping-Cheng Liu, Cheng-Chang Lu, Sung-Yen Lin, Shu-Chun Chuang, Shun-Cheng Wu, et al. 2022. “The Role of Autophagy in Osteoarthritic Cartilage.” *Biomolecules* 12 (10). https://doi.org/10.3390/biom12101357.

Kiel, Douglas P., John P. Kemp, Fernando Rivadeneira, Jennifer J. Westendorf, David Karasik, Emma L. Duncan, Yuuki Imai, et al. 2020. “The Musculoskeletal Knowledge Portal: Making Omics Data Useful to the Broader Scientific Community.” *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research* 35 (9): 1626–33.

Krueger, Felix. n.d. *Babraham Bioinformatics - Trim Galore!* Accessed April 6, 2021. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin Morgan, and Vincent Carey. 2013. “Software for Computing and Annotating Genomic Ranges.” *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1003118.

Loeser, Richard F. 2014. “Integrins and Chondrocyte-Matrix Interactions in Articular Cartilage.” *Matrix Biology: Journal of the International Society for Matrix Biology* 39 (October): 11–16.

Loeser, Richard F., Steven R. Goldring, Carla R. Scanzello, and Mary B. Goldring. 2012. “Osteoarthritis: A Disease of the Joint as an Organ.” *Arthritis and Rheumatism* 64 (6): 1697–1707.

Loeser, Richard F., Carol A. Pacione, and Susan Chubinskaya. 2003. “The Combination of Insulin-like Growth Factor 1 and Osteogenic Protein 1 Promotes Increased Survival of and Matrix Synthesis by Normal and Osteoarthritic Human Articular Chondrocytes.” *Arthritis and Rheumatism* 48 (8): 2188–96.

Loh, Po-Ru, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, et al. 2016. “Reference-Based Phasing Using the Haplotype Reference Consortium Panel.” *Nature Genetics* 48 (11): 1443–48.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Love, Michael I., Charlotte Soneson, Peter F. Hickey, Lisa K. Johnson, N. Tessa Pierce, Lori Shepherd, Martin Morgan, and Rob Patro. 2020. “Tximeta: Reference Sequence Checksums for Provenance Identification in RNA-Seq.” *PLoS Computational Biology* 16 (2): e1007664.

Lu, Xianghuai, Linda Gilbert, Xiaofei He, Janet Rubin, and Mark S. Nanes. 2006. “Transcriptional Regulation of the Osterix (Osx, Sp7) Promoter by Tumor Necrosis Factor Identifies Disparate Effects of Mitogen-Activated Protein Kinase and NFκB Pathways *.” *The Journal of Biological Chemistry* 281 (10): 6297–6306.

MacGregor, A. J., and T. D. Spector. 1999. “Twins and the Genetic Architecture of Osteoarthritis.” *Rheumatology* 38 (7): 583–88.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303.

Molnar, Vilim, Vid Matišić, Ivan Kodvanj, Roko Bjelica, Željko Jeleč, Damir Hudetz, Eduard Rod, et al. 2021. “Cytokines and Chemokines Involved in Osteoarthritis Pathogenesis.” *International Journal of Molecular Sciences* 22 (17). https://doi.org/10.3390/ijms22179208.

Mortier, Geert R., Daniel H. Cohn, Valerie Cormier-Daire, Christine Hall, Deborah Krakow, Stefan Mundlos, Gen Nishimura, et al. 2019. “Nosology and Classification of Genetic Skeletal Disorders: 2019 Revision.” *American Journal of Medical Genetics. Part A* 179 (12): 2393–2419.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19.

Patterson, Nick, Alkes L. Price, and David Reich. 2006. “Population Structure and Eigenanalysis.” *PLoS Genetics* 2 (12): e190.

Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. “Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies.” *Nature Genetics* 38 (8): 904–9.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75.

Rai, M. F., E. D. Tycksen, L. Cai, J. Yu, R. W. Wright, and R. H. Brophy. 2019. “Distinct Degenerative Phenotype of Articular Cartilage from Knees with Meniscus Tear Compared to Knees with Osteoarthritis.” *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 27 (6): 945–55.

Raine, Emma V. A., Andrew W. Dodd, Louise N. Reynard, and John Loughlin. 2013. “Allelic Expression Analysis of the Osteoarthritis Susceptibility Gene COL11A1 in Human Joint Tissues.” *BMC Musculoskeletal Disorders* 14 (March): 85.

Reed, K. S. M., V. Ulici, C. Kim, S. Chubinskaya, R. F. Loeser, and D. H. Phanstiel. 2021. “Transcriptional Response of Human Articular Chondrocytes Treated with Fibronectin Fragments: An in Vitro Model of the Osteoarthritis Phenotype.” *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society* 29 (2): 235–47.

Reynard, Louise N., Catherine Bui, Catherine M. Syddall, and John Loughlin. 2014. “CpG Methylation Regulates Allelic Expression of GDF5 by Modulating Binding of SP1 and SP3 Repressor Proteins to the Osteoarthritis Susceptibility SNP rs143383.” *Human Genetics* 133 (8): 1059–73.

Snelgrove, T. A., L. J. Peddle, C. Stone, F. Nofball, D. Peddle, D. Squire, P. Rockwood, et al. 2005. “Association of COL1A2, COL2A1 and COL9A1 and Primary Osteoarthritis in a Founder Population.” *Clinical Genetics* 67 (4): 359–60.

Southam, Lorraine, Julio Rodriguez-Lopez, James M. Wilkins, Manuel Pombo-Suarez, Sarah Snelling, Juan J. Gomez-Reino, Kay Chapman, Antonio Gonzalez, and John Loughlin. 2007. “An SNP in the 5’-UTR of GDF5 Is Associated with Osteoarthritis Susceptibility in Europeans and with in Vivo Differences in Allelic Expression in Articular Cartilage.” *Human Molecular Genetics* 16 (18): 2226–32.

Sun, Kai, Jiachao Guo, Xudong Yao, Zhou Guo, and Fengjing Guo. 2021. “Growth Differentiation Factor 5 in Cartilage and Osteoarthritis: A Possible Therapeutic Candidate.” *Cell Proliferation* 54 (3): e12998.

Tachmazidou, Ioanna, Konstantinos Hatzikotoulas, Lorraine Southam, Jorge Esparza-Gordillo, Valeriia Haberland, Jie Zheng, Toby Johnson, et al. 2019. “Identification of New Therapeutic Targets for Osteoarthritis through Genome-Wide Analyses of UK Biobank Data.” *Nature Genetics* 51 (2): 230–36.

Tenenbaum, Dan, and Bioconductor Package Maintainer. 2021. *KEGGREST: Client-Side REST Access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)*.

Umans, Benjamin D., Alexis Battle, and Yoav Gilad. 2021. “Where Are the Disease-Associated eQTLs?” *Trends in Genetics: TIG* 37 (2): 109–24.

Wilson, Gabrielle R., Jasmine Sunley, Katherine R. Smith, Kate Pope, Catherine J. Bromhead, Elizabeth Fitzpatrick, Maja Di Rocco, et al. 2014. “Mutations in SH3PXD2B Cause Borrone Dermato-Cardio-Skeletal Syndrome.” *European Journal of Human Genetics: EJHG* 22 (6): 741–47.

Wood, Scott T., David L. Long, Julie A. Reisz, Raghunatha R. Yammani, Elizabeth A. Burke, Chananat Klomsiri, Leslie B. Poole, Cristina M. Furdui, and Richard F. Loeser. 2016. “Cysteine-Mediated Redox Regulation of Cell Signaling in Chondrocytes Stimulated With Fibronectin Fragments.” *Arthritis & Rheumatology (Hoboken, N.J.)* 68 (1): 117–26.

Xie, D. L., R. Meyers, and G. A. Homandberg. 1992. “Fibronectin Fragments in Osteoarthritic Synovial Fluid.” *The Journal of Rheumatology* 19 (9): 1448–52.

Xie, Junxiong, Zhiqin Deng, Murad Alahdal, Jianquan Liu, Zhe Zhao, Xiaoqiang Chen, Guanghui Wang, et al. 2021. “Screening and Verification of Hub Genes Involved in Osteoarthritis Using Bioinformatics.” *Experimental and Therapeutic Medicine* 21 (4): 330.

Yerges-Armstrong, Laura M., Michelle S. Yau, Youfang Liu, Subha Krishnan, Jordan B. Renner, Charles B. Eaton, C. Kent Kwoh, et al. 2014. “Association Analysis of BMD-Associated SNPs with Knee Osteoarthritis.” *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research* 29 (6): 1373–79.

CHAPTER 4: RESPONSE EQTL ANALYSIS TRANSLATES OSTEOARTHRITIS GWAS LOCI INTO PUTATIVE RISK GENES

4.1 Introduction

In the monumental effort to study the genetic architecture of complex traits and diseases, genome-wide association studies (GWAS) have identified numerous genomic loci associated with one or more phenotypes (Uffelmann et al., 2021). However, translating these associated variants into molecular mechanisms of disease risk remains difficult in large part since the majority of GWAS variants reside in non-coding regions of the genome and most likely perturb gene expression through the disruption of regulatory elements (Alexander et al., 2010). Understanding this genetic regulation is further complicated by the presence of numerous variants within risk loci in high linkage disequilibrium (LD) and many genes, making it hard to identify the putative causal variants and affected genes (Cano-Gamez & Trynka, 2020). For many diseases, there is a clear and urgent need to resolve these genetic signals and understand which candidates can be targeted by actionable therapeutics.

Expression quantitative trait loci (eQTL) mapping is a powerful statistical tool used to connect genetic variation with gene expression variation to further annotate disease-associated variants and suggest potential gene regulatory mechanisms. Since gene expression is an intermediate link between an organism's DNA sequence and the observed phenotype, leveraging eQTL studies in combination with GWAS through colocalization can elucidate the potential gene regulatory mechanisms by which a variant affects a disease (Hormozdiari et al., 2016). Many studies, notably including the Genotype-Tissue Expression Project (GTEx) (GTEx Consortium, 2013), have mapped eQTLs in a vast array of tissues in an effort to contribute to the understanding of many traits and diseases such as obesity (Smemo et al., 2014) and Alzheimer's disease (Schwartzentruber et al., 2021). Despite the large collection of available eQTL datasets, many eQTLs do not colocalize with GWAS and have thus failed to explain a large proportion of disease-associated genetic variation (Manolio et al., 2009).

A possible explanation for the inability to map disease-associated variants with standard eQTL studies is a lack of datasets not only utilizing the appropriate cell type but also capturing the dynamic

nature of gene regulation, particularly in the context of disease-specific stimuli. Thus, response eQTL (reQTL) studies are necessary to capture context-specific effects of genetic variation on gene expression before and after a disease-relevant trigger or treatment (Umans et al., 2021). For example, one recent reQTL study in macrophages identified immune response eQTLs that colocalized with disease risk loci only after stimulation with IFNγ and/or Salmonella (Alasoo et al., 2018). The number of reQTL studies is limited, and reQTL studies have not been carried out for the vast array of tissues and diseases to the same extent that standard eQTL studies have been conducted.

Osteoarthritis (OA) is one such disease lacking a strong understanding in the genetic underpinnings and mechanisms of the disease (Aubourg et al., 2022). OA GWAS have identified 100 independently associated risk variants across 11 OA phenotypes (Boer et al., 2021; Tachmazidou et al., 2019), and recent OA QTL studies have identified methylation QTLs (Rice et al., 2022) and standard eQTLs. In particular, one study conducted a standard eQTL analysis in low-grade (intact, preserved), high-grade (degraded, lesioned), and synovial joint tissues, finding strong evidence for OA signals with non-coding eQTLs for *ALDH1A2*, *NPC1*, *SMAD3*, *FAM53A*, and *SLC44A2* (Steinberg et al., 2021). While this study identified “differential eQTLs” for various genes, it did not capture the dynamics of OA progression and only assessed disease endpoints. Here we leverage a validated model of OA transcription and inflammation using fibronectin fragment (FN-f) treated chondrocytes (Forsyth et al., 2002; Homandberg, 1999; Pulai et al., 2005; Reed et al., 2021) to perform the first response QTL analysis in an OA-relevant system. Using paired control and FN-f treated samples of articular chondrocytes extracted from human donors, we robustly map reQTLs to identify eGenes specific to resting chondrocytes and eGenes specific to FN-f treated chondrocytes and characterize dynamic genetic regulation. We identified 264 control reQTLs and 384 FN-f reQTLs and connected them to genes previously implicated in OA like *SMAD3* as well as novel response eGenes. Furthermore, colocalization with OA GWAS revealed strong evidence for *SMAD3* as a likely effector gene driving the signals in multiple OA GWAS phenotypes. These findings within a controlled in vitro model of OA provide translational opportunities for the functional probing and ultimate development of OA treatments.

4.2 Results

4.2.1 Study design and gene expression profiling

To establish a robust reQTL design, we collected primary articular chondrocytes from the talar cartilage of 79 human donors (see **Table 4.1** for donor characteristics) and extracted paired samples to be treated with either FN-f or PBS as a control. With both control and OA-stimulated samples coming from paired donors, we are able to perform analyses with matched genetic backgrounds as well as capture the dynamic transcriptomic changes of OA progression. We profiled genotype information for each donor and RNA-seq data for each sample, generating genome-wide data to define condition-specific reQTLs (**Figure 4.1A**).

Before performing eQTL mapping with our dataset, we first assessed the transcriptomic profiles of control and FN-f treated chondrocyte samples to assess which factors are driving major expression changes. First, we confirmed the validity of FN-f chondrocytes as an appropriate model in recapitulating OA-relevant transcriptomic changes (see Chapter 3.2.1), confirming upregulation and downregulation of genes implicated in OA including *WNT10B*, *MMP13*, *CXCL2*, *GDF5*, and *COL2A1*. Sample clustering of differential gene expression was driven by FN-f treatment as opposed to the donor-specific factors age, sex, and race (**Figure 4.1B**). Furthermore, PBS-treated and FN-f-treated samples clustered separately by principal component analysis (PCA) of global gene expression (**Figure 4.1C**, **Figure 4S.1A**), reinforcing that global transcriptomic differences are mainly condition-specific. Furthermore, we confirmed sufficient replication of gene expression results across donors from each condition (**Figure 4S.1B**), ensuring the robustness of our study system.

Number of donors

Sex	Male	Female			
	64	15			
Age	34 – 44	45 – 55	56 – 66	67 – 77	78 – 88
	7	16	27	28	1
Ancestry	African	American	East Asian	European	South Asian
	17	40	0	8	5

Table 4.1. Study donor characteristics.

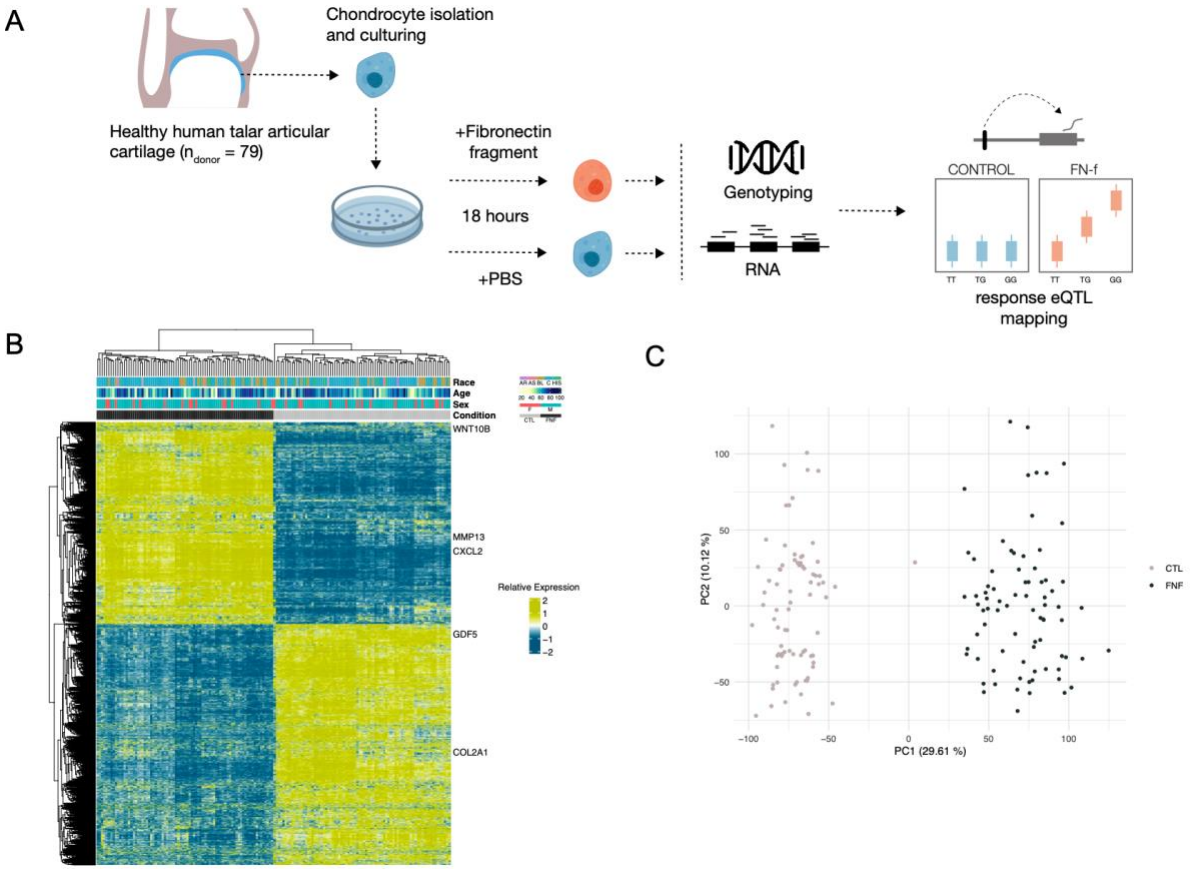

Figure 4.1. reQTL study design and gene expression profiling. (A) Human chondrocytes were isolated from 79 donors of healthy human talar articular cartilage and either treated with either PBS as a control or fibronectin fragment (FN-f) to simulate a cell-type-relevant osteoarthritis response. Imputed genotypes and gene expression were used to perform local response eQTL mapping to discover disease stimulus-specific reQTLs. **(B)** Gene expression profile of transcriptome-wide gene expression changes after FN-f treatment. Genes were clustered by their relative expression. Previously implicated OA genes are labeled. Samples are labeled for donor race, donor age, donor sex, and condition. **(C)** Principal component analysis of donor samples from each condition indicates treatment-specific clustering. Control samples are colored in light grey and FN-f treated samples are colored in dark grey.

4.2.2 Determining covariates for modeling

To ensure appropriate correction for confounding variation in our datasets prior to local eQTL mapping, we performed a methodical assessment of correlations and eGene calling power using a variety of covariates within our model. In both conditions, we first carried out correlation analyses between gene expression principal components and recorded technical factors for donors as well as sample preparations ranging from DNA or RNA extraction kit batches to FN-f treatment batch. Both sample groups showed similar, highly significantly correlated variables with the first principal component, suggesting some kind of related batch effects between RNA extraction kit batch, sequencing batch, DNA reagent batch, and genotyping batch (**Figure 4.2**). Assuming RNA extraction kit batch/sequencing batch

and DNA reagent batch/genotyping batch were almost redundant batch effects, we corrected our RNA-seq data for RNA extraction kit batch and DNA reagent batch and saw a reduction in significant correlations of batch effects with variation in our gene expression data (**Figure 4.S2A**). Thus, we determined RNA extraction kit batch and DNA reagent batch to be essential covariates in our eQTL model.

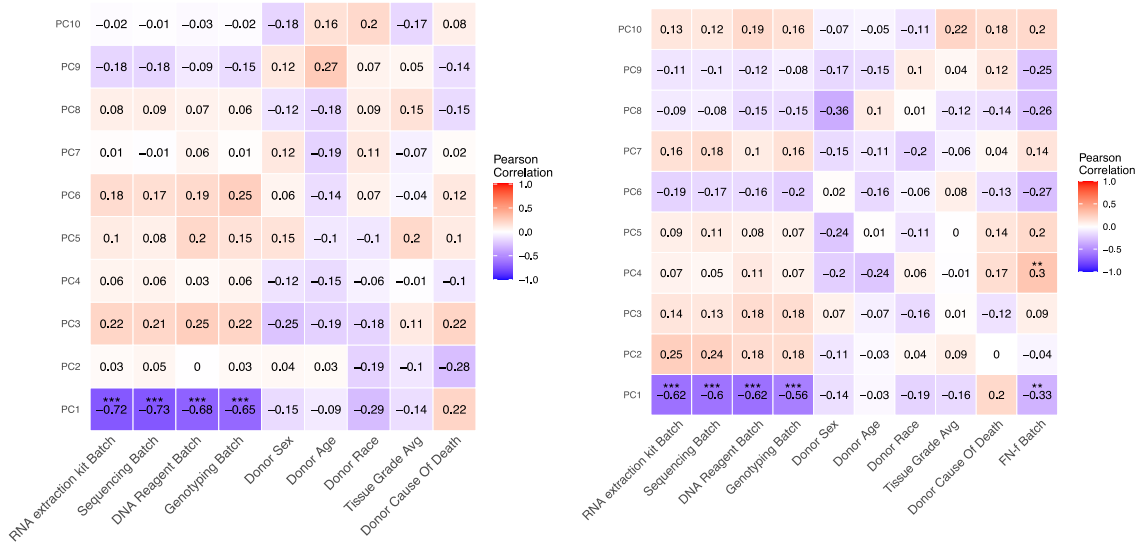

Figure 4.2. Correlation analysis between gene expression principal components and technical factors. Pearson's correlations were calculated between potential technical cofounders (x-axis) and the top 10 gene expression PCs (y-axis), separately for control (left) and FN-f treated (right) samples. Asterisks indicate significant correlation (** p-value <= 0.01; *** p-value < 0.001).

To correct for other potential hidden sources of variation within our data, we simultaneously explored the use of RNA-seq principal components versus PEER factors (Stegle et al., 2012). In one comparison, we calculated principal components (PCs) and determined which number to include based on the percent variance explained of each PC (**Figure 4S.2B**). We performed *cis* eQTL mapping with a permutation pass and calculated the number of significant eGenes obtained from mapping with correction for PC's and various combinations of other covariates. Simultaneously, we calculated 10 to 60 PEER factors in iterations of 10 and compared the number of significant eGenes obtained when using similar combinations of other covariates. For all combinations, we obtained within a range of 1250 to 2250 significant eGenes after global multiple testing correction with both the Storey q-value (Storey, 2003) and Benjamini-Hochberg false discovery rate (FDR) (Benjamini & Hochberg, 1995) (**Figure 4.3A**). As to be expected, we saw fewer significant eGenes when including batch effect, genotype ancestry, and higher numbers either of PCs or PEER factors. Although we observed variable trends when comparing the

inclusion of PCs versus PEER factors, PEER factors generally produced more conservative eGene results, so we proceeded with PEER factors in our subsequent analyses.

In a similar comparison, we refined the number of PEER factors included (using only 5, 10, 15, or 20), included 2 genotyping principal components to correct for ancestry, and further explored the effect of including various types of batch effects in our model on the number of significant eGenes obtained from local eQTL mapping (**Figure 4.3B**). Based on this analysis and the previous correlation analyses, we concluded that the use of 10 PEER factors with RNA extraction kit batch and DNA extraction kit batch yielded the appropriate amount of correction in our linear model without overcorrecting or including redundant batch effects. We also confirmed that our PEER factors did not correlate with each other (**Figure 4.S2C**). Thus, the final covariates included in our model include donor sex, 2 genotyping PCs, 10 PEER factors, RNA extraction kit batch, and DNA extraction kit batch.

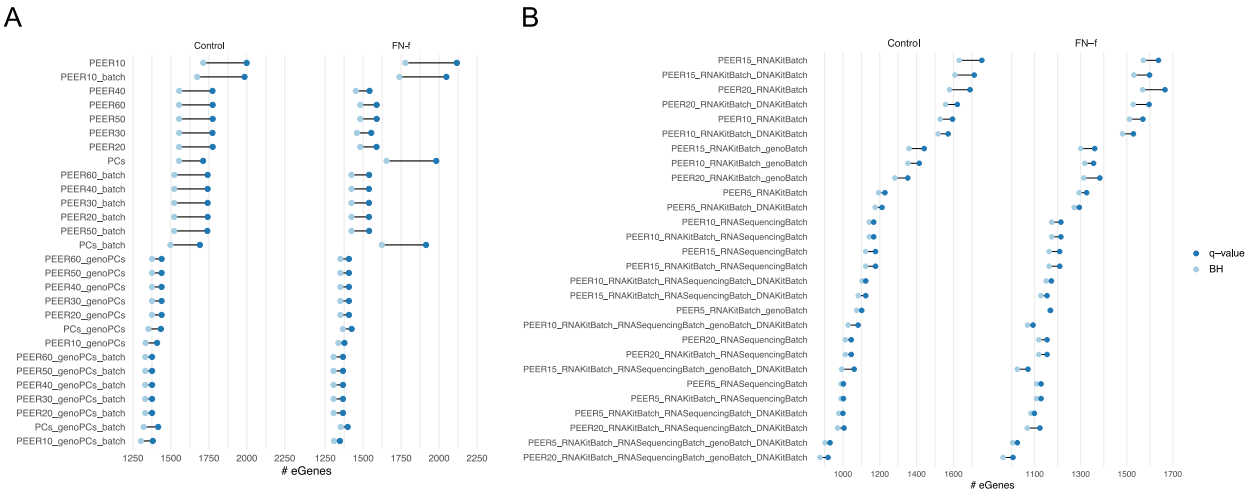

Figure 4.3. Covariate selection analysis for local eQTL mapping. (A) Number of eGenes called from eQTL mapping in separate conditions using different linear models comparing PC correction and inclusion of differing number of PEER factors. 15 gene expression PCs were used in any model with PC covariates. **(B)** Comparison of number of eGenes called from local eQTL mapping in separate conditions using 5 to 20 PEER factors and correcting for different potential batch effects. For each eGene dataset, the same dataset was corrected with either the Storey q-value (dark blue) or Benjamini-Hochberg FDR (light blue).

4.2.3 Local eQTL mapping

We identified *cis* eQTLs using the aforementioned covariates separately in control samples and FN-f-treated samples. We identified 1517 significant control eGenes (Benjamini-Hochberg FDR < 0.05) and 1482 significant FN-f eGenes, with 667 eGenes specific to control samples, 632 eGenes specific to FN-f samples, and 850 eGenes found in both conditions (**Figure 4.4A**). When looking at the effect sizes

of lead eGene-eSNP pairs that were detected as significant with FDR < 0.05 in both conditions, all show a concordant direction of effect (**Figure 4.4B**). Similarly, nominally significant eQTLs in both conditions also show concordant directions of effect. Nominally significant eQTLs detected in only one condition also show an overall concordance of directions of effect, though with less strong linear relationships which is to be expected when eQTLs are not detected as significant in the opposite condition (**Figure 4.4C**). In both conditions, the highest frequency of nominally significant eQTLs lie within 250 Kb of the eGene transcription start site (TSS) (**Figure 4.4D**).

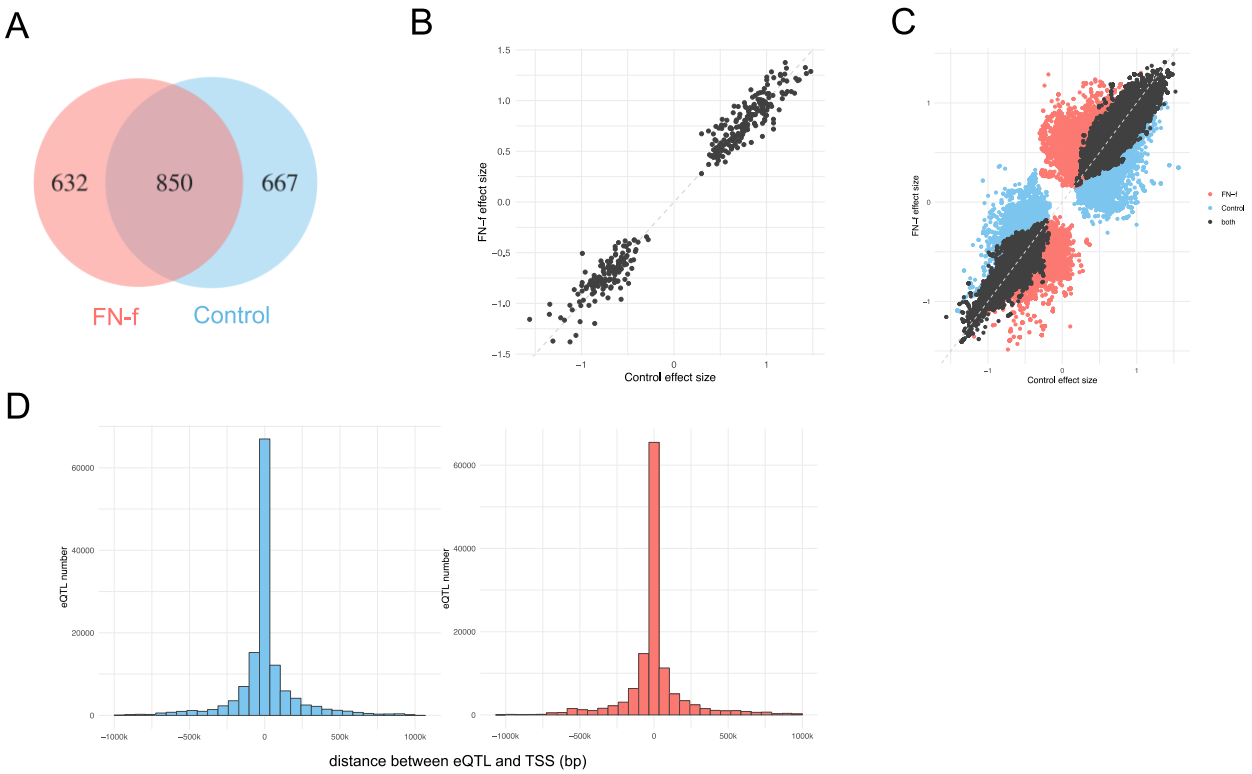

Figure 4.4. Features of condition-specific local eQTLs. (A) Venn diagram illustrating number of significant eGenes identified in the control eQTL dataset, FN-f eQTL dataset, or both datasets. **(B)** Effect sizes of significant lead eGene-eSNP pairs identified in both control and FN-f eQTL datasets. **(C)** Effect sizes of nominally significant eQTLs identified in both (grey points), control (blue points), or FN-f (orange) datasets. In **(B)** and **(C)**, x-coordinates represent effect size in control and y-coordinates represent effect size in FN-f. Dashed line represents y = x. **(D)** Frequency of distances between significant eGene TSS and their nominally significant eQTLs in control (left) and FN-f (right) datasets.

To confirm cell-type specificity and condition specificity of our identified eQTLs, we performed a series of π_1 comparisons (see Materials and Methods) to determine fractions of lead eGene-eSNP pairs that were true associations in other datasets. First, we assessed sharing with eQTLs identified in GTEx tissues for both our control and FN-f eQTLs (**Figure 4.5A**). GTEx does not include any OA-specific tissues or cell types, but highest sharing of eGene-eSNP pairs was observed in somewhat similar or

related tissues including tibial artery tissue, subcutaneous adipose, and cell cultured fibroblasts. However, among the top 10 GTEx tissues with highest sharing of identified eQTLs, 7 of these tissues (tibial artery, subcutaneous adipose, skin sun exposed lower leg, tibial nerve, thyroid, skin not sun exposed suprapubic, and skeletal muscle) were among the top 10 largest GTEx sample sizes, with sizes ranging from 517 to 706 RNA-seq and genotyped samples. Thus, the highest sharing with these tissues may be attributed to sample size and not necessarily relevance of tissue type. Next, we assessed concordance of our control and FN-f eQTLs with those identified in Steinberg et al. (2021) in low-grade and high-grade diseased cartilage (**Figure 4.5B**). Overall, our lead FN-f eGene-eSNP pairs showed slightly higher π_1 values than our control eGene-eSNP pairs against both cartilage grades, which suggests the FN-f treatment is able to capture more disease relevant eQTLs. Lastly, we compared our control and FN-f eGene-eSNP pairs against each other and only observed π_1 values below 0.4 (**Figure 4.5C**), perhaps suggesting the presence of many condition-specific eQTLs.

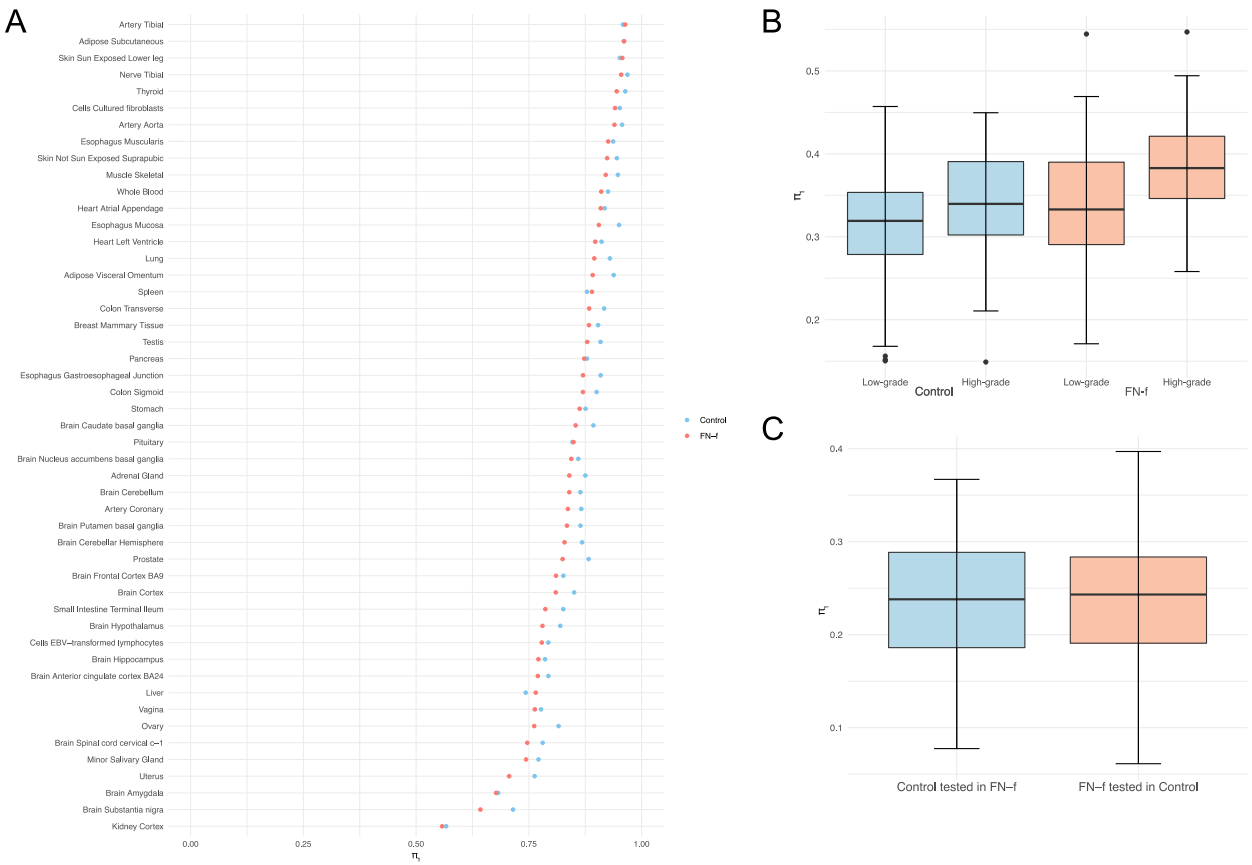

Figure 4.5. eQTL sharing between control and FN-f eQTLs and previously published datasets. (A) The fraction of control (blue) and FN-f (orange) eGene-eSNP pairs that are true associations (π_1) with eGene-eSNP pairs in GTEx tissues. **(B)** π_1 values for eGene-eSNP pairs in control and FN-f conditions with eQTLs identified in low-grade and high-grade cartilage in Steinberg et al. (2021). **(C)** eQTL sharing across conditions assessed via π_1. Box plots show π_1 values calculated from 100 samplings of the uniform distribution to account for eGene-eSNP pairs not found in both compared datasets.

4.2.4 Response eQTLs

To determine OA stimulus-specific effects of genetic variation on gene expression, we identified response eQTLs that only showed a significant interaction effect of donor genotype with condition either before or after FN-f treatment. We found 264 response eGenes specific to control samples and 384 response eGenes specific to FN-f treated samples. We observed FN-f reQTLs with both directions of effect, with either the risk allele being associated with higher gene expression, as with the G allele at rs12901081 being associated with higher *SMAD3* expression (log2aFC 1.24) **(Figure 4.6A)**, or the risk allele being associated with lower gene expression, as seen at rs8011143 where the C allele is associated with lower *ABHD4* expression (log2aFC -1.79) (**Figure 4.6B**). Numerous control and FN-f response eGenes had multiple significant reQTLs (**Figure 4.6C**). Furthermore, we overlapped our significant response eGenes with differential gene expression information from Chapter 3 and found 177

differentially regulated genes that intersected with our FN-f response eGenes, where 69 were downregulated and 108 were upregulated. One notable example was the differentially downregulated response eGene *DIO2*, which has been cited as an OA susceptibility gene (Bomer et al., 2015; Bos et al., 2012; Goldring, 2013).

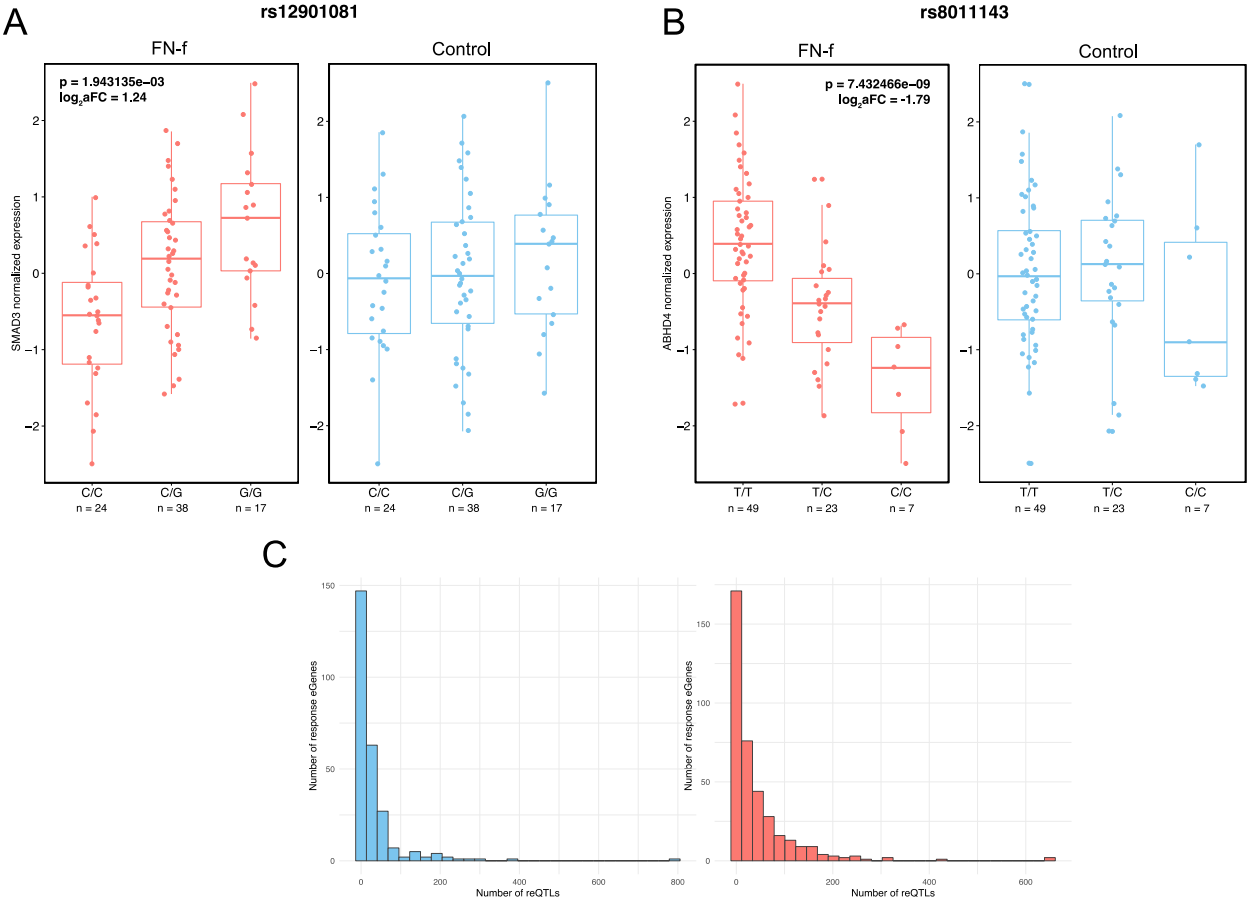

Figure 4.6. Response eQTLs with significant associations in control and FN-f conditions.. **(A)** Example of an FN-f reQTL in a previously implicated OA gene whereby the risk allele is associated with increasing gene expression. **(B)** Example of an FN-f reQTL whereby the risk allele is associated with decreasing gene expression of the response eGene. In **(A)** and **(B)** boxplots show normalized expression for each donor, separated by genotype. n: number of donors for each genotype; p: FDR-corrected p-value of condition-specific interaction term; log_2aFC: log_2 allelic fold change of the reQTL. The same expression is shown for the variant showing non-significant effects in control. **(C)** Frequency of number of nominally significant reQTLs in response eGenes discovered in control (left) and FN-f (right).

4.2.5 Colocalization of OA GWAS and FN-f reQTLs

In order to identify the likely effector genes driving OA GWAS signals, we performed colocalization analysis in regions where we identified significant FN-f reQTLs with the most recent OA GWAS meta-analysis, which comprises 13 international cohorts where 2 are of East Asian descent and 11 cohorts are of European descent. Colocalization is a powerful method to integrate eQTL and GWAS

datasets and determine the likelihood that the same variant or genetic signal underpins both the associations with gene expression and disease. We found strong evidence for colocalization of one FN-f response eQTL signal for *SMAD3* with 7 OA GWAS phenotypes, with the signals for All OA, KneeHip OA, Hip OA, total joint replacement (TJR), and total hip replacement (THR) reaching genome-wide significance (**Figure 4.7, Figure 4.S3**). In each of the GWAS signals, the index variant is rs12908498, which is in high LD (r^2 = 0.9) with the lead reQTL variant rs12901081. Both of these variants lie within non-coding, intronic regions of *SMAD3*. A recent cartilage eQTL study (Steinberg et al., 2021) also found strong evidence for colocalization at this region with the index SNP rs12901372, which is also in high LD (r^2 = 0.9) with both the GWAS lead SNP and reQTL lead SNP.

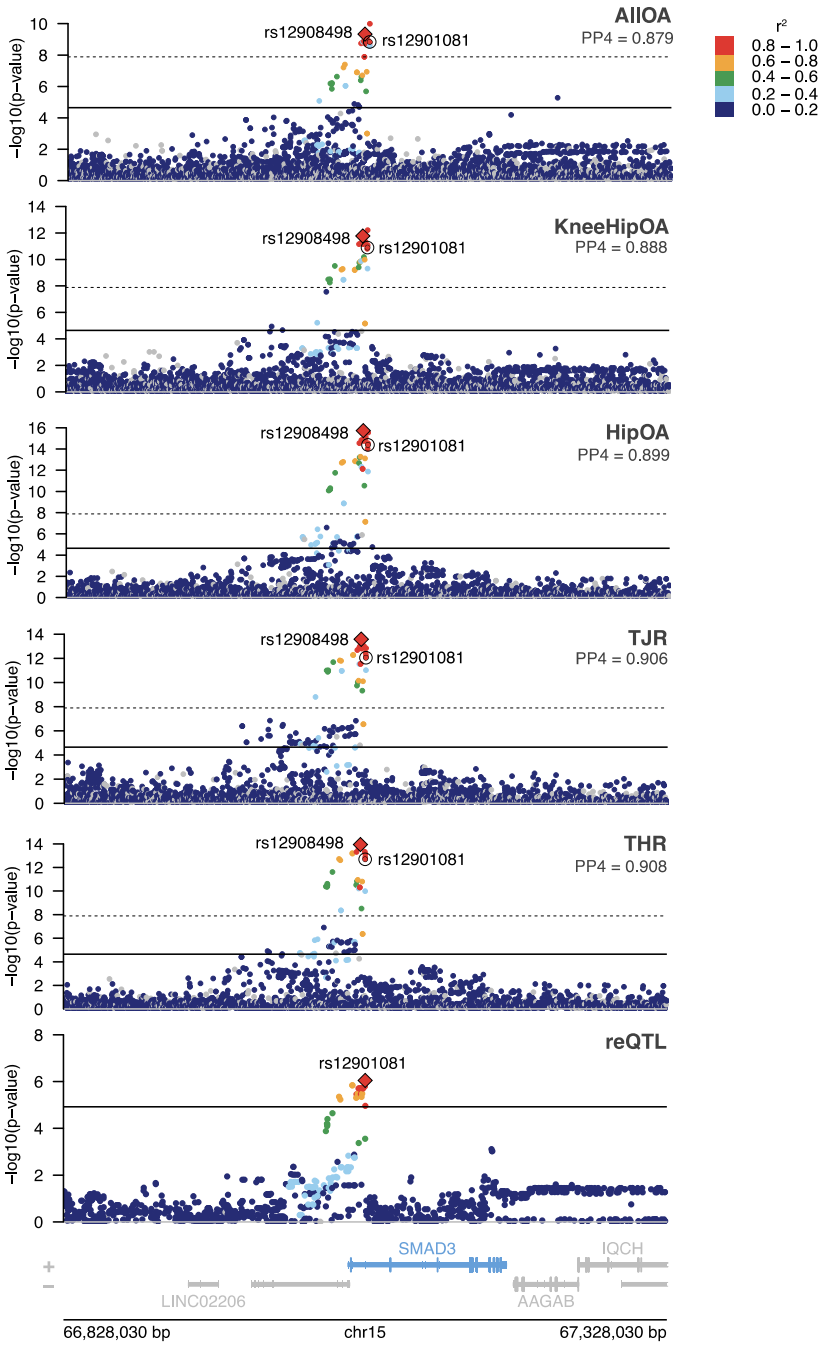

Figure 4.7. An FN-f reQTL shows strong evidence for colocalization with a genetic signal in multiple OA GWAS phenotypes. Manhattan plots of All OA, KneeHip OA, Hip OA, total joint replacement (TJR), and total hip replacement (THR) GWAS and an FN-f reQTL signal. Associations are depicted with -log10 p-values on the y-axis and the genomic location of each variant on the x-axis. GWAS plots are colored by pairwise linkage disequilibrium (LD, r^2) to the index SNP rs12908498 (red diamond). reQTL plot is colored by pairwise LD to the index reQTL SNP rs12901081 (red diamond). reQTL index SNP rs12901081 is circled in each GWAS plot. reQTL p-values show associations of SNPs with *SMAD3* gene expression. In GWAS tracks, dashed lines represent the genome-wide significance threshold at 1.3e-8 and solid lines represent the nominal significance threshold at 2.27e-5. Solid line in reQTL track represents the nominal significance threshold for the *SMAD3* eGene at.1.201e-5. PP4 = posterior probability of colocalization.

4.3 Discussion

Despite the recent rapid influx of standard eQTL studies, a high fraction of disease GWAS loci fail to colocalize with and be explained by known eQTLs. In particular, not only has eQTL mapping of OA-relevant systems been absent from large scale consortia, but standard QTL mapping in cartilage has still left gaps in addressing the functionality of OA-associated genetic loci. One possible explanation for not capturing all GWAS loci with current eQTLs is that these eQTLs must be captured in response to disease-appropriate stimuli. Here we perform the first OA response eQTL study, using 79 paired donor samples either unstimulated or stimulated with FN-f, a known OA trigger. By capturing the inflammatory response of chondrocytes, we identify eQTL signals that mark disease progression-dependent patterns of gene regulation that are driven by genetic variation.

To perform successful eQTL mapping, it is important to properly control any confounding factors to increase power and capture only the true genetic effects on gene expression. Our study followed a methodical approach to covariate selection, thereby ensuring robust modeling with results optimizing discovery and controlling false positives. Although there is no gold standard for appropriately determining necessary covariates, our framework attempts to transparently assess any possible recorded and hidden sources of confounding information. It is unclear how well previous studies controlled for covariates as these factors will vary from dataset to dataset, but this essential step in eQTL studies may serve as an additional checkmark for ensuring the detection of all accurate signals that could then colocalize with GWAS loci.

To capture cell-type-specific or disease-state-specific eQTLs, many standard eQTL studies map eQTLs separately and compare the overlaps of signals from each sample group. While this approach may be sufficient to capture strongly divergent patterns of genetic regulation, it may not be able to discern more subtle differences, especially using study designs that do not correct for inter-individual confounding or cell state heterogeneity. Before specifically testing for reQTLs, we identified 667 eGenes specific to control samples and 632 specific to FN-f samples, but these numbers decreased to 264 and 384 response eGenes, respectively, after reQTL mapping. Although standard eQTLs in each condition had high overlap with GTEx eQTLs in cartilage-related tissues, the large sample sizes of GTEx capture many signals without necessarily differentiating between context or deciphering regulatory mechanisms. Our

controlled study design allowed us to specifically test the combined effect of genotype and condition on gene expression, revealing the true condition-specific eQTLs. We hypothesize that we capture robust eQTL signals that are distinguished by OA progression while also observing many diseased cartilage signals, as seen in our dataset overlaps with a previous cartilage eQTL study.

Colocalization analysis identified a lead reQTL for the *SMAD3* gene with genome-wide significant GWAS signals for 5 OA phenotypes (All OA, KneeHip OA, Hip OA, TJR, and THR) and nominally significant GWAS signals for 2 OA phenotypes (Knee OA and Spine OA). Previous studies have either colocalized GTEx eQTLs, which do not include cartilage or chondrocytes, or standard diseased cartilage eQTLs with OA GWAS from smaller study sizes. Here we have colocalized FN-f reQTLs with OA GWAS from the latest and largest meta-analysis, investigating context-specific colocalizations with the most available GWAS associations to date. Although colocalization analysis alone is insufficient to determine causal genes at GWAS loci, we have strong evidence to prioritize *SMAD3* as a gene candidate in future studies. The reQTL signal is only strongly associated with *SMAD3* after FN-f treatment, with the G allele of rs12901081 associated with increasing expression of *SMAD3*. This signal's colocalization with GWAS signals in numerous OA phenotypes suggests the shared importance of genetic variation at this locus in potentially contributing to OA at many different joints. Furthermore, a colocalization at this signal was also identified in a previous study where an eQTL only identified in high-grade cartilage colocalized with OA GWAS. This agreement further confirms the validity of the chondrocyte FN-f model in identifying relevant eQTLs, and the robustness of this signal's colocalization across multiple OA disease-relevant eQTL and GWAS datasets highlights its potential relevance as an effector gene for OA. Lastly, *SMAD3* functions in the transforming growth factor-beta signaling pathway (TGF-beta), which has been shown to be required for maintaining adulthood cartilage homeostasis with TGF-beta signaling disrupted during OA progression (Finnson et al., 2012; van der Kraan et al., 2009), further confirming its disease relevance.

Although these data reveal insights into the context-dependent regulatory landscape of OA gene expression, this work has limitations. Our work uses a model of OA inflammation and thus may not capture all disease-specific gene expression changes affected by genetic variation, particularly at the later stages of disease. Furthermore, OA is a disease of the whole joint and may be affected by multiple tissues and cell types, whereas we are using a single cell type in cartilage to simulate the OA phenotype.

However, the use of the model system allows for a paired reQTL study that limits any inter-individual confounding effects and captures OA-relevant genetic changes that can elucidate the cell type specific mechanisms towards the beginning of the disease, which is when novel treatments would be most effective. With a sample size of 79 donors, we are underpowered to detect many eQTLs and reQTLs with smaller effect sizes, which may be more relevant to disease progression. These signals would also provide additional signals to test for colocalization with OA GWAS. Lastly, although we observe a colocalization to prioritize a likely effector gene at a GWAS locus, future studies will still be needed to confirm the causality of this gene and prove how the identified gene expression changes influence OA development. These results would be further strengthened by chromatin accessibility QTLs (caQTLs) to identify genetic variant associations with open chromatin regions, which may further resolve OA GWAS signals and FN-f reQTLs by suggesting potential mechanisms for variants at shared signals in influencing gene expression and OA phenotypes by altering chromatin accessibility. Ultimately, these results require experimental validation through studies like gene knockout to prove which gene expression changes are causal in disease progression and Hi-C to study DNA looping between putative causal variants and their target genes to confirm the physical interaction of distal variants and genes.

Here we have used a model of OA using FN-f perturbation to investigate the context-specific effects of genetic variation on gene expression and its influence on disease progression. We have generated the first reQTL map for an osteoarthritis-relevant cell type and have strengthened the evidence of a possible effector gene by colocalizing our results with OA GWAS. As we increase our study size, we will likely be able to identify more reQTL signals with smaller effect sizes and resolve even more genetic association signals. In combination with other studies, our findings contribute to the investigation of OA genetics and will provide evidence for the ultimate translation of genetic findings into novel drug targets for the treatment of OA.

4.4 Materials and Methods

Sample collection and treatment

The samples used in these analyses are the same as those used in Chapter 3. Human talar cartilage was obtained from 79 tissue donors without a history of arthritis through the Gift of Hope Organ and Tissue Donor Network (Elmhurt, IL). Primary articular chondrocytes were isolated by enzymatic

digestion and treated with either 42-kDa endotoxin-free recombinant FN-f (1 μM final concentration in PBS) or with PBS as a control after serum starvation. After 18 hours of either treatment, RNA was isolated using the RNeasy kit from Qiagen and samples were sent for library preparation and sequencing at the New York Genome Center.

RNA-sequencing data processing/quantification of RNA levels

The samples used in these analyses are the same as those used in Chapter 3. RNA-seq libraries were sequenced to an average depth of approximately 80 million reads per sample at the New York Genome Center. FASTQ files from the same library were merged, quality controlled with FastQC v0.11.8 (Andrews, 2010), and trimmed for low quality reads and adapters with TrimGalore! V0.6.2 (Krueger, n.d.). These trimmed reads were quantified with Salmon v1.4.0 (Patro et al., 2017) against the hg38 transcriptome with default settings.

Genotype processing

Genotype processing was carried out as described in Chapter 3. Genotyping was performed using the Illumina Human Infinium Global Diversity Array platform and exported into PLINK format with the GenomeStudio software from Illumiina. PLINK v1.9 (Purcell et al., 2007) was used to perform quality control, filtering out SNPs with missing genotype rate > 10% (--geno 0.1), deviations from Hardy-Weinberg equilibrium at a p-value < 1 x 10e-6 (--hwe 10^-6), and minor allele frequency < 1% (--maf 0.01). Reported sample sexes were confirmed based on heterozygosity on the X chromosome. After combining our data with data from the 1000 Genomes Project, we used EIGENSTRAT v7.2.1 (Patterson et al., 2006; Price et al., 2006) to estimate the population structure of our samples (**Figure 4S1.C**). The TOPMed Imputation Server was used for Eagle2 (v2.4) phasing (Loh et al., 2016) and imputation against the TOPMed reference panel (version R2 on GRC38) (Das et al., 2016). Additional quality control was performed with PLINK following imputation to retain SNPs with missing genotype rate < 10% (--geno 0.1), deviations from Hardy-Weinberg equilibrium at a p-value > 1 x 10e-6 (--hwe 10^-6), minor allele frequency > 1% (--maf 0.01), and sufficient imputation quality (R2 > 0.3). The final dataset contained 10419216 autosomal variants for 79 donor samples.

Sample quality control

VerifyBamID v1.1.3 (Jun et al., 2012) was used to detect sample swaps or mixing between samples. 2 genotyping sample swaps were detected and corrected. Samples were kept when satisfying [FREEMIX] > 0.04 and [CHIPMIX] < 0.04. 79 donors with genotyping and control and FN-f-treated RNA-seq data were retained.

Replicate correlation

To quantify cell culture-induced noise in our samples, we cultured additional control samples from 2 donors and additional FN-f samples from 3 donors. Pearson's correlations of gene expression were calculated between libraries from the same donors as well as between libraries across different donors. Pearson's correlations were transformed with Fisher's z and tested for significant difference between donor-self libraries and donor-other libraries with an unpaired, two-sided t test.

Condition-specific *cis* eQTL mapping

We performed local eQTL mapping separately for control and FN-f treated samples. We included genes with at least 10 reads in at least 5% of samples and normalized between samples using weighted trimmed mean of M values (TMM) (Robinson & Oshlack, 2010) with edgeR (Robinson et al., 2010). We then normalized gene expression data separated by condition with an inverse normal transformation across each gene.

We selected genetic variants for testing with at least 10 counts of the minor allele and at least 5 heterozygotes using GATK VariantFiltration (McKenna et al., 2010). For each gene, we considered variants within a 1 Mb window in either direction of autosomal gene transcription start sites (TSS). The TSS was defined as the transcription start site of the gene isoform with the most upstream exon, defined from the GRCh38 Ensembl genome assembly.

We performed QTL mapping with QTLtools (Delaneau et al., 2017) using a final model including 10 PEER factors, donor sex, RNA extraction kit batch, DNA extraction kit batch, and 2 genotyping principal components as covariates (See section 4.2.2). To perform a permutation-based analysis for adjusting associations, we employed the QTLtools cis permutation pass with 1000 permutations. Globally adjusted p-values were obtained both with the Storey q-value (Storey, 2003) and Benjamini-Hochberg false discovery rate (FDR) (Benjamini & Hochberg, 1995). Genes with significant eQTLs (eGenes) were

defined as eGenes with FDR less than 0.05. For each eGene, we defined a nominal threshold for significant eQTLs by calculating a p-value as the mean of the smallest p-value above the FDR threshold and the highest p-value above the FDR threshold and using the beta distribution (qbeta) with shape1 and shape2 parameters defined from QTLtools. Allelic fold change (aFC) of eQTLs was calculated using the aFC tool (Mohammadi et al., 2017).

QTL sharing

We quantified QTL sharing between our control and FN-f local eQTL datasets and compared our datasets against publicly available GTEx tissue eQTLs (GTEx Consortium, 2013) and recently published cartilage eQTLs (Steinberg et al., 2021) using the π_1 statistic (Storey & Tibshirani, 2003) from the R qvalue package (Storey et al., 2022). For each comparison, we selected primary eSNP-eGene pairs from our control and FN-f datasets and extracted nominal p values from other datasets for those corresponding eSNP-eGene pairs. For eSNP-eGene pairs that were not found in other datasets, we assigned a random p-value sampled from the uniform distribution (runif). To find the fraction of true associations in other datasets, we computed π_0 and defined π_1 as 1- π_0 for each set of p-values. These values were calculated 100 times with random p-value samplings.

Identification of reQTLs

Separately within the control eQTL and FN-f eQTL datasets, we first identified the significant (FDR < 0.05) lead eGene-eSNP pairs that were only found in either condition. We then assembled all pairs of eGenes and corresponding significant SNPs and tested whether the eQTL effect size was significantly different among conditions by comparing the following two linear models with lme4 (Bates et al., 2015):

$$H0: expression \sim genotype + covariates + condition + (1|donor)$$

$$H1: expression \sim genotype + covariates + condition + genotype:condition + (1|donor)$$

where (1|donor) accounts for any donor-specific random effects. Normalized expression data and the same covariates used in standard eQTL mapping were input to these models. For each eGene-eSNP pair, we tested the significance of the genotype:condition term using ANOVA. We used Benjamini-Hochberg FDR correction (Benjamini & Hochberg, 1995) and defined significant response eQTLs (reQTLs) with FDR values below 0.05.

Colocalization between reQTLs and osteoarthritis GWAS associations

To test for colocalization between reQTLs and GWAS hits, we used summary statistics for 11 OA phenotypes from the largest OA GWAS meta-analysis to date (Boer et al., 2021). We considered colocalizations between the lead variants from Boer et al. and their LD proxies. LD proxies were identified using the 1000 Genomes European reference panel (11 of 13 cohorts are of European descent). r^2 values were calculated with the –ld function in PLINK v1.9 (Purcell et al., 2007) using a window of 1 Mb for calculation, and lead variant GWAS proxies were those defined as those in high LD ($r^2 > 0.8$) with a lead variant. For compatibility with GRCh38-based eQTL data, GWAS summary statistics and LD proxies were lifted over to GRCh38 coordinates with UCSC liftOver (Hinrichs et al., 2006). Variant rsIDs were assigned with dbSNP155 based on variant positions. We analyzed all phenotypes separately with significant reQTLs identified after FN-f treatment. For each analysis, we ran coloc (Giambartolomei et al., 2014) using default priors considering a region of 250 Kb centered on the lead reQTL variant if the lead reQTL variant was in moderate LD ($r^2 > 0.5$) with the lead GWAS variant of that region. LD r^2 values of lead reQTL variants with other variants were computed with our in-study reference with the PLINK v1.9 (Purcell et al., 2007) –ld function using a window of 1 Mb. We considered a posterior probability (PP4) > 0.7 to be sufficient evidence of colocalized signals.

4.5 Supplemental Figures

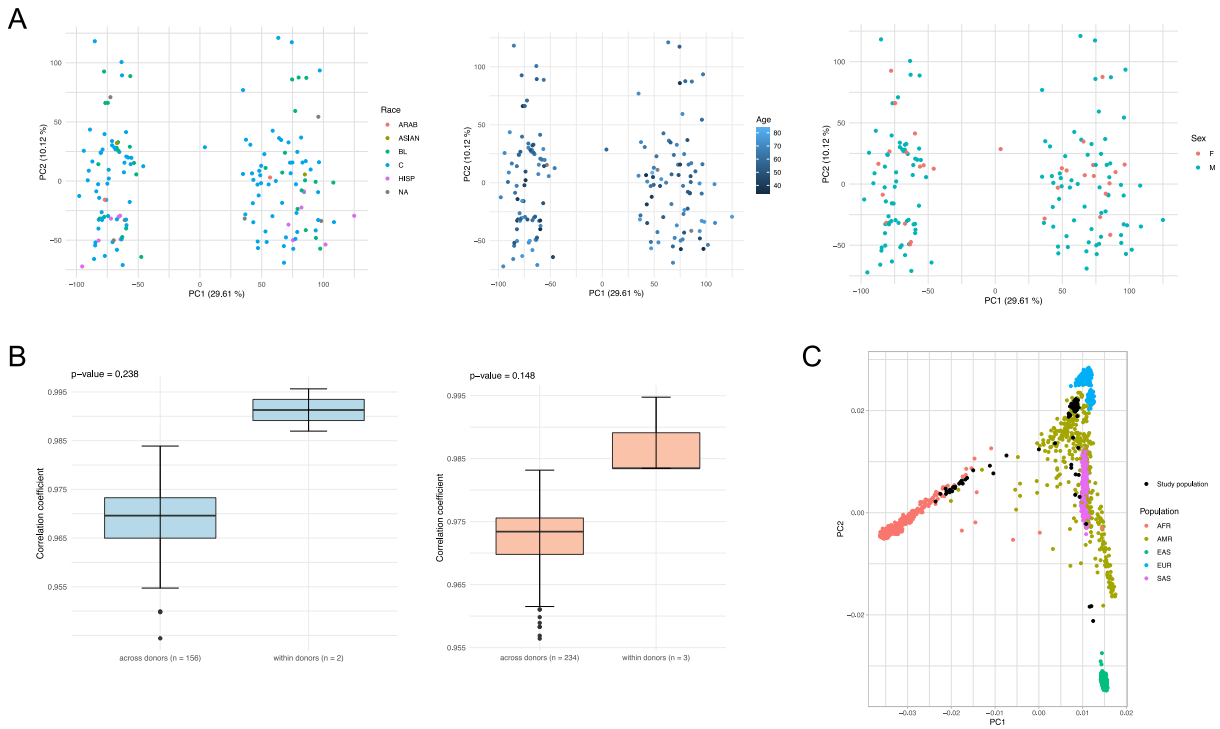

Figure 4.S1. Sample gene expression profiling and donor ancestry.(**A)** Principal component analysis (PCA) of control and FN-f gene expression colored by donor self-reported race (left), donor age (middle), or donor sex (right). These factors do not drive the variation seen in PC1. **(B)** Replicate correlation of RNA-seq libraries across donors and within donors for control (left) and FN-f (right) libraries. P-values between across donor and within donor groups were calculated using an unpaired, two-sided t-test on Fisher's z transformed Pearson's correlation coefficients. **(C)** Principal component analysis of sample genotypes overlaid with 1000 Genomes data. Data from 1000 Genomes are colored by superpopulation and samples from this study are colored in black. AFR: African, AMR: American, EAS: East Asian, EUR: European, SAS: South Asian.

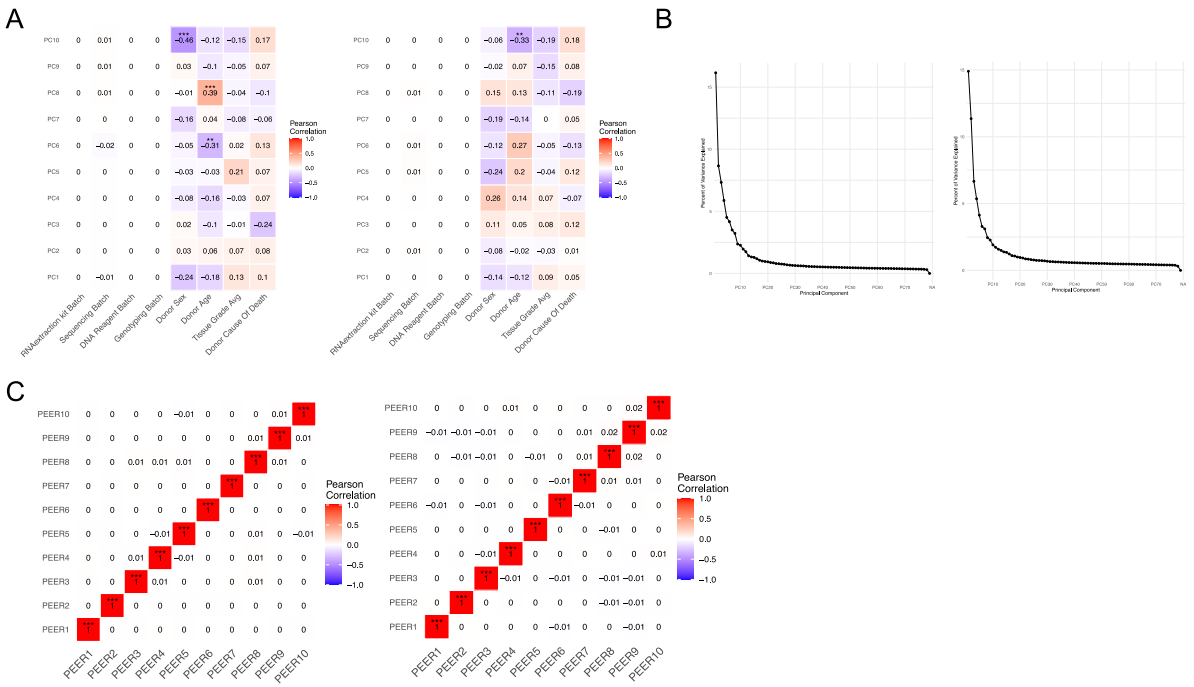

Figure 4.S2. Additional covariate selection analysis. (A) Pearson's correlation analysis of control (left) and FN-f (right) gene expression principal components after RNA extraction kit and DNA reagent batch corrections. **(B)** Percent of variance explained by principal components calculated for control (left) and FN-f (right) expression. **(C)** Pearson's correlation analysis between 10 PEER factors calculated for control (left) and FN-f (right) local eQTL mapping. PEER factors are not correlated with each other. Asterisks in correlation analyses indicate significant correlation (** p-value <= 0.01; *** p-value < 0.001).

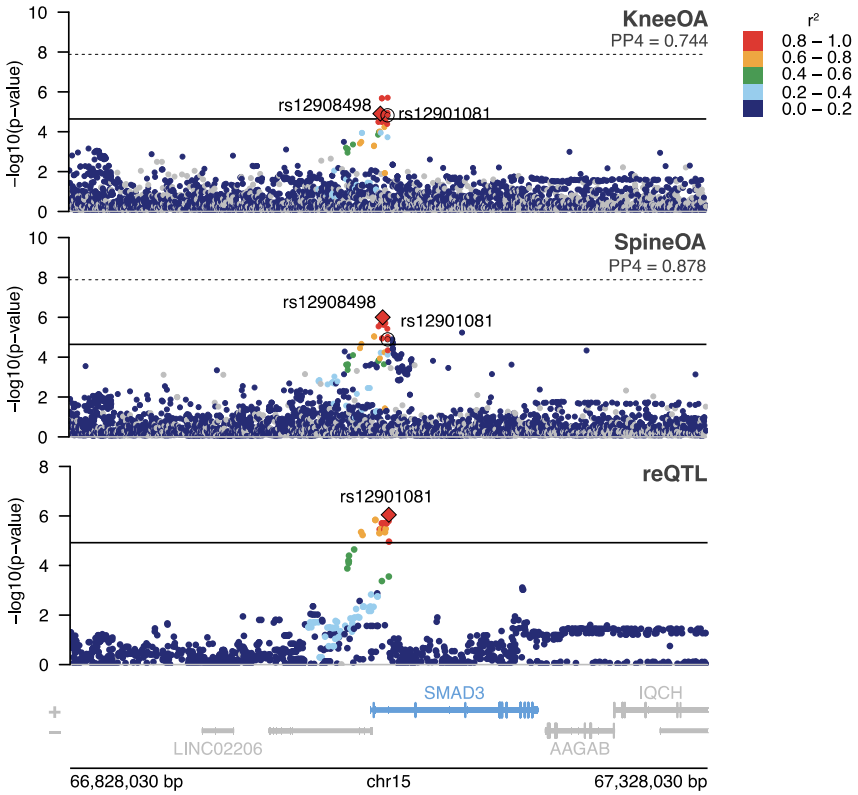

Figure 4.S3. Additional OA GWAS phenotype colocalizations where the lead GWAS variant did not reach genome-wide significance. Manhattan plots of Knee and Spine OA and the FN-f reQTL signal for *SMAD3* gene expression. Associations are depicted with -log10 p-values on the y-axis and variant genomic locations on the x-axis. The index variant of Knee OA and Spine OA GWAS is rs12908498 while the index variant of the reQTL signal is rs12901081, which are represented by red diamonds. reQTL index SNP rs12901081 is circled in GWAS Manhattan plots. In GWAS tracks, dashed lines represent the genome-wide significance threshold at 1.3e-8 and solid lines represent the nominal significance threshold at 2.27e-5. Solid line in reQTL track represents the nominal significance threshold for the *SMAD3* eGene at1.201e-5. PP4 = posterior probability of colocalization.

REFERENCES

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., HIPSCI Consortium, Hale, C., Dougan, G., & Gaffney, D. J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature Genetics*, *50*(3), 424–431.

Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M., & Gerstein, M. B. (2010). Annotating non-coding regions of the genome. *Nature Reviews. Genetics*, *11*(8), 559–571.

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Aubourg, G., Rice, S. J., Bruce-Wootton, P., & Loughlin, J. (2022). Genetics of osteoarthritis. *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society*, *30*(5), 636–649.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Boer, C. G., Hatzikotoulas, K., Southam, L., Stefánsdóttir, L., Zhang, Y., de Almeida, R. C., Wu, T. T., Zheng, J., Hartley, A., Teder-Laving, M., Skogholt, A. H., Terao, C., Zengini, E., Alexiadis, G., Barysenka, A., Bjornsdottir, G., Gabrielsen, M. E., Gilly, A., Ingvarsson, T., … Zeggini, E. (2021). Deciphering osteoarthritis genetics across 826,690 individuals from 9 populations. *Cell*, *0*(0). https://doi.org/10.1016/j.cell.2021.07.038

Bomer, N., den Hollander, W., Ramos, Y. F. M., Bos, S. D., van der Breggen, R., Lakenberg, N., Pepers, B. A., van Eeden, A. E., Darvishan, A., Tobi, E. W., Duijnisveld, B. J., van den Akker, E. B., Heijmans, B. T., van Roon-Mom, W. M., Verbeek, F. J., van Osch, G. J. V. M., Nelissen, R. G. H. H., Slagboom, P. E., & Meulenbelt, I. (2015). Underlying molecular mechanisms of DIO2 susceptibility in symptomatic osteoarthritis. *Annals of the Rheumatic Diseases*, *74*(8), 1571–1579.

Bos, S. D., Bovée, J. V. M. G., Duijnisveld, B. J., Raine, E. V. A., van Dalen, W. J., Ramos, Y. F. M., van der Breggen, R., Nelissen, R. G. H. H., Slagboom, P. E., Loughlin, J., & Meulenbelt, I. (2012). Increased type II deiodinase protein in OA-affected cartilage and allelic imbalance of OA risk polymorphism rs225014 at DIO2 in human OA joint tissues. *Annals of the Rheumatic Diseases*, *71*(7), 1254–1258.

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, *11*, 424.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287.

Delaneau, O., Ongen, H., Brown, A. A., Fort, A., Panousis, N. I., & Dermitzakis, E. T. (2017). A complete tool set for molecular QTL discovery and analysis. *Nature Communications*, *8*, 15452.

Finnson, K. W., Chi, Y., Bou-Gharios, G., Leask, A., & Philip, A. (2012). TGF-b signaling in cartilage homeostasis and osteoarthritis. *Frontiers in Bioscience* , *4*, 251–268.

Forsyth, C. B., Pulai, J., & Loeser, R. F. (2002). Fibronectin fragments and blocking antibodies to alpha2beta1 and alpha5beta1 integrins stimulate mitogen-activated protein kinase signaling and increase collagenase 3 (matrix metalloproteinase 13) production by human articular chondrocytes. *Arthritis and Rheumatism*, *46*(9), 2368–2376.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, *10*(5), e1004383.

Goldring, M. B. (2013). Insight into the function of DIO2, a susceptibility gene in human osteoarthritis, as an inducer of cartilage damage in a rat model: is there a role for chondrocyte hypertrophy? [Review of *Insight into the function of DIO2, a susceptibility gene in human osteoarthritis, as an inducer of cartilage damage in a rat model: is there a role for chondrocyte hypertrophy?*]. *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society*, *21*(5), 643–645.

GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., … Kent, W. J. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, *34*(Database issue), D590–D598.

Homandberg, G. A. (1999). Potential regulation of cartilage metabolism in osteoarthritis by fibronectin fragments. *Frontiers in Bioscience: A Journal and Virtual Library*, *4*, D713–D730.

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *American Journal of Human Genetics*, *99*(6), 1245–1260.

Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., Boehnke, M., & Kang, H. M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American Journal of Human Genetics*, *91*(5), 839–848.

Krueger, F. (n.d.). *Babraham Bioinformatics - Trim Galore!* Retrieved April 6, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, *48*(11), 1443–1448.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.

Mohammadi, P., Castel, S. E., Brown, A. A., & Lappalainen, T. (2017). Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Research*, *27*(11), 1872–1884.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909.

Pulai, J. I., Chen, H., Im, H.-J., Kumar, S., Hanning, C., Hegde, P. S., & Loeser, R. F. (2005). NF-kappa B mediates the stimulation of cytokine and chemokine expression by human articular chondrocytes in response to fibronectin fragments. *Journal of Immunology* , *174*(9), 5781–5788.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575.

Reed, K. S. M., Ulici, V., Kim, C., Chubinskaya, S., Loeser, R. F., & Phanstiel, D. H. (2021). Transcriptional response of human articular chondrocytes treated with fibronectin fragments: an in vitro model of the osteoarthritis phenotype. *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society*, *29*(2), 235–247.

Rice, S. J., Brumwell, A., Falk, J., Kehayova, Y. S., Casement, J., Parker, E., Hofer, I. M. J., Shepherd, C., & Loughlin, J. (2022). Genetic risk of osteoarthritis operates during human skeletogenesis. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddac251

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , *26*(1), 139–140.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25.

Schwartzentruber, J., Cooper, S., Liu, J. Z., Barrio-Hernandez, I., Bello, E., Kumasaka, N., Young, A. M. H., Franklin, R. J. M., Johnson, T., Estrada, K., Gaffney, D. J., Beltrao, P., & Bassett, A. (2021). Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nature Genetics*, *53*(3), 392–402.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311.

Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F., Lee, J. H., Puviindran, V., Tam, D., Shen, M., Son, J. E., Vakili, N. A., Sung, H.-K., Naranjo, S., Acemel, R. D., … Nóbrega, M. A. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, *507*(7492), 371–375.

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, *7*(3), 500–507.

Steinberg, J., Southam, L., Roumeliotis, T. I., Clark, M. J., Jayasuriya, R. L., Swift, D., Shah, K. M., Butterfield, N. C., Brooks, R. A., McCaskie, A. W., Bassett, J. H. D., Williams, G. R., Choudhary, J. S., Wilkinson, J. M., & Zeggini, E. (2021). A molecular quantitative trait locus map for osteoarthritis. *Nature Communications*, *12*(1), 1309.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, *31*(6), 2013–2035.

Storey, J. D., Bass, A. J., Dabney, A., & Robinson, D. (2022). *qvalue: Q-value estimation for false discovery rate control*. http://github.com/jdstorey/qvalue

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(16), 9440–9445.

Tachmazidou, I., Hatzikotoulas, K., Southam, L., Esparza-Gordillo, J., Haberland, V., Zheng, J., Johnson, T., Koprulu, M., Zengini, E., Steinberg, J., Wilkinson, J. M., Bhatnagar, S., Hoffman, J. D., Buchan, N., Süveges, D., arcOGEN Consortium, Yerges-Armstrong, L., Smith, G. D., Gaunt, T. R., … Zeggini, E.

(2019). Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nature Genetics*, *51*(2), 230–236.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., & Lappalainen, T. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 1–21.

Umans, B. D., Battle, A., & Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? *Trends in Genetics: TIG*, *37*(2), 109–124.

van der Kraan, P. M., Blaney Davidson, E. N., Blom, A., & van den Berg, W. B. (2009). TGF-beta signaling in chondrocyte terminal differentiation and osteoarthritis: modulation and integration of signaling pathways through receptor-Smads. *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society*, *17*(12), 1539–1545

CHAPTER 5: DISCUSSION

In this work, I have created the first response AI and eQTL datasets in chondrocytes and developed a novel visualization tool for programmatically generating complex, publication-quality, multi-panel genomic data figures. The findings and workflows/tool presented here do not only contribute to the understanding of the genetic basis of OA, but can also be broadly applied to other genomic cell types and contexts. There is still much to be answered about the regulatory molecular mechanisms that impact traits and diseases, but it is important to continue the rigorous research, insightful discussions, and development of innovative methods and analyses to investigate the mysteries of the non-coding genome.

5.1 The importance of computational methods in studying genomics

Since the sequencing of the human genome, there has been an exponential increase in the size and quantities of genomic data alongside a decrease in sequencing costs (Lander et al., 2001; Muir et al., 2016). Whereas studies once primarily focused on single genes or used observational laboratory methods to test hypotheses, we are now flooded with terabytes of high-throughput data that must be appropriately parsed and integrated in order to make any meaningful conclusions. The sequence information of FASTQ files must be translated into readouts of gene expression, chromatin accessible regions, transcription factor motifs, binding sites of DNA-associated proteins, regions of interacting chromatin, and potentially more data types that can inform DNA structure, regulation, and function (Cock et al., 2010). Furthermore, these data must be interpreted properly through integration, statistics, and visualization. Taken together, these factors highlight the crucial role that bioinformaticians and computational biologists play in studying the complex landscape of genomics.

Open-source tools like plotgardener, detailed in Chapter 2, are necessary for allowing scientists worldwide to conduct analyses with their own datasets. In particular, visualization of such data is often overlooked as a key step in scientific research, despite it usually being the final form of understanding and communicating results through presentations and publications. To many, not only is seeing believing, but it also conceptualizes complicated information in a more accessible format. Making the process of

scientific visualization programmatic has many advantages: it fluidly combines with upstream data analysis, it allows figures to easily be updated with slight data changes, and it makes figures reproducible and the process by which data was translated into the figure transparent. With plotgardener, plots do not need to be finished with graphic design software, so there is no risk of misleading manipulation of scientific findings. It is my hope that setting such a standard will extend the rigor and reproducibility of science all the way to end-stage visualization.

Creating visualizations from their primitive elements, while challenging, has provided a richer understanding of the way data types translate into plots. Plotgardener focuses on plotting and annotation functions for genomic data, but it would be further strengthened by incorporating functions for non-genomic plotting. Another useful future direction for plotgardener is the development of an interactive version of the package, making it more accessible for less experienced programmers while keeping the programmatic aspect of the tool intact. As genomic datasets continue to evolve in complexity, we will continue to develop and improve plotgardener for numerous use cases.

Beyond tools, publicly available and distributed code for bioinformatic workflows and analyses is also essential for scientific discovery, particularly in the study of genomics. The pipelines I developed for work in Chapters 3 and 4 have been assembled with snakemake (Mölder et al., 2021), which combines different command line tools and parsing scripts from different programming languages into one continuous workflow. I am the first to establish the response AI and eQTL workflows within my lab, which will advance studies building upon mine or related studies that can use adapted versions of my code. In addition, every analysis is available on GitHub, so that other researchers can learn from and/or contribute to my computational methods. Just as experimental assays are critical in science, computational methods are a necessary and important part of research.

5.2 Response eQTLs and AI for investigating non-coding GWAS loci

The datasets created and explored in Chapters 3 and 4 link genetic variation to gene expression, with each approach having their own benefits and drawbacks. Assessing allele-specific expression of genes allows for higher powered studies with smaller donor sample sizes, but is limited by the number of heterozygous donors for a given genetic variant. Mechanisms of ASE are driven by variants located in or near the gene, restricting us to *cis* genetic regulatory measures. eQTLs link regulatory variants to specific

genes, but must be studied in the correct cell types and disease-relevant contexts. Both of these techniques, though informative in understanding gene expression changes and their relationship with genetic variants, still leave many open questions in understanding disease-causal variants and their molecular mechanisms. Although in getting information about the local coordination of gene regulation we may capture the *cis* effects of many *trans* mechanisms, it is still difficult to understand the full extent of their mechanisms when using these data alone. It is unclear what downstream effects these variants may have on other aspects of the genome, including expression of non-coding RNAs and altering 3D genomic structures. With the possibility of one genetic variant influencing the expression of multiple genes in a network of fine-tuned regulation, we may also be losing information about relative effects of gene networks. Overall, the field needs these studies as well as orthogonal datasets like chromatin accessibility QTLs (QTLs) and Hi-C in numerous cell types and states to get a complete picture of genetic regulation.

With the dynamic nature of genetic regulation, it makes sense that disease-relevant regulation would occur over a course of time in response to a perturbation relevant to the disease context. In particular, studying such systems would reveal variants that display gene-environment interactions, which is a relevant feature to understanding the genetics of complex diseases. Response studies like those conducted in Chapters 3 and 4 seek to isolate these effects, but there are still many open avenues for investigating these aspects of gene regulation. Although the number of these studies are rapidly increasing, they are still in early stages in using controlled disease systems and specific cell types. Studies of the immune response and blood gene regulation have already gained some crucial insights into gene-environment interactions (Findley et al., 2019; Mangravite et al., 2013; Maranville et al., 2011; Moyerbrailean et al., 2016), but these systems are only the beginning of what could be done for disease genetics. There are still numerous questions associated with probing disease contexts – which cell types are most relevant? What is the appropriate stimulus? At what point should the response be assessed? Capturing the specific effects of genetic variation on disease progression would also require more developed systems to understand the fine-grained temporal dynamics of regulation that cannot be observed at singular time points.

Until now, the only “response” OA QTL and AI studies have centered on different states of diseased tissue. In these study designs, low-grade, preserved cartilage acts as a control while high-

grade, lesioned cartilage, usually from the same donor, acts as the “response” sample. This framework most likely loses genetic information specific to OA progression and cannot recapitulate the dynamics of OA gene regulation and could potentially be confounded by uncontrolled variables. This makes comparison with our results challenging, as limited overlap between identified genetic variants and gene expression changes may simply correspond to different aspects of relevance. The continued efforts of both kinds of studies may be required to generate a complete map of OA genetics at various stages, including predisposition to disease, dynamic changes as disease worsens, and end-stage disease. Comparison with these studies also brings into question which analytical methods are best suited to distinguish context-specific AI events and QTLs and how conservative these estimates should be. Chapter 3 explored two different methods for testing for AI, either at the variant level with DESeq2 (Love et al., 2014) or at the gene level using population information with ASEP (Fan et al., 2020). For reQTLs, Chapter 4 focused on testing the significance of the interaction term of genotype with condition within the linear model, but it may be worthwhile to explore other tools like Metasoft (Sul et al., 2013) or other measures of significantly different eQTLs like β comparisons (Kim-Hellmuth et al., 2017). Like standard QTL mapping, exploring these methods will require a balance of maximizing discovery and controlling false positives.

5.3 Insights gained from using fibronectin fragments to study OA regulatory genomics

Chapters 3 and 4 leverage an ex vivo model of OA whereby human chondrocytes are stimulated with fibronectin fragments (FN-f) to conduct response AI and eQTL studies and characterize aspects of OA gene regulation. In line with previous findings (Reed et al., 2021), we found evidence of the differential expression of many OA-relevant genes in response to FN-f treatment. We conclude that our use of the system is comparable to this study, but our increased donor number may benefit from a more direct comparison of gene expression to the RAAK (Research Arthritis and Articular Cartilage) study or other diseased cartilage samples (Ramos et al., 2014).

The use of an ex vivo model provided the platform to study response-specific genetic regulation against various human genetic backgrounds. The treatment of FN-f induces variable transcriptional changes that are consistent with what is seen in patient samples while also controlling for any potentially confounding variables and isolating specific gene-environment interactions. Furthermore, the candidate

genetic variants and genes identified from our reQTL and AI analyses can be functionally validated within this system to directly test the relevance of regulatory variants on the OA phenotype. Two works have already successfully used the FN-f OA model in conjunction with CRISPR-based editing (D'Costa et al., 2020; Thulson et al., 2022), so the variants and genes identified here would be promising targets to probe with this kind of validation.

Despite its numerous advantages in studying gene regulation in the context of OA, the FN-f chondrocyte model and the findings from the studies conducted with it in Chapters 3 and 4 leave some outstanding questions. The limited overlap of our identified putative SNPs and affected genes with previous OA eQTL and AI studies suggests that we might not capture the full range of OA phenotypes. Our results suggest that our study only focuses on one specific aspect of OA progression, particularly related to the immune response and inflammation in the cartilage. Future work could focus on the other key cell types within the entire joint, i.e. synovial macrophages and synovial fibroblasts, with similar response systems to characterize the distinct roles of different cell types of the joint in mediating OA-specific genetic regulation. However, such models have not been developed and validated to the extent of the FN-f chondrocyte model, and integrating information from all models may prove challenging as regulation likely involves the additive contributions of multiple gene regulatory networks from combinations of cell types. Another promising avenue to explore would be the identification of single cell QTLs within joint tissue to capture QTLs for all joint cell types, although it may be difficult to probe response effects at the single-cell level against variable genetic and environmental backgrounds. Thus, model systems provide the platform for well-designed, robust studies but will always face questions of disease relevance and should be explored in conjugation with other disease models and systems.

5.4 Colocalization: challenges and future directions

Colocalization of QTL signals with GWAS genetic loci is a powerful technique to connect putative causal variants associated with molecular traits with a complex phenotype, but it involves many challenges. In Chapter 4, we identified one reQTL colocalization at rs12901081 with GWAS signals in All OA, KneeHip OA, Hip OA, TJR, and THR. This SNP is in high LD ($r^2 = 0.9$) with the lead GWAS SNP rs12908498 and is associated with the expression of *SMAD3*, which has been previously cited as a likely effector gene from GWAS (Boer et al., 2021) and identified in a colocalization analysis done in a previous

cartilage eQTL study (Steinberg et al., 2021). Although we can make hypotheses about the gene and variants driving this signal from such a finding, we are still limited in our understanding of the mechanisms at this locus and would benefit from finding standard and response caQTLs that colocalize with this locus as well.

Numerous statistical methods have been developed to assess the colocalization of signals. In our colocalization analyses, we employed coloc (Giambartolomei et al., 2014) with an assumption of a single causal variant shared between reQTL data and GWAS summary statistics. However, this method does not consider differing LD structures between datasets. OA GWAS resulted from a meta-analysis of 9 populations (Boer et al., 2021) whereas donors in our reQTL study were primarily of European descent, making it more complicated to isolate a single causal variant amongst varying LD structures. Furthermore, we did not perform stepwise conditional analysis with our data to distinguish independent eQTLs, which may provide more accurate results to perform colocalization. These considerations are promising avenues to explore as we continue to increase our reQTL donor sample sizes, diversify donor ancestries, introduce caQTLs, and optimize available colocalization methods for our datasets. Nonetheless, colocalization will continue to be challenging as methods become more advanced in handling LD structure (Hormozdiari et al., 2016), analyzing multiple datasets simultaneously (Foley et al., 2021), and relaxing the single causal variant assumption (Wallace, 2021) as we battle complications with maintaining computational efficiency.

Beyond prioritizing a causal variant through colocalization which implies that the same variant affects a GWAS trait through the modulation of gene expression, mediation analysis would elucidate the potential mediating role of *cis* genetic variants on distal gene expression. This method could potentially disentangle more *trans* gene regulation mechanisms which may be influenced by one or many genes local to a SNP (Shan et al., 2019). The FN-f chondrocyte model would provide an interesting platform to identify the mediating factors involved in gene regulation in response to an OA-relevant stimulus. By disentangling these effects, we can hopefully begin translating knowledge of disease-related genetic variation into novel therapies and treatments.

REFERENCES

Boer, C. G., Hatzikotoulas, K., Southam, L., Stefánsdóttir, L., Zhang, Y., de Almeida, R. C., Wu, T. T., Zheng, J., Hartley, A., Teder-Laving, M., Skogholt, A. H., Terao, C., Zengini, E., Alexiadis, G., Barysenka, A., Bjornsdottir, G., Gabrielsen, M. E., Gilly, A., Ingvarsson, T., … Zeggini, E. (2021). Deciphering osteoarthritis genetics across 826,690 individuals from 9 populations. *Cell*, *0*(0). https://doi.org/10.1016/j.cell.2021.07.038

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771.

D'Costa, S., Rich, M. J., & Diekman, B. O. (2020). Engineered Cartilage from Human Chondrocytes with Homozygous Knockout of Cell Cycle Inhibitor p21. *Tissue Engineering. Part A*, *26*(7-8), 441–449.

Fan, J., Hu, J., Xue, C., Zhang, H., Susztak, K., Reilly, M. P., Xiao, R., & Li, M. (2020). ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genetics*, *16*(5), e1008786.

Findley, A. S., Richards, A. L., Petrini, C., Alazizi, A., Doman, E., Shanku, A. G., Davis, G. O., Hauff, N., Sorokin, Y., Wen, X., Pique-Regi, R., & Luca, F. (2019). Interpreting Coronary Artery Disease Risk Through Gene-Environment Interactions in Gene Regulation. *Genetics*, *213*(2), 651–663.

Foley, C. N., Staley, J. R., Breen, P. G., Sun, B. B., Kirk, P. D. W., Burgess, S., & Howson, J. M. M. (2021). A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature Communications*, *12*(1), 764.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, *10*(5), e1004383.

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *American Journal of Human Genetics*, *99*(6), 1245–1260.

Kim-Hellmuth, S., Bechheim, M., Pütz, B., Mohammadi, P., Nédélec, Y., Giangreco, N., Becker, J., Kaiser, V., Fricker, N., Beier, E., Boor, P., Castel, S. E., Nöthen, M. M., Barreiro, L. B., Pickrell, J. K., Müller-Myhsok, B., Lappalainen, T., Schumacher, J., & Hornung, V. (2017). Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nature Communications*, *8*(1), 266.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Mangravite, L. M., Engelhardt, B. E., Medina, M. W., Smith, J. D., Brown, C. D., Chasman, D. I., Mecham, B. H., Howie, B., Shim, H., Naidoo, D., Feng, Q., Rieder, M. J., Chen, Y.-D. I., Rotter, J. I., Ridker, P. M., Hopewell, J. C., Parish, S., Armitage, J., Collins, R., … Krauss, R. M. (2013). A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. *Nature*, *502*(7471), 377–380.

Maranville, J. C., Luca, F., Richards, A. L., Wen, X., Witonsky, D. B., Baxter, S., Stephens, M., & Di Rienzo, A. (2011). Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. *PLoS Genetics*, *7*(7), e1002162.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, *10*(33), 33.

Moyerbrailean, G. A., Richards, A. L., Kurtz, D., Kalita, C. A., Davis, G. O., Harvey, C. T., Alazizi, A., Watza, D., Sorokin, Y., Hauff, N., Zhou, X., Wen, X., Pique-Regi, R., & Luca, F. (2016). High-throughput allele-specific expression across 250 environmental conditions. *Genome Research*, *26*(12), 1627–1638.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F., Rozowsky, J., & Gerstein, M. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology*, *17*, 53.

Ramos, Y. F. M., den Hollander, W., Bovée, J. V. M. G., Bomer, N., van der Breggen, R., Lakenberg, N., Keurentjes, J. C., Goeman, J. J., Slagboom, P. E., Nelissen, R. G. H. H., Bos, S. D., & Meulenbelt, I. (2014). Genes involved in the osteoarthritis process identified through genome wide expression analysis in articular cartilage; the RAAK study. *PloS One*, *9*(7), e103056.

Reed, K. S. M., Ulici, V., Kim, C., Chubinskaya, S., Loeser, R. F., & Phanstiel, D. H. (2021). Transcriptional response of human articular chondrocytes treated with fibronectin fragments: an in vitro model of the osteoarthritis phenotype. *Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society*, *29*(2), 235–247.

Shan, N., Wang, Z., & Hou, L. (2019). Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics*, *20*(Suppl 3), 126.

Steinberg, J., Southam, L., Roumeliotis, T. I., Clark, M. J., Jayasuriya, R. L., Swift, D., Shah, K. M., Butterfield, N. C., Brooks, R. A., McCaskie, A. W., Bassett, J. H. D., Williams, G. R., Choudhary, J. S., Wilkinson, J. M., & Zeggini, E. (2021). A molecular quantitative trait locus map for osteoarthritis. *Nature Communications*, *12*(1), 1309.

Sul, J. H., Han, B., Ye, C., Choi, T., & Eskin, E. (2013). Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genetics*, *9*(6), e1003491.

Thulson, E., Davis, E. S., D'Costa, S., Coryell, P. R., Kramer, N. E., Mohlke, K. L., Loeser, R. F., Diekman, B. O., & Phanstiel, D. H. (2022). 3D chromatin structure in chondrocytes identifies putative osteoarthritis risk genes. *Genetics*, *222*(4). https://doi.org/10.1093/genetics/iyac141

Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genetics*, *17*(9), e1009440.