UNIVERSITY OF CALIFORNIA

Los Angeles

Exploring the roles of genetic regulation in human phenotypes

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Human Genetics

by

Malika Kumar Freund

2020

ABSTRACT OF THE DISSERTATION

Exploring the roles of genetic regulation

in human phenotypes

by

Malika Kumar Freund

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2020

Professor Bogdan Pasaniuc, Chair

Human phenotypes are influenced to varying extents by inherited genetic variation, although specific mechanisms through which this variation affects the phenotypes are not completely understood. In this dissertation I explore different modes of genetic regulation in the context of human complex traits and rare disorders. First, I examine the degree of shared genetic basis between complex traits and rare monogenic disorders across a wide range of phenotypes. Second, I explore the regulatory landscape of ovarian surface epithelial cells to identify putative pathways involved in the development of epithelial ovarian cancer. This work provides a foray into understanding the different ways that genetic variation can drive downstream phenotypes through direct and epigenetic regulation of target genes.

The dissertation of Malika Kumar Freund is approved.

Esteban C Dell'Angelica

Stanley F Nelson

Janet S Sinsheimer

Bogdan Pasaniuc, Committee Chair

University of California, Los Angeles

2020

DEDICATION

This dissertation is dedicated to my PhD sisters: Jazlyn Mooney and Chantle Edillor.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I'd like to first acknowledge my advisor Bogdan Pasaniuc, whose patient teaching and faith in me has pushed me forward every year. His vision for these projects and commitment to thorough, robust science is the standard of excellence to which I aspire.

The other members of my committee Esteban Dell'Angelica, Stan Nelson, and Janet Sinsheimer have been generous with their time, insightful feedback, and words of support. I have enjoyed working with them and look forward to continuing our relationship.

I next wish to acknowledge my coauthors, collaborators, labmates, and colleagues who have helped tremendously with guidance, materials, approaches, problem-solving, motivation, and camaraderie across the projects covered in these chapters. In particular, from the Bogdan Lab past and present: Valerie Arboleda, Kathryn S. Burch, Nicholas Mancuso, Tommer Schwarz, Ruth Johnson, Yi Ding, Igor Mandric, Huwenbo Shi, Claudia Giambartolomei, Megan Roytman, Megan Major, Kangcheng Hou, Arunabha Majumdar, Robert Brown, Gleb Kichaev, and Pagé Goddard; from UCLA: Kristina Garske, David Pan, Paivi Pajukanta, and Jennifer Zhou; from Cedars-Sinai: Kate Lawrenson, Simon Gayther, Jasmine Plummer, Forough Abbasi, Brian Davis, Ivetth Corona, and Marcos Abraao; and beyond: Sasha Gusev, Paul Pharoah, and Annique Claringbould.

Finally, I want to thank my personal support team: my husband Brendan; my PhD sisters Jazlyn Mooney and Chantle Edillor; my dear friends Cody Aros, Pooja Pradhan, and Andrew Stiles; my parents Veena and Kumar; and my sister Nandita. This work would not have been possible, or meaningful, if not for you.

Last, but not least, I thank Adriana Arneson for selflessly donating her thesis formatting so I didn't have to figure out how to implement UCLA's thesis guidelines myself.

VITA

EDUCATION

2022 (expected)	MS, Genetic Counseling Stanford University | Stanford, CA
2020	PhD Candidate, Human Genetics University of California, Los Angeles | Los Angeles, CA
2014	BA, Human Biology Stanford University | Stanford, CA

PEER-REVIEWED PUBLICATIONS

Freund MK, Burch K, Shi H, Mancuso N, Kichaev G, Garske KM, Pan DZ, Miao Z, Mohlke KL, Laakso M, Pajukanta P, Pasaniuc B*, Arboleda VA*. Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. Am J Hum Genet. 2018 Oct 4;103(4):535-552.

Major M, **Freund MK**, Burch KS, Mancuso N, Ng M, Furniss D, Pasaniuc B, Ophoff R. Integrative analysis of Dupuytren's disease identifies novel risk locus and reveals a shared genetic etiology with BMI. Genet Epidemiol. 2019 Sep;43(6):629-645.

Mancuso N, **Freund MK**, Johnson R, Shi H, Kichaev G, Gusev A, Pasaniuc B. Probabilistic fine-mapping of transcriptome-wide association studies. Nat Genet. 2019 Apr;51(4):675-682.

Shi H, Burch KS, Johnson R, **Freund MK,** Kichaev G, Mancuso N, Manuel AM, Dong N, Pasaniuc B. Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. Am J Hum Genet. 2020.

Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, **Freund MK**, Schoech A, Pasaniuc B, Price AL. Leveraging polygenic functional enrichment to improve GWAS power. Am J Hum Genet. 2019 Jan 3;104(1):65-75.

Chapter 1: Introduction

Human traits and diseases, or phenotypes, are broadly considered to result from a combination of genetic factors and non-genetic "environmental" factors, to varying degrees. A long-term goal of human medical genetic research has been to isolate and identify the genetic factors driving phenotypes, both towards the goals of better interpreting individuals' risk of developing a trait or disease and better understanding the biological basis of these varied phenotypes. In many cases, understanding the biological basis of phenotypes can directly inform therapeutic approaches.

The last few decades of medical genetic research have successfully identified myriad genetic factors influencing human phenotypes. Large scale family studies, and more recently exome sequencing studies, have identified single genes harboring rare genetic variation linked to a variety of rare genetic disorders[1]. In parallel, genome-wide association studies (GWAS) have identified common genetic variation associated with thousands of complex traits and diseases across many different ancestries and ethnicities[2]. However, there remains much to be understood about the way these genetic variants affect phenotypes. In particular, how "rare disease" genes contribute to broader phenotypes is not well characterized, and common genetic variation identified by GWAS tends to fall in the non-coding genome where interpretation of variant consequences is significantly harder to determine[3].

One path through which genetic variation can be linked to human complex traits and rare disorders is through epigenetics and regulation of gene expression. Various strategies have been proposed to link common genetic variation to more interpretable target genes, through the hypothesis that common genetic variation is directly regulating expression of target genes, and recent studies have provided evidence that genetic variation associated with complex traits can

act through epigenetic regulation of target genes as well[4]. The goal of this thesis is to explore different modes of genetic regulation in the context of human complex traits and rare disorders, and explore mechanisms of how genetic variation ultimately affects human phenotypes.

In **Chapter 2**, I quantify the shared genetic basis of complex traits and Mendelian disorders. This work establishes a systematic, phenotype-specific investigation of the overlap of Mendelian disease genes with GWAS risk genes for complex traits, specifically, to show (1) that GWAS risk genes show specific, significant overlap with phenotypically matched mendelian disorder genes; (2) that single nucleotide polymorphisms (SNPs) near phenotypically matched mendelian disorder genes show increased effect size on complex traits; and (3) there are examples of candidate causal variants for complex traits interacting with phenotypically matched Mendelian disorder genes. This work allows for comparison of the genetic architectures among complex traits, and also provides a baseline level of enrichment for relevant Mendelian genes at GWAS loci. My co-authors and I increase understanding of how SNPs identified by GWAS may be regulating phenotype-relevant known disease genes to contribute to complex trait phenotypes, and suggest candidate variants for functional follow-up studies. A version of Chapter 2 been published as:

Freund MK, Burch KS, Shi H, Mancuso N, Kichaev G, Garske KM, Pan DZ, Miao Z, Mohlke KL, Laakso M, Pajukanta P, Pasaniuc B*, Arboleda VA*. Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits. Am J Hum Genet. 2018;103(4):535-552.

In **Chapter 3**, I explore the regulatory landscape of ovarian surface epithelial cells (OSEC) to identify putative pathways involved in the development of epithelial ovarian cancer. This work characterizes the regulatory landscape of OSEC and explores the hypothesis that common variants in OSEC influence the development of ovarian cancer through epigenetic and

transcriptomic pathways. In particular, with support from collaborators named below, I (1) perform population-based profiling of active chromatin in ovarian surface epithelial cells, (2) construct global maps of gene regulation in ovarian surface epithelial cells, and (3) identify hQTLs and eQTLs in OSEC cell lines. Ultimately, this work aims to support identification of colocalized genetic drivers of H3K27ac peaks, gene expression, and GWAS risk for ovarian cancer. In addition to providing insight into the regulatory landscape of healthy ovarian surface epithelial cells and identifying putative tissue-specific ovarian cancer pathways, these results can be contrasted to similar QTL analyses in ovarian tumors to pinpoint the major changes in gene regulation and pathway activation that occur during neoplastic transformation. This work is a product of collaboration with Kate Lawrenson, Simon Gayther, Jasmine Plummer, Forough Abbasi, Brian Davis, Sasha Gusev, Paul Pharoah, Nicholas Mancuso, Tommer Schwarz, Claudia Giambartolomei, and Bogdan Pasaniuc.

References:

1. Katsanis, N. (2016). The continuum of causality in human genetic disorders. Genome Biol 17, 233.
2. Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19, 212-219.
3. Pasaniuc, B., and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet 18, 117-127.
4. Gusev A, Lawrenson K, Lin X, et al. A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. Nat Genet. 2019;51(5):815-823.

Chapter 2: Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits

2.1 Introduction

Genetic architectures of human traits have traditionally been classified into two major categories. Typically, complex traits demonstrate polygenic architectures arising from many low-effect common variants, whereas rare traits tend to have high-effect monogenic determinants[1]. The underlying and practical distinction between these classes has historically been based on the presence of highly penetrant, rare, single-gene disruptive mutations causing recognizable clinical monogenic diseases (e.g., cystic fibrosis, [MIM: 219700][2]), and the relative absence of such mutations in complex diseases such as diabetes and schizophrenia[3]. Evidence is accumulating that these two classes of phenotypes may not be as biologically distinct as previously thought[4]. Multiple exceptions to the "common disease, common variant" hypothesis[1] have been identified for complex traits[5-7] and their molecular phenotypes[8-11], and Mendelian disorders have also been found to be affected by multiple or common genetic variants[12-15]. This suggests that there exists a spectrum of genetic architectures rather than a dichotomous classification. Accordingly, the monogenic forms of complex traits (i.e., phenotypically-matched Mendelian disorders) are increasingly used as a starting point to identify genes relevant to complex traits for further study[16-18]. Furthermore, overlap has been identified between genes, common variants, and CNVs linked with Mendelian disorders and genetic determinants of complex traits and diseases such as Parkinson's disease [MIM: 68600][19], obesity[20] , height[21], ototoxicity[22], and others[23]. However, the overlap of each of these complex traits with Mendelian disorders has been examined individually, with different metrics of overlap. In a large study of patient medical records, Blair et al. identified systematic, significant comorbidities between Mendelian disorders and complex diseases, and that association signals from genome-wide association studies (GWAS) for complex diseases were enriched in genomic regions with

known roles in comorbid Mendelian disorders, suggesting a shared genetic basis[24]. However, the study focuses on Mendelian disorders comorbid with complex diseases in the same individual, rather than Mendelian disorders demonstrating similar phenotypes to complex traits. Furthermore, advances in sequencing technology have greatly expanded the phenotypic spectrum in known Mendelian syndromes, allowing for deconstruction of syndromic diseases into component medical phenotypes. As such, it is now possible to identify all the component-phenotype consequences of genes linked to Mendelian disorders, allowing for greater resolution in identifying gene-phenotype relationships. However, to the best of our knowledge, no study has taken advantage of this to identify genes linked to any related component-phenotype regardless of the Mendelian disorder's best-known or primary phenotype. Thus, a thorough quantification of the overlap between genes associated with complex traits and genes linked to Mendelian disorders in a phenotype-specific manner remains elusive.

Given that the majority of genome-wide association studies for complex traits and diseases have identified significant associations in non-coding genomic regions[25], we hypothesize that genes individually involved in Mendelian disease belong to the biological pathway(s) shared by both complex and Mendelian disease. Specifically, we hypothesize that large-effect coding variants disrupt individual genes, resulting in severe phenotypes (i.e., Mendelian disorders), while non-coding variants produce complex traits by collectively dysregulating expression of these same genes, allowing for nuanced or tissue-specific phenotypes. Based on this hypothesis, we expect to identify an enrichment of GWAS signal for a given complex trait near genes linked to Mendelian disorders demonstrating similar phenotypes, but no enrichment near genes linked to Mendelian disorders with phenotypes unrelated to the complex trait of interest. To test this hypothesis, we define "Mendelian disorder genes" as any genes linked to Mendelian disorders in the Online Mendelian Inheritance in Man (OMIM) database, and use the well-curated phenotypic breakdown of Mendelian disorders to

identify subsets of these genes linked to particular phenotypes (e.g., growth defects or immune dysregulation) expressed as part of any Mendelian disorder. We then examined publicly available GWAS across 62 complex traits (listed in **Table 1**, and detailed in **Table S1**) to identify risk genes (here called "GWAS gene sets") for each complex trait, and quantified the overlap of each GWAS gene set with 20 other sets of Mendelian disorder genes for particular phenotypes (detailed in **Table 1** and **Table S1**). We find a consistent, significant, and specific enrichment between GWAS gene sets for complex traits and Mendelian disorder genes for matched and related phenotypes (50/1,240 pairs; e.g., rheumatoid arthritis and immune dysregulation), supporting our hypothesis of a shared genetic basis between complex and Mendelian forms of disease. In addition, we observe instances of enrichments between GWAS gene sets for certain complex traits and Mendelian disorder genes for unrelated phenotypes (27/1,240 pairs; e.g., systemic lupus erythematosus and mature-onset diabetes of the young), suggestive of shared biological mechanisms yet to be examined. Furthermore, we find an increase in average effect size of GWAS variants near Mendelian disorder genes for matched phenotypes, and identify examples of associated SNPs found directly at the transcription start sites (TSSs) of these phenotypically-matched Mendelian disorder genes as candidates for functional follow-up. Finally, we report examples of significant body mass index (BMI)-associated variants directly interacting with phenotypically-related Mendelian disorder genes *CREBBP* [MIM: 180849] and *CYP19A1* [MIM: 139300 and 613546], using human primary white adipocyte-specific Hi-C data[26]. Leveraging the growing body of well-curated phenotypic data from studies of Mendelian disorders, we provide a phenotype-driven approach to identifying genetic pathways shared by Mendelian diseases and complex traits. Last, please note that although there are supplementary tables (referenced here in the text by **Table S***), the formatting guidelines for this thesis did not allow for the tables' inclusion in this document. Please instead refer to the published version of this chapter, Freund et al., *Am J Hum Genet* 2018, to access these tables.

2.2 Material and Methods

Gene coordinates and symbols

We downloaded gene body coordinates (NCBI build 37/hg19, UCSC Genes track) from the UCSC Table Browser[27] (see Web Resources) using the gene symbol from the *knownGene* table, transcription start and end sites for each gene from the *knownCanonical* table, and the longest transcript from the *knownGene* table for genes where no entry or multiple entries were listed in the *knownCanonical* table. We used these coordinates for all analyses in our study. Since many genes have been renamed over time, we standardized gene symbols across all analyses in our study by downloading a table of approved symbols, previous symbols, and locus group for each gene from HUGO Gene Nomenclature Committee at the European Bioinformatics Institute (HGNC) (see Web Resources) and renaming any genes identified by previous symbols with approved gene symbols. We restricted all analyses in our study to genes classified as protein-coding according to the HGNC locus group, from chromosomes 1-22. These processing steps resulted in a final single set of coordinates for 17,695 autosomal protein-coding genes (for data access, see Web Resources).

Mendelian disorder genes and loss-of-function (LOF) intolerant genes

To identify Mendelian disorder genes, we downloaded the Online Mendelian Inheritance in Man (OMIM) catalogue database and identified all genes linked to Mendelian disorders satisfying the following criteria: (1) disorder is Mendelian and fully penetrant, therefore excluding susceptibility phenotypes and (2) molecular basis of the Mendelian disorder is known (i.e., phenotype mapping key = 3). We defined loss-of-function (LOF) intolerant genes as any gene with greater than 90% probability of being loss-of-function intolerant, according to the pLI score (pLI > 0.9) from the Exome Aggregation Consortium (ExAC)[28]; this score is derived from the number of observed versus expected LOF variants in a given gene across approximately 60,000 healthy

exomes. Following the same restriction and gene symbol standardization criteria described above resulted in a final set of 3,446 Mendelian disorder genes and 2,978 LOF-intolerant genes.

Phenotype-specific Mendelian disorder gene sets

To identify subsets of Mendelian disorder genes linked to particular phenotypes, for each complex trait we curated a set of standardized clinical phenotype terms to describe the full range of relevant Mendelian phenotypes. We used these terms to search the OMIM database via API for all Mendelian disorders demonstrating these phenotypes, then extracted the gene(s) linked to each Mendelian disorder. We restricted gene-phenotype associations to those satisfying the same criteria (1) and (2) as described above, and with the following additional criteria: (3) gene-phenotype association description does not contain “genome-wide association study” or other GWAS synonyms unless: the description also contains any of the terms “missense”, “nonsense”, “nonsynonymous”, or “frameshift”; or the gene contains at least one pathogenic or likely pathogenic allele in the ClinVar database. We include a full list of phenotype-specific Mendelian disorder gene sets and clinical phenotype terms used in **Table S2**.

A comparison of all phenotype-specific Mendelian disorder gene sets revealed a high degree of overlap among the gene sets for clinically-related Mendelian phenotypes (**Figure S1**). Accordingly, we clustered gene sets based on pairwise overlap, and intersected gene sets clustering together by visual inspection at a hierarchical clustering threshold to create a single gene set for the representative group of Mendelian disorders. Each complex trait was thus matched with the single Mendelian disorder category in which the original specific Mendelian disorder gene set clustered, which ultimately best exemplified the phenotype.

Due to the systemic and pleiotropic nature of complex traits, some complex traits could conceivably be phenotypically-related to more than one Mendelian disorder gene set. For

example, we generated the Mendelian disorder gene set for Systemic Lupus Erythematosus (SLE, MIM: 152700) using clinical keywords for both the driving immunological event and the clinical manifestations associated with SLE autoimmunity across a large number of organ systems (kidney, brain, skin, pleura, joints, etc), such as "anemia". Although the substantial contribution of Mendelian disorder genes related to anemia resulted in SLE pairing with the Hematological Disorders group, the immunological component of SLE is central to the disease. Thus, we identified Immune Dysregulation as a "relevant phenotype" for SLE, and denoted it as such in **Figure 2** and **Table 2**; the same occurred with other traits and also appear in **Figure 2** and **Table 2**.

After combining similar gene sets, a total of 20 non-disjoint phenotype-specific Mendelian disorder gene sets remained with an average of 375 genes per set; we include a description of each cluster in **Table S3**.

Complex trait gene sets

We downloaded publicly available summary statistics (per-allele SNP effect sizes, or log-odds ratios for case–control traits, with standard errors[29]) for large-scale GWAS of 62 traits[26] (**Table 1** and **Table S1**; average N=83,170, minimum N=10,610, maximum N=298,420; some GWAS were imputed using the 1000 Genomes Project as a reference panel by their respective consortia while others were not.) For each trait, we identified a gene set by mapping each autosomal genome-wide significant SNP ($p < 5x10^{-8}$) to the closest up- and downstream protein-coding genes as defined above, resulting in a total of 62 non-disjoint GWAS gene sets. As GWAS regions often contain multiple genome-wide significant SNPs, and the relevant gene may not lie adjacent to the lead SNP in a region[30; 31], we defined GWAS gene sets by mapping genes with respect to every genome-wide significant SNP rather than only the index GWAS SNPs at each genomic risk region.

Quantifying overlap between complex trait and Mendelian disorder

For each complex trait-Mendelian disorder pair, we compared the GWAS gene set and phenotype-specific Mendelian disorder gene set using a 2x2 contingency table (counting whether each gene was in the GWAS gene set or not, and in the Mendelian disorder gene set or not), with the set of autosomal protein-coding genes (n=17,695) representing the total sample. We used Fisher's Exact Test[32] to determine significance. Phenotype-specificity of overlap significance was assessed by comparing the GWAS gene sets for each complex trait (n=62) to all phenotype-specific Mendelian disorder gene sets (n=20), a total of 1,240 pairs. Significance was assessed at an FDR < 5% threshold ($p < 0.00310$).

To assess the robustness and stability of our SNP-gene mapping approach for complex traits, we performed an overlap quantification with phenotype-specific Mendelian disorder genes using GWAS gene sets derived from three additional SNP-gene mapping methods: by mapping each genome-wide significant SNP to all genes within a 50Mb window, to all genes within a 500Mb window, and by mapping all SNPs in the credible set to the closest two genes. Comparison of the odds ratios produced by Fisher's Exact Test for the comparisons of GWAS gene sets (derived by each mapping method) and phenotype-specific Mendelian disorder gene sets demonstrates no major difference in outcomes from different mapping methods (**Table S4**) ; thus, we find that even more conservative gene sets, such as the GWAS gene sets derived from the credible set for each complex trait, still demonstrate the pattern of trait-specific enrichment.

Estimating enrichment of GWAS SNP association signal

We created genomic annotations to capture the regions spanning 50kb upstream through 50kb downstream of gene bodies for four categories of genes: all protein-coding genes (N=17,695),

all Mendelian disorder genes (N=3,446), all LOF-intolerant genes (N=2,978), and the phenotype-specific Mendelian disorder gene sets (average N=609). For each complex trait-gene category pair, we computed enrichment of GWAS signal within the category c with respect to the set of all protein-coding genes as

$$a_c = \frac{\frac{1}{N_c}\sum_{j=1}^{N_c}\sum_{i=1}^{M_j}\frac{Z_i^2}{M_j}}{\frac{1}{N_p}\sum_{j=1}^{N_p}\sum_{i=1}^{M_j}\frac{Z_i^2}{M_j}}$$

where N_c is the number of genes in category c, M_j is the number of SNPs within 50kb of gene j, Z_i = GWAS effect size of SNP i divided by standard error, with total number of protein-coding genes N_p. Thus, a_c is the enrichment in average SNP effect size (Z^2) per gene in category (compared to average Z^2 for any protein-coding gene). The percent increase in average SNP effect size per gene for category c, or $(a_c - 1) * 100$, is shown in **Figure 3**. We performed similar comparisons for median SNP effect size per gene for category c, and maximum SNP effect size per gene for category c (**Table S5**).

To ensure that this signal was not driven by linkage disequilibrium (LD), minor allele frequency (MAF), or average gene length per category, we compared these three properties across the gene categories for each complex trait. We calculated LD scores[33] reflecting the amount of LD tagged by each SNP in the HapMap 3 reference panel; then, for each gene category, we averaged the LD scores of SNPs falling within 50kb of each gene. Similar analyses were performed to examine average MAF per gene and average gene length per category across each complex trait (**Table S6**). For comparison, we additionally performed a permutation test by drawing 100 sets of random genes for each Mendelian disorder gene set, matched for number and length of genes, and computing the average effect size per gene for each phenotypically-matched complex trait across all 100 random sets.

Putative causal mechanisms at GWAS risk regions

We performed statistical fine-mapping of the genome-wide significant regions ($p < 5x10^{-8}$) for each GWAS using fgwas[34] with no functional annotations and default parameter settings. For each GWAS, we constructed a 95% credible set (defined as the minimum set of SNPs where 95% of the probability of causation at a region is accumulated) for each region of 500 SNPs containing a significant GWAS association. We achieved this by adding SNPs one at a time with a decreasing posterior probability of causation (posterior probability of association for the SNP, conditioned on there being an association in the region) until a cumulative 95% probability of causation is reached.

Identification of candidate regulatory variants

We intersected credible sets for each complex trait with genomic regions 1kb upstream of each phenotypically-relevant Mendelian disorder gene to identify SNPs localizing at the TSS. To identify genes whose expression the GWAS SNPs may regulate, we queried the UCSC GTEx combined-eQTL table (version 2017-10-25) and joined on SNP rsID. This table describes all gene/tissue pairs where a SNP has evidence of regulatory function. We restricted results to phenotype-matched Mendelian genes whose promoter contained a genome-wide significant SNP in our GWAS fine-mapped results. To identify candidate regulatory variants interacting with promoters of phenotype-matched Mendelian disorder genes, we used interactions from promoter capture Hi-C in human primary white adipocytes [26] for each complex trait, and filtered interactions to pairs of interacting regions where at least one region contained a promoter of a phenotype-specific Mendelian disorder gene. We then intersected interaction pairs for each of these regions with credible sets for each complex trait to identify credible SNPs interacting with regions containing promoters of phenotype-specific Mendelian disorder genes.

Estimating the enrichment of SNP-heritability of complex traits within Mendelian disorder gene set annotations

We used stratified LD score regression (s-LDSC)[35] to estimate the enrichment of SNP-heritability of 47 complex traits and diseases within each of 20 Mendelian disorder gene set annotations, corresponding to the regions spanning 50kb upstream through 50kb downstream of gene bodies for each Mendelian disorder gene set. The 47 complex traits and diseases are a subset of the 62 total GWAS traits analyzed in this study that meet the criteria for running s-LDSC (i.e., the GWAS did not use custom genotyping arrays). The annotation value for SNP *i* and gene set *k* is defined as a_{ik} = 1 if SNP *i* is within 50kb upstream or 50kb downstream of any of the gene bodies in gene set *k*, and a_{ik} = 0 otherwise. For each of the 20 annotations, LD scores were computed within 1cM blocks using default parameters and LD estimated from the European individuals in the 1000 Genomes Phase 3 reference panel. For each GWAS/annotation pair, we ran s-LDSC using the recommended "baseline model"[35] as covariates in the regression, for a total of 53 annotations per run (52 "baseline" annotations + the gene set annotation of interest).

2.3 Results

GWAS risk genes show specific, significant overlap with phenotypically-matched Mendelian disorder genes

We first sought to examine the degree of overlap between phenotype-matched Mendelian disorder genes with risk genes for complex traits as identified through GWAS. For each complex trait, we identified corresponding Mendelian forms, often as familial forms or rare phenotypic extremes, and curated Mendelian disorder gene sets composed of Mendelian disorder genes linked to those specific phenotypes from the OMIM database (see Methods, and **Figure 1**). We combined similar Mendelian disorder gene sets to create one gene set for the representative

Mendelian disorder(s) (for a total of 20 Mendelian disorder gene sets). We separately ascertained GWAS gene sets for each complex trait by identifying the closest up- and downstream genes to each GWAS SNP meeting genome-wide significance (see Methods, and **Figure 1**). Overlap between each phenotype-specific Mendelian disorder gene set (n=20) and each GWAS gene set (n=62) was assessed using Fisher's Exact Test, for a total of 1,240 comparisons (**Table 1** and **Table S7**). We hypothesized that GWAS gene sets would have a specific significant enrichment of Mendelian disorder genes for perfectly matched Mendelian disorders (as identified in Table 1; 62 of the 1,240 comparisons) or related Mendelian disorders (an additional 30 of the 1,240 comparisons; 92 of 1240 total), but no enrichment for unrelated Mendelian disorders (the remaining 1,148 of 1,240 comparisons). Among all 1,240 pairs of complex and Mendelian disorder gene sets assessed, we identified 77 pairs with significant overlap crossing an FDR < 5% cutoff at a $p < 0.00310$ (**Figure 2**). An examination of the log-odds ratios for each overlap comparison revealed more extreme enrichments among phenotypically-matched pairs compared to phenotypically-unmatched pairs (**Table 2**), which is consistent with our hypothesis. 50 out of the 77 significantly overlapping pairs showed perfectly matching phenotypes (as defined in **Table 1**; see Methods) or reflected known shared biology (identified in dark blue within **Figure 2**). Specifically, in many of these pairs, monogenic forms of the complex trait have been well established in the genetics literature; examples include Age-related Macular Degeneration (AMD) and cholesterol traits (high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol (TC), and triglycerides (TG))[7; 36-39]. We confirmed significant enrichment between many of these previously reported pairs such as the complex and monogenic forms of height[40] (OR=1.39, $p=1.43 \times 10^{-3}$) and HDL and Mendelian forms of cardiovascular disease[41] (OR=2.10, $p=3.45 \times 10^{-4}$). We also identified previously unreported enrichments; for example, we find a strong enrichment between inflammatory bowel disease (IBD) and Mendelian forms of immune dysregulation (OR=3.32, $p=1.58 \times 10^{-8}$) and between

hemoglobin (HB) and Mendelian hematologic disorders (OR=3.99, p=4.42 x10^{-4}). The remaining 27 pairs with significant overlap suggested shared biological mechanisms yet to be established between complex traits and Mendelian disorders (**Table 3**). For example, we observed an enrichment between height and renal disorders (OR=1.48, p=3.75 x 10^{-5}), and enrichment between Crohn's Disease and mature-onset diabetes of the young (OR=2.69, p=2.32 x 10^{-4}). Strikingly, a high proportion of phenotypically-matched or related pairs demonstrated significant overlap (n=50 of 92; 54%) compared to phenotypically-unmatched pairs (n=27 of 1,148; 2%), consistent with our hypothesis of a phenotype-specific enrichment pattern (**Table S7** and **Figure S2**).

To investigate whether proximal clustering within chromosomes of genes with similar functionality was confounding our results, we pruned our dataset of all protein-coding genes to include only one gene per 33.2 KB window across the genome (determined by the average distance to the next closest gene in our data set) and re-computed overlap odds ratios. We found highly similar results to our original approach (Pearson r=0.96). Moreover, we found the average distance to the next closest gene among the sets of genes shared by a phenotypically-matched pair of complex trait and Mendelian disorder to be 21.1 MB (**Table S8**).

SNPs near phenotypically-matched Mendelian disorder genes show increased effect size on complex traits

Because Mendelian disorder genes are linked with severe biological effects when either one or both alleles are disrupted, dysregulation of the gene through changes in expression or other mechanisms might have a more significant effect than dysregulation of another protein coding gene not linked to any Mendelian disorders. We hypothesized that SNPs near these phenotype-specific Mendelian disorder genes have further increased effects on complex traits due to the increased biological relevance of these gene categories. From the publicly available GWAS

summary statistics for each complex trait, we computed the average GWAS effect sizes of SNPs falling within each protein-coding gene, and compared the average effect sizes per gene across all Mendelian disorder genes and across phenotypically-relevant Mendelian disorder genes (see Methods). Across complex traits, we found an increased average effect size per gene for all Mendelian disorder genes and a further increased average effect size per gene for phenotypically-relevant Mendelian disorder genes (**Figure 3 and Table S5**). This suggests that the genomic regions containing the most biologically-relevant genes for each trait contribute most significantly to complex trait biology. We also confirmed that loss-of-function (LOF) intolerant genes (as defined by ExAC's pLI score > 0.9, see Methods) demonstrate a higher average effect size across most complex traits examined[28]. Given the extreme intolerance of deleterious mutations in these genes, it is possible that LOF-intolerant genes are linked with embryonic lethal mutant phenotypes, and are thus undiscovered as Mendelian disorder genes at this time.

We found no significant increase in linkage disequilibrium (LD) or decrease in average minor allele frequency (MAF) of the SNPs within each category compared to the SNPs within all protein-coding genes (**Table S6)**, suggesting that the observed signal is not driven by any of these confounders. In particular, we found the average LD (95% confidence interval) tagged for all protein coding genes to be 24.38 (24.35, 24.42); for none of the other three gene classes did the confidence intervals fall above the upper bound, including the average across all Mendelian disease gene sets (24.10 (23.47, 24.72)). Similarly, we found the average MAF (95% confidence interval) for all protein-coding genes to be 0.238 (0.238, 0.238); for none of the other gene classes did the confidence intervals fall below the lower bound, including the average across all Mendelian disease gene sets (0.238 (0.237, 0.240)). Details for each gene set are included in **Table S6**. Of note, we did observe a respective increase in average gene length (95% confidence interval) between all protein-coding genes, particularly between all protein-

coding genes (159.94 kb (0.91 kb)) and on average across all phenotype-specific Mendelian disorder gene sets (177.87 kb, (168.49 kb, 187.24 kb)) (**Table S6**). To ensure that our findings of enriched GWAS signal in these gene categories was not due to longer genes being more likely to tag causal variation, we performed a permutation test comparing the average effect size per gene for phenotype-matched Mendelian disorder genes to the same metric across 100 sets of random genes matched for number of genes and gene length (**Table S6**). We find no evidence of gene length confounding our results, as across 58/62 complex traits the average effect size per gene is higher for phenotype-matched Mendelian disorder genes than for random genes of the same length. Also of note, we did not find evidence of a pervasive phenotype-specific enrichment of SNP heritability within 50kb of Mendelian disorder genes, (**Table S9;** see Methods). Thus we can conclude that the average GWAS effect size per gene for phenotypically-relevant Mendelian disorder genes is increased relative to all protein coding genes, all Mendelian disorder genes, and LOF-intolerant genes, but not necessarily for unrelated sets of Mendelian disorder genes.

Examples of credible SNPs for GWAS regions near phenotypically-matched Mendelian disorder genes

We next sought to identify common non-coding variants that may causally impact complex trait phenotypes by dysregulating phenotypically-relevant Mendelian disorder genes. For each complex trait, we performed statistical fine-mapping of significant GWAS regions to construct 95% credible sets for each region (see Methods), and identified SNPs from the credible set located at the TSS of a gene from the phenotypically-relevant Mendelian disorder gene set. We found a total of 786 credible set SNPs (out of approximately 3.5 million) localizing at the TSS of a phenotypically-relevant Mendelian disorder gene (an average of 20 SNPs per trait, for 38 traits where at least one such SNP was found; **Tables S10 and S11**), and identified 25 promising

candidate SNPs (attaining genome-wide significance in GWAS) at TSSs that could be regulating the proximal Mendelian disorder gene (**Table 4**). We further examined the GTEx database to determine whether any of these SNPs were also eQTLs for the corresponding gene; we found 12 variants to be significant eQTLs for the corresponding gene in at least one tissue (**Table 4**). We highlight two examples: first, we found a significantly associated SNP from the credible set for coronary artery disease (CAD) (rs1332327, Z=6.798) at the promoter of *LIPA* [MIM: 278000], a Mendelian disorder gene linked to Wolman Disease and Cholesteryl Ester Storage Disease (both Lysosomal Acid Lipase Deficiencies, MIM: 278000) involving hypercholesterolemia and hypertriglyceridemia as part of cholesteryl ester- and triglyceride-filled macrophage infiltration syndromes (**Figure 4A**). Additional analyses identified rs1332327 (along with other SNPs in linkage disequilibrium with the variant), as a *cis*-eQTL for *LIPA* in the METSIM adipose RNA-sequencing dataset (**Table S13**); this finding is consistent with eQTL results reported in GTEx for these SNPs and *LIPA*. Second, from the credible set for red blood cell count (RBC), we found a significantly associated SNP (rs1010222, Z= -5.961) at the promoter of *CALR* [MIM: 109091], a Mendelian disorder gene linked to Myelofibrosis [MIM: 254450] involving generalized bone marrow fibrosis, reduced hemopoiesis, no hemophagocytosis, and myeloproliferative disease (**Figure 4B)**. In both cases, the putative causal SNP for the complex trait lies immediately upstream of the TSS of the phenotypically-relevant Mendelian disorder gene, in addition to falling within regions containing by regulatory epigenetic marks.

Putative causal SNPs for GWAS regions interacting with promoters of phenotypically-relevant Mendelian disorder genes

Functional genomic datasets, such as chromatin interactions identified through Hi-C, can give us insight into the functional interpretation of GWAS variants and how they might regulate

Mendelian disorder genes. Examination of chromatin interactions in human primary white adipocytes[26] revealed further candidate credible set SNPs for metabolic traits physically interacting with promoters of phenotypically-relevant Mendelian disorder genes (**Table S12)**. Specifically, we report that a genome-wide significant SNP for BMI (rs758747, Z=6.081) physically interacts with the promoter of *CREBBP*, a gene linked to Rubinstein-Taybi Syndrome 1 [MIM: 180849] in which obesity is one of the syndromic features (**Figure 4C**). These interactions can also identify the relevant isoforms of genes in disease. We identified a cluster of SNPs from the credible set of variants associated with BMI that physically interact with the promoter of a specific isoform of *CYP19A1*, a gene linked to Aromatase Excess Syndrome [MIM: 139300] involving short stature and excess fat storage in the chest (gynecomastia) **(Figure 4D)**. Although longer isoforms of *CYP19A1* are by default chosen to represent the gene, our data suggests that the shorter isoform is likely to be more relevant in obesity. Taken together, these results demonstrate examples of GWAS variants localizing in regulatory regions for phenotypically-relevant Mendelian disorder genes, consistent with the hypothesis that low-effect common variants contribute to complex traits by regulating genes known to cause Mendelian disorders.

2.4 Discussion

In this work we used GWAS summary statistics from 62 complex traits and genes linked to specific phenotypes within 20 Mendelian broad disorders to quantify the shared genetic basis of complex traits and Mendelian disorders. We identified a specific enrichment of phenotypically-matched and related Mendelian disorder genes in GWAS regions for complex traits; we also identified fewer pairs of complex traits and phenotypically-unmatched Mendelian disorders with similar significant enrichment. We further found that phenotypically-relevant Mendelian disorder genes are enriched for GWAS signal across complex traits, compared to all Mendelian disorder

genes and other protein-coding genes. Finally, we report examples of putative causal SNPs for GWAS regions in potentially regulating phenotypically-relevant Mendelian disorder genes. We conclude with four considerations about how our results contribute to understanding of genetic architectures and biological mechanisms across complex traits and Mendelian disorders.

First, our finding of a specific enrichment of phenotypically-matched and related Mendelian disorder genes in GWAS regions for complex traits suggests that, across complex trait architectures, many complex traits share the genetic bases (and by extension, biological mechanisms) with their Mendelian forms. This supports our hypothesis that the shared genes contribute to both extreme and common genetic phenotypes, and suggests an important role of gene regulation by non-coding variants in complex traits. However, we note that our findings are limited by the power of each GWAS to detect significant associations. As GWAS become better-powered, we anticipate being able to identify phenotype-specific enrichments of Mendelian disorder genes in GWAS regions for more complex traits.

Second, the subset of complex trait-Mendelian disorder pairs with no known shared biology that still demonstrated significant enrichment of Mendelian disorder genes in GWAS regions can offer us insight into the biological mechanisms of complex traits and Mendelian disorders. A high degree of co-morbidity between complex traits and Mendelian disorders has been previously observed, regardless of phenotype-similarity[24]; these findings together suggest that many complex traits and Mendelian disorders may also be linked by the pleiotropic properties of the underlying genes, in addition to regulatory differences. These observations are also consistent with a multigenic or oligogenic architecture of human disease; the pervasive pleiotropic effects that are seen observed across complex traits are consistent with the wide-spread prevalence of multi-system, syndromic phenotypes observed across a majority of Mendelian disorders. We also confirm that LOF-intolerant genes harbor an enrichment of GWAS signal[28]; because genes with pLI > 0.9 exhibit extreme intolerance of deleterious

mutation, it is possible that these genes demonstrate embryonic lethal mutant phenotypes, and are thus undiscovered as Mendelian disorder genes at this time. Our findings provide further motivation to explore phenotypic consequences of mutations in LOF-intolerant genes (particularly those enriched for GWAS signal for a particular complex trait) for phenotypically-relevant Mendelian disorders.

Third, linking Mendelian disorder genes with complex traits can help with characterization of the genetic architecture of complex traits – specifically, with genes and pathways that can be functionally characterized to identify molecular mechanisms[7]. Identifying causal variants from large-scale GWAS studies is particularly challenging given that most GWAS loci lie in non-coding regions of the genome; though thousands of genomic loci have been significantly associated with specific diseases, few casual SNPs have been functionally verified[42; 43]. Although many approaches have been used to tie a particular variant to a relevant gene or genes[44-46], including newer methods that directly link gene expression to a trait (e.g., TWAS[30], PrediXcan[47]), we find that leveraging GWAS findings with functional data to identify candidate regulatory variants for Mendelian disorder genes can potentially lead to better interpretation of relevant genes and isoforms. Here, we demonstrate the heterogeneity of mechanisms potentially underlying causal variation, showing roles for TSS promoter regions of Mendelian disorder genes and long-range interactions involving significant GWAS regions. We expand on recent work showing that BMI-associated variants interact with genes in GWAS regions to demonstrate similar findings for Mendelian disorder genes[26]. With the appropriate functional data from relevant tissues and cell types, this phenotype-driven approach can identify relevant candidate regulatory variants and their targets. Further, from the perspective of monogenic diseases, identifying common variants that might modify the expressivity of phenotypes can provide insights into gene function in addition to putative drug targets. Many drugs approved by the FDA and developed by pharmaceutical companies are targeted towards

the treatment of complex traits and diseases; by identifying underlying links between Mendelian disorders and complex traits through their effects on the same biological genes and pathways, we can systematically and rationally target existing drugs for complex traits and diseases towards those with rare Mendelian disorders which largely do not have any rationally targeted treatments[48-50].

Last, we note that our approach of examining traits and disorders at the component-phenotype level offers us valuable resolution into the specific pathways involved the overall trait or disorder. In clinical medicine, genome-wide sequencing has expanded the clinical phenotypic spectrum associated with a gene [51; 52] through identification of pleiotropic effects due to mutations in specific protein domains[53; 54], detected a genetic predisposition for diseases previously considered to be due to environment[13], uncovered variable penetrance for genetic mutations previously thought to be sufficient to cause disease, and has suggested that genetic background influences the phenotypic variability of monogenic diseases[55; 56]. The phenotypic characterizations of Mendelian syndromes are deconstructed by expert clinical geneticists into component phenotypes, labeled by standardized clinical terms that identify both the primary phenotypes and phenotypes that have variable penetrance and expressivity[57]. Recent work has demonstrated that incorporation of such dense phenotype information to rank putative disease-causing genetic mutations improves diagnostic rates in clinical exome sequencing tests[58; 59]; using component Mendelian phenotypes to identify Mendelian disorders that may be phenotypically-relevant to a variety of complex traits can be similarly impactful in identifying biological pathways for complex traits. Ultimately, identification of GWAS-significant regions with biologically relevant genes and pathways will enable effective utilization of GWAS data in medical settings.

2.5 Figures

Figure 2.1 GWAS gene sets and phenotype-specific Mendelian disorder gene sets.

For each complex trait (e.g., height), we first identified matched Mendelian phenotypes (e.g., undergrowth, short stature; **Table S10**). Using publicly available GWAS data, we defined the "GWAS genes" for a given complex trait to be the closest upstream and closest downstream protein-coding gene for every genome-wide significant variant in the GWAS. We selected phenotype-matched Mendelian disorder genes by first identifying Mendelian disorders expressing any of the matched Mendelian phenotypes, and then identifying all genes linked to any of those disorders.

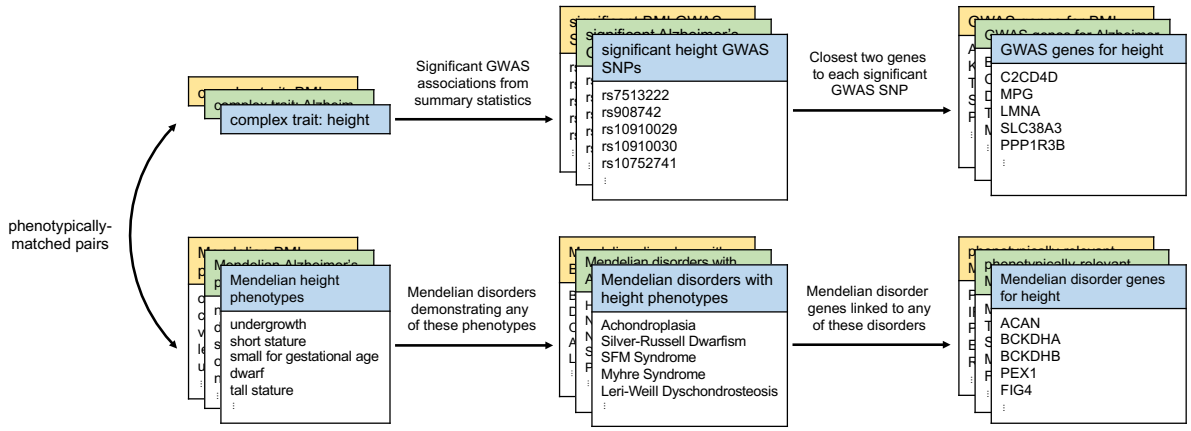


Figure 2.2 Overlap of GWAS genes with Mendelian disorder genes demonstrates trait-specificity.

Significant overlaps from phenotypically-matched pairs of complex traits and Mendelian disorders (blue) and pairs with unrelated phenotypes (grey) are shown. Phenotypically-matched pairs are subdivided into pairs with perfectly-matched phenotypes (light blue) and pairs with related phenotypes (dark blue). Complex traits and Mendelian disorders with no significant overlaps are excluded here; results from all traits are presented in Figure S2. Significance was assessed by controlling for FDR < 5% at $p < 0.00310$.

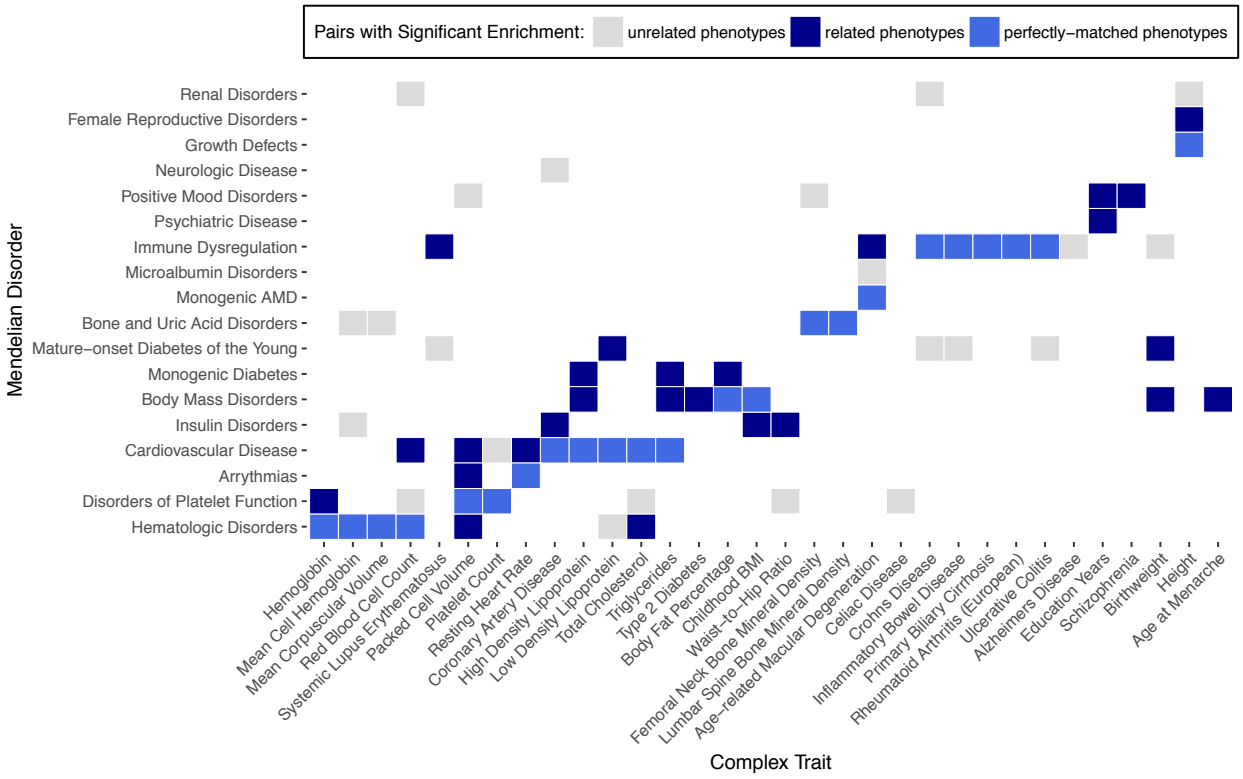


Figure 2.3 Effect sizes for SNPs on complex traits from GWAS are increased for genes that are loss-of-function intolerant and for phenotypically-relevant Mendelian disorder genes.

The increase in average SNP effect size per gene across gene categories. We averaged effect size (Z^2) across all SNPs falling within 50kb of a gene to obtain an average SNP effect size per gene, and averaged across all genes in each category (all protein coding genes, all Mendelian disorder genes, all LOF-intolerant genes, and all phenotypically-relevant Mendelian disorder genes for each trait). We normalized these averages to the average SNP effect per gene for any protein coding genes. The box plots represent the distribution of increase in average effect size per gene across all traits, and notches designate the confidence intervals. From left to right, confidence intervals read: (0.07, 1.24), (1.47, 3.54), (5.88, 12.19).

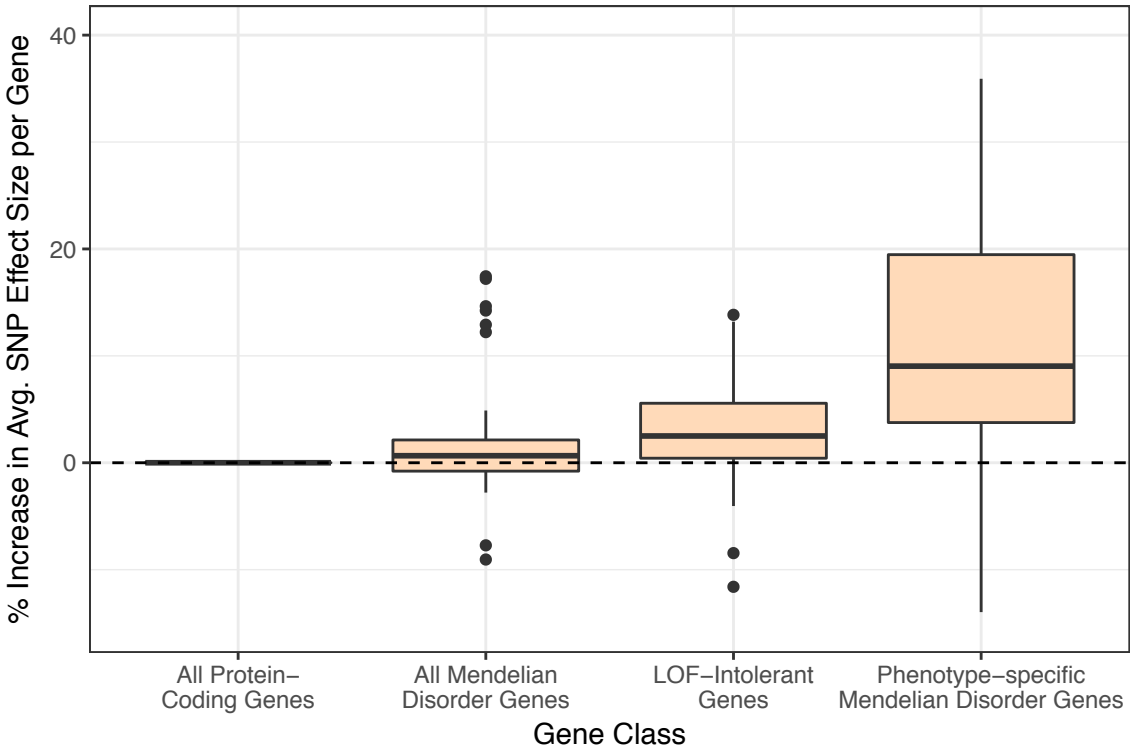

% Increase in Avg. SNP Effect Size per Gene
40
20
0
All Protein–
Coding Genes
All Mendelian
Disorder Genes
LOF–Intolerant
Genes
Phenotype–specific
Mendelian Disorder Genes
Gene Class

Figure 2.4 Candidate regulatory SNPs fall at transcription start sites and long-range promoters of phenotypically-relevant Mendelian disorder genes

A, B) Shown here are two examples of putative causal SNPs localizing at a TSS of a phenotypically-relevant Mendelian disorder gene. **A**) Putative causal SNP rs1332327, associated with coronary artery disease (Z = 6.796), lies at the TSS of *LIPA*. **B**) Putative causal SNP rs1010222, associated with red blood cell count with a Z-score of -5.961, lies at the TSS of *CALR*. **C, D)** Shown here are two representations of chromatin interactions in white adipose tissue. **C**) A cluster of SNPs from the credible set of variants associated with BMI (Z-score plotted in orange and blue) physically interacts with the promoter of a particular isoform of *CYP19A1*. **D**) A single SNP (rs758747) from the credible set, associated with BMI (Z = 6.081), physically interacts with the promoter of a distant gene *CREBBP*.

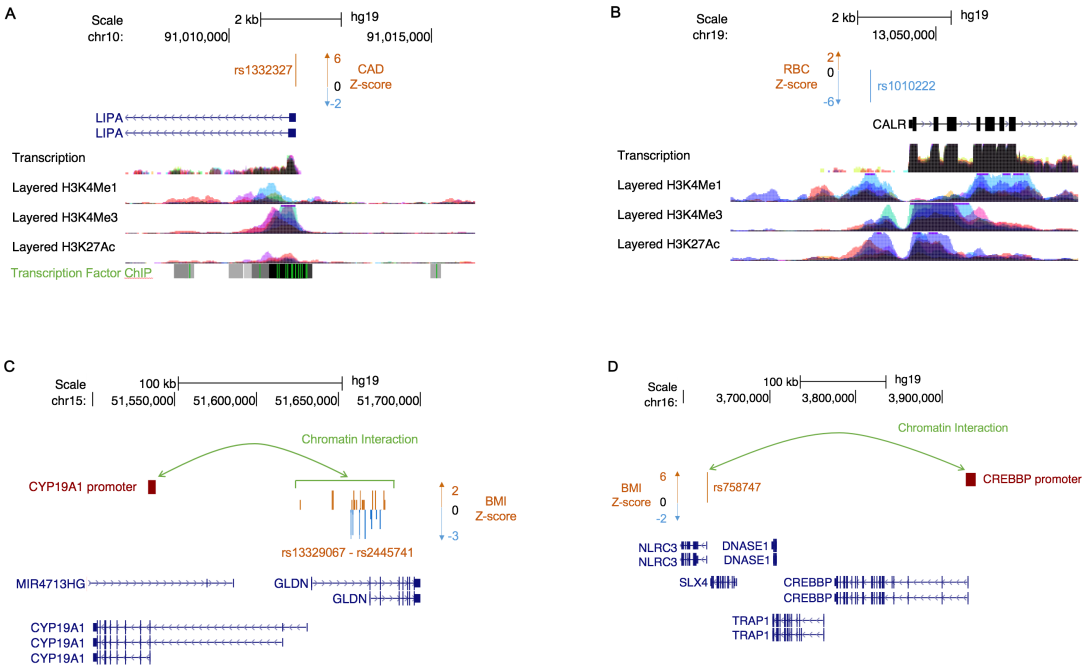


Figure 2.5 Similarity of Mendelian disorder gene sets

After generation of phenotype-specific Mendelian disorder gene sets, we performed pairwise comparisons of each gene set to determine proportions of genes shared. We performed Hierarchical clustering was performed, and gene sets sharing large proportions of genes (identified by visual clusters) were intersected to form a single representative Mendelian disorder gene set (see Supplementary Table 2 for these cluster descriptions).

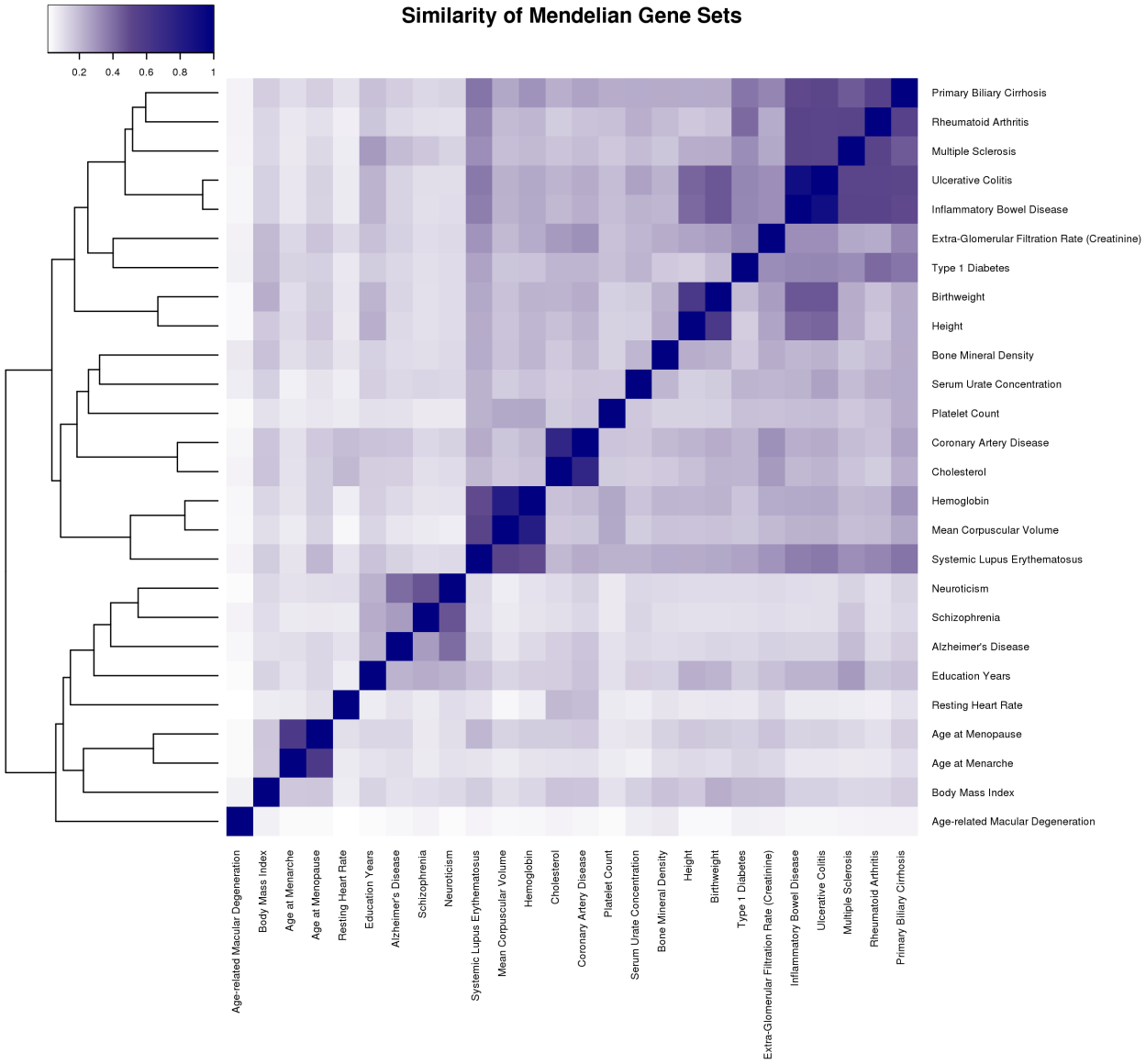
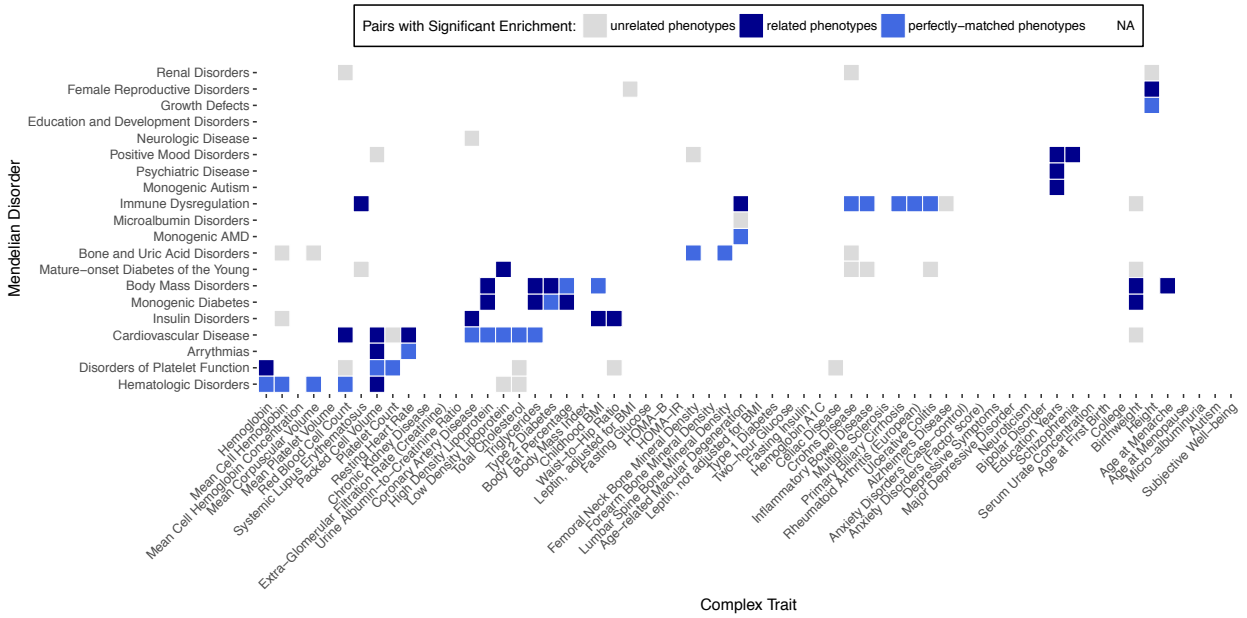


Figure 2.6 Overlap of GWAS genes with Mendelian disorder genes demonstrates trait-specificity

After generation of phenotype-specific Mendelian disorder gene sets, we performed pairwise comparisons of each gene set to determine proportions of genes shared. We performed hierarchical clustering, and gene sets sharing large proportions of genes (identified by visual clusters based on a hierarchical clustering threshold, indicated in red boxes and red dashed line respectively) were intersected to form a single representative Mendelian disorder gene set (see Table S3 for these cluster descriptions).



2.6 Tables

Table 2.1 Complex Traits and corresponding Mendelian disorders.

This table lists the phenotypically-matched pairs of complex traits (N=62) and groups of Mendelian disorders (N=20) examined in our study. More details on these traits, including mean GWAS sample size, number of significant GWAS loci reported from original GWAS publications, and number of significant GWAS SNPs are included in **Table S1**. GWAS genes for each complex trait were identified using the mapping approach described in Methods.

Complex Trait	Abbrev.	Number of GWAS Genes	Matched Mendelian Disorder(s)
Celiac Disease[60]	CEL	34	Immune Dysregulation
Crohn's Disease[61]	CD	239	
Inflammatory Bowel Disease[61]	IBD	368	
Ulcerative Colitis[67]	UC	202	
Primary Biliary Cirrhosis[62]	PBC	149	
Rheumatoid Arthritis (European)[63]	RA	297	
Multiple Sclerosis[64]	MS	160	
Autism[65]	AUT	2	Monogenic Autism
Hemoglobin[66]	HB	89	Hematologi Hematologic Disorders
Mean Cell Hemoglobin[66]	MCH	164	
Mean Cell Hemoglobin Concentration[66]	MCHC	12	
Mean Corpuscular Volume[66]	MCV	180	
Mean Platelet Volume[67]	MPV	102	
Red Blood Cell Count[66]	RBC	107	
Systemic Lupus Erythematosus[68]	SLE	286	
Birthweight[69]	BW	179	Growth Defects
Height[70]	HGT	2361	

Femoral Neck Bone Mineral Density[71]	FN	58	Bone and Uric Acid Disorders
Forearm Bone Mineral Density[71]	FA	8	
Lumbar Spine Bone Mineral Density[71]	LS	67	
Serum Urate Concentration[72]	URT	161	
Packed Cell Volume[66]	PCV	53	Disorders of Platelet Function
Platelet Count[67]	PLT	134	
Coronary Artery Disease[73]	CAD	132	Cardiovascular Disease
High Density Lipoprotein[74]	HDL	464	
Low Density Lipoprotein[74]	LDL	370	
Total Cholesterol[74]	TC	500	
Triglycerides[74]	TG	354	
Hemoglobin A1C[75]	HBA	33	Monogenic Diabetes
Type 2 Diabetes[76]	T2D	28	
Age-related Macular Degeneration[77]	AMD	215	Monogenic AMD
Age at Menarche[78]	MNR	207	Female Reproductive Disorders
Age at Menopause[79]	MNP	316	
Fasting Glucose[80]	FG	39	Insulin Disorders
HOMA-B[80]	HMB	12	
HOMA-IR[80]	HMIR	0	
Micro-albuminuria[81]	MA	2	Microalbumin Disorders
Fasting Insulin[80]	FI	23	Mature-onset Diabetes of the Young
Two-hour Glucose[82]	2HG	2	
Type 1 Diabetes[83]	T1D	144	
Alzheimer's Disease[84]	ALZ	58	Neurologic Disease
Anxiety Disorders (Case-control)[85]	ANXC	2	
Anxiety Disorders (Factor score)[85]	ANXF	3	
Major Depressive Disorder[86]	MDD	4	
Depressive Symptoms[87]	DS	10	
Neuroticism[87]	NRT	82	
Bipolar Disorder[88]	BIP	8	Psychiatric Disease
Schizophrenia[89]	SCZ	479	
Chronic Kidney Disease[90]	CKD	16	Renal Disorders
Glomerular Filtration Rate (CRN)[90]	EGFR	162	
Urine Albumin-to-Creatinine Ratio[81]	UACR	2	
Resting Heart Rate[91]	RHR	304	Arrythmias
Age at First Birth[92]	AFB	45	Education and Development Disorders
College[93]	COL	12	
Education Years[94]	EY	554	
Subjective Well-being[87]	SWB	9	Positive Mood Disorders
Body Fat Percentage[95]	BFP	22	Body Mass Disorders
Body Mass Index[96]	BMI	231	
Childhood BMI[97]	CBMI	49	
Leptin, adjusted for BMI[98]	LEPB	5	
Leptin, not adjusted for BMI[98]	LEP	0	
Waist-to-Hip Ratio[99]	WHR	74	

Table 2.2 Overlap of GWAS genes and phenotypically-matched Mendelian disorder genes.

For each pair of complex trait and Mendelian disorder, Fisher's exact test was used to quantify the enrichment of shared genes with an odds ratio and p-value (see Methods). Pairs with significant enrichment passed the cutoff of FDR < 5% at $p < 0.00310$. This table lists pairs of complex traits and phenotypically-matched or related Mendelian disorders with significant overlap. For comparison, the average odds ratio and 95% confidence interval for pairings of each complex trait with all unrelated Mendelian disorder gene sets is included.

Complex Trait (# genes)	Matched or Related Mendelian Disorder (# genes)	Shared Genes	Odds Ratio for Matched Pair (CI)	Raw P-value	Average Odds Ratio for Unmatched Pairs (CI)
AMD (215)	Monogenic AMD (104)	9	7.99 (3.50, 16.11)	4.94E-06	1.69 (1.21, 2.16)
	Immune Dysregulation (550)	17	2.73 (1.55, 4.52)	4.13E-04	
BFP (22)	Body Mass Disorders (128)	3	22.14 (4.14, 76.70)	5.15E-04	2.63 (0.75, 4.51)
	Monogenic Diabetes (182)	3	15.42 (2.90, 53.04)	1.43E-03	
BW (179)	Body Mass Disorders (128)	6	4.94 (1.76, 11.29)	1.93E-03	1.93 (1.48, 2.37)
	Monogenic Diabetes (182)	7	4.03 (1.57, 8.66)	2.56E-03	
CAD (132)	Cardiovascular Disease (598)	13	3.17 (1.63, 5.67)	5.36E-04	2.13 (1.65, 2.61)
	Insulin Disorders (623)	13	3.04 (1.56, 5.43)	7.83E-04	
CBMI (49)	Body Mass Disorders (128)	5	16.18 (4.92, 41.63)	2.71E-05	1.55 (1.13, 1.97)
	Insulin Disorders (623)	7	4.61 (1.74, 10.41)	1.54E-03	
CD (239)	Immune Dysregulation (550)	23	3.42 (2.10, 5.32)	1.70E-06	2.09 (-0.39, 4.58)
EY (554)	Positive Mood Disorders (69)	9	4.70 (2.04, 9.59)	2.88E-04	0.94 (0.81, 1.07)
	Monogenic Autism (111)	10	3.10 (1.44,	2.54E-03	

			5.98)		
	Psychiatric Disease (264)	19	2.45 (1.44, 3.95)	8.96E-04	
FN (58)	Bone and Uric Acid Disorders (220)	5	7.64 (2.36, 19.24)	7.61E-04	3.79 (2.36, 5.22)
HB (89)	Disorders of Platelet Function (443)	12	6.21 (3.05, 11.59)	2.17E-06	2.38 (1.68, 3.09)
	Hematologic Disorders (551)	10	3.99 (1.83, 7.79)	4.42E-04	
HDL (464)	Body Mass Disorders (128)	12	3.92 (1.95, 7.17)	1.39E-04	1.18 (1.00, 1.37)
	Monogenic Diabetes (182)	15	3.41 (1.85, 5.85)	9.36E-05	
	Cardiovascular Disease (598)	31	2.10 (1.40, 3.06)	3.45E-04	
HGT (2361)	Female Reproductive Disorders (288)	61	1.76 (1.30, 2.36)	2.18E-04	1.31 (1.19, 1.42)
	Growth Defects (723)	126	1.39 (1.13, 1.70)	1.43E-03	
IBD (368)	Immune Dysregulation (550)	34	3.32 (2.23, 4.79)	1.58E-08	1.38 (1.10, 1.66)
LDL (370)	Cardiovascular Disease (598)	31	2.70 (1.79, 3.95)	3.45E-06	1.66 (1.24, 2.08)
	Mature-onset Diabetes of the Young (561)	24	2.17 (1.36, 3.32)	8.54E-04	
LS (67)	Bone and Uric Acid Disorders (220)	6	8.00 (2.80, 18.72)	1.83E-04	2.53 (1.98, 3.08)
MCH (164)	Hematologic Disorders (551)	15	3.19 (1.73, 5.48)	1.92E-04	1.64 (1.16, 2.11)
MCV (180)	Hematologic Disorders (551)	20	4.00 (2.36, 6.44)	8.90E-07	1.61 (1.18, 2.04)
MNR	Body Mass	7	5.02	7.76E-04	1.03 (0.78, 1.27)

(207)	Disorders (128)		(1.95, 10.86)		
PBC (149)	Immune Dysregulation (550)	13	3.03 (1.56, 5.40)	7.77E-04	1.07 (0.76, 1.38)
PCV (53)	Disorders of Platelet Function (443)	10	9.24 (4.11, 18.83)	6.50E-07	4.04 (2.22, 5.87)
	Arrythmias (275)	5	6.70 (2.07, 16.94)	1.36E-03	
	Hematologic Disorders (551)	8	5.60 (2.27, 12.08)	2.17E-04	
	Cardiovascular Disease (598)	7	4.39 (1.66, 9.84)	1.94E-03	
PLT (134)	Disorders of Platelet Function (443)	12	3.91 (1.95, 7.15)	1.40E-04	1.42 (0.98, 1.87)
RA (297)	Immune Dysregulation (550)	25	2.95 (1.86, 4.50)	6.80E-06	0.83 (0.65, 1.01)
RBC (107)	Hematologic Disorders (551)	14	4.78 (2.50, 8.50)	5.82E-06	2.76 (2.20, 3.33)
	Cardiovascular Disease (598)	13	4.02 (2.05, 7.26)	6.49E-05	
RHR (304)	Arrythmias (275)	17	3.93 (2.23, 6.53)	5.87E-06	1.29 (0.95, 1.64)
	Cardiovascular Disease (598)	26	2.75 (1.75, 4.16)	1.47E-05	
SCZ (479)	Positive Mood Disorders (69)	9	5.47 (2.37, 11.19)	9.71E-05	1.11 (0.94, 1.28)
SLE (286)	Immune Dysregulation (550)	24	2.94 (1.83, 4.52)	1.09E-05	1.41 (1.10, 1.72)
T2D (28)	Body Mass Disorders (128)	4	23.55 (5.85, 69.99)	4.67E-05	2.43 (1.57, 3.30)
	Monogenic Diabetes (182)	3	11.72 (2.24, 38.93)	2.90E-03	
TC (500)	Cardiovascular Disease (598)	38	2.44 (1.69,	3.46E-06	1.40 (1.10, 1.70)

			3.45)		
TG (354)	Body Mass Disorders (128)	9	3.77 (1.67, 7.49)	1.10E-03	1.26 (1.01, 1.51)
	Monogenic Diabetes (182)	12	3.54 (1.78, 6.43)	3.10E-04	
	Cardiovascular Disease (598)	25	2.22 (1.41, 3.38)	5.06E-04	
UC (202)	Immune Dysregulation (550)	21	3.72 (2.23, 5.92)	1.39E-06	1.45 (1.12, 1.78)
WHR (74)	Insulin Disorders (623)	9	3.83 (1.67, 7.78)	1.13E-03	2.37 (1.71, 3.04)

Table 2.3 Instances of significant overlap of GWAS genes and unrelated Mendelian disorder genes.

As in Table 2, Fisher's exact test was used to quantify the enrichment of shared genes between complex traits and Mendelian disorders with an odds ratio and p-value (see Methods). Pairs with significant enrichment passed the cutoff of FDR < 5% at $p < 0.00310$. This table lists pairs of complex traits and phenotypically-unrelated Mendelian disorders that demonstrated significant overlap. For comparison, the average odds ratio and confidence interval for pairings of each complex trait with all remaining unrelated Mendelian disorder gene sets is included.

Complex Trait (# genes)	Matched or Related Mendelian Disorder (# genes)	Shared Genes	Odds Ratio (CI)	Raw P-value	Average Odds Ratio for Remaining Unrelated Pairs (CI)
ALZ (58)	Immune Dysregulation (550)	7	4.32 (1.65, 9.62)	2.06E-03	1.95 (1.36, 2.53)
AMD (215)	Microalbumin Disorders (159)	8	4.43 (1.86, 9.13)	7.37E-04	1.59 (1.03, 2.15)
BW (179)	Mature-onset Diabetes of the Young (561)	16	3.06 (1.69, 5.16)	1.91E-04	2.35 (1.94, 2.75)
	Immune Dysregulation (550)	14	2.69 (1.43, 4.68)	1.44E-03	
	Cardiovascular Disease (598)	14	2.46 (1.31, 4.28)	3.10E-03	
CAD (132)	Neurologic Disease (222)	7	4.52 (1.76, 9.75)	1.39E-03	2.34 (1.83, 2.86)
CD (239)	Bone and Uric Acid Disorders (220)	9	3.20 (1.42, 6.29)	3.10E-03	2.09 (-0.39, 4.58)
	Mature-onset Diabetes of the Young (561)	19	2.69 (1.58, 4.35)	2.32E-04	
	Renal Disorders (838)	23	2.17 (1.34, 3.37)	1.13E-03	

CEL (34)	Disorders of Platelet Function (443)	5	6.78 (2.04, 17.83)	1.47E-03	1.90 (1.28, 2.52)
FN (58)	Positive Mood Disorders (69)	3	14.51 (2.83, 46.53)	1.50E-03	3.79 (2.36, 5.22)
HGT (2361)	Renal Disorders (838)	153	1.48 (1.23, 1.78)	3.75E-05	1.31 (1.19, 1.42)
IBD (368)	Mature-onset Diabetes of the Young (561)	30	2.81 (1.85, 4.13)	2.40E-06	1.38 (1.10, 1.66)
LDL (370)	Hematologic Disorders (551)	25	2.31 (1.46, 3.51)	3.50E-04	1.66 (1.24, 2.08)
LEPB (5)	Female Reproductive Disorders (288)	2	40.52 (3.37, 353.67)	2.56E-03	10.83 (3.75, 17.91)
MCH (164)	Bone and Uric Acid Disorders (220)	8	4.19 (1.75, 8.61)	1.05E-03	1.65 (1.15, 2.16)
	Insulin Disorders (623)	15	2.80 (1.52, 4.81)	6.99E-04	
MCV (180)	Bone and Uric Acid Disorders (220)	8	3.80 (1.59, 7.79)	1.89E-03	1.74 (1.25, 2.22)
PCV (53)	Positive Mood Disorders (69)	3	15.97 (3.11, 51.37)	1.16E-03	4.04 (2.22, 5.87)
PLT (134)	Cardiovascular Disease (598)	12	2.85 (1.42, 5.20)	1.97E-03	1.42 (0.98, 1.87)
RBC (107)	Disorders of Platelet Function (443)	12	5.03 (2.49, 9.29)	1.51E-05	2.90 (2.13, 3.67)
	Renal Disorders (838)	18	4.14 (2.33, 6.96)	2.60E-06	
SLE (286)	Mature-onset Diabetes of the Young (561)	22	2.61 (1.59, 4.07)	1.24E-04	1.59 (1.24, 1.94)
TC (500)	Hematologic Disorders (551)	32	2.20 (1.47, 3.18)	1.18E-04	1.35 (1.05, 1.66)
	Disorders of Platelet Function (443)	25	2.11 (1.34, 3.20)	1.13E-03	
UC (202)	Mature-onset Diabetes of the Young (561)	19	3.25 (1.90, 5.27)	2.44E-05	1.41 (1.05, 1.77)
WHR (74)	Disorders of Platelet Function (443)	7	4.12 (1.59, 9.03)	2.50E-03	2.38 (1.58, 3.18)

Table 2.4 Genome-wide significant SNPs localizing at TSS of phenotypically-relevant Mendelian disorder genes.

GWAS SNPs from the credible set for each complex trait were intersected with transcription start site (TSS) regions 1kb upstream of phenotypically-matched Mendelian disorder genes. This table lists all genome-wide significant SNPs ($p < 5x10^{-8}$ from GWAS, with chromosomal location) from all complex traits localizing at the TSS of a phenotypically-matched Mendelian disorder gene (italicized).

Complex Trait	SNP ID	Chr: pos	Z score	Gene	Max. eQTL Effect	Max. -log10 P	Tissues
PBC	rs13239597	chr7:	9.85309	*TNPO3*			

		128695982					
HGT	rs8028537	chr15: 89345946	-9.333	*ACAN*			
HGT	rs10853751	chr19: 41903219	8.71	*BCKDHA*			
CD	rs59283234	chr5: 150225586	-8.454	*IRGM*	-0.042	6.733	wholeBlood
CD	rs751627	chr5: 150225112	-8.451	*IRGM*	-0.042	6.733	wholeBlood
CD	rs35707106	chr5: 150225376	-8.332	*IRGM*	-0.041	6.412	wholeBlood
HGT	rs2298307	chr6: 80816295	8.276	*BCKDHB*			
HGT	rs12386601	chr7: 92157885	8.2	*PEX1*			
BMI	rs17066842	chr18: 58040623	-7.542	*MC4R*			
HGT	rs12192268	chr6: 110011457	-7	*FIG4*			
CAD	rs1332327	chr10: 91011680	6.798	*LIPA*	12.541	5.132	adiposeSubcut, adiposeVisceral, adrenalGland, colonTransverse, lung, spleen, thyroid, wholeBlood
RA	rs13239597	chr7: 128695982	6.65672	*TNPO3*			
IBD	rs59283234	chr5: 150225586	6.51	*IRGM*	-0.042	6.733	wholeBlood
IBD	rs751627	chr5: 150225112	6.507	*IRGM*	-0.042	6.733	wholeBlood
HGT	rs7592246	chr2: 219926220	6.452	*IHH*	0.056	6.199	brainCerebellum
IBD	rs34005003	chr5: 150225198	6.427	*IRGM*	-0.043	6.637	wholeBlood
IBD	rs35707106	chr5: 150225376	6.326	*IRGM*	-0.041	6.412	wholeBlood
MNR	rs3775971	chr4: 104641919	6.20413	*TACR3*	-0.019	8.588	lung
IBD	rs27741	chr16: 28504180	6.109	*CLN3*			
HGT	rs4244808	chr11: 2163109	6.061	*IGF2*			
RBC	rs1010222	chr19: 13048607	-5.96154	*CALR*	17.142	6.851	lung
CD	rs27741	chr16: 28504180	-5.866	*CLN3*			
HGT	rs613924	chr11: 65769294	-5.862	*BANF1*			
AFB	rs4845357	chr1: 153896211	-5.775	*GATAD2B*	-0.108	4.443	skinNotExposed
HGT	rs6591226	chr11: 66675989	5.517	*PC*			

2.7 References

1. Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19, 212-219.
2. Cutting, G.R., Kasch, L.M., Rosenstein, B.J., Zielenski, J., Tsui, L.C., Antonarakis, S.E., and Kazazian, H.H., Jr. (1990). A cluster of cystic fibrosis mutations in the first nucleotide-binding fold of the cystic fibrosis conductance regulator protein. Nature 346, 366-369.
3. Henriksen, M.G., Nordgaard, J., and Jansson, L.B. (2017). Genetics of Schizophrenia: Overview of Methods, Findings and Limitations. Front Hum Neurosci 11, 322.
4. Katsanis, N. (2016). The continuum of causality in human genetic disorders. Genome Biol 17, 233.
5. Auer, P.L., Teumer, A., Schick, U., O'Shaughnessy, A., Lo, K.S., Chami, N., Carlson, C., de Denus, S., Dube, M.P., Haessler, J., et al. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nat Genet 46, 629-634.
6. Cohen, J., Pertsemlidis, A., Kotowski, I.K., Graham, R., Garcia, C.K., and Hobbs, H.H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet 37, 161-165.
7. Tsai, C.W., North, K.E., Tin, A., Haack, K., Franceschini, N., Saroja Voruganti, V., Laston, S., Zhang, Y., Best, L.G., MacCluer, J.W., et al. (2015). Both rare and common variants in PCSK9 influence plasma low-density lipoprotein cholesterol level in American Indians. J Clin Endocrinol Metab 100, E345-349.
8. Han, K., Holder, J.L., Jr., Schaaf, C.P., Lu, H., Chen, H., Kang, H., Tang, J., Wu, Z., Hao, S., Cheung, S.W., et al. (2013). SHANK3 overexpression causes manic-like behaviour with unique pharmacogenetic properties. Nature 503, 72-77.
9. Sztainberg, Y., and Zoghbi, H.Y. (2016). Lessons learned from studying syndromic autism spectrum disorders. Nat Neurosci 19, 1408-1417.
10. Lin, A., Ching, C.R.K., Vajdi, A., Sun, D., Jonas, R.K., Jalbrzikowski, M., Kushan-Wells, L., Pacheco Hansen, L., Krikorian, E., Gutman, B., et al. (2017). Mapping 22q11.2 Gene Dosage Effects on Brain Morphometry. J Neurosci 37, 6183-6199.
11. Toro, R., Konyukh, M., Delorme, R., Leblond, C., Chaste, P., Fauchereau, F., Coleman, M., Leboyer, M., Gillberg, C., and Bourgeron, T. (2010). Key role for gene dosage and synaptic homeostasis in autism spectrum disorders. Trends Genet 26, 363-372.
12. Posey, J.E., Harel, T., Liu, P., Rosenfeld, J.A., James, R.A., Coban Akdemir, Z.H., Walkiewicz, M., Bi, W., Xiao, R., Ding, Y., et al. (2017). Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. The New England journal of medicine 376, 21-31.
13. Corvol, H., Blackman, S.M., Boelle, P.Y., Gallins, P.J., Pace, R.G., Stonebraker, J.R., Accurso, F.J., Clement, A., Collaco, J.M., Dang, H., et al. (2015). Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. Nat Commun 6, 8382.
14. Emond, M.J., Louie, T., Emerson, J., Zhao, W., Mathias, R.A., Knowles, M.R., Wright, F.A., Rieder, M.J., Tabor, H.K., Nickerson, D.A., et al. (2012). Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. Nat Genet 44, 886-889.
15. Dorfman, R., Sandford, A., Taylor, C., Huang, B., Frangolias, D., Wang, Y., Sang, R., Pereira, L., Sun, L., Berthiaume, Y., et al. (2008). Complex two-gene modulation of lung disease severity in children with cystic fibrosis. J Clin Invest 118, 1040-1049.

16. Dron, J.S., and Hegele, R.A. (2017). Genetics of Triglycerides and the Risk of Atherosclerosis. Curr Atheroscler Rep 19, 31.
17. Hegele, R.A. (2018). Learning From Patients With Ultrarare Conditions: Cholesterol Hoof Beats. J Am Coll Cardiol 71, 289-291.
18. Peltonen, L., Perola, M., Naukkarinen, J., and Palotie, A. (2006). Lessons from studying monogenic disease for common disease. Hum Mol Genet 15 Spec No 1, R67-74.
19. Hernandez, D.G., Reed, X., and Singleton, A.B. (2016). Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. J Neurochem 139 Suppl 1, 59-74.
20. Lim, E.T., Liu, Y.P., Chan, Y., Tiinamaija, T., Karajamaki, A., Madsen, E., Go, T.D.C., Altshuler, D.M., Raychaudhuri, S., Groop, L., et al. (2014). A novel test for recessive contributions to complex diseases implicates Bardet-Biedl syndrome gene BBS10 in idiopathic type 2 diabetes and obesity. American journal of human genetics 95, 509-520.
21. Chan, Y., Salem, R.M., Hsu, Y.H., McMahon, G., Pers, T.H., Vedantam, S., Esko, T., Guo, M.H., Lim, E.T., Consortium, G., et al. (2015). Genome-wide Analysis of Body Proportion Classifies Height-Associated Variants by Mechanism of Action and Implicates Genes Important for Skeletal Development. American journal of human genetics 96, 695-708.
22. Wheeler, H.E., Gamazon, E.R., Frisina, R.D., Perez-Cervantes, C., El Charif, O., Mapes, B., Fossa, S.D., Feldman, D.R., Hamilton, R.J., Vaughn, D.J., et al. (2017). Variants in WFS1 and Other Mendelian Deafness Genes Are Associated with Cisplatin-Associated Ototoxicity. Clin Cancer Res 23, 3325-3333.
23. Amininejad, L., Charloteaux, B., Theatre, E., Liefferinckx, C., Dmitrieva, J., Hayard, P., Muls, V., Maisin, J.M., Schapira, M., Ghislain, J.M., et al. (2018). Analysis of Genes Associated with Monogenic Primary Immunodeficiency Identifies Rare Variants in XIAP in Patients With Crohn's disease. Gastroenterology.
24. Blair, D.R., Lyttle, C.S., Mortensen, J.M., Bearden, C.F., Jensen, A.B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam, E.V., Brunak, S., et al. (2013). A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. Cell 155, 70-80.
25. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42, D1001-1006.
26. Pan, D.Z., Garske, K.M., Alvarez, M., Bhagat, Y.V., Boocock, J., Nikkola, E., Miao, Z., Raulerson, C.K., Cantor, R.M., Civelek, M., et al. (2018). Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. Nat Commun 9, 1512.
27. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32, D493-496.
28. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291.
29. Pasaniuc, B., and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet 18, 117-127.
30. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. American journal of human genetics 100, 473-487.
31. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 48, 245-252.

32. Fisher, R.A. (1922). On the interpretation of χ2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85.
33. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics, C., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47, 291-295.
34. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. American journal of human genetics 94, 559-573.
35. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 47, 1228-1235.
36. Fritsche, L.G., Chen, W., Schu, M., Yaspan, B.L., Yu, Y., Thorleifsson, G., Zack, D.J., Arakawa, S., Cipriani, V., Ripke, S., et al. (2013). Seven new loci associated with age-related macular degeneration. Nat Genet 45, 433-439, 439e431-432.
37. Helgason, H., Sulem, P., Duvvari, M.R., Luo, H., Thorleifsson, G., Stefansson, H., Jonsdottir, I., Masson, G., Gudbjartsson, D.F., Walters, G.B., et al. (2013). A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. Nat Genet 45, 1371-1374.
38. Brunham, L.R., Singaraja, R.R., and Hayden, M.R. (2006). Variations on a gene: rare and common variants in ABCA1 and their impact on HDL cholesterol levels and atherosclerosis. Annu Rev Nutr 26, 105-129.
39. Kanoni, S., Masca, N.G., Stirrups, K.E., Varga, T.V., Warren, H.R., Scott, R.A., Southam, L., Zhang, W., Yaghootkar, H., Muller-Nurasyid, M., et al. (2016). Analysis with the exome array identifies multiple new independent variants in lipid loci. Hum Mol Genet 25, 4094-4106.
40. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42, 565-569.
41. Rosenson, R.S., Brewer, H.B., Jr., Barter, P.J., Bjorkegren, J.L.M., Chapman, M.J., Gaudet, D., Kim, D.S., Niesor, E., Rye, K.A., Sacks, F.M., et al. (2018). HDL and atherosclerotic cardiovascular disease: genetic insights into complex biology. Nat Rev Cardiol 15, 9-19.
42. Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., et al. (2016). Schizophrenia risk from complex variation of complement component 4. Nature 530, 177-183.
43. Smemo, S., Tena, J.J., Kim, K.H., Gamazon, E.R., Sakabe, N.J., Gomez-Marin, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 507, 371-375.
44. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467, 832-838.
45. de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 11, e1004219.
46. Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nat Commun 8, 1826.
47. Wang, J., Gamazon, E.R., Pierce, B.L., Stranger, B.E., Im, H.K., Gibbons, R.D., Cox, N.J., Nicolae, D.L., and Chen, L.S. (2016). Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx. American journal of human genetics 98, 697-708.

48. Pascual, V., Allantaz, F., Arce, E., Punaro, M., and Banchereau, J. (2005). Role of interleukin-1 (IL-1) in the pathogenesis of systemic onset juvenile idiopathic arthritis and clinical response to IL-1 blockade. J Exp Med 201, 1479-1486.
49. Pardoll, D.M. (2012). Immunology beats cancer: a blueprint for successful translation. Nat Immunol 13, 1129-1132.
50. Gerich, M.E., and McGovern, D.P. (2014). Towards personalized care in IBD. Nat Rev Gastroenterol Hepatol 11, 287-299.
51. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. American journal of human genetics 97, 199-215.
52. Reuter, C.M., Brimble, E., DeFilippo, C., Dries, A.M., Undiagnosed Diseases, N., Enns, G.M., Ashley, E.A., Bernstein, J.A., Fisher, P.G., and Wheeler, M.T. (2018). A New Approach to Rare Diseases of Children: The Undiagnosed Diseases Network. J Pediatr.
53. Goudie, D.R., D'Alessandro, M., Merriman, B., Lee, H., Szeverenyi, I., Avery, S., O'Connor, B.D., Nelson, S.F., Coats, S.E., Stewart, A., et al. (2011). Multiple self-healing squamous epithelioma is caused by a disease-specific spectrum of mutations in TGFBR1. Nat Genet 43, 365-369.
54. Arboleda, V.A., Lee, H., Parnaik, R., Fleming, A., Banerjee, A., Ferraz-de-Souza, B., Delot, E.C., Rodriguez-Fernandez, I.A., Braslavsky, D., Bergada, I., et al. (2012). Mutations in the PCNA-binding domain of CDKN1C cause IMAGe syndrome. Nat Genet 44, 788-792.
55. Born, H.A., Dao, A.T., Levine, A.T., Lee, W.L., Mehta, N.M., Mehra, S., Weeber, E.J., and Anderson, A.E. (2017). Strain-dependence of the Angelman Syndrome phenotypes in Ube3a maternal deficiency mice. Sci Rep 7, 8451.
56. Hensman Moss, D.J., Pardinas, A.F., Langbehn, D., Lo, K., Leavitt, B.R., Roos, R., Durr, A., Mead, S., investigators, T.-H., investigators, R., et al. (2017). Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. Lancet Neurol 16, 701-711.
57. Kohler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., Campbell, J., et al. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res 42, D966-974.
58. Yang, H., Robinson, P.N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods 12, 841-843.
59. Zemojtel, T., Kohler, S., Mackenroth, L., Jager, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., et al. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med 6, 252ra123.
60. Dubois, P.C., Trynka, G., Franke, L., Hunt, K.A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G.A., Adany, R., Aromaa, A., et al. (2010). Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 42, 295-302.
61. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet 47, 979-986.
62. Cordell, H.J., Han, Y., Mells, G.F., Li, Y., Hirschfield, G.M., Greene, C.S., Xie, G., Juran, B.D., Zhu, D., Qian, D.C., et al. (2015). International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. Nat Commun 6, 8019.

63. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376-381.
64. International Multiple Sclerosis Genetics, C., Wellcome Trust Case Control, C., Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z., et al. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476, 214-219.
65. Consortium., A.S.D.W.G.o.t.P.G. (2015). PGC-ASD summary statistics from a meta-analysis of 5,305 ASD-diagnosed cases and 5,305 pseudocontrols of European descent (based on similarity to CEPH reference genotypes). In. (http://www.med.unc.edu/pgc/results-anddownloads).
66. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature 492, 369-375.
67. Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. Nature 480, 201-208.
68. Bentham, J., Morris, D.L., Graham, D.S.C., Pinder, C.L., Tombleson, P., Behrens, T.W., Martin, J., Fairfax, B.P., Knight, J.C., Chen, L., et al. (2015). Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet 47, 1457-1464.
69. Horikoshi, M., Beaumont, R.N., Day, F.R., Warrington, N.M., Kooijman, M.N., Fernandez-Tajes, J., Feenstra, B., van Zuydam, N.R., Gaulton, K.J., Grarup, N., et al. (2016). Genome-wide associations for birth weight and correlations with adult disease. Nature 538, 248-252.
70. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46, 1173-1186.
71. Zheng, H.F., Forgetta, V., Hsu, Y.H., Estrada, K., Rosello-Diez, A., Leo, P.J., Dahia, C.L., Park-Min, K.H., Tobias, J.H., Kooperberg, C., et al. (2015). Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 526, 112-117.
72. Kottgen, A., Albrecht, E., Teumer, A., Vitart, V., Krumsiek, J., Hundertmark, C., Pistis, G., Ruggiero, D., O'Seaghdha, C.M., Haller, T., et al. (2013). Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. Nat Genet 45, 145-154.
73. Nikpay, M., Goel, A., Won, H.H., Hall, L.M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C.P., Hopewell, J.C., et al. (2015). A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet 47, 1121-1130.
74. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. Nat Genet 45, 1274-1283.
75. Soranzo, N., Sanna, S., Wheeler, E., Gieger, C., Radke, D., Dupuis, J., Bouatia-Naji, N., Langenberg, C., Prokopenko, I., Stolerman, E., et al. (2010). Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways. Diabetes 59, 3229-3239.
76. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segre, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al. (2012). Large-scale

association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44, 981-990.
77. Fritsche, L.G., Igl, W., Bailey, J.N., Grassmann, F., Sengupta, S., Bragg-Gresham, J.L., Burdon, K.P., Hebbring, S.J., Wen, C., Gorski, M., et al. (2016). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat Genet 48, 134-143.
78. Perry, J.R., Day, F., Elks, C.E., Sulem, P., Thompson, D.J., Ferreira, T., He, C., Chasman, D.I., Esko, T., Thorleifsson, G., et al. (2014). Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature 514, 92-97.
79. Day, F.R., Ruth, K.S., Thompson, D.J., Lunetta, K.L., Pervjakova, N., Chasman, D.I., Stolk, L., Finucane, H.K., Sulem, P., Bulik-Sullivan, B., et al. (2015). Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. Nat Genet 47, 1294-1303.
80. Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., Gloyn, A.L., et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42, 105-116.
81. Teumer, A., Tin, A., Sorice, R., Gorski, M., Yeo, N.C., Chu, A.Y., Li, M., Li, Y., Mijatovic, V., Ko, Y.A., et al. (2016). Genome-wide Association Studies Identify Genetic Loci Associated With Albuminuria in Diabetes. Diabetes 65, 803-817.
82. Saxena, R., Hivert, M.F., Langenberg, C., Tanaka, T., Pankow, J.S., Vollenweider, P., Lyssenko, V., Bouatia-Naji, N., Dupuis, J., Jackson, A.U., et al. (2010). Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nat Genet 42, 142-148.
83. Onengut-Gumuscu, S., Chen, W.M., Burren, O., Cooper, N.J., Quinlan, A.R., Mychaleckyj, J.C., Farber, E., Bonnie, J.K., Szpak, M., Schofield, E., et al. (2015). Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat Genet 47, 381-386.
84. Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 45, 1452-1458.
85. Otowa, T., Hek, K., Lee, M., Byrne, E.M., Mirza, S.S., Nivard, M.G., Bigdeli, T., Aggen, S.H., Adkins, D., Wolen, A., et al. (2016). Meta-analysis of genome-wide association studies of anxiety disorders. Mol Psychiatry 21, 1391-1399.
86. Major Depressive Disorder Working Group of the Psychiatric, G.C., Ripke, S., Wray, N.R., Lewis, C.M., Hamilton, S.P., Weissman, M.M., Breen, G., Byrne, E.M., Blackwood, D.H., Boomsma, D.I., et al. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. Mol Psychiatry 18, 497-511.
87. Okbay, A., Baselmans, B.M., De Neve, J.E., Turley, P., Nivard, M.G., Fontana, M.A., Meddens, S.F., Linner, R.K., Rietveld, C.A., Derringer, J., et al. (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. Nat Genet 48, 624-633.
88. Psychiatric, G.C.B.D.W.G. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet 43, 977-983.
89. Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421-427.
90. Pattaro, C., Teumer, A., Gorski, M., Chu, A.Y., Li, M., Mijatovic, V., Garnaas, M., Tin, A., Sorice, R., Li, Y., et al. (2016). Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. Nat Commun 7, 10023.

91. Eppinga, R.N., Hagemeijer, Y., Burgess, S., Hinds, D.A., Stefansson, K., Gudbjartsson, D.F., van Veldhuisen, D.J., Munroe, P.B., Verweij, N., and van der Harst, P. (2016). Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. Nat Genet 48, 1557-1563.
92. Barban, N., Jansen, R., de Vlaming, R., Vaez, A., Mandemakers, J.J., Tropf, F.C., Shen, X., Wilson, J.F., Chasman, D.I., Nolte, I.M., et al. (2016). Genome-wide analysis identifies 12 loci influencing human reproductive behavior. Nat Genet 48, 1462-1472.
93. Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science 340, 1467-1471.
94. Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.B., Emilsson, V., Meddens, S.F., et al. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature 533, 539-542.
95. Lu, Y., Day, F.R., Gustafsson, S., Buchkovich, M.L., Na, J., Bataille, V., Cousminer, D.L., Dastani, Z., Drong, A.W., Esko, T., et al. (2016). New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. Nat Commun 7, 10495.
96. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature 518, 197-206.
97. Felix, J.F., Bradfield, J.P., Monnereau, C., van der Valk, R.J., Stergiakouli, E., Chesi, A., Gaillard, R., Feenstra, B., Thiering, E., Kreiner-Moller, E., et al. (2016). Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. Hum Mol Genet 25, 389-403.
98. Kilpelainen, T.O., Carli, J.F., Skowronski, A.A., Sun, Q., Kriebel, J., Feitosa, M.F., Hedman, A.K., Drong, A.W., Hayes, J.E., Zhao, J., et al. (2016). Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels. Nat Commun 7, 10494.
99. Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Magi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E., et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. Nature 518, 187-196.

Chapter 3: The regulatory landscape of ovarian surface epithelial cells in relation to gene expression and epithelial ovarian cancer risk

3.1 Introduction

Epithelial ovarian cancer (EOC) is a rare but highly lethal malignancy, expected to cause ~14,000 deaths in the US in 2020[1]. EOC is a heterogeneous disease with a major heritable component[2]. Apart from highly penetrant germline mutations in genes such as BRCA1, BRCA2, and mismatch repair genes, the remaining heritability of EOC has been estimated at 5.6%, ranging from 3.2% to 8.8% across EOC subtypes[3]. Genome wide association studies (GWAS), primarily in European populations, have established 39 independent genomic risk regions for EOC[4], which account for approximately 40% of this heritability[3]. However, interpretation of these common risk variants remains challenging because they fall predominantly in the noncoding genome; this suggests a regulatory role toward target susceptibility genes, possibly through epigenomic and chromatin conformation mechanisms[2]. Though recent studies have provided insight into the mechanisms of EOC risk with respect to susceptibility genes[4], the regulatory landscape in ovarian surface epithelial cells (OSEC), a primary EOC precursor cell type, remains largely unexplored. Previous studies show that both expression profiles and regulatory features such as histone-3-lysine-27-acetylation (H3K27ac) in OSEC are implicated in the cellular origins of EOC[5,6]; however, little is known about how allelic variation affects gene expression, the epigenomic landscape, and gene regulation in OSEC. Understanding the specific regulatory pathways in OSEC is crucial for interpreting mechanisms of EOC risk arising in this cell type.

This chapter is the product of a collaboration between co-authors named in Chapter 1, and is a draft of a manuscript in preparation. In this work, we examine the largest ovarian cancer precursor cell dataset to date, comprising H3K27ac ChIP-sequencing, RNA-sequencing, and GWAS genotyping data in primary normal ovarian surface epithelial cells (OSEC) from 105

individuals. We characterize the enhancer landscape of OSEC, explore the genomic control of H3K27ac and gene expression in OSEC, and integrate EOC risk to identify putative tissue-specific EOC pathways. We first identify 62,106 regions of active chromatin in OSEC; in aggregate, we find these regions to be enriched for epithelial ovarian cancer GWAS risk. Second, we identify 947 putative gene-enhancer pairs through correlation of gene expression and H3K27ac peak intensity across our population. Third, we identify 20,693 expression QTLs (eQTLs) for 580 eGenes and 143 H3K27ac QTLs (hQTLs) for 30 peaks. Of the 29 known risk regions for ovarian cancer, we find 10 containing eQTL associations for 13 eGenes; one of these, *SLC12A7*, was also linked to a putative enhancer in our earlier analysis. In addition to providing insight into the regulatory landscape of healthy ovarian surface epithelial cells and identifying putative tissue-specific EOC pathways, these results can be contrasted to the regulatory landscape of ovarian cancer to pinpoint the major changes in gene regulation and pathway activation that occur during neoplastic transformation.

3.2 Results

We obtained high-throughput RNA-sequencing and high-density genotyping in a set of primary normal ovarian surface epithelial cells (OSEC) from 121 individuals, and H3K27ac ChIP-sequencing in a subset of these OSEC (54 individuals; Figure 1). After quality control, outlier filtering, normalization, variance filtering, and consensus peak region identification (see Methods), this dataset contained expression measurements of 19,923 genes and genotypes of approximately 6 million SNPs single nucleotide polymorphisms (SNPs) across 105 individuals, and H3K27ac peak calls in 62,106 regions across 52 of these individuals. The OSEC donors in our study were primarily of European descent. We explore the hypothesis that genetic variation drives changes in chromatin activity and gene expression, which ultimately affects ovarian cancer risk (Figure 2).

Population-based profiling of active chromatin in OSEC

H3K27ac marks active enhancers and promoters[7]. We first examined the H3K27ac ChIP-sequencing metrics to determine sequencing quality for population-level comparison. The H3K27ac ChIP-sequencing read depth per sample varied from approximately 70 million reads to 157 million reads (Table 1); after outlier filtering, retained samples had a minimum of approximately 10 million uniquely mapped reads (Figure 3). To assess whether sequencing quality was driving peak discovery within our samples, we computed correlations between number of peaks called per sample and total read depth per sample, and between number of peaks called per sample and unique map rate per sample. Neither correlation was significant (r^2 = 0.06 and 0.01 respectively; Figure 4a and Figure 4b respectively.)

To explore variation in the OSEC enhancer and promoter landscape, we then identified 62,106 H3K27ac peak regions ("peaks") across 52 individuals and compared peak calls within these regions across individuals. This was performed by taking the union of all peak boundaries across samples and merging any regions within 147 base pairs of each other (Figure 5; see Methods for more detail.) The median peak length was 3.4kb, ranging from 280bp to 571kb (Figure 6). Each peak was called in an average of 22 individuals (Figure 7); on average, Y peaks were called per individual. For each of the 62,106 peaks across each of the 52 individuals, peak scores were constructed by dividing the average number of reads over only the covered bases within the peak boundaries by the total number of uniquely mapped reads for each individual. As expected, the variance of scores show that most peaks are common (with low variance) and a few peaks show dramatically increased variance in peak scores (Figure 8). The attrition in new peaks identified per sample (Figure 9) indicates that our peak calls comprehensively reflect the active chromatin landscape in OSEC.

Enrichment of ovarian cancer GWAS risk in H3K27ac peak regions

To quantify the relevance of H3K27ac peak regions to ovarian cancer risk, we performed an enrichment analysis using a genomic annotation corresponding to all 62,106 H3K27ac consensus peak regions and the GWAS summary statistics from 6 subtypes of epithelial ovarian cancer (see Methods). We identified significant enrichment of GWAS signal in our peak regions for the high-grade serous subtype (2.59 fold enrichment, s.e. 0.35; $p = 3.29 \times 10^{-4}$), and for the meta-analysis of all non-mucinous histotypes (2.71 fold enrichment, s.e. 0.31; $p = 4.62 \times 10^{-5}$). These results are consistent with our hypothesis that H3K27ac in OSEC are relevant for the development of ovarian cancer risk.

Global maps of gene regulation in ovarian surface epithelial cells

Although many genes have been identified as important in the development of ovarian cancer, their regulatory landscape in OSEC remains unknown; specifically, the enhancers regulating each gene have not been clearly mapped. This is a critical step in understanding the pathway from genetic variation to the phenotype of ovarian cancer development. Here, we statistically linked enhancers and genes by computing a correlation between H3K27ac peak intensity and gene expression across the 52 individuals with both H3K27ac ChIP-seq and RNA-seq; we treat H3K27ac peaks as markers of enhancers in OSEC. With this sample size, and the total numbers of genes (19,923) and peaks (61,206), we are best powered to identify proximal enhancers (vs. distal enhancers), so we limited correlations to pairs of genes and H3K27ac peaks within 1 Mb of each other. Of 702,057 peak-gene pairs tested (see Methods), we identified 923 peak-gene pairs with significant correlation. This included 729 unique genes and 907 unique peaks; on average, we identified 1 significantly correlated gene per peak, and 1-2 significantly correlated peaks per gene (Figure 10a and Figure 10b). The average distance between peak and gene among significantly correlated pairs was 492 kb, with Pearson r less

than -0.4 and greater than 0.4 (Figure 11). For the genes in each of these pairs, we evaluated Gene Ontology (GO) enrichments, which revealed specific enrichment among cellular biosynthetic processes and detection of chemical stimuli, especially related to sensory smell perception (Table 2).

As the 923 peak-gene pairs identified represent putative gene-enhancer pairs, intersecting this map of gene regulation with known ovarian cancer risk in the genome could prioritize genes and pathways relevant to the development of ovarian cancer. Across the 29 identified ovarian cancer genomic risk regions, we identify 20 putative gene-enhancer pairs across 19 unique genes, 20 unique peaks, and 13 unique genomic risk regions (Table 3). We identify one gene, *SLC12A7*, in an ovarian cancer risk region on chromosome 5, to be correlated with 2 putative enhancers. To identify further evidence of any gene's relevance to ovarian cancer, we compared this set of 19 genes to the 23 genes identified from a recent transcriptome-wide association study in ovarian cancer[4]. However, none of these genes had been specifically identified as associated with ovarian cancer in the TWAS or other reported literature as far as we can tell.

Cis-eQTLs identified for 580 eGenes in OSEC

An important step in understanding the regulatory landscape in a cell type is to identify the genetic drivers of gene expression and overall ovarian cancer risk. We start by identifying expression QTLs (eQTLs) across 105 individuals, for expression measurements of 13,334 genes, and genotype measurements of close to 6 million single nucleotide polymorphisms (SNPs); these were selected based on outlier and variance filtering (see Methods); PEER factors and genotype PCs were regressed out to correct for covariates and ancestry confounding. Permutations were used to disrupt linkage disequilibrium (LD) between genotypes and phenotypes, and control for multiple testing using a 5% FDR.

In total, we identified 20,693 significant eQTL SNP-gene pairs, corresponding to 580 unique eGenes and 20,402 unique SNPs. As expected, nominal eQTL signal was enriched near the transcription start site of eGenes (Figure 12). To identify any biological processes or pathways enriched within these 580 genes, we performed another GO enrichment analysis (Table 4). This identified enrichment of protein targeting processes, as well as detection of chemical stimuli also particularly involved in sensory smell perception. In addition, one of the eGenes, *STK11*, has been linked to ovarian cancer through Peutz-Jaeger syndrome[8].

As before, intersecting the eQTL map with known ovarian cancer risk in the genome could prioritize genes and genetic variation relevant to the development of ovarian cancer. Across the 29 identified ovarian cancer genomic risk regions, we identify 362 significant eQTL SNP-eGene pairs representing 12 unique eGenes, 289 unique peaks, and 10 unique genomic risk regions (Table 5). To identify further evidence of any gene's relevance to ovarian cancer, as before, we compared this set of 12 genes to the 23 genes identified from the recent ovarian cancer TWAS[4]; none of these genes were implicated. However, one of the 12 eGenes, *SLC12A7*, was also identified in our study to be correlated with 2 putative enhancers. These findings are consistent with previous studies of eQTLs in ovarian cancer tissue types[4].

Few cis-hQTLs identified in OSEC

We next attempted to identify H3K27ac chromatin QTLs (hQTLs) in the 52 individuals with H3K27ac ChIP-seq and genotyping. We started with the peak intensities across all 62,106 consensus peaks in each individual and the 6 million SNP genotypes; however, we only identified 101 hQTL SNP-peak pairs, corresponding to 20 unique peaks and 101 unique SNPs. There was no noticeable enrichment of hQTL signal near the center of peaks, even at reduced significance thresholds (Figure 13). Furthermore, we found no hQTL associations in any of the

29 ovarian cancer genomic risk regions. These results did not change significantly when using binarized peak phenotypes or with differing numbers of PEER factors corrected (see Methods).

No statistical colocalization of eQTL and GWAS associations

Last, we attempted to statistically quantify whether any eQTL signal in ovarian cancer risk regions was driven by the same variants associated with ovarian cancer through GWAS. In each of the 29 risk regions, we performed a colocalization test between each gene's nominal eQTL associations and GWAS risk for each of the 6 subtypes of ovarian cancer. Using a PP4 threshold of 0.8, signifying a posterior probability that both eQTL and GWAS associations are driven by the same variant, we did not identify any significant colocalization in any of the GWAS risk regions for any of the ovarian cancer subtypes. We could not perform this analysis between hQTL and GWAS associations, as no significant hQTL associations were identified in any ovarian cancer risk regions.

3.3 Methods

Generation of primary normal OSEC and selection of individuals for our study

Obtention of primary normal OSEC was as previously described in [4] : "OSECs were collected from histologically normal ovaries and fallopian tubes removed from women diagnosed with ovarian, uterine or cervical cancer. Short-term cultures were established[9,10]. OSECs were collected using a Cytobrush and cultured in NOSE-CM media containing 15% fetal bovine serum (FBS; HyClone, Fisher Scientific), 34 µg ml−1 bovine pituitary extract, 10 ng ml−1 epidermal growth factor (Thermo Fisher Scientific), 5 µg ml−1 insulin and 500 ng ml−1 hydrocortisone (Sigma-Aldrich)." These samples were obtained as part of a larger study; for our analyses, we selected all individuals where matched IDs could be identified across genotype

data and RNA-sequencing data (described below). This resulted in a set of 121 individuals. All analyses described in our study relate to these individuals only.

Genotyping and data processing

High-density imputed genotypes of primary normal OSEC from approximately 272 individuals were obtained were obtained from our collaborators Paul Pharoah, Simon Gayther, and Kate Lawrenson as part of a larger study. We restricted our analysis to the 121 individuals mentioned above, where genotype IDs could be matched to RNA-sequencing IDs (described below). Given the imputed genotypes (an output of the IMPUTE2 software) from our collaborators, we first converted imputed genotype probabilities into dosages in VCF format using custom scripts. We then restricted all analyses to biallelic SNPs with minor allele frequency (MAF) > 0.05, no missingness, and Hardy-Weinberg equilibrium > 0.0001. These restrictions resulted in measurements of 6 million SNPs across 121 individuals.

RNA sequencing and data processing

The RNA sequencing used in this study was previously described in [4]: "At approximately 80% confluency, cells were lysed using the QIAzol lysis reagent and RNA extracted using the RNeasy Mini Kit (both QIAGEN). RNA-seq was performed by the University of Southern California Epigenome Core Facility using 50 bp single-end reads." RNA-seq gene-level quantification, GC normalization, and batch correction were performed by our collaborators at Cedars-Sinai; we ultimately obtained TPM counts per gene per individual for 19,923 genes from these collaborators and restricted to the 121 individuals whose RNA-sequencing ID could be linked to a genotype ID.

Using the TPM counts per gene per individual, we restricted our analysis to genes with >0.1 TPM in at least 20% of individuals (resulting N = 13,334 genes). We computed and

regressed out 15 PEER factors along with 3 principal components (PCs) computed from the corresponding sample genotypes to obtain residuals, and normalized the residuals across individuals using a rank inverse normal transformation.

H3K27ac ChIP sequencing and data processing

H3K27ac ChIP-sequencing was performed by our collaborators at Cedars-Sinai on a subset of the 121 individuals in our study; this included 54 individuals, plus 4 additional replicate individuals, for 58 experiments total. The libraries were then sent to the UCLA Sequencing Core (care of Giovanni Coppola and Yue Qin), who called narrow and broad H3K27ac peaks for each individual using HOMER. We then obtained the output peak calls and bigWig files from the UCLA Sequencing core.

We first examined the sequencing statistics and removed outlier samples with coverage < X; this threshold was determined by preliminarily identifying population-level peak regions and performing a principal components analysis to visualize outliers. After outlier and replicate removal, we re-identified population-level peak regions by taking the union of peak regions across all remaining individuals and merging all peak regions within 147bp of another. This resulted in 62,106 peak regions across 52 individuals. All further analyses using H3K27ac ChIP-sequencing were restricted to these 52 individuals. We then constructed 2 phenotype matrices of individuals vs. population-level peak regions: one with continuous values and one with binarized values. Each cell of the binarized phenotype matrix contained either 1 or 0 depending on whether any peak was originally called in that individual in that region; each cell of the continuous phenotype matrix contained, per individual per region, the average number of reads over only the covered bases in the region divided by the total number of uniquely mapped reads for each individual. Across both matrices, we computed and regressed out between 2 and 10 PEER factors along with 3 genotype PCs and batch covariates to obtain residuals. Ultimately,

the binarized phenotype matrix with 2 PEER factors, 3 genotype PCs, and batch covariates regressed out was used for hQTL calling (see "Determining and troubleshooting parameters" section below). This data pre-processing follows the GTEx pipeline steps, modified for the different data type (ChIP-seq vs. RNA-seq) and for the reduced sample size. The inclusion of 3 genotype PCs follows the process used by the GEUVADIS study and controls for genetic background and ancestry in our study.

Obtention of genome-wide association summary statistics for epithelial ovarian cancer

Summary statistics were obtained from our collaborators Paul Pharoah, Simon Gayther, and Kate Lawrenson from the latest, largest genome-wide association study of ovarian cancer (N > 61,000 individuals), published in [11]. These summary statistics contained separate associations analyzed for 5 histotypes of epithelial ovarian cancer: endometrioid, clear cell, mucinous ("`mucinous_all`"), low-grade serious ("`ser_lg_lmp`"), and high-grade serous ("`serous_hg_extra`"). Additionally, a meta-analysis of associations for all non-mucinous histotypes was included ("`all_non_mucinous`").

Quantifying enrichment of SNP heritability

We used stratified LD score regression (s-LDSC)[12] to estimate the enrichment of SNP heritability of the 6 ovarian cancer subtypes within a single genomic annotation containing the 62,106 H3K27ac population-level peak regions in OSEC defined above. The annotation value for SNP i is defined as $a_i = 1$ if SNP i is within an H3K27ac peak region and ai = 0 otherwise. We computed LD scores within 1 cM blocks with default parameters and LD estimated from the European individuals in the 1000 Genomes Phase 3 reference panel. For each of the 6 ovarian cancer subtypes, we ran s-LDSC by using the recommended "baseline model"[12] as covariates

in the regression for a total of 53 annotations per run (52 “baseline” annotations plus the OSEC H3K27ac annotation).

Identifying putative gene-enhancer pairs

To identify putative gene-enhancer pairs, we computed correlations across the 52 individuals with both H3K27ac ChIP-sequencing and RNA-sequencing. First, across all 19,923 genes and 62,106 peaks, we identified all peak-gene pairs within 1 Mb of each other. This resulted in 702,057 pairs. For each pair, we computed a Pearson correlation of gene expression values as described above and continuous peak phenotype values as described above across the 52 individuals. Significance was assessed at a Bonferroni-corrected threshold per gene, correcting for the number of peak-gene pairs tested for each gene (mean = 54 tests per gene, resulting in a Bonferroni-corrected threshold of 0.05/54 corresponding to $p < 0.0009$).

Genome-wide cis-QTL calling

To identify cis-hQTLs, we used the pipeline previously described by the eQTLgen consortium as well. Before calling hQTLs, we filter genotypes by MAF > 5% to exclude rare variants, require 0% missingness to maintain sample size across each QTL test, and restrict to biallelic variants. We used the permutation setting to disrupt LD between genotypes and phenotypes, and control for multiple testing using a 10% FDR. All SNP-peak pairs were tested for association where the center position of the peak falls within 2 kb of the SNP. A false discovery rate (FDR) was determined based on 10 permutations of sample labels in either the genotype or expression dataset; cis-hQTLs with FDR > 0.10 were considered significant. These parameters follow [13].

To identify cis-eQTLs, we similarly used the pipeline previously described by the eQTLgen consortium; briefly, all SNP-gene pairs were tested for association where the center

position of the gene falls within 250 kb of the SNP. A false discovery rate (FDR) was determined based on 10 permutations of sample labels in either the genotype or expression dataset; cis-eQTLs with FDR > 0.05 were considered significant. These pre-processing and QTL-calling steps followed the GTEx and GEUVADIS pipelines.

Determining and troubleshooting parameters for genome-wide cis-hQTL calling

To determine the optimal number of PEER factors to correct for, we performed a cis-hQTL run correcting one PEER factor at a time from 2 to 10 PEER factors, with 1 permutation per run. Maximal QTLs were identified correcting for 2 PEER factors, although results were similar across all PEER trials. Similarly, to determine whether to use continuous or binarized phenotypes, we performed cis-hQTL using both phenotype matrices; maximal QTLs were identified using binarized phenotypes, although results were similar across both trials. To determine the optimal QTL-calling pipeline, we additionally performed a cis-hQTL analysis using Matrix eQTL, restricting to SNPs within 2kb of each peak. 3 genotype PCs were included as covariates, along with age, batch, and 4 technical covariates. Cis-hQTL significance was assessed at FDR < 0.1. Equivalent results were obtained as compared to using the eQTLgen pipeline.

Colocalization analysis between eQTL associations and GWAS associations

Colocalization analysis was performed using the Coloc R package[14] using nominal eQTL association summary statistics and nominal GWAS summary statistics for each of the 6 ovarian cancer subtypes. Colocalization tests were performed between GWAS signal and eQTL associations for eGenes in each of the 29 genomic risk regions identified by [11]. A minimum of 2 SNPs were required to be present in each dataset for each colocalization test; significance was assessed at a threshold of PP4 > 0.8.

3.4 Discussion

In this study, we integrated gene expression, H3K27ac ChIP-sequencing, and genotyping data in OSEC from a population of individuals with the largest GWAS dataset available for epithelial ovarian cancer to characterize the regulatory landscape of this important ovarian cancer precursor cell type. We were able to compare H3K27ac ChIP-sequencing across 52 individuals to identify common and variable active chromatin regions, demonstrated enrichment of multiple ovarian cancer subtype GWAS signal within the population-level active chromatin regions, identified pairs of genes and putative enhancers to understand the global gene regulation landscape in OSEC, and identified eQTLs for 580 eGenes in OSEC. We identtified *SLC12A7*, which has been previously linked to ovarian and cervical cancer cell invasion[15], to have 2 putative enhancers and a significant eQTL in OSEC, as well as localizing in an ovarian cancer risk region, suggesting its potential as a candidate gene for follow-up functional analyses. None of the genes identified by our analyses have previously been identified as specifically linked to ovarian cancer at the population level; thus, we offer a selection of novel genes for future studies. We also identified 102 hQTLs for 20 ePeaks, although we were unable to identify any evidence of colocalization of causal signal between GWAS associations for any subtype and eQTL associations or hQTL associations. Taken together, these findings indicate that the regulatory landscape of OSEC remains partially understood, and increased sample size may offer new insight into the specific pathways and mechanisms at play in ovarian cancer risk.

We observed a striking overlap between the gene ontology enrichment findings for genes with putative enhancers and eGenes; in particular, detection of chemical stimuli and sensory perception of smell were biological processes found to be significantly enriched in both sets of genes. These findings suggest a potential benefit of further exploration into the links between OSEC and sensory perception, particularly in the case of smell. In addition, a majority of eGenes identified in OSEC that also fell within ovarian cancer risk regions have been

previously reported in the context of cancer. Specifically, *FES* has been identified as an oncogene in acute promyelocytic leukemia; *GJC1* has been associated with promoter hypermethylation-driven silencing in colorectal cancer[16]; *INA* has been found to have high RNA expression in urothelial & testicular cancer; *KCNAB1* is differentially expressed in ovarian cancer[17] and has been associated with ovarian cancer survival[18]; *NDUFS6* is a therapeutic target for HER2+ breast cancer[19] and has been linked to cervical cancer[20]; *NFE2L1* has been associated with cellular toxicity in breast cancer and showed similar effects as c-MYC overexpression[21]; *PGLS* is a driver gene in cervical non-keratinized squamous cell carcinoma[22]; *TMEM38A* is an unfavorable prognostic marker in ovarian cancer; and *TRIAP1* is a microRNA target in inhibiting ovarian cancer growth[23].

It remains likely that our analyses missed an unquantifiable proportion of true associations and gene-enhancer pairs as our sample sizes were limited to 52 H3K27ac ChIP-sequencing samples and 105 RNA-sequencing samples. It is possible for long-range enhancers to regulate genes, and our current approach will miss these. However, removing the distance limit between genes and H3K27ac peaks would likely result in the multiple testing correction being too stringent to observe signal at our sample size. Furthermore, it is possible that our sample size would be too small to observe any correlation signal even within a 1MB window. To address these limitations, future studies can restrict the number of genes tested to only those previously established to have relevance for ovarian cancer, such as those conferring hereditary risk like BRCA1 and BRCA2, or the susceptibility genes identified in the recent TWAS of ovarian cancer[4]. This would substantially reduce the multiple testing correction burden and allow for possible identification of enhancer-gene pairs. Second, we could instead compute the genetic covariance between the genetic variants associated with H3K27ac peaks and genes. Though this would not link pairs of genes and enhancers, it would provide genome-level evidence that

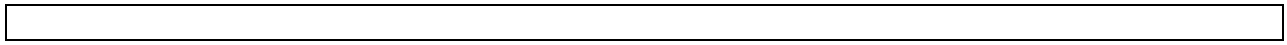
the genetic variants driving chromatin activity (as measured by H3K27ac peak intensity) similarly contribute to gene expression.

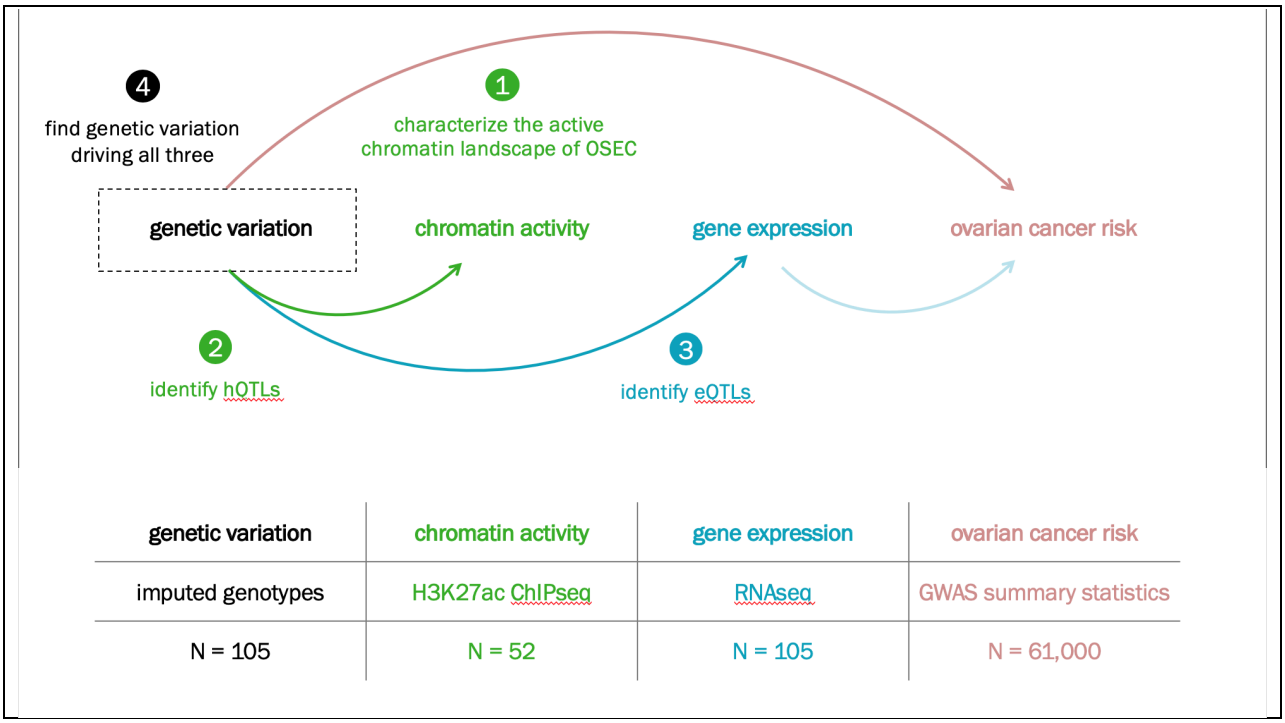
Similar studies with slightly larger sample sizes identified significantly more QTL associations[13,24], suggesting the value of larger sample sizes to identify hQTLs in particular. Furthermore, our samples demonstrated strong batch and covariate effects, making it difficult to correct for confounders while maximizing signal. It's also possible our analyses were limited due to overinclusion of covariates with excess missing data. Furthermore, although histone modifications such as H3K27ac reflect genetic control, calling hQTLs with a sample size of 52 is challenging regardless. Future studies may benefit from an alternate approach to link genetic variation with H3K27ac, by identifying instances of allelic imbalance; this would require restricting to peaks with SNPs (specifically heterozygous genotypes) contained within peak boundaries, and measuring whether an imbalance of ChIP-sequencing reads carry one allele vs. the other. Recent studies have shown the benefit of this approach in cases of low sample size.

Last, integration of more functional genomic data would provide significant value in this endeavor. In particular, Hi-C and H3K27ac HiChIP experiments which identify physical chromatin interactions would provide additional avenues to link genetic variation to enhancers, enhancers to genes, and genes to significant GWAS associations. In particular, identifying cases where GWAS associations interact with enhancers and/or gene promoters would provide direct evidence of molecular pathways which QTL analyses are not powered to detect.

3.5 Figures

Figure 3.1 Schematic of data and analyses to explore regulatory landscape of OSEC in relation to ovarian cancer risk.





genetic variation	chromatin activity	gene expression	ovarian cancer risk
imputed genotypes	H3K27ac ChIPseq	RNAseq	GWAS summary statistics
N = 105	N = 52	N = 105	N = 61,000

Figure 3.2 Regulatory hypothesis linking genetic variation, chromatin activity, gene expression, and inherited ovarian cancer risk.

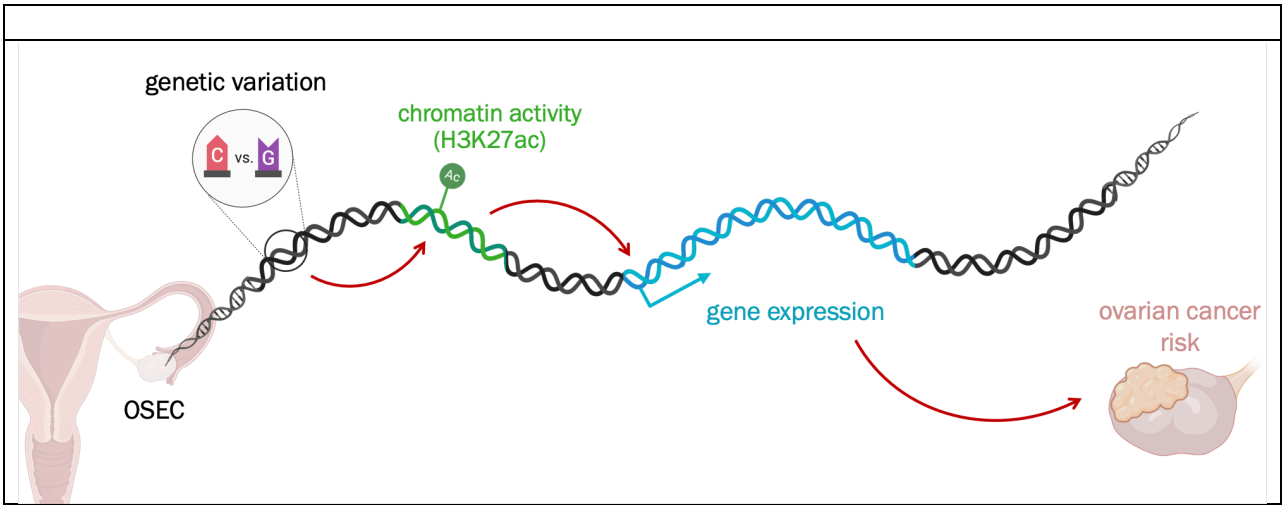


Figure 3.3 Mapped vs. uniquely mapped reads for H3K27ac ChIP-sequencing alignment in 52 retained samples.

This plot shows counts of mapped reads (light pink) vs uniquely mapped reads (dark pink) for H3K27ac ChIP-sequencing alignment across 52 retained samples. Sample IDs are sorted along the X axis from lowest number of mapped reads to highest; read counts are reflected on the Y axis.

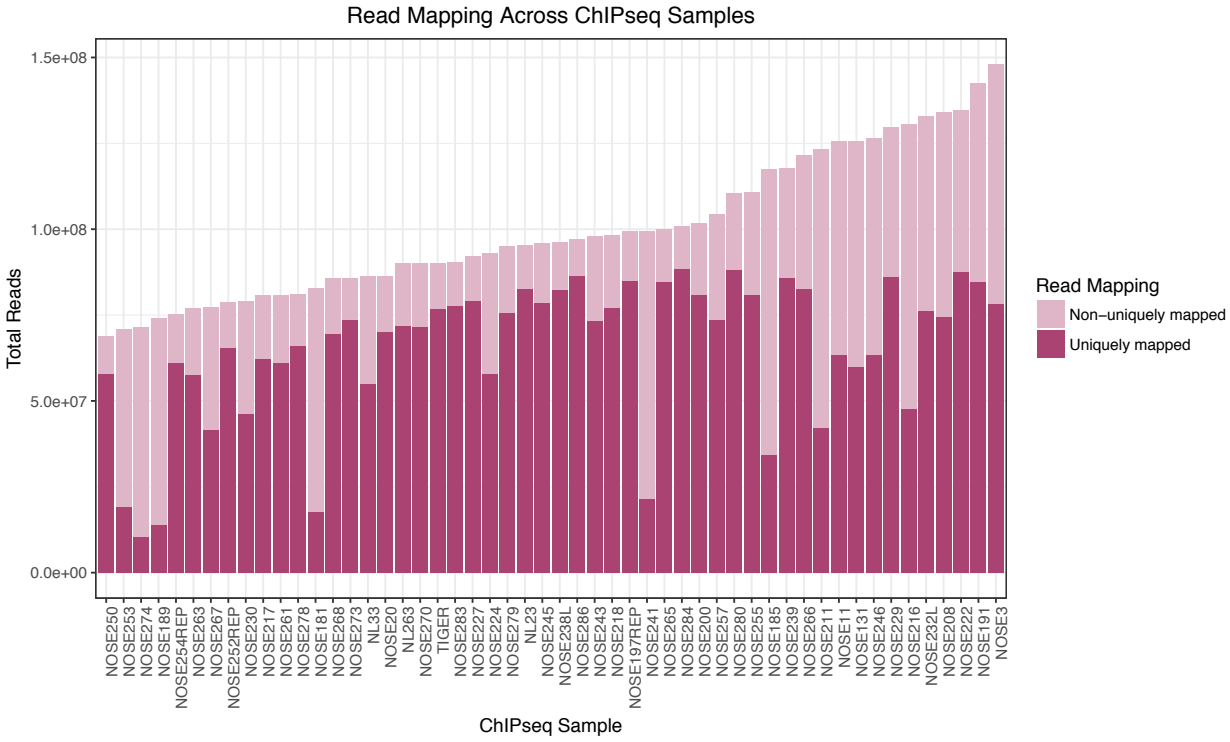


Figure 3.4 Read count and map rate are uncorrelated with number of peaks called per sample.

Shown here are correlations between number of peaks called per sample and (left) total read count, and (right) unique map rate. Each dot represents one sample; black line represents line of best fit.

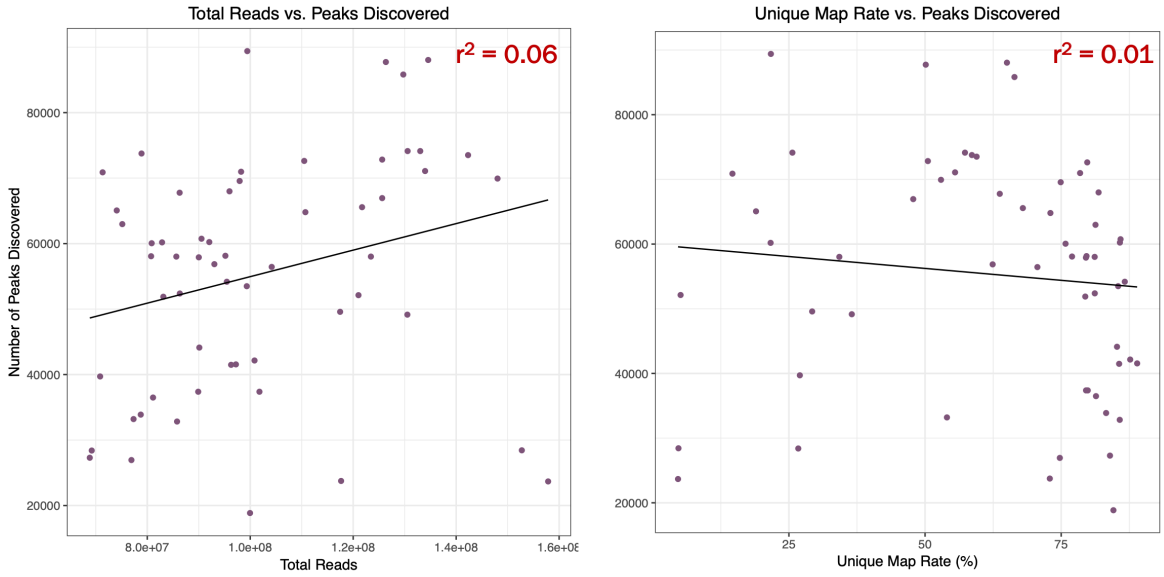


Figure 3.5 Diagram of of consensus peak region identification across 52 H3K27ac ChIP-sequencing samples.

Consensus peaks were identified by taking the union of peak regions across all remaining individuals and merging all peak regions within 147bp of another.

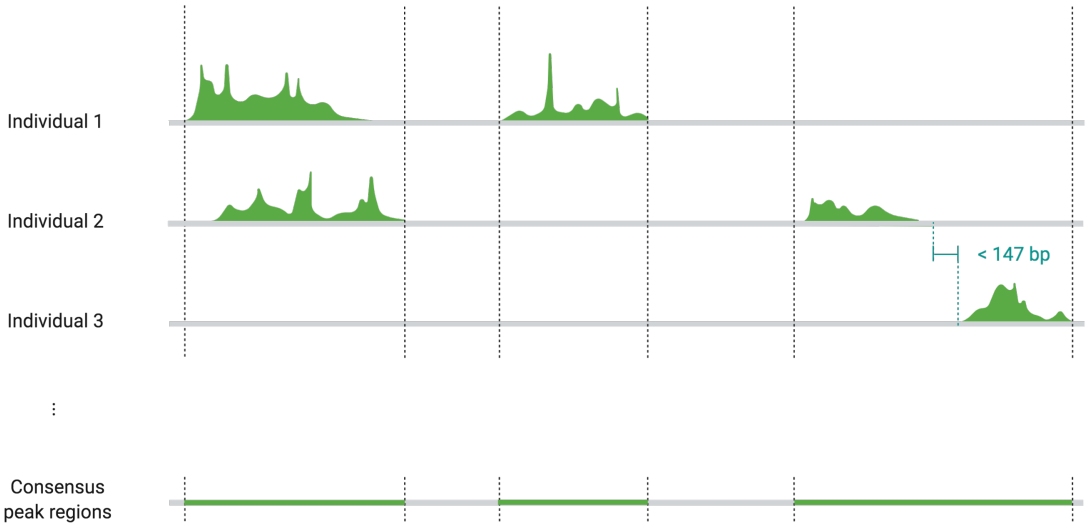


Figure 3.6 Histogram of H3K27ac consensus peak sizes (lengths).

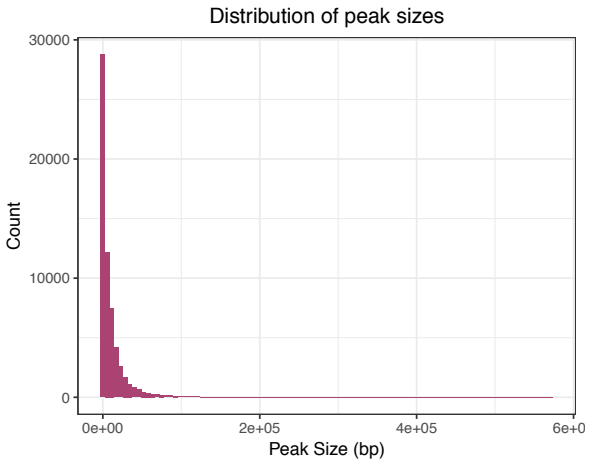


Figure 3.7 Peak frequency across individuals.

Shown here is the distribution of peaks across individuals; for each of the 62,106 consensus peak regions, we counted how many individuals had a peak originally called within that region. The median number of individuals identified with a particular peak was 22; the quartiles ranged from 0-3 individuals, 3-22 individuals, 22-51 individuals, and 51-52 individuals in order.

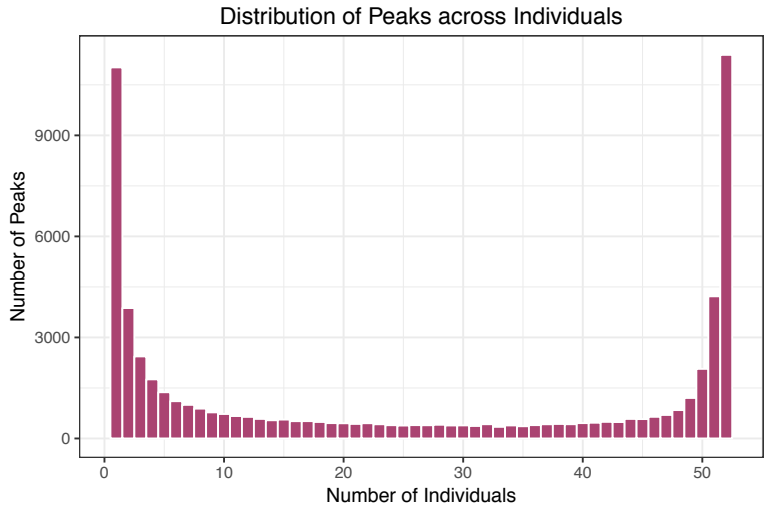


Figure 3.8 Variation of peak scores across individuals.

For each of the 62,106 peaks across each of the 52 individuals, a peak score was constructed by dividing the average number of reads over only the covered bases in the region by the total number of uniquely mapped reads for each individual. Shown here is the log variance of peak scores.

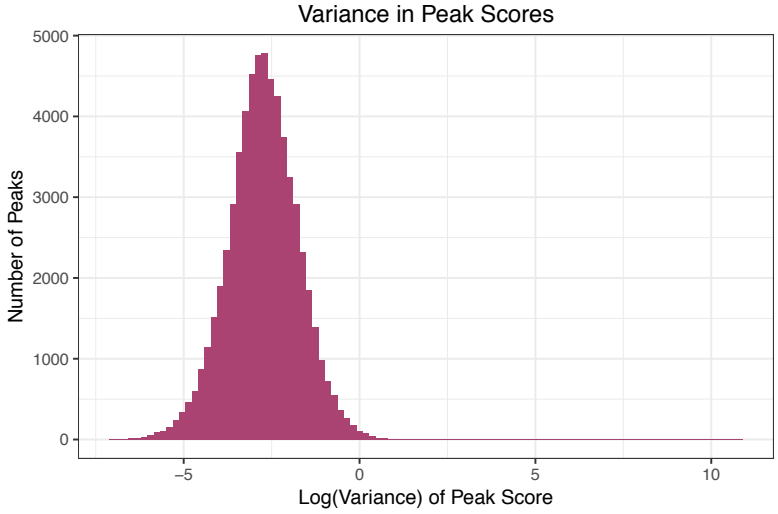


Figure 3.9 Attrition of new peaks discovered with each new individual considered.

Shown here are the number of peaks discovered in each sample, out of the total consensus 62,106 peaks. This was constructed by randomly selecting one individual at a time without replacement and identifiying how many of the 62,106 peak regions contained a peak originally called in that individual. With each new sample, only new consensus peak regions were considered.

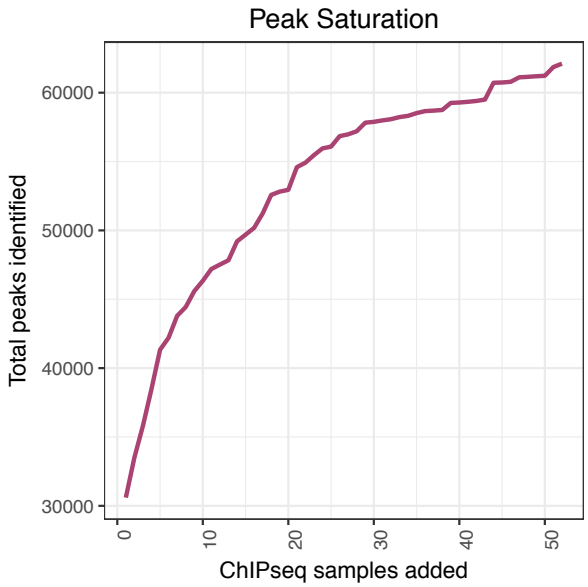


Figure 3.10 Histograms of significantly correlated peaks per gene and genes per peak.

Of the 923 significantly correlated pairs of peaks and genes, 729 unique genes and 907 unique peaks were identified. Shown here are the distributions of genes per peak (left) and peaks per gene (right) across the significant pairs.

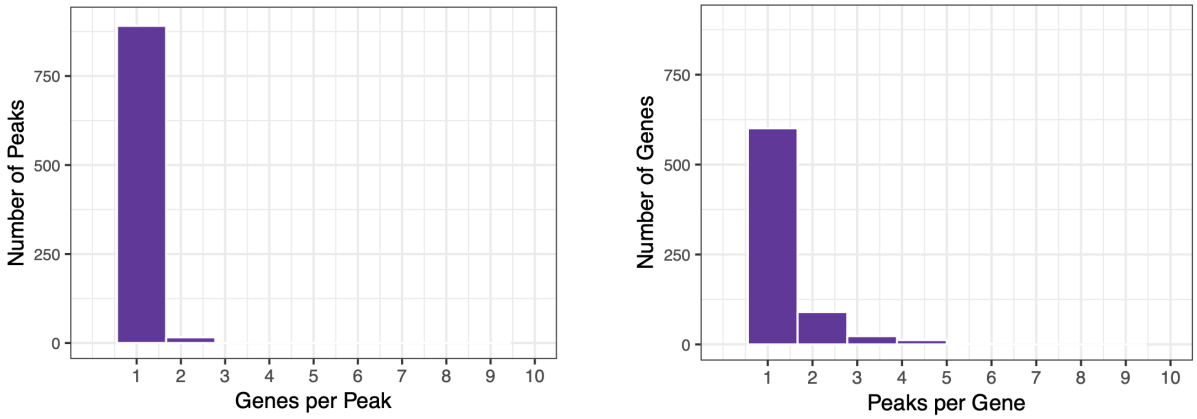


Figure 3.11 Comparison of Pearson r values across significant and non-significant peak-gene pairs tested.

Of the 702,057 peak-gene pairs tested, 923 were significantly correlated. Shown here are the Pearson r values across the significant and non-significant pairs.

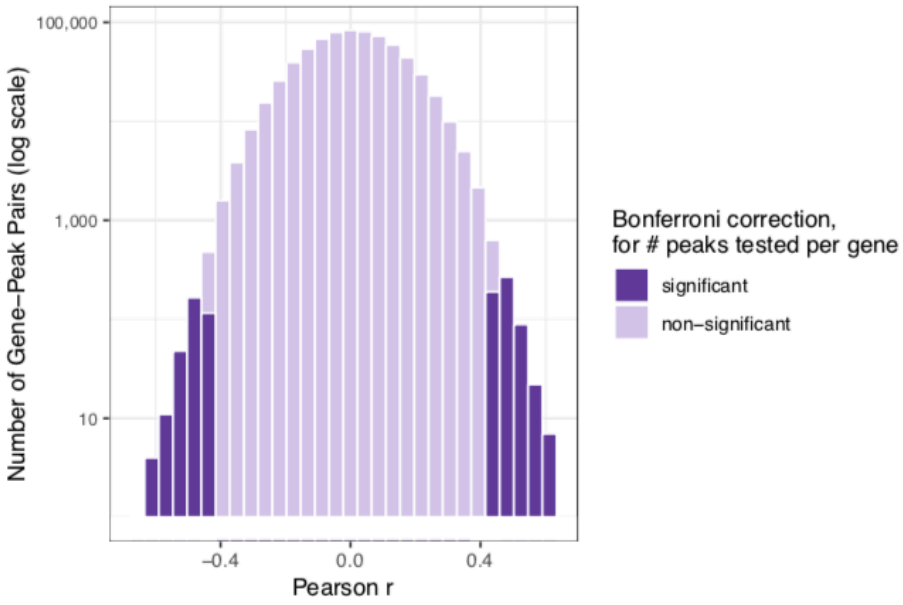


Figure 3.12 Enrichment of eQTL signal near transcription start site of eGenes.

For the top eQTL across each of the 580 eGenes, we show the distance to the eGene's transcription start site.

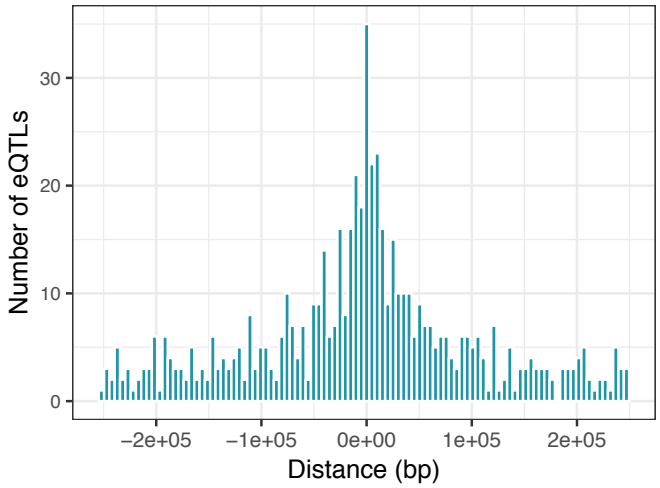
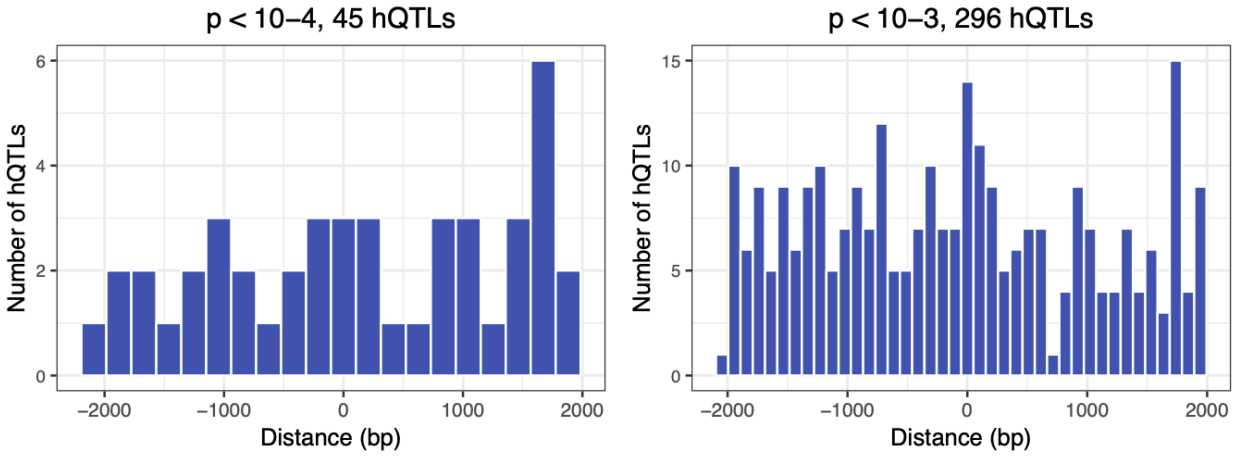


Figure 3.13 No enrichment of hQTL signal near centers of peaks.

For the top eQTL at various significance thresholds, we show the distance to the ePeaks's center.



3.6 Tables

Table 3.1 Sequencing and alignment statistics for 58 H3K27ac ChIP-sequencing samples.

sample	total reads	reads mapped	uniq mapped reads	mapped rate	uniq map rate	average length	mean quality
NL23	95457790	94463189	82746991	98.96%	86.68%	133	35.4
NL263	89897963	88520476	71809065	98.47%	79.88%	134	35.5
NL33	86289146	83920913	54985643	97.26%	63.72%	131	35.5
NL49	83122519	81105414	66013336	97.57%	79.42%	129	35.6
NOSE11	125631962	117761795	63474597	93.74%	50.52%	123	35.5
NOSE131	125632951	117332688	60093365	93.39%	47.83%	125	35.5
NOSE181	82873544	75061891	17936152	90.57%	21.64%	123	34.8
NOSE185	117461830	101596836	34382311	86.49%	29.27%	117	35.5
NOSE189	74070664	70291720	14043233	94.90%	18.96%	127	35.1
NOSE191	142318337	137012047	84640238	96.27%	59.47%	127	35.5
NOSE197	69213954	64138790	18496064	92.67%	26.72%	132	34.7
NOSE197REP	99328227	98292710	84889255	98.96%	85.46%	134	35.6
NOSE200	101791431	99865500	80985586	98.11%	79.56%	131	35.6
NOSE208	133969516	128541857	74360582	95.95%	55.51%	126	35.6
NOSE20	86325230	84693063	70070745	98.11%	81.17%	132	35.4
NOSE211	123439534	120325167	42289301	97.48%	34.26%	126	35.7
NOSE216	130523184	116206027	47704811	89.03%	36.55%	115	35.6
NOSE216REP	130573114	106304228	33480775	81.41%	25.64%	116	35.4
NOSE217	80744614	79740985	62168531	98.76%	76.99%	138	35.5
NOSE218	98227357	95686518	77075077	97.41%	78.47%	129	35.5
NOSE222	134570607	131745209	87523679	97.90%	65.04%	131	35.5
NOSE224	93032672	88281729	58057282	94.89%	62.41%	127	35.6
NOSE227	92048815	91037845	78957028	98.90%	85.78%	134	35.4
NOSE229	129734032	124953771	86164680	96.32%	66.42%	128	35.5
NOSE230	78885490	77059405	46223986	97.69%	58.60%	136	35.7
NOSE232L	133024026	126145278	76267502	94.83%	57.33%	127	35.4
NOSE238L	96288783	94256998	82459033	97.89%	85.64%	132	35.4
NOSE239	117651205	116023155	85777984	98.62%	72.91%	126	35.6
NOSE241	99428272	97454374	21564964	98.01%	21.69%	129	35.6
NOSE243	97943588	96073832	73370257	98.09%	74.91%	132	35.5
NOSE245	95972175	93914357	78574409	97.86%	81.87%	131	35.5
NOSE246	126347685	117834993	63336765	93.26%	50.13%	124	35.6
NOSE250	68834492	68252193	57790878	99.15%	83.96%	134	35.4
NOSE252	152755529	38160193	7216006	24.98%	4.72%	57	20.2
NOSE252REP	78738009	77755259	65546574	98.75%	83.25%	139	35.4
NOSE253	70844410	67053290	19132791	94.65%	27.01%	133	35.1
NOSE254	157867856	36182710	7326599	22.92%	4.64%	55	19.5
NOSE254REP	75148837	73890048	61093540	98.32%	81.30%	134	35.4
NOSE255	110716155	109059945	80842488	98.50%	73.02%	132	35.6
NOSE257	104172009	102282120	73576624	98.19%	70.63%	136	35.4
NOSE261	80854457	78987798	61284644	97.69%	75.80%	131	35.4
NOSE263	76923459	72812300	57499344	94.66%	74.75%	129	35.5
NOSE265	99961309	98345964	84553738	98.38%	84.59%	137	35.3
NOSE266	121726110	116756693	82723429	95.92%	67.96%	130	35.5
NOSE267	77337344	73514247	41775628	95.06%	54.02%	127	35.5
NOSE268	85645203	84307452	69500922	98.44%	81.15%	131	35.5
NOSE270	90008666	87531875	71592011	97.25%	79.54%	130	35.4
NOSE273	85787236	84357054	73559053	98.33%	85.75%	135	35.4
NOSE274	71328481	68304810	10432178	95.76%	14.63%	131	35.3
NOSE278	81120817	79571034	66025012	98.09%	81.39%	132	35.4
NOSE279	95176090	93286221	75808228	98.01%	79.65%	131	35.6
NOSE280	110488700	108310390	88147505	98.03%	79.78%	129	35.6
NOSE283	90534281	89048192	77765291	98.36%	85.90%	133	35.4
NOSE284	100835677	99537722	88398924	98.71%	87.67%	134	35.4
NOSE286	97214846	96302960	86476597	99.06%	88.95%	136	35.3
NOSE3	148038877	142221498	78351222	96.07%	52.93%	120	35.7
NOSE9	121028308	58098321	6179154	48.00%	5.11%	77	28.4
TIGER	90111778	88975524	76817835	98.74%	85.25%	134	35.4

Table 3.2 Gene ontology enrichments for genes with putative enhancers.

Of the 923 signficantly-correlated peak gene pairs, 729 unique genes were identified; 726 of these were identifiable in the GO ontology enrichment dataset. Shown here are the enrichments in biological processes of those 726 genes.

GO biological process complete	Homo sapiens - REFLIST (20996)	Our list (726)	expected	over/under	fold Enrich.	raw P-value	FDR
cellular biosynthetic process (GO:0044249)	2764	138	95.57	+	1.44	1.40E-05	3.19E-02
cellular process (GO:0009987)	14514	561	501.87	+	1.12	1.90E-06	6.04E-03
biological_process (GO:0008150)	17815	661	616.01	+	1.07	1.31E-06	6.92E-03
Unclassified (UNCLASSIFIED)	3181	65	109.99	-	0.59	1.31E-06	5.19E-03
detection of chemical stimulus involved in sensory perception (GO:0050907)	480	1	16.6	-	0.06	2.15E-06	5.71E-03
detection of chemical stimulus (GO:0009593)	517	1	17.88	-	0.06	6.95E-07	1.11E-02
detection of chemical stimulus involved in sensory perception of smell (GO:0050911)	430	0	14.87	-	< 0.01	8.63E-07	6.86E-03

Table 3.3 Putative gene-enhancer pairs in ovarian cancer risk regions.

Of the 923 signficantly-correlated peak gene pairs, 20 pairs were identified across 13 ovarian cancer risk regions. Listed here are the genes, putative enhancers, and genomic risk regions.

Gene	PeakChr	PeakStart	PeakEnd	r	p	Distance (kb)	gwasChr	gwasStart	gwasEnd
COPZ2	chr17	46452582	46461840	-0.53	5.3E-05	348	chr17	45920806	46920806
GMFG	chr19	39748917	39750271	-0.55	2.5E-05	73	chr19	39231783	40231783
HNRNPA3	chr2	178623056	178624056	0.44	1.0E-03	541	chr2	176537342	178537342
TTC30B	chr2	179266529	179292319	0.45	9.1E-04	863	chr2	176537342	178537342
TIPARP	chr3	155840054	155841947	-0.45	8.7E-04	567	chr3	155906997	156906997
PXN	chr12	121286740	121294609	0.45	8.4E-04	615	chr12	120903724	121903724
NR2F6	chr19	17176767	17264742	-0.45	8.2E-04	129	chr19	16890291	17890291
SLC12A7	chr5	1881751	1882844	0.45	7.2E-04	801	chr5	780830	1780830
PHOSPHO1	chr17	46428972	46430348	0.45	7.2E-04	875	chr17	45920806	46920806
C1QL1	chr17	43560290	43579290	0.45	7.2E-04	528	chr17	43000000	45500000
DYRK1B	chr19	39521351	39524033	0.46	6.5E-04	798	chr19	39231783	40231783
RINL	chr19	39108162	39112245	0.46	6.3E-04	253	chr19	39231783	40231783
SNX16	chr8	83271186	83273061	0.46	5.7E-04	539	chr8	82153644	83153644
NMT1	chr17	43441441	43457094	0.46	5.7E-04	287	chr17	43000000	45500000
PNPO	chr17	46483984	46485260	0.46	5.4E-04	462	chr17	45920806	46920806
BUB1	chr2	111068289	111069289	0.47	4.9E-04	347	chr2	111318658	112318658
TLL1	chr4	166825934	166827911	-0.48	3.4E-04	83	chr4	166687046	167687046
SLC12A7	chr5	648490	675221	-0.49	2.1E-04	419	chr5	780830	1780830
COPRS	chr17	29537883	29552281	0.50	1.4E-04	638	chr17	28721277	29721277
DNAJC1	chr10	22147375	22153382	0.51	1.1E-04	19	chr10	21330104	22330104

Table 3.4 Gene ontology enrichments for eGenes.

Of the 13,334 genes tested in the eQTL analysis, 580 eGenes were identified; 565 of these were identifiable in the GO ontology enrichment dataset. Shown here are the enrichments in biological processes of those 565 genes.

GO biological process complete	Homo sapiens - REFLIST (20996)	Our list (565)	expected	over/ under	fold Enrich.	raw P-value	FDR
protein targeting (GO:0006605)	365	26	9.82	+	2.65	1.53E-05	2.70E-02
establishment of protein localization to organelle (GO:0072594)	440	29	11.84	+	2.45	2.31E-05	3.34E-02
peptide metabolic process (GO:0006518)	536	33	14.42	+	2.29	2.35E-05	3.12E-02
protein localization to organelle (GO:0033365)	748	44	20.13	+	2.19	3.54E-06	1.13E-02
cellular amide metabolic process (GO:0043603)	790	44	21.26	+	2.07	1.16E-05	2.31E-02
detection of stimulus (GO:0051606)	691	1	18.59	-	0.05	3.28E-07	5.21E-03
detection of chemical stimulus involved in sensory perception of smell (GO:0050911)	430	0	11.57	-	< 0.01	2.05E-05	3.25E-02
detection of chemical stimulus involved in sensory perception (GO:0050907)	480	0	12.92	-	< 0.01	4.05E-06	1.07E-02
detection of stimulus involved in sensory perception (GO:0050906)	538	0	14.48	-	< 0.01	8.30E-07	6.60E-03
detection of chemical stimulus (GO:0009593)	517	0	13.91	-	< 0.01	1.90E-06	7.56E-03
sensory perception of smell (GO:0007608)	460	0	12.38	-	< 0.01	9.37E-06	2.13E-02
sensory perception of chemical stimulus (GO:0007606)	532	0	14.32	-	< 0.01	1.28E-06	6.81E-03

Table 3.5 Top eQTLs identified in ovarian cancer genomic risk regions.

Shown here are the eQTL summary statistics for the top eQTLs for the 12 eGenes within ovarian cancer risk regions.

eGene	p	effect size	top eQTL		
			chr	start	end
DNALI1	6.7E-05	3.99	chr1	37586578	38586578
FES	9.1E-06	-4.44	chr15	91009215	92009215
GATC	9.0E-09	-5.75	chr12	120903724	121903724
GJC1	6.7E-05	3.99	chr17	43000000	45500000
INA	4.1E-05	4.10	chr10	105194301	106194301
KCNAB1	8.2E-12	6.84	chr3	155906997	156906997
NDUFS6	4.5E-05	4.08	chr5	780830	1780830
NFE2L1	9.6E-06	4.43	chr17	45920806	46920806
PGLS	5.0E-05	4.06	chr19	16890291	17890291
SLC12A7	5.3E-05	4.04	chr5	780830	1780830
TMEM38A	5.0E-05	-4.06	chr19	16890291	17890291
TRIAP1	9.6E-06	4.43	chr12	120903724	121903724

3.7 References

1. American Cancer Society, 2020. https://cancerstatisticscenter.cancer.org/?_ga=2.164978303.1956443394.1581112601-1343696158.1581112601#!/cancer-site/Ovary
2. Jones MR, Kamara D, Karlan BY, Pharoah PDP, Gayther SA. Genetic epidemiology of ovarian cancer and prospects for polygenic risk prediction. Gynecol Oncol. 2017;147(3):705-713.
3. Cuellar-partida G, Lu Y, Dixon SC, et al, et al. Assessing the genetic architecture of epithelial ovarian cancer histological subtypes. Hum Genet. 2016;135(7):741-56.
4. Gusev A, Lawrenson K, Lin X, et al. A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. Nat Genet. 2019;51(5):815-823.
5. Adler E, Mhawech-fauceglia P, Gayther SA, Lawrenson K. PAX8 expression in ovarian surface epithelial cells. Hum Pathol. 2015;46(7):948-56.
6. Coetzee SG, Shen HC, Hazelett DJ, et al. Cell-type-specific enrichment of risk-associated regulatory elements at ovarian cancer susceptibility loci. Hum Mol Genet. 2015;24(13):3595-607.
7. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43-9.
8. Jenne DE, Reimann H, Nezu J, et al. Peutz-Jeghers syndrome is caused by mutations in a novel serine threonine kinase. Nat Genet. 1998;18(1):38-43.
9. Lawrenson K, Benjamin E, Turmaine M, Jacobs I, Gayther S, Dafou D. In vitro three-dimensional modelling of human ovarian surface epithelial cells. Cell Prolif. 2009;42(3):385-93.
10. Karst AM, Levanon K, Drapkin R. Modeling high-grade serous ovarian carcinogenesis from the fallopian tube. Proc Natl Acad Sci USA. 2011;108(18):7547-52.
11. Pharoah PD, Tsai YY, Ramus SJ, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. Nat Genet. 2013;45(4):362-70, 370e1-2.
12. Finucane HK, Bulik-sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47(11):1228-35.
13. Grubert F, Zaugg JB, Kasowski M, et al. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell. 2015;162(5):1051-65.
14. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383.
15. Brown TC, Murtha TD, Rubinstein JC, Korah R, Carling T. SLC12A7 alters adrenocortical carcinoma cell adhesion properties to promote an aggressive invasive behavior. Cell Commun Signal. 2018;16(1):27.
16. Sirnes S, Honne H, Ahmed D, et al. DNA methylation analyses of the connexin gene family reveal silencing of GJC1 (Connexin45) by promoter hypermethylation in colorectal cancer. Epigenetics. 2011;6(5):602-9.
17. Ramakrishna M, Williams LH, Boyle SE, et al. Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis. PLoS ONE. 2010;5(4):e9983.
18. Bonome T, Levine DA, Shih J, et al. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. Cancer Res. 2008;68(13):5478-86.

19. Liu Y, Yin X, Zhong J, et al. Systematic Identification and Assessment of Therapeutic Targets for Breast Cancer Based on Genome-Wide RNA Interference Transcriptomes. Genes (Basel). 2017;8(3)
20. Gallagher MF, Heffron CC, Laios A, et al. Suppression of cancer stemness p21-regulating mRNA and microRNA signatures in recurrent ovarian cancer patient samples. J Ovarian Res. 2012;5(1):2.
21. Végran F, Boidot R, Coudert B, et al. Gene expression profile and response to trastuzumab-docetaxel-based treatment in breast carcinoma. Br J Cancer. 2009;101(8):1357-64.
22. Wang M, Li L, Liu J, Wang J. A gene interaction network-based method to measure the common and heterogeneous mechanisms of gynecological cancer. Mol Med Rep. 2018;18(1):230-242.
23. Liu P, Qi X, Bian C, et al. MicroRNA-18a inhibits ovarian cancer growth via directly targeting TRIAP1 and IPMK. Oncol Lett. 2017;13(6):4039-4046.
24. Pelikan RC, Kelly JA, Fu Y, et al. Enhancer histone-QTLs are enriched on autoimmune risk haplotypes and influence gene expression within chromatin networks. Nat Commun. 2018;9(1):2905.

Chapter 4: Conclusion

This thesis has explored different mechanisms of genetic regulation contributing to the development of complex human genetic traits; specifically, this work has provided a foray into understanding the different ways that genetic variation can drive downstream phenotypes through direct and epigenetic regulation of target genes. Chapter 2 explored the shared genetic basis between complex traits and Mendelian disorders, primarily from the perspective of leveraging phenotypic information from genes linked to Mendelian disorders to identify genetic regulatory mechanisms driving complex trait risk, and Chapter 3 examined the molecular genomic mechanisms related to epithelial ovarian cancer risk.

In this final chapter, I suggest ways these findings can be applied toward the goals of isolating and identifying the genetic factors driving phenotypes, both towards the goals of better interpreting individuals' risk of developing a trait or disease and better understanding the biological basis of these varied phenotypes. First, my co-authors and I show that genetic risk variants identified by GWAS have higher effects for complex traits when located near genes with any Mendelian phenotypes, and even higher effects when located next to genes where the phenotype is related to the complex trait. A significant contribution of this work is the resulting conclusion that these rare-disease-linked genes are broadly important for complex traits. Our findings complement recent work on the omnigenic model in complex trait genetics[1], and further suggest that Mendelian genes may be the core genes underlying complex traits in the model. Future work focused on identifying core genes for various traits would benefit from examining and prioritizing evolutionary constrained genes, Mendelian genes, and specifically Mendelian genes with similar phenotypes as compared to the complex trait of interest. Furthermore, the findings of Chapter 2 have the potential to reveal differences in genetic architecture across populations with different ancestries. Although our study included GWAS from non-European

populations, the majority of genomic associations were derived from European populations, which are not necessarily representative of global genomic diversity. Based on the support for our hypothesis that Mendelian genes underlie complex traits, we can extend this hypothesis across populations. If future studies were to examine the shared genetic basis of complex traits and Mendelian disorders using GWAS from non-European populations, and/or genes identified to be linked to Mendelian disorders in global populations, we might develop a better understanding of the genetic architecture of complex traits, as well as gain insight into how well we have identified Mendelian genes and GWAS associations across populations.

This work has important implications for rare diseases as well. Given the support we find for the genetic regulatory hypothesis outlined in Chapters 2 and 3, as well as the large body of work linking regulatory genetic variation to changes in gene expression, the findings presented in this thesis have the potential to lend insight into understanding the mechanisms of variable presentation and penetrance in Mendelian disorders. One such example, from recent work on the KAT6A group of disorders, highlights how individuals with the same genetic variant driving KAT6A Syndrome can have a wide range of phenotypic severity[2], and the factors influencing this have yet to be pinpointed. Our work suggests that regulatory variation may underlie patterns like these, where common variation outside of Mendelian genes affect changes in epigenetic markers such as promoters or enhancers, or 3D chromatin conformation affecting physical interaction of different genomic regions, which together may affect disease gene expression leading to phenotypic differences unrelated to the “causal” rare variant for the syndrome. Importantly, the work presented in Chapter 3 indicates the importance of identifying the relevant tissue for each disease in which to study molecular mechanisms, and the importance of sufficiently powered population-level association studies to detect links from genetic to epigenetic to transcriptomic changes, and ultimately to phenotypes.

References:

1. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169(7):1177-1186.
2. Kennedy J, Goudie D, Blair E, et al. KAT6A Syndrome: genotype-phenotype correlation in 76 patients with pathogenic KAT6A variants. Genet Med. 2019;21(4):850-860.