# Looking beyond the standard genome-wide association study: Biologically-motivated methodological approaches to discover novel genetic variants associated with complex human traits and disease

by

Cindy Im

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Epidemiology

School of Public Health

University of Alberta

**ABSTRACT**

The standard approach for testing associations between common single nucleotide genetic variants (referred to as single nucleotide polymorphisms or SNPs) and disease entails testing disease associations for each SNP in the genome individually. This "hypothesis-free" approach has identified thousands of statistically significant associations between single SNPs and a wide range of diseases. However, complex forms of genetic variation – which include epistatic interactions, gene-environment interactions, inheritance patterns, rare variants, and structural variants – represent a tremendous potential source of transcriptional complexity in the human genome and may contribute substantially to disease risk. These complex forms of genetic variation are not explored in conventional single-SNP genome-wide association studies, largely due to computational, methodological, and statistical constraints.

In this dissertation, we look beyond the contributions of single SNPs to the genetic architecture of disease and consider novel approaches to investigate how untested classes of genomic variation present in the human genome may advance our understanding of the genetic basis of disease. More specifically, the studies presented in this thesis describe a novel methodological framework to detect patterns of epistasis (multiple SNP interactions) and haplotypes (SNP alleles arranged on the same chromosome) associated with complex disease traits that may also potentially model the regulation of trait-related gene transcription events, thereby elucidating central biomolecular mechanisms that influence disease trait pathogenesis. This methodological framework may be summarized as follows: first, a "filter" is employed to restrict the set of investigated SNPs to those with putative biological functions; subsequently, a novel, non-exhaustive statistical approach is implemented to discover candidate epistatic

interaction and haplotype associations with disease traits among filtered SNPs. As a final step, replication and biological inference analyses are conducted to assess the credibility of complex genetic variant discoveries. Under this framework, we increase the prior probability of identifying epistatic interactions or haplotypes that are transcriptionally relevant, and facilitate searches of the large space of interactions/haplotypes without limiting the number of tested associations using computational burden-based criteria to improve power. Our results demonstrate the relevance of studies of epistasis in explaining the variability of bone mineral density (an integral determinant of bone health) in adult survivors of pediatric cancer exposed to bone-diminishing treatments, and the effects of haplotypes on risk for primary biliary cholangitis (an incurable autoimmune disease of the liver) in Japanese. We suggest that the discovered genetic targets from these analyses be considered for future basic research into biological mechanisms influencing bone mineral density and primary biliary cholangitis, under the expectation that such research will support the eventual objective of developing potential health applications for the prevention, diagnosis, or treatment of these health conditions.

## PREFACE

This thesis is an original work by Cindy Im, and is part of a larger research project led by Prof. Yutaka Yasui that received research ethics approval from the University of Alberta Health Research Ethics Board (HREB) under project name "Statistical analyses – Genome Wide Association Study", No. Pro00042122, on August 30, 2013. As my PhD advisor, Prof. Yasui was responsible for supervising the ethical conduct of research and the overall direction of methodological approaches implemented as a part of this thesis.

Chapter 2 of this thesis has been published as C. Im, K.K. Ness, S.C. Kaste, W. Chemaitilly, W. Moon, Y. Sapkota, R.J. Brooke, M.M. Hudson, L.L. Robison, Y. Yasui, and C.L. Wilson, "Genome-wide search for higher order epistasis as modifiers of treatment effects on bone mineral density in childhood cancer survivors," *European Journal of Human Genetics*, vol. 26, pp. 275-286. The data used for the analyses presented in Chapter 2 come from the "St. Jude Lifetime Cohort Study" (SJLIFE), which was conceived, designed, and implemented under the supervision of L.L. Robison, M.M. Hudson, K.K. Ness, S.C. Kaste, W. Chemaitilly, and C.L. Wilson (funded by the National Cancer Institute, #U01 CA195547). C.L. Wilson and W. Moon provided technical assistance by providing access to SJLIFE data. Under the supervision of Prof. Yasui, I was responsible for developing the study hypothesis, designing the analytic method, performing genetic data quality checks, phasing/imputation, and annotation, conducting the analysis, interpreting and summarizing the results, and composing the manuscript. C.L. Wilson provided clinical expertise throughout all project stages. All co-authors contributed critical revisions to the final manuscript.

Chapter 3 of this thesis was submitted as C. Im, W. Moon, R.J. Brooke, Y. Sapkota, and Y. Yasui, "Genome-wide search for higher order epistasis as modifiers of treatment effects on bone mineral density in childhood cancer survivors" to *Advances in Neural Information Processing Systems 29*. Under Prof. Yasui's guidance, I contributed to the design of the simulation study, interpreted and summarized the results, and composed the manuscript. W. Moon provided technical assistance by conducting simulation study iterations. All co-authors contributed critical revisions to the final manuscript.

Chapter 4 of this thesis has been published as C. Im, Y. Sapkota, W. Moon, M. Kawashima, M. Nakamura, K. Tokunaga, and Y. Yasui, "Genome-wide haplotype association analysis of primary biliary cholangitis risk in Japanese", *Scientific Reports*, vol. 8, issue 1. The data used for the analyses presented in Chapter 4 come from the Japan PBC-GWAS (PBC: Primary Biliary Cirrhosis; GWAS: Genome-Wide Association Study) Consortium. Conception of the original study and initial data collection was led M. Nakamura (primarily funded by the Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science program, #20590800, #23591006, #26293181). M. Kawashima provided technical assistance by providing access to the data, while Y. Sapkota and W. Moon supported the preliminary processing and phasing of the genotype data. Under the supervision of Prof. Yasui, I was responsible for developing the study hypothesis, performing genetic data quality checks and annotation, designing the analytic methods, conducting the analysis, interpreting and summarizing the results, and composing the manuscript. Y. Sapkota, M. Nakamura, and K. Tokunaga provided critical input in the interpretation and presentation of study results. All co-authors contributed revisions to the final manuscript.

**DEDICATIONS**

*In loving memory of my father, Myo Soon Im.*

## ACKNOWLEDGEMENTS

There are many people that I would like to thank for their contributions to this thesis. First, I would like to express my deepest appreciation for having had the opportunity to work with Prof. Yutaka Yasui, my Ph.D. program supervisor. The thesis would not have been possible without his brilliant insights, profound research perspective, and generous and thoughtful mentorship. I also sincerely thank the members of my PhD Supervisory Committee, Prof. Sambasivarao Damaraju and Prof. Irina Dinu, for their invaluable feedback. Their collaborative efforts were integral for this work; I am thankful for their contributions to my professional development. I would also like to thank the wonderful faculty members and researchers at the University of Alberta (Profs. Keumhee Carriere Chough and Michael Kouritzin), St. Jude Children's Research Hospital (Drs. Leslie Robison, Kiri Ness, Carmen Wilson, Yadav Sapkota, Russell J. Brooke, Wonjong Moon), University of Alabama at Birmingham (Dr. Noha Sharafeldin), and the University of Tokyo (Prof. Katsushi Tokunaga) who have not only influenced this work, but inspired me as a researcher and lifelong learner.

This work was supported by several institutions and funding agencies. I am indebted to St. Jude Children's Research Hospital for providing me access to unparalleled research resources. I thank the Alberta Machine Intelligence Institute for providing me funding support throughout the duration of my PhD program. I would like to specifically acknowledge that the research described in Chapter 2 was funded by the St. Jude Lifetime Cohort Study (U01 CA195547), American Lebanese Syrian Associated Charities, Rally Foundation for Childhood Cancer Research, and National Institutes of Health Grant R01CA216354, while the study described in Chapter 4 was supported by funding provided by Grants-in-Aid for Scientific

Research from the Japan Society for the Promotion of Science, Grant-in-Aid for Clinical Research from the National Hospital Organization, Health Labor Science Research Grants from Research on Measures for Intractable Diseases, Intractable Hepato-Biliary Diseases Study Group in Japan, and the Japan Agency for Medical Research and Development.

Finally, I cannot express the depth of love and gratitude I have for my family: returning to school at this stage in my life would have been impossible without their unwavering support. To my incredible parents, Susan Myungsoo Im and Myo Soon Im – I will never be able to thank you enough for your sacrifices to give me the life that I have today, and for reminding me that as a citizen of the world, I should strive to contribute to a better tomorrow. And to my amazing husband, David Rudy Favero – there is not a day that goes by where I do not feel tremendously lucky that I managed to meet my soulmate on a Chinatown bus between Philadelphia and New York in 2005, and that you are now my partner in life. Thank you for being my most enthusiastic cheerleader and for perpetually inspiring me to challenge myself.

**TABLE OF CONTENTS**

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1

# Introduction

## 1.1 Genetic association analyses and the "missing heritability" of complex disease

Over the last decade, genome-wide association studies, or GWAS, have identified thousands of genetic susceptibility factors significantly associated with one or more complex human traits or diseases (1). The conventional approach for identifying these genetic susceptibility factors entails testing associations between common germline genetic variants (referred to as single nucleotide polymorphisms or SNPs, consisting of a variation at a single position in the DNA sequence that arises in at least 1% of the population) and a disease trait in a population-based sample of unrelated individuals. Since all assayed SNPs in the genome are tested indiscriminately for association with the trait of interest, this approach is frequently described as a "hypothesis-free" methodology. After conducting an association test for each SNP separately, a conservative multiple testing correction procedure is applied to control the Type I probability of making at least one false discovery.

Despite the tremendous success of GWAS, consideration of the sum total of GWA study findings to date for most complex disease traits suggests that these efforts serve as a starting point for future exploration. SNPs associated with complex disease traits are largely characterized by small effect sizes and reside in non-protein coding regions of the genome, providing relatively meager immediate translational clinical/public health opportunities (1). In addition, for most complex traits, the proportion of variation of a trait due to variation in genetic

factors (referred to as "heritability" or the $h^2$ statistic) explained by SNPs discovered in GWAS is far lower than the heritability that is expected to exist when estimated from family-based pedigree studies (2). This "missing heritability" of complex disease problem has become a call to action in the exploration of genetic susceptibility factors, since the identification of novel genetic susceptibility factors that explain more of the missing heritability of complex traits is anticipated to increase the clinical/public health utility of GWA study findings.

A popular theoretical framework that is commonly cited as one of the most plausible explanations for the missing heritability of complex disease is the "infinitesimal model", which hypothesizes that the majority of common causal SNPs, especially those with small effect sizes, remain undiscovered and will additively explain most of the variance of a trait (3, 4). The infinitesimal model further suggests that embracing the conventional single-SNP association testing strategy may partially resolve the problem of missing heritability. To contextualize this conceptual framework, consider the classical height GWAS. Height is a highly heritable trait ($h^2$=80%). Prior to Yang *et al.*'s (4) influential investigation of the missing heritability of height, GWA studies had detected ~50 SNPs associated with height that collectively explained ~5% of the variance of height. Assuming the infinitesimal model, Yang *et al.* (4) hypothesized that the additive effects of many common SNPs with small effects, likely in imperfect linkage disequilibrium (LD; association of alleles at two or more loci in a given population) with causal variants, could explain most of the heritability of height. By fitting a linear mixed effects model with a set of thousands of common SNPs simultaneously (~300K SNPs with genome-wide coverage) to estimate the variance explained by *all* SNPs while accounting for imperfect LD between tag and causal SNPs, this set of common SNPs was estimated to explain most (~67%) of the heritability of height.

While Yang *et al.* (4) demonstrates that most of the heritability of height could be captured by common variants of small effect, the method described in this paper does not prioritize specific loci that meaningfully contribute to height. Consequently, it is unclear whether this result truly enhances our understanding of the genetic determinants that meaningfully contribute to height. If nearly *every* locus in the genome contributes to a given complex disease (while contributing to numerous other traits and biological functions), how do we determine which genetic variants are the most meaningful contributors to the prevention, diagnosis, and treatment of a specific trait or disease? Under the infinitesimal model analytic strategy, the answer lies in performing additional conventional GWAS with even larger sample sizes and denser SNP arrays that provide higher levels of genome-wide coverage. While such investigations will inevitably uncover many novel SNP associations with increasingly smaller effect sizes, it is less certain whether these future SNP discoveries will independently offer greater clinical or biological insight than SNP discoveries reported in previous GWAS.

### 1.2 Exploring complex genetic variants and their associations with disease traits

While conducting large-scale conventional single-SNP GWAS and meta-analyses may identify many novel SNP associations of small effect to explain more of the missing heritability of complex traits and diseases, this dissertation considers resolving the problem of missing heritability to be a secondary goal. In this thesis, we primarily consider the development and use of methodological approaches that aim to advance our understanding of specific genetic determinants of complex traits and diseases by identifying genetic variants that may offer novel insights into biological mechanisms that underpin complex disease pathogenesis. To this end, we

look beyond single SNPs as genetic susceptibility factors and consider largely untested classes of complex genomic variation present in the human genome.

Complex forms of genetic variation represent an abundant potential source of transcriptional complexity in the human genome and may contribute substantially to disease risk (5). These complex forms of genetic variation are not evaluated specifically in conventional single-SNP GWA analyses. Examples of under-explored complex genetic variation include epistatic interactions (e.g., SNP-SNP interactions), gene-environment interactions (e.g., SNP-drug interaction), inheritance patterns (e.g., haplotypes), rare variants (e.g., SNPs with allele frequencies of <1%), and structural variants (e.g., copy number variants) (5).

Exploration of these untapped sources of genetic variation is appealing on many levels. Complex genetic variants potentially contribute to the *total* ("broad-sense") heritability of complex traits, and include dominance and non-additive effects (i.e., epistatic, gene-environment, and epigenetic effects) (5, 6). Given that heritability estimates that are strictly attributed to additive genetic variant effects ("narrow-sense" heritability) may be overinflated (7, 8), untested classes of complex genetic variants may ultimately offer greater insight into the broader missing heritability problem by explaining more of the total missing heritability of disease traits. Similar to the skewed distribution of common SNP allele frequencies favoring lower-frequency variant discoveries, lower frequency and rarer forms of complex genetic variation are anticipated to have larger effects on phenotypes (5), which is consistent with prevailing population genetics theory that suggests disease-causing genetic variation is unlikely to be common due to negative selection pressures (9). Unlike single SNPs, complex genetic variations may also potentially signal the involvement of multiple SNPs and/or genes that influence disease risk. As a result, some forms of complex genetic variation may better

contextualize the relative contributions or roles of specific loci, including those identified by previous single-SNP analyses, which can subsequently be targeted for clinical/public health actions to prevent or treat complex disease. Most importantly, complex genetic variants play a role in the tissue-specific expression of genes: for example, in a recent validation study of breast cancer-associated copy number variations (CNVs), germline copy number status for a subset of breast cancer-associated CNVs was reported to be correlated with the expressions of nine genes in breast tumor tissue (10).

For the remainder of this thesis, we focus on two specific classes of complex genetic variation: epistatic interactions and haplotypes. We further describe the importance of epigenetic modifications on the regulation of gene expression, and underscore the importance of exploring the interplay of SNP variations in regulatory regions of the genome under both epistatic and/or haplotypic contexts. Instead of conducting conventional analyses that assess the contributions of genetic discoveries to missing heritability estimates, candidate epistatic interactions and haplotypes associated with traits of interest will primarily be interrogated for plausible biological insights.

### 1.2.1 Epistasis

Epistasis, defined as the phenomenon where the effect of a genetic variant on a trait depends on the genotypes of other variants in the genome, is biologically essential: genetic interactions play a major role in gene regulation, signal transduction, biochemical networks, and pathways for homeostasis and development in multiple organisms (11-14). The epistasis phenomenon is theoretically supported by canalization, an evolutionary genetics concept that

describes how traits become robust to the knockout of one genetic element and require knockouts of multiple genetic elements for effect. Canalization is a posited selective force that can generate widespread epistasis (8, 12, 15), and evidence of canalization and epistasis are frequently observed in molecular studies with model organisms (16-18).

Genetic studies of epistatic SNP interactions and their associations with human traits are under-explored. One reason for discounting studies of epistasis is that such investigations are perceived to have little impact on the missing heritability problem. Narrow-sense heritability estimates, representing the relative contributions of additive genetic variation to the total variation of a trait, are estimated to be large for many complex traits. This implies that a search for epistasis, a source of non-additive genetic variation, would be relatively pointless for a highly heritable trait. Interestingly, much of the genetic variation from common SNPs that persists under selection is expected to be non-additive (8). This is consistent with reports that show estimates of narrow-sense heritability are inflated when complex traits are affected by epistasis (7). Thus, the contribution of non-additive genetic variation to the total (broad-sense) heritability of a complex trait should be much larger. Under this paradigm, studies of epistasis may contribute substantially to the overall missing heritability of complex traits.

### 1.2.2 Haplotypes

Haplotypes, or the arrangement of multiple (SNP) alleles on the same chromosome, may not only be more powerful for mapping novel disease genes (19), but may also be uniquely informative about known single SNP associations. Haplotype frequencies vary considerably between human demographic populations and bear signatures of positive selection; more

importantly, this class of complex genetic variants may advance our understanding of the role genetic variants play, singly or in tandem, in disease pathogenesis (19-21).

There is growing evidence that combinations of SNP genotypes that incorporate multiple *cis*-acting (acting on the same haplotype) allelic variants may affect common disease traits (22). Essentially, the expression of a given mRNA transcript is controlled by *cis*-acting factors (e.g., SNP allele variations) in coding regions and/or the flanking DNA sequences surrounding genes, as well as *trans*-acting factors (acting on the opposing haplotype; e.g., transcription factors) (23). These combinations of *cis*- and *trans*-acting factors have wide-ranging effects on the stability, processing, or isoform expression of mRNA transcripts; in particular, *cis*-acting variation is conservatively estimated to explain up to 35% of interindividual differences in gene expression (23). A functional validation study of the interleukin-1 gene family (previously linked with diseases with an inflammatory response in multiple genetic association studies) in human monocyte cells demonstrated the importance of haplotypes by showing that individual SNPs in the *IL1B* promoter region contribute to allele-specific expressions and affect promoter function in a haplotype-specific manner (24). These results suggest that without haplotype information, we likely have incomplete knowledge of the functional consequences of the unique distribution of variants among two homologous chromosomes in genic regions on disease pathogenesis (22, 24).

### 1.2.3 Relevance of epigenetic modifications and gene regulation in studying epistasis and haplotypes

Epigenetic modifications, or biomolecular or chemical changes to DNA that do not alter the DNA sequence, broadly impact the regulation of genes, as well as cellular development and differentiation (11, 25). One of the major molecular mechanisms that mediate epigenetic phenomena is the biochemical modification of the histone proteins that dictate the spatial structure of DNA to form chromatin (26); as such, global patterns of tissue- and context-specific histone modification marks are indicators for functional regulatory regions of the genome (25, 27). An important mechanism for regulating gene transcription involves the formation of chromatin loops, enabling physical interactions between regulatory genomic regions (11, 28); epigenetic modifications may therefore allow common SNPs in genomic regulatory regions to influence physical interactions between regulatory regions in a tissue-specific manner to affect phenotypes (29, 30), while allele-specific variants may also affect gene expression in a chromosome-/haplotype-specific manner (22).

Two major classes of genomic regulatory elements known to modulate gene transcription through interaction-based or *cis*-acting mechanisms include "promoters" (DNA sequences that are upstream of transcription start sites, and define where transcription begins) and "enhancers" (DNA sequences that stimulate transcription of target genes) (11). For many genes, the promoter is insufficient to drive gene expression: gene transcription frequently also depends on additional distal sequences of DNA (i.e., enhancers) that are *cis*-acting (31).

An example of the potential utility of investigations that consider how epistasis, haplotypes, and epigenetic variants can influence gene function and disease pathogenesis comes from genetic studies of obesity. Multiple GWAS of obesity-related traits have observed that SNPs residing in the first intron of the *FTO* gene have the strongest associations with obesity-related traits (32, 33). However, direct connections between obesity-associated SNPs and *FTO*

expression have never been confirmed (29). Smemo *et al.* (29) hypothesized that the obesity-associated *FTO* SNPs affected an alternative target in the genome, given that the genomic region of interest within *FTO* was enriched with enhancer-associated epigenetic marks. Using chromosome conformation capture methods to detect long-range physical interactions between the putative *FTO* enhancer region and other regions of the genome, Smemo *et al.* (29) revealed that the *FTO* enhancer region not only physically interacts with the *IRX3* gene promoter region, but confirmed that the obesity-associated SNPs in *FTO* contributed to an enhancer that distally regulates *IRX3* activity. Based on these results, the authors hypothesized that allelic variants in the *FTO* enhancer region disrupt binding with *IRX3*, leading to altered *IRX3* expression and disrupted production of a transcription factor in the brain known to play a role in the regulation of body mass.

Lastly, environmental exposures can also affect cell- or tissue-specific epigenetic modifications that result in differential gene expressions that persist over time (26). In considering lines of public health research inquiry that can explain the genetic basis of disease, it is crucial to consider investigations that have the potential to dissect the nexus between epistatic and haplotypic genetic variants and the environment. With the understanding that modifiable epigenetic processes allow organisms to respond to the environment, it is worthwhile to exploit growing insights surrounding epigenetic phenomena to identify novel complex genetic variants and environmental exposures that collectively influence disease risk.

### 1.2.4 Limitations of current methods to study epistatic interaction and haplotype associations

There are several excellent reviews of available methodologies to study epistasis (34-36) and haplotypes (19). Broadly speaking, computational approaches employed to detect SNP interaction and haplotype associations fall into two broad categories: exhaustive or non-exhaustive. Exhaustive searches test all possible statistical interactions or haplotypes. Non-exhaustive searches, on the other hand, perform a selective search of the total interaction space.

The most common approach to study epistasis is to exhaustively test two-way SNP interaction associations using a likelihood ratio test (LRT) (37) after restricting SNPs to those with significant marginal effects. Specifically, for a *k*-way interaction, each of *k* SNP loci are coded for either additive or dominant genetic inheritance effects, and the full generalized linear model (GLM) includes 2**k* main effects and $2^k$ parameterized interactions between the *k* SNPs. In the SNP pair case, the full GLM between two loci is:

$$g(E[Y|X]) = \mu + a_1x_1 + d_1z_1 + a_2x_2 + d_2z_2 + i_{aa}x_1x_2 + i_{ad}x_1z_2 + i_{da}z_1x_2 + i_{dd}z_1z_2,$$

where $a_j$ is the regression coefficient for the additive effect at loci *j*, $d_j$ is the regression coefficient for the dominant effect at loci *j*, and the $i_{xy}$ are four non-redundant interaction parameters between the two loci under additive/dominant effects. Thus, for a two-SNP model, this four degree-of-freedom LRT would be performed exhaustively for all possible SNP pairs.

Interpretations of this specific statistical method to study SNP interactions may be conflated with true biological epistasis. Most published studies of SNP interactions report epistasis as a departure from additivity using this four degree-of-freedom LRT for pairs of SNPs (36-39), often with the prerequisite that SNPs have statistically significant marginal effects (36, 40). This specific interpretation of interaction, however, is inconsistent with a broader understanding of biological epistasis. First, failure to observe departure from additivity does not imply that biological epistasis does not exist (8, 41-43). Second, biological interaction is

plausible without exhibiting marginal effects, in that trait effects that require perturbation of a network of genes may include single genes with no trait effects (14). In addition, tests that examine a limited number of interaction types between loci preclude study of the many other epistatic models that are hypothesized to exist (44). Thus, using a specific statistical definition of epistasis may obscure findings of biological epistasis even when it truly exists.

Regarding haplotype analysis, the most significant challenge in conducting large-scale haplotype association analyses is that SNP array data typically does not include haplotype or "phase" information given the high cost of large-scale experimental phasing. As a result, resolution of which of the two parental chromosomes a given SNP allele is located is generally resolved with statistical and computational phasing methods. For haplotype association testing, the most popular strategy is to take an exhaustive approach that entails splitting the genome indiscriminately into small overlapping "sliding" windows (e.g., 50- to 500-kb in size), and simultaneously inferring and testing haplotype frequencies formed with a small, fixed number of SNPs in each window under a regression-based framework (19, 45, 46). Associations between haplotypes and disease risk may be tested with a global test of haplotypes ($h$-1 degrees of freedom for $h$ haplotypes, treating the most common haplotype configuration as a reference), with variance estimates that account for the uncertainty in the haplotype estimation (45). The main drawback of this approach to study haplotype associations is that the computational burden of simultaneously inferring and testing haplotypes constrain the size of windows for haplotype formation, and constructed haplotypes frequently only consider contiguous SNPs.

A considerable barrier in conducting studies of epistasis or haplotypes is that all available methods suffer from insufficient power. For example, an exhaustive search of only two-way SNP interactions with a 500,000 SNP array yields a per-test significance threshold of $<5\text{x}10^{-13}$ after

Bonferroni correction. Power is further eroded when investigating larger *n*-way SNP combinations, sparse SNP combinations, or SNPs involved in interactions that are not highly correlated with the "causal" SNPs they tag (34). Similarly, power to discover novel haplotype associations is also adversely affected as the number of SNPs considered in haplotype formation increase (e.g., the number of degrees of freedom increase as the number of considered haplotypes increase) and when haplotypes are rare.

## 1.3 Objectives and Aims

The central objective of the research presented in this dissertation is to detect transcriptionally-relevant epistatic interactions and haplotype patterns associated with complex disease traits to advance our understanding of both the genetic factors and biomolecular mechanisms that influence disease pathogenesis. As a secondary goal, we intend to advance methods used to study associations between complex disease traits and epistatic interactions and haplotypes on a genome-wide scale, and adopt a general framework to improve scientific inference for such investigations.

An important consequence of the various scientific challenges described in Section 1.2.4 is that large-scale explorations of epistatic SNP interactions and haplotypes as models for the regulation of genes have been limited by existing methods. Given the potential for false discovery, each of the large-scale genome-wide association analyses presented in this thesis include three general components to improve scientific inference. First, we emphasize a biologically-motivated perspective to detect complex genetic variants associated with a phenotype/disease that may plausibly reflect mechanisms for transcriptional regulation. We

accomplish this goal by focusing on genetic variant classes that can potentially model regulatory activity in the genome (i.e., epistatic interactions and haplotypes), and by applying SNP "filters" to increase the prior probability of detecting transcriptionally-relevant SNP interactions or haplotypes during discovery. Second, for each analysis, we use two separate cohorts: one cohort is reserved for the discovery of candidate association signals, while the other cohort is strictly used to replicate candidate association signals. Lastly, we leverage several publicly accessible bioinformatics resources to assess the biological plausibility of implicated SNPs that contribute to replicated interaction and haplotype associations. Under this framework, we aim to build a body of evidence for discovered findings and help guard against false positives.

Study-specific hypotheses and aims are provided in the subsequent paragraphs for each of the three projects presented in this thesis.

*Hypothesis 1: Regulatory SNP interactions may modify the effects of cancer treatments known to diminish bone mineral density in adult survivors of pediatric acute lymphoblastic leukemia.*

Nearly every child treated for acute lymphoblastic leukemia (ALL) is exposed to high cumulative doses of anti-leukemic treatments for prolonged periods and experiences a decline in bone mineral density (BMD), a clinical predictor of osteoporosis and long-term bone health, immediately after treatment (47). Although many pediatric ALL survivors recover, some have far lower BMD in comparison to age- and sex-matched reference populations during adulthood (47-50). The extent of variation in response to the treatments clearly indicates that the treatment exposures are not universally toxic to bone development, and further suggests that genetic predisposition can worsen treatment effects. Since the heritability of BMD is estimated to be 60-

80% (51) and no prior GWA study of BMD has identified single SNPs exhibiting Mendelian effects (e.g., physical expression of a trait depends on the presence or absence of a single gene or SNP), epistasis may explain some of the variation in BMD in response to treatment among ALL survivors.

We hypothesize that common SNPs that reside in interacting genomic regulatory regions may not necessarily have significant individual effects, but may modify gene transcription through interaction to affect treatment-related deficits in BMD. This genome-wide association analysis of epistasis aims to:

1. Restrict the pool of tested SNPs to those mapped to genomic regions strongly predicted to regulate gene function, specifically regions with predicted enhancer or promoter function in any of nine diverse human cell types; and
2. Identify combinations of regulatory SNPs (epistatic interactions) associated with BMD in adult survivors of pediatric ALL with a novel "biologically-motivated" statistical algorithm.

*Hypothesis 2: Statistical learning algorithms tailored to study the unique features of genomic datasets may provide novel methodological solutions for future genetic association studies.*

Discovery of novel genetic susceptibility factors associated with complex disease traits are limited by existing analytic methodologies employed in genetic epidemiology. In particular, the methods applied in standard single-SNP GWAS do not address all of the unique challenges that the study of genomic datasets pose in the genetic epidemiology field. This study therefore aims to:

1. Describe the genomic data structure in mathematical terms and contextualize the challenges associated with analyzing such data to a machine learning audience;
2. Propose a biologically-motivated machine learning approach to identify SNP interactions that potentially influence gene regulation events underpinning complex disease traits; and
3. Evaluate the performance of the proposed approach by assessing the precision of a key aspect of this approach in a simulation study.

*Hypothesis 3: Gene-based haplotype patterns that potentially influence gene transcription events may contribute to primary biliary cholangitis risk in Japanese.*

Primary biliary cholangitis (PBC) is a progressive autoimmune disease of the liver and is characterized by the irreversible destruction of the bile ducts of the liver. PBC has a strong hereditary component, with an estimated concordance rate of 63 percent in monozygotic twins (52). PBC GWAS in Japanese cohorts have revealed novel features of the PBC genetic architecture that have not been observed in European populations, with *TNFSF15*, *POU2AF1*, and *PRKCB* emerging as major susceptibility loci among Japanese (53, 54). We hypothesize that gene-specific haplotype associations may not only identify novel PBC risk loci but may shed additional insights on previously reported susceptibility loci by potentially modeling gene transcription events. This genome-wide haplotype association study aims to:

1. Restrict the pool of tested SNPs to those mapped to a transcriptionally-relevant genomic region, or an extended gene-centered window that includes flanking DNA regions surrounding each gene to capture corresponding gene regulatory regions; and

2. Detect gene-specific haplotype-based combinations of SNP alleles associated with PBC risk in Japanese with the implementation of a novel statistical algorithm.

## 1.4 General methods

### 1.4.1 St. Jude Lifetime Cohort Study (SJLIFE)

Initiated in 2007, SJLIFE is a single institution-based retrospective cohort study supported by the National Cancer Institute (NCI), and aims to establish a lifetime cohort of childhood cancer survivors treated at St. Jude Children's Research Hospital (SJCRH) to support the prospective study of long-term health outcomes in this population (55). SJLIFE participants eligible for this analysis include individuals who: were 18 years or older at enrollment; were treated for a pediatric malignancy at SJCRH; and survived ≥10 years post-diagnosis. All enrolled participants undergo medical, physical, psychosocial, and neurocognitive assessments. The SJLIFE research protocol seeks to maximize participation by offering monetary compensation for missed days at work and childcare expenses, and providing cost-free transport, housing, and clinical evaluation. Non-participation bias is likely limited for SJLIFE (56), and response rates continue to be high: as of April 2016, 85% of the 4,963 survivors in the source population agreed to participate, of which 3,186 survivors have completed the initial clinical assessment.

The most important distinguishing feature of SJLIFE is that all late effects (adverse health conditions due to the effects of curative therapies for cancer on healthy tissues) are clinically ascertained. Previously, Phillips *et al.* (57) showed that survivors face an escalating burden of morbidity as they age, and estimated that nearly half (48%) of survivors aged 40-49 years have a severe chronic condition in a US population-level prevalence dataset with nearly

110,000 childhood cancer survivors. These estimates, however, are largely based on self-reported outcomes and fail to capture subclinical events, likely underestimating the true burden of late effects morbidity. Under the National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE), the cumulative incidence of disabling or life-threatening chronic health conditions (CTCAE grades 3-5) in the SJLIFE study population was estimated to be 96% at age 50 years (58).

The analysis presented in this thesis is restricted to 2,284 SJLIFE participants with both BMD measurements and SNP genotype data. This sample was split into two separate cohorts: (1) the discovery cohort, comprised of 856 pediatric ALL survivors, and (2) the replication cohort, with 1,428 survivors of pediatric non-ALL cancers.

Bone mineral density (BMD): BMD was ascertained in SJLIFE participants using quantitative computed tomography (QCT), which is the optimal method for BMD measurement in this population (47). The QCT method takes direct measurements of trabecular volumetric BMD, and is reported to be more sensitive to disease-related bone change and is less likely to overestimate BMD in obese individuals compared to another commonly used method, dual energy X-ray absorptiometry or DXA (47). The measure of BMD used is the BMD Z-score, expressed in units of standard deviation. BMD Z-scores were computed for each participant by taking the difference between the average of their respective two vertebral BMD measurements and the age- and sex-matched mean of a reference population, divided by the standard deviation of BMD in the reference population.

Relevant clinical data: Information about past treatment exposures and other relevant demographic and clinical factors (e.g., sex, age, cancer diagnosis) were obtained by trained medical record abstractors using standardized research protocols.

Processing genotype data: The Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA) was used to genotype DNA samples for consenting SJLIFE participants. This SNP array is capable of simultaneously genotyping HapMap European, Asian, and African populations with comparable genome coverage, and has an overall per-sample call rate of at least 97% for >900,000 SNP probes.

Preliminary quality control thresholds for SNP or sample exclusion in the discovery cohort were as follows: <95% per-sample call rate across markers; <95% SNP call rate across samples; and SNP minor allele frequency (MAF) <1%. Violations of Hardy-Weinberg equilibrium (HWE) were evaluated among remaining SNPs and samples (indicator of genotyping error and stability of SNP genotype frequencies from generation to generation), using a SNP exclusion threshold of $P<1x10^{-6}$ from genotype chi-squared tests. After performing these quality control steps using the GenABEL package in R (version 3.1.2), no participant samples were excluded and a total of 770,471 autosomal SNPs were retained. Imputation of missing measured SNP genotypes in the discovery cohort was completed using BEAGLE version 4.0 (59). All SNPs retained for analysis met the allelic $R^2$ cutoff of >0.5 (indicator of imputation quality score). Quality control thresholds in the replication cohort were consistent with the discovery cohort to assure per-SNP and per-sample genotyping quality control.

### 1.4.2 Japan Primary Biliary Cirrhosis – Genome-Wide Association Study (PBC-GWAS) Consortium Study

The Japan PBC-GWAS Consortium Study is a case-control study coordinated by the member hospitals of the National Hospital Organization Study Group for Liver Disease in Japan (NHOSLJ) and the gp210 Working Group in the Intractable Liver Disease Research Project Team of the Ministry of Health and Welfare in Japan. A total of 4,324 participants were originally recruited over two study enrollment time periods to conduct two PBC GWAS (53, 54) in Japanese.

Primary biliary cholangitis (PBC; also known as primary biliary cirrhosis): Healthy controls reporting no apparent disease were recruited from the medical staff at the participating hospitals of the Japan PBC-GWAS Consortium. PBC cases were defined by laboratory or histological evidence of at least two of the following criteria: cholestasis, ascertained by elevated alkaline phosphatase; serum anti-mitochondrial antibodies; and non-suppurative destructive cholangitis and interlobular bile duct destruction.

Processing genotype data (Japan PBC-GWAS): All DNA samples were genotyped using the Affymetrix Axiom Genome-Wide ASI 1 Array (Affymetrix, Santa Clara, CA). This SNP array is reported to be the first array with high genomic coverage of rare alleles in a consensus East Asian genome that considers both Han Chinese and Tokyo Japanese HapMap genomes, and has an average SNP call rate of >99% for >600,000 SNP probes.

Participant samples with excess heterozygosity rates and cryptic relatedness were removed from analysis, as well as samples of non-Japanese ancestry as determined by principal components analysis using HapMap-JPT (Japanese in Tokyo, Japan) samples as a reference to control for population stratification. A total of 425,290 autosomal SNPs were retained under the following quality control criteria: SNP call rate ≥95%; MAF ≥5%; and Hardy-Weinberg equilibrium (HWE) P ≥0.001 in controls. Samples with <97% sample call rate among retained SNPs were excluded, resulting in a study sample of 2,886 individuals. Haplotype phase on the quality-controlled, unphased SNP genotype data was computationally estimated using SHAPEIT v2.79 (60) for whole chromosomes for all 2,886 samples simultaneously.

Prior to conducting the haplotype association analysis, the study sample of 2,886 participants was split *a priori* into two separate cohorts to correspond with the timing of sample collections for the two previous GWAS conducted by the Japan PBC-GWAS Consortium. A cohort of 1,937 participants (901 cases, 1,036 controls) was reserved for the discovery of haplotype signals associated with PBC risk, while a cohort of 949 participants (480 cases, 469 controls) was available for the replication of candidate haplotype signals.

### 1.4.3 Bioinformatics databases

Data from several major web-based bioinformatics resources were used to annotate SNPs prior to analysis and evaluate the biological plausibility of complex genetic variants of interest associated with BMD and PBC risk. Brief descriptions of each resource are provided in **Table 1.1**.

**Table 1.1: List of major web-based bioinformatics resources**

| Resource | Description |
|---|---|
| *Encyclopedia of DNA Elements (ENCODE)* | ENCODE (25) hosts thousands of epigenome (global pattern of epigenetic marks) and transcriptome (complete set of gene transcripts) datasets for hundreds of human cell and tissue types, and is the most comprehensive resource for characterizing functional elements of the human genome. **URL**: www.encodeproject.org |
| *Ensembl* | Ensembl (61) includes genomic annotations for humans along with >80 vertebrate species, and is useful for examining genomic variations (e.g., SNPs) and their consequences on genes and genotypes in HapMap populations. **URL**: www.ensembl.org |
| *Roadmap Epigenomics Mapping Consortium Web Portal (REMC)* | The REMC Web Portal provides access to estimated chromatin state annotations (classifications of genomic regulatory states based on learned patterns of epigenetic marks) for 127 consolidated human cells or tissues (62). REMC also hosts experimental epigenetic mark datasets used to develop chromatin state annotation models. **URL**: egg2.wustl.edu |
| *Genotype-Tissue Expression Project (GTEx)* | The GTEx Portal (63) enables access to summary-level association data between SNPs and gene expression levels or expression quantitative trait loci (eQTL), in over 40 major cell and tissue types. **URL**: www.gtexportal.org |
| *WashU EpiGenome Browser* | The WashU EpiGenome Browser (64) supports powerful visualizations of long-range chromatin interactions (regions of chromatin that may be linearly far away, but are in close physical proximity based on the spatial configuration of DNA) in many different human cell and tissue types. **URL**: http://epigenomegateway.wustl.edu/ |

### 1.4.4 Annotation of SNPs: Enhancer and promoter regulatory regions

To identify SNPs that "tag" putative enhancer or promoter regions of the genome, we use annotations derived from a statistical model learned on ChIP-seq data (chromatin immunoprecipitation followed by sequencing, to profile DNA-binding proteins genome-wide) that characterizes chromatin state changes reflecting enhancer or promoter function (27). We used an external ENCODE database of "ChromHMM" chromatin state annotations estimated by a Hidden Markov Model (probabilistic model that can be used to learn patterns of observed data to label states that are not directly observable or "hidden") (27) to identify SNPs that "tag" putative "strong enhancer" and "active promoter" regions in any of nine major human cell lines.

### 1.4.5 Annotation of SNPs: Extended gene regions

The search for haplotype signals was centered on annotated protein-coding and non-translated RNA-encoding genes annotated by the RefSeq gene model (release 74, GRCh37/hg19 build) (65). ANNOVAR (66) was employed to map SNPs in our dataset to introns, exons, and 3'/5' untranslated regions. After SNP-gene annotation, 500-kb flanking regions before and after transcription start and stop sites were identified for each RefSeq transcript to capture potential regulatory elements embedded in gene-flanking regions (25).

### 1.4.6 Analytic methods

#### 1.4.6.1 Adapting logic regression to study SNP interactions and haplotype patterns

Gene expression is modulated by genetic variants comprised of both proximal and distal SNPs with individual or interacting effects (67, 68) and is often influenced by larger networks of SNPs (i.e., three SNPs or more) (38). Given these biological contexts, our goal was to implement a non-exhaustive search method to detect transcriptionally-relevant interactions or haplotypes between three SNPs associated with complex disease traits without requiring SNPs to be proximal, have marginal effects, or follow specific models of interaction. We chose to base the development of our statistical algorithm on the logic regression methodology (69) because of its suitability with our research aims.

Logic regression is an adaptive regression methodology that combines generalized linear models (GLMs) with a stochastic search algorithm to detect higher order interactions of binary

predictors associated with an outcome, and has been successfully applied in both genome-wide and candidate gene association analyses (70-73). Detected interactions, or "logic trees", are binary Boolean variables (true/false statements) that join binary predictor variables with "and"/"or" statements at nodes and genetic variables at "leaves" (terminal nodes).

Our analyses use the "simulated annealing" stochastic search algorithm implementation of logic regression. Essentially, simulated annealing finds interaction predictors based on stochastic-process theories of Markov chains and uses a move set defined by six permissible moves to "grow/trim" a logic tree: as a result, any constructed logic tree is in a neighborhood of trees that are within one of six basis moves of the current tree under consideration. The stochastic search algorithm utilizes the GLM regression framework to score candidate models to select the best-fitting logic tree. Given a maximum number of leaves and iterations ($n_{iter}$), the simulated annealing algorithm for selecting a single logic expression may be summarized as follows:

1. Initialize with a logic tree $L_0$ with one leaf.
2. For $k = 1, \dots, n_{iter}$:
    a. Propose a new tree, $L_{new}$, by randomly selecting a permissible move.
    b. Accept the new tree with acceptance probability $\min\{1, \exp\left(\frac{M_{k-1} - M_{new}}{T}\right)\}$.

$M_{k-1}$ and $M_{new}$ are the respective GLM scores (e.g., residual sum of squares for continuous traits or deviance score for binary outcomes) for $L_{k-1}$ and $L_{new}$. The temperature $T$ is controlled by a simulated annealing "cooling scheme"; as the temperature decreases, the probability of accepting a new model with a worse score relative to the current model decreases.

For the studies included in this thesis, 3-SNP interactions were defined as logic trees that combine three SNPs in genomic regulatory regions, e.g., ((promoter SNP A and enhancer SNP

B) or enhancer SNP C) = {True, False}, while 3-SNP haplotypes were defined as logic trees that combine SNP alleles on the same chromosome in extended gene regions, e.g., ((SNP1=reference allele) and (SNP2=alternative allele) and (SNP3=alternative allele)) = {True, False}. In general, the logic regression search for each epistatic interaction or haplotype tree was conducted under at least 100 randomly selected initialization values. Among these logic regression fits, the best-fitting linear or logistic regression models were selected by comparing decrements in the model residual sum of squares or deviance scores, respectively, to assure algorithm performance stability.

We applied different logic regression-based algorithms for the study of regulatory SNP interaction associations with BMD and gene-based haplotype associations with PBC risk based on the potential size of the interaction/haplotype search space. For the genome-wide interaction association analysis of BMD, we considered chromosome-wide interactions between SNPs mapped to enhancer/promoter regions and therefore applied a novel "sequential conditioning" algorithm to each of the 22 autosomes separately. Each 3-SNP interaction tree was detected via logic regression one at a time for each chromosome, using forward addition to form a linear predictor of up to ten 3-SNP interaction trees. This algorithm has certain advantages compared to a marginal search for chromosome-wide 3-way SNP interactions: (1) a conditioned search can guide the stochastic search in different directions from previously identified best interaction trees in the current model; and as a result, (2) subsequently identified trees are less likely to be correlated with previously detected trees in the model. Logic regression models with SNP interactions took the following form:

$$E[Y|X,Z] = \beta_0 + \sum_{j=1}^{p} \beta_j Z_j + \sum_{k=p+1}^{m} \beta_k L_k,$$

where $Y$ is the BMD Z-score, $X$ is the vector of binary SNP variables, $Z$ is the vector of non-genetic covariates associated with BMD Z-score, and $L_k$ is the vector of 3-SNP interaction trees, combining multiple regulatory SNPs in $X$.

To contrast, for the genome-wide haplotype association analysis of PBC risk, we evaluated smaller interaction search spaces given our interest in 3-SNP haplotypes mapped to extended gene-based windows (e.g., ~1-2 Mb in size). After detecting the best-fitting 3-SNP haplotype, we addressed the fact that the haplotype logic regression models consider two observations with the same case/control status from each subject (i.e., two haplotypes from two homologous chromosomes). We therefore used the following logistic regression model for the best detected 3-SNP haplotype associations, treating each subject as an independent observation to satisfy key GLM assumptions for valid statistical inference:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 L$$

where $p$ is PBC risk and $L$ represents additive haplotype effects of the best-fitting 3-SNP haplotype logic tree.

#### 1.4.6.2 Permutation-based inference for the discovery of association signals

To identify candidate SNP interactions and haplotypes to follow-up in replication cohorts, we proposed the use of a permutation-based "evaluation statistic" instead of an overly conservative Bonferroni-corrected p-value threshold. While the implementation details differ across studies, we essentially employed $m$ permutations of the disease trait (i.e., BMD Z-score or PBC case/control status) to compute an empirically derived median from relevant model statistics under permutation for each of the candidate 3-SNP interaction or haplotype trees, along

with its corresponding median absolute deviation (MAD, a robust measure of variability, defined as the median of absolute deviations from the permutation-derived median). The permutation-based evaluation statistic was then defined as $\frac{s_{obs}-s_{med}}{MAD_{s_{med}}}$, where $s_{obs}$ is the observed model statistic for a given candidate SNP interaction or haplotype, and $s_{med}$ and $MAD_{s_{med}}$ are the median and MAD, respectively, from the corresponding model statistics from $m$ permutations. To select candidate SNP interactions for replication follow-up in our BMD analysis, we considered 3-SNP interaction trees with permutation-based evaluation statistics two or more absolute deviations away from its corresponding permutation-derived median to explain unusually large proportions of BMD Z-score variation. For the haplotype association study of PBC risk, candidate 3-SNP haplotypes with the top 1% of permutation-based evaluation statistics were selected for replication follow-up.

#### 1.4.6.3 Ancillary analyses to support biological inference

Chromatin state and histone modification mark enrichment analysis: We assessed whether the set of SNPs identified as members of replicated SNP interaction or haplotype associations were enriched in either enhancer, promoter, or open chromatin states in cell or tissue types relevant to the disease trait of interest relative to a chosen comparison SNP set. Chromatin state annotation or experimental histone modification mark databases for 127 cell/tissue types were utilized to evaluate SNP overlaps with regulatory elements (62). Strength of evidence for regulatory element enrichments in each cell/tissue type was evaluated using a 2-sided Fisher's exact test.

Gene expression enrichment analysis: We tested whether there were excesses of significant gene expressions associated with SNPs in replicated SNP interaction or haplotype associations in cell or tissue types relevant to the disease trait of interest. Data from the GTEx Project (63), specifically significant *cis*-eQTLs (expression quantitative trait loci, defined as SNPs within +/-1 Mb of gene transcription start sites with expression associations meeting a q-value threshold of 0.05) in tissue types with N≥70 samples, were primarily used for this purpose. The observed proportion of significant eQTLs in selected cells/tissues for SNPs of interest were compared to the proportion of significant eQTLs for comparison SNPs using a 2-sided Fisher's exact test.

Plausibility of physical chromatin interactions: Evidence of physical chromatin contact enhances the plausibility of interaction between regulatory regions that contain SNPs of interest. The WashU EpiGenome Browser was used to examine evidence of physical chromatin contacts (64). A long-range chromatin interaction data library generated from lymphoblastoid cells was used to assess evidence of physical chromatin contact (74); to this end, we considered chromatin interactions with at least +4-fold observed contact frequency over expected between regions bearing SNPs involved in replicated SNP interactions.

### 1.5 Ethics statement

The ethics committees of participating institutions and the Human Research Ethics Board of the University of Alberta approved the study methods described in this thesis. All research participants provided informed consent. The studies included in this thesis are analyses of de-

identified databases extracted from stored clinical and genetic data and therefore pose no additional risks to study participants.

# CHAPTER 2

## Genome-wide search for higher order epistasis as modifiers of treatment effects on bone mineral density in childhood cancer survivors

### 2.1 Introduction

Survivors of pediatric acute lymphoblastic leukemia (ALL) are at risk for long-term deficits in bone mineral density (BMD) due to childhood cancer treatment exposures, including cranial radiation, antimetabolites (e.g., methotrexate), and glucocorticoids (47-49). Cranial radiation diminishes BMD through injury to the hypothalamic-pituitary axis, affecting sex and growth hormone secretions that play an important role in bone metabolism (49). Methotrexate and glucocorticoids decrease BMD by influencing factors that control osteoblast and osteoclast cell activity (49). Despite common past treatment exposures, pediatric ALL survivors exhibit substantial variation in BMD later in life. An unexplored explanation for some of this uncharacterized variation in BMD is epistasis, where the effect of a locus on a trait is conditional on genotypes observed at other loci.

While studies have investigated pairs of SNPs in select candidate genes with BMD (75, 76), higher order epistasis involving three or more SNPs is also likely to play a vital role in the genetic architecture of BMD. BMD reflects the cumulative effects of interacting genetic and environmental factors on peak bone mass and bone remodeling (77). Signaling pathways requiring both spatiotemporal cues and epigenetic modifications of genetic loci guide the differentiation of bone cells from cells of mesenchymal and hematopoietic origin (78). In a

recent genome-wide scan of SNP pair interactions, over half of gene expressions in peripheral blood significantly associated with SNP pairs were influenced by networks involving three SNPs or more (38).

To our knowledge, no studies have explored higher order epistasis and BMD. In general, searches for epistasis are challenged in identifying true interactions between SNPs on a genome-wide scale, largely due to insufficient statistical power. Novel strategies have been applied to increase power and identify reliable interactions. One strategy is to restrict the search for epistasis to SNPs that are likely to contribute to biological interactions, reducing the number of tested interactions (39, 71). Another strategy is to search for interactions with large effects on phenotypes (38). Lastly, some epistatic interactions failing to meet conservative genome-wide significance thresholds have been shown to be reliable signals through replication (79).

In this study, we combined all of these strategies to identify higher order epistatic interactions that explain some of the variability of treatment effects on BMD among adult survivors of childhood ALL exposed to BMD-diminishing treatments. We leveraged knowledge that SNPs in interacting enhancer and promoter regions modulate gene expression and thus affect phenotypes (28, 31). We applied chromatin state annotations (27) to restrict the search for epistasis to SNPs mapped to putative enhancer or promoter regions. To detect interactions between regulatory regions carrying SNPs associated with BMD (hereafter referred to as "SNP interactions") as potential modifiers of treatment effects, a novel, non-exhaustive statistical algorithm was implemented. Our specific focus was to identify regulatory 3-way SNP interactions associated with BMD in ALL survivors. An independent cohort of cancer survivors was used to replicate candidate regulatory SNP interaction signals as modifiers of treatment effects on BMD.

Supplemental bioinformatics analyses were conducted to characterize replicated SNP interactions.

## 2.2 Subjects and Methods

Study cohorts

Individuals included in this analysis are participants in the St. Jude Lifetime Cohort Study (SJLIFE) (80). Eligible survivors were divided into two cohorts: a discovery cohort of 856 adult survivors of pediatric ALL and a replication cohort consisting of 1428 adult survivors of any non-ALL pediatric cancer (a second cohort of ALL survivors with comparable BMD measurements and genotype data was unavailable). BMD was ascertained using quantitative computed tomography (QCT) from the mid-bodies of the first and second lumbar vertebra. A BMD Z-score was computed for each survivor by taking the difference between the average of their two vertebral BMD measurements and the age- and sex-matched mean of a reference population, divided by the standard deviation in the reference population. Cumulative doses of cranial radiation (none, >0 to <2400, ≥2400 cGy), methotrexate (<5100, ≥5100 to <20000, ≥20000 mg/m$^2$), and glucocorticoid (<2000, ≥2000 to <11000, ≥11000 mg/m$^2$) treatment exposures were considered as risk factors for BMD deficiency among ALL survivors (47-49). We built a multiple linear regression model for BMD Z-scores including sex, categorical treatment exposures, and genetic ancestry estimated using STRUCTURE software (81) (to control for population stratification in our multi-ethnic cohorts) for adjustment in subsequent

genetic association analyses. Additional study cohort details are provided in Supplementary Methods.

Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA) was used to genotype DNA samples in SJLIFE. Genotyping quality control thresholds for exclusion from the analysis in the discovery cohort were as follows: <95% per-sample call rate across markers, <95% SNP call rate across samples, MAF <1%, and Hardy-Weinberg equilibrium p-value $<1x10^{-6}$. Imputation of missing measured genotypes was completed using BEAGLE (59). An allelic $R^2$ imputation quality score cutoff of >0.5 was applied. For replication, a per-sample call rate of >90% was used to avoid sample exclusion due to the smaller set of SNPs selected for follow-up. Otherwise, genotyping quality control thresholds were identical for discovery and replication cohorts. Imputation was not employed for replication to limit data uncertainty associated with imputation. Per-SNP missingness rates were comparable between cohorts (**Table S1**). Genetic data is available in the European Genome-phenome Archive under study accession number EGAS00001002645 (https://www.ebi.ac.uk/ega/studies/EGAS00001002645).

Statistical methodology: Discovery analysis

ChromHMM chromatin state annotations (27) were used to map SNPs to putative enhancer or promoter regions. We retained SNPs mapped to "strong enhancer" or "active promoter" elements in any of nine ChromHMM-annotated human cell types, since it was unknown which types would be most relevant. This restriction limited the search for SNP interactions to 75523 SNPs. Each of these SNPs was dichotomized to create two binary variables, or indicators for carrying at least one non-reference allele or homozygous non-

reference alleles. Any binary-encoded SNP variable with frequency <5% was removed to limit evaluations of sparse 3-SNP interactions. A total of 115800 binary SNP variables were retained genome-wide.

Despite this SNP restriction, exhaustively testing enhancer/promoter 3-SNP interactions would entail >260 trillion tests. To decrease the number of effective tests without compromising the search quality, we developed a non-exhaustive, sequential conditioning algorithm based on logic regression (69) to conduct an effective search of the 3-SNP interaction search space. Briefly, logic regression is an adaptive regression methodology that combines generalized linear models (GLMs) with a stochastic search algorithm to identify best-fitting models that include interaction variables ("logic trees") comprised of binary predictors. Logic regression selects best-fitting models by comparing decrements in model scores. The search for 3-SNP interactions was also restricted by chromosome, reducing the computational burden to a manageable level.

Our algorithm identified 3-SNP interactions sequentially via logic regression for each chromosome, using forward addition to form a linear predictor that included ten 3-SNP interaction trees per chromosome. Models for each chromosome took the following form:

$$E[Y] = \mu + \delta S + \sum_{j=1}^{3} \alpha_j A_j + \sum_{k=1}^{6} \gamma_k T_k + \sum_{p=1}^{m} \beta_p L_p, \quad \text{(Eqn. 1)}$$

where $Y$ is BMD Z-score, $S$ is sex, $A_{1\text{-}3}$ are the three STRUCTURE genetic ancestry covariates, $T_{1\text{-}6}$ are indicator variables for the three categorical treatment variables (i.e., three 3-level variables), and $L_p$ are the 3-SNP interaction trees ($m$=1, 2, …, 10 trees). By identifying 3-SNP interaction trees conditioned on previously identified trees, the algorithm guides the stochastic search in different directions, yielding 3-SNP interactions that are unlikely to be correlated.

We applied a permutation-based approach to identify candidate 3-SNP interactions for replication follow-up. For each of the algorithm-identified 3-SNP interaction trees, 1000

permutations of BMD Z-scores were used to compute the corresponding empirically derived median for the absolute value of the t-statistic and its median absolution deviation (MAD, a robust measure of variability). Permutations of BMD Z-score values were conditioned on 50 quantiles of the fitted BMD Z-score from the clinical baseline model to approximately preserve relationships between adjustment covariates and BMD Z-score. The search algorithm was applied to these conditionally-permuted BMD Z-scores in the exact same manner as the unpermuted case. To select candidate 3-SNP interactions for replication follow-up, we compared the observed t-statistic of a given tree with the corresponding empirically-derived median, similar to the Significance Analysis of Microarray method (82). Our evaluation statistic is $\frac{t_{obs}-t_{med}}{MAD_{t_{med}}}$, where $t_{obs}$ is the absolute value of the t-statistic for the $p^{\text{th}}$ tree given ($p$-*1*) observed trees, and $t_{med}$ and $MAD_{t_{med}}$ are the median and MAD, respectively, of the corresponding 1000 conditioned t-statistics from 1000 permutations. If the evaluation statistic was >2 (i.e., $t_{obs}$ was >2 median absolute deviations away from its corresponding permutation-derived median), the 3-SNP interaction tree was selected as a candidate interaction for replication, as the tree explains an unusually large proportion of BMD Z-score variation than expected by chance alone.

After selecting candidate regulatory 3-SNP interaction trees ("original" trees), a "neighborhood" analysis was conducted to identify "proximal" 3-SNP interaction trees with stronger associations with BMD Z-score than original trees. The reasons for conducting this analysis were twofold: (1) our non-exhaustive logic regression-based algorithm may have missed proximal SNP interactions with stronger associations with BMD; and (2) these strongly associated neighborhood SNP interactions may include SNPs that "tag" additional regulatory regions relevant for BMD. Neighborhood trees were constructed with binary-encoded SNP variables (same filtering criteria as the discovery analysis) from SNPs +/-100 kb of SNPs in the

original tree, with the same Boolean logic structure as the corresponding original tree. Neighborhood trees selected for follow-up in the replication cohort explained larger proportions of BMD Z-score variation than their corresponding original trees.

Statistical methodology: Replication analysis

Since every participant in the discovery cohort received substantial cumulative doses of at least one of the three treatments known to affect BMD, we expected that interaction signals observed in the discovery cohort were potential modifiers of treatment effects on BMD. We therefore defined evidence of replication as significant modification of treatment effects by 3-SNP trees in the replication cohort. We assessed modification of treatment effects using two different approaches: (1) 3-SNP tree interactions with each of the three treatments, and (2) 3-SNP tree main effects among those exposed to each of the three treatments. If the 3-SNP tree had a significant interaction ($P<0.05$) with at least one of the three treatments or a significant main effect ($P<0.05$) among those exposed to one of the treatments, we deemed the interaction to be replicated. We further required treatment modification effects in the replication cohort to have the same direction and similar magnitude as the discovery cohort.

Additional explanation of the statistical methodology is given in Supplementary Methods.

Comparison of the proposed method to a benchmark 2-SNP interaction analysis method

We conducted an exhaustive, within-chromosome 2-way SNP interaction analysis among enhancer/promoter SNPs with the linear regression-based epistasis module in PLINK v1.90, a benchmark methodology for epistasis analysis (34). We also performed a simulation study under three sample size scenarios (N=1000, 1500, and 2000) to compare the performance, measured by power and positive predictive value (PPV), of our proposed method and the benchmark method's detection of component SNP pairs for replicated 3-SNP interactions. Details for both analyses are provided in Supplementary Methods.

Biological characterization of replicated interactions

We evaluated whether there was an excess of significant gene expressions (expression quantitative trait loci or eQTLs) for SNPs in replicated 3-SNP interactions in bone-related cells/tissues using *cis*-eQTLs achieving study-wide significance from the Genotype-Tissue Expression (GTEx) Project (63) and GHS-Express monocyte transcriptome (83) databases. Using the BMD biology literature, we defined 16 cell or tissue groups to be related to bone out of 45 available cell/tissue groups. Counts of significant eQTLs in bone-related cells/tissues for SNPs of interest were compared to all other SNPs genome-wide with at least one significant eQTL in these databases (~2.6 million SNPs with ~26.4 million eQTLs) using a 2-sided Fisher's exact test.

To investigate the cell- and tissue-specificity of enhancer and promoter states for SNPs contributing to replicated interactions, we conducted enrichment analyses using the 15-state chromatin state annotation data for 127 consolidated human cell types from the Roadmap Epigenomics Mapping Consortium (REMC) (62). For each cell type, we compared the set of

SNPs in replicated interactions with the set of non-overlapping SNPs originally mapped to enhancers/promoters. Frequencies of overlap between SNPs in each set and REMC enhancer or promoter regions were counted in each cell type. Strength of evidence for enrichments was evaluated using a 2-sided Fisher's exact test.

Assays based on chromosome conformation capture (3C) enable study of physical interactions between chromatin regions (28, 31). We evaluated the likelihood of physical interaction between SNP regions participating in replicated 3-SNP interactions using a publicly available Hi-C data library generated in lymphoblastoid cells (74), visualized with the WashU EpiGenome Browser resource (64).

Details for bioinformatics analyses are available in Supplementary Methods.

## 2.3 Results

The discovery cohort included 856 adult survivors of pediatric ALL. Cohort clinical characteristics are provided in **Table 2.1**. Every ALL survivor was exposed to cranial radiation therapy (CRT), methotrexate, and/or glucocorticoids during childhood. Our linear regression model with sex, ancestry, and treatment covariates demonstrated that decreases in adjusted mean BMD Z-scores were significantly associated with increasing cumulative dosages for each of these treatments (**Table S2**).

Using the proposed logic regression-based algorithm, we identified 220 3-SNP interactions (10 interactions per chromosome) associated with BMD Z-score. Consistent with previous observations of regulatory complexes involving enhancer-promoter, enhancer-enhancer, or promoter-promoter interactions (25), no restrictions were made on the composition of 3-SNP

interactions. Six distinct (uncorrelated) 3-SNP interactions were selected as candidate interactions for replication follow-up using our permutation-based evaluation statistic threshold (values >2). We considered each of these six distinct 3-SNP interactions separately as genomic "interaction neighborhoods" associated with BMD and looked for other 3-SNP interactions in these "neighborhoods" that were more strongly associated with BMD than the original 3-SNP interactions in the discovery cohort. All 3-way SNP interactions using any SNP located within 100-kb of regulatory loci contributing to the originally selected 3-SNP interactions were assessed. We identified ten additional "neighborhood" 3-SNP interactions that explained larger proportions of BMD Z-score variation than their corresponding original interactions for four of the six selected 3-SNP interactions: this yielded a total of 16 candidate 3-SNP interactions for replication follow-up.

The replication cohort of SJLIFE participants (N=1428) with a range of non-ALL pediatric cancer diagnoses (**Table S3**) was comparable to the discovery cohort with respect to age, sex, and ancestry distributions (**Table 2.1**). Participants in the replication cohort exposed to either CRT or methotrexate received, on average, higher cumulative doses of these treatments compared to the discovery cohort (**Table S4**). Applying our replication definition, 12 of the 16 3-SNP interactions were replicated as modifiers of treatment effects (**Tables S5-S7**). Considering the six originally selected 3-SNP interactions, each reflecting a distinct interaction neighborhood, at least one original or neighborhood 3-SNP interaction candidate was replicated for five of the six selected 3-SNP interaction neighborhoods.

**Table 2.2** shows the best replicated original or neighborhood 3-SNP interaction (defined by replication p-value) detected among the five genomic neighborhoods with replicated interactions. Adjusted changes in mean BMD Z-scores for these five best replicated 3-SNP

interactions in the discovery cohort ranged from -1.30 to +1.77 SD, with regression coefficient t-test-based (naïve) p-values ranging from $2.9 \times 10^{-13}$ to $3.5 \times 10^{-11}$. Four of these 3-SNP interactions included at least one SNP that was not nominally significant. No component SNP pair fully recovered the entire magnitude of association of its respective 3-SNP interaction. In the discovery cohort, the breakdown of the proportions of variance in BMD Z-score explained by the non-genetic covariates (14.5%) and the five best replicated 3-SNP interactions (14.1%) were comparable (**Table S8**).

To compare the performance of our proposed algorithm to a benchmark SNP interaction association testing method, we conducted an exhaustive, within-chromosome pairwise SNP interaction analysis using the 75523 SNPs mapped to putative regulatory regions. Of the nearly 158 million SNP pair combinations considered, seven pairs achieved genome-wide significance (Bonferroni-adjusted $P<3.2 \times 10^{-10}$). None were contributing pairs to any of the 220 3-SNP interactions detected with our search algorithm. Considering all SNP pair results with $P<1.0 \times 10^{-9}$ and the SNP pairs formed by their LD proxy SNPs, none of the 967 original or LD proxy SNP pairs were contributing pairs for any of the 220 3-SNP interactions (**Table S9**). To further distinguish differences in performance between our novel method and the benchmark SNP pair testing method, we conducted a simulation study. Assuming effect sizes observed in our discovery analysis (**Table 2.2**), our proposed method has 18-60% power and 17-49% positive predictive value (PPV) to detect "true" (replicated) 3-SNP interactions in smaller samples (N=1000), with marked improvements in both statistics with modest increments in sample size (**Table S10**). In comparison, the benchmark SNP pair method is appreciably less powerful and has low PPV for detecting component 2-SNP interactions in underlying true 3-SNP interactions,

even with larger sample sizes and under a liberal p-value threshold ($P<1x10^{-5}$) to select top SNP pairs (**Table S11**).

The overall biological plausibility of association with BMD was assessed for the set of 22 unique SNPs contributing to the 12 replicated original and neighborhood interactions. First, we examined gene expression data, specifically eQTL associations achieving study-wide significance in GTEx Project (63) and GHS-Express (83) databases. Our 22-SNP set had a total of 51 significant eQTLs in 17 cells/tissues, of which 40 were observed among 16 cell/tissue types related to bone (enrichment $P=3.6x10^{-4}$, relative to the set of non-overlapping SNPs genome-wide with at least one significant eQTL in any of the 45 queried cell/tissue types) (**Figure 2.1a**; **Tables S12, S13**). Second, we used REMC chromatin state annotation data (62) to examine whether our 22-SNP set was enriched in enhancer or promoter states in each of 127 consolidated cell/tissue groups. We observed suggestive enrichment in overlap between SNPs in our 22-SNP set and putative enhancer states in four cell types relevant to bone biology ($P<0.05$, no Bonferroni adjustment), relative to a background set of 75508 non-overlapping enhancer/promoter SNPs in our original SNP restriction set (**Figure 2.1b**). Consideration of weakly significant enhancer and promoter enrichment analysis results ($P<0.10$, no Bonferroni adjustment; **Tables S14, S15**) suggests the 22-SNP set is relatively enriched for both regulatory states in monocytes and hematopoietic stem cells, which are related to bone metabolism (78). For each of the distinct replicated 3-SNP interactions, chromatin contacts between putative regulatory regions containing the three SNPs of interest appeared supported: at least two chromatin contacts connecting the three target loci were observed, each with proximity scores ≥2 (**Table S16**; **Figures S17-S20**).

The 3-SNP interaction with the strongest evidence of association with BMD was observed between rs1020745 (hg19 chr12:g.53692955G>A; *PFDN5* intronic and *C12orf10* promoter region), rs2110167 (hg19 chr12:g.5734319A>G; *ANO2*, intronic region), and rs10444471 (hg19 chr12:g.4677211G>T; *DYRK4* synonymous coding variant) with an adjusted mean increase in BMD Z-score of 1.72 SD (95% CI: 1.27, 2.17). Both rs10444471 and rs2110167 were more frequently observed in enhancer states in bone-related cell types, whereas rs1020745 overlapped both enhancer and promoter states with relatively high frequencies (**Table 2.3**). Hi-C chromatin interaction maps in lymphoblastoid cells connecting the three SNP regions showed contact selectivity for the rs1020745 locus, with proximity scores indicating nearly 13-fold interaction enrichment with the rs10444471 locus, and over 6-fold interaction enrichment with the rs2110167 locus. Enhancer regions including rs10444471 and rs2110167 may interact distally with a promoter or enhancer region bearing rs1020745, in cell types known to play a role in osteoblast or osteoclast differentiation (**Figure 2.2**). Notably, the rs1020745 locus is known to reside in a region of high linkage disequilibrium (84), implicating several potential gene targets including *SP7*.

## 2.4 Discussion

Previous studies of epistasis have successfully used exhaustive testing methods to assess SNP pair interactions. To detect 3-SNP interactions associated with a complex trait on a genome-wide scale, we implemented a novel, non-exhaustive logic regression-based algorithm among SNPs mapped to regulatory genomic regions. Specifically, our algorithm: (a) focuses on 3-way interactions that plausibly reflect gene regulation events using SNPs mapped to enhancers or

promoters; and (b) considers many epistatic candidates, but only allocates 1 degree-of-freedom for a 3-SNP interaction. The strength of our method is that we use logic regression combined with a conditioning strategy to encourage a multi-directional, stochastic search, bypassing an exhaustive search for 3-way interactions that may miss a true interaction due to lack of statistical power.

Despite known limitations of logic regression (e.g., non-exhaustive searches may miss the "best" interaction solution), we propose our method as a complementary approach to existing exhaustive 2-SNP search methods to detect higher order epistasis. We observed no overlap between top 2-way regulatory SNP interactions identified using a benchmark exhaustive testing method and 3-way regulatory SNP interactions detected with our proposed method. Furthermore, our simulation results revealed that SNP pair searches are ineffective for detecting 3-SNP interaction patterns associated with variations in BMD, unless component 2-SNP interactions have strong associations with phenotype without the inclusion of an additional SNP. These results suggest exhaustive searches for 2-SNP interactions are not universally effective for detecting higher order epistasis, and novel methods to conduct deliberate searches for higher order epistasis are needed.

To safeguard against the reporting of false positive results, we used a permutation-based evaluation statistic to identify candidate 3-SNP interactions, performed a replication analysis, and conducted additional bioinformatics analyses. We identified six regulatory 3-SNP interactions that potentially modify treatment effects on BMD among adult survivors of pediatric ALL. Five of these 3-SNP interactions were replicated as treatment modification effects in an independent sample. Our bioinformatics analyses indicated that SNPs contributing to replicated interactions had both an excess of gene expressions and an enrichment of enhancer states in cell

and tissue types important for bone biology. The plausibility of interactions between regulatory regions bearing target SNP variants was supported by observations of chromatin contacts that occurred in greater frequencies than expected between regions that overlapped SNPs of interest in lymphoblastoid cells. Although these 3-SNP interactions were not functionally validated, our findings represent viable leads in identifying epistatic interactions with cancer treatment-related effects on BMD.

There are multiple ways to interpret these 3-SNP interactions. Given the long-range chromatin interaction data, it is plausible that epistatic networks consisting of three SNPs embedded in regulatory regions that physically interact jointly affect gene expressions that modify BMD in pediatric cancer survivors exposed to specific cytotoxic treatments. For example, among those exposed to methotrexate, the genomic regulatory region bearing rs1020745 could act as a "hub" for the 3-way chromosome 12 genetic interaction, with rs2110167 and rs10444471 acting as supportive regulatory elements to influence the *SP7* locus (rs1020745). *SP7* has previously been reported as a candidate gene affecting bone biology in both adult and pediatric populations (84, 85), and is known to encode an osteogenic transcription factor, Osterix (Osx) (86).

Although a second independent cohort of ALL survivors would be desirable for replication analyses, the availability of a replication cohort of non-ALL survivors, which consisted predominantly of survivors of solid tumors or lymphoma, provided the opportunity to assess whether genetic interactions associated with BMD Z-score in the discovery cohort plausibly modified cancer treatment effects on BMD. Our replication results support the discovery findings and underscore the relative importance of treatment exposures, as these epistatic interactions do not appear to be pathological artifacts specific to ALL. To contextualize

these treatment effect modifications, consider the chromosome 12 interaction. This putative epistatic interaction may modulate the effects of *SP7* and as a consequence, Osx expression levels. Exposure to methotrexate has been linked to decreased Osx expression and significant reductions in osteocyte precursor cells and metaphyseal trabecular bone volume in rats (87). As such, this interaction may counter BMD loss in cancer survivors exposed to methotrexate.

In conclusion, our results demonstrate the feasibility of detecting and replicating higher order interactions between SNPs within putative regulatory regions associated with a complex quantitative trait, using a hypothesis-driven approach. Similar searches can be implemented in other contexts, using known biological interaction mechanisms. Although power to assess larger $n^{th}$-order interactions decreases as the number of participatory SNPs increases, biologically-motivated searches for SNP interaction networks involving more than three SNPs at a time are warranted.

**Table 2.1: Participant Characteristics**

| Characteristic | **Discovery cohort[1] (N=856)** n (%) | **Replication cohort[2] (N=1428)** n (%) |
|---|---|---|
| *Age at BMD measurement (years)* | | |
| Median (range) | 31.3 (18.4-59.7) | 31.6 (18.5-65.9) |
| *Age at diagnosis (years)* | | |
| Median (range) | 5.0 (0.2-19.5) | 9.2 (0-24.8) |
| *Sex* | | |
| Male | 427 (49.9) | 767 (53.7) |
| Female | 429 (50.1) | 661 (46.3) |
| *Treatment profile[3]* | | |
| *Cranial radiation (cGy)* | | |
| Median cumulative dose (range) | 1800 (0-5100) | 0 (0-10600) |
| None | 348 (41.2) | 1235 (86.9) |
| >0 to < 2400 | 215 (25.4) | 14 (1.0) |
| ≥ 2400 | 282 (33.4) | 172 (12.1) |
| *Methotrexate (mg/m$^2$)* | | |
| Median cumulative dose (range) | 5462 (85-83350) | 0 (0-211900) |
| < 5100 | 340 (39.9) | 1306 (91.5) |
| ≥ 5100 to < 20000 | 327 (38.3) | 22 (1.5) |
| ≥ 20000 | 186 (21.8) | 99 (6.9) |
| *Glucocorticoids (mg/m$^2$)* | | |
| Median cumulative dose (range) | 9520 (0-27360) | 0 (0-14460) |
| < 2000 | 328 (38.6) | 1239 (86.8) |
| ≥ 2000 to < 11000 | 355 (41.8) | 160 (11.2) |
| ≥ 11000 | 166 (19.6) | 28 (2.0) |
| *BMD Z-score (expressed in SD)* | | |
| Median (range) | -0.4 (-3.5, 5.4) | -0.2 (-5.5, 6.0) |
| ≤ -1 | 256 (29.9) | 349 (24.4) |
| ≥ 1 | 104 (12.1) | 249 (17.4) |

1. Adult survivors of pediatric acute lymphoblastic leukemia (ALL).
2. Adult survivors of pediatric non-ALL cancers.
3. Discovery cohort: Missing cranial radiation, methotrexate, and glucocorticoid cumulative dosage information for 11, 3, and 7 participants, respectively. Replication cohort: Missing cranial radiation, methotrexate, and glucocorticoid cumulative dosage information for 7, 1, and 1 participant(s), respectively.

**Table 2.2: Replicated 3-SNP interactions associated with BMD Z-score identified by the novel logic regression-based algorithm**

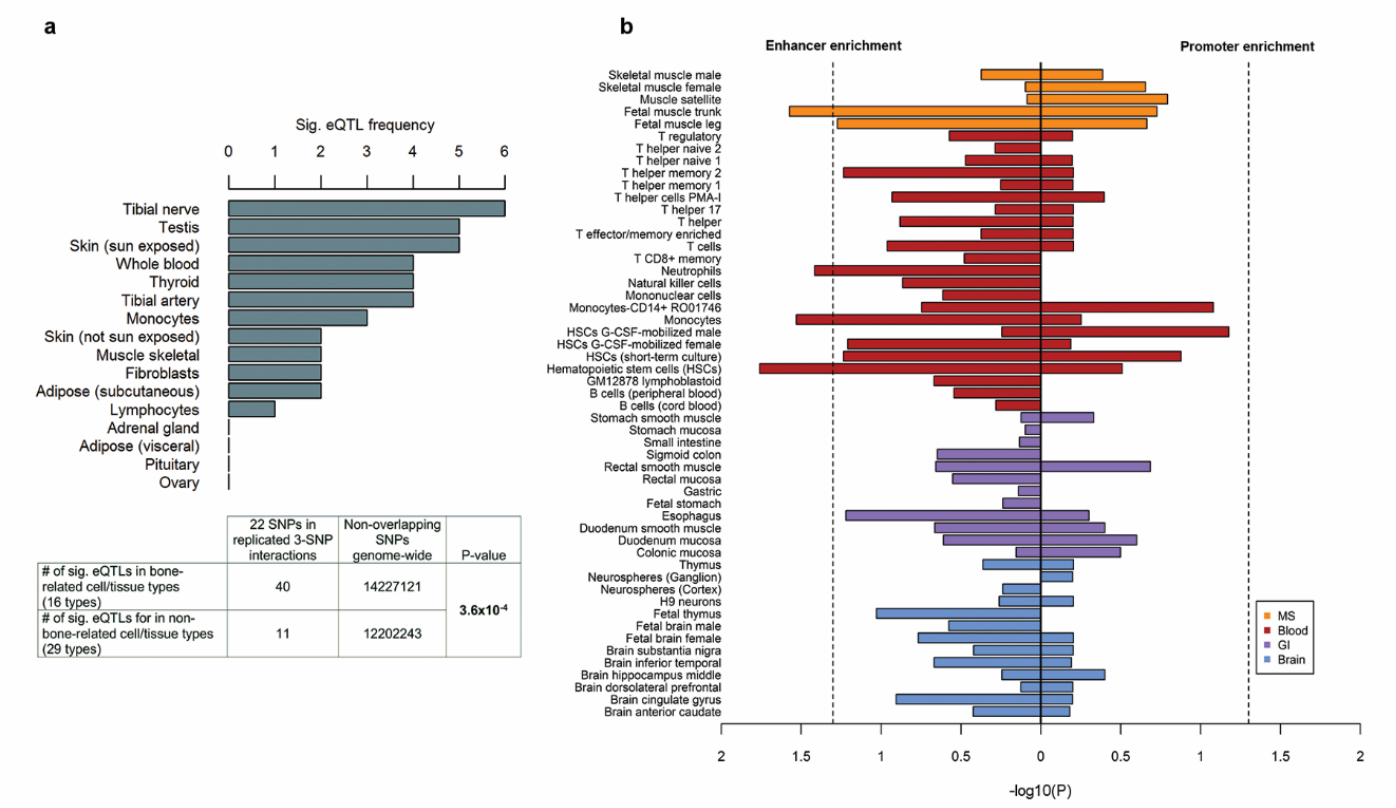
| Chr | Regulatory 3-SNP Interaction[1] | Single SNP rsID (*hg19 HGVS identifier*) | Marginal effect[2] SNP β (P) | Component 2-SNP interaction | Pair effect[2] Pair β (P) | Discovery Cohort SNP interaction β[3] (95% CI) P-value (Tree frequency) | Permutation-based evaluation statistic[4] | Replication Cohort[5] Replicated $\beta_{int}$ P-value (Modified treatment effect) |
|---|---|---|---|---|---|---|---|---|
| 2 | rs901466={CC,CG} or (rs7569568={GG} or rs921319={CC}) | rs901466<br>*chr2:g.114724122C>G* | -0.457<br>(0.002) | rs901466 or rs7569568 | -0.832<br>($5.5x10^{-5}$) | -1.304<br>(-1.668, -0.939)<br>$P=4.7x10^{-12}$<br>(810) | 2.063 | -1.769<br>P=0.004<br>(Methotrexate) |
| | | rs7569568<br>*chr2:g.225264598G>A* | -0.139<br>(0.091) | rs901466 or rs921319 | -0.437<br>(0.004) | | | |
| | | rs921319<br>*chr2:g.62367720T>C* | 0.088<br>(0.405) | rs7569568 or rs921319 | -0.181<br>(0.036) | | | |
| 12 | rs1020745={AG,GG} and (rs2110167={GA,AA} and rs10444471={GG}) | rs1020745<br>*chr12:g.53692955G>A* | 0.589<br>(0.001) | rs1020745 and rs2110167 | 1.215<br>($2.8x10^{-8}$) | 1.719<br>(1.265, 2.174)<br>$P=2.9x10^{-13}$<br>(22) | 2.775 | 1.402<br>P=0.013<br>(Methotrexate) |
| | | rs2110167<br>*chr12:g.5734319A>G* | 0.055<br>(0.461) | rs1020745 and rs10444471 | 0.790<br>($3.5x10^{-5}$) | | | |
| | | rs10444471<br>*chr12:g.4677211G>T* | 0.081<br>(0.538) | rs2110167 and rs10444471 | 0.091<br>(0.220) | | | |
| 12 | (rs1894331={TT} or rs10773093={TC,CC}) and rs4768783={TT,TC} | rs1894331<br>*chr12:g.11930889G>T* | -0.137<br>(0.091) | rs1894331 or rs10773093 | -0.409<br>($1.7x10^{-5}$) | -0.508<br>(-0.649, -0.367)<br>$P=3.1x10^{-12}$<br>(522) | 2.514 | -0.514<br>P=0.042<br>(Methotrexate) |
| | | rs10773093<br>*chr12:g.125046036T>C* | -0.208<br>(0.011) | rs1894331 and rs4768783 | -0.236<br>(0.005) | | | |
| | | rs4768783<br>*chr12:g.47592945C>T* | -0.263<br>(0.005) | rs10773093 and rs4768783 | -0.294<br>($1.0x10^{-4}$) | | | |
| 13 | rs7321815={CC} and (rs9315069={TC,CC} and rs913071={TT,TC}) | rs7321815<br>*chr13:g.101701427C>A* | 0.095<br>(0.211) | rs7321815 and rs9315069 | 1.403<br>($2.5x10^{-9}$) | 1.774<br>(1.283, 2.265)<br>$P=2.9x10^{-12}$<br>(20) | 2.115 | 1.414<br>P=0.046<br>(Methotrexate) |
| | | rs9315069<br>*chr13:g.31371544T>C* | 0.767<br>($2.4x10^{-4}$) | rs7321815 and rs913071 | 0.135<br>(0.076) | | | |
| | | rs913071<br>*chr13:g.36553105C>T* | 0.051<br>(0.666) | rs9315069 and rs913071 | 0.958<br>($2.0x10^{-5}$) | | | |
| 14 | (rs887890={TG,GG} or rs7142110={AA}) or rs1884632={GG} | rs887890<br>*chr14:g.75699438T>G* | 0.199<br>(0.014) | rs887890 or rs7142110 | 0.340<br>($1.1x10^{-5}$) | 0.498<br>(0.352, 0.644)<br>$P=3.5x10^{-11}$<br>(604) | 2.144 | 0.568<br>P=0.008<br>(Cranial radiation) |
| | | rs7142110<br>*chr14:g.51808403G>A* | 0.191<br>(0.016) | rs887890 or rs1884632 | 0.259<br>(0.001) | | | |
| | | rs1884632<br>*chr14:g.69418173C>G* | 0.198<br>(0.013) | rs7142110 or rs1884632 | 0.303<br>($5.5x10^{-5}$) | | | |

1. Shows the single best original or neighborhood interaction detected for each replicated 3-SNP interaction (based on interaction term replication p-values).
2. The marginal SNP and SNP pair linear regression models include the same set of adjustment covariates as the main SNP interaction analysis. Single SNP and SNP pair genotype classes are consistent with genetic effect codings observed in 3-SNP interactions.
3. Estimated mean changes in BMD Z-scores and 95% confidence intervals for 3-SNP interactions in the discovery cohort of ALL survivors, conditioned on previously identified interactions for a given autosome (N=835; participants with missing treatment values were excluded from discovery analysis).
4. Permutation-based evaluation statistics >2 were interpreted as explaining an unusually large proportion of BMD Z-score variation.
5. Any 3-SNP interaction that modified treatment effects (P<0.05) in the replication cohort was considered to be replicated. Methotrexate modification effects were detected among non-ALL diagnosis groups exposed to methotrexate (N=804). CRT modification effects were detected among non-ALL diagnosis groups exposed to CRT (N=1195).

**Table 2.3: Annotations of replicated regulatory 3-SNP interaction trees**

| Tree | dbSNP ID | Allele freq (alt)[1] | Chr position, hg19 (locus) | SNP context | Closest genes (+/- 50kb) | Significant eQTLs[2] | Enhancer in 127 cells/tissues[3] (bone ratio) | Promoter in 127 cells/tissues[3] (bone ratio) | Osteoblasts (E129)[4] | Monocytes (E029)[4] | Lymphoblastoid cells (E116)[4] | Fetal muscle trunk (E089)[4] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs901466:C>G | 0.26 (G) | 114724122 (2q14.1) | intergenic | *LINC01191*; *ACTR3* | *ACTR3*[d], *SLC35F5*[d], *RPL23AP7*[a-d], *DDX11L2*[b-d], *AC024704.2*[b-d], *AC010982.1*[b,d], *AC104653.1*[d] | 30 (2:3) | 0 (NA) | Enhancer | Enhancer | Weak Transcription | Weak Transcription |
| | rs7569568:G>A | 0.16 (A) | 225264598 (2q36.2) | intron | *FAM124B* | *FAM124B*[a] | 6 (1:5) | 20 (7:13) | Weak Repressed PC | Enhancer | Weak Repressed PC | Active TSS |
| | rs921319:T>C | 0.37 (C) | 62367720 (2p15) | intergenic | *COMMD1* | *FAM161A*[d] | 13 (3:10) | 2 (1:1) | Quiescent/Low | Quiescent/Low | Flanking Active TSS | Weak Transcription |
| 2 | rs1020745:G>A | 0.04 (G) | 53692955 (12q13.13) | intron (*PFDN5*); promoter (*C12orf10*) | *PFDN5*; *C12orf10*; *MFSD5*; *ESPL1*; *AAAS*; *SP7* | *AAAS*[c], *C12orf10*[d] | 56 (23:33) | 7 (5:2) | Flanking Active TSS | Flanking Active TSS | Genic Enhancers | Genic Enhancers |
| | rs2110167:A>G | 0.32 (A) | 5734319 (12p13.31) | intron | *ANO2* | NA | 6 (1:1) | 0 (NA) | Quiescent/Low | Quiescent/Low | Quiescent/Low | Enhancer |
| | rs10444471:G>T | 0.05 (T) | 4677211 (12p13.32) | coding (*DYRK4*) | *DYRK4*; *AKAP3*; *RAD51AP1*; *C12orf4* | NA | 13 (7:6) | 2 (1:1) | Quiescent/Low | Enhancer | Enhancer | Weak Repressed PC |
| 3 | rs1894331:G>T | 0.44 (G) | 11930889 (12p13.2) | intron | *ETV6* | NA | 49 (19:30) | 0 (NA) | Enhancer | Genic Enhancers | Genic Enhancers | Enhancer |
| | rs10773093:T>C | 0.43 (C) | 125046036 (12q24.31) | intron | *NCOR2* | NA | 68 (19:49) | 1 (1:0) | Weak Transcription | Enhancer | Flanking Active TSS | Enhancer |
| | rs4768783:C>T | 0.43 (C) | 47592945 (12q13.11) | intron (*PCED1B*) | *PCED1B*; *PCED1B-AS1* | *RP11-493L12.4*[d] | 7 (1:0) | 0 (NA) | Quiescent/Low | Weak Transcription | Quiescent/Low | Quiescent/Low |
| 4 | rs7321815:C>A | 0.35 (A) | 101701427 (13q33.1) | intron (*NALCN-AS1*) | *NALCN-AS1*; *NALCN* | NA | 3 (0:1) | 0 (NA) | Quiescent/Low | Heterochr. | Quiescent/Low | Quiescent/Low |
| | rs9315069:T>C | 0.03 (C) | 31371544 (13q12.3) | intergenic | *ALOX5AP*; *LINC00398* | NA | 38 (5:14) | 11 (2:9) | Flanking Active TSS | Weak Repressed PC | Enhancer | Enhancer |
| | rs913071:C>T | 0.36 (C) | 36553105 (13q13.3) | intron (*DCLK1*) | *DCLK1*; *MIR548F5* | *MAB21L1*[c] | 36 (1:5) | 7 (5:2) | Flanking Active TSS | Quiescent/Low | Quiescent/Low | Weak Transcription |
| 5 | rs887890:T>G | 0.18 (G) | 75699438 (14q24.3) | intergenic | *FOS* | *EIF2B2*[c], *RP11-293M10.1*[d] | 28 (17:11) | 0 (NA) | Weak Transcription | Enhancer | Enhancer | Weak Transcription |
| | rs7142110:G>A | 0.43 (G) | 51808403 (14q22.1) | intron | *LINC00640* | NA | 16 (1:7) | 1 (0:1) | Quiescent/Low | Quiescent/Low | Quiescent/Low | Quiescent/Low |
| | rs1884632:C>G | 0.41 (C) | 69418173 (14q24.1) | Intron (*ACTN1*) | *ACTN1*; *ACTN1-AS1* | *ACTN1*[d] | 83 (27:56) | 11 (3:8) | Genic Enhancers | Enhancer | Weak Transcription | Enhancer |

1. Allele frequency for the alternative/minor allele in the discovery cohort of ALL survivors (N=856); sample alternative/minor allele designations were used for binary SNP variable codings.
2. Significant eQTLs in bone-related cell/tissue groups for a given SNP, obtained from GTEx Portal and GHS-Express databases. Transcript superscripts reflect cell/tissue groups: a. Monocytes; b. Whole blood; c. Muscle skeletal; d. Other endocrinological tissue/pathway.
3. Frequency of overlap between SNP and putative enhancer or promoter state in 127 REMC-annotated epigenomes. "Bone ratio" = (Frequency of overlap with enhancer/promoter state in a bone-related cell/tissue) : (Frequency of overlap with enhancer/promoter state in a cell/tissue not related to bone).
4. Chromatin state annotation for SNPs in the specified cell/tissue type (E*** = epigenome ID), obtained from REMC. Abbreviations: PC = PolyComb; Heterochr. = Heterochromatin; TSS = Transcription Start Site.




| | 22 SNPs in replicated 3-SNP interactions | Non-overlapping SNPs genome-wide | P-value |
|---|---|---|---|
| # of sig. eQTLs in bone-related cell/tissue types (16 types) | 40 | 14227121 | $3.6 \times 10^{-4}$ |
| # of sig. eQTLs for in non-bone-related cell/tissue types (29 types) | 11 | 12202243 | |

**Figure 2.1: Biological plausibility of association between identified SNPs in replicated SNP interactions with BMD**

a. Counts of significant gene expressions (eQTLs) for the 22 unique SNPs in replicated 3-SNP interactions, grouped by the 16 cell or tissue types related to bone (above); the corresponding enrichment analysis result using ~2.6 million non-overlapping genome-wide SNPs with ~26.4 million eQTLs for comparison (below).

b. Plot of Fisher's exact test p-values ($\log_{10}(P)$) from enhancer (left) and promoter (right) enrichment state analyses for the 22 unique SNPs in replicated 3-SNP interactions, using Roadmap Epigenomics Mapping Consortium chromatin state annotations for two BMD-related human cell categories (Musculoskeletal [MS], Blood) and two comparison categories (Gastrointestinal [GI], Brain). Dashed lines correspond to $P<0.05$.

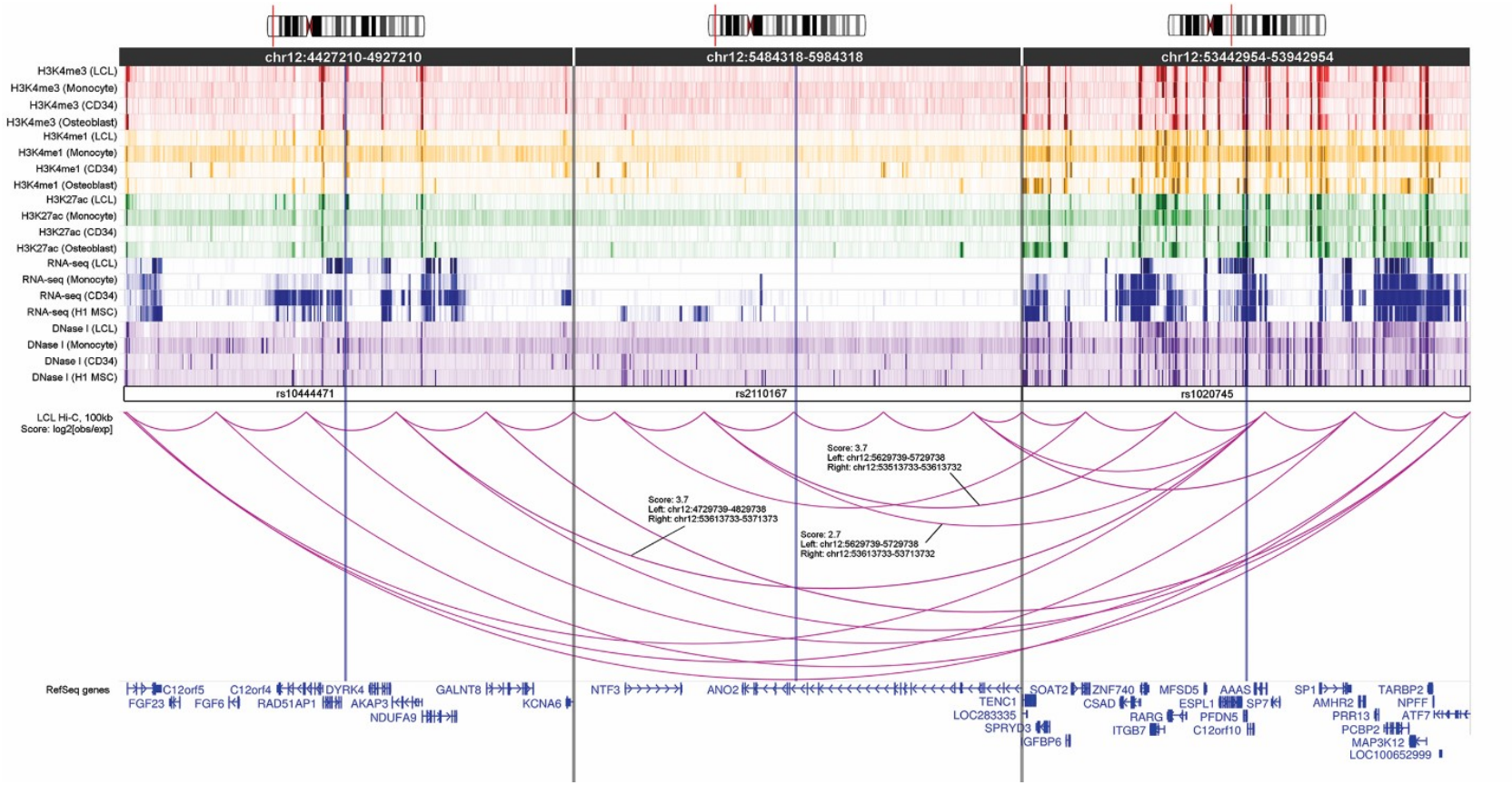



**Figure 2.2: Chromatin interactions for the chromosome 12 SNP interaction: (rs1020745={AG,GG} and (rs2110167={GA,AA} and rs10444471={GG}))**

The WashU EpiGenome Browser was used to visualize long-range chromatin interactions within and across three 500-kb windows centered at implicated SNPs. SNP locations are contextualized using ideograms at the top of regional windows and highlighted with vertical lines in the center of each window. Histone modification (H3K4me3, H3K4me1, H3K27ac), RNA-seq, and DNase I hypersensitivity heatmap data tracks were reviewed. Four data tracks per assay for each of four cell/tissue samples are shown: lymphoblastoid cells (LCLs), peripheral blood mononuclear cells or monocytes, mobilized CD34 cells, and osteoblasts or an osteoblastic precursor proxy (H1 mesenchymal cells). Hi-C data generated with GM06990 LCLs was used to assess evidence for long-range chromatin interactions between SNPs in 3-SNP interaction trees (100-kb bin resolution, log2[observed contact/expected contact] scores). Minimum Hi-C interaction scores were set such that interaction arcs represent chromatin interactions with at least +4-fold observed contact frequency over expected (scores >2).

## 2.5 Supplementary Information

### Chapter 2 Supplementary Methods

Study cohorts

Eligibility criteria for participation in SJLIFE include prior treatment for childhood cancer at St. Jude Children's Research Hospital, ≥10 years post-diagnosis, and age ≥18 years at follow-up. Details regarding SJLIFE and its design are documented elsewhere (Hudson *et al.*, 2014). In the current study, a BMD assessment and DNA sample were also required. Institutional Review Board approval for the current study was received by both St. Jude Children's Research Hospital and the University of Alberta. BMD was measured by quantitative computed tomography (QCT). Specifically, GE VCT Lightspeed 64 detector (GE Healthcare, Milwaukee, WI) and Mindways QCT calibration phantoms and software (Mindways Software Inc., Austin, TX) were used to measure trabecular BMD from the mid-bodies of the first and second lumbar vertebra and compute BMD Z-scores.

An exploratory analysis was conducted to evaluate associations between BMD Z-scores and sex, age at BMD assessment, cranial radiation, methotrexate, glucocorticoids, and ancestry (estimated using STRUCTURE software), and define categorical cancer treatment variables. We assessed adjusted model fits with systematic 2-level and 3-level cumulative treatment dosage cuts for the three treatments of interest. We chose 3-level factor variable definitions for each of these three treatments after observing appreciable decreases in adjusted mean BMD Z-scores within groups of ALL survivors with increasing levels of treatment exposure. Definitions of categorical treatment variables with the most significant adjusted associations with BMD Z-score were chosen. Age was excluded from our set of adjustment covariates; similar to observations made by Gurney *et al.,* 2014, age did not significantly improve model fit after including treatment variables.

Statistical methodology: ChromHMM annotations for SNPs

We used the 15-state ChromHMM annotations of nine primary human ENCODE cell lines (H1 ES, K562, GM12878, HepG2, HUVEC, HSMM, NHLF, NHEK, HMEC) to map SNPs to putative enhancer (states 4 and 5, "strong enhancer") or promoter (state 1, "active promoter") regions (Ernst *et al.*, 2012). The NIH Roadmap Epigenomics Mapping Consortium recently provided chromatin state annotations for 127 human cell types using an updated version of the original 15-state ChromHMM methodology (Kundaje *et al.*, 2015), with comparable chromatin states for promoters (states 1 [Active transcription start site or TSS] and 2 [Flanking active TSS]) and enhancers (states 6 [Genic enhancers] and 7 [Enhancers]). We chose to use the 9-cell annotation for our discovery analysis since the original ChromHMM is trained on a larger complete core set of nine chromatin marks (versus five marks for the update). Additionally, upon comparing the two sets of ChromHMM annotations, we observed that >99% of SNPs (75142) mapped to potential promoters/enhancers using the original ChromHMM annotation

methodology in any of the nine human cell lines were also mapped to potential promoters/enhancers in any of the 127 human cell lines annotated with the updated ChromHMM.

Statistical methodology: Logic regression and detection of regulatory 3-SNP interactions

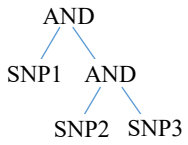
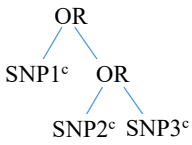
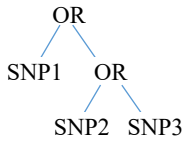
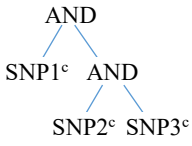
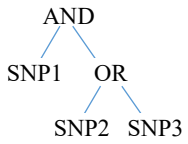
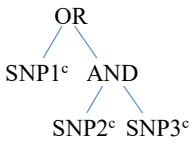
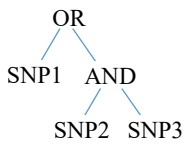
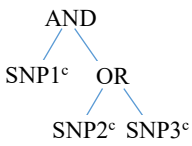
Logic regression (Ruczinski *et al.*, 2003) is a statistical learning method that supports the detection of higher order interactions among binary predictors associated with an outcome of interest within a generalized linear model (GLM) framework. Logic regression has been successfully applied in both genome-wide and candidate gene association analyses to detect SNP interactions associated with a range of complex traits (Dinu *et al.*, 2012).

To efficiently search the interaction search space, logic regression employs a "simulated annealing" stochastic search algorithm to find interaction predictors (implemented in the R "LogicReg" package, version 1.5.8). Starting from a single binary predictor variable, simulated annealing uses a pre-specified set of permissible "moves" (switch, add, delete binary predictor variables, etc.) to build the final interaction predictor or "logic tree". To score candidate models that include logic trees, the simulated annealing algorithm utilizes regression model scores (e.g., residual sum of squares for linear regression). This search algorithm builds and scores models with logic trees until the probability of accepting a new model with a worse model score relative to the current model is low.

In this study, we use logic regression to specifically detect interactions that are combinations of three SNP variables, binary encoded as indicator variables for either carrying at least one non-reference allele or homozygous non-reference alleles, among SNPs mapped to genomic regulatory regions. These binary-encoded SNP variables and their complements (essentially "not" versions of the binary SNP variable, e.g., treating "0" values as "1") are joined by "and" or "or" statements to form 3-SNP interaction predictors or logic trees. The resulting 3-SNP interaction tree is treated in our analysis as a binary Boolean expression: for example, ((enhancer SNP1 or enhancer SNP2) and promoter SNP3) = {True, False}.

Supplementary Table A (below) provides additional context for the types of 3-SNP interactions that are detected by our application of logic regression. Consider SNP1, SNP2, and SNP3 as binary-encoded SNP variables as previously described, with 0="reference genotype" and 1="alternative genotype". There are four major 3-SNP interaction types under the Boolean logic framework; each type has a complement tree. To obtain adjusted estimates of mean BMD Z-score increases/decreases using linear regression for the 3-SNP interaction, we compare individuals with the 3-SNP interaction (logic tree = True) against individuals who carry the complement 3-SNP interaction tree (logic tree = False).

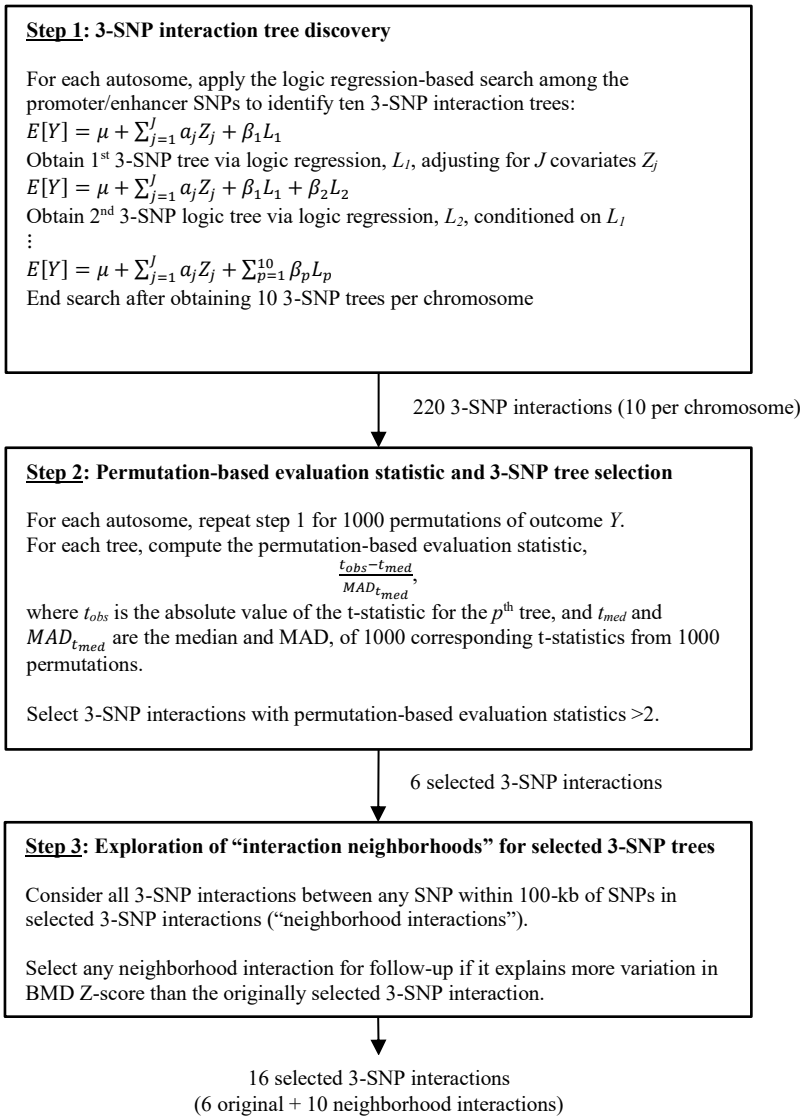
Supplementary Table A: Contextualization of 3-SNP interaction logic trees

| Tree type | 3-SNP interaction tree | Tree type interpretation | Complement tree | Complement interpretation |
|---|---|---|---|---|
| AND/AND |  | All 3 SNPs with "alternative" SNP genotypes jointly contribute to observed effect |  | Same as OR/OR, but with "reference" SNP genotypes (noted by SNP$^c$) |
| OR/OR |  | Any of the 3 SNPs with "alternative" SNP genotypes contributes to observed effect |  | Same as AND/AND, but with "reference" SNP genotypes (noted by SNP$^c$) |
| AND/OR |  | Either (SNP1 and SNP2) or (SNP1 and SNP3) with "alternative" SNP genotypes contribute to observed effect |  | Same as OR/AND, but with "reference" SNP genotypes (noted by SNP$^c$) |
| OR/AND |  | Either SNP1 or (SNP2 and SNP3) with "alternative" SNP genotypes contribute to observed effect |  | Same as AND/OR, but with "reference" SNP genotypes (noted by SNP$^c$) |

All of the diverse SNP interaction logic tree structures (above) are consistent with the notion of biological epistasis, especially when we consider tree complements. For example, for replicated 3-SNP interaction logic trees with only "and" operators (e.g., chromosome 13's 3-SNP tree), all three SNPs' genetic effects are necessary for the observed interaction effect, implying that physical interactions between the corresponding SNP regions may also be necessary. However, evidence of physical proximity between SNP regions for interactions with only "or" operators (e.g., chromosome 2's tree) also supports the plausibility of biological epistasis through physical interaction. Individuals without such "or"-only trees would instead carry the complementary interaction, e.g., an "and"-only tree with SNPs with complement genetic effect encodings (see Supplementary Table A, complement tree for the OR/OR tree type). Thus, SNP interaction logic trees should be considered in both original and complement formulations when assessing supportive evidence of epistasis.

Our proposed method enhances the efficient search of the large 3-SNP interaction space by logic regression. Specifically, we search for interactions among SNPs that have a higher prior probability of jointly regulating gene expression within the same chromosome. We also use a sequential conditioning strategy, where the search for new logic trees is conditioned on previously identified trees, to guide the stochastic search in different directions and yield logic trees that are unlikely to be correlated. To avoid overly rare interaction trees, those observed among <20 subjects were excluded. Lastly, the logic regression search for each tree was conducted under 200 randomly selected initialization values, for which we selected the best model (lowest residual sum of squares) to assure algorithm performance stability. An overview of our proposed method is provided in Supplementary Figure B.

Supplementary Figure B: Overview of the proposed method to detect regulatory 3-SNP interactions




Statistical methodology: Replication analysis

In the replication analysis, we first examined interactions between each of the three treatments and the candidate 3-SNP trees. For each of three treatments of interest, a linear regression model was fit for BMD Z-score ($Y$) in diagnosis groups that included survivors exposed to the treatment of interest within the replication cohort (N=1195 for CRT; N=804 for methotrexate; N=723 for glucocorticoids):

$$E[Y] = \mu + \delta S + \sum_{j=1}^{3} a_j A_j + \sum_{k=1}^{5} \gamma_k T_k + \gamma_{T^*} T^* + \beta_L L + \beta_{int} T^* L. \quad \text{(Eqn. 2)}$$

The same adjustment covariates used in discovery (Eqn. 1) were used for replication; terms for 3-SNP tree main effects ($L$) and the interaction between $L$ and the highest treatment level ($T^*$) for the treatment of interest were included. Second, we fit the same linear regression model (Eqn. 2) without the interaction term, but only among individuals in the replication cohort with any exposure to a given treatment, for each of the three treatments.

Comparison of the proposed method to a benchmark 2-SNP interaction analysis method: Exhaustive SNP pair analysis

The epistasis module implemented in PLINK v1.90 has been cited as a benchmark method for SNP interaction association analyses. Specifically, this implementation can be used to exhaustively assess 2-SNP interaction associations with a quantitative trait $y$ using the linear regression model

$$E[y] = \beta_0 + \beta_1 SNP_1 + \beta_2 SNP_2 + \beta_3 SNP_1 SNP_2,$$

where $SNP_1$ and $SNP_2$ are minor allele counts for SNPs of interest. Strength of evidence for SNP pair association with a quantitative phenotype is based on testing the null hypothesis $\beta_3 = 0$.

Since the benchmark method implemented in PLINK v1.90 does not permit covariate adjustment while exhaustively testing 2-way SNP interactions, we used a 2-stage strategy to allow covariate adjustment: the first stage entailed obtaining residuals by regressing BMD Z-scores on all of the covariates used in our main analysis (sex, treatment exposures, and ancestry) and using these residuals for the exhaustive SNP pair association analysis in the second stage. The final top results ($P<1x10^{-9}$) were confirmed to correspond with top results from the (single-stage) analysis using R v3.1.2, with models with the same covariate adjustment procedure used in our original analysis. We assessed all valid within-chromosome SNP pair combinations between SNPs mapped to putative enhancer/promoter regions within the ALL survivor discovery cohort (N=856).

As a final step, we sought to broaden comparisons between 3-SNP interactions identified by our proposed method and top SNP pairs identified with the benchmark method. In addition to top results (pairs with $P<1x10^{-9}$), we also examined whether SNP pairs that included "LD proxy SNPs" (any SNPs within 25-kb of the queried top pair SNPs with $r^2>0.8$ in 1000 Genomes CEU or YRI) overlapped the algorithm-identified 3-SNP interactions.

Comparison of the proposed method to a benchmark 2-SNP interaction analysis method: Simulation study

We performed simulation analyses to assess the power to identify underlying 3-SNP interactions with our proposed method relative to a benchmark method to detect 2-SNP interactions (PLINK epistasis) which, in this application, is intended to detect 2-SNP components of the underlying 3-SNP interactions. We performed 100 iterations of simulation. In each iteration, we first created SNP sets by randomly sampling a total of 500 SNPs (inclusive of SNPs contributing to the replicated 3-SNP interactions in Table 2.2) mapped to putative promoter and enhancer regions for each chromosome with replicated 3-SNP interactions (chromosomes 2, 12, 13, 14). Using these SNP sets, we created SNP genotype datasets by bootstrapping our subjects' observed genotypes from our ALL survivor discovery cohort for N=1000, 1500, and 2000 sample sizes; bootstrapping the subjects' observed SNP genotypes retains the LD structures of the 500 SNPs in each of the four chromosomes. We then simulated BMD Z-score values for each chromosome based on the replicated 3-SNP interaction logic tree(s), created with the SNP genotype dataset and its effect estimate(s) (Table 2.2), using the following Gaussian model:

$$E[Y] = \beta_0 + \beta_1 L_1 (+\beta_2 L_2) \text{ and } Y \sim Gaussian(0, \sigma^2 = 1),$$

where $Y$ is the BMD Z-score, $\beta$'s are regression coefficients (effect estimates of the 3-SNP interaction trees), and $L_i$'s are 0/1 indicator values for the 3-SNP interaction tree(s) (up to two 3-SNP trees were included in the model for simulated BMD Z-score, since chr12 had two replicated 3-SNP trees). Since $Y$ is a Z-score, the variance of the Gaussian model was set to 1.

We then applied our proposed method to each simulated dataset to sequentially identify three 3-SNP logic trees per chromosome and used the same permutation-based evaluation statistic to select best 3-SNP interactions (values >2), with 50 permutations per tree. For comparison, we used PLINK to conduct the exhaustive 2-SNP interaction analyses among the 500 sampled SNPs for each chromosome and considered two levels of significance to select best 2-SNP interactions: suggestive ($P<1x10^{-5}$) and genome-wide ($P<3.2x10^{-10}$, based on a hypothetical exhaustive analysis of 75523 regulatory SNPs, Bonferroni-adjusted).

One hundred iterations of this simulation were conducted under each of the three sample size scenarios (N=1000, 1500, 2000). Power for the proposed and benchmark methods was calculated as the proportion of the 100 iterations in which the underlying replicated 3-SNP interaction of interest was detected (exact logic tree match exceeding the evaluation statistic threshold) or any of its contributing 2-SNP components was detected under the two pre-specified levels of significance. Positive predictive value (PPV) for each of these methods was computed as the proportion of 3-SNP or component 2-SNP interaction detections among selected or "significant" interactions.

Biological characterization of replicated interactions: eQTL enrichment in bone-related cell/tissue types

We assessed the relevance of the 45 cell/tissue types available in the two queried eQTL databases to BMD initially with landmark BMD GWAS literature (e.g., Estrada *et al*., 2012) and key bone biology literature (e.g., Seeman 2002; Takayanagi 2007; Wasilewski-Masker *et al.*, 2008). We examined citing articles for these selected papers and used PubMed searches with a limited set of BMD-related search terms ("bone mineral density", "bone mass", "osteoblast", "osteoclast", "skeletal homeostasis") in conjunction with the cell/tissue type of interest to assess the relationship between various cell/tissue types and bone. For references used to identify bone-related cells/tissues, see Supplementary Table C.

Supplementary Table C: Sample references for 16 cells/tissues related to bone biology

| **Tissue/Cell Line** | **PubMed references (PMID)** |
|---|---|
| Adipose (subcutaneous and visceral) | 21676245, 18854943, 10660043 |
| Adrenal | 15180950, 24418120 |
| Tibial artery and nerve | 22473330, 9041062 |
| Lymphocytes | 17380158, 18455228, 17202317 |
| Fibroblasts | 19710666, 7816067 |
| Muscle skeletal | 26453495, 26453500 |
| Ovary | 20637179, 7816067 |
| Pituitary | 9494780, 18310191 |
| Skin (sun exposed and not sun exposed) | 15585788, 22414785 |
| Testis | 15180950, 7816067 |
| Thyroid | 22634735 |
| Whole blood | 9032749, 17380158 |
| Monocytes | 7816067, 17380158, 2169622 |

To evaluate the enrichment of significant eQTLs for SNPs in replicated 3-SNP interactions in bone-related cells/tissues, we considered 26429415 significant *cis*-eQTLs for 2552384 and 30140 non-overlapping genome-wide SNPs in the GTEx and GHS-Express databases, respectively.

Biological characterization of replicated interactions: Promoter-/enhancer-state enrichment in 127 human cell types

Under the NIH Roadmap Epigenomics Mapping Consortium's 15-state ChromHMM methodology (Kundaje *et al.*, 2015), the comparable chromatin states applied in enrichment analyses for promoters were states 1 (Active transcription start site or TSS) and 2 (Flanking active TSS), and states 6 (Genic enhancers) and 7 (Enhancers) for enhancers.

Biological characterization of replicated interactions: Chromatin interactions

Using the WashU EpiGenome Browser, we examined Hi-C data generated from Broad/MIT/UMass GM06990 lymphoblastoid cells (100-kb bin resolution) (Lieberman-Aiden *et al.*, 2009) for evidence of physical chromatin interactions within and across 500-kb windows centered around SNPs in replicated 3-SNP interaction trees. For all figures provided, minimum Hi-C interaction scores ($\log_2$[observed contact/expected contact] scores) were set to +2 to retain interaction arcs representing chromatin interactions with at least +4-fold observed contact frequency over expected. In conjunction with chromatin interactions, we visualized epigenome and transcriptome heatmap data, using data populated from cell and tissue samples previously identified as relevant in prior studies of BMD (Supplementary Table C). Specifically, we examined RNA-seq, histone modification (H3K4me1, H3K27ac, and H3K4me3 marks), and DNase I hypersensitivity assay data tracks from ENCODE, using data populated from lymphoblastoid cells, osteoclastic precursors (peripheral blood monocytes and mobilized CD34 cells), and osteoblasts or an osteoblastic precursor (H1 mesenchymal cells).

## Chapter 2 Supplementary Tables and Figures

**Table S1**: Summary of per-SNP proportion missingness, in ALL discovery and non-ALL replication cohorts, without BEAGLE imputation

| | Summary statistics, per-SNP call rates | | |
|---|---|---|---|
| | Enhancer/promoter SNPs ALL discovery (N=856) | 22 unique SNPs in replicated 3-SNP interactions, ALL discovery (N=856) | 22 unique SNPs in replicated 3-SNP interactions, non-ALL replication (N=1428) |
| Median, proportion missing | 0.0012 | 0.0012 | 0.0007 |
| IQR, proportion missing | 0.0035 | 0.0047 | 0.0053 |
| Max, proportion missing | 0.0491 | 0.0292 | 0.0364 |

**Table S2**: Linear regression models for BMD Z-score with adjustment covariates

| Characteristic | **Discovery cohort (N=856; ALL)** Est (P)[a] | **Replication cohort (N=1428; Non-ALL)** Est (P)[b] |
|---|---|---|
| *Sex* | | |
| Male | | |
| Female | 0.59 ($4.5 \times 10^{-15}$) | 0.33 ($1.4 \times 10^{-7}$) |
| *Ancestry* | | |
| White (CEU) | | |
| Black (AFR) | 1.03 ($1.2 \times 10^{-9}$) | 1.13 ($1.6 \times 10^{-39}$) |
| East Asian (CHB/JPT) | 0.06 (0.91) | -0.51 (0.48) |
| South Asian (SAS) | -0.35 (0.59) | -0.24 (0.66) |
| *Treatment profile[(a)]* | | |
| *Cranial radiation (cGy)* | | |
| None | | |
| >0 to < 2400 | -0.09 (0.40) | -0.18 (0.57) |
| ≥ 2400 | -0.30 (0.01) | -0.78 ($1.8 \times 10^{-15}$) |
| *Methotrexate ($mg/m^2$)* | | |
| < 5100 | | |
| ≥ 5100 to < 20000 | -0.23 (0.03) | 0.36 (0.15) |
| ≥ 20000 | -0.44 ($2.9 \times 10^{-3}$) | -0.18 (0.16) |
| *Glucocorticoids ($mg/m^2$)* | | |
| < 2000 | | |
| ≥ 2000 to < 11000 | -0.29 ($7.2 \times 10^{-4}$) | -0.03 (0.78) |
| ≥ 11000 | -0.33 ($4.3 \times 10^{-3}$) | -0.17 (0.46) |

a. OLS parameter estimates for adjustment covariates in the discovery cohort.
b. OLS parameter estimates for adjustment covariates in the replication cohort.

**Table S3**: Distribution of childhood cancer diagnoses in the replication cohort of adult survivors (N=1428)

| Diagnosis groups | n (%) |
|---|---|
| Acute myeloid leukemia | 72 (5.0) |
| Central nervous system tumors | 208 (14.6) |
| Ewing sarcoma | 61 (4.3) |
| Hodgkin lymphoma | 240 (16.8) |
| Neuroblastoma | 90 (6.3) |
| Non-Hodgkin lymphoma | 172 (12.0) |
| Osteosarcoma | 93 (6.5) |
| Retinoblastoma | 75 (5.3) |
| Rhabdomyosarcoma | 76 (5.3) |
| Wilms tumor | 132 (9.2) |
| Other | 209 (14.6) |

**Table S4**: Treatment distributions in discovery and replication cohorts

| Characteristic | Discovery Cohort (ALL survivors, N=856) | Replication Cohort (Non-ALL survivors, N=1428) |
|---|---|---|
| *Cranial radiation (CRT)* | | |
| Received any cranial radiation (n) | 497 | 186 |
| Mean cumulative dose (cGy) among those who received CRT (range) | 2219 (590-5100) | 4807 (600-10600) |
| *Methotrexate (MTX)* | | |
| Received any methotrexate (n) | 853 | 416 |
| Mean cumulative dose ($mg/m^2$) among those who received MTX (range) | 10390 (85-83350) | 19950 (6-211900) |
| *Glucocorticoids (GC)* | | |
| Received any glucocorticoids (n) | 848 | 336 |
| Mean cumulative dose *(*$mg/m^2$*)* among those who received GC (range) | 6948 (33-27360) | 3099 (106-14460) |

**Table S5**: dbSNP and HGVS-based SNP identifiers for 28 unique SNPs contributing to 3-SNP interaction trees selected for replication follow-up

| rsID* | dbSNP/HGVS-based identifier |
|---|---|
| **rs901466** | hg19 chr2:g.114724122C>G |
| **rs7569568** | hg19 chr2:g.225264598G>A |
| **rs921319** | hg19 chr2:g.62367720T>C |
| **rs7569573** | hg19 chr2:g.225264619G>A |
| **rs2122382** | hg19 chr2:g.62326484C>A |
| **rs17407839** | hg19 chr2:g.225332677A>C |
| **rs6708208** | hg19 chr2:g.62274123G>A |
| rs10893935 | hg19 chr11:g.128811507C>T |
| rs7114794 | hg19 chr11:g.107993500A>G |
| rs10896438 | hg19 chr11:g.68906570T>G |
| rs2924528 | hg19 chr11:g.68926593T>C |
| **rs1020745** | hg19 chr12:g.53692955G>A |
| **rs2110167** | hg19 chr12:g.5734319A>G |
| **rs10444471** | hg19 chr12:g.4677211G>T |
| **rs1894331** | hg19 chr12:g.11930889G>T |
| **rs10773093** | hg19 chr12:g.125046036T>C |
| **rs4768783** | hg19 chr12:g.47592945C>T |
| rs10881072 | hg19 chr12:g.47585115G>A |
| **rs7321815** | hg19 chr13:g.101701427C>A |
| **rs9315069** | hg19 chr13:g.31371544T>C |
| **rs913071** | hg19 chr13:g.36553105C>T |
| **rs887890** | hg19 chr14:g.75699438T>G |
| **rs7142110** | hg19 chr14:g.51808403G>A |
| **rs1884632** | hg19 chr14:g.69418173C>G |
| rs4899553 | hg19 chr14:g.75698304C>T |
| **rs4901111** | hg19 chr14:g.51822151A>T |
| **rs1884633** | hg19 chr14:g.69434083G>A |
| **rs4901112** | hg19 chr14:g.51822356T>G |

*__Bolded__: 22 unique SNPs contributing to replicated 3-SNP interactions

**Table S6**: Discovery results for all 16 original and neighborhood 3-SNP interaction trees

| Chr | Tree Number (Original/ Neighborhood)[a] | 3-SNP interaction trees | Permutation-based evaluation statistic | Tree Frequency[b] | Conditioned change in BMD Z-score (95% CI)[c] | P | Unconditioned change in BMD Z-score (95% CI)[d] | P | Replicated[e] |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 8 (Original) | rs901466={CC,CG} or (rs7569568={GG} or rs921319={CC}) | 2.063 | 831 | -1.304 (-1.668,-0.939) | $4.7x10^{-12}$ | -1.030 (-1.472,-0.588) | $5.4x10^{-6}$ | Yes |
| 2 | (Neighborhood) | rs901466={CC,CG} or (rs7569568={GG} or rs2122382={AA}) | | 834 | | | -1.077 (-1.539,-0.616) | $5.3x10^{-6}$ | Yes |
| 2 | (Neighborhood) | rs901466={CC,CG} or (rs7569573={GG} or rs2122382={AA}) | | 834 | | | -1.077 (-1.539,-0.616) | $5.3x10^{-6}$ | Yes |
| 2 | (Neighborhood) | rs901466={CC,CG} or (rs7569573={GG} or rs921319={CC}) | | 831 | | | -1.030 (-1.472,-0.588) | $5.4x10^{-6}$ | Yes |
| 2 | (Neighborhood) | rs901466={CC,CG} or (rs17407839={AA} or rs6708208={AA}) | | 831 | | | -1.028 (-1.461,-0.596) | $3.6x10^{-6}$ | Yes |
| 2 | (Neighborhood) | rs901466={CC,CG} or (rs17407839={AA} or rs2122382={AA}) | | 831 | | | -1.028 (-1.461,-0.596) | $3.6x10^{-6}$ | Yes |
| 11 | 5 (Original) | (rs10893935={TT} and rs7114794={AA}) and rs10896438={TG,GG} | 2.433 | 24 | 1.463 (1.058,1.869) | $3.0x10^{-12}$ | 1.466 (1.013,1.919) | $3.5x10^{-10}$ | No |
| 11 | (Neighborhood) | (rs10893935={TT} and rs7114794={AA}) and rs2924528={TC,CC} | | 27 | | | 1.383 (0.967,1.799) | $1.2x10^{-10}$ | No |
| 12 | 2 (Original) | rs1020745={AG,GG} and (rs2110167={GA,AA} and rs10444471={GG}) | 2.775 | 22 | 1.719 (1.265,2.174) | $2.9x10^{-13}$ | 1.771 (1.301,2.241) | $3.4x10^{-13}$ | Yes |
| 12 | 5 (Original) | (rs1894331={TT} or rs10773093={TC,CC}) and rs10881072={AA,AG} | 2.514 | 529 | -0.508 (-0.649,-0.367) | $3.1x10^{-12}$ | -0.438 (-0.599,-0.277) | $1.2x10^{-7}$ | No |
| 12 | (Neighborhood) | (rs1894331={TT} or rs10773093={TC,CC}) and rs4768783={TT,TC} | | 522 | | | -0.422 (-0.576,-0.269) | $8.8x10^{-8}$ | Yes |
| 13 | 2 (Original) | rs7321815={CC} and (rs9315069={TC,CC} and rs913071={TT,TC}) | 2.115 | 20 | 1.774 (1.283,2.265) | $2.9x10^{-12}$ | 1.727 (1.220,2.235) | $4.3x10^{-11}$ | Yes |
| 14 | 7 (Original) | (rs4899553={CT,TT} or rs7142110={AA}) or rs1884632={GG} | 2.144 | 654 | 0.498 (0.352,0.644) | $3.5x10^{-11}$ | 0.442 (0.270,0.614) | $5.6x10^{-7}$ | No |
| 14 | (Neighborhood) | (rs887890={TG,GG} or rs7142110={AA}) or rs1884632={GG} | | 604 | | | 0.438 (0.279,0.598) | $9.3x10^{-8}$ | Yes |
| 14 | (Neighborhood) | (rs887890={TG,GG} or rs4901111={AA}) or rs1884633={AA} | | 561 | | | 0.410 (0.255,0.565) | $2.7x10^{-7}$ | Yes |
| 14 | (Neighborhood) | (rs887890={TG,GG} or rs4901112={TT}) or rs1884633={AA} | | 562 | | | 0.400 (0.245,0.556) | $5.1x10^{-7}$ | Yes |

Ten neighborhood trees that explained larger proportions of BMD Z-score variation than their original corresponding tree were identified for four out of six 3-SNP interactions with permutation-based evaluation statistics >2.

a. "Original" 3-SNP interaction trees were detected using the sequential logic regression algorithm. The tree number (column 2) indicates the order in which the tree was identified using the sequential conditioning algorithm, e.g., tree number 8 = 8th detected tree, conditioned on 7 previously detected trees for the same autosome.

b. 3-SNP interaction frequencies observed in entire ALL discovery cohort (N=856).

c. Adjusted OLS betas representing mean changes in BMD Z-scores (in standard deviations) and respective 95% confidence intervals for the $p^{th}$ tree given *(p-1)* observed trees for a given autosome, using individuals with complete treatment data and genotype data with imputed missing measured SNPs (N=835).

d. Same estimates as (b), but without conditioning on previously observed trees for a given autosome, using genotype data without imputation of missing measured SNPs (N=856).

e. Replication result; see Supplementary Table S7 for details.

**Table S7**: Replication results for all 16 original and neighborhood 3-SNP interaction trees

| | | | | | | | Replication, Cranial Radiation (CRT) | | | | | Replication, Methotrexate (MTX) | | | | | Replication, Glucocorticoids (GC) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Discovery cohort (N=835) | | a. Tree main effects in replication (N=1428) | | | b. CRT dx groups only (N=1195) Tree x High CRT | | | c. CRT only (N=186) Tree main effects | | d. MTX dx groups only (N=804) Tree x High MTX | | | e. MTX only (N=121) Tree main effects | | f. GC dx groups only (N=723) Tree x High GC | | | g. GC only (N=188) Tree main effects | |
| Chr | 3-SNP interaction tree | Cond Beta | Cond P | Freq | Beta (Tree) | P | Beta (Tree) | Beta (TxCRT) | $P_{int}$ | Beta (Tree) | P | Beta (Tree) | Beta (TxMTX) | $P_{int}$ | Beta (Tree) | P | Beta (Tree) | Beta (TxGC) | $P_{int}$ | Beta (Tree) | P |
| 2 | rs901466={CC,CG} or (rs7569568={GG} or rs921319={CC}) | -1.304 | $4.7x10^{-12}$ | 1362 | -0.138 | 0.385 | 0.053 | -0.217 | 0.643 | 0.102 | 0.854 | **0.097** | **-1.769** | **0.004** | **-1.920** | **$1.3x10^{-4}$** | 0.193 | 0.006 | 0.983 | 0.197 | 0.713 |
| 2 | rs901466={CC,CG} or (rs7569568={GG} or rs2122382={AA}) | | | **1374** | **-0.418** | **0.015** | -0.184 | -0.879 | 0.078 | -1.023 | 0.078 | -0.128 | -0.785 | 0.266 | **-1.316** | **0.017** | -0.358 | 0.020 | 0.945 | 0.024 | 0.969 |
| 2 | rs901466={CC,CG} or (rs7569573={GG} or rs2122382={AA}) | | | **1374** | **-0.418** | **0.015** | -0.184 | -0.879 | 0.078 | -1.023 | 0.078 | -0.128 | -0.785 | 0.266 | **-1.316** | **0.017** | -0.358 | 0.020 | 0.945 | 0.024 | 0.969 |
| 2 | rs901466={CC,CG} or (rs7569573={GG} or rs921319={CC}) | | | 1363 | -0.177 | 0.266 | 0.047 | -0.540 | 0.274 | -0.336 | 0.561 | **0.097** | **-1.769** | **0.004** | **-1.920** | **$1.3x10^{-4}$** | 0.100 | 0.008 | 0.978 | 0.197 | 0.713 |
| 2 | rs901466={CC,CG} or (rs17407839={AA} or rs6708208={AA}) | | | **1364** | **-0.388** | **0.018** | -0.202 | -0.917 | 0.082 | -1.061 | 0.100 | -0.312 | -0.503 | 0.419 | **-1.153** | **0.024** | -0.403 | 0.023 | 0.938 | 0.003 | 0.997 |
| 2 | rs901466={CC,CG} or (rs17407839={AA} or rs2122382={AA}) | | | **1359** | **-0.401** | **0.012** | -0.224 | -0.648 | 0.146 | -0.744 | 0.165 | -0.349 | -0.479 | 0.441 | **-1.153** | **0.024** | -0.522 | 0.028 | 0.925 | -0.407 | 0.518 |
| 11 | (rs10893935={TT} and rs7114794={AA}) and rs10896438={TG,GG} | 1.463 | $3.0x10^{-12}$ | 35 | -0.155 | 0.443 | -0.143 | 0.741 | 0.316 | 0.687 | 0.377 | -0.492 | 0.064 | 0.924 | -0.349 | 0.495 | -0.007 | -0.477 | 0.717 | -0.306 | 0.561 |
| 11 | (rs10893935={TT} and rs7114794={AA}) and rs2924528={TC,CC} | | | 50 | -0.278 | 0.100 | -0.226 | -0.071 | 0.896 | -0.224 | 0.688 | -0.409 | -0.017 | 0.979 | -0.349 | 0.495 | -0.224 | -0.261 | 0.841 | -0.192 | 0.683 |
| 12 | rs1020745={AG,GG} and (rs2110167={GA,AA} and rs10444471={GG}) | 1.719 | $2.9x10^{-13}$ | 76 | 0.069 | 0.643 | -0.037 | 0.636 | 0.117 | 0.272 | 0.553 | **-0.322** | **1.402** | **0.013** | 0.847 | 0.097 | -0.016 | 0.225 | 0.808 | -0.229 | 0.570 |
| 12 | (rs1894331={TT} or rs10773093={TC,CC}) and rs10881072={AA,AG} | -0.508 | $3.1x10^{-12}$ | 849 | 0.050 | 0.491 | 0.143 | -0.279 | 0.180 | 0.143 | 0.555 | 0.144 | -0.484 | 0.064 | -0.124 | 0.599 | 0.157 | 0.206 | 0.688 | 0.043 | 0.810 |
| 12 | (rs1894331={TT} or rs10773093={TC,CC}) and rs4768783={TT,TC} | | | 828 | 0.099 | 0.159 | **0.200** | **-0.414** | **0.049** | -0.062 | 0.792 | **0.211** | **-0.514** | **0.042** | -0.162 | 0.483 | 0.209 | 0.151 | 0.763 | 0.178 | 0.303 |
| 13 | rs7321815={CC} and (rs9315069={TC,CC} and rs913071={TT,TC}) | 1.774 | $2.9x10^{-12}$ | 57 | -0.016 | 0.928 | -0.018 | 0.573 | 0.318 | 0.388 | 0.580 | **-0.138** | **1.414** | **0.046** | 0.840 | 0.153 | NA | NA | NA | 0.201 | 0.628 |
| 14 | (rs4899553={CT,TT} or rs7142110={AA}) or rs1884632={GG} | 0.498 | $3.5x10^{-11}$ | 1095 | -0.105 | 0.167 | -0.106 | 0.432 | 0.070 | 0.293 | 0.230 | -0.179 | -0.058 | 0.835 | -0.139 | 0.552 | -0.021 | -0.084 | 0.866 | 0.025 | 0.886 |
| 14 | (rs887890={TG,GG} or rs7142110={AA}) or rs1884632={GG} | | | 983 | -0.091 | 0.184 | **-0.138** | **0.568** | **0.008** | **0.480** | **0.029** | -0.159 | -0.143 | 0.587 | -0.110 | 0.612 | -0.058 | 0.121 | 0.805 | -0.055 | 0.742 |
| 14 | (rs887890={TG,GG} or rs4901111={AA}) or rs1884633={AA} | | | 904 | -0.021 | 0.755 | -0.022 | 0.374 | 0.067 | **0.462** | **0.030** | -0.045 | -0.372 | 0.147 | -0.279 | 0.191 | 0.123 | -0.787 | 0.106 | -0.039 | 0.818 |
| 14 | (rs887890={TG,GG} or rs4901112={TT}) or rs1884633={AA} | | | 909 | -0.031 | 0.647 | -0.035 | 0.386 | 0.059 | **0.462** | **0.030** | -0.055 | -0.362 | 0.157 | -0.279 | 0.191 | 0.103 | -0.767 | 0.114 | -0.081 | 0.625 |

**Replication result a**: Mean changes in BMD Z-scores for each 3-SNP interaction tree (OLS beta, main effects) in the entire non-ALL replication cohort, adjusted for the same covariates used in the discovery analysis.

**Replication results b, d, f**: 3-SNP tree modification of ALL treatment effects as (Tree x High treatment) interactions in treatment-exposed diagnosis groups, adjusted for the same covariates used in the discovery analysis. "High" treatment levels correspond to: ≥ 2400 cGy of CRT; ≥ 20000 mg/m$^2$ of methotrexate (MTX); and ≥ 11000 mg/m$^2$ of glucocorticoids (GC).

**Replication results c, e, g**: 3-SNP tree main effects within replication subsamples restricted to participants exposed to any dose of the cancer treatment of interest, adjusted for the same covariates used in the discovery analysis.

**Table S8**: Variance of BMD Z-score explained by non-genetic covariates and replicated 3-SNP trees in discovery cohort (N=800[a])

| **Model variables**[b] | **Df** | **Sum Sq.** | **F value** | **F-test p-value** | **% Var(BMD) explained** |
|---|---|---|---|---|---|
| Non-genetic covariates | 10 | 145.8 | 15.9 | $2.9\times10^{-26}$ | 14.49 |
| Chr 2: rs901466={CC,CG} or (rs7569568={GG} or rs921319={CC}) | 1 | 27.0 | 29.5 | $7.4\times10^{-8}$ | 2.69 |
| Chr 14: (rs887890={TG,GG} or rs7142110={AA}) or rs1884632={GG} | 1 | 28.8 | 31.5 | $2.8\times10^{-8}$ | 2.86 |
| Chr 12: rs1020745={AG,GG} and (rs2110167={GA,AA} and rs10444471={GG}) | 1 | 39.2 | 42.8 | $1.1\times10^{-10}$ | 3.89 |
| Chr 12: (rs1894331={TT} or rs10773093={TC,CC}) and rs4768783={TT,TC} | 1 | 28.3 | 30.9 | $3.8\times10^{-8}$ | 2.81 |
| Chr 13: rs7321815={CC} and (rs9315069={TC,CC} and rs913071={TT,TC}) | 1 | 18.8 | 206 | $6.7\times10^{-6}$ | 1.87 |
| Residuals | 784 | 718.3 | - | - | 71.38 |

a. Participants with missing values were excluded.
b. Order of model entry for genetic covariates corresponds with level of statistical significance observed for replication test results.

**Table S9**: Top adjusted SNP pair associations ($P<1x10^{-9}$) with BMD Z-score in discovery cohort (N=856), using a benchmark exhaustive SNP pair testing method (PLINK epistasis)

| Chr | SNP1 | SNP1 risk allele | SNP2 | SNP2 risk allele | $Beta_{interaction}$ | $P_{interaction}$ | Number of LD SNP proxy pairs | Number of SNP pairs (original or proxy) included in any of 10 algorithm-detected 3-SNP interactions per autosome |
|---|---|---|---|---|---|---|---|---|
| 5 | rs2676240 | G | rs12188727 | C | 2.219 | **$2.6x10^{-11}$** | 50 | 0 |
| 8 | rs16873180 | C | rs17700442 | T | 2.961 | **$2.9x10^{-11}$** | 0 | 0 |
| 8 | rs11784193 | C | rs1019960 | A | 4.932 | **$4.3x10^{-11}$** | 15 | 0 |
| 7 | rs7781067 | G | rs17158763 | T | 2.206 | **$5.8x10^{-11}$** | 363 | 0 |
| 7 | rs7780759 | G | rs17158763 | T | 2.208 | **$5.9x10^{-11}$** | 363 | 0 |
| 8 | rs965670 | A | rs17700442 | T | 2.638 | **$1.6x10^{-10}$** | 4 | 0 |
| 9 | rs10810585 | G | rs12115310 | G | 1.757 | **$2.3x10^{-10}$** | 23 | 0 |
| 7 | rs12532970 | A | rs983926 | T | 1.925 | $3.3x10^{-10}$ | 8 | 0 |
| 2 | rs16843822 | A | rs10166654 | A | 1.588 | $4.6x10^{-10}$ | 71 | 0 |
| 2 | rs16985851 | C | rs2168369 | A | 1.280 | $4.9x10^{-10}$ | 26 | 0 |
| 10 | rs4148920 | T | rs4751904 | C | 1.149 | $5.4x10^{-10}$ | 14 | 0 |
| 1 | rs506290 | T | rs883125 | G | 0.868 | $5.7x10^{-10}$ | 1 | 0 |
| 8 | rs16870304 | G | rs1019960 | A | 2.297 | $8.9x10^{-10}$ | 0 | 0 |
| 3 | rs9837986 | C | rs2019082 | C | 1.275 | $9.8x10^{-10}$ | 15 | 0 |

**Bolded**: SNP pair associations with $P<(0.05/157603906 \text{ tests}) = 3.2x10^{-10}$

**Table S10**: Proposed method's power and positive predictive value (PPV) for identifying replicated 3-SNP interactions

| | a. # of selected* 3-SNP trees (100 iterations, 3 logic trees per iteration) | b. # of iterations in which the underlying 3-SNP interaction was detected** (100 iterations) | c. Power: % of iterations in which the underlying 3-SNP interaction was detected (b/100 iterations) | d. PPV: % of underlying 3-SNP interaction detections, among the selected trees (b/a) |
|---|---|---|---|---|
| N=1000 Chr2, tree8 | 108 | 18 | 18.0% | 16.7% |
| Chr12, tree2 | 233[α] | 60 | 60.0% | 48.5%[α] |
| Chr12, tree5 | | 53 | 53.0% | |
| Chr13, tree2 | 124 | 47 | 47.0% | 37.9% |
| Chr14, tree7 | 92 | 43 | 43.0% | 46.7% |
| N=1500 Chr2, tree8 | 136 | 51 | 51.0% | 37.5% |
| Chr12, tree2 | 248 | 86 | 86.0% | 69.8% |
| Chr12, tree5 | | 87 | 87.0% | |
| Chr13, tree2 | 142 | 86 | 86.0% | 60.6% |
| Chr14, tree7 | 134 | 86 | 86.0% | 64.2% |
| N=2000 Chr2, tree8 | 128 | 64 | 64.0% | 50.0% |
| Chr12, tree2 | 242 | 90 | 90.0% | 77.7% |
| Chr12, tree5 | | 98 | 98.0% | |
| Chr13, tree2 | 133 | 96 | 96.0% | 72.2% |
| Chr14, tree7 | 132 | 96 | 96.0% | 72.7% |

* Selected: Permutation-based evaluation statistic>2

** Detected: Permutation-based evaluation statistic>2 and exact 3-SNP interaction tree match

[α] Two replicated trees were observed for chromosome 12. Since it is not possible to determine *a priori* which of the 3 trees identified in the simulation for a given iteration corresponds to a specific underlying replicated 3-SNP tree, we combined the total number of selections and detections such that PPV = (total detections)/(total selections).

**Table S11**: Benchmark method's power and positive predictive value (PPV) for identifying component 2-SNP interactions in replicated 3-SNP interactions

| | a. # of selected pairs meeting P<threshold (100 iterations, ~125K tests per iteration) | b. # of iterations for which at least one of the 2-SNP interactions in the underlying 3-SNP interactions met P<threshold (100 iterations) | c. Power: % of iterations in which at least one of the 2-SNP interactions in the underlying 3-SNP interactions met P<threshold (b/100 iterations) | d. # of 2-SNP interactions (all possible sub-pairs) in underlying 3-SNP interactions (P<threshold) | e. PPV: % of 2-SNP interactions (all possible sub-pairs) in underlying 3-SNP interactions, among selected pairs (P<threshold) (d/a) |
|---|---|---|---|---|---|
| P-value threshold: $P<3.2x10^{-10}$ | | | | | |
| N=1000 Chr2, tree8 | 0 | 0 | 0.0% | 0 | 0.0% |
| Chr12, tree2 | 28[α] | 12 | 12.0% | 12 | 42.9%[α] |
| Chr12, tree5 | | 0 | 0.0% | 0 | |
| Chr13, tree2 | 10 | 0 | 0.0% | 0 | 0.0% |
| Chr14, tree7 | 0 | 0 | 0.0% | 0 | 0.0% |
| N=1500 Chr2, tree8 | 3 | 0 | 0.0% | 0 | 0.0% |
| Chr12, tree2 | 101 | 31 | 31.0% | 31 | 30.7% |
| Chr12, tree5 | | 0 | 0.0% | 0 | |
| Chr13, tree2 | 34 | 2 | 2.0% | 2 | 5.9% |
| Chr14, tree7 | 0 | 0 | 0.0% | 0 | 0.0% |
| N=2000 Chr2, tree8 | 10 | 0 | 0.0% | 0 | 0.0% |
| Chr12, tree2 | 246 | 52 | 52.0% | 52 | 21.5% |
| Chr12, tree5 | | 1 | 1.0% | 1 | |
| Chr13, tree2 | 111 | 8 | 8.0% | 8 | 7.2% |
| Chr14, tree7 | 0 | 0 | 0.0% | 0 | 0.0% |
| P-value threshold: $P<1.0x10^{-5}$ | | | | | |
| N=1000 Chr2, tree8 | 544 | 0 | 0.0% | 0 | 0.0% |
| Chr12, tree2 | 4640 | 60 | 60.0% | 64 | 1.6% |
| Chr12, tree5 | | 8 | 8.0% | 8 | |
| Chr13, tree2 | 2302 | 23 | 23.0% | 23 | 1.0% |
| Chr14, tree7 | 248 | 0 | 0.0% | 0 | 0.0% |
| N=1500 Chr2, tree8 | 987 | 0 | 0.0% | 0 | 0.0% |
| Chr12, tree2 | 10891 | 84 | 84.0% | 94 | 0.9% |
| Chr12, tree5 | | 8 | 8.0% | 8 | |
| Chr13, tree2 | 4495 | 47 | 47.0% | 47 | 1.0% |
| Chr14, tree7 | 264 | 1 | 1.0% | 1 | 0.4% |
| N=2000 Chr2, tree8 | 1528 | 0 | 0.0% | 0 | 0.0% |
| Chr12, tree2 | 19979 | 98 | 98.0% | 119 | 0.7% |
| Chr12, tree5 | | 21 | 21.0% | 21 | |
| Chr13, tree2 | 7919 | 67 | 67.0% | 67 | 0.8% |
| Chr14, tree7 | 317 | 12 | 12.0% | 12 | 3.8% |

[α] Two replicated trees were observed for chromosome 12. Since it is not possible to determine *a priori* which of the 3 trees identified in the simulation for a given iteration corresponds to a specific underlying replicated 3-SNP tree, we combined the total number of selections (P<threshold) and detections (component SNP pair in an underlying 3-SNP interaction tree with P<threshold) such that PPV = (total detections)/(total selections).

**Table S12**: Significant eQTL associations in bone-related cells/tissues for SNPs in replicated 3-SNP interactions (GTEx Portal and GHS-Express monocyte *cis*-eQTL data)

| Chr | SNP | Allele | Tissue | Assoc. Gene | Beta Trend | P | P-value threshold (study) |
|---|---|---|---|---|---|---|---|
| chr2 | rs901466 | C/G | Adipose Subcutaneous | *RPL23AP7* | - | $4.5x10^{-6}$ | $6.5x10^{-5}$ |
| chr2 | rs901466 | C/G | Artery Tibial | *DDX11L2* | - | $7.8x10^{-7}$ | $8.4x10^{-5}$ |
| chr2 | rs901466 | C/G | Artery Tibial | *RPL23AP7* | - | $1.0x10^{-12}$ | $6.9x10^{-5}$ |
| chr2 | rs901466 | C/G | Artery Tibial | *AC024704.2* | + | $7.2x10^{-9}$ | $7.3x10^{-5}$ |
| chr2 | rs901466 | C/G | Cells EBV-transformed lymphocytes | *AC024704.2* | + | $1.1x10^{-5}$ | $3.8x10^{-5}$ |
| chr2 | rs901466 | C/G | Cells Transformed fibroblasts | *RPL23AP7* | - | $3.4x10^{-6}$ | $1.0x10^{-4}$ |
| chr2 | rs901466 | C/G | Cells Transformed fibroblasts | *DDX11L2* | - | $1.7x10^{-8}$ | $1.1x10^{-4}$ |
| chr2 | rs901466 | G/C | Monocytes | *RPL23AP7* | - | $3.9x10^{-20}$ | $5.8x10^{-12}$ |
| chr2 | rs901466 | C/G | Muscle Skeletal | *DDX11L2* | - | $8.6x10^{-7}$ | $6.6x10^{-5}$ |
| chr2 | rs901466 | C/G | Muscle Skeletal | *RPL23AP7* | - | $8.8x10^{-6}$ | $6.1x10^{-5}$ |
| chr2 | rs901466 | C/G | Nerve Tibial | *DDX11L2* | - | $7.8x10^{-5}$ | $9.3x10^{-5}$ |
| chr2 | rs901466 | C/G | Nerve Tibial | *RPL23AP7* | - | $3.8x10^{-7}$ | $9.3x10^{-5}$ |
| chr2 | rs901466 | C/G | Nerve Tibial | *AC024704.2* | + | $1.1x10^{-7}$ | $9.2x10^{-5}$ |
| chr2 | rs901466 | C/G | Skin Not Sun Exposed Suprapubic | *AC024704.2* | + | $2.3x10^{-6}$ | $4.6x10^{-5}$ |
| chr2 | rs901466 | C/G | Skin Sun Exposed Lower leg | *DDX11L2* | - | $1.7x10^{-5}$ | $7.9x10^{-5}$ |
| chr2 | rs901466 | C/G | Skin Sun Exposed Lower leg | *ACTR3* | - | $4.0x10^{-5}$ | $7.9x10^{-5}$ |
| chr2 | rs901466 | C/G | Skin Sun Exposed Lower leg | *AC010982.1* | - | $5.1x10^{-8}$ | $7.0x10^{-5}$ |
| chr2 | rs901466 | C/G | Testis | *AC024704.2* | + | $5.5x10^{-13}$ | $7.8x10^{-5}$ |
| chr2 | rs901466 | C/G | Testis | *AC104653.1* | - | $5.2x10^{-5}$ | $7.6x10^{-5}$ |
| chr2 | rs901466 | C/G | Thyroid | *SLC35F5* | - | $2.3x10^{-11}$ | $1.0x10^{-4}$ |
| chr2 | rs901466 | C/G | Whole Blood | *DDX11L2* | - | $2.3x10^{-6}$ | $5.9x10^{-5}$ |
| chr2 | rs901466 | C/G | Whole Blood | *RPL23AP7* | - | $1.5x10^{-8}$ | $6.0x10^{-5}$ |
| chr2 | rs901466 | C/G | Whole Blood | *AC024704.2* | + | $1.2x10^{-5}$ | $5.8x10^{-5}$ |
| chr2 | rs901466 | C/G | Whole Blood | *AC010982.1* | - | $1.0x10^{-7}$ | $6.0x10^{-5}$ |
| chr2 | rs7569568 | A/G | Monocytes | *FAM124B* | + | $1.7x10^{-15}$ | $5.8x10^{-12}$ |
| chr2 | rs921319 | T/C | Thyroid | *FAM161A* | + | $3.0x10^{-5}$ | $1.6x10^{-4}$ |
| chr2 | rs6708208 | G/A | Thyroid | *FAM161A* | + | $1.3x10^{-6}$ | $1.6x10^{-4}$ |
| chr2 | rs7569573 | A/G | Monocytes | *FAM124B* | + | $2.4x10^{-15}$ | $5.8x10^{-12}$ |
| chr2 | rs2122382 | C/A | Thyroid | *FAM161A* | + | $7.9x10^{-5}$ | $1.6x10^{-4}$ |
| chr2 | rs17407839 | A/C | Testis | *CUL3* | - | $2.5x10^{-5}$ | $9.2x10^{-5}$ |
| chr12 | rs1020745 | G/A | Artery Tibial | *AAAS* | + | $3.5x10^{-7}$ | $7.3x10^{-5}$ |
| chr12 | rs1020745 | G/A | Nerve Tibial | *AAAS* | + | $1.5x10^{-5}$ | $9.1x10^{-5}$ |
| chr12 | rs1020745 | G/A | Testis | *C12orf10* | - | $6.6x10^{-6}$ | $6.6x10^{-5}$ |
| chr12 | rs4768783 | C/T | Adipose Subcutaneous | *RP11-493L12.4* | + | $1.4x10^{-5}$ | $8.4x10^{-5}$ |
| chr12 | rs4768783 | C/T | Testis | *RP11-493L12.4* | + | $9.6x10^{-6}$ | $8.0x10^{-5}$ |
| chr13 | rs913071 | C/T | Nerve Tibial | *MAB21L1* | - | $3.6x10^{-5}$ | $1.0x10^{-4}$ |
| chr14 | rs1884632 | C/G | Skin Not Sun Exposed Suprapubic | *ACTN1* | - | $3.9x10^{-7}$ | $5.4x10^{-5}$ |
| chr14 | rs1884632 | C/G | Skin Sun Exposed Lower leg | *ACTN1* | - | $5.8x10^{-5}$ | $7.6x10^{-5}$ |
| chr14 | rs887890 | T/G | Nerve Tibial | *EIF2B2* | - | $3.6x10^{-6}$ | $1.1x10^{-4}$ |
| chr14 | rs887890 | T/G | Skin Sun Exposed Lower leg | *RP11-293M10.1* | - | $4.2x10^{-6}$ | $8.1x10^{-5}$ |

**Table S13**: Total frequencies of significant *cis*-eQTL associations available in GTEx and GHS-Express gene expression databases for 16 bone-related cells or tissues

| Tissue | Total significant eQTL association frequencies, by cell/tissue |
|---|---|
| Adipose Subcutaneous | 1311216 |
| Adipose Visceral Omentum | 593623 |
| Adrenal Gland | 405728 |
| Artery Tibial | 1245107 |
| Cells EBV-transformed lymphocytes | 369167 |
| Cells Transformed fibroblasts | 1315975 |
| Muscle Skeletal | 1124399 |
| Nerve Tibial | 1491673 |
| Ovary | 185414 |
| Pituitary | 269321 |
| Skin Not Sun Exposed Suprapubic | 741587 |
| Skin Sun Exposed Lower leg | 1336160 |
| Testis | 1148143 |
| Thyroid | 1592982 |
| Whole Blood | 1060536 |
| Monocyte | 36130 |
| Bone-related (all) | 14227161 |

**Table S14**: Enhancer enrichment analysis for 22 SNPs in all replicated 3-SNP trees (P<0.1)

| EID | Epigenome name | Enhancer state in replicated tree SNPs (N=22 SNPs) | Enhancer state in background SNPs (N=75508 SNPs) | OR | P |
|---|---|---|---|---|---|
| E035 | Primary hematopoietic stem cells (HSCs) | 8 | 12176 | 2.972 | 0.017 |
| E089 | Fetal muscle trunk | 7 | 10608 | 2.855 | 0.027 |
| E029 | Primary monocytes (from peripheral blood, PB) | 9 | 15416 | 2.699 | 0.030 |
| E030 | Primary neutrophils (from PB) | 6 | 8936 | 2.794 | 0.038 |
| E090 | Fetal muscle leg | 10 | 20004 | 2.312 | 0.053 |
| E036 | Primary HSCs short term culture | 8 | 14747 | 2.354 | 0.058 |
| E040 | Primary T helper memory cells (from PB) | 6 | 9868 | 2.494 | 0.058 |
| E079 | Esophagus | 6 | 9954 | 2.470 | 0.060 |
| E050 | Primary HSCs G-CSF-mobilized female | 8 | 15001 | 2.305 | 0.062 |
| E009 | H9 derived neuronal progenitor cultured cells | 0 | 10132 | 0.000 | 0.063 |
| E022 | iPS DF 19.11 cells | 5 | 8184 | 2.419 | 0.082 |
| E093 | Fetal thymus | 7 | 13387 | 2.166 | 0.094 |

**Table S15**: Promoter enrichment analysis for 22 SNPs in all replicated 3-SNP trees (P<0.1)

| EID | Epigenome name | Promoter state in replicated tree SNPs (N=22 SNPs) | Promoter state in background SNPs (N=75508 SNPs) | OR | P |
|---|---|---|---|---|---|
| E051 | Primary HSCs G-CSF-mobilized male | 5 | 7699 | 2.590 | 0.067 |
| E124 | Monocytes-CD14+ RO01746 | 5 | 8209 | 2.411 | 0.083 |

**Table S16**: Summary of Hi-C chromatin interaction evidence for select replicated 3-SNP interactions in lymphoblastoid cells

| 3-SNP interaction chr (SNPs)[a] | Interacting regions[b] Locus 1 : Locus 2 | Interacting regions Locus 2 : Locus 3 | Interacting regions Locus 1 : Locus 3 |
|---|---|---|---|
| Chr 2<br><br>SNP 1: rs921319<br>SNP 2: rs901466<br>SNP 3: rs7569568 | Locus 1: chr2:62446496-62546495<br>Locus 2: chr2:114583530-114683529<br>score: 2.7 | **Locus 2: chr2:114683530-114783529**<br>**Locus 3: chr2:225191756-225291755**<br>score: 3.6 | Locus 1: chr2:62446496-62546495<br>Locus 3: chr2:225091756-225191755<br>score: 4.7 |
| Chr 12<br><br>SNP 1: rs10444471<br>SNP 2: rs2110167<br>SNP 3: rs1020745 | None | Locus 2: chr12:5629739-5729738<br>**Locus 3: chr12:53613733-53713732**<br>score: 2.7 | Locus 1: chr12:4729739-4829738<br>**Locus 3: chr12:53613733-5371373**<br>score: 3.7 |
| Chr 12<br><br>SNP 1: rs1894331<br>SNP 2: rs4768783<br>SNP 3: rs10773093 | None | Locus 2: chr12:47613733-47713732<br>Locus 3: chr12:124734047-124834046<br>score: 3.2 | **Locus 1: chr12:11908733-12008732**<br>**Locus 3: chr12:125034047-125134046**<br>score: 3.6 |
| Chr 13<br><br>SNP 1: rs9315069<br>SNP 2: rs913071<br>SNP 3: rs7321815 | **Locus 1: chr13:31302000-31401999**<br>Locus 2: chr13:36402000-36501999<br>score: 2.0<br><br>Locus 1: chr13:31402000-31501999<br>**Locus 2: chr13:36502000-36601999**<br>score: 2.0 | None | **Locus 1: chr13:31302000-31401999**<br>**Locus 3: chr13:101601999-101701998**<br>score: 3.4 |
| Chr 14<br><br>SNP 1: rs7142110<br>SNP 2: rs1884632<br>SNP 3: rs887890 | None | **Locus 2: chr14:69330247-69430246**<br>Locus 3: chr14:75530247-75630246<br>score: 2.1 | Locus 1: chr14:51830250-51930249<br>**Locus 3: chr14:75630247-75730246**<br>score: 2.2 |

a. Numbered SNPs are SNPs in best replicated 3-SNP interaction trees and correspond with numbered loci in other table columns.

b. **Bolded** genomic coordinates: SNPs in the replicated 3-SNP interaction are in 100-kb regions with evidence of direct chromatin interaction between 2 distal regions of interest in lymphoblastoid cells (Broad/MIT/UMass GM06990). Score = $\log_2$(observed contact frequency/expected contact frequency).

**Figures S17-S20**: Long-range chromatin interaction visualizations for select replicated 3-SNP interactions (WashU EpiGenome Browser)

Description of Figures S17-S20: To examine evidence for interactions between regulatory regions with implicated SNPs in 3-SNP interactions, we used the WashU EpiGenome Browser to visualize long-range chromatin interactions within and across 500-kb windows (separated by gray boundary lines) centered around SNPs of interest (specific locations shown by purple vertical lines, with broader context given by ideograms), in conjunction with epigenome and transcriptome data. Histone modification (H3K4me3: promoter mark [red], H3K4me1 and H3K27ac: enhancer marks [yellow and green, respectively]), RNA-seq (blue), and DNase I hypersensitivity (purple) data tracks were reviewed. Four data tracks per assay for cell/tissue samples previously identified as relevant in the BMD literature were examined (in this order, top to bottom): lymphoblastoid cells (LCLs), peripheral blood mononuclear cells or monocytes, mobilized CD34 cells, and osteoblasts or an osteoblastic precursor proxy (H1 mesenchymal cells). Broad/MIT/UMass Hi-C data generated with GM06990 LCLs was used to assess evidence for long-range chromatin interactions between SNPs involved in 3-SNP interaction trees (100-kb bin resolution, $\log_2$[observed contact/expected contact] scores). Minimum Hi-C interaction scores were set such that interaction arcs represent chromatin interactions with at least +4-fold observed contact frequency over expected (interaction scores >2).

Figure S17: Chromatin interaction evidence for the chromosome 2 interaction (rs921319, rs901466, rs7569568).

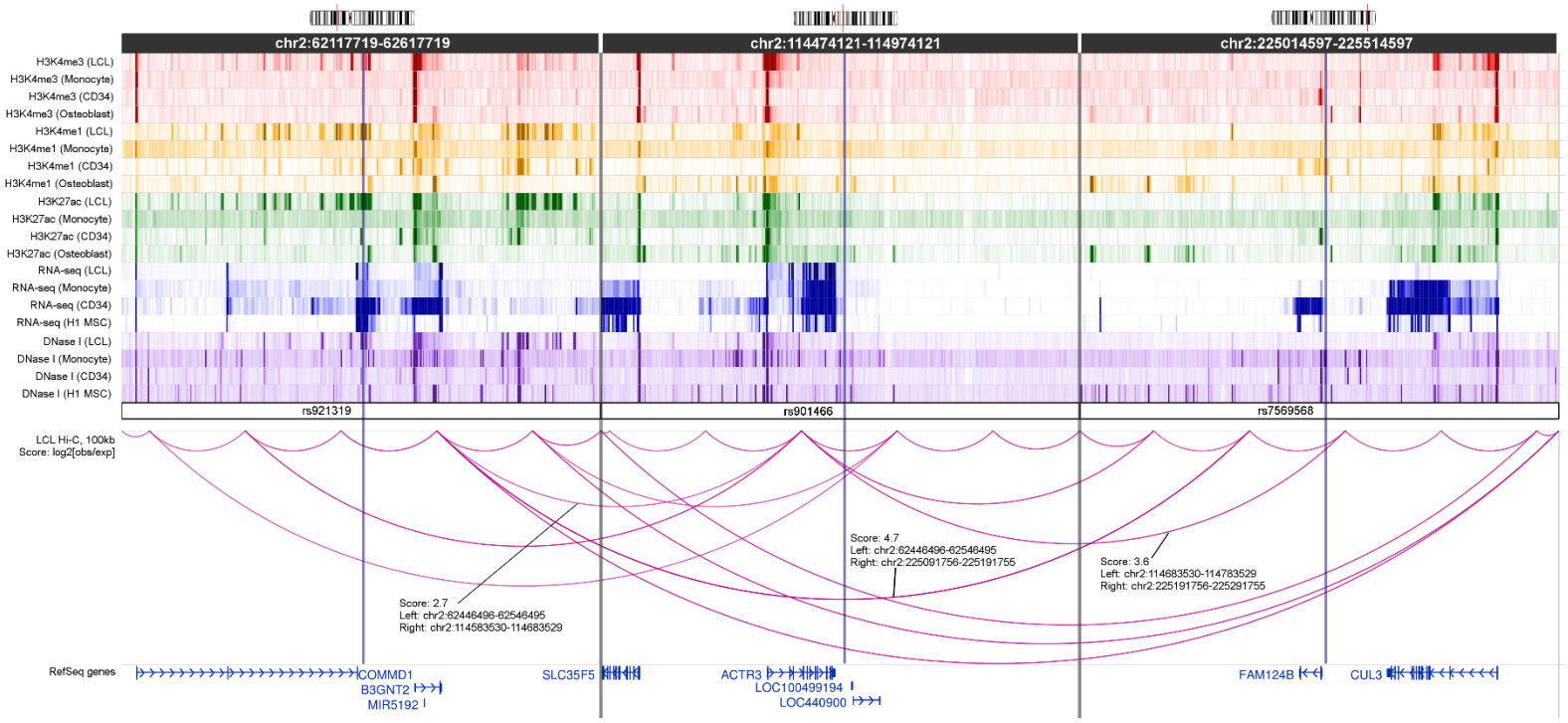


Figure S18: Chromatin interaction evidence for the chromosome 12 interaction (rs1894331, rs4768783, rs10773093).

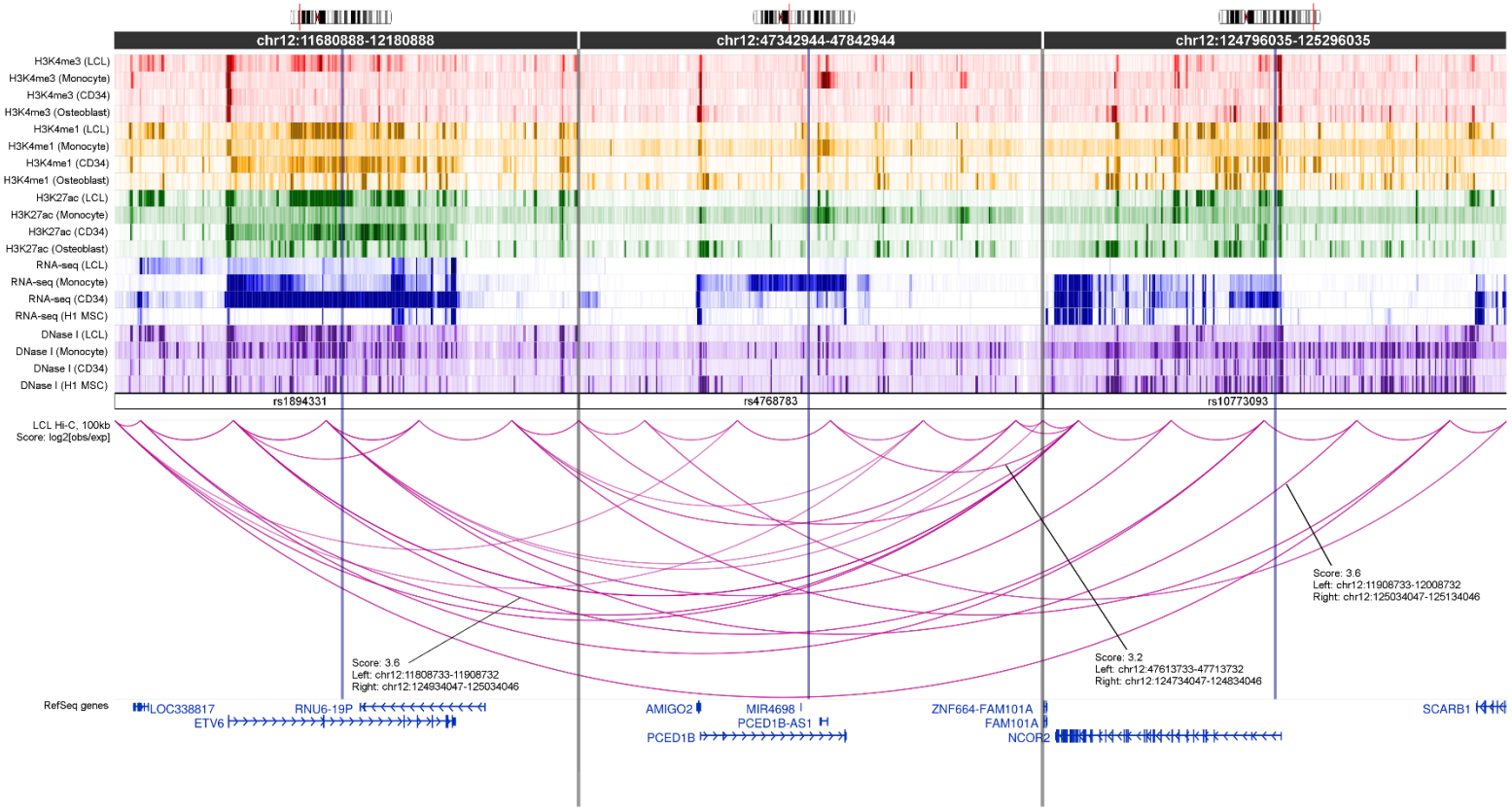


Figure S19: Chromatin interaction evidence for the chromosome 13 interaction (rs9315069, rs913071, rs7321815).

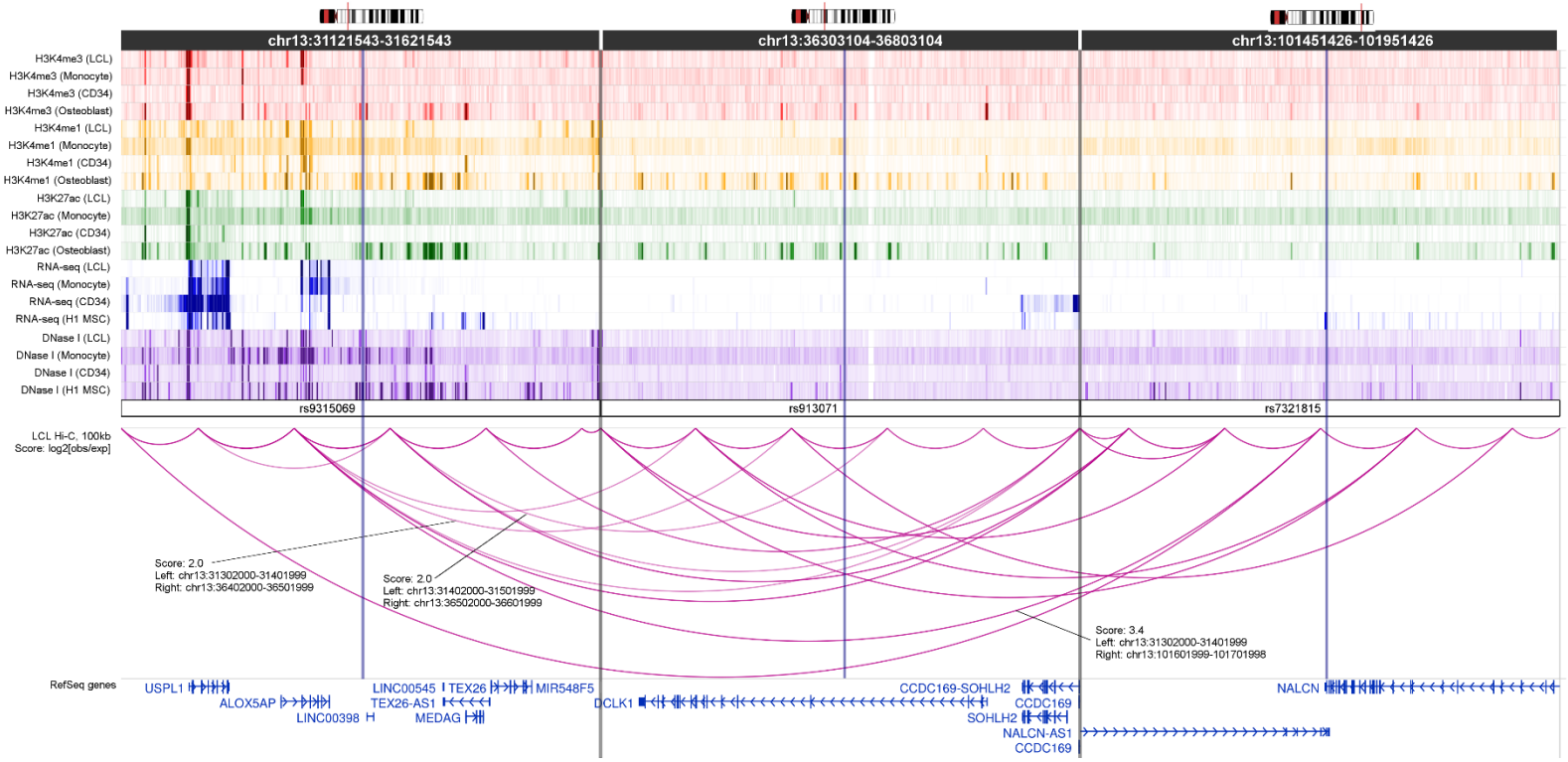
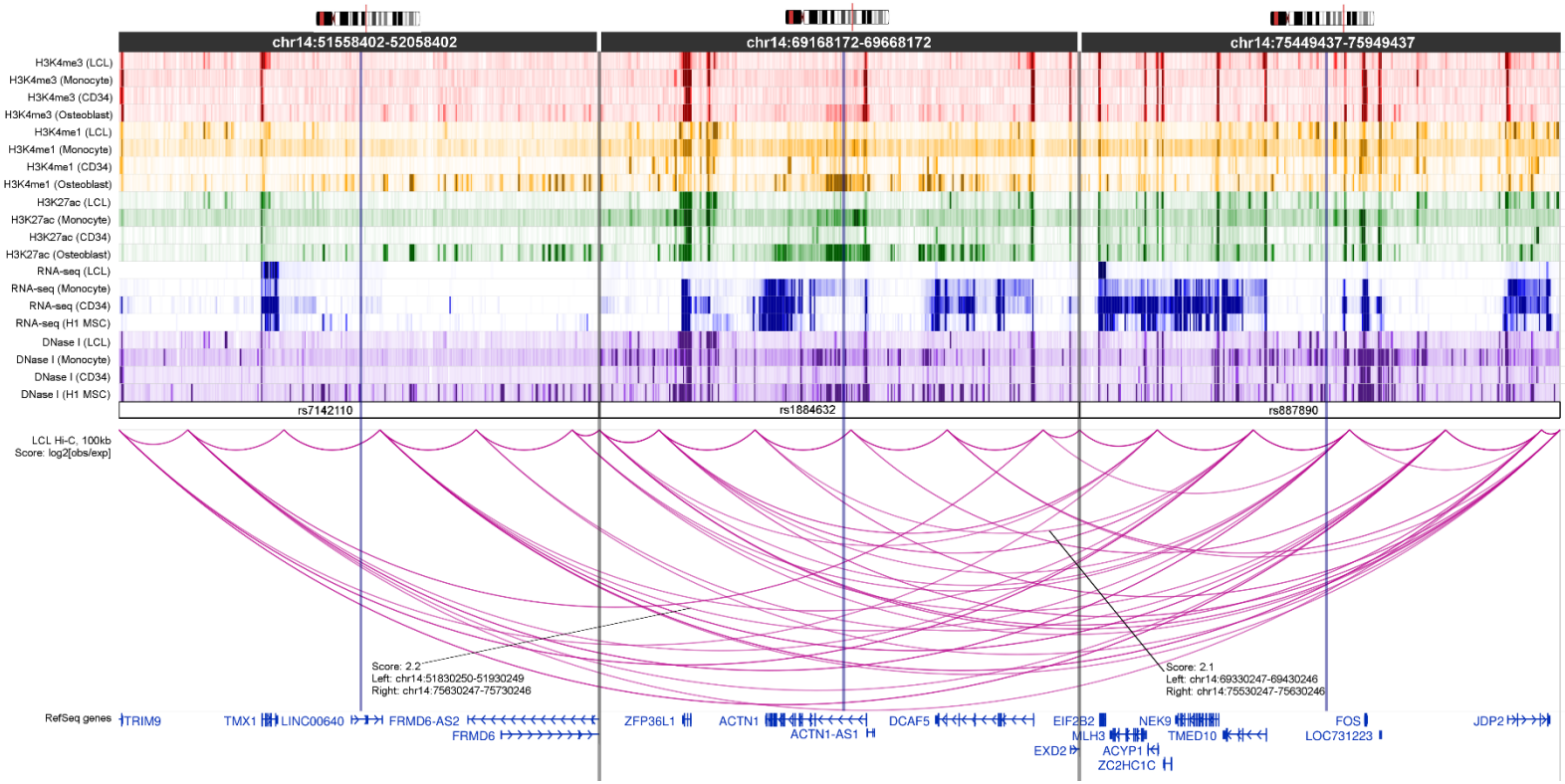


Figure S20: Chromatin interaction evidence for the chromosome 14 interaction (rs7142110, rs1884632, rs887890).



# CHAPTER 3

## Biologically-Motivated Learning from the Whole Human Genome using Logic Regression

### 3.1 Introduction

This paper describes a major challenge in modern genetics research for which machine learning may provide effective solutions. Simply put, a common challenge that biomedical research faces today is to discover, with a limited number of observations, patterns in the 0.1% of billions of genetic measurements that vary across individuals which predict, or are associated with, physically-expressed characteristics of interest, or "phenotypes" (e.g., binary indicators for developing a disease or side effects from treatment, or quantitative characteristics such as blood pressure). The data analytic capacity of researchers remains limited, however, with respect to finding meaningful associations: far more associations than what have been reported to date are expected.

In this paper, our goals are to: (1) mathematically describe the genomic data structure and the learning problem without requiring knowledge of genetics; and (2) propose a machine learning approach to identify associations utilizing external biological information available online. Specifically, we represent the data contained in the whole human genome as a three-dimensional array, introduce the learning problem in these terms, and discuss limitations of the standard analytic method. We then propose a biologically-motivated logic regression algorithm (69) that uses sequential conditioning and permutation-based inference, and evaluate its performance by assessing its precision in a simulation study. Our proposed machine learning

approach has successfully identified genetic patterns associated with a physically-expressed characteristic using real data. While effective, our approach does not provide a comprehensive solution by any means. Our hope is that by introducing this analytic challenge critical to the progress of biomedicine to experts in the field of machine learning, more complete solutions will be developed, eventually moving association discoveries to predictions.

### 3.2 A representation of the whole human genome as an array

- The human genome can be thought of as a 3-dimensional array $X[j = 1{:}\,2, k = 1{:}\,23, l = 1{:}\,L_k]$. Each cell of an individual carries two 2-dimensional arrays $X[1, k = 1{:}\,23, l = 1{:}\,L_k]$ and $X[2, k = 1{:}\,23, l = 1{:}\,L_k]$ in *X*, with one originating from his/her father and the other from his/her mother.

- Each of the two 2-dimensional arrays consists of 23 one-dimensional arrays, e.g., $X[1, 1, l = 1{:}\,L_1]$, $X[1, 2, l = 1{:}\,L_2]$, ..., $X[1, 23, l = 1{:}\,L_{23}]$. For a given *k*, a pair of the equal-length one-dimensional arrays ($X[1, k = 1{:}\,23, l = 1{:}\,L_k]$, $X[2, k = 1{:}\,23, l = 1{:}\,L_k]$, or $X[1{:}\,2, k = 1{:}\,23, l = 1{:}\,L_k]$), represents a pair of "chromosomes". Most cells of the human body contain an identical set of 23 pairs of chromosomes.

- The elements of *X*, called "bases", are represented by one of the following four letters, A, G, C, or T: $X[j, k, l] \in \{A, C, T, G\}$. Typically, for any given position (*k*, *l*), $X[1, k, l]$ and $X[2, k, l]$ can be only one of two possibilities, say, C and T, among the four letters. The pair of these two elements $X[1{:}\,2, k, l]$ is referred to as a "base pair". Specific letter

combinations, without distinguishing maternal or paternal origin (i.e., $j$=1 or $j$=2), are called "genotypes" (e.g., (C,C), (C,T), (T,T)).

- The total number of elements in $X$ is approximately 6 billion (a pair of 3 billion): $\sum_{k=1}^{23} L_k \approx 3\text{x}10^9$.With today's biotechnology, we can measure all 6 billion elements in $X$ relatively quickly and inexpensively. Each individual has his/her own $X$ (which we will denote as $X_i$ for $i^{\text{th}}$ individual), but the vast majority of its elements, ~99.9%, are identical for all human beings: $X_i[j,k,l] = X_{i'}[j,k,l]$ for 99.9% of $\{(k,l): k = 1, \dots, 23, l = 1, \dots, L_k\}$ for any $j$.

### 3.3 Difficulties in learning about subsets of X that influence Y's of interest

Family pedigree studies have shown that certain phenotypes $Y$'s, such as height and developing heart disease, are influenced, to varying degrees, by variations in the 0.1% of the human genome $X$. If one or a small number of $X$'s elements strongly determine a phenotype $Y$ (e.g., a genetic disorder such as sickle cell disease), researchers can successfully identify causal element(s) in $X$ using standard methods. However, phenotypes $Y$'s are frequently not entirely determined by the additive effects of independent single genetic variants (e.g., many types of cancer). Physical and social environments and lifestyle factors, as well as common complex forms of genetic variation like interacting elements in $X$, significantly influence physically-expressed characteristics. Unfortunately, we do not necessarily know all non-genetic factors that influence a given phenotype $Y$. Even if we could enumerate all relevant non-genetic factors, we may not be capable of measuring all such factors precisely and accurately, or obtain them for use

in our genetic investigations. Difficulties in ascertaining and measuring key non-genetic factors hinder our abilities to identify subsets of $X$ that influence $Y$'s of interest.

Another difficulty is "the curse of dimensionality". The number of observations (individuals) in a single study from whom both $X$ and $Y$ are measured is limited and much smaller than the number of non-identical elements in $X$, by a few to several orders of magnitude. In addition, if detecting associations of interacting elements of $X$ with $Y$ are of primary interest, the number of potential interactions that would need to be evaluated increases exponentially as the number of elements in $X$ increases.

To give a concrete example of applications: at St. Jude Children's Research Hospital, we are following a cohort of adult long-term (10+ years) survivors of childhood cancer, which includes approximately 3,000 individuals. Because treatments that kill cancerous cells could also damage normal cells, these survivors are at elevated risk of developing various medical problems many years after their treatment for cancer concludes. To understand who is at higher risk for different medical problems and to what degrees, our study collects comprehensive data on various phenotypes $Y$'s, along with potential predictors of $Y$'s, such as cancer treatment information from medical records and genetic data $X$. While the measurements of the whole genome $X$ have been recently completed, this comprehensive genetic dataset is undergoing extensive quality-control checks. To conduct conventional genetic association analyses, a reduced subset of approximately one million ($k$, $l$)'s at which the elements of $X$ are commonly known to vary across individuals (e.g., vary in at least 1% of a reference population) are currently available for analysis. Even with a reduced subset of genetic features, the number of observations, 3,000, is three orders of magnitude smaller than the number of commonly occurring variations in $X$.

## 3.4 Two goals of learning from whole genome data

A major goal of medical research with genetic data $X$ is to predict $Y$ precisely and accurately so that proper prevention and intervention strategies could be developed based on that prediction. This framework is often referred to as "precision medicine." In our cohort study of adult survivors of childhood cancer, we would like to know survivors' risk of having cardiac problems, for example, using both treatment exposure information and genetics data $X$ of each survivor so that he/she can receive individually-tailored follow-up care. Depending on how strongly a specific $Y$ of interest is associated with a small subset of $X$ and other predictors, predicting $Y$ for each individual with precision required by clinical applications may not be a realistic goal for certain phenotypes $Y$'s at this time.

An alternative, more achievable goal for many phenotypes $Y$'s at this time is to identify subsets of $X$ that are associated with $Y$, i.e., detecting differences in the distribution of $Y$ by subsets of $X$. Because $X$ can be annotated based on observed or predicted biological functions, the discoveries of associations between “biologically-meaningful” subsets of $X$ and $Y$ give clues to biomedical scientists investigating biological mechanisms underlying these associations, even if they are insufficient for individual prediction. Such investigations advance science and could lead to broader biological implications with respect to the cause, prevention, and/or treatment of $Y$ and associated conditions. Thus, the learning from the human genome has two goals: one is to predict $Y$ with $X$ along with other predictors; and the other is to evaluate associations of $Y$ with subsets of $X$. In this paper, we focus on the latter.

### 3.5 Standard analyses to identify subsets of X associated with Y

The current standard approach to evaluate associations between subsets of $X$ and $Y$ examines $X[1{:}\,2, k, l]$ in relation to the distribution of $Y$, for each position ($k$, $l$) one at a time. Specifically, since $X$'s elements at a given ($k$, $l$) typically takes one of two possibilities, say, C and T, from {A, G, C, T}, $X[1{:}\,2, k, l]$ can be one of three possibilities or genotypes, e.g., (C, C), (C, T), or (T, T). The information regarding maternal or paternal origin provided by $j$=1 vs. $j$=2 (called "haplotype phase") may not be distinguishable (i.e., the measurement given as (C, T) may be $X[1{:}\,2, k, l]$ or $X[2{:}\,1, k, l]$, but this knowledge is not necessarily required to conduct an association analysis). The association can be statistically evaluated by assessing whether the distribution of $Y$ differs across the three $X[1{:}\,2, k, l]$ genotype groups, e.g., by Chi-square test if $Y$ is binary and ANOVA if $Y$ is continuous. Because there are many ($k$, $l$)'s, the problem of multiple testing arises and the standard practice is to control Type I error by Bonferroni correction, dividing the statistical significance by the number of ($k$, $l$)'s tested, e.g., one million.

This standard approach is effective when single $X[1{:}\,2, k, l]$'s influence $Y$ independently. For many complex phenotypes $Y$'s, however, single $X[1{:}\,2, k, l]$'s identified by the standard approach are far from sufficient in explaining the amount of variation in $Y$'s that are estimated to be attributable to genetic variations in $X$. This phenomenon is referred to as the "missing heritability" problem (2). For example, for Type I diabetes, the amount of variation in disease development attributable to hereditary genetic components is three times larger when estimated from studies of twins (88) than when estimated from the sum of discoveries made by the standard analysis to date (1).

## 3.6 Two key aspects of biologically-motivated learning

Representing biologically-plausible joint effects of multiple ($k$, $l$)'s

An approach that is explicitly complementary to the standard analysis of single $X[1{:}\,2, k, l]$'s one at a time is to examine multiple $X[1{:}\,2, k, l]$'s jointly.

Suppose there are necessary conditions on elements of $\{X[1{:}\,2, k, l]{:}\ (k, l) \in S\}$, all of which are required to cause an influence on $Y$, where $S$ is a biologically-meaningful set. Then, the association with $Y$ can be seen fully only when we examine the elements of the relevant set $\{X[1{:}\,2, k, l]{:}\ (k, l) \in S\}$ jointly and when all the necessary conditions are met. Similarly, when there is a set of sufficient conditions on elements of $\{X[1{:}\,2, k, l]{:}\ (k, l) \in S\}$, any one of which can cause an influence on $Y$, then the effective approach is to examine the elements of the relevant set $\{X[1{:}\,2, k, l]{:}\ (k, l) \in S\}$ jointly.

There are a number of biological reasons to consider the existence of multiple necessary and/or sufficient conditions. For example, developing cancer requires multiple necessary conditions in a cell (e.g., destroying a checkpoint that flags abnormality AND activating promotion of growth AND deactivating suppression of growth), but there are multiple sufficient ways to develop the same type of cancer (i.e., any one of the different sets of necessary conditions is sufficient for developing the same cancer). "Genetic heterogeneity" is a more direct example of multiple sufficient causes. For example, cystic fibrosis is a disorder associated with a part of the genome - specifically, in a "gene" located at $S = \{(k, l){:}\ k = 7, l = 116{,}907{,}253, \ldots, 117{,}095{,}955\}$. A "gene" is a set of adjacent elements ($k$, $l$)'s that encodes a specific "protein", and proteins carry out certain biological functions. A number of subsets of this

set *S* could cause a defect in a key protein and lead to the same disease, cystic fibrosis, representing multiple sufficient causes.

Boolean logic trees can express these necessary and sufficient causes and their combinations mathematically. AND links necessary causes, while OR links sufficient causes. Because each element of *X* at a given (*k*, *l*) typically takes one of two possibilities from {A, G, C, T}, Boolean logic trees are particularly suited to expressing a specific pattern of multiple $X[1{:}\,2, k, l]$'s. A Boolean logic tree may involve the operator NOT ($^{C}$): e.g., $X[1{:}\,2, k, l] = (C, C)\ AND\ (X[1{:}\,2, k, l'] = (T, T))^{C}$. A Boolean logic tree may have each of its leaves defined for a given (*k*, *l*), or for one of the paired one-dimensional arrays ("haplotypes") if *j*=1 and *j*=2 are distinguishable (i.e., the "phase" is known): $((X[1, k, l], X[1, k, l'], X[1, k, l'']) = (C, C, T))\ AND\ ((X[2, k, l], X[2, k, l'], X[2, k, l'']) = (C, C, T))$.

Restriction of the search space through biological considerations

Given "the curse of dimensionality" problem, searching the entire space of *X* for its subsets associated with a given *Y* may be ineffective. There are a number of ways biologists may annotate the human genome *X*, some of which may be used to restrict the search space. Many of these biological annotations of the human genome are made available publicly online though bioinformatics resources such as ENCODE (encodeproject.org) (25), ENSEMBL (ensembl.org) (89), KEGG (genome.jp/kegg) (90) and GO (geneontology.org) (91) to name a few.

### 3.7 A proposed learning method for the whole genome

Logic regression

Logic regression (69) uses the regression framework of Generalized Linear Models with one change: its predictors are Boolean logic trees. Specifically, the systematic component of the model takes the form:

$$h(E[Y]) = b_0 + b_1 BL_1 + b_2 BL_2 + \cdots + b_p BL_p$$

where $h(.)$ is a link function, E[$Y$] is the expected value of $Y$, $b$'s are regression coefficients, and $BL$'s are Boolean logic trees. The random component of the model specifies the probability distribution of $Y$ as one of the exponential family distributions. By employing logic regression as a learning method, we can incorporate the biological concept of joint effects of multiple $X[1{:}\,2, k, l]$'s.

The objective function to be maximized is the likelihood function. Since an exhaustive search of all potential interaction models is not feasible in terms of power or computational resources, and use of a greedy algorithm implementation may not necessarily lead to a globally optimal solution, logic regression was implemented with simulated annealing by its developers. We use their implementation in the current work.

Focus on joint effects of "enhancers" and "promoters": A restriction of the search space

We propose the use of one of the annotations available in ENCODE (25) to restrict the search space. Evidence from a large number of "single nucleotide polymorphism" studies that measured a subset (in the order of millions) of $X$ whose genotypes $X[1{:}\,2, k, l]$'s are known to differ in at least 1% of a reference population indicates that, regardless of the phenotype $Y$, the

majority of subsets of *X* that have been found to be associated with *Y*'s appear to be "regulatory elements" of the genome (27, 92). Biologically, these "regulatory elements" are known to interact with each other to jointly regulate the expression of genes, which in turn, can affect the levels and rate at which gene products, e.g., proteins, are produced. Thus, one possible restriction of the search space, which is also consistent with the consideration of joint effects of multiple elements of *X*, is to search subsets of *X* that strictly include (*k*, *l*)'s that are annotated as regulatory elements. Ernst *et al.*, 2011 estimated from experimental data using Hidden Markov Models (*k*, *l*)'s that are likely regulatory elements in nine different types of human cells (27). Our search space restriction used their labeling of specific "regulatory elements" (called "enhancers" and "promoters") in any of the nine human cell types they studied.

Accounting for non-genetic factors that influence Y

In biomedical studies, there are often non-genetic factors that are known to influence *Y*. In evaluating the association of genetic factors with *Y*, it is essential to account for their effects. To do so, we modify the systematic part of the logic regression model described previously to:

$$h(E[Y]) = b_0 + a_1 Z_1 + \cdots + a_p Z_p + b_1 BL_1 + b_2 BL_2 + \cdots + b_B BL_B$$

where *Z*'s are non-genetic factors and *a*'s are their corresponding regression coefficients.

The inclusion of non-genetic factors *Z* must be accounted for in determining the statistical significance of a *BL* in logic regression. Since our null hypothesis is no association between *Y* and *BL* conditioned on the association of *Y* with *Z*, we considered permutation-based inference conditioned on the inclusion of *Z* in the statistical model described above (in our empirical example, we refer to the statistical model with *Z* only as the “base model”). If we

regress *Y* on *Z*, obtain predicted values of *Y* given measured values of *Z*, and permute *Y* within strata defined by the fitted values of *Y* from the base model, we effectively account for the effects of non-genetic factors *Z* on phenotype *Y* in permutation. This is the permutation process we used to account for non-genetic factors in assessing the statistical significance of *BL*'s in our empirical example.

Sequential conditioning of logic regression models to find "causal" Boolean logic trees

Even with the use of a two-stage process to preliminarily restrict the number of elements of *X* to those annotated as "regulatory elements", the size of the search space for Boolean logic trees is enormous. We propose a sequential conditioning logic regression algorithm (**Figure 3.1**). The algorithm selects 10 *BL*'s sequentially where each *BL* has three leaves (elements of *X*) and the selection is conditioned on the previously identified trees and *Z* (the base model). This sequential strategy is a way of learning restrictively rather than learning a large Boolean logic tree from the enormous search space, relying solely on logic regression's simulated annealing methodology. The selection of 10 *BL*'s is conducted for each $k$=1, …, 22 (chromosomes), resulting in 220 *BL*'s which we assessed for statistical significance.

A total of *m* conditional permutations form the basis for assessing statistical significance. Each of the *m* permutations undergoes the same procedure for selecting 220 *BL*'s as the original unpermuted analysis. The Wald-test p-value of the $b^{th}$ *BL* for a given *k* for the original unpermuted *Y* is compared against its *m* counterparts, where the ranking of the original *Y*'s Wald-test p-value among its *m* counterparts divided by *m* is the permutation-based p-value. We used

$p<0.05$ as our threshold to call the *BL* in question a "discovery". We used $m$=1,000 in the real data analysis (see Section 3.8) and $m$=100 for the simulation study (see Section 3.9).

Reasons for using permutation in assessing statistical significance

If a set of Boolean logic trees of interest is given a priori, then standard statistical inference (e.g., likelihood ratio or Wald tests) could be used to assess its statistical significance in association with *Y*. However, we stochastically search for Boolean logic trees in a very large space in relation to *Y*, and, therefore, the standard statistical inference would not provide correct statistical significance. The proposed permutation-based statistical inference addresses this issue. Any test statistic obtained from the unpermuted original dataset can be compared against its permutation-based null distribution obtained by applying the exact same method to each of the permuted datasets to estimate the statistical significance. This feature is also effective in reducing the scale of the multiple testing issue. Specifically, we do not exhaustively test millions of possible combinations of $X[1{:}\,2, k, l]$'s. Rather, we apply logic regression to each $k$ (chromosome), which contains ($k$, $l$)'s in the order of hundred thousands, to obtain a relatively small number test statistics (e.g., test statistics for 220 *BL*'s) in both the unpermuted and permuted datasets.

## 3.8 Application to a Real Data Example

One of the goals of our cohort study is to evaluate genetic variation in *X* associated with bone mineral density (*Y*) among acute lymphoblastic leukemia (ALL) survivors. After quality-

control, a total of 770,471 ($k$, $l$)'s were retained for analysis. The standard analysis identified one of these ($k$, $l$)'s as being associated with $Y$. Upon annotating ($k$, $l$)'s using the annotation framework described in (see Section 3.7), our final restricted subset of $X$ included 75,523 ($k$, $l$)'s. For each of these, two binary variables were created to represent two typical genetic effects (dominant and recessive genetic inheritance models). The sequentially conditioned logic regression algorithm was then applied, using binary sex, continuous ancestry, and categorical cancer treatments as non-genetic factors ($Z$), to perform stochastic searches of the space of three-leaf Boolean logic trees. After the first three-leaf Boolean logic tree was identified, nine more trees were sequentially identified conditioned on the preceding trees for each of the 22 non-sex chromosomes. Using the algorithm described in **Figure 3.1** with m=1,000 permutations, conditional permutation-based p-values were calculated for each of the 22x10=220 three-leaf Boolean logic trees.

Of the 220 three-leaf Boolean logic trees, only eight had conditional permutation-based p-values <0.05, less than what would be expected by chance if the 220 tests were independent. To validate, each of the eight three-leaf Boolean logic trees was evaluated for its association with $Y$ in an independent separate sample of N=1,428 childhood cancer survivors with pediatric cancers other than ALL. To our surprise, five of the eight trees were validated in the non-ALL sample for association with $Y$ (bone mineral density), as a modifier of the effect of high-dose methotrexate (a cancer chemotherapy) or cranial radiation on $Y$, with a Wald-test p-value <0.05. Literature reviews of the elements in these five Boolean logic trees also clearly support the biological plausibility of their involvement as influencers of bone mineral density.

## 3.9 Simulation experiment: Evaluation of mean precision

Simulation model motivation and objectives

The sequential conditioning logic regression algorithm with permutation applied to real data successfully identified five "causal" Boolean logic tree signals that were also replicated in an independent dataset. It was surprising to observe that this algorithm would be able to identify five causal 3-leaf trees, while calling only eight trees statistically significant in discovery.

We conducted a simulation study to determine the level of precision of the algorithm. Specifically, the objective of our simulation study was to assess whether a sequential conditioning algorithm with permutation applied to simulated phenotype $Y$ and genetic data $X$ for a given chromosome $k$ could successfully identify "causal" signals with comparable precision in contrast to an algorithm without sequential conditioning (hereafter referred to as the "marginal" algorithm).

Description of the simulation model and results

A continuous phenotype $Y$ was simulated under a linear model, $Y|b_l, X_l, Y = \sum_{l=1}^{50} b_l X_l + e$ where $X_l$'s represent binary causal genetic variables with $l$=1, …, 50, their additive effects on $Y$ are $b_l$'s, and $e$ is random error generated from the standard normal distribution. To generate $b_l$ and $X_l$, we first generated a parameter $p_l$ from the triangular probability distribution with range 0.05 to 0.40 and mode 0.10, with which each $X_l|p_l$ was generated from the Bernoulli distribution with its parameter $p_l$. Each effect size $b_l|p_l$ was generated from the uniform

distribution using two different ranges: one had a range from 0.05 to $0.05/\sqrt{p_l}$ (weaker effects); and the other had a range from 0.05 to $0.10/\sqrt{p_l}$ (stronger effects). In addition to the 50 causal genetic variables, 1,000 binary random $X_l$, $l = 51, \ldots, 1050$, with no association with the phenotype $Y$, were generated from the Bernoulli distribution with parameter 0.10. Two sample sizes were considered: 1,000 and 3,000.

For each iteration of the 1,000 simulation iterations, we applied the sequential conditioning and marginal algorithms to the simulated dataset for the iteration. Specifically, the sequential conditioning algorithm would find the best $X_l$ of the 1,050 $X_l$'s (i.e., the $X_l$ with the smallest Wald-test p-value) associated with $Y$, conditioning on the previously identified best $X_l$'s, until ten best $X_l$'s are identified ($b$=10). The marginal algorithm would simply find the top ten $X_l$'s of the 1,050 $X_l$'s (i.e., the 10 $X_l$'s with the 10 smallest Wald-test p-values) associated with $Y$. To calculate the statistical significance of the best $X_l$'s, we permuted $Y$ 100 times ($m$=100) and applied the two algorithms for each permuted $Y$ in the exact same way as the unpermuted $Y$. The Wald-test p-value of the top 10 $X_l$'s identified with the original unpermuted $Y$ by each algorithm were compared against their 100 counterparts with the permuted $Y$, where the original $Y$'s first best was compared against the 100 permuted $Y$'s first bests, the original $Y$'s second best was compared against the 100 permuted $Y$'s second bests, and so on. The ranking of the original $Y$'s $X_l$ among its 100 counterparts from the 100 permuted $Y$s divided by 100 is the permutation-based p-value, and we used $p<0.05$ to indicate that the $X_l$ was designated by the algorithm as a "discovery". Simulation results are summarized in **Table 3.1**.

Precision is defined as the proportion of "true" or causal discoveries out of the total number of claimed discoveries. This was estimated by the number of causal $X_l$'s with $p<0.05$ in the best (top) ten divided by the number of all $X_l$'s with $p<0.05$ in the top ten, averaged over the

1,000 iterations excluding the iterations with no $X$ with $p<0.05$ in the top ten. With the exception of the simulation scenario with a smaller sample size and weaker expected causal effects, where approximately 600 iterations resulted in no $X_l$ with $p<0.05$ in the top ten, the sequential conditioning algorithm had the same mean precision as the marginal algorithm.

The marginal algorithm is more difficult to implement, especially in the setting of stochastic searches in a large space, which is the case for our application with logic regression. This is because the marginal algorithm has to find $B$ best patterns, while the sequential conditioning algorithm has to find the single best pattern. This is particularly relevant when the patterns we search for contain multiple elements of $X$ such as our Boolean logic trees: we cannot remove all elements involved in previously identified patterns, as one or more of these elements may contribute to subsequent best patterns. By observing the same level of precision between the marginal approach (the standard permutation inference) and the conditional approach in our simulation study, we assert that our proposed sequential conditioning approach has validity even though some of the previously identified $X_l$'s are not causal (random $X_l$'s generated from the standard normal distribution).

## 3.10 Conclusion

Our proposed approach was to:

- Restrict the search space biologically by focusing on elements of $X$ which regulate how genes are expressed;
- Consider Boolean logic trees to reflect "biologically realistic" necessary and/or sufficient conditions on the restricted set $X$ as causal factors associated with phenotype $Y$;

- Apply logic regression to select features (elements of *X*) in the form of Boolean logic trees under the Generalized Linear Model framework; and
- Implement a sequential conditioning algorithm with permutation-based inference, which, when compared to the standard marginal algorithm with permutation-based inference that is difficult to implement in a high-dimensional space, showed equivalent performance with respect to precision in a simulation study.

This approach was effective in our empirical analysis, in which five out of eight claimed discoveries were validated in an independent dataset.

There are a number of limitations with our investigation. First, the biological restriction of the search space can be done in many other ways. Second, the genetic organization of elements in *X* that influence a given phenotype *Y* is unknown. In addition to the true size of causal Boolean logic trees being unknown, with a limited number of observations, we are unable to learn large Boolean logic trees. The ten three-element trees that were selected sequentially for each of 22 chromosomes may not represent the extent of patterns underlying genetic interactions. Third, given the stochastic nature of the simulated annealing algorithm and the dimensionality of the search space, the global optimum may not have been achieved in selecting each tree. While the simulated annealing procedure was started with 200 different initial values for each tree search, the larger the search space (e.g., increasing elements of *X* and/or searching for larger-sized Boolean logic trees), the more difficult it is to search for the global optimum.

In light of these limitations, we hope that by presenting this challenge as it pertains to genetics research in programming/mathematical terms, machine learning experts gain fresh perspective and develop innovative solutions that meet this challenge.

**Table 3.1: Simulation study results, with 1,000 simulation iterations**

| | | | Performance | | | |
|---|---|---|---|---|---|---|
| Effect size generation ($p_I$=prevalence of causal X) | Sample Size | Method | Average # of causal $X_I$'s in top 10 | Average # of $X_I$'s with p<0.05 in top 10 | Average # of causal $X_I$'s with p<0.05 in top 10 | Mean precision[a] |
| $U(0.05,\frac{0.05}{\sqrt{p_I}})$ | 1,000 | Marginal | 2.50 | 1.47 | 0.42 | 0.32[b] |
| | | Conditional | 2.44 | 1.03 | 0.28 | 0.28[b] |
| | 3,000 | Marginal | 6.39 | 7.30 | 4.73 | 0.64 |
| | | Conditional | 6.36 | 6.65 | 4.32 | 0.64 |
| $U(0.05,\frac{0.10}{\sqrt{p_I}})$ | 1,000 | Marginal | 6.29 | 7.36 | 4.73 | 0.64 |
| | | Conditional | 6.21 | 6.21 | 4.01 | 0.64 |
| | 3,000 | Marginal | 9.52 | 9.97 | 9.49 | 0.95 |
| | | Conditional | 9.59 | 9.95 | 9.54 | 0.96 |

Abbreviations: U=uniform distribution; #=number.

[a] Defined as the average of the number of causal $X_I$'s with p<0.05 in the top ten divided by the number of $X_I$'s with p<0.05 in the top 10, excluding runs where there was no $X_I$ with p<0.05 in the top 10.

[b] Approximately 60% of iterations had no $X_I$ with p<0.05.

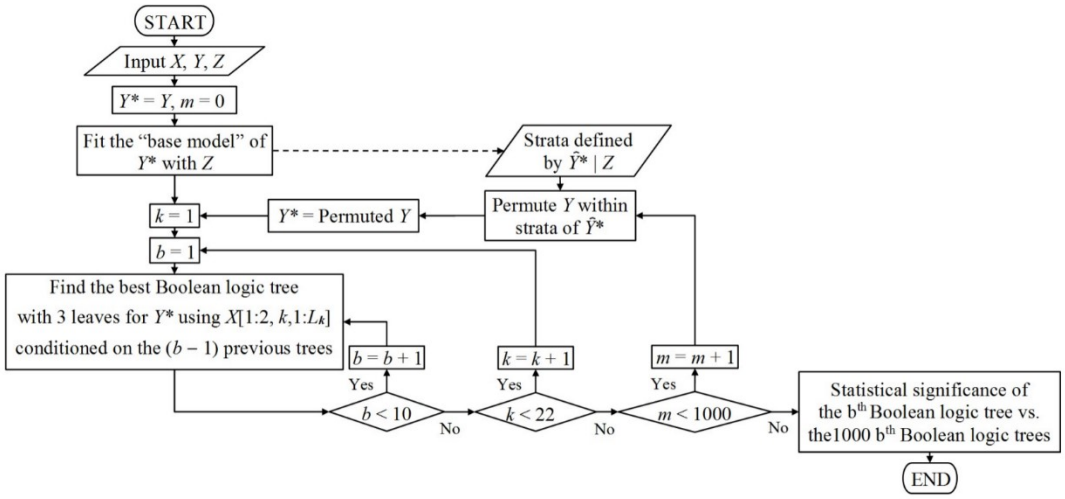



**Figure 3.1: Overview of the sequential conditioning logic regression algorithm with permutation**

The proposed sequential conditioning algorithm for finding $b$=1, …, 10 three-leaf Boolean logic trees for each of $k$=1, …, 22 chromosomes, with $m$=1,000 permutations that accounts for the base model with non-genetic factors $Z$ for determining statistical significance.

# CHAPTER 4

## Genome-wide haplotype association analysis of primary biliary cholangitis risk in Japanese

### 4.1 Introduction

Primary biliary cholangitis (PBC; also known as primary biliary cirrhosis, MIM 109720) is a progressive autoimmune disease of the liver, leading to the destruction of the bile ducts, and in end-stage cases, liver failure. While the factors that underlie PBC susceptibility and increase risk for disease progression remain enigmatic (93, 94), a concordance rate of 63% for PBC has been observed in monozygotic twins, suggesting that PBC has a strong hereditary component (52).

Recent genome-wide association studies (GWAS) in European cohorts have confirmed associations between the human leukocyte antigen (HLA) locus and identified over two dozen non-HLA PBC susceptibility loci (95-98). PBC GWAS in Japanese cohorts have only replicated a minor number of risk loci identified in European populations (e.g., *IL7R*, *IKZF3*, *CD80*), and revealed features of the PBC genetic architecture that may be dissimilar between populations, with *TNFSF15*, *POU2AF1*, and *PRKCB* emerging as major susceptibility loci among Japanese (53, 54). While these results suggest that additional single-SNP GWAS with larger sample sizes are warranted, it is also worthwhile to consider complementary analytic approaches to gain further insights. One such approach involves the exploration of the effects of haplotypes, or the arrangement of multiple SNP alleles on the same chromosome. Haplotype patterns may not only be more powerful for mapping disease genes, but may also be uniquely informative about known

single SNP associations, since haplotypes are also known to vary considerably between populations, bear signatures of selection, and may contextualize the role genetic variants play, singly or in tandem, in disease pathogenesis (19-21).

Among existing methods to conduct genome-wide haplotype association studies with unphased genotype data, the most popular strategy is to split the genome indiscriminately into overlapping "sliding" windows, and simultaneously infer frequencies for all possible haplotypes among a small, fixed number of SNPs within each window and test global associations under a regression-based framework (19, 45, 46). Due to the computational burden of simultaneously inferring and testing haplotypes, the selected windows for haplotype formation are small, and constructed haplotypes consider a limited number of contiguous SNPs. An important consequence of these analytic restrictions is that a comprehensive exploration of haplotypes as models for the *cis*-regulation of gene expression is typically not feasible, given that gene expression is modulated by haplotypes comprised of both proximal and distal SNPs with individual or interacting effects (67, 68) and may be influenced by larger networks of SNPs (i.e., three SNPs or more) (38).

In the current study, we conducted a genome-wide gene-based haplotype association analysis to uncover relationships between SNPs that are potentially transcriptionally-relevant and are also associated with PBC risk. We applied a regression-based methodology to find haplotypes consisting of three SNPs associated with PBC risk among sets of SNPs mapped to extended gene-centered windows in a discovery cohort of 1937 Japanese, without restricting formed 3-SNP haplotypes to contiguous SNPs. Phased haplotypes inferred with whole chromosome SNP data were treated as observed in our downstream association analysis. We used a permutation-based approach to select top haplotypes and replicated selected findings in a

second independent Japanese cohort (N=949). Ancillary bioinformatics analyses were conducted to assess the biological plausibility of associations between PBC risk and detected 3-SNP haplotypes.

## 4.2 Results

Discovery and replication of genic haplotypes associated with PBC risk

We developed a novel approach to detect haplotypes associated with PBC risk by leveraging the logic regression (69) methodology, a statistical learning method that employs a stochastic search algorithm to detect combinations of binary predictors (i.e., haplotypes) associated with an outcome of interest (i.e., PBC risk) within a generalized linear model (GLM) framework. After computationally resolving haplotype phase for the 2886 participants in the combined study cohort, a total of 272131 SNPs were mapped to at least one of 15137 gene analytic windows, or broadly-defined gene regions that include RefSeq genes and the flanking 500-kb regions before and after transcription start and stop sites (**Figure S1**). Using the proposed methodology, we identified the most strongly associated haplotype consisting of three SNPs for every gene analytic window in our discovery cohort (N=1937). A total of 7317 unique candidate 3-SNP haplotype associations were detected: 7272 genic haplotypes included no SNPs in the HLA region (chr6:29645000–33365000, hg19 build), while 45 genic haplotypes included at least one HLA region SNP.

We used a permutation-based evaluation statistic (99) to assess whether the candidate genic 3-SNP haplotypes had stronger associations with PBC risk than expected. Applying a pre-

defined cut-off (top one percentile of permutation-based evaluation statistic values), 74 candidate 3-SNP haplotype associations were selected for replication follow-up. We employed a Bonferroni-corrected per-test significance threshold ($P<0.05/74=6.8x10^{-4}$) in our replication analysis with our independent cohort (N=949). Under this p-value threshold, nearly two-thirds of selected 3-SNP haplotypes were replicated (49/74 haplotypes). All genic haplotypes with HLA region SNPs were replicated (37/37 trees), while 32.4% of genic haplotypes without HLA region SNPs (12/37 trees) were replicated. To contextualize the efficacy of our permutation-based evaluation statistic cut-off in capturing the strongest haplotype signals detected with logic regression, **Figure 4.1** contrasts the distribution of p-values for association tests in the replication cohort (N=949) for the 74 selected 3-SNP haplotypes to the relatively uniform replication p-value distribution for the 7243 dropped 3-SNP haplotypes.

Discovery and replication analysis results for the 49 replicated 3-SNP genic haplotypes are provided in **Table S1**. Among replicated haplotypes, 69.4% (34) included at least one SNP that was not nominally significant, while 24.5% (12) contained no SNPs that individually achieved genome-wide significance ($P<5x10^{-8}$) (**Table S2**). The magnitude of estimated ORs for replicated 3-SNP haplotypes in the combined cohort (N=2,886) assuming additive haplotype effects ranged from 1.672 to 15.246 (inverting protective associations), with p-values ranging from $1.3x10^{-35}$ to $3.9x10^{-9}$. **Tables 4.1 and 4.2** highlight selected example results for five replicated 3-SNP haplotypes. Among these examples, (rs9295704=C) and ((rs2451752=A) and (rs2575174=C)) on chromosome 6 (OR=0.365, $P=8.9x10^{-15}$), and ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G) on chromosome 7 (OR=0.066, $P=3.9x10^{-9}$) represent novel non-HLA loci associated with PBC risk (**Table 4.2**). Other examples provided correspond

to genomic regions carrying single variants with the strongest associations with PBC risk in previous PBC GWAS in Japanese (near *HLA-DRA* and *TNFSF15*) (53, 54).

Missingness rates for replicated 3-SNP haplotypes in the unphased data among all available controls (median: 1.13%, IQR: 0.86%; N=1505) and cases (median: 1.09%, IQR: 0.80%; N=1381) were low and comparable between groups (**Table S3**). Pre-phasing counts of unambiguous homozygous carriers and potential carriers of at least one haplotype copy were consistent with the estimated corresponding replicated haplotype counts in the phased data for all available controls (N=1,505, **Table S4**). Lastly, frequency distributions for each of the replicated 3-SNP haplotypes among all phased study controls (N=1505) and the phased 1000G JPT reference panel (N=104) demonstrated that these distributions were comparable for each haplotype (**Table S4**).

Comparing proposed and benchmark methodologies for haplotype detection

We applied a benchmark haplotype association methodology to conduct global tests of association (45) for estimated haplotypes formed within all available sliding window sets comprised of three contiguous SNPs in each of the 15137 gene analytic windows in our discovery cohort (N=1937). Using this benchmark method, we identified 1425 haplotypes with p-values meeting a Bonferroni-corrected p-value threshold ($P<0.05/1567361$ tests$=3.2\times10^{-8}$) across 205 gene analytic windows (**Table 4.3**). Nearly two-thirds of the gene windows (135/205) with a top haplotype association detected by the benchmark method was a window that also contained a replicated 3-SNP haplotype detected with our proposed method, suggesting the proposed and benchmark methods found many of the same gene analytic windows to be

important for haplotype exploration. No exact matches for 3-SNP haplotypes detected by logic regression were observed among top haplotypes identified by the benchmark method.

Given that SNPs contributing to 3-SNP haplotypes detected by logic regression are not necessarily contiguous and estimated haplotypes were treated as observed in this analysis, we computed global test score statistics with variance estimates that account for the uncertainty in the haplotype estimation (45) for estimated haplotypes consisting of the three SNPs in each of the replicated 3-SNP haplotype trees in the combined cohort (N=2886). The scale of each haplotype tree's global test p-value appeared to be consistent with the haplotype p-value obtained with the proposed method (**Table S5**).

Functional annotation of replicated haplotype SNPs

We conducted enrichment analyses to broadly investigate the biological plausibility of replicated 3-SNP haplotypes' associations with PBC risk by comparing the set of 106 unique SNPs contributing to replicated 3-SNP haplotypes ("haplotype SNPs") against an unpruned comparison set of 16036 SNPs mapped to gene analytic windows with nominal univariate associations with PBC risk ($P<0.05$). An examination of the number of gene expressions (eQTLs) significantly associated with haplotype SNPs in three blood and liver cell/tissue types (Genotype-Tissue Expression, GTEx v7 (63)) demonstrated that haplotype SNPs were significantly enriched for eQTLs in lymphoblastoid cells relative to the comparison SNP set ($P=2.7x10^{-3}$) (**Table S6**). The set of haplotype SNPs was also significantly enriched for overlaps with ChiP-seq histone modification peaks linked with enhancer (H3K4me1) or promoter (H3K4me3) activity in at least 20 of the 29 consolidated blood and liver cell types available in

Roadmap Epigenomics Mapping Consortium (62) (REMC) data (Bonferroni-corrected $P<0.05/29=1.7x10^{-3}$; **Tables S7, S8**), respectively. Haplotype SNPs were also enriched for an indicator of open chromatin (DNase I peaks) in four blood cell types among 11 blood/liver cell types with REMC assay data (Bonferroni-corrected $P<0.05/11=4.5x10^{-3}$; **Table S9**). Haplotype SNPs were jointly enriched for enhancer and promoter peaks in 15 blood/liver cell types, and simultaneously enriched for all three chromatin state indicators in three cell types in peripheral blood: primary B cells, primary T cells, and monocytes (**Figure 4.2**). These results are consistent with reported immune-related PBC disease mechanisms that implicate B cell and T cell differentiation pathways (100), and observations of significant inflammatory cell infiltration (including B cells, T cells, and macrophages) associated with the loss of biliary epithelial cells in the portal tract (101).

**Figure 4.3** highlights two replicated 3-SNP haplotypes in the HLA region near rs3129887, the variant with the strongest single-SNP association with PBC risk in Japanese (53, 54). Association testing results for haplotypes detected with logic regression and the benchmark method are shown in the top data track in Figure 3 across chr6:32156782-32585905, followed by a visual summary of relevant functional annotations for SNPs in the replicated 3-SNP haplotypes. One of the replicated 3-SNP haplotypes shown is (rs3129881=C) or ((rs375244=A) and (rs3132947=G)) (OR=3.665, $P=2.3x10^{-24}$), with SNPs in *NOTCH4* and *HLA-DRA* introns; the other is ((rs9268831=T) or (rs9269190=T)) or (rs9270652=C) (OR=3.075, $P=7.3x10^{-29}$), comprised of intergenic SNPs near *HLA-DRA* and *HLA-DRB1*. These haplotypes connect SNPs that may be linearly far apart, can consist of SNPs that do not achieve genome-wide significance ($P<5x10^{-8}$), and combine SNPs across multiple top 3-SNP haplotype windows tagged by the benchmark method. Additionally, all six haplotype SNPs overlap H3K4me1 and/or H3K4me3

peaks in primary B or T cells, while five SNPs have at least one significant eQTL in whole blood, lymphoblastoid, or liver cells.

Replicated 3-SNP haplotypes in chromosomes 7 and 9 include SNPs that overlap genomic regions tagged with the smallest p-values identified by the benchmark method across chr7:7667281-8026742 and chr9:116727079-118438852, respectively, while corresponding functional annotations contextualize the relative contributions of each SNP (**Figures S2, S3**). To clarify, two intergenic SNPs in the chromosome 9 (rs4979484=C) or ((rs13300483=T) and (rs7028891=G)) haplotype with strong univariate associations with PBC risk ($P<5x10^{-8}$) also have significant associations with *TNFSF8* expression in whole blood. However, this haplotype also includes rs4979484, an intergenic SNP with a relatively weak marginal association with PBC risk (P=0.005) (**Table 4.2**). Interestingly, rs4979484 not only overlaps H3K4me1 and H3K4me3 peaks in multiple blood/liver cell types, but is also in a region with evidence of binding to nuclear factor kappa B (NF-kB), a transcription factor reported to play a critical role in inflammation and immunity processes (102), in a lymphoblastoid cell line of Japanese origin (GM18951; **Table 4.4, Table S10**).

## 4.3 Discussion

In the current study, we propose a novel approach for the detection of haplotype associations to detect PBC risk haplotypes genome-wide. Previous studies have successfully combined agnostic sliding windows with an exhaustive testing strategy to identify haplotypes associated with disease risk. To contrast, our proposed method uses a logic regression-based stochastic search to detect best 3-SNP haplotype associations among phased SNP alleles mapped

to broadly-defined gene regions. This approach has several strengths. First, the chosen analytic windows encourage searches for genic haplotypes, thereby finding combinations of variably-spaced SNPs that may influence *cis*-regulatory mechanisms for gene expression. Second, logic regression considers multiple models of risk (e.g., presence of risk alleles at *either* of two loci) that can better reflect regulatory redundancies that may exist in controlling transcription. Lastly, the proposed method avoids exhaustive testing, potentially capturing true haplotype associations that would otherwise be missed due to lack of statistical power. Using the proposed method, a total of 74 3-SNP haplotypes on chromosomes 6, 7, and 9 were considered as having stronger associations with PBC risk than expected under a permutation-based approach in a discovery cohort of 1,937 Japanese individuals. Nearly two-thirds of these selected haplotypes (49 haplotypes) were replicated in a second independent Japanese cohort (N=949) after applying a Bonferroni-corrected p-value threshold ($P<6.8x10^{-4}$).

Haplotype association analyses using inferred haplotypes in downstream analyses may be vulnerable to Type I error inflation and biased estimates of genetic effects due to misclassification of haplotype states (103). Several aspects of this analysis mitigate these concerns. Phasing was conducted for cases and controls simultaneously, which provides greater control of Type I error than phasing these groups separately (103). Differential misclassification of haplotype states is unlikely, since haplotype phasing was conducted without knowledge of disease status. Non-differential misclassification is more plausible; thus, reported effect estimates may be biased towards the null (haplotype has no effect on PBC risk). To safeguard against false positives, we conducted a replication study and applied a Bonferroni-corrected p-value threshold to address multiple testing in replication. Our benchmark method analysis that considered the uncertainty of haplotype phase also tagged ~82% of the gene analytic windows containing

replicated 3-SNP haplotypes. Lastly, the frequency distributions for replicated 3-SNP haplotypes among controls in our study were consistent with the 1000G JPT reference panel. These results suggest that our method can reliably detect credible 3-SNP genic haplotypes associated with PBC risk.

Limitations of our proposed method include not finding the haplotype with the strongest association (due to non-exhaustive testing) and missing risk haplotypes outside of genic regions. However, we replicated 49 out of 74 genic haplotype associations selected in discovery; replicated 3-SNP haplotypes also frequently included SNPs that overlapped top haplotype associations identified with the exhaustive testing-based benchmark method. Yet, exact matches between top haplotypes detected with the proposed and benchmark methods were not observed. Instead, replicated 3-SNP haplotypes detected by logic regression linked SNPs ~335 kb apart on average, with many contributing SNPs overlapping functional annotations in cell/tissue types relevant to PBC. Thus, the haplotypes detected with the proposed method: (a) frequently include SNPs that overlap genomic regions with top haplotype associations identified by an exhaustive testing-based benchmark method; (b) are unlikely to be detected with existing association methods; and (c) combine the effects of variably-spaced and potentially functional SNPs mapped to regions of the genome that are more likely to be transcribed.

Similar to recent PBC GWAS (54, 95, 96), we did not examine sex chromosome variants. However, sex chromosome-related defects and haplotype deficiencies likely play an important role in PBC etiology (104, 105). Specifically, higher rates of X monosomy in peripheral blood cells in PBC-affected women have been observed, suggesting X monosomy influences PBC pathogenesis (106). Although skewed X-chromosome inactivation (XCI) may contribute to autoimmune disease risk (107, 108), preferential X loss that involves particular X-linked

haplotypes may better explain the increased X monosomy in women with PBC (109). Analogous to this X haploinsufficiency observed in women with PBC, increased Y chromosome loss has been associated with PBC in men (110). Further exploration of sex chromosome-specific haplotype associations with PBC risk is needed.

While we did not functionally validate replicated 3-SNP haplotypes, results from ancillary bioinformatics analyses suggest that identified haplotypes may be considered in future functional investigations of genetic susceptibility factors for PBC. SNPs in replicated 3-SNP haplotypes were significantly enriched for gene expressions in lymphoblastoid cells and indicators of enhancer, promoter, and open chromatin states in blood and liver cell types compared to SNPs in extended genic regions that were marginally associated with PBC risk. Specific functional annotations of SNPs in PBC-associated haplotypes also indicate that identified haplotypes may enhance understanding of posited disease mechanisms for both novel and known genetic associations. For example, the chromosome 7 haplotype, ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G), represents a novel candidate PBC susceptibility locus, with SNPs mapped to intronic regions of *UMAD1*. Two SNPs in this haplotype (rs12702656, rs11768586) overlap enhancer peaks in liver cells. A suggestive association with insulin-like growth factor 1 (IGF1) levels, a hormone predominantly produced by the liver, has been reported for a proximal variant mapped to *UMAD1* (rs7780564, $p=3.9x10^{-7}$) (111); IGF1 is hypothesized to regulate biliary epithelial cell proliferation (112). On the other hand, *TNFSF15* is reported to be the most strongly associated non-HLA genetic susceptibility factor for PBC in Japanese. *TNFSF15* is anticipated to play a role in the inflammation response, activating Th1/Th17 cell differentiation and cytokine production (53). The (rs4979484=C) or ((rs13300483=T) and (rs7028891=G)) haplotype, comprised of intergenic SNPs near *TNFSF15*, implicate *TNFSF8* as

another possible contributor to PBC risk. This 3-SNP haplotype includes two SNPs significantly associated with *TNFSF8* expression and a third SNP (rs4979484) with evidence of NF-kB binding in GM18951 (JPT) lymphoblastoid cells. *TNFSF8* specifically encodes a cytokine (CD30L) expressed in T cells and monocytes; like *TNFSF15*, *TNFSF8* is also a known activator of NF-kB, a transcription factor implicated in inflammation and immunity pathways (102).

In conclusion, this study presents a novel approach to conduct haplotype association analyses genome-wide, with a deliberate focus on interrogating regions of the genome that are likely to be transcribed. With this method, we identified novel candidate PBC susceptibility loci (e.g., *UMAD1*) and detected haplotype patterns that potentially contribute to hypothesized disease pathways that include previously reported PBC susceptibility genes (*HLA-DRA*, *TNFSF15*) in the Japanese population. Broader explorations of haplotypes may increase understanding of the genetic basis of PBC and potentially inform new interventions that improve disease prognosis.

## 4.4 Methods

Study population

The current study combines data collected for two previous PBC GWAS in Japanese, both coordinated by the Japan PBC-GWAS Consortium (53, 54). All research participants provided informed consent. The methods/protocols implemented in this analysis were approved by the ethics committees of the Nagasaki Medical Center and all participating institutions. All methods were performed in accordance with relevant guidelines and regulations.

The combined study population has been previously described in detail (54). Briefly, DNA samples were obtained from either healthy controls reporting no apparent disease, or PBC cases, defined by laboratory or histological evidence of at least two of the following criteria: cholestasis, ascertained by elevated alkaline phosphatase; serum anti-mitochondrial antibodies; and non-suppurative destructive cholangitis and interlobular bile duct destruction. DNA samples were genotyped using the Affymetrix Axiom Genome-Wide ASI 1 Array (Affymetrix, Santa Clara, CA). As previously reported (54), samples with excess heterozygosity rates, cryptic relatedness, or of non-Japanese ancestry were removed. A total of 425290 autosomal SNPs were retained under the following quality control criteria: SNP call rate ≥95%; MAF ≥5%; and Hardy-Weinberg equilibrium (HWE) P ≥0.001 in controls. Samples with <97% sample call rate among retained SNPs were excluded, resulting in a study sample of 2886 individuals. We split this sample *a priori* into two cohorts to correspond with the timing of sample collections for the two previous PBC GWAS: 1937 participants (901 cases, 1036 controls) for haplotype signal discovery, and 949 participants (480 cases, 469 controls) for haplotype signal replication.

Haplotype phasing

Haplotype phase was computationally estimated using SHAPEIT v2.79 (60) for whole chromosomes with unphased SNP genotypes for all 2886 unrelated samples in the combined study cohort. To improve phasing accuracy, we used 1000 Genomes Phase 3 genetic map recombination rates, increased the number of conditioning states for the haplotype estimation to 600 states, and increased the numbers of burn-in, pruning, and main iterations (10, 10, and 50 iterations, respectively) of the SHAPEIT MCMC algorithm. Recommended parameters for

haplotype estimation mean window size (2 Mb) and effective population size (15000) for GWAS data were employed. Upon phasing, each sample was assigned its most likely haplotype phase configuration with binary-encoded phased genotypes at each SNP locus (0|0, 0|1, 1|0, or 1|1, where 0=reference allele, 1=alternative allele).

Statistical Analyses

*Restricting haplotype formation within gene regions*

We restricted our search for haplotype signals within regions centered on annotated protein-coding and non-translated RNA-encoding genes. Specifically, we used gene transcripts annotated by the RefSeq gene model (release 74, GRCh37/hg19 build) (65), employed ANNOVAR (66) to map SNPs in our dataset to introns, exons, and 3'/5' untranslated regions, and identified flanking 500-kb regions for each RefSeq transcript, as flanking gene regions are critical for transcriptional events (25). SNPs mapped to a gene and flanking regions were considered as a single "gene analytic window" (**Figure S1**). To reduce each set of SNPs in a given gene analytic window to "tagging SNPs" in formulating haplotypes, we removed SNPs within each window sequentially so that no pair of SNPs was in high linkage disequilibrium ($r^2 \leq 0.8$). Gene analytic windows with a minimum of two SNPs were retained for analysis.

*Detecting 3-SNP haplotypes with logic regression*

We identified gene-based haplotype patterns consisting of three SNPs associated with PBC risk with an adapted logic regression (69) algorithm. Briefly, logic regression is a statistical learning method that utilizes a stochastic search algorithm to support the detection of higher order interactions associated with an outcome within a generalized linear model (GLM) framework. For our analysis, 3-SNP haplotypes are expressed mathematically as Boolean (true/false) variables that combine SNP alleles on the same chromosome, e.g., ((SNP1=reference allele) and (SNP2=alternative allele) and (SNP3=alternative allele)) = {True, False}.
We utilized logic regression's "simulated annealing" search algorithm (R "LogicReg" package, version 1.5.8) to build many possible haplotype predictors stochastically and evaluate them based on GLM model scores (e.g., deviance scores for logistic regression). Logic regression was applied to each gene analytic window, considering the contributions of two haplotypes per subject in our discovery cohort (N=1937), to identify candidate 3-SNP haplotype patterns associated with PBC risk. To stabilize the performance of the stochastic algorithm, we utilized 100 different initialization values for each gene analytic window and chose the model with the lowest deviance score among the 100 fits.

To select candidate 3-SNP haplotypes for replication, we used a permutation-based evaluation statistic based on previous work (99). Specifically, for each gene analytic window, 20 permutations of PBC disease status were used to obtain model deviance scores under each permutation. We then calculated the 20 possible null distribution deviations between a deviance score under a given permutation and the median estimated with the remaining 19 permuted datasets, along with the corresponding median absolute deviation (MAD, a robust measure of variability) for each gene window's set of 20 null distribution deviations. We defined our permutation-based evaluation statistic as $\frac{D_{obs}-D_{med}}{MAD_{D_{med}}}$, where $D_{obs}$ is the observed deviance score

for the gene analytic window's best-fitting 3-SNP haplotype, $D_{med}$ is its respective empirically-derived median under 19 permutations, and $MAD_{D_{med}}$ is the MAD of the null distribution deviations under 20 permutations.

Given gene analytic windows potentially overlap, the same best-fitting 3-SNP haplotype could be identified for multiple windows. In these cases, the 3-SNP haplotype with the best permutation-based evaluation statistic across multiple windows was retained to contribute to a set of unique 3-SNP haplotypes for replication follow-up (e.g., no haplotype tree had an exact SNP and logic tree structure match with another tree). We set an *a priori* threshold to select candidate 3-SNP haplotypes: 3-SNP haplotypes with the top 1% of permutation-based evaluation statistics among the set of unique 3-SNP haplotypes were selected for replication, translating to statistic values of less than -11.4. This cut-off translates to the selection of 3-SNP haplotypes with logistic regression model deviance scores at least 11.4 median absolute deviations away from corresponding medians estimated under the null distribution.

Since haplotype logic regression models consider two observations with the same case/control status from each subject (i.e., two haplotypes from two homologous chromosomes), we used the following logistic regression model for the best detected 3-SNP haplotype associations, treating each subject as an independent observation to satisfy GLM assumptions for valid statistical inference:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 L \qquad (1)$$

where $p$ is PBC risk and $L$ represents additive haplotype effects of the best-fitting 3-SNP haplotype logic tree. The reported model deviance score, odds ratio, and p-value for each detected haplotype in the discovery cohort were derived from equation 1.

To replicate selected haplotype associations, we tested the selected 3-SNP haplotypes from the discovery cohort with the same model described in equation 1 in our independent replication cohort (N=949) and applied a Bonferroni-corrected per-test significance threshold ($P<0.05$/number of selected haplotypes). We further required odds ratios in the replication cohort to have the same direction as the discovery cohort to consider a 3-SNP haplotype as replicated.

*Detecting 3-SNP haplotypes with a benchmark method*

We used the R "haplo.stats" package (45), commonly employed in secondary explorations of haplotypes in single-SNP GWAS (113, 114), as our benchmark haplotype association analysis method. This method is used to obtain maximum likelihood estimates for haplotype frequencies using the expectation-maximization (EM) algorithm. Associations between haplotypes and disease risk are tested with a score statistic for a global test of haplotypes ($h$-1 degrees of freedom for $h$ haplotypes, treating the most common haplotype as a reference), with variance estimates that account for the uncertainty in the haplotype estimation. We used this method to exhaustively test 3-SNP haplotypes consisting of contiguous SNPs (with estimated haplotype counts of at least 20), using a 3-SNP sliding window strategy with a skip length of one SNP, among SNPs in gene analytic windows.

*Ancillary bioinformatics analyses*

We performed functional annotations and enrichment analyses with HaploReg v4 (115), Roadmap Epigenomics Mapping Consortium (REMC) (62), and Genotype-Tissue Expression

(GTEx Analysis v7) (63) resources for SNPs in replicated 3-SNP haplotypes. We conducted enrichment analyses using REMC histone modification ChIP-seq peak (gappedPeak algorithm) data for H3K4me3 (promoter) and H3K4me1 (enhancer) marks, and DNase I hypersensitivity peak (narrowPeak algorithm) data for all available consolidated blood/liver human cell types (29 and 11 cell types available, respectively). For each cell type, we compared the set of SNPs in replicated haplotypes with a set of 16036 non-overlapping SNPs mapped to gene analytic windows, each with nominal single-SNP associations ($P<0.05$) with PBC risk. Frequencies of overlap between SNPs in each set and epigenomic peaks were counted in each cell type. We evaluated enrichment for each assay using Bonferroni-corrected p-values obtained from 2-sided Fisher's exact tests. To investigate enrichments in gene expressions for SNPs in replicated haplotypes among the three blood- and liver-related tissue types available in GTEx, counts of significant *cis*-eQTLs (SNPs within +/-1 Mb of transcription start sites, q-value<0.05) for haplotype SNPs were compared to the aforementioned comparison SNP set using a 2-sided Fisher's exact test. Visualizations highlighting functional annotations of selected 3-SNP haplotypes were created with the "Gviz" R/Bioconductor package (116).

**Table 4.1: Selected examples of replicated 3-SNP haplotypes**

| | | | | Discovery (N=1937) | | | Replication (N=949) | | Combined (N=2886) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr | 3-SNP haplotype or logic tree | Gene analytic windows with 3-SNP haplotype | # SNPs in gene windows | Permutation-based selection statistic | OR | P | OR | P | OR | P | # with 0 haplotype copies (% cases) | # with 1 haplotype copy (% cases) | # with 2 haplotype copies (% cases) |
| 6 | (rs3129881=C) or ((rs375244=A) and (rs3132947=G)) | *CFB; NELFE; C2; C2-AS1* | 153-163 | -69.730 | 4.937 | $1.8x10^{-20}$ | 2.324 | $1.6x10^{-5}$ | 3.665 | $2.3x10^{-24}$ | 17 (17.6%) | 360 (21.7%) | 2509 (51.8%) |
| 6 | ((rs9268831=T) or (rs9269190=T)) or (rs9270652=C) | *ATF6B; FKBPL; PPT2; PPT2-EGFL8; AGPAT1* | 158-169 | -43.565 | 3.640 | $3.2x10^{-23}$ | 2.315 | $1.9x10^{-7}$ | 3.075 | $7.3x10^{-29}$ | 46 (15.2%) | 496 (26.0%) | 2344 (53.1%) |
| 6 | (rs9295704=C) and ((rs2451752=A) and (rs2575174=C))[a] | *BTN3A2; BTN2A2* | 46-52 | -12.263 | 0.362 | $1.7x10^{-10}$ | 0.374 | $1.3x10^{-5}$ | 0.365 | $8.9x10^{-15}$ | 2571 (50.4%) | 304 (27.3%) | 11 (9.1%) |
| 7 | ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G)[a] | *GLCCI1; LOC100505921; ICA1; COL28A1; MIOS; RPA3; LOC100505938; UMAD1; LOC101927391* | 204-292 | -16.424 | 0.040 | $8.7x10^{-6}$ | 0.112 | $3.6x10^{-4}$ | 0.066 | $3.9x10^{-9}$ | 2802 (49.1%) | 84 (6.0%) | 0 (NA) |
| 9 | (rs4979484=C) or ((rs13300483=T) and (rs7028891=G))[a] | *LOC100505478; TNFSF15; C9orf91; TNFSF8* | 121-156 | -28.431 | 1.746 | $4.2x10^{-17}$ | 1.528 | $8.6x10^{-6}$ | 1.672 | $3.0x10^{-21}$ | 775 (35.4%) | 1434 (48.7%) | 677 (60.4%) |

Abbreviation: #, number.

[a] Contains no HLA region SNPs.

**Table 4.2: Single SNP and component 2-SNP haplotype effects for example replicated 3-SNP haplotypes**

| Chr | 3-SNP haplotype or logic tree | Tree OR[b] | Tree P[b] | Single SNP | Alternative allele | SNP OR[c] | SNP P[c] | Component 2-SNP haplotype | Pair OR[d] | Pair P[d] |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | (rs3129881=C) or ((rs375244=A) and (rs3132947=G)) | 3.665 | $2.3\times10^{-24}$ | rs3129881 | C | 2.227 | $1.8\times10^{-23}$ | rs375244=A and rs3132947=G | 1.238 | $7.9\times10^{-5}$ |
| | | | | rs375244 | A | 0.980 | 0.715 | | | |
| | | | | rs3132947 | G | 2.117 | $1.4\times10^{-16}$ | | | |
| 6 | ((rs9268831=T) or (rs9269190=T)) or (rs9270652=C) | 3.075 | $7.3\times10^{-29}$ | rs9268831 | T | 1.324 | $1.6\times10^{-7}$ | rs9268831=T or rs9269190=T | 1.633 | $2.8\times10^{-16}$ |
| | | | | rs9269190 | T | 1.252 | $1.3\times10^{-4}$ | rs9268831=T or rs9270652=C | 1.321 | $2.0\times10^{-5}$ |
| | | | | rs9270652 | C | 1.109 | 0.057 | rs9269190=T or rs9270652=C | 1.925 | $1.2\times10^{-18}$ |
| 6 | (rs9295704=C) and ((rs2451752=A) and (rs2575174=C)) [a] | 0.365 | $8.9\times10^{-15}$ | rs9295704 | C | 0.669 | $1.5\times10^{-7}$ | rs9295704=C and rs2451752=A | 0.399 | $3.2\times10^{-14}$ |
| | | | | rs2451752 | A | 0.951 | 0.457 | rs9295704=C and rs2575174=C | 0.638 | $1.2\times10^{-7}$ |
| | | | | rs2575174 | C | 0.940 | 0.393 | rs2451752=A and rs2575174=C | 0.933 | 0.228 |
| 7 | ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G) [a] | 0.066 | $3.9\times10^{-9}$ | rs12671658 | T | 1.023 | 0.682 | rs12671658=T and rs11768586=G | 0.104 | $1.4\times10^{-6}$ |
| | | | | rs12702656 | A | 1.039 | 0.585 | rs12702656=A and rs11768586=G | 0.000 | 0.953 |
| | | | | rs11768586 | G | 0.865 | 0.010 | | | |
| 9 | (rs4979484=C) or ((rs13300483=T) and (rs7028891=G)) [a] | 1.672 | $3.0\times10^{-21}$ | rs4979484 | C | 1.365 | 0.005 | rs13300483=T and rs7028891=G | 1.637 | $1.2\times10^{-19}$ |
| | | | | rs13300483 | T | 1.584 | $1.4\times10^{-17}$ | | | |
| | | | | rs7028891 | G | 1.574 | $2.8\times10^{-17}$ | | | |

[a] Contains no HLA region SNPs.
[b] 3-SNP haplotype OR and p-value in the combined sample (N=2886).
[c] Single SNP ORs and p-values, assuming an additive genetic effect model for the specified alternative allele.
[d] 2-SNP haplotype ORs and p-values, assuming an additive genetic effect model for the specified haplotype pattern.

**Table 4.3: Comparison of logic regression and benchmark methods to detect 3-SNP haplotypes in the discovery cohort, N=1937**

| | Method A (proposed): Logic regression | | | Method B (benchmark): 3-SNP sliding windows | | | | Comparison |
|---|---|---|---|---|---|---|---|---|
| Chr | # Gene windows with replicated 3-SNP haplotype (Method A) | Gene window with best p-value | Best p-value | # Tests with $P<3.2\times10^{-8}$ (Bonferroni) | # Gene windows with at least one haplotype with $P<3.2\times10^{-8}$ | Gene window with best p-value | Best p-value | # Gene window matches between methods A and B |
| 2 | 0 | NA | NA | 1 | 1 | *LRP1B* | $1.1\times10^{-9}$ | 0 |
| 3 | 0 | NA | NA | 5 | 3 | *NEK10* | $3.6\times10^{-18}$ | 0 |
| 6 | 143 | *NOTCH4* | $6.1\times10^{-27}$ | 1352 | 173 | *TAAR2* | $2.2\times10^{-30}$ | 123 |
| 7 | 9 | *GLCCI1* | $8.7\times10^{-6}$ | 6 | 3 | *HGF* | $2.1\times10^{-15}$ | 0 |
| 8 | 0 | NA | NA | 4 | 4 | *CYP11B1* | $1.1\times10^{-8}$ | 0 |
| 9 | 12 | *DEC1* | $2.5\times10^{-17}$ | 74 | 16 | *DEC1* | $3.2\times10^{-13}$ | 12 |
| 18 | 0 | NA | NA | 10 | 5 | *MTCL1* | $1.6\times10^{-10}$ | 0 |

Abbreviation: #, number.

**Table 4.4: Functional annotations of SNPs in selected replicated 3-SNP haplotypes**

| Chr | 3-SNP haplotype | SNP | Chr position (hg19) | Ontology | Mapped gene | DHS overlap[a], # EIDs (# PBC EIDs) | H3K4me1 overlap[a], # EIDs (# PBC EIDs) | H3K4me3 overlap[a], # EIDs (# PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | (rs3129881=C) or ((rs375244=A) and (rs3132947=G)) | rs3129881 | 32409484 | intronic | *HLA-DRA* | 0 (0) | 42 (20) | 50 (17) | GM12878 (OCT2, POL2, POL24H8, POU2F2); GM12891 (OCT2, POL2, POL24H8, POU2F2); GM12892 (POL2, POL24H8) | 3 | Whole Blood (*C4A, C4B, HLA-DQA1, HLA-DQA2, HLA-DQB2, HLA-DRB5*); Lymphoblastoid (*HLA-DQA2, HLA-DRB5*); Liver (*HLA-DRB5*) |
| | | rs375244 | 32191457 | intronic | *NOTCH4* | 0 (0) | 69 (15) | 38 (3) | NA | 3 | Whole Blood (*GPSM3, NOTCH4*) |
| | | rs3132947 | 32176782 | intronic | *NOTCH4* | 0 (0) | 2 (1) | 4 (0) | NA | 3 | Whole Blood (*AGPAT1, C4A, C4B, CYP21A1P, HLA-DRA, PBX2, SKIV2L*); Lymphoblastoid (*C4A, HLA-DQA1, RNF5*); Liver (*AGPAT1, RNF5*) |
| 6 | ((rs9268831=T) or (rs9269190=T)) or (rs9270652=C) | rs9268831 | 32427748 | intergenic | *HLA-DRA* (dist=14922), *HLA-DRB5* (dist=57406) | 5 (1) | 31 (19) | 71 (16) | GM18951 (POL2) | 1 | Whole Blood (*HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DQB2, HLA-DRB1, HLA-DRB6, HLA-DRB9*); Lymphoblastoid (*HLA-DQA2, HLA-DQB2, HLA-DRB6, HLA-DRB9, NOTCH4*); Liver (*HLA-DQA2, HLA-DQB2*) |
| | | rs9269190 | 32448500 | intergenic | *HLA-DRA* (dist=35674), *HLA-DRB5* (dist=36654) | 2 (2) | 4 (4) | 2 (2) | NA | 2 | Whole Blood *(HLA-DRA*) |
| | | rs9270652 | 32565905 | intergenic | *HLA-DRB1* (dist=8292), *HLA-DQA1* (dist=39278) | 1 (1) | 1 (1) | 2 (0) | NA | 0 | NA |
| 6 | (rs9295704=C) and ((rs2451752=A) and (rs2575174=C)) | rs9295704[e] | 26704816 | intergenic | *ZNF322* (dist=44836), *GUSBP2* (dist=134450) | 0 (0) | 13 (1) | 8 (1) | NA | 4 | Whole Blood (*ABT1*) |
| | | rs2451752[e] | 26648013 | intronic | ZNF322 | 0 (0) | 1 (1) | 3 (1) | NA | 0 | Whole Blood (*BTN3A1, BTN3A2, HMGN4, ZNF322*) |
| | | rs2575174[e] | 25885552 | intergenic | *SLC17A3* (dist=11081), *SLC17A2* (dist=27432) | 0 (0) | 5 (0) | 0 (0) | NA | 2 | Whole Blood (*HIST1H1T, HIST1H4A*) |
| 7 | ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G) | rs12671658[e] | 7842281 | intronic | *UMAD1* | 1 (0) | 6 (0) | 2 (0) | NA | 7 | NA |
| | | rs12702656[e] | 7851742 | intronic | *UMAD1* | 0 (0) | 9 (3) | 0 (0) | NA | 4 | NA |
| | | rs11768586[e] | 7849806 | intronic | *UMAD1* | 1 (0) | 6 (2) | 1 (1) | NA | 0 | NA |
| 9 | (rs4979484=C) or ((rs13300483=T) and (rs7028891=G)) | rs4979484[e] | 117751450 | intergenic | *TNFSF8* (dist=58575), *TNC* (dist=30404) | 22 (6) | 33 (20) | 11 (7) | GM12878 (BATF, NFKB); GM12891 (NFKB); GM15510 (NFKB); GM18951 (NFKB) | 3 | NA |
| | | rs13300483[e] | 117643362 | intergenic | *TNFSF15* (dist=74954), *TNFSF8* (dist=12261) | 0 (0) | 11 (6) | 2 (0) | NA | 2 | Whole Blood (*TNFSF8*) |
| | | rs7028891[e] | 117645015 | intergenic | *TNFSF15* (dist=76607), *TNFSF8* (dist=10608) | 0 (0) | 4 (3) | 1 (1) | NA | 3 | Whole Blood (*TNFSF8*) |

Abbreviations: EID, epigenome identifier; dist, distance; #, number; DHS, DNase I hypersensitivity site; eQTL, expression quantitative trait loci.

[a] Counts of the number of consolidated cell types (EIDs) for which the SNP of interest overlaps the queried epigenomic assay peak (Roadmap Epigenomics Mapping Consortium). "PBC EID": Separately considers peak overlap counts among the 29 blood/liver cell types available in Roadmap Epigenomics.

[b] Bound protein: Regulatory protein-binding ChIP-seq peak overlaps for specified proteins are provided for blood- or liver-related cell lines only (HaploReg v4).

[c] Altered motifs: The number of regulatory motifs predicted to be affected by the SNP based on position weight matrices (PWM) score changes (HaploReg v4).

[d] eQTLs: Reported significant eQTLs for whole blood, lymphoblastoid, and liver cell types only (GTEx Consortium; HaploReg v4).

[e] Signifies non-HLA SNPs.

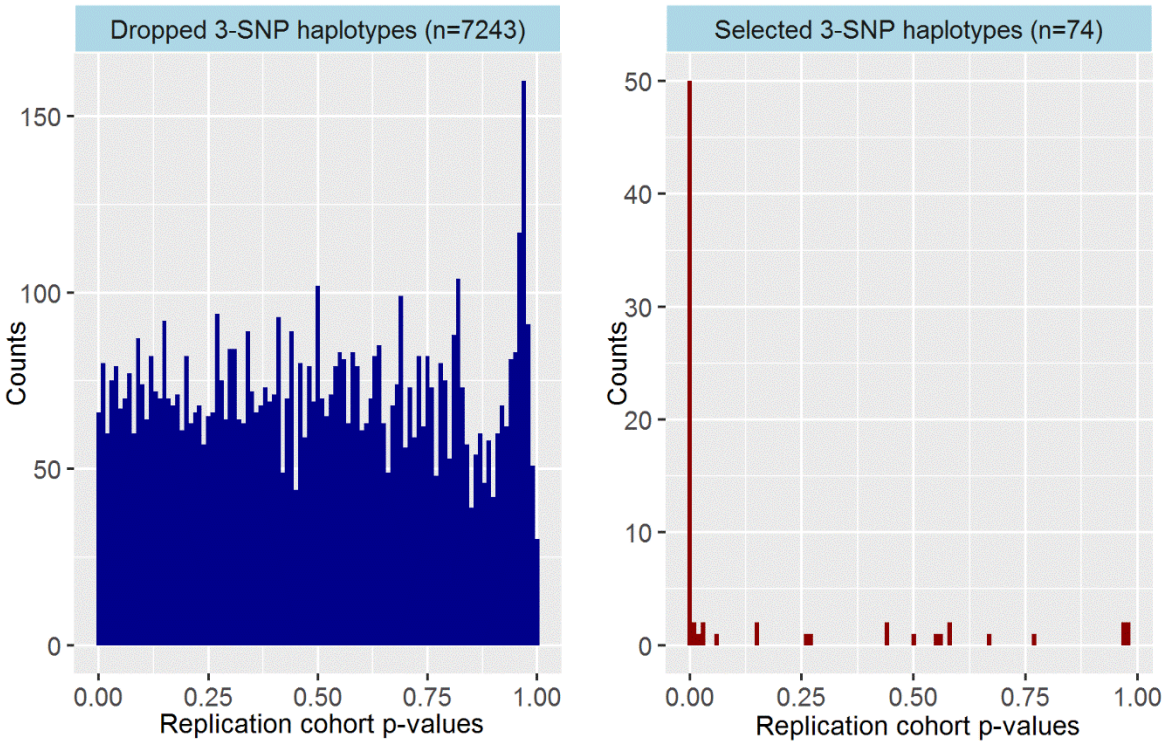



**Figure 4.1: Distributions of haplotype association test p-values for dropped versus selected 3-SNP haplotypes in the replication cohort**

Side-by-side histograms of haplotype association p-values for dropped haplotypes and selected haplotypes in our replication cohort (N=949) are provided for comparison. Selected 3-SNP haplotypes have the top percentile of permutation-based evaluation statistics (values less than -11.4).

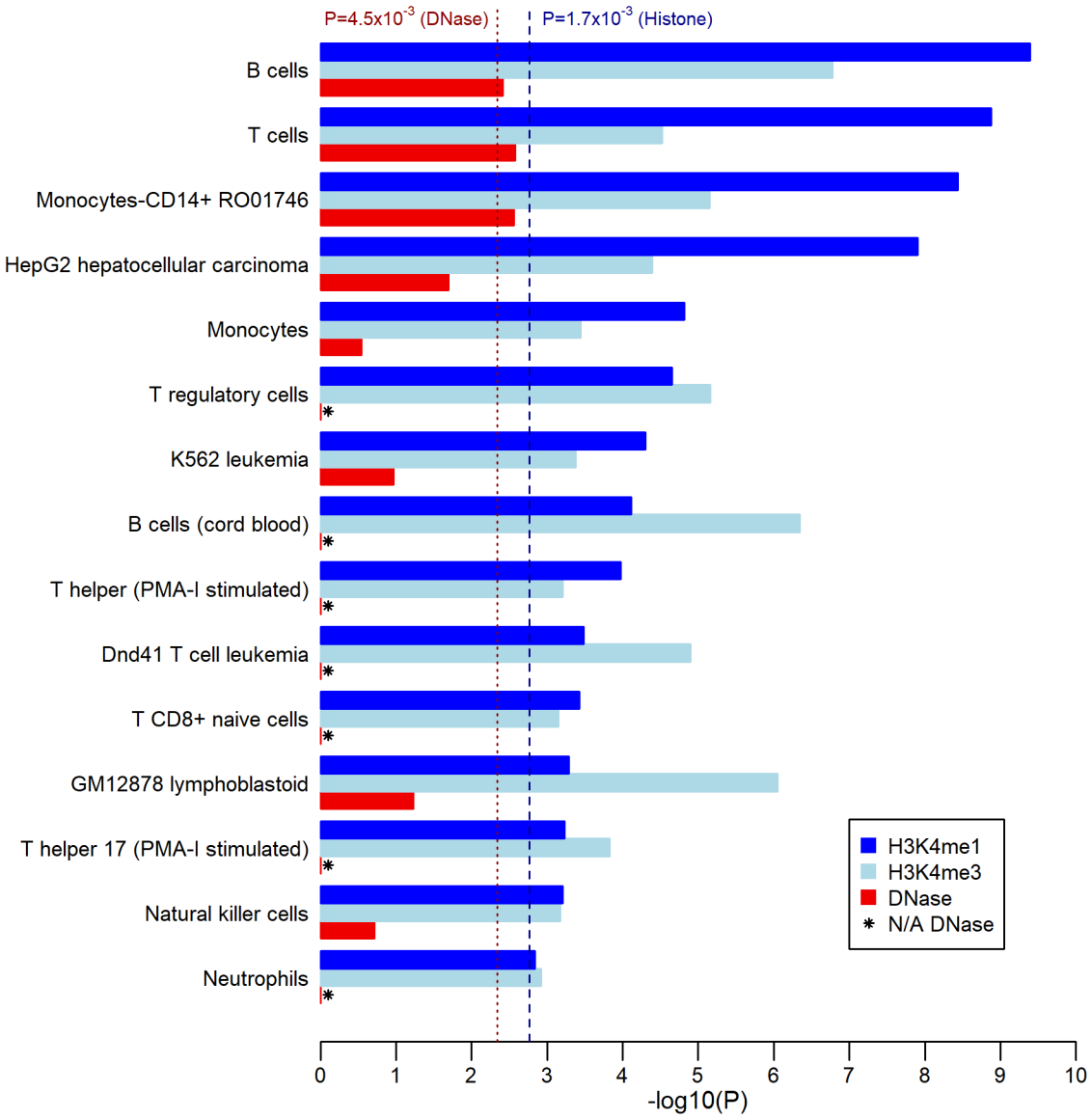



**Figure 4.2: Selection of histone modification and DNase peak enrichment analysis results**

The enrichment analyses compared 106 SNPs in replicated 3-SNP haplotypes to nominally associated single SNPs in gene regions. Figure shows enrichment test p-values that are log-transformed ($-\log_{10}(P)$) for the 15 blood/liver cell types for which haplotype SNPs are significantly enriched for both H3K4me1 and H3K4me3 peaks. Dashed and dotted vertical lines show Bonferroni-corrected p-value thresholds based on the number of blood/liver cell types for which Roadpmap Epigenomics assay data was available (29 and 11 types for histone mark and DNase peaks, respectively).

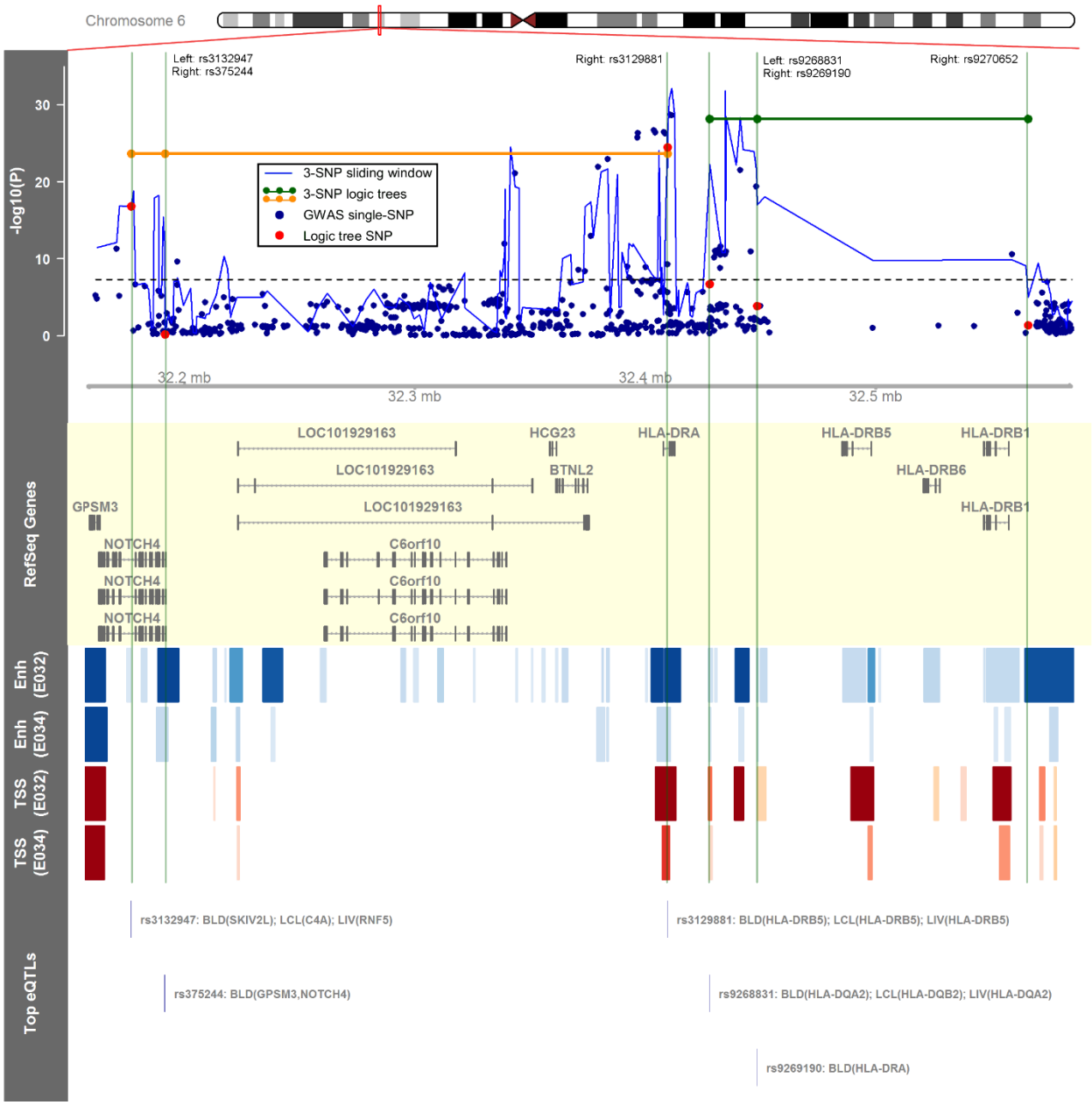


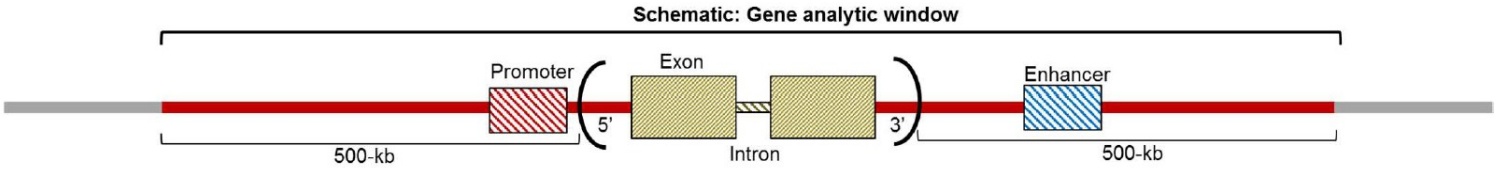

**Figure 4.3: Visualization of two replicated 3-SNP haplotype logic trees containing SNPs in the HLA region**

The plotted region spans chr6:32156782-32585905 (hg19), which corresponds with the red rectangle in the chromosomal ideogram. The top data track shows single-SNP association results and haplotype association results using benchmark ("3-SNP sliding window") and proposed ("3-SNP logic tree") methods for the selected genomic region. The subsequent annotation tracks show RefSeq genes, a heatmap corresponding to Roadmap Epigenomics H3K4me1 (Enh) and H3K4me3 (TSS) ChIP-seq peaks in primary B cells (E032) and primary T cells (E034), and the top significant blood (BLD), lymphoblastoid (LCL), and liver (LIV) eQTLs (expression quantitative trait loci) associated with SNPs in replicated haplotypes (GTEx Consortium).

## 4.5 Supplementary Information

### Chapter 4 Supplementary Figures and Tables

**Figure S1:** Example schematic of RefSeq gene-based analytic windows used for haplotype formation in the main analysis. The bracketed region highlighted in red, which includes the RefSeq gene transcript and flanking 500-kb regions before and after transcription start/stop sites, spans typical genomic elements included in any given gene analytic window.



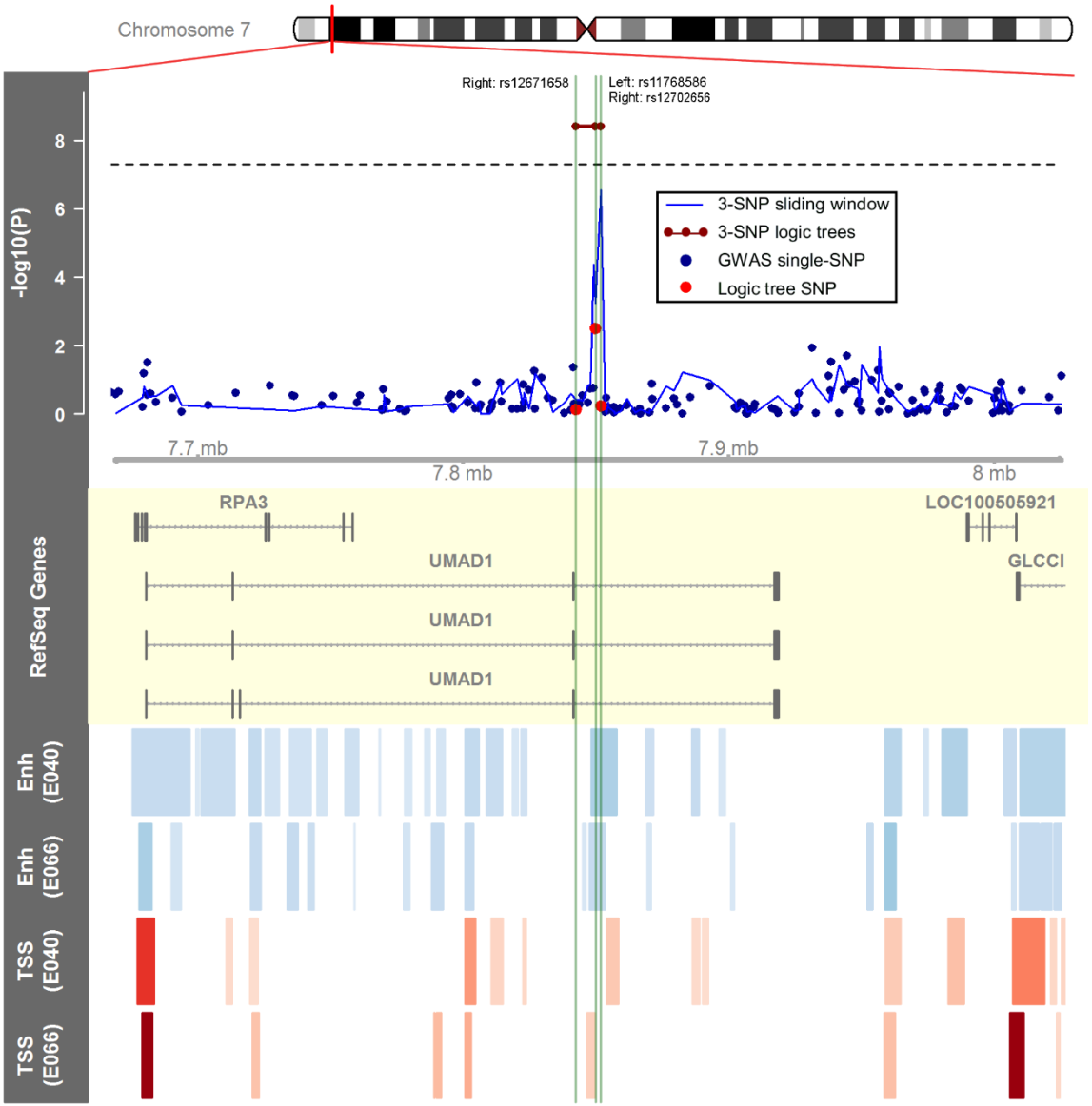



**Figure S2:** Visualization of the region surrounding the replicated 3-SNP haplotype logic tree for chromosome 7 (chr7:7667281-8026742, hg19; corresponds to the red band in the chromosomal ideogram). The top data track shows single-SNP association results and haplotype association results using benchmark ("3-SNP sliding window") and proposed ("3-SNP logic tree") methods for the selected genomic region. The annotation tracks beneath show RefSeq genes, a heatmap corresponding to Roadmap Epigenomics H3K4me1 (Enh or enhancer) and H3K4me3 (TSS or promoter) ChIP-seq peaks in primary T helper memory cells (E040) and liver cells (E066). No significant eQTLs were associated with SNPs in replicated haplotypes.

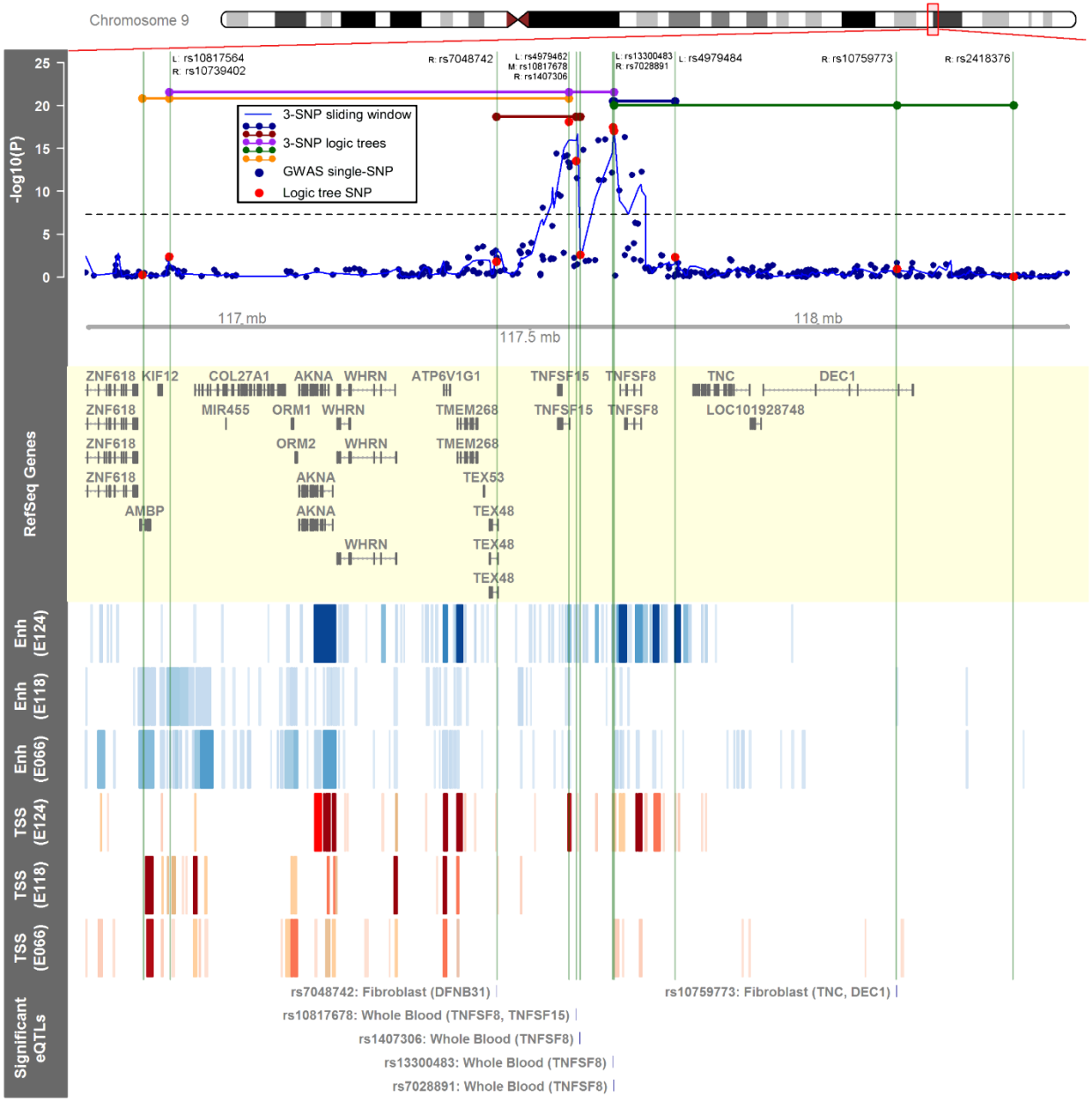



**Figure S3:** Visualization of the region surrounding the five replicated 3-SNP haplotype logic trees in chromosome 9 (chr9:116727079-118438852, hg19; corresponds to the red rectangle in the chromosomal ideogram). The top data track shows single-SNP association results and haplotype association results using benchmark ("3-SNP sliding window") and proposed ("3-SNP logic tree") methods for the selected genomic region. The annotation tracks beneath show RefSeq genes, a heatmap corresponding to Roadmap Epigenomics H3K4me1 (Enh or enhancer) and H3K4me3 (TSS or promoter) ChIP-seq peaks in monocytes (E124), hepatocellular carcinoma cells or HepG2 (E118), and liver cells (E066), and all significant eQTLs associated with SNPs in replicated haplotypes in blood, liver, and fibroblast cell types (GTEx Consortium).

**Table S1**: Summary of discovery and replication analysis results for the 49 replicated 3-SNP haplotypes

| | | | | DISCOVERY (N=1937) | | | REPLICATION (N=949) | | COMBINED (N=2886) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr | 3-SNP haplotype tree | RefSeq gene windows with 3-SNP haplotype | # SNPs in gene windows | Permutation-based selection statistic | OR | P | OR | P | OR | P | # with 0 haplotype copies (% cases) | # with 1 haplotype copy (% cases) | # with 2 haplotype copies (% cases) |
| 6 | (rs3117106=C) and ((rs206018=G) or (rs9501179=A)) | *SLC44A4; EHMT2* | 148-158 | -82.148 | 0.169 | $3.3x10^{-17}$ | 0.257 | $2.7x10^{-7}$ | 0.196 | $3.0x10^{-23}$ | 2609 (51.2%) | 264 (17.4%) | 13 (0.0%) |
| 6 | (rs3129881=C) or ((rs375244=A) and (rs3132947=G)) | *CFB; NELFE; C2; C2-AS1* | 153-163 | -69.730 | 4.937 | $1.8x10^{-20}$ | 2.324 | $1.6x10^{-5}$ | 3.665 | $2.3x10^{-24}$ | 17 (17.6%) | 360 (21.7%) | 2509 (51.8%) |
| 6 | ((rs35372932=T) and (rs9269190=C)) or (rs9270493=C) | *TNXB* | 157 | -58.438 | 0.325 | $4.3x10^{-25}$ | 0.451 | $1.7x10^{-8}$ | 0.365 | $9.6x10^{-32}$ | 2162 (54.3%) | 656 (29.7%) | 68 (16.2%) |
| 6 | (rs9268977=T) or ((rs3135395=T) or (rs550513=T)) | *SKIV2L* | 159 | -57.402 | 3.741 | $1.7x10^{-21}$ | 2.991 | $1.0x10^{-9}$ | 3.448 | $1.2x10^{-29}$ | 43 (16.3%) | 439 (23.0%) | 2404 (53.0%) |
| 6 | ((rs9268634=G) or (rs35344500=C)) and (rs9275175=G) | *NOTCH4; GPSM3* | 177-193 | -47.911 | 0.417 | $6.1x10^{-27}$ | 0.500 | $6.8x10^{-11}$ | 0.448 | $1.3x10^{-35}$ | 1606 (57.8%) | 1069 (38.2%) | 211 (21.3%) |
| 6 | ((rs9268831=T) or (rs9269190=T)) or (rs9270652=C) | *ATF6B; FKBPL; PPT2; PPT2-EGFL8; AGPAT1* | 158-169 | -43.565 | 3.640 | $3.2x10^{-23}$ | 2.315 | $1.9x10^{-7}$ | 3.075 | $7.3x10^{-29}$ | 46 (15.2%) | 496 (26.0%) | 2344 (53.1%) |
| 6 | (rs3132947=G) or ((rs9501179=G) and (rs41546114=C)) | *C6orf48; SNORD48* | 162 | -42.196 | 6.316 | $1.2x10^{-16}$ | 3.533 | $7.9x10^{-7}$ | 5.030 | $2.4x10^{-22}$ | 11 (0.0%) | 257 (17.5%) | 2618 (51.0%) |
| 6 | (rs9268831=T) or ((rs2395194=G) or (rs387608=A)) | *STK19* | 160 | -38.720 | 3.869 | $8.0x10^{-22}$ | 3.169 | $4.3x10^{-10}$ | 3.601 | $2.2x10^{-30}$ | 42 (16.7%) | 431 (22.0%) | 2413 (53.0%) |
| 6 | (rs3132947=G) or ((rs9268014=C) and (rs41546114=C)) | *MSH5* | 181 | -38.294 | 6.221 | $5.3x10^{-17}$ | 3.533 | $7.9x10^{-7}$ | 5.012 | $1.0x10^{-22}$ | 12 (0.0%) | 261 (17.6%) | 2613 (51.1%) |
| 6 | (rs3132947=G) or ((rs9268055=T) and (rs41546114=C)) | *MSH5-SAPCD1* | 181 | -38.294 | 6.221 | $5.3x10^{-17}$ | 3.533 | $7.9x10^{-7}$ | 5.012 | $1.0x10^{-22}$ | 12 (0.0%) | 261 (17.6%) | 2613 (51.1%) |
| 6 | ((rs241437=G) or (rs3129299=C)) or (rs9276909=C) | *PFDN6; WDR46; VPS52; DAXX; RPS18; HCG25* | 125-139 | -37.551 | 6.468 | $5.0x10^{-16}$ | 2.912 | $2.2x10^{-5}$ | 4.716 | $1.8x10^{-20}$ | 16 (0.0%) | 237 (18.6%) | 2633 (50.8%) |
| 6 | ((rs2395194=G) and (rs16870908=G)) or (rs9268831=T) | *HLA-DMA; HLA-DQB2; LOC100294145; C6orf10; HLA-DOB; TAP2; HLA-DQA2; HCG23; HLA-DMB; BTNL2; PSMB9; TAP1; HLA-DQA1; HLA-DQB1; HLA-DRA; HLA-DRB6* | 197-283 | -35.650 | 2.995 | $1.7x10^{-25}$ | 2.292 | $4.4x10^{-9}$ | 2.732 | $5.9x10^{-33}$ | 72 (11.1%) | 690 (31.2%) | 2124 (54.5%) |
| 6 | ((rs9268213=A) and (rs41546114=C)) or (rs3132947=G) | *HSPA1L* | 158 | -35.026 | 6.221 | $5.3x10^{-17}$ | 3.533 | $7.9x10^{-7}$ | 5.012 | $1.0x10^{-22}$ | 12 (0.0%) | 261 (17.6%) | 2613 (51.1%) |
| 6 | ((rs4538748=C) or (rs2064476=G)) and (rs35344500=A) | *HLA-DOA; BRD2; HLA-DPB2; HLA-DPA1; HLA-DPB1* | 187-192 | -34.656 | 2.406 | $9.3x10^{-26}$ | 2.136 | $5.4x10^{-11}$ | 2.313 | $3.3x10^{-35}$ | 173 (20.8%) | 941 (35.8%) | 1772 (56.9%) |
| 6 | ((rs2244027=A) or (rs2242665=C)) or (rs2071591=G) | *MICB; LY6G5B; PRRC2A; ATP6V1G2-DDX39B; DDX39B; MCCD1; NFKBIL1; AIF1; ATP6V1G2; DDX39B-AS1; LST1; LTA; APOM; CSNK2B; GPANK1; BAG6; HCP5; MICA* | 152-203 | -27.403 | 2.943 | $3.4x10^{-19}$ | 2.659 | $9.3x10^{-8}$ | 2.873 | $8.7x10^{-26}$ | 33 (21.2%) | 509 (26.7%) | 2344 (52.8%) |
| 6 | ((rs9296088=A) and (rs12207818=C)) and (rs181997=A) | *KIFC1; SYNGAP1* | 118-123 | -26.700 | 0.142 | $1.3x10^{-14}$ | 0.377 | $6.1x10^{-4}$ | 0.204 | $6.6x10^{-18}$ | 2671 (50.4%) | 205 (17.6%) | 10 (0.0%) |
| 6 | ((rs2071287=C) and (rs3094596=C)) or (rs185819=C) | *ABHD16A* | 152 | -22.999 | 3.720 | $6.5x10^{-18}$ | 2.882 | $3.2x10^{-8}$ | 3.370 | $1.5x10^{-24}$ | 29 (3.4%) | 379 (25.1%) | 2478 (51.9%) |
| 6 | (rs3132947=T) and ((rs404860=C) or (rs41546114=T)) | *C6orf25* | 169 | -21.846 | 0.186 | $8.2x10^{-17}$ | 0.382 | $7.7x10^{-6}$ | 0.253 | $3.2x10^{-21}$ | 2584 (51.0%) | 288 (21.5%) | 14 (0.0%) |
| 6 | (rs11754586=T) or ((rs3131932=G) or (rs9262537=G)) | *PPP1R10; ABCF1; ATAT1* | 138-148 | -21.366 | 3.889 | $1.0x10^{-15}$ | 2.396 | $2.6x10^{-5}$ | 3.241 | $1.8x10^{-19}$ | 13 (7.7%) | 327 (24.5%) | 2546 (51.1%) |

| | | | | DISCOVERY (N=1937) | | | REPLICATION (N=949) | | COMBINED (N=2886) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr | 3-SNP haplotype tree | RefSeq gene windows with 3-SNP haplotype | # SNPs in gene windows | Permutation-based selection statistic | OR | P | OR | P | OR | P | # with 0 haplotype copies (% cases) | # with 1 haplotype copy (% cases) | # with 2 haplotype copies (% cases) |
| 6 | ((rs10947251=A) or (rs41546114=T)) and (rs3132947=T) | *VARS* | 173 | -20.943 | 0.166 | $1.8x10^{-16}$ | 0.278 | $5.4x10^{-7}$ | 0.203 | $2.8x10^{-22}$ | 2616 (51.0%) | 258 (17.8%) | 12 (0.0%) |
| 6 | ((rs3132947=G) or (rs9266632=G)) or (rs185819=C) | *LY6G6E* | 154 | -20.783 | 5.008 | $3.8x10^{-17}$ | 3.426 | $1.0x10^{-7}$ | 4.321 | $2.0x10^{-23}$ | 14 (0.0%) | 298 (20.1%) | 2574 (51.3%) |
| 6 | (rs805273=A) and ((rs805267=G) or (rs9266774=T)) | *HCG27* | 222 | -20.015 | 0.229 | $8.2x10^{-17}$ | 0.337 | $7.2x10^{-7}$ | 0.264 | $3.1x10^{-22}$ | 2555 (51.3%) | 317 (22.4%) | 14 (0.0%) |
| 6 | (rs3828901=A) or ((rs707921=A) and (rs9266774=T)) | *CCHCR1; TCF19* | 222 | -19.687 | 0.231 | $6.9x10^{-17}$ | 0.382 | $6.0x10^{-6}$ | 0.279 | $2.5x10^{-21}$ | 2551 (51.2%) | 322 (23.3%) | 13 (0.0%) |
| 6 | (rs3828901=G) and ((rs805268=A) or (rs9266774=C)) | *POU5F1* | 222 | -19.444 | 4.354 | $1.1x10^{-16}$ | 2.618 | $6.9x10^{-6}$ | 3.586 | $4.4x10^{-21}$ | 12 (0.0%) | 319 (23.2%) | 2555 (51.2%) |
| 6 | (rs3828901=A) or ((rs3828919=A) and (rs2523497=T)) | *MUC22* | 219 | -19.050 | 0.231 | $4.2x10^{-17}$ | 0.479 | $1.6x10^{-4}$ | 0.313 | $9.0x10^{-20}$ | 2531 (51.1%) | 345 (25.2%) | 10 (0.0%) |
| 6 | (rs707922=G) or ((rs1046089=A) and (rs9266774=C)) | *PSORS1C3* | 222 | -17.984 | 4.314 | $1.6x10^{-16}$ | 2.967 | $7.2x10^{-7}$ | 3.748 | $6.2x10^{-22}$ | 14 (0.0%) | 315 (22.5%) | 2557 (51.2%) |
| 6 | ((rs3131932=G) or (rs28360042=C)) or (rs885950=C) | *NRM; MDC1; MDC1-AS1; DHX16; PPP1R18; TUBB; FLOT1; IER3; DPCR1; DDR1; GTF2H4; LINC00243* | 146-211 | -17.890 | 3.907 | $7.4x10^{-16}$ | 2.650 | $2.3x10^{-6}$ | 3.356 | $1.1x10^{-20}$ | 14 (0.0%) | 334 (24.6%) | 2538 (51.2%) |
| 6 | ((rs9266774=T) and (rs2255741=A)) or (rs3828901=A) | *PSORS1C2* | 221 | -17.684 | 0.234 | $1.4x10^{-16}$ | 0.371 | $3.4x10^{-6}$ | 0.279 | $2.6x10^{-21}$ | 2551 (51.2%) | 323 (23.2%) | 12 (0.0%) |
| 6 | (rs3749946=A) and ((rs2517506=T) and (rs7741091=A)) | *MUC21* | 209 | -17.342 | 0.234 | $3.1x10^{-16}$ | 0.394 | $9.8x10^{-6}$ | 0.287 | $1.9x10^{-20}$ | 2554 (51.1%) | 322 (23.6%) | 10 (0.0%) |
| 6 | (rs3828901=G) and ((rs12110785=T) or (rs2523644=T)) | *PSORS1C1; HCG22* | 218-223 | -15.815 | 4.572 | $2.9x10^{-16}$ | 2.427 | $3.0x10^{-5}$ | 3.544 | $6.3x10^{-20}$ | 10 (0.0%) | 306 (23.2%) | 2570 (51.0%) |
| 6 | ((rs2239888=T) or (rs3134769=C)) and (rs3828901=G) | *CDSN* | 221 | -14.656 | 4.240 | $1.2x10^{-16}$ | 2.578 | $7.7x10^{-6}$ | 3.510 | $5.5x10^{-21}$ | 11 (0.0%) | 327 (23.5%) | 2548 (51.2%) |
| 6 | (rs2517681=T) or ((rs4148248=C) and (rs2735078=A)) | *TRIM10; TRIM26; TRIM31; TRIM31-AS1; TRIM40; TRIM15* | 93-96 | -13.047 | 3.254 | $5.1x10^{-13}$ | 3.065 | $1.1x10^{-6}$ | 3.196 | $2.6x10^{-18}$ | 13 (7.7%) | 309 (24.6%) | 2564 (50.9%) |
| 6 | (rs17195733=G) or ((rs13201129=C) and (rs1737069=T)) | *HLA-L* | 93 | -13.034 | 0.361 | $2.0x10^{-14}$ | 0.364 | $2.2x10^{-7}$ | 0.360 | $1.8x10^{-20}$ | 2444 (51.7%) | 419 (27.2%) | 23 (17.4%) |
| 6 | (rs3130785=T) and ((rs9261301=G) or (rs1264570=C)) | *RPP21* | 100 | -12.876 | 0.291 | $6.0x10^{-14}$ | 0.441 | $6.7x10^{-5}$ | 0.340 | $2.7x10^{-17}$ | 2549 (50.8%) | 323 (25.7%) | 14 (14.3%) |
| 6 | (rs2517681=C) and ((rs3130785=T) or (rs2735078=G)) | *TRIM39; TRIM39-RPP21* | 103 | -11.587 | 0.289 | $4.3x10^{-14}$ | 0.376 | $5.4x10^{-6}$ | 0.317 | $1.2x10^{-18}$ | 2552 (51.0%) | 320 (24.4%) | 14 (14.3%) |
| 6 | ((rs1345229=A) or (rs1233387=T)) or (rs1003581=G) | *HCG4; LOC554223; HLA-F; HLA-G* | 113-117 | -11.537 | 3.168 | $6.9x10^{-13}$ | 2.885 | $1.7x10^{-6}$ | 3.070 | $5.4x10^{-18}$ | 14 (7.1%) | 318 (25.5%) | 2554 (50.9%) |
| 6 | ((rs4713429=G) or (rs12110785=T)) or (rs3131932=G) | *GNL1; PRR3* | 139 | -11.517 | 3.661 | $8.4x10^{-16}$ | 2.344 | $2.0x10^{-5}$ | 3.093 | $1.0x10^{-19}$ | 17 (0.0%) | 346 (26.3%) | 2523 (51.1%) |
| 6 | ((rs1003581=G) and (rs16894681=T)) or (rs1233387=T)[a] | *OR2H2; UBD; OR10C1; OR11A1; OR12D2; OR12D3; OR5V1; OR2H1; GABBR1; OR2J2; OR2J3* | 78-110 | -14.375 | 2.697 | $6.4x10^{-13}$ | 2.255 | $1.9x10^{-5}$ | 2.541 | $5.5x10^{-17}$ | 18 (5.6%) | 401 (29.9%) | 2467 (51.1%) |
| 6 | ((rs2281043=T) or (rs7751451=G)) and (rs1635=A)[a] | *ZSCAN12; ZKSCAN3; PGBD1; ZSCAN31; ZSCAN23; ZBED9; GPX5; GPX6* | 50-53 | -14.306 | 0.377 | $2.7x10^{-11}$ | 0.360 | $1.7x10^{-6}$ | 0.371 | $2.1x10^{-16}$ | 2527 (50.8%) | 341 (27.9%) | 18 (11.1%) |

| | | | | DISCOVERY (N=1937) | | | REPLICATION (N=949) | | COMBINED (N=2886) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr | 3-SNP haplotype tree | RefSeq gene windows with 3-SNP haplotype | # SNPs in gene windows | Permutation-based selection statistic | OR | P | OR | P | OR | P | # with 0 haplotype copies (% cases) | # with 1 haplotype copy (% cases) | # with 2 haplotype copies (% cases) |
| 6 | ((rs3117192=C) or (rs16894216=T)) and (rs7773193=T)[a] | *ZNF311; LINC01556; OR2W1* | 61-67 | -14.047 | 2.709 | $2.6x10^{-11}$ | 2.728 | $1.5x10^{-6}$ | 2.711 | $2.0x10^{-16}$ | 17 (11.8%) | 340 (27.6%) | 2529 (50.8%) |
| 6 | (rs9295704=C) and ((rs2451752=A) and (rs2575174=C))[a] | *BTN3A2; BTN2A2* | 46-52 | -12.263 | 0.362 | $1.7x10^{-10}$ | 0.374 | $1.3x10^{-5}$ | 0.365 | $8.9x10^{-15}$ | 2571 (50.4%) | 304 (27.3%) | 11 (9.1%) |
| 6 | (rs6939576=G) or ((rs2859365=G) or (rs6930033=A))[a] | *TRIM27* | 64 | -12.215 | 2.681 | $9.9x10^{-11}$ | 3.120 | $3.3x10^{-7}$ | 2.820 | $1.7x10^{-16}$ | 20 (15.0%) | 312 (26.3%) | 2554 (50.7%) |
| 6 | (rs7773193=C) or ((rs17280818=T) and (rs2394100=T))[a] | *NKAPL; ZSCAN26; ZKSCAN4; ZSCAN9* | 52-53 | -11.424 | 0.383 | $4.3x10^{-11}$ | 0.383 | $3.2x10^{-6}$ | 0.382 | $6.3x10^{-16}$ | 2524 (50.8%) | 342 (28.7%) | 20 (10.0%) |
| 7 | ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G)[a] | *GLCCI1; LOC100505921; ICA1; COL28A1; MIOS; RPA3; LOC100505938; UMAD1; LOC101927391* | 204-292 | -16.424 | 0.040 | $8.7x10^{-6}$ | 0.112 | $3.6x10^{-4}$ | 0.066 | $3.9x10^{-9}$ | 2802 (49.1%) | 84 (6.0%) | 0 (NA) |
| 9 | (rs4979484=C) or ((rs13300483=T) and (rs7028891=G))[a] | *LOC100505478; TNFSF15; C9orf91; TNFSF8* | 121-156 | -28.431 | 1.746 | $4.2x10^{-17}$ | 1.528 | $8.6x10^{-6}$ | 1.672 | $3.0x10^{-21}$ | 775 (35.4%) | 1434 (48.7%) | 677 (60.4%) |
| 9 | (rs7028891=G) and ((rs4979462=T) or (rs10739402=T))[a] | *ATP6V1G1; AKNA; DFNB31* | 116-133 | -26.160 | 1.749 | $3.1x10^{-17}$ | 1.597 | $9.8x10^{-7}$ | 1.696 | $2.7x10^{-22}$ | 870 (36.0%) | 1413 (49.3%) | 603 (61.7%) |
| 9 | (rs7028891=A) and ((rs2418376=A) or (rs10759773=G))[a] | *DEC1; TNC; LOC101928748* | 153-189 | -22.728 | 0.569 | $2.5x10^{-17}$ | 0.674 | $2.3x10^{-5}$ | 0.604 | $9.3x10^{-21}$ | 873 (59.1%) | 1398 (46.7%) | 615 (34.5%) |
| 9 | (rs10817678=G) or ((rs1407306=T) and (rs7048742=A))[a] | *ORM2* | 111 | -21.730 | 0.586 | $6.6x10^{-16}$ | 0.672 | $3.2x10^{-5}$ | 0.613 | $2.1x10^{-19}$ | 960 (57.7%) | 1380 (46.7%) | 546 (33.5%) |
| 9 | (rs4979462=T) or ((rs10739402=T) and (rs10817564=C))[a] | *COL27A1* | 128 | -12.015 | 1.687 | $4.7x10^{-15}$ | 1.699 | $3.8x10^{-8}$ | 1.687 | $1.5x10^{-21}$ | 532 (33.1%) | 1390 (46.0%) | 964 (58.6%) |

[a] Contains no HLA region SNPs

Abbreviations: #, number

**Table S2**: Component single SNPs and SNP pair haplotype effects for replicated 3-SNP haplotypes

| Chr | 3-SNP haplotype tree | Haplotype OR[a] | Haplotype P[a] | Single SNP | Alt. allele | SNP OR[b] | SNP P[b] | Component haplotype pair | Pair OR[c] | Pair P[c] |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | (rs3117106=C) and ((rs206018=G) or (rs9501179=A)) | 0.196 | $3.0x10^{-23}$ | rs3117106 | C | 0.433 | $1.2x10^{-20}$ | rs3117106=C and rs206018=G | 0.154 | 0.002 |
| | | | | rs206018 | G | 0.924 | 0.276 | rs3117106=C and rs9501179=A | 0.203 | $5.2x10^{-21}$ |
| | | | | rs9501179 | A | 0.903 | 0.117 | | | |
| 6 | (rs3129881=C) or ((rs375244=A) and (rs3132947=G)) | 3.665 | $2.3x10^{-24}$ | rs3129881 | C | 2.227 | $1.8x10^{-23}$ | rs375244=A and rs3132947=G | 1.238 | $7.9x10^{-5}$ |
| | | | | rs375244 | A | 0.980 | 0.715 | | | |
| | | | | rs3132947 | G | 2.117 | $1.4x10^{-16}$ | | | |
| 6 | ((rs35372932=T) and (rs9269190=C)) or (rs9270493=C) | 0.365 | $9.6x10^{-32}$ | rs35372932 | T | 1.054 | 0.399 | rs35372932=T and rs9269190=C | 0.313 | $8.7x10^{-21}$ |
| | | | | rs9269190 | C | 0.798 | $1.3x10^{-4}$ | | | |
| | | | | rs9270493 | C | 0.488 | $8.0x10^{-11}$ | | | |
| 6 | (rs9268977=T) or ((rs3135395=T) or (rs550513=T)) | 3.448 | $1.2x10^{-29}$ | rs9268977 | T | 1.493 | $2.1x10^{-11}$ | rs9268977=T or rs3135395=T | 3.136 | $3.8x10^{-29}$ |
| | | | | rs3135395 | T | 1.149 | 0.010 | rs9268977=T or rs550513=T | 1.494 | $4.2x10^{-11}$ |
| | | | | rs550513 | T | 0.858 | 0.112 | rs3135395=T or rs550513=T | 1.132 | 0.019 |
| 6 | ((rs9268634=G) or (rs35344500=C)) and (rs9275175=G) | 0.448 | $1.3x10^{-35}$ | rs9268634 | G | 0.750 | $7.5x10^{-8}$ | rs9268634=G and rs9275175=G | 0.467 | $7.6x10^{-31}$ |
| | | | | rs35344500 | C | 0.497 | $1.2x10^{-9}$ | rs35344500=C and rs9275175=G | 0.450 | $4.5x10^{-11}$ |
| | | | | rs9275175 | G | 0.579 | $8.3x10^{-24}$ | | | |
| 6 | ((rs9268831=T) or (rs9269190=T)) or (rs9270652=C) | 3.075 | $7.3x10^{-29}$ | rs9268831 | T | 1.324 | $1.6x10^{-7}$ | rs9268831=T or rs9269190=T | 1.633 | $2.8x10^{-16}$ |
| | | | | rs9269190 | T | 1.252 | $1.3x10^{-4}$ | rs9268831=T or rs9270652=C | 1.321 | $2.0x10^{-5}$ |
| | | | | rs9270652 | C | 1.109 | 0.057 | rs9269190=T or rs9270652=C | 1.925 | $1.2x10^{-18}$ |
| 6 | (rs3132947=G) or ((rs9501179=G) and (rs41546114=C)) | 5.030 | $2.4x10^{-22}$ | rs3132947 | G | 2.117 | $1.4x10^{-16}$ | rs9501179=G and rs41546114=C | 1.110 | 0.086 |
| | | | | rs9501179 | G | 1.107 | 0.117 | | | |
| | | | | rs41546114 | C | 0.957 | 0.685 | | | |
| 6 | (rs9268831=T) or ((rs2395194=G) or (rs387608=A)) | 3.601 | $2.2x10^{-30}$ | rs9268831 | T | 1.324 | $1.6x10^{-7}$ | rs9268831=T or rs2395194=G | 3.107 | $3.5x10^{-29}$ |
| | | | | rs2395194 | G | 1.934 | $8.1x10^{-19}$ | rs9268831=T or rs387608=A | 1.302 | $1.2x10^{-6}$ |
| | | | | rs387608 | A | 1.095 | 0.170 | rs2395194=G or rs387608=A | 2.106 | $2.7x10^{-20}$ |
| 6 | (rs3132947=G) or ((rs9268014=C) and (rs41546114=C)) | 5.012 | $1.0x10^{-22}$ | rs3132947 | G | 2.117 | $1.4x10^{-16}$ | rs9268014=C and rs41546114=C | 1.120 | 0.060 |
| | | | | rs9268014 | C | 1.119 | 0.078 | | | |
| | | | | rs41546114 | C | 0.957 | 0.685 | | | |
| 6 | (rs3132947=G) or ((rs9268055=T) and (rs41546114=C)) | 5.012 | $1.0x10^{-22}$ | rs3132947 | G | 2.117 | $1.4x10^{-16}$ | rs9268055=T and rs41546114=C | 1.134 | 0.036 |
| | | | | rs9268055 | T | 1.134 | 0.049 | | | |
| | | | | rs41546114 | C | 0.957 | 0.685 | | | |
| 6 | ((rs241437=G) or (rs3129299=C)) or (rs9276909=C) | 4.716 | $1.8x10^{-20}$ | rs241437 | G | 1.441 | $6.1x10^{-12}$ | rs241437=G or rs3129299=C | 1.903 | $6.7x10^{-15}$ |
| | | | | rs3129299 | C | 1.225 | $4.9x10^{-4}$ | rs241437=G or rs9276909=C | 3.413 | $4.2x10^{-21}$ |
| | | | | rs9276909 | C | 1.438 | $1.7x10^{-8}$ | rs3129299=C or rs9276909=C | 2.798 | $1.7x10^{-18}$ |
| 6 | ((rs2395194=G) and (rs16870908=G)) or (rs9268831=T) | 2.732 | $5.9x10^{-33}$ | rs2395194 | G | 1.934 | $8.1x10^{-19}$ | rs2395194=G and rs16870908=G | 1.948 | $5.0x10^{-23}$ |
| | | | | rs16870908 | G | 1.514 | 0.001 | | | |
| | | | | rs9268831 | T | 1.324 | $1.6x10^{-7}$ | | | |
| 6 | ((rs9268213=A) and (rs41546114=C)) or (rs3132947=G) | 5.012 | $1.0x10^{-22}$ | rs9268213 | A | 1.129 | 0.058 | rs9268213=A and rs41546114=C | 1.130 | 0.042 |
| | | | | rs41546114 | C | 0.957 | 0.685 | | | |
| | | | | rs3132947 | G | 2.117 | $1.4x10^{-16}$ | | | |
| 6 | ((rs4538748=C) or (rs2064476=G)) and (rs35344500=A) | 2.313 | $3.3x10^{-35}$ | rs4538748 | C | 1.519 | $5.1x10^{-13}$ | rs4538748=C and rs35344500=A | 1.723 | $1.1x10^{-22}$ |
| | | | | rs2064476 | G | 1.521 | $1.8x10^{-14}$ | rs2064476=G and rs35344500=A | 1.623 | $2.0x10^{-19}$ |
| | | | | rs35344500 | A | 2.013 | $1.2x10^{-9}$ | | | |
| 6 | ((rs2244027=A) or (rs2242665=C)) or (rs2071591=G) | 2.873 | $8.7x10^{-26}$ | rs2244027 | A | 1.316 | $4.7x10^{-7}$ | rs2244027=A or rs2242665=C | 1.500 | $7.6x10^{-11}$ |
| | | | | rs2242665 | C | 1.157 | 0.008 | rs2244027=A or rs2071591=G | 1.601 | $2.3x10^{-12}$ |
| | | | | rs2071591 | G | 1.311 | $1.1x10^{-6}$ | rs2242665=C or rs2071591=G | 1.787 | $5.2x10^{-16}$ |
| 6 | ((rs9296088=A) and (rs12207818=C)) and (rs181997=A) | 0.204 | $6.6x10^{-18}$ | rs9296088 | A | 0.752 | $6.1x10^{-8}$ | rs9296088=A and rs12207818=C | 0.482 | $1.6x10^{-14}$ |
| | | | | rs12207818 | C | 0.708 | $1.7x10^{-6}$ | rs9296088=A and rs181997=A | 0.437 | $9.7x10^{-16}$ |
| | | | | rs181997 | A | 0.619 | $3.3x10^{-10}$ | rs12207818=C and rs181997=A | 0.284 | $1.3x10^{-16}$ |
| 6 | ((rs2071287=C) and (rs3094596=C)) or (rs185819=C) | 3.370 | $1.5x10^{-24}$ | rs2071287 | C | 1.482 | $7.1x10^{-12}$ | rs2071287=C and rs3094596=C | 1.497 | $9.4x10^{-13}$ |
| | | | | rs3094596 | C | 2.104 | $2.5x10^{-16}$ | | | |
| | | | | rs185819 | C | 1.302 | $4.1x10^{-6}$ | | | |
| 6 | (rs3132947=T) and ((rs404860=C) or (rs41546114=T)) | 0.253 | $3.2x10^{-21}$ | rs3132947 | T | 0.472 | $1.4x10^{-16}$ | rs3132947=T and rs404860=C | 0.261 | $4.0x10^{-20}$ |
| | | | | rs404860 | C | 1.115 | 0.036 | rs3132947=T and rs41546114=T | 0.333 | 0.055 |
| | | | | rs41546114 | T | 1.045 | 0.685 | | | |

| Chr | 3-SNP haplotype tree | Haplotype OR[a] | Haplotype P[a] | Single SNP | Alt. allele | SNP OR[b] | SNP P[b] | Component haplotype pair | Pair OR[c] | Pair P[c] |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | (rs11754586=T) or ((rs3131932=G) or (rs9262537=G)) | 3.241 | $1.8x10^{-19}$ | rs11754586 | T | 1.090 | 0.125 | rs11754586=T or rs3131932=G | 1.205 | 0.002 |
| | | | | rs3131932 | G | 1.181 | 0.003 | rs11754586=T or rs9262537=G | 1.232 | $1.1x10^{-4}$ |
| | | | | rs9262537 | G | 1.181 | 0.005 | rs3131932=G or rs9262537=G | 1.885 | $4.2x10^{-13}$ |
| 6 | ((rs10947251=A) or (rs41546114=T)) and (rs3132947=T) | 0.203 | $2.8x10^{-22}$ | rs10947251 | A | 0.891 | 0.075 | rs10947251=A and rs3132947=T | 0.199 | $1.5x10^{-21}$ |
| | | | | rs41546114 | T | 1.045 | 0.685 | rs41546114=T and rs3132947=T | 0.333 | 0.055 |
| | | | | rs3132947 | T | 0.472 | $1.4x10^{-16}$ | | | |
| 6 | ((rs3132947=G) or (rs9266632=G)) or (rs185819=C) | 4.321 | $2.0x10^{-23}$ | rs3132947 | G | 2.117 | $1.4x10^{-16}$ | rs3132947=G or rs9266632=G | 2.253 | $1.9x10^{-17}$ |
| | | | | rs9266632 | G | 1.229 | 0.002 | rs3132947=G or rs185819=C | 3.733 | $1.2x10^{-22}$ |
| | | | | rs185819 | C | 1.302 | $4.1x10^{-6}$ | rs9266632=G or rs185819=C | 1.338 | $8.1x10^{-7}$ |
| 6 | (rs805273=A) and ((rs805267=G) or (rs9266774=T)) | 0.264 | $3.1x10^{-22}$ | rs805273 | A | 0.554 | $1.8x10^{-12}$ | rs805273=A and rs805267=G | 0.275 | $2.4x10^{-19}$ |
| | | | | rs805267 | G | 1.097 | 0.372 | rs805273=A and rs9266774=T | 0.214 | $2.1x10^{-4}$ |
| | | | | rs9266774 | T | 0.984 | 0.774 | | | |
| 6 | (rs3828901=A) or ((rs707921=A) and (rs9266774=T)) | 0.279 | $2.5x10^{-21}$ | rs3828901 | A | 0.292 | $2.0x10^{-18}$ | rs707921=A and rs9266774=T | 0.201 | $3.2x10^{-4}$ |
| | | | | rs707921 | A | 0.912 | 0.372 | | | |
| | | | | rs9266774 | T | 0.984 | 0.774 | | | |
| 6 | (rs3828901=G) and ((rs805268=A) or (rs9266774=C)) | 3.586 | $4.4x10^{-21}$ | rs3828901 | G | 3.429 | $2.0x10^{-18}$ | rs3828901=G and rs805268=A | 1.858 | $1.8x10^{-12}$ |
| | | | | rs805268 | A | 1.085 | 0.466 | rs3828901=G and rs9266774=C | 1.293 | $3.6x10^{-6}$ |
| | | | | rs9266774 | C | 1.016 | 0.774 | | | |
| 6 | (rs3828901=A) or ((rs3828919=A) and (rs2523497=T)) | 0.313 | $9.0x10^{-20}$ | rs3828901 | A | 0.292 | $2.0x10^{-18}$ | rs3828919=A and rs2523497=T | 0.535 | 0.029 |
| | | | | rs3828919 | A | 1.028 | 0.667 | | | |
| | | | | rs2523497 | T | 1.064 | 0.234 | | | |
| 6 | (rs707922=G) or ((rs1046089=A) and (rs9266774=C)) | 3.748 | $6.2x10^{-22}$ | rs707922 | G | 1.794 | $2.9x10^{-12}$ | rs1046089=A and rs9266774=C | 1.173 | 0.007 |
| | | | | rs1046089 | A | 1.059 | 0.306 | | | |
| | | | | rs9266774 | C | 1.016 | 0.774 | | | |
| 6 | ((rs3131932=G) or (rs28360042=C)) or (rs885950=C) | 3.356 | $1.1x10^{-20}$ | rs3131932 | G | 1.181 | 0.003 | rs3131932=G or rs28360042=C | 2.826 | $2.2x10^{-19}$ |
| | | | | rs28360042 | C | 1.165 | 0.005 | rs3131932=G or rs885950=C | 1.635 | $6.0x10^{-11}$ |
| | | | | rs885950 | C | 1.165 | 0.004 | rs28360042=C or rs885950=C | 1.255 | 0.001 |
| 6 | ((rs9266774=T) and (rs2255741=A)) or (rs3828901=A) | 0.279 | $2.6x10^{-21}$ | rs9266774 | T | 0.984 | 0.774 | rs9266774=T and rs2255741=A | 0.207 | $4.4x10^{-4}$ |
| | | | | rs2255741 | A | 0.904 | 0.370 | | | |
| | | | | rs3828901 | A | 0.292 | $2.0x10^{-18}$ | | | |
| 6 | (rs3749946=A) and ((rs2517506=T) and (rs7741091=A)) | 0.287 | $1.9x10^{-20}$ | rs3749946 | A | 0.774 | $4.9x10^{-5}$ | rs3749946=A and rs2517506=T | 0.363 | $7.8x10^{-17}$ |
| | | | | rs2517506 | T | 0.921 | 0.157 | rs3749946=A and rs7741091=A | 0.709 | $6.1x10^{-7}$ |
| | | | | rs7741091 | A | 0.852 | 0.004 | rs2517506=T and rs7741091=A | 0.831 | $4.9x10^{-4}$ |
| 6 | (rs3828901=G) and ((rs12110785=T) or (rs2523644=T)) | 3.544 | $6.3x10^{-20}$ | rs3828901 | G | 3.429 | $2.0x10^{-18}$ | rs3828901=G and rs12110785=T | 1.322 | $8.1x10^{-6}$ |
| | | | | rs12110785 | T | 1.315 | $1.3x10^{-5}$ | rs3828901=G and rs2523644=T | 1.876 | $7.0x10^{-15}$ |
| | | | | rs2523644 | T | 1.243 | 0.022 | | | |
| 6 | ((rs2239888=T) or (rs3134769=C)) and (rs3828901=G) | 3.510 | $5.5x10^{-21}$ | rs2239888 | T | 0.894 | 0.119 | rs2239888=T and rs3828901=G | 1.278 | $1.8x10^{-4}$ |
| | | | | rs3134769 | C | 1.091 | 0.272 | rs3134769=C and rs3828901=G | 1.578 | $1.4x10^{-10}$ |
| | | | | rs3828901 | G | 3.429 | $2.0x10^{-18}$ | | | |
| 6 | (rs2517681=T) or ((rs4148248=C) and (rs2735078=A)) | 3.196 | $2.6x10^{-18}$ | rs2517681 | T | 0.956 | 0.386 | rs4148248=C and rs2735078=A | 1.274 | $4.4x10^{-6}$ |
| | | | | rs4148248 | C | 2.975 | $1.1x10^{-16}$ | | | |
| | | | | rs2735078 | A | 1.030 | 0.578 | | | |
| 6 | (rs17195733=G) or ((rs13201129=C) and (rs1737069=T)) | 0.360 | $1.8x10^{-20}$ | rs17195733 | G | 0.333 | $6.0x10^{-17}$ | rs13201129=C and rs1737069=T | 0.480 | $1.2x10^{-4}$ |
| | | | | rs13201129 | C | 0.879 | 0.049 | | | |
| | | | | rs1737069 | T | 0.728 | $5.1x10^{-7}$ | | | |
| 6 | (rs3130785=T) and ((rs9261301=G) or (rs1264570=C)) | 0.340 | $2.7x10^{-17}$ | rs3130785 | T | 0.454 | $8.8x10^{-17}$ | rs3130785=T and rs9261301=G | 0.421 | 0.053 |
| | | | | rs9261301 | G | 1.286 | $7.0x10^{-6}$ | rs3130785=T and rs1264570=C | 0.333 | $3.1x10^{-17}$ |
| | | | | rs1264570 | C | 0.813 | $1.3x10^{-4}$ | | | |
| 6 | (rs2517681=C) and ((rs3130785=T) or (rs2735078=G)) | 0.317 | $1.2x10^{-18}$ | rs2517681 | C | 1.046 | 0.386 | rs2517681=C and rs3130785=T | 0.330 | $2.9x10^{-17}$ |
| | | | | rs3130785 | T | 0.454 | $8.8x10^{-17}$ | rs2517681=C and rs2735078=G | 0.154 | 0.014 |
| | | | | rs2735078 | G | 0.971 | 0.578 | | | |
| 6 | ((rs1345229=A) or (rs1233387=T)) or (rs1003581=G) | 3.070 | $5.4x10^{-18}$ | rs1345229 | A | 1.341 | 0.001 | rs1345229=A or rs1233387=T | 1.159 | 0.006 |
| | | | | rs1233387 | T | 1.102 | 0.066 | rs1345229=A or rs1003581=G | 1.293 | $5.1x10^{-6}$ |
| | | | | rs1003581 | G | 1.226 | $1.8x10^{-4}$ | rs1233387=T or rs1003581=G | 2.653 | $4.6x10^{-16}$ |

| Chr | 3-SNP haplotype tree | Haplotype OR[a] | Haplotype P[a] | Single SNP | Alt. allele | SNP OR[b] | SNP P[b] | Component haplotype pair | Pair OR[c] | Pair P[c] |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | ((rs4713429=G) or (rs12110785=T)) or (rs3131932=G) | 3.093 | $1.0x10^{-19}$ | rs4713429 | G | 1.092 | 0.237 | rs4713429=G or rs12110785=T | 1.388 | $2.0x10^{-6}$ |
| | | | | rs12110785 | T | 1.315 | $1.3x10^{-5}$ | rs4713429=G or rs3131932=G | 1.337 | $1.4x10^{-5}$ |
| | | | | rs3131932 | G | 1.181 | 0.003 | rs12110785=T or rs3131932=G | 1.911 | $1.4x10^{-13}$ |
| 6 | ((rs1003581=G) and (rs16894681=T)) or (rs1233387=T) | 2.541 | $5.5x10^{-17}$ | rs1003581 | G | 1.226 | $1.8x10^{-4}$ | rs1003581=G and rs16894681=T | 1.253 | $2.3x10^{-5}$ |
| | | | | rs16894681 | T | 1.009 | 0.928 | | | |
| | | | | rs1233387 | T | 1.102 | 0.066 | | | |
| 6 | ((rs2281043=T) or (rs7751451=G)) and (rs1635=A) | 0.371 | $2.1x10^{-16}$ | rs2281043 | T | 0.746 | $8.7x10^{-5}$ | rs2281043=T and rs1635=A | 0.358 | $7.5x10^{-15}$ |
| | | | | rs7751451 | G | 0.974 | 0.789 | rs7751451=G and rs1635=A | 0.472 | 0.008 |
| | | | | rs1635 | A | 0.801 | $4.5x10^{-5}$ | | | |
| 6 | ((rs3117192=C) or (rs16894216=T)) and (rs7773193=T) | 2.711 | $2.0x10^{-16}$ | rs3117192 | C | 0.849 | 0.006 | rs3117192=C and rs7773193=T | 1.080 | 0.177 |
| | | | | rs16894216 | T | 0.856 | 0.077 | rs16894216=T and rs7773193=T | 1.310 | $2.2x10^{-4}$ |
| | | | | rs7773193 | T | 2.697 | $1.5x10^{-14}$ | | | |
| 6 | (rs9295704=C) and ((rs2451752=A) and (rs2575174=C)) | 0.365 | $8.9x10^{-15}$ | rs9295704 | C | 0.669 | $1.5x10^{-7}$ | rs9295704=C and rs2451752=A | 0.399 | $3.2x10^{-14}$ |
| | | | | rs2451752 | A | 0.951 | 0.457 | rs9295704=C and rs2575174=C | 0.638 | $1.2x10^{-7}$ |
| | | | | rs2575174 | C | 0.940 | 0.393 | rs2451752=A and rs2575174=C | 0.933 | 0.228 |
| 6 | (rs6939576=G) or ((rs2859365=G) or (rs6930033=A)) | 2.820 | $1.7x10^{-16}$ | rs6939576 | G | 1.091 | 0.137 | rs6939576=G or rs2859365=G | 2.265 | $5.4x10^{-14}$ |
| | | | | rs2859365 | G | 1.165 | 0.004 | rs6939576=G or rs6930033=A | 1.759 | $3.2x10^{-10}$ |
| | | | | rs6930033 | A | 1.143 | 0.015 | rs2859365=G or rs6930033=A | 1.435 | $1.0x10^{-6}$ |
| 6 | (rs7773193=C) or ((rs17280818=T) and (rs2394100=T)) | 0.382 | $6.3x10^{-16}$ | rs7773193 | C | 0.371 | $1.5x10^{-14}$ | rs17280818=T and rs2394100=T | 0.481 | 0.013 |
| | | | | rs17280818 | T | 0.988 | 0.904 | | | |
| | | | | rs2394100 | T | 1.034 | 0.602 | | | |
| 7 | ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G) | 0.066 | $3.9x10^{-9}$ | rs12671658 | T | 1.023 | 0.682 | rs12671658=T and rs11768586=G | 0.104 | $1.4x10^{-6}$ |
| | | | | rs12702656 | A | 1.039 | 0.585 | rs12702656=A and rs11768586=G | 0.000 | 0.953 |
| | | | | rs11768586 | G | 0.865 | 0.010 | | | |
| 9 | (rs4979484=C) or ((rs13300483=T) and (rs7028891=G)) | 1.672 | $3.0x10^{-21}$ | rs4979484 | C | 1.365 | 0.005 | rs13300483=T and rs7028891=G | 1.637 | $1.2x10^{-19}$ |
| | | | | rs13300483 | T | 1.584 | $1.4x10^{-17}$ | | | |
| | | | | rs7028891 | G | 1.574 | $2.8x10^{-17}$ | | | |
| 9 | (rs7028891=G) and ((rs4979462=T) or (rs10739402=T)) | 1.696 | $2.7x10^{-22}$ | rs7028891 | G | 1.574 | $2.8x10^{-17}$ | rs7028891=G and rs4979462=T | 1.634 | $1.5x10^{-19}$ |
| | | | | rs4979462 | T | 1.599 | $2.6x10^{-18}$ | rs7028891=G and rs10739402=T | 1.387 | $1.5x10^{-5}$ |
| | | | | rs10739402 | T | 1.193 | 0.003 | | | |
| 9 | (rs7028891=A) and ((rs2418376=A) or (rs10759773=G)) | 0.604 | $9.3x10^{-21}$ | rs7028891 | A | 0.635 | $2.8x10^{-17}$ | rs7028891=A and rs2418376=A | 0.628 | $7.5x10^{-18}$ |
| | | | | rs2418376 | A | 1.003 | 0.972 | rs7028891=A and rs10759773=G | 0.735 | $4.5x10^{-7}$ |
| | | | | rs10759773 | G | 0.921 | 0.118 | | | |
| 9 | (rs10817678=G) or ((rs1407306=T) and (rs7048742=A)) | 0.613 | $2.1x10^{-19}$ | rs10817678 | G | 0.652 | $8.3x10^{-14}$ | rs1407306=T and rs7048742=A | 0.710 | $3.6x10^{-4}$ |
| | | | | rs1407306 | T | 0.785 | 0.003 | | | |
| | | | | rs7048742 | A | 1.166 | 0.014 | | | |
| 9 | (rs4979462=T) or ((rs10739402=T) and (rs10817564=C)) | 1.687 | $1.5x10^{-21}$ | rs4979462 | T | 1.599 | $2.6x10^{-18}$ | rs10739402=T and rs10817564=C | 1.258 | 0.004 |
| | | | | rs10739402 | T | 1.193 | 0.003 | | | |
| | | | | rs10817564 | C | 1.034 | 0.526 | | | |

[a] 3-SNP haplotype OR and p-value in the combined sample (N=2886).
[b] Single SNP ORs and p-values, assuming an additive genetic effect model for the specified alternative allele.
[c] 2-SNP haplotype ORs and p-values, assuming an additive genetic effect model for the specified haplotype pattern.
Abbreviations: Alt, alternative.

**Table S3**: Pre-phasing missingness rates for replicated 3-SNP haplotypes

| | Entire sample (N=2886) | All controls (N=1505) | All cases (N=1381) |
|---|---|---|---|
| Median, proportion missing | 0.0104 | 0.0113 | 0.0109 |
| IQR, proportion missing | 0.0069 | 0.0086 | 0.0080 |
| Max, proportion missing | 0.0596 | 0.0585 | 0.0608 |

Per-sample tree missingness was defined as having at least 1 SNP genotype missing among the 3 SNPs constituting the haplotype for a given sample.

**Table S4**: Comparison of distribution frequencies for replicated 3-SNP haplotypes in unphased and phased control groups in the Japan-PBC GWAS data and the phased 1000 Genomes Japanese reference panel

| | | PBC UNPHASED (controls, N=1505) | | PBC PHASED (controls, N=1505) | | | | | | | | | 1000 GENOMES, JPT (N=104) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr | 3-SNP haplotype tree (tree SNP order: rs1, rs2, rs3) | # Homozygous carriers | # Potential carriers of at least 1 copy | rs1 AF | rs2 AF | rs3 AF | h0 | % (h0/N) | h1 | % (h1/N) | h2 | % (h2/N) | rs1 AF | rs2 AF | rs3 AF | h0 | % (h0/N) | h1 | % (h1/N) | h2 | % (h2/N) |
| 6 | (rs3117106=C) and ((rs206018=G) or (rs9501179=A)) | 11 | 277 | C=0.072 | G=0.152 | A=0.199 | 1274 | 84.7% | 218 | 14.5% | 13 | 0.9% | C=0.173 | G=0.159 | A=0.192 | 86 | 82.7% | 17 | 16.3% | 1 | 1.0% |
| 6 | (rs3129881=C) or ((rs375244=A) and (rs3132947=G)) | 1112 | 1494 | C=0.903 | A=0.633 | G=0.928 | 14 | 0.9% | 282 | 18.7% | 1209 | 80.3% | C=0.784 | A=0.615 | G=0.875 | 3 | 2.9% | 21 | 20.2% | 80 | 76.9% |
| 6 | ((rs35372932=T) and (rs9269190=C)) or (rs9270493=C) | 36 | 755 | T=0.243 | C=0.704 | C=0.047 | 987 | 65.6% | 461 | 30.6% | 57 | 3.8% | T=0.231 | C=0.697 | C=0.072 | 75 | 72.1% | 27 | 26.0% | 2 | 1.9% |
| 6 | (rs9268977=T) or ((rs3135395=T) or (rs550513=T)) | 841 | 1469 | T=0.760 | T=0.407 | T=0.074 | 36 | 2.4% | 338 | 22.5% | 1131 | 75.1% | T=0.635 | T=0.389 | T=0.072 | 2 | 1.9% | 26 | 25.0% | 76 | 73.1% |
| 6 | ((rs9268634=G) or (rs35344500=C)) and (rs9275175=G) | 143 | 966 | G=0.510 | C=0.043 | G=0.373 | 678 | 45.0% | 661 | 43.9% | 166 | 11.0% | G=0.615 | C=0.067 | G=0.476 | 50 | 48.1% | 44 | 42.3% | 10 | 9.6% |
| 6 | ((rs9268831=T) or (rs9269190=T)) or (rs9270652=C) | 857 | 1466 | T=0.579 | T=0.296 | C=0.653 | 39 | 2.6% | 367 | 24.4% | 1099 | 73.0% | T=0.500 | T=0.303 | C=0.630 | NA | NA | 17 | 16.3% | 87 | 83.7% |
| 6 | (rs3132947=G) or ((rs9501179=G) and (rs41546114=C)) | 1247 | 1493 | G=0.928 | G=0.801 | C=0.936 | 11 | 0.7% | 212 | 14.1% | 1282 | 85.2% | G=0.875 | G=0.808 | C=0.938 | 1 | 1.0% | 15 | 14.4% | 88 | 84.6% |
| 6 | (rs9268831=T) or ((rs2395194=G) or (rs387608=A)) | 1039 | 1468 | T=0.579 | G=0.883 | A=0.199 | 35 | 2.3% | 336 | 22.3% | 1134 | 75.3% | T=0.500 | G=0.760 | A=0.183 | 2 | 1.9% | 26 | 25.0% | 76 | 73.1% |
| 6 | (rs3132947=G) or ((rs9268014=C) and (rs41546114=C)) | 1240 | 1493 | G=0.928 | C=0.790 | C=0.936 | 12 | 0.8% | 215 | 14.3% | 1278 | 84.9% | G=0.875 | C=0.798 | C=0.938 | 1 | 1.0% | 15 | 14.4% | 88 | 84.6% |
| 6 | (rs3132947=G) or ((rs9268055=T) and (rs41546114=C)) | 1242 | 1493 | G=0.928 | T=0.793 | C=0.936 | 12 | 0.8% | 215 | 14.3% | 1278 | 84.9% | G=0.875 | T=0.798 | C=0.938 | 1 | 1.0% | 15 | 14.4% | 88 | 84.6% |
| 6 | ((rs241437=G) or (rs3129299=C)) or (rs9276909=C) | 1218 | 1489 | G=0.593 | C=0.727 | C=0.808 | 16 | 1.1% | 193 | 12.8% | 1296 | 86.1% | G=0.500 | C=0.692 | C=0.798 | 1 | 1.0% | 13 | 12.5% | 90 | 86.5% |
| 6 | ((rs2395194=G) and (rs16870908=G)) or (rs9268831=T) | 897 | 1455 | G=0.883 | G=0.958 | T=0.579 | 64 | 4.3% | 475 | 31.6% | 966 | 64.2% | G=0.760 | G=0.947 | T=0.500 | 3 | 2.9% | 39 | 37.5% | 62 | 59.6% |
| 6 | ((rs9268213=A) and (rs41546114=C)) or (rs3132947=G) | 1240 | 1493 | A=0.791 | C=0.936 | G=0.928 | 12 | 0.8% | 215 | 14.3% | 1278 | 84.9% | A=0.798 | C=0.938 | G=0.875 | 1 | 1.0% | 15 | 14.4% | 88 | 84.6% |
| 6 | ((rs4538748=C) or (rs2064476=G)) and (rs35344500=A) | 670 | 1412 | C=0.736 | G=0.674 | A=0.957 | 137 | 9.1% | 604 | 40.1% | 764 | 50.8% | C=0.601 | G=0.558 | A=0.933 | 4 | 3.8% | 48 | 46.2% | 52 | 50.0% |
| 6 | ((rs2244027=A) or (rs2242665=C)) or (rs2071591=G) | 875 | 1478 | A=0.632 | C=0.386 | G=0.668 | 26 | 1.7% | 373 | 24.8% | 1106 | 73.5% | A=0.587 | C=0.351 | G=0.615 | 3 | 2.9% | 21 | 20.2% | 80 | 76.9% |
| 6 | ((rs9296088=A) and (rs12207818=C)) and (rs181997=A) | 9 | 209 | A=0.539 | C=0.138 | A=0.110 | 1326 | 88.1% | 169 | 11.2% | 10 | 0.7% | A=0.630 | C=0.202 | A=0.159 | 88 | 84.6% | 16 | 15.4% | NA | NA |
| 6 | ((rs2071287=C) and (rs3094596=C)) or (rs185819=C) | 1001 | 1477 | C=0.730 | C=0.928 | C=0.700 | 28 | 1.9% | 284 | 18.9% | 1193 | 79.3% | C=0.615 | C=0.846 | C=0.688 | 2 | 1.9% | 16 | 15.4% | 86 | 82.7% |
| 6 | (rs3132947=T) and ((rs404860=C) or (rs41546114=T)) | 14 | 295 | T=0.072 | C=0.542 | T=0.064 | 1265 | 84.1% | 226 | 15.0% | 14 | 0.9% | T=0.125 | C=0.481 | T=0.062 | 87 | 83.7% | 16 | 15.4% | 1 | 1.0% |
| 6 | (rs11754586=T) or ((rs3131932=G) or (rs9262537=G)) | 811 | 1493 | T=0.346 | G=0.703 | G=0.298 | 12 | 0.8% | 247 | 16.4% | 1246 | 82.8% | T=0.322 | G=0.702 | G=0.221 | 1 | 1.0% | 16 | 15.4% | 87 | 83.7% |
| 6 | ((rs10947251=A) or (rs41546114=T)) and (rs3132947=T) | 12 | 255 | A=0.197 | T=0.064 | T=0.072 | 1281 | 85.1% | 212 | 14.1% | 12 | 0.8% | A=0.192 | T=0.062 | T=0.125 | 88 | 84.6% | 15 | 14.4% | 1 | 1.0% |
| 6 | ((rs3132947=G) or (rs9266632=G)) or (rs185819=C) | 1210 | 1490 | G=0.928 | G=0.214 | C=0.700 | 14 | 0.9% | 238 | 15.8% | 1253 | 83.3% | G=0.875 | G=0.139 | C=0.688 | 2 | 1.9% | 15 | 14.4% | 87 | 83.7% |
| 6 | (rs805273=A) and ((rs805267=G) or (rs9266774=T)) | 10 | 403 | A=0.088 | G=0.936 | T=0.304 | 1245 | 82.7% | 246 | 16.3% | 14 | 0.9% | A=0.173 | G=0.904 | T=0.284 | 89 | 85.6% | 13 | 12.5% | 2 | 1.9% |
| 6 | (rs3828901=A) or ((rs707921=A) and (rs9266774=T)) | 9 | 303 | A=0.025 | A=0.064 | T=0.304 | 1245 | 82.7% | 247 | 16.4% | 13 | 0.9% | A=0.082 | A=0.096 | T=0.284 | 88 | 84.6% | 14 | 13.5% | 2 | 1.9% |
| 6 | (rs3828901=G) and ((rs805268=A) or (rs9266774=C)) | 1209 | 1494 | G=0.975 | A=0.946 | C=0.696 | 12 | 0.8% | 245 | 16.3% | 1248 | 82.9% | G=0.918 | A=0.933 | C=0.716 | 1 | 1.0% | 15 | 14.4% | 88 | 84.6% |

| | | PBC UNPHASED (controls, N=1505) | | PBC PHASED (controls, N=1505) | | | | | | | | | 1000 GENOMES, JPT (N=104) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr | 3-SNP haplotype tree (tree SNP order: rs1, rs2, rs3) | # Homozygous carriers | # Potential carriers of at least 1 copy | rs1 AF | rs2 AF | rs3 AF | h0 | % (h0/N) | h1 | % (h1/N) | h2 | % (h2/N) | rs1 AF | rs2 AF | rs3 AF | h0 | % (h0/N) | h1 | % (h1/N) | h2 | % (h2/N) |
| 6 | (rs3828901=A) or ((rs3828919=A) and (rs2523497=T)) | 10 | 508 | A=0.025 | A=0.205 | T=0.483 | 1237 | 82.2% | 258 | 17.1% | 10 | 0.7% | A=0.082 | A=0.197 | T=0.452 | 86 | 82.7% | 17 | 16.3% | 1 | 1.0% |
| 6 | (rs707922=G) or ((rs1046089=A) and (rs9266774=C)) | 1128 | 1493 | G=0.912 | A=0.352 | C=0.696 | 14 | 0.9% | 244 | 16.2% | 1247 | 82.9% | G=0.827 | A=0.370 | C=0.716 | 2 | 1.9% | 13 | 12.5% | 89 | 85.6% |
| 6 | ((rs3131932=G) or (rs28360042=C)) or (rs885950=C) | 1067 | 1491 | G=0.703 | C=0.620 | C=0.395 | 14 | 0.9% | 252 | 16.7% | 1239 | 82.3% | G=0.702 | C=0.577 | C=0.317 | 1 | 1.0% | 17 | 16.3% | 86 | 82.7% |
| 6 | ((rs9266774=T) and (rs2255741=A)) or (rs3828901=A) | 9 | 295 | T=0.304 | A=0.054 | A=0.025 | 1245 | 82.7% | 248 | 16.5% | 12 | 0.8% | T=0.284 | A=0.067 | A=0.082 | 88 | 84.6% | 15 | 14.4% | 1 | 1.0% |
| 6 | (rs3749946=A) and ((rs2517506=T) and (rs7741091=A)) | 10 | 520 | A=0.198 | T=0.707 | A=0.638 | 1249 | 83.0% | 246 | 16.3% | 10 | 0.7% | A=0.250 | T=0.668 | A=0.678 | 87 | 83.7% | 16 | 15.4% | 1 | 1.0% |
| 6 | (rs3828901=G) and ((rs12110785=T) or (rs2523644=T)) | 1218 | 1494 | G=0.975 | T=0.789 | T=0.927 | 10 | 0.7% | 235 | 15.6% | 1260 | 83.7% | G=0.918 | T=0.736 | T=0.909 | 1 | 1.0% | 15 | 14.4% | 88 | 84.6% |
| 6 | ((rs2239888=T) or (rs3134769=C)) and (rs3828901=G) | 1200 | 1494 | T=0.834 | C=0.887 | G=0.975 | 11 | 0.7% | 250 | 16.6% | 1244 | 82.7% | T=0.856 | C=0.856 | G=0.918 | 1 | 1.0% | 17 | 16.3% | 86 | 82.7% |
| 6 | (rs2517681=T) or ((rs4148248=C) and (rs2735078=A)) | 797 | 1489 | T=0.493 | C=0.971 | A=0.624 | 12 | 0.8% | 233 | 15.5% | 1260 | 83.7% | T=0.505 | C=0.913 | A=0.606 | 1 | 1.0% | 17 | 16.3% | 86 | 82.7% |
| 6 | (rs17195733=G) or ((rs13201129=C) and (rs1737069=T)) | 13 | 396 | G=0.029 | C=0.202 | T=0.192 | 1181 | 78.5% | 305 | 20.3% | 19 | 1.3% | G=0.091 | C=0.173 | T=0.255 | 83 | 79.8% | 20 | 19.2% | 1 | 1.0% |
| 6 | (rs3130785=T) and ((rs9261301=G) or (rs1264570=C)) | 12 | 323 | T=0.065 | G=0.377 | C=0.419 | 1253 | 83.3% | 240 | 15.9% | 12 | 0.8% | T=0.149 | G=0.279 | C=0.438 | 85 | 81.7% | 18 | 17.3% | 1 | 1.0% |
| 6 | (rs2517681=C) and ((rs3130785=T) or (rs2735078=G)) | 12 | 706 | C=0.507 | T=0.065 | G=0.376 | 1251 | 83.1% | 242 | 16.1% | 12 | 0.8% | C=0.495 | T=0.149 | G=0.394 | 85 | 81.7% | 18 | 17.3% | 1 | 1.0% |
| 6 | ((rs1345229=A) or (rs1233387=T)) or (rs1003581=G) | 964 | 1489 | A=0.123 | T=0.561 | G=0.680 | 13 | 0.9% | 237 | 15.7% | 1255 | 83.4% | A=0.115 | T=0.519 | G=0.639 | 1 | 1.0% | 16 | 15.4% | 87 | 83.7% |
| 6 | ((rs4713429=G) or (rs12110785=T)) or (rs3131932=G) | 1058 | 1487 | G=0.150 | T=0.789 | G=0.703 | 17 | 1.1% | 255 | 16.9% | 1233 | 81.9% | G=0.120 | T=0.736 | G=0.702 | 1 | 1.0% | 19 | 18.3% | 84 | 80.8% |
| 6 | ((rs1003581=G) and (rs16894681=T)) or (rs1233387=T) | 901 | 1484 | G=0.680 | T=0.927 | T=0.561 | 17 | 1.1% | 281 | 18.7% | 1207 | 80.2% | G=0.639 | T=0.918 | T=0.519 | 1 | 1.0% | 22 | 21.2% | 81 | 77.9% |
| 6 | ((rs2281043=T) or (rs7751451=G)) and (rs1635=A) | 12 | 386 | T=0.129 | G=0.074 | A=0.346 | 1243 | 82.6% | 246 | 16.3% | 16 | 1.1% | T=0.168 | G=0.087 | A=0.365 | 91 | 87.5% | 12 | 11.5% | 1 | 1.0% |
| 6 | ((rs3117192=C) or (rs16894216=T)) and (rs7773193=T) | 1177 | 1488 | C=0.719 | T=0.895 | T=0.969 | 15 | 1.0% | 246 | 16.3% | 1244 | 82.7% | C=0.716 | T=0.889 | T=0.918 | 1 | 1.0% | 18 | 17.3% | 85 | 81.7% |
| 6 | (rs9295704=C) and ((rs2451752=A) and (rs2575174=C)) | 10 | 385 | C=0.114 | A=0.806 | C=0.849 | 1274 | 84.7% | 221 | 14.7% | 10 | 0.7% | C=0.173 | A=0.798 | C=0.837 | 84 | 80.8% | 19 | 18.3% | 1 | 1.0% |
| 6 | (rs6939576=G) or ((rs2859365=G) or (rs6930033=A)) | 1141 | 1488 | G=0.735 | G=0.544 | A=0.640 | 17 | 1.1% | 230 | 15.3% | 1258 | 83.6% | G=0.707 | G=0.524 | A=0.601 | 1 | 1.0% | 17 | 16.3% | 86 | 82.7% |
| 6 | (rs7773193=C) or ((rs17280818=T) and (rs2394100=T)) | 14 | 300 | C=0.031 | T=0.075 | T=0.209 | 1243 | 82.6% | 244 | 16.2% | 18 | 1.2% | C=0.082 | T=0.087 | T=0.212 | 88 | 84.6% | 15 | 14.4% | 1 | 1.0% |
| 7 | ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G) | 0 | 554 | T=0.337 | A=0.175 | G=0.333 | 1426 | 94.8% | 79 | 5.2% | NA | NA | T=0.303 | A=0.173 | G=0.370 | 104 | 100.0% | NA | NA | NA | NA |
| 9 | (rs4979484=C) or ((rs13300483=T) and (rs7028891=G)) | 242 | 1008 | C=0.071 | T=0.543 | G=0.568 | 501 | 33.3% | 736 | 48.9% | 268 | 17.8% | C=0.058 | T=0.438 | G=0.452 | 33 | 31.7% | 52 | 50.0% | 19 | 18.3% |
| 9 | (rs7028891=G) and ((rs4979462=T) or (rs10739402=T)) | 221 | 982 | G=0.568 | T=0.580 | T=0.292 | 557 | 37.0% | 717 | 47.6% | 231 | 15.3% | G=0.452 | T=0.495 | T=0.308 | 38 | 36.5% | 46 | 44.2% | 20 | 19.2% |
| 9 | (rs7028891=A) and ((rs2418376=A) or (rs10759773=G)) | 373 | 1179 | A=0.432 | A=0.863 | G=0.458 | 357 | 23.7% | 745 | 49.5% | 403 | 26.8% | A=0.548 | A=0.851 | G=0.462 | 29 | 27.9% | 48 | 46.2% | 27 | 26.0% |
| 9 | (rs10817678=G) or ((rs1407306=T) and (rs7048742=A)) | 224 | 1137 | G=0.292 | T=0.108 | A=0.774 | 406 | 27.0% | 736 | 48.9% | 363 | 24.1% | G=0.385 | T=0.101 | A=0.755 | 34 | 32.7% | 47 | 45.2% | 23 | 22.1% |
| 9 | (rs4979462=T) or ((rs10739402=T) and (rs10817564=C)) | 334 | 1205 | T=0.580 | T=0.292 | C=0.565 | 356 | 23.7% | 750 | 49.8% | 399 | 26.5% | T=0.495 | T=0.308 | C=0.596 | 19 | 18.3% | 52 | 50.0% | 33 | 31.7% |

Abbreviations: rs1 AF, allele frequency (AF) for 1st SNP listed in 3-SNP haplotype; rs2 AF, AF for 2nd SNP in haplotype; rs3 AF, AF for 3rd SNP in haplotype; h0, number (#) carriers of 0 haplotype tree copies; h1, # carriers of 1 haplotype tree copy; h2, # carriers with 2 haplotype tree copies.

**Table S5**: Application of the benchmark method to SNPs in replicated 3-SNP haplotypes identified with the proposed method

| | | | | Benchmark method (applied to logic tree SNPs only)[b] | | |
|---|---|---|---|---|---|---|
| Chr | 3-SNP haplotype tree | OR[a] | P[a] | Global score statistic | Degrees of freedom | P |
| 6 | (rs3117106=C) and ((rs206018=G) or (rs9501179=A)) | 0.196 | $3.0x10^{-23}$ | 127.389 | 6 | $4.6x10^{-25}$ |
| 6 | (rs3129881=C) or ((rs375244=A) and (rs3132947=G)) | 3.665 | $2.3x10^{-24}$ | 160.689 | 6 | $4.2x10^{-32}$ |
| 6 | ((rs35372932=T) and (rs9269190=C)) or (rs9270493=C) | 0.365 | $9.6x10^{-32}$ | 158.755 | 5 | $1.8x10^{-32}$ |
| 6 | (rs9268977=T) or ((rs3135395=T) or (rs550513=T)) | 3.448 | $1.2x10^{-29}$ | 119.847 | 7 | $8.2x10^{-23}$ |
| 6 | ((rs9268634=G) or (rs35344500=C)) and (rs9275175=G) | 0.448 | $1.3x10^{-35}$ | 163.123 | 7 | $7.1x10^{-32}$ |
| 6 | ((rs9268831=T) or (rs9269190=T)) or (rs9270652=C) | 3.075 | $7.3x10^{-29}$ | 140.289 | 6 | $8.7x10^{-28}$ |
| 6 | (rs3132947=G) or ((rs9501179=G) and (rs41546114=C)) | 5.030 | $2.4x10^{-22}$ | 113.509 | 6 | $3.8x10^{-22}$ |
| 6 | (rs9268831=T) or ((rs2395194=G) or (rs387608=A)) | 3.601 | $2.2x10^{-30}$ | 124.925 | 7 | $7.2x10^{-24}$ |
| 6 | (rs3132947=G) or ((rs9268014=C) and (rs41546114=C)) | 5.012 | $1.0x10^{-22}$ | 113.035 | 6 | $4.7x10^{-22}$ |
| 6 | (rs3132947=G) or ((rs9268055=T) and (rs41546114=C)) | 5.012 | $1.0x10^{-22}$ | 112.163 | 6 | $7.2x10^{-22}$ |
| 6 | ((rs241437=G) or (rs3129299=C)) or (rs9276909=C) | 4.716 | $1.8x10^{-20}$ | 127.570 | 7 | $2.0x10^{-24}$ |
| 6 | ((rs2395194=G) and (rs16870908=G)) or (rs9268831=T) | 2.732 | $5.9x10^{-33}$ | 146.693 | 6 | $3.9x10^{-29}$ |
| 6 | ((rs9268213=A) and (rs41546114=C)) or (rs3132947=G) | 5.012 | $1.0x10^{-22}$ | 111.456 | 6 | $1.0x10^{-21}$ |
| 6 | ((rs4538748=C) or (rs2064476=G)) and (rs35344500=A) | 2.313 | $3.3x10^{-35}$ | 168.448 | 6 | $9.6x10^{-34}$ |
| 6 | ((rs2244027=A) or (rs2242665=C)) or (rs2071591=G) | 2.873 | $8.7x10^{-26}$ | 76.256 | 7 | $8.0x10^{-14}$ |
| 6 | ((rs9296088=A) and (rs12207818=C)) and (rs181997=A) | 0.204 | $6.6x10^{-18}$ | 97.965 | 7 | $2.8x10^{-18}$ |
| 6 | ((rs2071287=C) and (rs3094596=C)) or (rs185819=C) | 3.370 | $1.5x10^{-24}$ | 119.583 | 7 | $9.4x10^{-23}$ |
| 6 | (rs3132947=T) and ((rs404860=C) or (rs41546114=T)) | 0.253 | $3.2x10^{-21}$ | 104.903 | 6 | $2.4x10^{-20}$ |
| 6 | (rs11754586=T) or ((rs3131932=G) or (rs9262537=G)) | 3.241 | $1.8x10^{-19}$ | 76.098 | 6 | $2.3x10^{-14}$ |
| 6 | ((rs10947251=A) or (rs41546114=T)) and (rs3132947=T) | 0.203 | $2.8x10^{-22}$ | 112.040 | 6 | $7.6x10^{-22}$ |
| 6 | ((rs3132947=G) or (rs9266632=G)) or (rs185819=C) | 4.321 | $2.0x10^{-23}$ | 100.854 | 7 | $7.2x10^{-19}$ |
| 6 | (rs805273=A) and ((rs805267=G) or (rs9266774=T)) | 0.264 | $3.1x10^{-22}$ | 101.022 | 5 | $3.2x10^{-20}$ |
| 6 | (rs3828901=A) or ((rs707921=A) and (rs9266774=T)) | 0.279 | $2.5x10^{-21}$ | 94.802 | 5 | $6.6x10^{-19}$ |
| 6 | (rs3828901=G) and ((rs805268=A) or (rs9266774=C)) | 3.586 | $4.4x10^{-21}$ | 92.605 | 5 | $1.9x10^{-18}$ |
| 6 | (rs3828901=A) or ((rs3828919=A) and (rs2523497=T)) | 0.313 | $9.0x10^{-20}$ | 92.502 | 5 | $2.0x10^{-18}$ |
| 6 | (rs707922=G) or ((rs1046089=A) and (rs9266774=C)) | 3.748 | $6.2x10^{-22}$ | 95.094 | 7 | $1.1x10^{-17}$ |
| 6 | ((rs3131932=G) or (rs28360042=C)) or (rs885950=C) | 3.356 | $1.1x10^{-20}$ | 85.291 | 7 | $1.1x10^{-15}$ |
| 6 | ((rs9266774=T) and (rs2255741=A)) or (rs3828901=A) | 0.279 | $2.6x10^{-21}$ | 93.411 | 5 | $1.3x10^{-18}$ |
| 6 | (rs3749946=A) and ((rs2517506=T) and (rs7741091=A)) | 0.287 | $1.9x10^{-20}$ | 76.474 | 7 | $7.2x10^{-14}$ |
| 6 | (rs3828901=G) and ((rs12110785=T) or (rs2523644=T)) | 3.544 | $6.3x10^{-20}$ | 97.538 | 4 | $3.3x10^{-20}$ |
| 6 | ((rs2239888=T) or (rs3134769=C)) and (rs3828901=G) | 3.510 | $5.5x10^{-21}$ | 95.089 | 5 | $5.7x10^{-19}$ |
| 6 | (rs2517681=T) or ((rs4148248=C) and (rs2735078=A)) | 3.196 | $2.6x10^{-18}$ | 92.509 | 4 | $3.9x10^{-19}$ |
| 6 | (rs17195733=G) or ((rs13201129=C) and (rs1737069=T)) | 0.360 | $1.8x10^{-20}$ | 92.753 | 5 | $1.8x10^{-18}$ |
| 6 | (rs3130785=T) and ((rs9261301=G) or (rs1264570=C)) | 0.340 | $2.7x10^{-17}$ | 92.694 | 6 | $8.3x10^{-18}$ |
| 6 | (rs2517681=C) and ((rs3130785=T) or (rs2735078=G)) | 0.317 | $1.2x10^{-18}$ | 98.277 | 5 | $1.2x10^{-19}$ |
| 6 | ((rs1345229=A) or (rs1233387=T)) or (rs1003581=G) | 3.070 | $5.4x10^{-18}$ | 73.057 | 7 | $3.6x10^{-13}$ |
| 6 | ((rs4713429=G) or (rs12110785=T)) or (rs3131932=G) | 3.093 | $1.0x10^{-19}$ | 75.618 | 7 | $1.1x10^{-13}$ |
| 6 | ((rs1003581=G) and (rs16894681=T)) or (rs1233387=T) | 2.541 | $5.5x10^{-17}$ | 64.413 | 6 | $5.7x10^{-12}$ |
| 6 | ((rs2281043=T) or (rs7751451=G)) and (rs1635=A) | 0.371 | $2.1x10^{-16}$ | 55.683 | 7 | $1.1x10^{-9}$ |
| 6 | ((rs3117192=C) or (rs16894216=T)) and (rs7773193=T) | 2.711 | $2.0x10^{-16}$ | 71.075 | 5 | $6.1x10^{-14}$ |
| 6 | (rs9295704=C) and ((rs2451752=A) and (rs2575174=C)) | 0.365 | $8.9x10^{-15}$ | 53.502 | 7 | $3.0x10^{-9}$ |
| 6 | (rs6939576=G) or ((rs2859365=G) or (rs6930033=A)) | 2.820 | $1.7x10^{-16}$ | 78.111 | 7 | $3.3x10^{-14}$ |
| 6 | (rs7773193=C) or ((rs17280818=T) and (rs2394100=T)) | 0.382 | $6.3x10^{-16}$ | 67.438 | 5 | $3.5x10^{-13}$ |
| 7 | ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G) | 0.066 | $3.9x10^{-9}$ | 31.899 | 6 | $1.7x10^{-5}$ |
| 9 | (rs4979484=C) or ((rs13300483=T) and (rs7028891=G)) | 1.672 | $3.0x10^{-21}$ | 90.263 | 6 | $2.7x10^{-17}$ |

| | | | | Benchmark method (applied to logic tree SNPs only)[b] | | |
|---|---|---|---|---|---|---|
| Chr | 3-SNP haplotype tree | OR[a] | P[a] | Global score statistic | Degrees of freedom | P |
| 9 | (rs7028891=G) and ((rs4979462=T) or (rs10739402=T)) | 1.696 | $2.7 \times 10^{-22}$ | 100.579 | 7 | $8.2 \times 10^{-19}$ |
| 9 | (rs7028891=A) and ((rs2418376=A) or (rs10759773=G)) | 0.604 | $9.3 \times 10^{-21}$ | 85.233 | 7 | $1.2 \times 10^{-15}$ |
| 9 | (rs10817678=G) or ((rs1407306=T) and (rs7048742=A)) | 0.613 | $2.1 \times 10^{-19}$ | 91.584 | 6 | $1.4 \times 10^{-17}$ |
| 9 | (rs4979462=T) or ((rs10739402=T) and (rs10817564=C)) | 1.687 | $1.5 \times 10^{-21}$ | 91.479 | 7 | $6.1 \times 10^{-17}$ |

[a] 3-SNP haplotype OR and p-value in the combined sample (N=2886), under the proposed logic regression method.

[b] Benchmark haplotype association testing method implemented in R 'haplo.stats', with 3 SNPs specified in the haplotype tree (minimum haplotype frequency=20).

**Table S6:** GTEx Analysis v7 eQTL enrichment analysis results for blood- and liver-related cell/tissue types

| Cell/Tissue Type | Significant eQTLs, PBC SNPs (N=5207) | Significant eQTLs, Comparison (N=76136) | OR | Fisher P |
|---|---|---|---|---|
| EBV-transformed lymphocytes | 64 | 618 | 1.521 | $2.7x10^{-3}$ |
| Whole blood | 179 | 2415 | 1.087 | 0.289 |
| Liver | 51 | 672 | 1.111 | 0.446 |

Abbreviations: EBV, Epstein-Barr virus; sig, significant; eQTL, expression quantitative trait loci.
Enrichment defined as any result meeting the Bonferroni-corrected p-value threshold (.05/3=0.017)

**Table S7:** H3K4me1 histone mark enrichment analysis results for all 29 blood- and liver-related epigenomes

| EID | EID Grouping | Epigenome Name | H3K4me1 Peak Overlap, PBC SNPs (N=106) | H3K4me1 Peak Overlap, Comparison (N=16036) | OR | Fisher P |
|---|---|---|---|---|---|---|
| E032 | HSC & B-cell | Primary B cells (from PB) | 35 | 1680 | 4.212 | $4.1x10^{-10}$ |
| E034 | Blood & T-cell | Primary T cells from primary blood (from PB) | 34 | 1665 | 4.076 | $1.3x10^{-9}$ |
| E124 | ENCODE2012 | Monocytes-CD14+ RO01746 | 34 | 1732 | 3.899 | $3.7x10^{-9}$ |
| E118 | ENCODE2012 | HepG2 hepatocellular carcinoma | 31 | 1538 | 3.895 | $1.2x10^{-8}$ |
| E029 | HSC & B-cell | Primary monocytes (from PB) | 24 | 1412 | 3.031 | $1.5x10^{-5}$ |
| E044 | Blood & T-cell | Primary T regulatory cells (from PB) | 25 | 1546 | 2.892 | $2.2x10^{-5}$ |
| E043 | Blood & T-cell | Primary T helper cells (from PB) | 28 | 1863 | 2.731 | $2.6x10^{-5}$ |
| E038 | Blood & T-cell | Primary T helper naive cells (from PB) | 26 | 1694 | 2.751 | $4.2x10^{-5}$ |
| E123 | ENCODE2012 | K562 leukemia | 22 | 1320 | 2.919 | $5.0x10^{-5}$ |
| E031 | HSC & B-cell | Primary B cells from cord blood | 20 | 1163 | 2.974 | $7.6x10^{-5}$ |
| E041 | Blood & T-cell | Primary T helper cells PMA-I stimulated | 28 | 2016 | 2.496 | $1.1x10^{-4}$ |
| E048 | Blood & T-cell | Primary T CD8+ memory cells (from PB) | 25 | 1704 | 2.596 | $1.1x10^{-4}$ |
| E039 | Blood & T-cell | Primary T helper naive cells (from PB) | 25 | 1751 | 2.518 | $2.3x10^{-4}$ |
| E037 | Blood & T-cell | Primary T helper memory cells (from PB) | 25 | 1783 | 2.467 | $2.7x10^{-4}$ |
| E115 | ENCODE2012 | Dnd41 T cell leukemia | 19 | 1198 | 2.705 | $3.3x10^{-4}$ |
| E047 | Blood & T-cell | Primary T CD8+ naive cells (from PB) | 24 | 1707 | 2.456 | $3.7x10^{-4}$ |
| E116 | ENCODE2012 | GM12878 lymphoblastoid | 21 | 1441 | 2.502 | $5.1x10^{-4}$ |
| E042 | Blood & T-cell | Primary T helper 17 cells PMA-I stimulated | 23 | 1654 | 2.409 | $5.8x10^{-4}$ |
| E046 | HSC & B-cell | Primary natural killer cells (from PB) | 21 | 1466 | 2.455 | $6.3x10^{-4}$ |
| E040 | Blood & T-cell | Primary T helper memory cells (from PB) | 23 | 1672 | 2.380 | $6.5x10^{-4}$ |
| E030 | HSC & B-cell | Primary neutrophils (from PB) | 18 | 1233 | 2.456 | $1.4x10^{-3}$ |
| E062 | Blood & T-cell | Primary mononuclear cells (from PB) | 14 | 913 | 2.520 | 0.005 |
| E050 | HSC & B-cell | Primary HSCs G-CSF-mobilized female | 24 | 2077 | 1.967 | 0.006 |
| E051 | HSC & B-cell | Primary HSCs G-CSF-mobilized male | 21 | 1766 | 1.996 | 0.008 |
| E045 | Blood & T-cell | Primary T cells effector/memory enriched (PB) | 17 | 1399 | 1.998 | 0.014 |
| E066 | Other | Liver | 21 | 1895 | 1.844 | 0.016 |
| E036 | HSC & B-cell | Primary HSCs short term culture | 20 | 1751 | 1.897 | 0.018 |
| E035 | HSC & B-cell | Primary HSCs | 13 | 1056 | 1.983 | 0.029 |
| E033 | Blood & T-cell | Primary T cells from cord blood | 9 | 971 | 1.439 | 0.302 |

Abbreviations: H3K4me1, histone H3 lysine 4 monomethylation; EID, epigenome identifier; PB, peripheral blood; HSC, hematopoietic stem cell.
Enrichment defined as any result meeting the Bonferroni-corrected p-value threshold ($.05/29=1.7x10^{-3}$ for histone modification marks).

**Table S8:** H3K4me3 histone mark enrichment analysis results for all 29 blood- and liver-related epigenomes

| EID | EID Grouping | Epigenome Name | H3K4me3 Peak Overlap, PBC SNPs (N=106) | H3K4me3 Peak Overlap, Comparison (N=16036) | OR | Fisher P |
|---|---|---|---|---|---|---|
| E032 | HSC & B-cell | Primary B cells (from PB) | 13 | 306 | 7.182 | $1.7x10^{-7}$ |
| E031 | HSC & B-cell | Primary B cells from cord blood | 13 | 335 | 6.550 | $4.5x10^{-7}$ |
| E116 | ENCODE2012 | GM12878 lymphoblastoid | 18 | 699 | 4.487 | $9.0x10^{-7}$ |
| E062 | Blood & T-cell | Primary mononuclear cells (from PB) | 12 | 335 | 5.982 | $2.9x10^{-6}$ |
| E044 | Blood & T-cell | Primary T regulatory cells (from PB) | 15 | 575 | 4.431 | $7.0x10^{-6}$ |
| E124 | ENCODE2012 | Monocytes-CD14+ RO01746 | 19 | 896 | 3.690 | $7.1x10^{-6}$ |
| E115 | ENCODE2012 | Dnd41 T cell leukemia | 12 | 389 | 5.134 | $1.3x10^{-5}$ |
| E034 | Blood & T-cell | Primary T cells from primary blood (from PB) | 11 | 357 | 5.084 | $3.0x10^{-5}$ |
| E118 | ENCODE2012 | HepG2 hepatocellular carcinoma | 12 | 440 | 4.524 | $4.1x10^{-5}$ |
| E042 | Blood & T-cell | Primary T helper 17 cells PMA-I stimulated | 14 | 669 | 3.495 | $1.5x10^{-4}$ |
| E050 | HSC & B-cell | Primary HSCs G-CSF-mobilized female | 13 | 610 | 3.534 | $2.2x10^{-4}$ |
| E029 | HSC & B-cell | Primary monocytes (from PB) | 7 | 194 | 5.772 | $3.6x10^{-4}$ |
| E123 | ENCODE2012 | K562 leukemia | 11 | 486 | 3.704 | $4.2x10^{-4}$ |
| E051 | HSC & B-cell | Primary HSCs G-CSF-mobilized male | 14 | 769 | 3.021 | $5.9x10^{-4}$ |
| E041 | Blood & T-cell | Primary T helper cells PMA-I stimulated | 14 | 772 | 3.008 | $6.2x10^{-4}$ |
| E046 | HSC & B-cell | Primary natural killer cells (from PB) | 8 | 283 | 4.543 | $6.6x10^{-4}$ |
| E035 | HSC & B-cell | Primary HSCs | 9 | 358 | 4.063 | $7.0x10^{-4}$ |
| E047 | Blood & T-cell | Primary T CD8+ naive cells (from PB) | 10 | 437 | 3.718 | $7.1x10^{-4}$ |
| E066 | Other | Liver | 14 | 786 | 2.952 | $7.3x10^{-4}$ |
| E030 | HSC & B-cell | Primary neutrophils (from PB) | 10 | 469 | 3.457 | 0.001 |
| E037 | Blood & T-cell | Primary T helper memory cells (from PB) | 11 | 589 | 3.036 | 0.002 |
| E038 | Blood & T-cell | Primary T helper naive cells (from PB) | 11 | 618 | 2.888 | 0.003 |
| E033 | Blood & T-cell | Primary T cells from cord blood | 8 | 368 | 3.475 | 0.003 |
| E043 | Blood & T-cell | Primary T helper cells (from PB) | 10 | 554 | 2.911 | 0.004 |
| E048 | Blood & T-cell | Primary T CD8+ memory cells (from PB) | 11 | 656 | 2.714 | 0.004 |
| E036 | HSC & B-cell | Primary HSCs short term culture | 9 | 517 | 2.785 | 0.008 |
| E045 | Blood & T-cell | Primary T cells effector/memory enriched (PB) | 10 | 634 | 2.530 | 0.010 |
| E039 | Blood & T-cell | Primary T helper naive cells (from PB) | 9 | 615 | 2.326 | 0.022 |
| E040 | Blood & T-cell | Primary T helper memory cells (from PB) | 10 | 770 | 2.065 | 0.037 |

Abbreviations: H3K4me3, histone H3 lysine 4 trimethylation; EID, epigenome identifier; PB, peripheral blood; HSC, hematopoietic stem cell.
Enrichment defined as any result meeting the Bonferroni-corrected p-value threshold ($.05/29=1.7x10^{-3}$ for histone modification marks).

**Table S9:** DNase enrichment analysis results for all 11 blood- and liver-related epigenomes

| EID | EID Grouping | Epigenome Name | DNase Peak Overlap, PBC SNPs (N=106) | DNase Peak Overlap, Comparison (N=16036) | OR | Fisher P |
|---|---|---|---|---|---|---|
| E033 | Blood & T-cell | Primary T cells from cord blood | 8 | 337 | 3.802 | 0.002 |
| E034 | Blood & T-cell | Primary T cells from primary blood (from PB) | 9 | 437 | 3.311 | 0.003 |
| E124 | ENCODE2012 | Monocytes-CD14+ RO01746 | 9 | 439 | 3.296 | 0.003 |
| E032 | HSC & B-cell | Primary B cells (from PB) | 9 | 463 | 3.120 | 0.004 |
| E118 | ENCODE2012 | HepG2 hepatocellular carcinoma | 6 | 319 | 2.956 | 0.020 |
| E050 | HSC & B-cell | Primary HSCs G-CSF-mobilized female | 7 | 428 | 2.578 | 0.025 |
| E116 | ENCODE2012 | GM12878 lymphoblastoid | 8 | 581 | 2.171 | 0.059 |
| E123 | ENCODE2012 | K562 leukemia | 7 | 572 | 1.911 | 0.107 |
| E046 | HSC & B-cell | Primary natural killer cells (from PB) | 5 | 397 | 1.950 | 0.194 |
| E029 | HSC & B-cell | Primary monocytes (from PB) | 4 | 331 | 1.861 | 0.286 |
| E051 | HSC & B-cell | Primary HSCs G-CSF-mobilized male | 4 | 352 | 1.747 | 0.300 |

Abbreviations: EID, epigenome identifier; PB, peripheral blood; HSC, hematopoietic stem cell.
Enrichment defined as any result meeting the Bonferroni-corrected p-value threshold ($.05/11=4.5x10^{-3}$ for DNAse).

Table S10: Functional annotations of SNPs in all replicated 3-SNP haplotypes

| Chr | 3-SNP Haplotype | SNP | Chr position (hg19) | Ontology | Mapped Gene | DHS overlap[a], # EIDs (PBC EIDs) | H3K4me1 overlap[a], # EIDs (PBC EIDs) | H3K4me3 overlap[a], # EIDs (PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | (rs3117106=C) and ((rs206018=G) or (rs9501179=A)) | rs3117106 | 32343369 | intergenic | *C6orf10*(dist=3680), *HCG23*(dist=14918) | 0 (0) | 11 (0) | 3 (0) | NA | 1 | Whole Blood (*C4A, C4B, CYP21A1P, CYP21A2, HLA-DRB5, PBX2*); Lymphoblastoid (*C4A*); Liver (*C4A, CYP21A1P, HLA-DMA, STK19P, TNXA*) |
| 6 | (rs3117106=C) and ((rs206018=G) or (rs9501179=A)) | rs206018 | 32177880 | intronic | *NOTCH4* | 0 (0) | 8 (2) | 2 (0) | NA | 2 | Whole Blood (*GPSM3\|NOTCH4, HLA-DRA, LY6G5C, SKIV2L*) |
| 6 | (rs3117106=C) and ((rs206018=G) or (rs9501179=A)) | rs9501179 | 32292993 | intronic | *C6orf10* | 0 (0) | 2 (1) | 2 (0) | NA | 4 | Whole Blood (*HLA-DRA*); Lymphoblastoid (*C2*) |
| 6 | (rs3129881=C) or ((rs375244=A) and (rs3132947=G)) | rs3129881 | 32409484 | intronic | *HLA-DRA* | 0 (0) | 42 (20) | 50 (17) | GM12878 (OCT2, POL2, POL24H8, POU2F2); GM12891 (OCT2, POL2, POL24H8, POU2F2); GM12892 (POL2, POL24H8) | 3 | Whole Blood (*C4A, C4B, HLA-DQA1, HLA-DQA2, HLA-DQB2, HLA-DRB5*); Lymphoblastoid (*HLA-DQA2, HLA-DRB5*); Liver (*HLA-DRB5*) |
| 6 | (rs3129881=C) or ((rs375244=A) and (rs3132947=G)) | rs375244 | 32191457 | intronic | *NOTCH4* | 0 (0) | 69 (15) | 38 (3) | NA | 3 | Whole Blood (*GPSM3\|NOTCH4*) |
| 6 | (rs3129881=C) or ((rs375244=A) and (rs3132947=G)) | rs3132947 | 32176782 | intronic | *NOTCH4* | 0 (0) | 2 (1) | 4 (0) | NA | 3 | Whole Blood (*AGPAT1, C4A, C4B, CYP21A1P, HLA-DRA, PBX2, SKIV2L*); Lymphoblastoid (*C4A, HLA-DQA1, RNF5*); Liver (*AGPAT1, RNF5*) |
| 6 | ((rs35372932=T) and (rs9269190=C)) or (rs9270493=C) | rs35372932 | 32564985 | intergenic | *HLA-DRB1*(dist=7372), *HLA-DQA1*(dist=40198) | 0 (0) | 1 (1) | 0 (0) | NA | 1 | NA |
| 6 | ((rs35372932=T) and (rs9269190=C)) or (rs9270493=C) | rs9269190 | 32448500 | intergenic | *HLA-DRA*(dist=35674), *HLA-DRB5*(dist=36654) | 2 (2) | 4 (4) | 2 (2) | NA | 2 | Whole Blood (*HLA-DRA*) |
| 6 | ((rs35372932=T) and (rs9269190=C)) or (rs9270493=C) | rs9270493 | 32559110 | intergenic | *HLA-DRB1*(dist=1497), *HLA-DQA1*(dist=46073) | 0 (0) | 12 (9) | 2 (1) | NA | 1 | NA |
| 6 | (rs9268977=T) or ((rs3135395=T) or (rs550513=T)) | rs9268977 | 32434939 | intergenic | *HLA-DRA*(dist=22113), *HLA-DRB5*(dist=50215) | 0 (0) | 0 (0) | 0 (0) | NA | 0 | Whole Blood (*C2, C4B, HLA-DRB5*); Liver (*C4A, HLA-DMA, HLA-DRB5*) |
| 6 | (rs9268977=T) or ((rs3135395=T) or (rs550513=T)) | rs3135395 | 32405192 | intergenic | *BTNL2*(dist=30285), *HLA-DRA*(dist=2427) | 8 (8) | 45 (19) | 16 (8) | GM10847 (NFKB); GM12878 (NFKB); GM12891 (NFKB, POL2); GM12892 (NFKB, POL24H8); GM15510 (NFKB); GM18505 (NFKB); GM18951 (NFKB); GM19099 (NFKB); GM19193 (NFKB) | 1 | Whole Blood (*HLA-DQB1-AS1, HLA-DRB5, HLA-DRB6, HLA-DRB9*); Lymphoblastoid (*HLA-DRB5, HLA-DRB6, HLA-DRB9*); Liver (*HLA-DRB5*) |
| 6 | (rs9268977=T) or ((rs3135395=T) or (rs550513=T)) | rs550513 | 31920687 | intronic; downstream | *NELFE; CFB* | 0 (0) | 42 (7) | 24 (2) | HepG2 (POL2) | 0 | Whole Blood (*HSPA1B, LY6G6F, RDBP, SKIV2L*); Lymphoblastoid (*SKIV2L*) |
| 6 | ((rs9268634=G) or (rs35344500=C)) and (rs9275175=G) | rs9268634 | 32406530 | intergenic | *BTNL2*(dist=31623), *HLA-DRA*(dist=1089) | 1 (1) | 30 (13) | 14 (5) | NA | 5 | Whole Blood (*HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRA, HLA-DRB1, HLA-DRB6*); Lymphoblastoid (*HLA-DQA2, HLA-DQB2, HLA-DRB6, HLA-DRB9, NOTCH4*); Liver (*HLA-DQA2*) |

| Chr | 3-SNP Haplotype | SNP | Chr position (hg19) | Ontology | Mapped Gene | DHS overlap[a], # EIDs (PBC EIDs) | H3K4me1 overlap[a], # EIDs (PBC EIDs) | H3K4me3 overlap[a], # EIDs (PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | ((rs9268634=G) or (rs35344500=C)) and (rs9275175=G) | rs35344500 | 32609525 | intronic | *HLA-DQA1* | 1 (1) | 20 (15) | 8 (6) | NA | 0 | Whole Blood (*HLA-DQA1, HLA-DQB1, HLA-DQB2, HLA-DRB1, HLA-DRB5, PSMB9*); Lymphoblastoid (*C4A, HLA-DQB2, HLA-DRB1, HLA-DRB9*); Liver (*C4A, HLA-DMA, HLA-DQB1, STK19P, TNXA*) |
| 6 | ((rs9268634=G) or (rs35344500=C)) and (rs9275175=G) | rs9275175 | 32654147 | intergenic | *HLA-DQB1*(dist=19681), *HLA-DQA2*(dist=55016) | 0 (0) | 26 (24) | 7 (6) | NA | 2 | NA |
| 6 | ((rs9268831=T) or (rs9269190=T)) or (rs9270652=C) | rs9268831 | 32427748 | intergenic | *HLA-DRA*(dist=14922), *HLA-DRB5*(dist=57406) | 5 (1) | 31 (19) | 71 (16) | GM18951 (POL2); MCF-7 (CMYC, HAE2F1, POL2) | 1 | Whole Blood (*HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DQB2, HLA-DRB1, HLA-DRB6, HLA-DRB9*); Lymphoblastoid (*HLA-DQA2, HLA-DQB2, HLA-DRB6, HLA-DRB9, NOTCH4*); Liver (*HLA-DQA2, HLA-DQB2*) |
| 6 | ((rs9268831=T) or (rs9269190=T)) or (rs9270652=C) | rs9269190 | 32448500 | intergenic | *HLA-DRA*(dist=35674), *HLA-DRB5*(dist=36654) | 2 (2) | 4 (4) | 2 (2) | NA | 2 | Whole Blood (*HLA-DRA*) |
| 6 | ((rs9268831=T) or (rs9269190=T)) or (rs9270652=C) | rs9270652 | 32565905 | intergenic | *HLA-DRB1*(dist=8292), *HLA-DQA1*(dist=39278) | 1 (1) | 1 (1) | 2 (0) | NA | 0 | NA |
| 6 | (rs3132947=G) or ((rs9501179=G) and (rs41546114=C)) | rs3132947 | 32176782 | intronic | *NOTCH4* | 0 (0) | 2 (1) | 4 (0) | NA | 3 | Whole Blood (*AGPAT1, C4A, C4B, CYP21A1P, HLA-DRA, PBX2, SKIV2L*); Lymphoblastoid (*C4A, HLA-DQA1, RNF5*); Liver (*AGPAT1, RNF5*) |
| 6 | (rs3132947=G) or ((rs9501179=G) and (rs41546114=C)) | rs9501179 | 32292993 | intronic | *C6orf10* | 0 (0) | 2 (1) | 2 (0) | NA | 4 | Whole Blood (*HLA-DRA*); Lymphoblastoid (*C2*) |
| 6 | (rs3132947=G) or ((rs9501179=G) and (rs41546114=C)) | rs41546114 | 31382831 | exonic; UTR3 | *MICA*; *MICA*(NM_001289152:c.*72T>C, NM_001289153:c.*72T>C, NM_001289154:c.*72T>C, NM_001177519:c.*72T>C) | 3 (1) | 15 (8) | 9 (3) | NA | 0 | NA |
| 6 | (rs9268831=T) or ((rs2395194=G) or (rs387608=A)) | rs9268831 | 32427748 | intergenic | *HLA-DRA*(dist=14922), *HLA-DRB5*(dist=57406) | 5 (1) | 31 (19) | 71 (16) | GM18951 (POL2); MCF-7 (CMYC, HAE2F1, POL2) | 1 | Whole Blood (*HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DQB2, HLA-DRB1, HLA-DRB6, HLA-DRB9*); Lymphoblastoid (*HLA-DQA2, HLA-DQB2, HLA-DRB6, HLA-DRB9, NOTCH4*); Liver (*HLA-DQA2, HLA-DQB2*) |
| 6 | (rs9268831=T) or ((rs2395194=G) or (rs387608=A)) | rs2395194 | 32447953 | intergenic | *HLA-DRA*(dist=35127), *HLA-DRB5*(dist=37201) | 0 (0) | 0 (0) | 0 (0) | NA | 0 | NA |
| 6 | (rs9268831=T) or ((rs2395194=G) or (rs387608=A)) | rs387608 | 31941557 | intronic | *STK19* | 0 (0) | 102 (23) | 82 (14) | NA | 4 | Whole Blood (*HSPA1B, LSM2, RDBP, SKIV2L*); Lymphoblastoid (*SKIV2L*) |
| 6 | (rs3132947=G) or ((rs9268014=C) and (rs41546114=C)) | rs3132947 | 32176782 | intronic | *NOTCH4* | 0 (0) | 2 (1) | 4 (0) | NA | 3 | Whole Blood (*AGPAT1, C4A, C4B, CYP21A1P, HLA-DRA, PBX2, SKIV2L*); Lymphoblastoid (*C4A, HLA-DQA1, RNF5*); Liver (*AGPAT1, RNF5*) |
| 6 | (rs3132947=G) or ((rs9268014=C) and (rs41546114=C)) | rs9268014 | 32224852 | intergenic | *NOTCH4*(dist=33008), *C6orf10*(dist=35623) | 0 (0) | 8 (2) | 12 (0) | NA | 0 | Whole Blood (*C4B, HLA-DRB5*) |

| Chr | 3-SNP Haplotype | SNP | Chr position (hg19) | Ontology | Mapped Gene | DHS overlap[a], # EIDs (PBC EIDs) | H3K4me1 overlap[a], # EIDs (PBC EIDs) | H3K4me3 overlap[a], # EIDs (PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | (rs3132947=G) or ((rs9268014=C) and (rs41546114=C)) | rs41546114 | 31382831 | exonic; UTR3 | *MICA*; *MICA*(NM_001289152:c.*72T>C, NM_001289153:c.*72T>C, NM_001289154:c.*72T>C, NM_001177519:c.*72T>C) | 3 (1) | 15 (8) | 9 (3) | NA | 0 | NA |
| 6 | (rs3132947=G) or ((rs9268055=T) and (rs41546114=C)) | rs3132947 | 32176782 | intronic | *NOTCH4* | 0 (0) | 2 (1) | 4 (0) | NA | 3 | Whole Blood (*AGPAT1, C4A, C4B, CYP21A1P, HLA-DRA, PBX2, SKIV2L*); Lymphoblastoid (*C4A, HLA-DQA1, RNF5*); Liver (*AGPAT1, RNF5*) |
| 6 | (rs3132947=G) or ((rs9268055=T) and (rs41546114=C)) | rs9268055 | 32230608 | intergenic | *NOTCH4*(dist=38764), *C6orf10*(dist=29867) | 0 (0) | 0 (0) | 0 (0) | NA | 1 | Whole Blood (*C4B, HLA-DRB5*); Liver (*C4A, CYP21A1P*) |
| 6 | (rs3132947=G) or ((rs9268055=T) and (rs41546114=C)) | rs41546114 | 31382831 | exonic; UTR3 | *MICA*; *MICA*(NM_001289152:c.*72T>C, NM_001289153:c.*72T>C, NM_001289154:c.*72T>C, NM_001177519:c.*72T>C) | 3 (1) | 15 (8) | 9 (3) | NA | 0 | NA |
| 6 | ((rs241437=G) or (rs3129299=C)) or (rs9276909=C) | rs241437 | 32797684 | intronic | *TAP2* | 0 (0) | 10 (7) | 5 (1) | T-REx-HEK293 (ZNF263) | 1 | Whole Blood (*HLA-DMA, HLA-DOB, HLA-DRB5, PSMB9, TAP2*); Liver (*HLA-DRB5*) |
| 6 | ((rs241437=G) or (rs3129299=C)) or (rs9276909=C) | rs3129299 | 32900787 | intergenic | *LOC100294145*(dist=29252), *HLA-DMB*(dist=1619) | 0 (0) | 28 (11) | 9 (2) | NA | 2 | Whole Blood (*HLA-DMA, TAP2*) |
| 6 | ((rs241437=G) or (rs3129299=C)) or (rs9276909=C) | rs9276909 | 32850839 | intergenic | *PSMB9*(dist=23211), *LOC100294145*(dist=11114) | 13 (0) | 29 (2) | 9 (1) | MCF10A-Er-Src (STAT3) | 0 | Whole Blood (*HLA-DMA, HLA-DPB1, PSMB9, PSMB9\|TAP1, TAP2*); Lymphoblastoid (*PSMB9*) |
| 6 | ((rs2395194=G) and (rs16870908=G)) or (rs9268831=T) | rs2395194 | 32447953 | intergenic | *HLA-DRA*(dist=35127), *HLA-DRB5*(dist=37201) | 0 (0) | 0 (0) | 0 (0) | NA | 0 | NA |
| 6 | ((rs2395194=G) and (rs16870908=G)) or (rs9268831=T) | rs16870908 | 32790089 | exonic | *TAP2* | 0 (0) | 5 (3) | 2 (2) | NA | 1 | Whole Blood (*HLA-DMB, HLA-DOA, HLA-DOB, PSMB8, PSMB9, TAP2*) |
| 6 | ((rs2395194=G) and (rs16870908=G)) or (rs9268831=T) | rs9268831 | 32427748 | intergenic | *HLA-DRA*(dist=14922), *HLA-DRB5*(dist=57406) | 5 (1) | 31 (19) | 71 (16) | GM18951 (POL2); MCF-7 (CMYC, HAE2F1, POL2) | 1 | Whole Blood (*HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB1-AS1, HLA-DQB2, HLA-DRB1, HLA-DRB6, HLA-DRB9*); Lymphoblastoid (*HLA-DQA2, HLA-DQB2, HLA-DRB6, HLA-DRB9, NOTCH4*); Liver (*HLA-DQA2, HLA-DQB2*) |
| 6 | ((rs9268213=A) and (rs41546114=C)) or (rs3132947=G) | rs9268213 | 32282081 | intronic | *C6orf10* | 0 (0) | 7 (0) | 0 (0) | NA | 2 | Whole Blood (*C4A, C4B, HLA-DRB5*); Lymphoblastoid (*HLA-DQB1*); Liver (*C4A, CYP21A1P, HLA-DMA*) |
| 6 | ((rs9268213=A) and (rs41546114=C)) or (rs3132947=G) | rs41546114 | 31382831 | exonic; UTR3 | *MICA*; *MICA*(NM_001289152:c.*72T>C, NM_001289153:c.*72T>C, NM_001289154:c.*72T>C, NM_001177519:c.*72T>C) | 3 (1) | 15 (8) | 9 (3) | NA | 0 | NA |
| 6 | ((rs9268213=A) and (rs41546114=C)) or (rs3132947=G) | rs3132947 | 32176782 | intronic | *NOTCH4* | 0 (0) | 2 (1) | 4 (0) | NA | 3 | Whole Blood (*AGPAT1, C4A, C4B, CYP21A1P, HLA-DRA, PBX2, SKIV2L*); Lymphoblastoid (*C4A, HLA-DQA1, RNF5*); Liver (*AGPAT1, RNF5*) |

| Chr | 3-SNP Haplotype | SNP | Chr position (hg19) | Ontology | Mapped Gene | DHS overlap[a], # EIDs (PBC EIDs) | H3K4me1 overlap[a], # EIDs (PBC EIDs) | H3K4me3 overlap[a], # EIDs (PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | ((rs4538748=C) or (rs2064476=G)) and (rs35344500=A) | rs4538748 | 32657505 | intergenic | *HLA-DQB1*(dist=23039), *HLA-DQA2*(dist=51658) | 5 (5) | 26 (21) | 6 (4) | GM10847 (NFKB); GM12878 (BATF, BCL11A, BCLAF1, IRF4, MEF2A, NFKB, OCT2, PAX5C20, PAX5N19, POL2, POL24H8, POU2F2, RFX5, SP1, SRF, TAF1, TBP, YY1); GM12891 (NFKB, OCT2, PAX5C20, POL2, POL24H8, POU2F2, TAF1, YY1); GM12892 (NFKB, PAX5C20, POL2, POL24H8, TAF1, YY1); GM15510 (NFKB); GM18505 (NFKB, POL2); GM18951 (NFKB); GM19099 (NFKB, POL2); Raji (POL2) | 2 | Whole Blood (*CYP21A1P, HLA-DOB, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2*); Lymphoblastoid (*HLA-DQA2, HLA-DQB2, NOTCH4*); Liver (*HLA-DQB2*) |
| 6 | ((rs4538748=C) or (rs2064476=G)) and (rs35344500=A) | rs2064476 | 33073322 | intergenic | *HLA-DPB1*(dist=15849), *HLA-DPB2*(dist=6971) | 0 (0) | 8 (3) | 3 (0) | NA | 3 | Whole Blood (*HLA-DPB2, RPL32P1*); Lymphoblastoid (*RPL32P1*) |
| 6 | ((rs4538748=C) or (rs2064476=G)) and (rs35344500=A) | rs35344500 | 32609525 | intronic | *HLA-DQA1* | 1 (1) | 20 (15) | 8 (6) | NA | 0 | Whole Blood (*HLA-DQA1, HLA-DQB1, HLA-DQB2, HLA-DRB1, HLA-DRB5, PSMB9*); Lymphoblastoid (*C4A, HLA-DQB2, HLA-DRB1, HLA-DRB9*); Liver (*C4A, HLA-DMA, HLA-DQB1, STK19P, TNXA*) |
| 6 | ((rs2244027=A) or (rs2242665=C)) or (rs2071591=G) | rs2244027 | 31347566 | intergenic | *HLA-B*(dist=22577), *MICA*(dist=19995) | 1 (0) | 16 (0) | 8 (1) | NA | 2 | Whole Blood (*HLA-C, MICA, MICB*) |
| 6 | ((rs2244027=A) or (rs2242665=C)) or (rs2071591=G) | rs2242665 | 31839309 | exonic | *SLC44A4* | 0 (0) | 55 (8) | 22 (3) | NA | 0 | Whole Blood (*C6orf48, CSNK2B, CYP21A1P, HLA-DQA2, HLA-DQB2, HLA-DRB6, HSPA1B, RDBP, SKIV2L*); Lymphoblastoid (*SKIV2L*) |
| 6 | ((rs2244027=A) or (rs2242665=C)) or (rs2071591=G) | rs2071591 | 31515799 | intronic | *NFKBIL1* | 1 (1) | 120 (29) | 123 (26) | NA | 0 | Whole Blood (*AIF1, BAT1, CSNK2B, DDX39B, HCP5, HLA-DRB5, LST1, TNF*) |
| 6 | ((rs9296088=A) and (rs12207818=C)) and (rs181997=A) | rs9296088 | 33125742 | intergenic | *HLA-DPB2*(dist=28852), *COL11A2*(dist=4727) | 0 (0) | 6 (2) | 4 (2) | NA | 6 | NA |
| 6 | ((rs9296088=A) and (rs12207818=C)) and (rs181997=A) | rs12207818 | 33809805 | intergenic | *MLN*(dist=38012), *LINC01016*(dist=47483) | 0 (0) | 42 (10) | 11 (1) | K562 (TFIIIC110) | 1 | NA |
| 6 | ((rs9296088=A) and (rs12207818=C)) and (rs181997=A) | rs181997 | 32900718 | intergenic | *LOC100294145*(dist=29183), *HLA-DMB*(dist=1688) | 0 (0) | 28 (11) | 10 (2) | NA | 4 | Whole Blood (*HLA-DMA, TAP2*) |
| 6 | ((rs2071287=C) and (rs3094596=C)) or (rs185819=C) | rs2071287 | 32170433 | intronic | *NOTCH4* | 1 (0) | 17 (1) | 8 (1) | NA | 1 | Whole Blood (*GPSM3\|NOTCH4, HLA-DRA, SKIV2L*); Lymphoblastoid (*HLA-DQA1*); Liver (*AGPAT1*) |
| 6 | ((rs2071287=C) and (rs3094596=C)) or (rs185819=C) | rs3094596 | 31350579 | intergenic | *HLA-B*(dist=25590), *MICA*(dist=16982) | 0 (0) | 3 (0) | 2 (0) | NA | 3 | Whole Blood (*AIF1, ATP6V1G2\|BAT1, HCP5, LST1, LTA, MICB*) |
| 6 | ((rs2071287=C) and (rs3094596=C)) or (rs185819=C) | rs185819 | 32050067 | exonic | *TNXB* | 2 (0) | 70 (1) | 34 (2) | NA | 4 | Whole Blood (*CYP21A1P, GPSM3\|NOTCH4, HLA-DQA2, HLA-DQB2, HLA-DRB6, SKIV2L*); Liver (*CYP21A1P*) |
| 6 | (rs3132947=T) and ((rs404860=C) or (rs41546114=T)) | rs3132947 | 32176782 | intronic | *NOTCH4* | 0 (0) | 2 (1) | 4 (0) | NA | 3 | Whole Blood (*AGPAT1, C4A, C4B, CYP21A1P, HLA-DRA, PBX2, SKIV2L*); Lymphoblastoid (*C4A, HLA-DQA1, RNF5*); Liver (*AGPAT1, RNF5*) |

| Chr | 3-SNP Haplotype | SNP | Chr position (hg19) | Ontology | Mapped Gene | DHS overlap[a], # EIDs (PBC EIDs) | H3K4me1 overlap[a], # EIDs (PBC EIDs) | H3K4me3 overlap[a], # EIDs (PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | (rs3132947=T) and ((rs404860=C) or (rs41546114=T)) | rs404860 | 32184345 | intronic | *NOTCH4* | 0 (0) | 5 (1) | 3 (0) | NA | 0 | NA |
| 6 | (rs3132947=T) and ((rs404860=C) or (rs41546114=T)) | rs41546114 | 31382831 | exonic; UTR3 | *MICA*; *MICA*(NM_001289152:c.*72T>C, NM_001289153:c.*72T>C, NM_001289154:c.*72T>C, NM_001177519:c.*72T>C) | 3 (1) | 15 (8) | 9 (3) | NA | 0 | NA |
| 6 | (rs11754586=T) or ((rs3131932=G) or (rs9262537=G)) | rs11754586 | 31001911 | intronic | *MUC22* | 0 (0) | 1 (0) | 1 (0) | NA | 3 | Whole Blood (*VARSL*); Lymphoblastoid (*VARS2*) |
| 6 | (rs11754586=T) or ((rs3131932=G) or (rs9262537=G)) | rs3131932 | 30940328 | intergenic | *DPCR1*(dist=18330), *MUC21*(dist=11157) | 0 (0) | 9 (0) | 5 (0) | NA | 2 | Whole Blood (*CCHCR1, FLOT1, IER3, LINC00243, VARSL*) |
| 6 | (rs11754586=T) or ((rs3131932=G) or (rs9262537=G)) | rs9262537 | 30990224 | intronic | *MUC22* | 0 (0) | 1 (0) | 1 (0) | NA | 1 | Whole Blood (*HLA-S*) |
| 6 | ((rs10947251=A) or (rs41546114=T)) and (rs3132947=T) | rs10947251 | 32261952 | intronic | *C6orf10* | 0 (0) | 1 (0) | 0 (0) | NA | 1 | Lymphoblastoid (*C2*) |
| 6 | ((rs10947251=A) or (rs41546114=T)) and (rs3132947=T) | rs41546114 | 31382831 | exonic; UTR3 | *MICA*; *MICA*(NM_001289152:c.*72T>C, NM_001289153:c.*72T>C, NM_001289154:c.*72T>C, NM_001177519:c.*72T>C) | 3 (1) | 15 (8) | 9 (3) | NA | 0 | NA |
| 6 | ((rs10947251=A) or (rs41546114=T)) and (rs3132947=T) | rs3132947 | 32176782 | intronic | *NOTCH4* | 0 (0) | 2 (1) | 4 (0) | NA | 3 | Whole Blood (*AGPAT1, C4A, C4B, CYP21A1P, HLA-DRA, PBX2, SKIV2L*); Lymphoblastoid (*C4A, HLA-DQA1, RNF5*); Liver (*AGPAT1, RNF5*) |
| 6 | ((rs3132947=G) or (rs9266632=G)) or (rs185819=C) | rs3132947 | 32176782 | intronic | *NOTCH4* | 0 (0) | 2 (1) | 4 (0) | NA | 3 | Whole Blood (*AGPAT1, C4A, C4B, CYP21A1P, HLA-DRA, PBX2, SKIV2L*); Lymphoblastoid (*C4A, HLA-DQA1, RNF5*); Liver (*AGPAT1, RNF5*) |
| 6 | ((rs3132947=G) or (rs9266632=G)) or (rs185819=C) | rs9266632 | 31346902 | intergenic | *HLA-B*(dist=21913), *MICA*(dist=20659) | 1 (1) | 17 (1) | 11 (1) | NA | 2 | Whole Blood (*HLA-C, PSORS1C3*); Lymphoblastoid (*PSORS1C3*) |
| 6 | ((rs3132947=G) or (rs9266632=G)) or (rs185819=C) | rs185819 | 32050067 | exonic | *TNXB* | 2 (0) | 70 (1) | 34 (2) | NA | 4 | Whole Blood (*CYP21A1P, GPSM3\|NOTCH4, HLA-DQA2, HLA-DQB2, HLA-DRB6, SKIV2L*); Liver (*CYP21A1P*) |
| 6 | (rs805273=A) and ((rs805267=G) or (rs9266774=T)) | rs805273 | 31665452 | intronic | *ABHD16A* | 1 (0) | 52 (17) | 24 (2) | NA | 5 | Whole Blood (*AIF1, LY6G5B, LY6G5C*) |
| 6 | (rs805273=A) and ((rs805267=G) or (rs9266774=T)) | rs805267 | 31639757 | exonic | *LY6G5B* | 1 (0) | 23 (8) | 16 (2) | HepG2 (POL2) | 3 | Whole Blood (*AIF1, LY6G5B*) |
| 6 | (rs805273=A) and ((rs805267=G) or (rs9266774=T)) | rs9266774 | 31352880 | intergenic | *HLA-B*(dist=27891), *MICA*(dist=14681) | 1 (1) | 7 (0) | 10 (1) | NA | 5 | Whole Blood (*HLA-C, LTA, MICA, MICB*) |

| Chr | 3-SNP Haplotype | SNP | Chr position (hg19) | Ontology | Mapped Gene | DHS overlap[a], # EIDs (PBC EIDs) | H3K4me1 overlap[a], # EIDs (PBC EIDs) | H3K4me3 overlap[a], # EIDs (PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | (rs3828901=A) or ((rs707921=A) and (rs9266774=T)) | rs3828901 | 31463718 | intronic | *MICB* | 8 (0) | 101 (29) | 94 (29) | HUVEC (CFOS) | 1 | NA |
| 6 | (rs3828901=A) or ((rs707921=A) and (rs9266774=T)) | rs707921 | 31625541 | intronic | *APOM* | 4 (4) | 112 (28) | 70 (15) | HepG2 (POL2) | 1 | Whole Blood (*AIF1, LY6G5B*) |
| 6 | (rs3828901=A) or ((rs707921=A) and (rs9266774=T)) | rs9266774 | 31352880 | intergenic | *HLA-B*(dist=27891), *MICA*(dist=14681) | 1 (1) | 7 (0) | 10 (1) | NA | 5 | Whole Blood (*HLA-C, LTA, MICA, MICB*) |
| 6 | (rs3828901=G) and ((rs805268=A) or (rs9266774=C)) | rs3828901 | 31463718 | intronic | *MICB* | 8 (0) | 101 (29) | 94 (29) | HUVEC (CFOS) | 1 | NA |
| 6 | (rs3828901=G) and ((rs805268=A) or (rs9266774=C)) | rs805268 | 31638178 | upstream; downstream | *LY6G5B; CSNK2B* | 0 (0) | 42 (14) | 37 (5) | NA | 0 | Whole Blood (*AIF1*) |
| 6 | (rs3828901=G) and ((rs805268=A) or (rs9266774=C)) | rs9266774 | 31352880 | intergenic | *HLA-B*(dist=27891), *MICA*(dist=14681) | 1 (1) | 7 (0) | 10 (1) | NA | 5 | Whole Blood (*HLA-C, LTA, MICA, MICB*) |
| 6 | (rs3828901=A) or ((rs3828919=A) and (rs2523497=T)) | rs3828901 | 31463718 | intronic | *MICB* | 8 (0) | 101 (29) | 94 (29) | HUVEC (CFOS) | 1 | NA |
| 6 | (rs3828901=A) or ((rs3828919=A) and (rs2523497=T)) | rs3828919 | 31466057 | intronic | *MICB* | 12 (5) | 111 (28) | 125 (29) | GM12878 (POL2, ZEB1); GM12892 (POL2); GM19099 (POL2); H1-hESC (TAF1, TBP); HepG2 (HEY1, TAF1); K562 (CCNT2, ELF1, GABP, NRSF, POL2, ZBTB7A) | 4 | Whole Blood (*MICB*) |
| 6 | (rs3828901=A) or ((rs3828919=A) and (rs2523497=T)) | rs2523497 | 31376928 | intronic | *MICA* | 0 (0) | 60 (18) | 5 (1) | NA | 3 | Whole Blood (*AIF1, HCG27, MICB, NOTCH4*); Lymphoblastoid (*MICA*); Liver (*MICA*) |
| 6 | (rs707922=G) or ((rs1046089=A) and (rs9266774=C)) | rs707922 | 31625507 | intronic | *APOM* | 3 (3) | 112 (28) | 69 (15) | HepG2 (POL2) | 5 | Whole Blood (*AIF1, LY6G5B, LY6G5C*) |
| 6 | (rs707922=G) or ((rs1046089=A) and (rs9266774=C)) | rs1046089 | 31602967 | exonic | *PRRC2A* | 0 (0) | 38 (9) | 23 (2) | HepG2 (POL2) | 0 | Whole Blood (*AIF1, C4A, C4B, HCP5, HLA-DRB5, HSPA1B, LY6G5B, LY6G5C*); Lymphoblastoid (*HLA-DRB5*); Liver (*HLA-DRB5*) |
| 6 | (rs707922=G) or ((rs1046089=A) and (rs9266774=C)) | rs9266774 | 31352880 | intergenic | *HLA-B*(dist=27891), *MICA*(dist=14681) | 1 (1) | 7 (0) | 10 (1) | NA | 5 | Whole Blood (*HLA-C, LTA, MICA, MICB*) |
| 6 | ((rs3131932=G) or (rs28360042=C)) or (rs885950=C) | rs3131932 | 30940328 | intergenic | *DPCR1*(dist=18330), *MUC21*(dist=11157) | 0 (0) | 9 (0) | 5 (0) | NA | 2 | Whole Blood (*CCHCR1, FLOT1, IER3, LINC00243, VARSL*) |
| 6 | ((rs3131932=G) or (rs28360042=C)) or (rs885950=C) | rs28360042 | 31001781 | intronic | *MUC22* | 0 (0) | 1 (0) | 1 (0) | NA | 4 | Whole Blood (*HCG27, HLA-C, POU5F1, PSORS1C3*); Lymphoblastoid (*CCHCR1, HLA-C, PSORS1C3, TCF19*); Liver (*HLA-C, PSORS1C3*) |
| 6 | ((rs3131932=G) or (rs28360042=C)) or (rs885950=C) | rs885950 | 31140152 | intergenic | *POU5F1*(dist=1682), *PSORS1C3*(dist=1360) | 5 (0) | 40 (0) | 25 (2) | H1-hESC (POL2) | 6 | Whole Blood (*HCG27, VARSL*) |

| Chr | 3-SNP Haplotype | SNP | Chr position (hg19) | Ontology | Mapped Gene | DHS overlap[a], # EIDs (PBC EIDs) | H3K4me1 overlap[a], # EIDs (PBC EIDs) | H3K4me3 overlap[a], # EIDs (PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | ((rs9266774=T) and (rs2255741=A)) or (rs3828901=A) | rs9266774 | 31352880 | intergenic | *HLA-B*(dist=27891), *MICA*(dist=14681) | 1 (1) | 7 (0) | 10 (1) | NA | 5 | Whole Blood (*HLA-C, LTA, MICA, MICB*) |
| 6 | ((rs9266774=T) and (rs2255741=A)) or (rs3828901=A) | rs2255741 | 31605167 | intronic | *PRRC2A* | 12 (3) | 45 (7) | 14 (0) | GM12878 (POL2); K562 (POL2, POL24H8) | 2 | Whole Blood (*AIF1*) |
| 6 | ((rs9266774=T) and (rs2255741=A)) or (rs3828901=A) | rs3828901 | 31463718 | intronic | *MICB* | 8 (0) | 101 (29) | 94 (29) | HUVEC (CFOS) | 1 | NA |
| 6 | (rs3749946=A) and ((rs2517506=T) and (rs7741091=A)) | rs3749946 | 31448862 | intergenic | *HCG26*(dist=8677), *MICB*(dist=13796) | 0 (0) | 8 (3) | 3 (1) | NA | 3 | Whole Blood (*HCP5, LST1, MICB*) |
| 6 | (rs3749946=A) and ((rs2517506=T) and (rs7741091=A)) | rs2517506 | 31031680 | intergenic | *HCG22*(dist=4025), *C6orf15*(dist=47320) | 0 (0) | 13 (0) | 1 (1) | NA | 1 | Whole Blood (*CCHCR1, FLOT1, HLA-L, MICB*) |
| 6 | (rs3749946=A) and ((rs2517506=T) and (rs7741091=A)) | rs7741091 | 31352631 | intergenic | *HLA-B*(dist=27642), *MICA*(dist=14930) | 0 (0) | 7 (0) | 10 (1) | NA | 2 | Whole Blood (*ATP6V1G2\|BAT1, HCG27, HLA-C, HLA-S, MICA, MICB, NOTCH4, ZBTB12*); Liver (*MICA*) |
| 6 | (rs3828901=G) and ((rs12110785=T) or (rs2523644=T)) | rs3828901 | 31463718 | intronic | *MICB* | 8 (0) | 101 (29) | 94 (29) | HUVEC (CFOS) | 1 | NA |
| 6 | (rs3828901=G) and ((rs12110785=T) or (rs2523644=T)) | rs12110785 | 30997824 | exonic | *MUC22* | 0 (0) | 3 (0) | 3 (0) | NA | 1 | NA |
| 6 | (rs3828901=G) and ((rs12110785=T) or (rs2523644=T)) | rs2523644 | 31342484 | intergenic | *HLA-B*(dist=17495), *MICA*(dist=25077) | 0 (0) | 2 (0) | 2 (0) | HeLa-S3 (CTCF) | 2 | Whole Blood (*AIF1, ATP6V1G2\|BAT1, HCG27, HCP5, LTA, MICB*); Lymphoblastoid (*HCG27*); Liver (*HLA-C*) |
| 6 | ((rs2239888=T) or (rs3134769=C)) and (rs3828901=G) | rs2239888 | 30649912 | intronic | *PPP1R18* | 20 (6) | 125 (29) | 120 (29) | GM12878 (PU1); GM12891 (PU1); HUVEC (CFOS); HeLa-S3 (INI1, JUND, STAT1, TBP); HepG2 (JUND); K562 (CJUN, CMYC, FOSL1, JUNB, POL2, STAT1, STAT2, TAF1, ZBTB7A); MCF10A-Er-Src (STAT3) | 2 | NA |
| 6 | ((rs2239888=T) or (rs3134769=C)) and (rs3828901=G) | rs3134769 | 31205754 | intergenic | *HCG27*(dist=34009), *HLA-C*(dist=30772) | 0 (0) | 0 (0) | 0 (0) | NA | 1 | Whole Blood (*CCHCR1, HCG27*); Lymphoblastoid (*C4B*); Liver (*HCG27*) |
| 6 | ((rs2239888=T) or (rs3134769=C)) and (rs3828901=G) | rs3828901 | 31463718 | intronic | *MICB* | 8 (0) | 101 (29) | 94 (29) | HUVEC (CFOS) | 1 | NA |
| 6 | (rs2517681=T) or ((rs4148248=C) and (rs2735078=A)) | rs2517681 | 29932330 | intergenic | *HLA-A*(dist=18669), *HCG9*(dist=10562) | 2 (0) | 122 (29) | 94 (25) | HTB-11 (NRSF) | 1 | Whole Blood (*HCG4P3, HCG4P5, HLA-A, HLA-F, HLA-G, HLA-J, HLA-V, IFITM4P, MICD, PPP1R11, ZFP57, ZNRD1*); Lymphoblastoid (*HCG4P5, IFITM4P, MICE*); Liver (*ZFP57*) |
| 6 | (rs2517681=T) or ((rs4148248=C) and (rs2735078=A)) | rs4148248 | 30557566 | intronic | *ABCF1* | 0 (0) | 39 (6) | 12 (0) | NA | 5 | NA |

| Chr | 3-SNP Haplotype | SNP | Chr position (hg19) | Ontology | Mapped Gene | DHS overlap[a], # EIDs (PBC EIDs) | H3K4me1 overlap[a], # EIDs (PBC EIDs) | H3K4me3 overlap[a], # EIDs (PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | (rs2517681=T) or ((rs4148248=C) and (rs2735078=A)) | rs2735078 | 29941400 | intergenic | *HLA-A*(dist=27739), *HCG9*(dist=1492) | 0 (0) | 23 (9) | 10 (2) | NA | 0 | Whole Blood (*DDX39BP2, GABBR1, HCG4P3, HCG4P5, HLA-A, HLA-H, HLA-W, RANP1, ZFP57*); Lymphoblastoid (*HCG4P5, HCG4P7, HLA-A, HLA-H, HLA-K, ZFP57*); Liver (*HCG4B, HCG4P3, HLA-K, ZFP57*) |
| 6 | (rs17195733=G) or ((rs13201129=C) and (rs1737069=T)) | rs17195733 | 30716991 | intergenic | *IER3*(dist=4664), *LINC00243*(dist=63652) | 17 (1) | 96 (11) | 41 (5) | A549 (GR, POL2); HCT-116 (POL24H8); HUVEC (CJUN, GATA2, POL2); HeLa-S3 (POL2); HepG2 (ELF1, FOSL2, FOXA1, FOXA2, HDAC2, HEY1, P300, POL2, RXRA, SP1) | 5 | NA |
| 6 | (rs17195733=G) or ((rs13201129=C) and (rs1737069=T)) | rs13201129 | 30601067 | intronic | *ATAT1* | 0 (0) | 8 (1) | 6 (1) | NA | 1 | Whole Blood (*HLA-E, IFITM4P, MRPS18B, NRM*) |
| 6 | (rs17195733=G) or ((rs13201129=C) and (rs1737069=T)) | rs1737069 | 29730730 | intergenic | *IFITM4P*(dist=11805), *HCG4*(dist=28078) | 3 (0) | 50 (2) | 10 (1) | GM15510 (NFKB) | 3 | Whole Blood (*AL645939.6-3, GABBR1, HCG9, HLA-F, HLA-F-AS1, HLA-G, HLA-H, HLA-J, IFITM4P*); Lymphoblastoid (*HLA-F, HLA-F-AS1, IFITM4P*); Liver (*HLA-A, HLA-F*) |
| 6 | (rs3130785=T) and ((rs9261301=G) or (rs1264570=C)) | rs3130785 | 30796738 | ncRNA_intronic | *LINC00243* | 10 (2) | 121 (29) | 82 (25) | NA | 5 | Whole Blood (*DDR1, FLOT1, HCG9, HLA-H, HLA-J, HLA-L, IER3, VARS2, VARSL*); Lymphoblastoid (*HLA-J*); Liver (*HLA-H*) |
| 6 | (rs3130785=T) and ((rs9261301=G) or (rs1264570=C)) | rs9261301 | 30041559 | intronic | *RNF39* | 0 (0) | 67 (11) | 47 (3) | NA | 2 | Whole Blood (*HCG4P3, HLA-A, HLA-G, HLA-L, HLA-V, PPP1R11, RPL23AP1, ZFP57*); Lymphoblastoid (*RPL23AP1*); Liver (*HLA-V, MICE*) |
| 6 | (rs3130785=T) and ((rs9261301=G) or (rs1264570=C)) | rs1264570 | 30365210 | intergenic | *TRIM39-RPP21*(dist=50575), *HLA-E*(dist=91973) | 1 (1) | 37 (12) | 14 (1) | NA | 2 | Whole Blood (*HLA-E, MRPS18B, RPP21*); Liver (*HCG4B*) |
| 6 | (rs2517681=C) and ((rs3130785=T) or (rs2735078=G)) | rs2517681 | 29932330 | intergenic | *HLA-A*(dist=18669), *HCG9*(dist=10562) | 2 (0) | 122 (29) | 94 (25) | HTB-11 (NRSF) | 1 | Whole Blood (*HCG4P3, HCG4P5, HLA-A, HLA-F, HLA-G, HLA-J, HLA-V, IFITM4P, MICD, PPP1R11, ZFP57, ZNRD1*); Lymphoblastoid (*HCG4P5, IFITM4P, MICE*); Liver (*ZFP57*) |
| 6 | (rs2517681=C) and ((rs3130785=T) or (rs2735078=G)) | rs3130785 | 30796738 | ncRNA_intronic | *LINC00243* | 10 (2) | 121 (29) | 82 (25) | NA | 5 | Whole Blood (*DDR1, FLOT1, HCG9, HLA-H, HLA-J, HLA-L, IER3, VARS2, VARSL*); Lymphoblastoid (*HLA-J*); Liver (*HLA-H*) |
| 6 | (rs2517681=C) and ((rs3130785=T) or (rs2735078=G)) | rs2735078 | 29941400 | intergenic | *HLA-A*(dist=27739), *HCG9*(dist=1492) | 0 (0) | 23 (9) | 10 (2) | NA | 0 | Whole Blood (*DDX39BP2, GABBR1, HCG4P3, HCG4P5, HLA-A, HLA-H, HLA-W, RANP1, ZFP57*); Lymphoblastoid (*HCG4P5, HCG4P7, HLA-A, HLA-H, HLA-K, ZFP57*); Liver (*HCG4B, HCG4P3, HLA-K, ZFP57*) |
| 6 | ((rs1345229=A) or (rs1233387=T)) or (rs1003581=G) | rs1345229 | 30182395 | intergenic | *TRIM26*(dist=1124), *HCG17*(dist=19421) | 35 (10) | 118 (27) | 127 (29) | GM12878 (TBP) | 4 | Whole Blood (*TRIM10, ZNRD1*) |
| 6 | ((rs1345229=A) or (rs1233387=T)) or (rs1003581=G) | rs1233387 | 29555864 | exonic | *OR2H2* | 0 (0) | 1 (0) | 3 (0) | NA | 0 | Whole Blood (*HLA-F, HLA-G, TRIM27*); Liver (*HLA-F*) |

| Chr | 3-SNP Haplotype | SNP | Chr position (hg19) | Ontology | Mapped Gene | DHS overlap[a], # EIDs (PBC EIDs) | H3K4me1 overlap[a], # EIDs (PBC EIDs) | H3K4me3 overlap[a], # EIDs (PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | ((rs1345229=A) or (rs1233387=T)) or (rs1003581=G) | rs1003581 | 29540204 | intergenic | *UBD*(dist=12502), *SNORD32B*(dist=9825) | 0 (0) | 0 (0) | 0 (0) | NA | 3 | Whole Blood (*HLA-F*) |
| 6 | ((rs4713429=G) or (rs12110785=T)) or (rs3131932=G) | rs4713429 | 31021017 | upstream | *HCG22* | 2 (2) | 18 (4) | 11 (4) | GM12878 (EBF1) | 7 | Whole Blood (*VARSL*); Lymphoblastoid (*TCF19*) |
| 6 | ((rs4713429=G) or (rs12110785=T)) or (rs3131932=G) | rs12110785 | 30997824 | exonic | *MUC22* | 0 (0) | 3 (0) | 3 (0) | NA | 1 | NA |
| 6 | ((rs4713429=G) or (rs12110785=T)) or (rs3131932=G) | rs3131932 | 30940328 | intergenic | *DPCR1*(dist=18330), *MUC21*(dist=11157) | 0 (0) | 9 (0) | 5 (0) | NA | 2 | Whole Blood (*CCHCR1, FLOT1, IER3, LINC00243, VARSL*) |
| 6 | ((rs1003581=G) and (rs16894681=T)) or (rs1233387=T) | rs1003581 | 29540204 | intergenic | *UBD*(dist=12502), *SNORD32B*(dist=9825) | 0 (0) | 0 (0) | 0 (0) | NA | 3 | Whole Blood (*HLA-F*) |
| 6 | ((rs1003581=G) and (rs16894681=T)) or (rs1233387=T) | rs16894681 | 29232072 | intergenic | *OR2J2*(dist=89721), *OR14J1*(dist=42395) | 0 (0) | 0 (0) | 0 (0) | NA | 4 | NA |
| 6 | ((rs1003581=G) and (rs16894681=T)) or (rs1233387=T) | rs1233387 | 29555864 | exonic | *OR2H2* | 0 (0) | 1 (0) | 3 (0) | NA | 0 | Whole Blood (*HLA-F, HLA-G, TRIM27*); Liver (*HLA-F*) |
| 6 | ((rs2281043=T) or (rs7751451=G)) and (rs1635=A) | rs2281043 | 28268497 | intronic | *PGBD1* | 0 (0) | 3 (1) | 2 (1) | NA | 3 | Whole Blood (*TRIM27, ZKSCAN3, ZNF193*) |
| 6 | ((rs2281043=T) or (rs7751451=G)) and (rs1635=A) | rs7751451 | 28752883 | intergenic | *ZBED9*(dist=197771), *LINC01623*(dist=74519) | 1 (1) | 6 (0) | 6 (0) | NA | 4 | Whole Blood (*ZFP57*) |
| 6 | ((rs2281043=T) or (rs7751451=G)) and (rs1635=A) | rs1635 | 28227604 | exonic; upstream | *NKAPL; ZKSCAN4* | 0 (0) | 45 (6) | 72 (11) | NA | 0 | NA |
| 6 | ((rs3117192=C) or (rs16894216=T)) and (rs7773193=T) | rs3117192 | 29401416 | intergenic | *OR11A1*(dist=5907), *OR10C1*(dist=6300) | 0 (0) | 0 (0) | 0 (0) | NA | 5 | NA |
| 6 | ((rs3117192=C) or (rs16894216=T)) and (rs7773193=T) | rs16894216 | 28664213 | intergenic | *ZBED9*(dist=109101), *LINC01623*(dist=163189) | 4 (0) | 33 (6) | 47 (1) | NA | 3 | NA |
| 6 | ((rs3117192=C) or (rs16894216=T)) and (rs7773193=T) | rs7773193 | 28611334 | intergenic | *ZBED9*(dist=56222), *LINC01623*(dist=216068) | 31 (4) | 6 (0) | 24 (0) | A549 (USF1); GM12878 (TBP); H1-hESC (TBP); HEK293(b) (KAP1); HeLa-S3 (AP2GAMMA, BRCA1, CEBPB, RFX5, RPC155, STAT1, TBP, TFIIIC110); HepG2 (CEBPB, HSF1, TBP); K562 (RPC155, TBP, TFIIIC110) | 2 | NA |
| 6 | (rs9295704=C) and ((rs2451752=A) and (rs2575174=C)) | rs9295704 | 26704816 | intergenic | *ZNF322*(dist=44836), *GUSBP2*(dist=134450) | 0 (0) | 13 (1) | 8 (1) | NA | 4 | Whole Blood (*ABT1*) |
| 6 | (rs9295704=C) and ((rs2451752=A) and (rs2575174=C)) | rs2451752 | 26648013 | intronic | *ZNF322* | 0 (0) | 1 (1) | 3 (1) | NA | 0 | Whole Blood (*BTN3A1, BTN3A2, HMGN4, ZNF322*) |

| Chr | 3-SNP Haplotype | SNP | Chr position (hg19) | Ontology | Mapped Gene | DHS overlap[a], # EIDs (PBC EIDs) | H3K4me1 overlap[a], # EIDs (PBC EIDs) | H3K4me3 overlap[a], # EIDs (PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | (rs9295704=C) and ((rs2451752=A) and (rs2575174=C)) | rs2575174 | 25885552 | intergenic | *SLC17A3*(dist=11081), *SLC17A2*(dist=27432) | 0 (0) | 5 (0) | 0 (0) | NA | 2 | Whole Blood (*HIST1H1T*\|*HIST1H4A*) |
| 6 | (rs6939576=G) or ((rs2859365=G) or (rs6930033=A)) | rs6939576 | 28669315 | intergenic | *ZBED9*(dist=114203), *LINC01623*(dist=158087) | 0 (0) | 2 (1) | 1 (0) | NA | 0 | NA |
| 6 | (rs6939576=G) or ((rs2859365=G) or (rs6930033=A)) | rs2859365 | 28391465 | intergenic | *ZSCAN12*(dist=23921), *ZSCAN23*(dist=8967) | 0 (0) | 4 (2) | 0 (0) | NA | 3 | Whole Blood (*TRIM27, ZNF187, ZNF193*); Lymphoblastoid (*ZSCAN23*) |
| 6 | (rs6939576=G) or ((rs2859365=G) or (rs6930033=A)) | rs6930033 | 29323905 | exonic | *OR5V1* | 0 (0) | 4 (0) | 1 (0) | NA | 6 | Whole Blood (*TRIM27*) |
| 6 | (rs7773193=C) or ((rs17280818=T) and (rs2394100=T)) | rs7773193 | 28611334 | intergenic | *ZBED9*(dist=56222), *LINC01623*(dist=216068) | 31 (4) | 6 (0) | 24 (0) | A549 (USF1); GM12878 (TBP); H1-hESC (TBP); HEK293(b) (KAP1); HeLa-S3 (AP2GAMMA, BRCA1, CEBPB, RFX5, RPC155, STAT1, TBP, TFIIIC110); HepG2 (CEBPB, HSF1, TBP); K562 (RPC155, TBP, TFIIIC110) | 2 | NA |
| 6 | (rs7773193=C) or ((rs17280818=T) and (rs2394100=T)) | rs17280818 | 28697751 | intergenic | *ZBED9*(dist=142639), *LINC01623*(dist=129651) | 7 (1) | 65 (19) | 100 (21) | NA | 2 | Whole Blood (*ZFP57*) |
| 6 | (rs7773193=C) or ((rs17280818=T) and (rs2394100=T)) | rs2394100 | 28422906 | intergenic | *ZSCAN23*(dist=11627), *GPX6*(dist=48167) | 0 (0) | 2 (0) | 3 (0) | NA | 0 | NA |
| 7 | ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G) | rs12671658 | 7842281 | intronic | *UMAD1* | 1 (0) | 6 (0) | 2 (0) | NA | 7 | NA |
| 7 | ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G) | rs12702656 | 7851742 | intronic | *UMAD1* | 0 (0) | 9 (3) | 0 (0) | NA | 4 | NA |
| 7 | ((rs12671658=T) or (rs12702656=A)) and (rs11768586=G) | rs11768586 | 7849806 | intronic | *UMAD1* | 1 (0) | 6 (2) | 1 (1) | NA | 0 | NA |
| 9 | (rs4979484=C) or ((rs13300483=T) and (rs7028891=G)) | rs4979484 | 117751450 | intergenic | *TNFSF8*(dist=58575), *TNC*(dist=30404) | 22 (6) | 33 (20) | 11 (7) | GM12878 (BATF, NFKB); GM12891 (NFKB); GM15510 (NFKB); GM18951 (NFKB); HeLa-S3 (AP2GAMMA, BAF155, CEBPB, CJUN, GTF2F1, JUND, P300, RAD21, RFX5, STAT3) | 3 | NA |
| 9 | (rs4979484=C) or ((rs13300483=T) and (rs7028891=G)) | rs13300483 | 117643362 | intergenic | *TNFSF15*(dist=74954), *TNFSF8*(dist=12261) | 0 (0) | 11 (6) | 2 (0) | NA | 2 | Whole Blood (*TNFSF8*) |
| 9 | (rs4979484=C) or ((rs13300483=T) and (rs7028891=G)) | rs7028891 | 117645015 | intergenic | *TNFSF15*(dist=76607), *TNFSF8*(dist=10608) | 0 (0) | 4 (3) | 1 (1) | NA | 3 | Whole Blood (*TNFSF8*) |

| Chr | 3-SNP Haplotype | SNP | Chr position (hg19) | Ontology | Mapped Gene | DHS overlap[a], # EIDs (PBC EIDs) | H3K4me1 overlap[a], # EIDs (PBC EIDs) | H3K4me3 overlap[a], # EIDs (PBC EIDs) | Bound protein[b]: Cell line (protein) | # Altered motifs[c] | Significant eQTLs[d]: Tissue (gene) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | (rs7028891=G) and ((rs4979462=T) or (rs10739402=T)) | rs7028891 | 117645015 | intergenic | *TNFSF15*(dist=76607), *TNFSF8*(dist=10608) | 0 (0) | 4 (3) | 1 (1) | NA | 3 | Whole Blood (*TNFSF8*) |
| 9 | (rs7028891=G) and ((rs4979462=T) or (rs10739402=T)) | rs4979462 | 117567013 | intronic | *TNFSF15* | 36 (3) | 79 (8) | 31 (2) | ECC-1 (ERALPHA_A); HUVEC (CFOS, GATA2); HeLa-S3 (ELK4); HepG2 (FOXA1); T-47D (ERALPHA_A, FOXA1, GATA3) | 3 | NA |
| 9 | (rs7028891=G) and ((rs4979462=T) or (rs10739402=T)) | rs10739402 | 116873231 | intergenic | *KIF12*(dist=11894), *COL27A1*(dist=44594) | 1 (1) | 22 (1) | 1 (0) | NA | 3 | NA |
| 9 | (rs7028891=A) and ((rs2418376=A) or (rs10759773=G)) | rs7028891 | 117645015 | intergenic | *TNFSF15*(dist=76607), *TNFSF8*(dist=10608) | 0 (0) | 4 (3) | 1 (1) | NA | 3 | Whole Blood (*TNFSF8*) |
| 9 | (rs7028891=A) and ((rs2418376=A) or (rs10759773=G)) | rs2418376 | 118338852 | intergenic | *DEC1*(dist=173929), *LOC101928775*(dist=163097) | 0 (0) | 0 (0) | 0 (0) | NA | 1 | NA |
| 9 | (rs7028891=A) and ((rs2418376=A) or (rs10759773=G)) | rs10759773 | 118136224 | intronic | *DEC1* | 8 (0) | 34 (1) | 9 (0) | NA | 7 | NA |
| 9 | (rs10817678=G) or ((rs1407306=T) and (rs7048742=A)) | rs10817678 | 117579457 | intergenic | *TNFSF15*(dist=11049), *TNFSF8*(dist=76166) | 0 (0) | 4 (0) | 2 (0) | NA | 11 | Whole Blood (*TNFSF15, TNFSF8*) |
| 9 | (rs10817678=G) or ((rs1407306=T) and (rs7048742=A)) | rs1407306 | 117586409 | intergenic | *TNFSF15*(dist=18001), *TNFSF8*(dist=69214) | 1 (0) | 23 (2) | 3 (0) | NA | 9 | Whole Blood (*TNFSF8*) |
| 9 | (rs10817678=G) or ((rs1407306=T) and (rs7048742=A)) | rs7048742 | 117441568 | intronic | *LOC100505478* | 0 (0) | 33 (2) | 10 (1) | NA | 0 | NA |
| 9 | (rs4979462=T) or ((rs10739402=T) and (rs10817564=C)) | rs4979462 | 117567013 | intronic | *TNFSF15* | 36 (3) | 79 (8) | 31 (2) | ECC-1 (ERALPHA_A); HUVEC (CFOS, GATA2); HeLa-S3 (ELK4); HepG2 (FOXA1); T-47D (ERALPHA_A, FOXA1, GATA3) | 3 | NA |
| 9 | (rs4979462=T) or ((rs10739402=T) and (rs10817564=C)) | rs10739402 | 116873231 | intergenic | *KIF12*(dist=11894), *COL27A1*(dist=44594) | 1 (1) | 22 (1) | 1 (0) | NA | 3 | NA |
| 9 | (rs4979462=T) or ((rs10739402=T) and (rs10817564=C)) | rs10817564 | 116827079 | intronic | *AMBP* | 0 (0) | 2 (2) | 2 (0) | NA | 2 | NA |

[a] Counts of the number of consolidated cell types (EIDs) for which the SNP of interest overlaps the queried epigenomic assay peak (Roadmap Epigenomics Mapping Consortium processed data, Kundaje et al.[20]). "PBC EID": Separately considers peak overlap counts among the 29 blood/liver cell types available in Roadmap Epigenomics.

[b] Bound protein: Regulatory protein-binding ChIP-seq peak overlaps for specified proteins are provided for blood- or liver-related cell lines only (HaploReg v4, Ward and Kellis[32]).

[c] Altered motifs: The number of regulatory motifs predicted to be affected by the SNP based on position weight matrices (PWM) score changes (HaploReg v4, Ward and Kellis[32]).

[d] eQTLs: Reported significant eQTLs for whole blood, lymphoblastoid, and liver cell types only (GTEx Consortium[19]; HaploReg v4, Ward and Kellis[32]).

Abbreviations: EID, epigenome identifier; dist, distance; #, number; DHS, DNase I hypersensitivity site; eQTL, expression quantitative trait loci.

# CHAPTER 5

## Discussion

The genetic architecture provides a complete picture of the genetic basis of a trait (117). Given that elucidating the genetic architecture of human traits is the immediate goal of genetic association studies, in-depth explorations of complex genetic variants – such as epistatic interactions and haplotypes – may offer unique insights. Studies of epistasis and haplotypes may not only identify novel genetic factors that influence trait expression, but are also likely to provide new clues as to the necessary biological conditions for trait expression, thereby contextualizing the functional roles of implicated loci.

The predominant methodology applied in studies of genetic risk factors conducted on a genome-wide scale examines univariate trait associations for all available assayed SNPs without prior contextualization of tested SNPs. In recent years, deliberate investigations of more complex genetic variants have also been undertaken, but most have adopted a "hypothesis-free" approach similar to single-SNP GWAS. Yet, over 80% of the human genome can be annotated with a biochemical regulatory function in at least one human cell type (25). Moreover, while disease-associated SNPs identified in GWAS largely reside outside of non-coding regions, these GWA study "hits" are significantly enriched for functional elements (25). Thus, considering complex genetic variants that potentially model gene regulation events while failing to integrate functional annotations of the human genome under a "hypothesis-free" approach may adversely affect power to detect true associations with disease. The major strength of the analyses presented in this thesis come from the methodological innovations implemented to search for associations between

disease traits and specific classes of complex genetic variants (epistatic interactions and haplotypes) consisting of combinations of SNPs that may collectively influence the regulation of genes.

## 5.1 Overview of main findings

### 5.1.1 Regulatory epistatic SNP interactions influence complex disease traits

Previous studies of epistatic interaction associations and complex disease traits have largely focused on exhaustively testing interactions between pairs of SNPs. While it is computationally feasible to extend this exhaustive testing strategy to investigate higher order epistatic interactions, the use of conservative methods to control Type I error dramatically limits statistical power to detect true higher order interaction associations as the number of participating SNPs increases. In this dissertation, we present a novel methodological approach that effectively bypasses an exhaustive genome-wide search for 3-way interactions to identify 3-SNP interactions that plausibly contribute to gene regulation events associated with a complex trait. Specifically, we implemented a stochastic logic regression-based algorithm among SNPs mapped to regulatory genomic regions (enhancers or promoters) that encourages a broader search of potential epistatic candidates while only expending one degree-of-freedom to test a 3-SNP interaction. To address the potential for reporting false positive results, we developed a permutation-based evaluation statistic to identify candidate 3-SNP interactions, performed a replication analysis, and conducted additional bioinformatics analyses.

Under this methodological framework, we identified six regulatory 3-SNP interactions on chromosomes 2, 12, 13, and 14 that potentially modify cancer treatment effects on BMD among adult survivors of pediatric ALL (N=856), and replicated five of these epistatic interactions as treatment modification effects in an independent sample of adult survivors of non-ALL pediatric cancers (N=1,428). All estimated interaction effects were relatively large (-1.30 to +1.77 SD) in comparison to typical reported single-SNP effect sizes from BMD GWAS (e.g., ~|0.5| SD). Our bioinformatics analyses revealed that SNPs contributing to replicated interactions had both an excess of gene expressions and an enrichment of enhancer states in cell and tissue types important for bone biology in comparison to the entire set of SNPs mapped to enhancer/promoter regions, and that interactions between regulatory regions bearing target SNP variants were plausible.

To assess the performance of our novel logic regression-based algorithm, we conducted two different simulation studies. The first simulation study aimed to investigate the comparative power and positive predictive value of our proposed method overall relative to a benchmark method. Assuming effect sizes observed in our discovery analysis, our proposed method has up to 60% power and 49% PPV to detect "causal" (replicated) 3-SNP interactions in smaller samples, with marked improvements in both statistics with modest increments in sample size. In comparison, a benchmark method that exhaustively tests 2-SNP interactions in order to detect component regulatory SNP pairs in causal 3-SNP interactions was appreciably less powerful and had lower PPV, even with larger sample sizes. We observed no overlap between top 2-way regulatory SNP interactions identified using the benchmark method and the 3-way regulatory SNP interactions detected with our proposed method. The second simulation study evaluated the mean precision of the sequential conditioning component of our proposed statistical algorithm

against a marginal approach to identify causal SNP associations. In our study of epistatic interactions, the sequential conditioning strategy may be advantageous in conducting stochastic searches of the large interaction space since this approach does not remove genetic elements that could contribute to subsequently identified best epistatic interactions, unlike a marginal approach. Our results demonstrated that the sequential conditioning approach had the same mean precision as the marginal approach for detecting causal SNP associations with a quantitative trait in most simulation scenarios. Collectively, these results suggest that exhaustive searches for 2-SNP interactions are not universally effective for detecting higher order epistasis, and sequential conditioning can be useful a tool in a genome-wide association analysis of epistasis.

Our replication results underscore the relative importance of cancer treatment exposures in investigations of regulatory SNP interactions associated with chronic health conditions in adult survivors of pediatric cancer. These results suggest that epistatic networks consisting of three SNPs embedded in regulatory regions that physically interact may modify BMD in pediatric cancer survivors exposed to specific cytotoxic treatments, presumably by jointly affecting gene expressions that influence BMD. For example, our findings suggest that the genomic regulatory region bearing rs1020745 could act in a promoter "hub" for a 3-way SNP interaction on chromosome 12, with rs2110167 and rs10444471 affecting regulatory enhancer elements to influence the *SP7* locus. *SP7* has previously been reported as a candidate gene affecting bone biology in both adult and pediatric populations (84, 85), and is known to encode an osteogenic transcription factor, Osterix (Osx) (86). Exposure to methotrexate has been linked to decreased Osx expression and significant reductions in osteocyte precursors and bone volume in rats (87). As such, this epistatic interaction may counter BMD loss in cancer survivors exposed to methotrexate.

### 5.1.2 Gene-based haplotypes contribute to disease risk

The general strategy implemented in previous genome-wide haplotype association investigations entails combining sliding SNP windows with exhaustive testing to identify haplotypes consisting of proximal SNPs associated with disease risk. While this method has been successful in identifying risk haplotypes for complex disease, exhaustive global association testing of haplotypes assembled within small, agnostically-selected windows has limited power to detect higher order haplotype associations and constrains deeper explorations of haplotypes as *cis*-acting allelic variants that regulate gene transcription. In this dissertation, we proposed a complementary method to detect haplotypes associated with complex traits, with an emphasis on identifying haplotypes consisting of variably-spaced SNPs that also potentially reflect *cis*-regulatory mechanisms for gene expression. Specifically, the proposed methodological approach relies on logic regression to detect 3-SNP haplotypes associated with PBC risk among phased SNP alleles mapped to extended genic regions in Japanese and avoids exhaustive testing, potentially capturing true haplotype associations that would otherwise be missed due to lack of statistical power. To safeguard against false positives, we applied a permutation-based evaluation statistic to select candidate 3-SNP haplotype associations, conducted a replication study for selected haplotypes in a second independent cohort, and performed bioinformatics analyses to assess the biological plausibility of haplotype associations with PBC risk.

Using our proposed haplotype association testing method, a total of 74 gene-based 3-SNP haplotypes associated with PBC risk in chromosomes 6, 7, and 9 were considered as having stronger associations with PBC risk than expected under a permutation-based approach in our

Japanese discovery cohort (N=1,937). Nearly two-thirds of these selected haplotypes (49 haplotypes) were replicated in a second independent Japanese cohort (N=949) under a Bonferroni-corrected p-value threshold ($P<6.8x10^{-4}$). The magnitude of estimated ORs observed for replicated 3-SNP haplotypes in the combined cohort (N=2,886) under the logistic regression model assuming additive haplotype effects ranged from 1.67 to 15.25 (inverting protective associations), with p-values ranging from $1.3x10^{-35}$ to $3.9x10^{-9}$. Upon comparing haplotype associations detected with a benchmark method (exhaustive global association testing of haplotypes formed within all available sliding window sets consisting of three contiguous SNPs in each gene window), we observed both methods identified many of the same gene regions to be important for further haplotype association investigation, but top haplotype associations between methods did not overlap. These findings suggest that our proposed method detects credible haplotype associations that may be missed by conventional analytic methods.

Overall, our bioinformatics analyses suggest that replicated haplotype associations combine the effects of variably-spaced and potentially functional SNPs mapped to regions of the genome that are more likely to be transcribed. The replicated 3-SNP haplotypes detected by logic regression linked SNPs ~335 kb apart on average, with many contributing SNPs overlapping functional annotations in cell/tissue types relevant to PBC biology. In addition, the set of SNPs contributing to replicated haplotype associations were significantly enriched for gene expressions in lymphoblastoid cells and indicators of enhancer, promoter, and open chromatin states in blood and liver cell types in comparison to the set of gene-based SNPs with at least marginal associations with PBC risk.

Replicated 3-SNP haplotype associations revealed both novel PBC susceptibility loci and provided further contextualization for known PBC risk loci. For example, the ((rs12671658=T)

or (rs12702656=A)) and (rs11768586=G) haplotype mapped to the introns of *UMAD1* represents a novel candidate non-HLA PBC susceptibility locus; we hypothesize that this haplotype may contribute to the regulation of biliary epithelial cell proliferation. On the other hand, the (rs4979484=C) or ((rs13300483=T) and (rs7028891=G)) haplotype, comprised of intergenic SNPs near *TNFSF15*, adds to the credibility of previous reports suggesting *TNFSF15* locus as a major genetic susceptibility factor for PBC in Japanese. However, annotation of the SNPs in this haplotype also implicates *TNFSF8* as another possible contributor to PBC risk. Two SNPs in this replicated 3-SNP haplotype are significantly associated with *TNFSF8* expression, while the remaining SNP resides in a genomic region that shows binding affinity for the NF-kB transcription factor in lymphoblastoid cells obtained from Japanese individuals. Interestingly, *TNFSF8* is a known activator of NF-kB, and may play a role in inflammation and immunity pathways.

## 5.2 Strengths and limitations

The major methodological innovations that we proposed in this thesis to study biologically meaningful epistatic interaction and haplotype associations may be summarized as follows: (1) we first applied a "biological filter" to restrict the set of investigated SNPs; (2) we then employed a novel, non-exhaustive statistical approach to identify epistatic interaction and haplotype associations with traits among filtered SNPs; and (3) we conducted replication and biological inference analyses to assess the credibility of our findings. Our proposed methodological approach has clear strengths. By preliminarily filtering the set of assayed SNPs based on biological functions, we increase the prior probability of identifying epistatic

interactions or haplotypes that are transcriptionally relevant. In addition, combining a biological filter with a non-exhaustive testing method enables searches of the large space of interactions/haplotypes without limiting the number of tested associations using computational burden-based criteria to improve power, i.e., restricting investigated SNPs to those with marginal associations with the trait, or to SNPs that are contiguous.

For each of the genetic association studies presented in this thesis, we applied a biological filter that was appropriate for the research questions of interest. For example, we employed ChromHMM chromatin state annotations (27) to map SNPs to putative enhancer or promoter regions in order to detect SNP combinations contributing to enhancer-promoter interactions that potentially affect gene regulation events that are relevant for bone biology. To study transcriptionally-relevant haplotype associations with PBC risk, we examined haplotypes consisting of SNPs mapped to gene transcripts annotated by the RefSeq reference gene set and flanking 500-kb regions that correspond to a broader definition of "gene". However, the use of any bioinformatics-based biological filter introduces a potential source of measurement error. For example, there are multiple bioinformatics resources that may used to map SNPs to putative enhancers or promoters (e.g., experimental histone modification data instead of ChromHMM estimates); each resource would generate slightly different SNP annotations. However, ChromHMM annotations have strong internal validity: (1) the model is trained on multiple ChIP-seq generated marks (e.g., a pattern of histone modification mark peaks is used to label a region rather than a single peak); and (2) validation studies of ChromHMM-annotated regions have shown these regions bear biological characteristics that are consistent with their annotation classification. Similarly, different reference gene sets yield minor differences in SNP-gene annotations; also, the set of known protein-coding genes is expected to expand and change over

time (25). Our choice of examining expanded gene windows in our haplotype association analysis better accommodates current and anticipated differences between reference gene sets, and is consistent with recent observations that nearly 75% of the human genome contributes to either processed or primary transcripts (118).

In our analyses, the discovery of signals entailed the use of a logic regression-based stochastic search and permutation-based inference to detect epistatic interaction and haplotype associations consisting of biologically-filtered SNPs. The primary strength of logic regression in this research context is that it considers multiple models of risk that are not traditionally assessed in interaction/haplotype association studies (e.g., presence of risk alleles at *either* of two loci) that can better reflect regulatory redundancies that may exist in SNP networks that control transcription. A known limitation of stochastic searches, however, is that finding globally optimal solutions are not assured. An alternative perspective of this limitation is that exhaustive testing may find the "best" association signal, yet miss many true association signals under conservative methods to control false discovery rates. To that end, our proposed method utilizes permutation-based inference to select top association signals instead of relying strictly on Bonferroni-corrected p-values; our method therefore captures association signals that would be missed under an exhaustive testing strategy. With evidence from replication studies, we are also able to assess the credibility of candidate association signals.

Other general limitations of the analyses presented in this thesis include our choice to conduct studies of epistasis and haplotypes with two different traits. Investigating both types of genetic variants for a single phenotype could offer further biological insights. Moreover, imputation of whole genomes may provide richer contextualization for our association analyses, but would be accompanied by heavy computational costs. A significant limitation of our analyses

is the lack of a comprehensive study of epistatic interaction and haplotype variants consisting of larger networks of SNPs. While the study of 3-SNP interactions and haplotypes represent an incremental improvement over previous studies of pairs of SNPs or SNP alleles, true SNP interaction and haplotype associations that incorporate more than three SNPs may exist. Another important limitation of these studies is that we have not functionally validated any of the replicated epistatic interaction and haplotype association signals. The overall functions of discovered interactions or haplotypes are challenging to interpret, since each discovered variant plausibly reflects multiple phenomena (e.g., defects in physical interactions between regulatory genomic regions, DNA-binding proteins and regulatory genomic regions, or among multiple DNA-binding proteins). We note that our use of bioinformatics analyses provide supportive evidence for the biological plausibility of associations between discovered variants and disease traits, and may generate viable leads for functional studies in the future. Lastly, the presence of random and systematic errors in phenotype/genotype measurement cannot be ruled out; unmeasured confounders or residual confounding may also distort the magnitude of association signals. Analyses included in this thesis minimize the impact of such biases with the use of clinically-assessed phenotypes and non-genetic data, and extensive quality control procedures to process genotype data.

## 5.3 Conclusions and clinical/public health implications

In this dissertation, we proposed a novel methodological framework motivated by biological phenomena, specifically the regulation of gene transcription, to investigate the associations between specific classes of complex genetic variants and disease traits. Under this

framework, we explored associations between bone mineral density and epistatic interactions between SNPs in enhancer and promoter elements in adult survivors of pediatric cancer, and gene-based haplotype associations with primary biliary cholangitis risk in Japanese. Our analyses revealed that our proposed framework not only successfully identifies credible associations between disease traits and these types of complex genetic variants, but in doing so, we may also gain insights into foundational biomolecular mechanisms that underpin disease pathogenesis. In summary, these findings strongly suggest that we: (1) contribute to future knowledge production in genetic and molecular epidemiology that incorporates biological theory; and (2) continue to explore novel approaches to conduct deliberate searches for complex genetic variants associated with human disease traits. For example, one potential line of future research to consider includes the study of combinations of defects in either epistatic or phase-dependent contexts in regulatory regions due to common SNPs, presumably with neutral or positive effects on fitness for disease traits, and rare single nucleotide variants, which are anticipated to have negative effects on fitness for traits.

While methodological development is necessary to advance the study of genetic susceptibility factors, the overarching goal of genetic association studies is to contribute to the growing body of public health knowledge that may be translated to reduce the burden of morbidity and mortality attributable to complex disease among genetically susceptible individuals. However, with few exceptions, the only appropriate translational research endpoint for SNP discoveries GWAS is knowledge generation (119). In the studies presented in this thesis, we have demonstrated that explorations of epistasis and haplotypes have the potential to provide greater mechanistic insight into the genetic architecture of complex traits; as such, our study findings may help identify novel etiological mechanisms behind treatment-related bone

loss in adult survivors of pediatric cancer, or primary biliary cholangitis in Japanese. As such, we suggest that the discovered genetic targets discussed in this thesis be considered for future basic research into biological mechanisms influencing bone mineral density and primary biliary cholangitis to support the eventual objective of developing potential health applications for the prevention, diagnosis, or treatment of disease.

**REFERENCES**


1. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *American Journal of Human Genetics*. 2012; **90**(1): 7-24.

2. Manolio TA, Collins FS, Cox NJ*, et al.* Finding the missing heritability of complex diseases. *Nature*. 2009; **461**(7265): 747-53.

3. Gibson G. Hints of hidden heritability in GWAS. *Nature Genetics*. 2010; **42**(7): 558-60.

4. Yang J, Benyamin B, McEvoy BP*, et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*. 2010; **42**(7): 565-9.

5. Eichler EE, Flint J, Gibson G*, et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*. 2010; **11**(6): 446-50.

6. Gibson G. Rare and common variants: twenty arguments. *Nature Reviews Genetics*. 2012; **13**(2): 135-45.

7. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*. 2012; **109**(4): 1193-8.

8. Hemani G, Knott S, Haley C. An evolutionary perspective on epistasis and the missing heritability. *PLoS genetics*. 2013; **9**(2): e1003295.

9. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease–common variant… or not? *Human Molecular Genetics*. 2002; **11**(20): 2417-23.

10. Kumaran M, Cass CE, Graham K*, et al.* Germline copy number variations are associated with breast cancer risk and prognosis. *Scientific Reports*. 2017; **7**(1): 14621.

11. Levine M, Cattoglio C, Tjian R. Looping back to leap forward: transcription enters a new era. *Cell*. 2014; **157**(1): 13-25.

12. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*. 2008; **9**(11): 855-67.

13. Carlborg Ö, Haley CS. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*. 2004; **5**(8): 618-25.

14. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*. 2003; **56**(1-3): 73-82.

15. Waddington CH. Canalization of development and the inheritance of acquired characters. *Nature*. 1942; **150**(3811): 563-5.

16. Dixon SJ, Costanzo M, Baryshnikova A, Andrews B, Boone C. Systematic mapping of genetic interaction networks. *Annual Review of Genetics*. 2009; **43**: 601-25.

17. Tyler AL, Asselbergs FW, Williams SM, Moore JH. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays*. 2009; **31**(2): 220-7.

18. Gibson G. Decanalization and the origin of complex disease. *Nature Reviews Genetics*. 2009; **10**(2): 134-40.

19. Liu N, Zhang K, Zhao H. Haplotype-association analysis. *Advances in Genetics*. 2008; **60**: 335-405.

20. Sabeti PC, Reich DE, Higgins JM*, et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 2002; **419**(6909): 832-7.

21. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*. 2011; **12**(10): 703-14.

22. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ. The importance of phase information for human genomics. *Nature Reviews Genetics*. 2011; **12**(3): 215.

23. Pastinen T, Hudson TJ. Cis-acting regulatory variation in the human genome. *Science*. 2004; **306**(5696): 647-50.

24. Chen H, Wilkins LM, Aziz N*, et al.* Single nucleotide polymorphisms in the human interleukin-1B gene affect transcription according to haplotype context. *Human Molecular Genetics*. 2006; **15**(4): 519-29.

25. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; **489**(7414): 57-74.

26. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*. 2003; **33**: 245-54.

27. 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.

28. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*. 2013; **14**(6): 390-403.

29. Smemo S, Tena JJ, Kim K-H*, et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*. 2014; **507**(7492): 371-5.

30. Harismendy O, Notani D, Song X*, et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature*. 2011; **470**(7333): 264-8.

31. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. *Nature*. 2009; **461**(7261): 199-205.

32. Frayling TM, Timpson NJ, Weedon MN*, et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*. 2007; **316**(5826): 889-94.

33. Dina C, Meyre D, Gallina S*, et al.* Variation in FTO contributes to childhood obesity and severe adult obesity. *Nature Genetics*. 2007; **39**(6): 724-6.

34. Wei W-H, Hemani G, Haley CS. Detecting epistasis in human complex traits. *Nature Reviews Genetics*. 2014; **15**(11): 722-33.

35. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *American Journal of Human Genetics*. 2010; **86**(1): 6-22.

36. Cordell HJ. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*. 2009; **10**(6): 392-404.

37. Cockerham CC, Zeng Z-B. Design III with marker loci. *Genetics*. 1996; **143**(3): 1437-56.

38. Hemani G, Shakhbazov K, Westra HJ*, et al.* Detection and replication of epistasis influencing transcription in humans. *Nature*. 2014; **508**(7495): 249-53.

39. Emily M, Mailund T, Hein J, Schauser L, Schierup MH. Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*. 2009; **17**(10): 1231-40.

40. Fish AE, Capra JA, Bush WS. Are Interactions between cis-Regulatory Variants Evidence for Biological Epistasis or Statistical Artifacts? *American Journal of Human Genetics*. 2016; **99**(4): 817-30.

41. Rothman KJ. Epidemiology: an introduction: Oxford university press; 2012.

42. Mackay TF. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics*. 2014; **15**(1): 22-33.

43. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*. 2002; **11**(20): 2463-8.

44. Li W, Reich J. A complete enumeration and classification of two-locus disease models. *Human Heredity*. 2000; **50**(6): 334-49.

45. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *The American Journal of Human Genetics*. 2002; **70**(2): 425-34.

46. Tregouet DA, Konig IR, Erdmann J*, et al.* Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nature Genetics*. 2009; **41**(3): 283-5.

47. Gurney JG, Kaste SC, Liu W*, et al.* Bone mineral density among long-term survivors of childhood acute lymphoblastic leukemia: results from the St. Jude Lifetime Cohort Study. *Pediatric Blood & Cancer*. 2014; **61**(7): 1270-6.

48. Kaste SC, Rai SN, Fleming K*, et al.* Changes in bone mineral density in survivors of childhood acute lymphoblastic leukemia. *Pediatric Blood & Cancer*. 2006; **46**(1): 77-87.

49. Wasilewski-Masker K, Kaste SC, Hudson MM, Esiashvili N, Mattano LA, Meacham LR. Bone mineral density deficits in survivors of childhood cancer: long-term follow-up guidelines and review of the literature. *Pediatrics*. 2008; **121**(3): e705-13.

50. Davies JH, Evans BA, Jenney ME, Gregory JW. Skeletal morbidity in childhood acute lymphoblastic leukaemia. *Clinical Endocrinology*. 2005; **63**(1): 1-9.

51. Peacock M, Turner CH, Econs MJ, Foroud T. Genetics of osteoporosis. *Endocrine Reviews*. 2002; **23**(3): 303-26.

52. Selmi C, Mayo MJ, Bach N*, et al.* Primary biliary cirrhosis in monozygotic and dizygotic twins: genetics, epigenetics, and environment. *Gastroenterology*. 2004; **127**(2): 485-92.

53. Nakamura M, Nishida N, Kawashima M*, et al.* Genome-wide association study identifies TNFSF15 and POU2AF1 as susceptibility loci for primary biliary cirrhosis in the Japanese population. *American Journal of Human Genetics*. 2012; **91**(4): 721-8.

54. Kawashima M, Hitomi Y, Aiba Y*, et al.* Genome-wide association studies identify PRKCB as a novel genetic susceptibility locus for primary biliary cholangitis in the Japanese population. *Human Molecular Genetics*. 2017; **26**(3): 650-9.

55. Hudson MM, Ness KK, Gurney JG*, et al.* Clinical ascertainment of health outcomes among adults treated for childhood cancer. *JAMA*. 2013; **309**(22): 2371-81.

56. Ojha RP, Oancea SC, Ness KK*, et al.* Assessment of potential bias from non-participation in a dynamic clinical cohort of long-term childhood cancer survivors: results from the St. Jude Lifetime Cohort Study. *Pediatric Blood & Cancer*. 2013; **60**(5): 856-64.

57. Phillips SM, Padgett LS, Leisenring WM*, et al.* Survivors of childhood cancer in the United States: prevalence and burden of morbidity. *Cancer Epidemiology and Prevention Biomarkers*. 2015; **24**(4): 653-63.

58. Bhakta N, Liu Q, Ness KK*, et al.* The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE). *The Lancet*. 2017; **390**(10112): 2569-82.

59. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*. 2007; **81**(5): 1084-97.

60. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*. 2013; **10**(1): 5-6.

61. Aken BL, Achuthan P, Akanni W*, et al.* Ensembl 2017. *Nucleic Acids Research*. 2016: gkw1104.

62. Kundaje A, Meuleman W, Ernst J*, et al.* Integrative analysis of 111 reference human epigenomes. *Nature*. 2015; **518**(7539): 317-30.

63. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*. 2015; **348**(6235): 648-60.

64. Zhou X, Lowdon RF, Li D*, et al.* Exploring long-range genome interactions using the WashU Epigenome Browser. *Nature methods*. 2013; **10**(5): 375-6.

65. O'Leary NA, Wright MW, Brister JR*, et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. 2016; **44**(D1): D733-45.

66. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. 2010; **38**(16): e164-e.

67. Tao H, Cox DR, Frazer KA. Allele-specific KRT1 expression is a complex trait. *PLoS Genet*. 2006; **2**(6): e93.

68. Rockman MV, Kruglyak L. Genetics of global gene expression. *Nature Reviews Genetics*. 2006; **7**(11): 862.

69. Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. *Journal of Computational and Graphical Statistics*. 2003; **12**(3): 475-511.

70. Sharafeldin N, Slattery ML, Liu Q*, et al.* A candidate-pathway approach to identify gene-environment interactions: analyses of colon cancer risk and survival. *Journal of the National Cancer Institute*. 2015; **107**(9): djv160.

71. Dinu I, Mahasirimongkol S, Liu Q*, et al.* SNP-SNP interactions discovered by logic regression explain Crohn's disease genetics. *PloS One*. 2012; **7**(10): e43035.

72. Suehiro Y, Wong CW, Chirieac LR*, et al.* Epigenetic-genetic interactions in the APC/WNT, RAS/RAF, and P53 pathways in colorectal carcinoma. *Clinical Cancer Research*. 2008; **14**(9): 2560-9.

73. Justenhoven C, Hamann U, Schubert F*, et al.* Breast cancer: a candidate gene approach across the estrogen metabolic pathway. *Breast Cancer Research and Treatment*. 2008; **108**(1): 137-49.

74. Lieberman-Aiden E, Van Berkum NL, Williams L*, et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009; **326**(5950): 289-93.

75. Kumar J, Swanberg M, McGuigan F, Callreus M, Gerdhem P, Åkesson K. LRP4 association to bone properties and fracture and interaction with genes in the Wnt-and BMP signaling pathways. *Bone*. 2011; **49**(3): 343-8.

76. Yang TL, Guo Y, Li J*, et al.* Gene-gene interaction between RBMS3 and ZNF516 influences bone mineral density. *Journal of Bone and Mineral Research*. 2013; **28**(4): 828-37.

77. Seeman E. Pathogenesis of bone fragility in women and men. *The Lancet*. 2002; **359**(9320): 1841-50.

78. Takayanagi H. Osteoimmunology: shared mechanisms and crosstalk between the immune and bone systems. *Nature Reviews: Immunology*. 2007; **7**(4): 292.

79. Wei W-H, Hemani G, Gyenesei A*, et al.* Genome-wide analysis of epistasis in body mass index using multiple human populations. *European Journal of Human Genetics*. 2012; **20**(8): 857-62.

80. Hudson MM, Ness KK, Nolan VG*, et al.* Prospective medical assessment of adults surviving childhood cancer: study design, cohort characteristics, and feasibility of the St. Jude Lifetime Cohort study. *Pediatric Blood & Cancer*. 2011; **56**(5): 825-36.

81. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000; **155**(2): 945-59.

82. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*. 2001; **98**(9): 5116-21.

83. Zeller T, Wild P, Szymczak S*, et al.* Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PloS One*. 2010; **5**(5): e10693.

84. Timpson NJ, Tobias JH, Richards JB*, et al.* Common variants in the region around Osterix are associated with bone mineral density and growth in childhood. *Human Molecular Genetics*. 2009; **18**(8): 1510-7.

85. Estrada K, Styrkarsdottir U, Evangelou E*, et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature Genetics*. 2012; **44**(5): 491-501.

86. Nakashima K, Zhou X, Kunkel G*, et al.* The novel zinc finger-containing transcription factor osterix is required for osteoblast differentiation and bone formation. *Cell*. 2002; **108**(1): 17-29.

87. Georgiou KR, Scherer MA, Fan CM*, et al.* Methotrexate chemotherapy reduces osteogenesis but increases adipogenic potential in the bone marrow. *Journal of Cellular Physiology*. 2012; **227**(3): 909-18.

88. Hyttinen V, Kaprio J, Kinnunen L, Koskenvuo M, Tuomilehto J. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. *Diabetes*. 2003; **52**(4): 1052-5.

89. Flicek P, Amode MR, Barrell D*, et al.* Ensembl 2014. *Nucleic Acids Research*. 2013; **42**(D1): D749-D55.

90. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. 2000; **28**(1): 27-30.

91. Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research*. 2014; **43**(D1): D1049-D56.

92. Welter D, MacArthur J, Morales J*, et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*. 2013; **42**(D1): D1001-D6.

93. Carey EJ, Ali AH, Lindor KD. Primary biliary cirrhosis. *The Lancet*. 2015; **386**(10003): 1565-75.

94. Kaplan MM, Gershwin ME. Primary biliary cirrhosis. *New England Journal of Medicine*. 2005; **353**(12): 1261-73.

95. Cordell HJ, Han Y, Mells GF*, et al.* International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nature communications*. 2015; **6**: 8019.

96. Mells GF, Floyd JA, Morley KI*, et al.* Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nature Genetics*. 2011; **43**(4): 329-32.

97. Liu X, Invernizzi P, Lu Y*, et al.* Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. *Nature Genetics*. 2010; **42**(8): 658-60.

98. Hirschfield GM, Liu X, Xu C*, et al.* Primary biliary cirrhosis associated with HLA, IL12A, and IL12RB2 variants. *New England Journal of Medicine*. 2009; **360**(24): 2544-55.

99. Im C, Ness KK, Kaste SC*, et al.* Genome-wide search for higher order epistasis as modifiers of treatment effects on bone mineral density in childhood cancer survivors. *European Journal of Human Genetics*. 2018; **26**: 275-86.

100. Nakamura M. Genetic Factors in the Pathogenesis of Primary Biliary Cirrhosis. Autoimmune Liver Diseases: Springer; 2014. p. 157-69.

101. Jones DE. Pathogenesis of primary biliary cirrhosis. *Postgraduate Medical Journal*. 2008; **84**(987): 23-33.

102. Gaur U, Aggarwal BB. Regulation of proliferation, survival and apoptosis by members of the TNF superfamily. *Biochemical Pharmacology*. 2003; **66**(8): 1403-8.

103. Lin DY, Huang BE. The use of inferred haplotypes in downstream analyses. *American Journal of Human Genetics*. 2007; **80**(3): 577-9.

104. Invernizzi P, Selmi C, Gershwin ME. Update on primary biliary cirrhosis. *Digestive and Liver Disease*. 2010; **42**(6): 401-8.

105. Webb G, Siminovitch K, Hirschfield G. The immunogenetics of primary biliary cirrhosis: a comprehensive review. *Journal of Autoimmunity*. 2015; **64**: 42-52.

106. Invernizzi P, Miozzo M, Battezzati PM*, et al.* Frequency of monosomy X in women with primary biliary cirrhosis. *The Lancet*. 2004; **363**(9408): 533-5.

107. Özbalkan Z, Bağışlar S, Kiraz S*, et al.* Skewed X chromosome inactivation in blood cells of women with scleroderma. *Arthritis & Rheumatology*. 2005; **52**(5): 1564-70.

108. Brix TH, Knudsen GPS, Kristiansen M, Kyvik KO, Ørstavik KH, Hegedüs L. High frequency of skewed X-chromosome inactivation in females with autoimmune thyroid disease: a possible explanation for the female predisposition to thyroid autoimmunity. *The Journal of Clinical Endocrinology & Metabolism*. 2005; **90**(11): 5949-53.

109. Miozzo M, Selmi C, Gentilin B*, et al.* Preferential X chromosome loss but random inactivation characterize primary biliary cirrhosis. *Hepatology*. 2007; **46**(2): 456-62.

110. Lleo A, Oertelt-Prigione S, Bianchi I*, et al.* Y chromosome loss in male patients with primary biliary cirrhosis. *Journal of Autoimmunity*. 2013; **41**: 87-91.

111. Kaplan RC, Petersen AK, Chen MH*, et al.* A genome-wide association study identifies novel loci associated with circulating IGF-I and IGFBP-3. *Human Molecular Genetics*. 2011; **20**(6): 1241-51.

112. Alvaro D, Metalli VD, Alpini G*, et al.* The intrahepatic biliary epithelium is a target of the growth hormone/insulin-like growth factor 1 axis. *Journal of Hepatology*. 2005; **43**(5): 875-83.

113. Turnbull C, Ahmed S, Morrison J*, et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature Genetics*. 2010; **42**(6): 504-7.

114. Cargill M, Schrodi SJ, Chang M*, et al.* A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes. *American Journal of Human Genetics*. 2007; **80**(2): 273-90.

115. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*. 2012; **40**(Database issue): D930-4.

116. Hahne F, Ivanek R. Visualizing genomic data using Gviz and Bioconductor. *Statistical Genomics: Methods and Protocols*. 2016: 335-51.

117. Mackay TF. The genetic architecture of quantitative traits. *Annual Review of Genetics*. 2001; **35**(1): 303-39.

118. Djebali S, Davis CA, Merkel A*, et al.* Landscape of transcription in human cells. *Nature*. 2012; **489**(7414): 101.

119. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genetics in Medicine*. 2007; **9**(10): 665-74.