IDENTIFYING GENETIC MECHANISMS OF CARDIOMETABOLIC TRAITS AND DISEASES USING QUANTITATIVE SEQUENCE DATA

Martin L. Buchkovich

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

Chapel Hill
2015

Approved by:

Praveen Sethupathy

Yun Li

Gregory E. Crawford

Terrence S. Furey

Karen L. Mohlke

ABSTRACT

Martin L. Buchkovich: Identifying genetic mechanisms of cardiometabolic traits and diseases using quantitative sequence data
(Under the direction of Karen L. Mohlke and Terrence S. Furey)

Cardiometabolic diseases are a worldwide health concern. Genetics studies have identified hundreds of genetic loci associated with these diseases and other cardiometabolic risk factors, but gaps remain in the understanding of the biological mechanisms responsible for these associations. Sequence data from quantitative experiments, such as DNase-seq and ChIP-seq, that identify genomic regions regulating gene transcription are helping to fill these gaps. Allelic imbalance at heterozygous sites, or enrichment of one allele, in this data can indicate allelic differences in transcriptional regulation, but reference mapping biases present in sequence alignments prevent accurate allelic imbalance detection.

We describe a pipeline, AA-ALIGNER, that removes mapping biases at heterozygous sites and increases allelic imbalance detection accuracy in samples with any amount of genotype data available. When complete genotype information is not available, AA-ALIGNER more accurately detects allelic imbalance at imputed heterozygous sites than heterozygous sites predicted using the sequence data. At predicted heterozygous sites, imbalance detection is more accurate at common variants than other variants. Additionally, imbalance detection with AA-ALIGNER is robust to a variety of experimental and analytical parameters.

Using AA-ALIGNER, we detected evidence of allelic imbalance at 22,414 heterozygous sites in data from samples with relevance to cardiometabolic disease and risk factors. We have identified protein binding motifs for one of the imbalanced proteins at a majority of these sites, and evidence that imbalance in data for this protein is associated with imbalance in data for other proteins. Additionally, a subset of sites of allelic imbalance are located at expression quantitative trait loci and/or genome-wide association loci for cardiometabolic traits and diseases. These sites are strong candidates to be studied experimentally and we report experimental evidence of allelic differences in protein binding, enhancer activity and/or the regulation of specific genes for a handful of these sites.

Using allelic imbalance detection, we have detected differences in protein binding across the genome providing valuable insight into mechanisms of transcriptional regulation. Focusing on cardiometabolic diseases and risk factors, this work demonstrates the utility of allelic imbalance detection in studying genetic effects on the regulation of gene transcription at complex disease- and trait-associated loci.

To my family, friends, and colleagues who have fostered my love of genetics and made this work possible.

ACKNOWLEDGEMENTS

This work is the direct result of a journey that started well before I enrolled in graduate school and subsequently became a Ph.D. candidate. Successful completion of this journey would not be possible without the aid, advice, encouragement and support of many individuals along the way. I value the friendship of all that I have encountered on this journey and take a moment here to acknowledge those who have hand a direct hand in the completion of this work.

While a paragraph is inadequate to fully acknowledge their contributions to this work, I would first like to thank and acknowledge my mentors Karen Mohlke and Terry Furey. They provided a positive environment that allowed me to engage in exciting research and fostered my growth as an individual and scientist. I would especially like to thank Karen for providing me with an abundance of opportunities to deepen my understanding of human genetics and complex phenotypes through interactions with collaborators and participation in their research. I am grateful to Terry for broadening my understanding of genomics and computational biology, by continually encouraging me to delve deep into biological questions being asked and find the most appropriate solution for answering these questions.

I would also like to acknowledge the contributions of my other committee members, Praveen Sethupathy, Yun Li, and Greg Crawford and thank them for their support and encouragement. Whether it was by advice given from miles away, during unannounced visits to their offices, or through participation in their courses, their advice, ideas, and knowledge were critical to this work.

As a member of two labs, I have been privileged to interact with many individuals, who are too many to name. I am especially thankful to members of the Mohlke and Furey labs, both past and present, as I don’t know where I would be without their willingness to provide feedback

on my work and share ideas. I would also like to acknowledge the many collaborators I have worked with who have expanded my horizons and pushed me to be a better researcher, as well as all others I have interacted with during my time as a graduate student. I will always be grateful for the footprints that my scientific mentors and colleagues have made in my life and will values the friendship we have made.

This work would not have been possible without the support and encouragement of many individuals outside of science. I am especially grateful for members of my family. I would specifically like to acknowledge my aunt and uncle, Karen Buchkovich-Sass and Phil Sass, who have been very supportive of my scientific pursuits and facilitated my first extended exposure to molecular biology experiments, and Nicholas Buchkovich, the first member of my immediate family to receive a Ph.D. I thank my wife's parents, Paul and Sally Baird, and would like to publically acknowledge them for the many hours of service given and miles driven to support my family as I have pursued my degree. Likewise, words cannot express enough my thanks to my parents Terry and Joan Buchkovich, for their continuous support and encouragement. They have given sacrificed much to foster my scientific career and I am sure that they would have sacrificed more had there been fewer miles between us.

Finally and most importantly, I acknowledge my wife, Jennie Buchkovich, for her love, encouragement and support. Between working the night shift and giving up her weekends to support our family to acting as a single mother when deadlines neared, her sacrifices have been great. Without them, this work would not have been possible. I cannot thank her enough and will always be grateful to have her at my side. I am also grateful for my sons Kelten and Caleb, for providing me with greater motivation to succeed, than anything else.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AA-ALIGNER	Allele-Aware Aligner for Investigating GeNetic Effects on Regulation
bp	base pairs
ChIP	Chromatin Immunoprecipitation
EMSA	Electrophoretic mobility shift assay
ENA	European Nucleotide Archive
ENCODE	ENCyclopedia of DNA Elements
eQTL	expression quantitative trait loci
FAIRE	Formaldehdye Assisted Isolation of Regulatory Elements
GIANT	Genetic Investigation of ANthropometric Traits
GLGC	Global Lipid Genetics Consortium
GWAS	genome-wide association study
HDL-C	High-density lipoprotein cholesterol
LD	Linkage Disequilibrium
LDL-C	Low-density lipoprotein cholesterol
SRA	Sequence Nucleotide Archive
T2D	Type 2 diabetes
TBP	TATA binding protein
TC	Total cholesterol
TG	Triglycerides

CHAPTER 1: INTRODUCTION

1.1 Introduction

Complex, cardiometabolic diseases (i.e. type 2 diabetes and coronary artery disease), and related traits (i.e. lipid levels) present a major public health concern worldwide[1]. These cardiometabolic phenotypes are influenced by both genetic and environmental factors. Understanding how genetic factors influence biological mechanisms contributing to phenotypes, such as gene transcription, is the first step in identifying novel, personalized treatment. While hundreds of genomic locations, or loci, have been associated with cardiometabolic phenotypes, the biological mechanisms responsible for these associations are understood for only a small proportion of loci. Analyzing short sequence reads from next generation sequencing technologies is a powerful method of studying the influence of genetic factors on biological mechanisms. These analyses can identify genetic variants influencing gene transcription and provide an important first step in bridging the gap between genetic variation and cardiometabolic phenotypes.

1.2 Genetic variation contributes to cardiometabolic traits and diseases

Genome-wide association studies (GWAS) are an active area of research and effective tool for studying genetic factors influencing cardiometabolic traits and diseases. These studies find statistical associations between genetics variants and the presence of disease or differences in trait measurements[2]. To date, GWAS have identified over 75 genetic loci significantly associated with type 2 diabetes[3], 46 loci associated with coronary artery disease[4], and over three hundred more loci associated with other cardiometabolic traits, such as lipid levels (HDL-C, LDL-C, triglycerides, and total cholesterol)[5,6], adiponectin levels[7–13], and obesity (body mass index, waist-to-hip ratio, and body fat percentage)[14–16]. Large consortia, such as the Genetic Investigation of Anthropometric Traits (GIANT) consortium and the Global Lipids Genetics Consortium (GLGC) are actively identifying additional variants associated with metabolic traits. In addition, the Metabolic Syndrome in Men (METSIM) study is identifying variants associated with detailed phenotypic traits, including ~200 traits related to diabetes status, including measures of lipids, lipoprotein particles, glucose, insulin, proinsulin, free fatty acids, body composition, cytokines, hormones, and metabolites, most collected for 10,000 subjects[17].

1.3 Genetic variants at GWAS loci likely influence gene transcription regulation

While the variants identified by these studies can influence phenotypes by altering protein structure[3] or influencing post transcriptional gene regulation[18], a majority of variants (90%) are located in non-coding regions at GWAS loci and likely influence gene transcription regulation[19]. Identifying genes differentially regulated by these variants is critical for understanding the biological pathways influenced by these variants, and their role in disease susceptibility.

Variants at GWAS loci have also been associated with the expression of nearby genes. Expression quantitative trait loci (eQTLs) are genetic variants that are associated with gene transcript levels[20]. Similar to GWAS, eQTL analyses utilize genotype data from individuals in a population, but test for associations between these genotypes and gene transcript levels rather than diseases and traits. As many as 12% of variants reported in the NHGRI GWAS catalog[21] are either the variant reported at an eQTL locus, or in linkage disequilibrium (LD) with the reported variant[22]. This co-occurrence of GWAS and eQTL associations provides additional evidence that at these loci, and possibly others, differential regulation of gene transcript levels may be an intermediary between genetic variants and observed differences in phenotype.

Genetic variation can influence the regulation of gene transcription by altering the DNA binding affinity of proteins important to transcription. DNA-binding proteins often bind preferentially to a specific DNA sequence, or binding motif. Sequence changes caused by genetic variation may result in changes in protein-DNA binding affinity and consequently alter the amount of DNA-bound protein. Altered binding of transcription factors can directly change the amount of gene transcription[23]. In other cases, differential binding of proteins, such as proteins from FOXA family, can indirectly alter gene transcription by influencing chromatin accessibility[24], defined as accessibility of DNA for binding by transcriptional machinery, including transcription factors[25]. Likewise, other proteins, such as CTCF, are important in chromatin looping which controls gene transcription by bringing regulatory regions in close proximity to each other to promote gene transcription, or by marking the boundaries between active and inactive regulatory regions[26]. Understanding the influence of genetic variants on transcription binding, chromatin accessibility and chromatin looping is important for understanding their overall influence on gene transcription.

While genome-wide association and eQTL studies identify genetic regions associated with phenotypes and gene expression, it is difficult, using only these studies, to identify the precise variants responsible for the associations. Groups of genetic variants that are inherited together, or haplotypes, are considered to be in linkage disequilibrium (LD). Generally the variant with the highest significance is reported as the eQTL or GWAS marker variant, but any number of the variants in LD with this marker variant may be responsible for or contribute to the association signal. Further study of GWAS and eQTL loci is required to identify the precise variant(s) responsible for the association signals and elucidate their influence on gene transcription.

Experimental studies are able to identify genetic variants influencing the regulation of gene transcription. These studies have experimentally identified the variants contributing to differential gene transcription at GWAS loci[18,22]. For example, at the low-density lipoprotein cholesterol (LDL-C) level-associated *SORT1* locus, on chromosome 1, one of the alleles at rs12740374 creates a C/EBPa binding site in a regulatory region. In mice, increased C/EBPa binding in this region, increases expression of the Sort1 protein, ultimately reducing LDL-C levels[23]. Like *SORT1*, variants altering protein binding, chromatin accessibility, and gene transcription continue to be reported for other loci associated with cardiometabolic phenotypes [27–33]. Experimental evidence identifying the variants regulating gene transcription at known loci is being discovered at a much slower rate than novel GWAS loci. Improved analytical and experimental techniques are necessary to identify regulatory variants at these established and novel loci.

1.4 Next-generation sequencing data is a powerful tool in genetic studies

Next-generation sequencing data from functional genomics experiments is a powerful tool in identifying variants regulating gene transcription and contributing to complex phenotypes. The advent of these technologies, capable of generating nucleotide sequence data for millions of DNA or RNA fragments[34], has led to scientific advancements in genetic studies. Whole-genome and whole-exome studies of large populations of individuals[35] continue to increase understanding of genetic variation and the effects of this variation on gene expression and complex phenotypes. Additionally, RNA-seq, in which RNA is isolated and sequenced, can be used for genome-wide measurement of transcription levels, and other quantitative sequence data, such as ChIP-seq[36], FAIRE-seq[37], DNase-seq[38] and ATAC-seq[39], can be used for genome-wide identification of gene transcription regulatory regions.

In each of these quantitative sequence data experiments, DNA is isolated from targeted genome-wide regions and sequenced. Sequence reads are mapped to a reference genome, and the regions of signal enrichment are identified based on the number of mapped sequence reads. ChIP-seq can be used to identify the location of transcription factor binding or histone modifications across the genome. DNase-seq, ATAC-seq and FAIRE-seq use different assays to identify regions of open chromatin typically indicative of regulatory regions. In DNase-seq, the

DNA is digested with the DNaseI enzyme to identify DNase hypersensitive regions, while in FAIRE-seq DNA fragments are isolated from regions nucleosomes-depleted regions. In ATAC-seq, the most recent open chromatin assay developed, transposase activity is used to identify regions of open chromatin.

To facilitate genome-wide identification of these regulatory regions, massive amounts of these quantitative sequence read data have been generated. For example, the ENCyclopedia of DNA Elements (ENCODE) consortium has generated more than 1,640 quantitative sequence datasets in more than 147 different cells lines and tissues[40]. These data, including transcription factor and histone modification ChIP-seq, RNA-seq, FAIRE-seq, and DNase-seq, can be found in online repositories at the UCSC Genome Browser[41] and the ENCODE Portal[42]. Likewise, the Roadmap Epigenomics project has generated over 2,500 datasets from 111 different cell types and tissues, including histone modification ChIP-seq, DNase-seq, and RNA-seq[43]. These data are available at the Human Epigenome Atlas[44]. Additionally the sequence data generated by these large consortia, as well as data generated by smaller collaborations or individual labs, are also available online in less specialized repositories, such as the Sequence Read Archive (SRA[45]) and the European Nucleotide Archive (ENA[46]).

The available quantitative sequence data is especially useful in identifying GWAS variants located in genomic regions that regulate gene transcription. Several published studies have used these data to predict regulatory regions genome-wide by integrating ChIP-seq data that identify protein binding sites[47,48] or histone modifications[43,49], with FAIRE-seq[50] or DNase-seq[19,51] that identify regions of chromatin accessibility. Additionally, multiple bioinformatics tools, such as HaploReg[52], RegulomeDB[53] and Annovar[54], have been created to annotate variants using quantitative sequence data and in some cases to predict the likelihood of a variant being in a regulatory region. Some methods also integrate RNA-seq data to identify direct correlations between protein binding or chromatin accessibility with gene transcription levels[43,51,55,56]. While GWAS variants are often annotated with previously predicted regulatory regions, some analyses directly consider the effects of genetic variation during regulatory element identification.

Allelic imbalance in quantitative sequence data can identify differences in the functional activity between two alleles at a variant. Allelic imbalance is indicated by a difference in the number of sequence reads containing each allele, or enrichment of one allele, at heterozygous sites in the quantitative sequence data. Allelic imbalance analyses can identify allelic differences in transcription factor binding[57,58], chromatin accessibility[19,55,59], histone modifications[56] and gene transcription [60–64]. While allelic imbalance identification does not require knowledge of heterozygous sites in the sequenced sample, many studies are limited to samples with full, or complete, genotype information available from whole-genome sequencing. Allelic imbalance can also be detected in samples with limited genotype information by estimating the presence of variants based on populations of individuals such as HapMap[65] and 1000 Genomes[35], and predicting heterozygous sites using the sequence data[19,57,62,64]. Regardless of genotype availability, sequence mapping biases introduced during sequence alignment can complicate and decrease accuracy of regulatory element detection, especially allelic imbalance identification.

1.5 Sequence mapping biases influence regulatory element identification

Read mapping biases occur at heterozygous sites when sequence reads containing one allele are more likely to map to the site than the other allele. While reads originating from a heterozygous site will contain one of the two alleles, reads are commonly mapped to a reference genome containing only a single base, the reference allele, at each heterozygous site. Mapping to a single allele commonly results in reference mapping biases because reads containing the reference allele better match the reference sequence and have a higher mapping quality than reads containing the non-reference allele, which are penalized for a mismatch at the heterozygous site. As a result, reads with non-reference allele are less likely to map correctly, decreasing the accuracy of sequence alignment and downstream analyses, particularly allelic imbalance detection. Many commonly used alignment software, such as BWA[66], MAQ[67], Bowtie[68], and STAR[69], are only built to map to single-allele reference genomes while other software, such as GSNAP[70], allows sequence mapping to multiple-allele references. Regardless of the software used, reference mapping biases will be introduced if reads are aligned to a reference containing a single allele.

A second type of bias can influence both reference and non-reference alleles and occurs when reads containing one allele map to multiple genomics locations and reads containing the other allele map to a single location. Filtering reads based on the number of genomic locations, a common practice, may preferentially remove reads containing the allele that maps to multiple locations, resulting in mapping bias. This bias and reference mapping bias can be corrected during post-alignment steps, but reference mapping biases are most commonly corrected during sequence mapping. Current methods of correcting reference mapping biases are described below.

1.6 Several methods have been reported to remove reference mapping biases

Methods for correcting reference mapping biases vary and are dependent on the underlying data structure of the mapping software. Differences in each sequence mapping software determine the method of reference bias correction associated with each. Many commonly used alignment software, such as BWA, Bowtie, and STAR, utilize a suffix array during sequence mapping. A suffix array contains a list of all possible suffixes of the reference genome, and their corresponding positon in the genome. Traversing suffix arrays takes little time, resulting in very fast sequence alignment. Using compressed suffix arrays such as the Burrows-Wheeler transformation can also reduce the amount of computational resources required, making suffix array-based software desirable when considering both computational time and resources[71]. Other alignment software, such as GSNAP, and MAQ, utilize hash-tables to store either the genomic locations or sequence reads associated with all possible sequences of length *k*. Using these *k*-mers, sequence reads can be matched to genomic locations, although alignments using hash tables are generally slower than alignments using suffix arrays. Most reference bias correction strategies can be implemented with software based on either suffix arrays or hash tables, although many bias correction methods have preferred the speedier suffix-array software.

Five strategies to correct mapping biases include using a biased mismatch threshold[19], variant masking[72], dual reference genomes[57,58,60–62,73], modified dual reference genomes[57], or creating an extended reference genome that includes sequence containing alternate alleles[64]. Alternatively, a slight modification of software utilizing hash-tables can create allele-aware aligners which also correct for reference mapping biases[55,56,74]. Advantages and disadvantages of each reference mapping bias correction strategy are discussed in more detail below.

1.6.1 Biased mismatch threshold

The simplest proposed correction for the reference mapping bias is to use a biased mismatch threshold. To overcome the reference bias, this approach uses a stricter mismatch threshold for reads containing the reference allele when calculating allelic imbalance. Reads containing the non-reference allele are allowed an additional mismatch compared to reads containing the reference. For example, if reads containing the reference allele are allowed one mismatch then reads containing the non-reference allele are allowed to have one or two mismatches to account for the mismatching non-reference allele. The main advantage of this approach is that sequence mapping can be performed with any mapping software using a single reference sequence and without requiring any modifications to the reference. This approach overcomes mapping biases at a majority of heterozygous sites, but is at a disadvantage when removing bias in regions containing more heterozygous sites than the relaxed mismatch threshold. In these cases, reads containing the reference allele may match perfectly and map to the reference while reads containing non-reference allele are not mapped at all. Additionally, the extra mismatches in these reads containing the non-reference allele may cause the reads to map equally as well to incorrect locations, which is especially problematic when only considering reads mapping to a single location.

1.6.2 Variant masking

Masking known variant sites during sequence alignment is also a fairly simple strategy. Any bases mapping to masked variants mismatch the genome, and the reads are equally likely to be mapped correctly regardless of the allele present. The masked variants can be known heterozygous sites in the sample, or simply sites from a database of known variants, making this approach advantageous for removing bias in the absence of sample-specific genotype information. Other advantages of this approach are that only a single alignment is required per dataset and in many cases only a single reference genome is required to map sequences from multiple samples. A disadvantage of this approach, particularly when mapping samples with low sequencing depth, is its effect on the number of reads aligned. Masking variants introduces mismatches into reads originating from the heterozygous site, and could cause reads containing sequencing errors to fail to map to the masked genome because of these additional mismatches. The reduced number of reads aligning to these sites decreases power to identify allelic imbalance. When masking all known variants in the absence of genotype information this approach also unnecessarily reduces the number of reads mapped to homozygous sites that have no need for bias correction.

1.6.3 Dual reference genomes

The dual reference strategy to remove reference mapping biases utilizes complete sample genotype information. Most often, sample genotypes are generated using whole-genome DNA sequencing and subsequent variant calling[35,75]. In these cases, phased genotypes from the sample are used to create two haploid reference genomes representing the sequence of the maternal and paternal chromosomes. Sequence reads are aligned to each reference separately and the alignments are merged to ensure that each sequence read is only represented once. In the absence of phased genotypes, each reference can contain one of the alleles at each known heterozygous sites identified by imputation, or simply at known variants sites identified within a population of individuals. When using phased genotypes, the dual reference strategy is particularly accurate at regions where a single read overlaps multiple variant sites, because the phased genomes best account for variants in LD and most accurately represent the sequence of the chromosome of origin. Additionally, once created, the reference sequence can be used to align sequences of any length, and with any alignment software. A disadvantage of this strategy is the added computation time of the second sequence alignment and the complexity of merging the alignments to create a single consensus alignment. Additionally, any changes in the genotype information require recreating the dual references. Finally, for the most accurate bias correction, this strategy is limited to processing the small number of samples that also have complete genotype information from whole-genome sequencing.

1.6.4 Modified dual reference genome

A modified dual reference strategy extends the standard dual reference strategy to allow for mapping bias correction in samples lacking complete genotype information. In this method, an initial non-allele-aware alignment of the quantitative sequence data is used to identify heterozygous sites in the sequenced sample[57]. As with the dual reference strategy, two reference genomes are then created to represent each allele at these heterozygous sites and sequences are aligned separately to each reference. While this modified method is advantageous in correcting reference mapping biases at heterozygous sites without prior genotype knowledge, it does so at the cost of an additional alignment, bringing the total number of required alignments to three. Adding to this disadvantage, while the references in this strategy capture both alleles at heterozygous sites, they are less successful at capturing LD and are not representative of the maternal and paternal chromosomes. Additionally, this method relies on detecting variants in the sequence data, which can be difficult, especially at sites of allelic imbalance[76].

1.6.5 Allele-aware reference creation

Mapping bias can also be removed during a single alignment by creating an allele-aware reference sequence. In this strategy, reference sequence containing the non-reference allele is appended to the reference genome, allowing for reads to be compared to regions containing each allele using a single extended reference genome. An advantage of his strategy is that mapping biases can corrected using any sequence mapping software, a single reference genome, and a single alignment. A disadvantage of this approach is that reference creation depends directly on sequence length and the genotypes or variant information used. Any changes in either of these parameters require recreation of the reference sequence. Additionally, more computation is required to reconcile reads aligning to the normal and extended regions of the reference sequence. Combined with the added time required to extend the reference sequence, this strategy is particularly disadvantageous when mapping sequence reads using different allele-aware sites (i.e. heterozygous sites from different individuals) or sequence read lengths (i,e. sequences from different assays).

1.6.6 Allele-aware aligner

Similar to the biases mismatch, variant masking and extended reference strategies, allele-aware aligners can remove mapping biases with a single alignment. Utilizing the hash table structure during mapping, these aligners compare sequence reads to reference sequence containing each of the two alleles at specified sites. These sites, can be known variants or heterozygous sites from the sequenced individual. While changes in the heterozygous sites and sequence read length may require the reference sequence to be reprocessed before alignment, the actual reference sequence itself does not need to change. This is advantageous for reference bias correction using the same reference sequence and different sets of allele-aware sites, or sequence reads with varying lengths. Additionally, allele-aware aligners only require a single alignment, without the need for additional computation to create a consensus alignment. This strategy is currently limited to software utilizing hash-tables and read mapping can be considerably slower than the suffix array-based alternatives. The reduction in speed becomes disadvantageous if the computation time required for a single allele-aware alignment exceeds the time required for the reference creation, sequence alignment, and consensus alignment creation steps of other strategies.

1.7 Overview of this work

Quantitative sequence data from functional genomics experiments has been instrumental in understanding the regulatory mechanisms influencing gene transcription and the influences of genetic variation on these mechanisms and, ultimately, cardiometabolic traits and diseases. Despite the massive effort to generate and interpret quantitative sequence data, large gaps remain in the understanding of gene transcription regulation at the cardiometabolic phenotype-associated loci, especially at the variant level. Limited understanding of the contributions of genetic variants to transcriptional regulation is caused in part by reference mapping biases, which are present in a majority of existing analyses. Removing these mapping biases allows for more accurate regulatory element identification, and a clearer understanding of how genetic variants influence gene transcription regulatory mechanisms.

In Chapter 2 of this work, I investigate the effects of these reference mapping biases and describe a pipeline, Allele-Aware ALignments for the Investigation of GeNetic Effects on Regulation (AA-ALIGNER), which can be used to remove mapping biases in any quantitative sequence dataset. AA-ALIGNER uses an allele-aware aligner to remove reference mapping bias in either the presence or absence of genotype data. In addition to sample genotype availability, I present a thorough exploration of how experimental conditions and analytical parameters influence the accuracy of allelic imbalance detection. Additionally, I describe sites of allelic imbalance located at inflammatory bowel disease-associated variants, demonstrating the utility of AA-ALIGNER in predicting allelic differences in protein binding at disease-associated loci.

In Chapter 3, I expand allelic imbalance detection with AA-ALIGNER into quantitative sequence data from cell lines and primary cells relevant to cardiometabolic traits and diseases. I summarize imbalance detection in adipose and liver cell lines, and primary pancreatic islets. Additionally, I discuss biological insights gained from imbalance detection in 70 experiments from HepG2 cells. I also describe sites of allelic imbalance identified at loci associated with gene transcription levels and/or cardiometabolic phenotypes.

Lastly, in Chapter 4 I summarize conclusions gained from this work and offer a glimpse into the future of using quantitative sequence data, specifically allelic imbalance identification, to understand the influence of genetic factors on cardiometabolic traits and diseases.

CHAPTER 2: REMOVING REFERENCE MAPPING BIASES USING LIMITED OR NO GENOTYPE DATA IDENTIFIES ALLELIC DIFFERENCES IN PROTEIN BINDING AT DISEASE-ASSOCIATED LOCI

2.1 Background[1]

Genetic studies of complex traits and diseases have been increasing their focus on the contribution of gene transcriptional regulation. The majority of complex trait-associated variants are in non-coding regions [22], suggesting many contribute by altering regulatory activity. Variants can alter transcription factor binding affinity, subsequently affecting transcription levels of target genes [22]. For example, the T allele of rs12740374 increases C/EBPa binding and transcription of *SORT1*, a gene influencing LDL cholesterol level [23]. Identifying precisely which genetic variants are responsible for changing regulatory activity can be difficult.

Quantitative short-read sequence data generated from experiments such as ChIP-seq [77], DNase-seq [38], FAIRE-seq [37], and ATAC-seq [39] broadly identify genomic regions that regulate gene transcription. Sequence information from these experiments can be used to detect allele-specific activity in samples where heterozygous variants are present in or near a regulatory element. For example, an uneven distribution in the number of reads containing each allele at a heterozygous site, referred to as allelic imbalance, provides evidence for differential regulatory activity due to genetic variation. Previous studies have also used quantitative short-read data to correlate genetic variation in regulatory regions with nearby gene expression [55,56] and to show the heritability of allelic regulatory effects [56,59,78–80].

[1] This chapter has been accepted for publication in *BMC Medical Genomics***.** Citation: Buchkovich ML, Ekland K, Duan Q, Li Y, Mohlke KL, and Furey TS. Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci. *BMC Medical Genomics*. 2015. *in press.*

Quantitative sequence data have been generated in hundreds of cell types and tissues by the ENCODE (Encyclopedia of DNA elements) Consortium [40] and Roadmap Epigenomics Project [43], offering a valuable source of genetic regulatory information. Exploration of allelic imbalance in this data is hindered by a lack of complete genotype information for individuals from which these data are derived, and the well-established alignment bias that arises when both alleles at a heterozygous site are not considered during alignment to a reference genome. Sequence reads containing the allele not represented in the reference genome are penalized as an additional mismatch compared to reads containing the reference allele [72], and are less likely to map to the correct genomic location. This can result in false detection of allelic imbalance favoring the reference allele, or failure to detect imbalance favoring the non-reference allele. Several methods for removing this alignment bias have been proposed, including masking known variants in the reference genome [72], aligning reads to two haplotype reference genomes [57,58,60–62,73], using known variants with allele-aware aligners [55,56,74] or creating an extended reference genome that included alternate alleles [64]. For these methods, full genotype information leads to the best results, but this data is rarely available. The performance of these methods using limited or no sample genotype data, compared to full genotype information has not been thoroughly investigated.

To evaluate detection of altered regulatory activity due to genetic variation in quantitative sequence data using full, limited or no genotype information, we created a computational analysis pipeline, called AA-ALIGNER (Allele-Aware ALignments for the Investigation of GeNetic Effects on Regulation). AA-ALIGNER strategically incorporates existing, publicly available tools to accurately annotate regions containing heterozygous variants given varying levels of genotype information, including no genotypes. To remove alignment biases at heterozygous variants, AA-ALIGNER uses the allele-aware aligner GSNAP [70] which has been previously shown to remove mapping biases using complete genotype information [74]. AA-ALIGNER also attempts to correct other biases that can influence imbalance detection, such as incorrect heterozygous site annotations in reference genome sequences and incorrectly detected imbalances due to differences in mappability between reads containing each of the alleles or due to PCR duplications introduced during sequencing [81].

We demonstrate that GSNAP also removes mapping biases using partial genotype data or common variants allowing for accurate identification of allelic imbalances. Using AA-ALIGNER, we determined the effect of experimental and analytical variables such as sequence read length, sequencing depth, number of mismatches allowed during alignment, and imputation quality thresholds on accurate allelic imbalance detection. Our analyses used data from one DNase-seq and thirteen ChIP-seq experiments generated in the GM12878 lymphoblastoid cell line, for which both complete, sequencing-based genotype and partial, array-based genotype information is available. We experimentally detected differential protein binding at six of nine tested imbalance predictions from AA-ALIGNER for CREB1 (Cyclic-AMP Responsive Element Binding protein 1) binding in GM12878 ChIP-seq data, including imbalances at two disease-associated loci. Overall, our results provide important empirical data that can be used to guide the design of and interpretation of similar studies using AA-ALIGNER to accurately annotate heterozygous sites and detect genetically-driven changes in regulatory element activity.

2.2 Results

2.2.1 Overview of AA-ALIGNER

The AA-ALIGNER pipeline is designed to maximize short-read sequence alignment accuracy at sites of DNA variation regardless of genotype availability. These alignments can be used to identify potential sites of regulatory activity, indicated by an enrichment of aligned reads and referred to as peaks, and of allelic imbalance at these sites (**Figure 2.1**). We first construct a sample-specific custom reference genome in a two-step process. To increase the likelihood that the allele in our starting reference genome matches the genotype of any sample, alleles of common variants in the standard reference are modified as needed to the most common allele from a particular population, such as the 1000 Genomes European samples [26]. In a second step, all available genotype information from the sequenced sample is used to further customize this reference sequence such that: (i) at homozygous variants, the sample allele is present; and (ii) at heterozygous sites, one of the two sample alleles is present. Alternate alleles at heterozygous sites are recorded in a separate file during this process. When no genotype information is available, this alternate allele file contains all common minor alleles (MAF > 0.05) for the selected population.

Next, we filter sequence reads to remove low quality sequences and align them to the custom reference genome using GSNAP [70], an allele-aware aligner. GSNAP takes as input the file containing reference and non-reference alternate alleles to equally consider alignments to both alleles. After alignment, we filter (i) sequences aligned to more than one genomic location; (ii) sequences aligned to regions underrepresented in the reference sequence (ENCODE blacklisted regions); and (iii) duplicate reads to correct for PCR artifacts. These final alignments are used to identify peaks and sites of allelic imbalance.

When testing for imbalances, AA-ALIGNER includes predicted heterozygous sites not included in the initial custom reference during sequence alignment. New heterozygous sites are predicted based on having a minimum number of reads containing each of two alleles. In addition, a minimum read threshold per allele can be applied to all heterozygous sites during imbalance detection to guard against incorrectly annotated heterozygous sites. While predicted heterozygous sites are not included in the initial reference genome customization (**Figure 2.1, Box 1**) or sequence alignment steps (**Figure 2.1, Box 3**), they can be added in a second round of reference customization and alignment if desired.

AA-ALIGNER is designed to correct for multiple sources of bias in the data whenever possible. Increasing the minimum read threshold required to test for an imbalance can guard against incorrect heterozygous site identification. Mappability biases, where reads containing one allele map uniquely while reads containing the other allele map to multiple locations and are filtered, may result in an artificial imbalance. AA-ALIGNER only considers reads that map uniquely to the same position in the genome regardless of the allele present. Post alignment filtering of duplicate reads corrects for biases that can arise from PCR duplication during library preparation.

AA-ALIGNER allows key parameters to be specified that influence sequence alignment and post-alignment steps, such as imbalance detection. The minimum read threshold for each allele is one of these parameters. In addition, allowed mismatches can be restricted to predicted heterozygous sites to increase confidence in evidence for multiple alleles. By default, significance of allelic imbalances is determined using a standard binomial test, but the AA-ALIGNER pipeline can be easily modified to incorporate alternative statistical methods of detecting imbalance. Peaks are determined here using SPP [27]. Additional details for individual steps can be found in the Methods. Unless otherwise indicated, the following results are based on alignments allowing for one mismatch, with a minimum of five reads required for each allele, and a nominal binomial p-value threshold of 0.01 for allelic imbalance detection. Each of these parameters is evaluated in detail in the following sections.

2.2.2 Using GSNAP removes alignment biases at heterozygous sites

We first evaluated the ability of GSNAP to overcome the reference alignment bias. We used 50 base pair (bp) CREB1 ChIP-seq reads generated in the GM12878 lymphoblastoid cell line by the HudsonAlpha Institute of Biotechnology as part of the ENCODE project. We created a custom GM12878 reference sequence based on a complete set of genotypes generated by the Broad Institute [83], and we created a GSNAP input file with non-reference alleles for each heterozygous site. To examine whether both alleles at heterozygous sites were equally considered during alignments, we also created a “complement” reference sequence by swapping the allele at each heterozygous site in the initial custom reference with the alternate allele from the input file. We compared sequence alignments to these two reference sequences using three metrics: reads mapped to heterozygous sites; sequence enrichment peaks called at heterozygous sites; and sites of allelic imbalance (**Table 2.1**). Only 120 of the 33.6 million (0.0003%) reads were aligned differently between the two alignments. Manual inspection indicated that these discrepancies were due to GSNAP failing to remove alignment bias when aligning sequences to regions containing more than 5 and as many as 16 heterozygous sites. These 120 differences did not affect the number of peaks or the predicted sites of allelic imbalances identified (**Table 2.1**). These data demonstrate that using GSNAP, AA-ALIGNER overcomes the alignment bias.

To quantify the importance of removing the alignment bias, we used the same metrics to compare allele-aware and non-allele-aware alignments using the same reference sequences. We used BWA for non-allele-aware alignments with the same alignment parameters as GSNAP. By considering alternate alleles, GSNAP (1.3M reads) aligned 8% more reads to heterozygous sites than BWA (1.2M reads; **Table 2.1**). As expected, GSNAP aligned a larger percentage of reads containing the non-reference allele compared to BWA (48% to 43%), more closely reflecting the expectation that each allele should be present in equal numbers of reads. Additionally, we aligned sequence reads to the complement reference using BWA. In contrast to GSNAP, we found that BWA aligned 344K (1.0%) reads differently to the complement and reference genomes. Greater than 54% of reads mapped to the reference allele at heterozygous sites in both BWA alignments (**Table 2.1**), demonstrating the effect of alignment bias on non-allele-aware alignments.

We examined, separately, the effect of biased alignments at heterozygous sites on peak and allelic imbalance detection. Among the top 10,000 peaks with the greatest signal enrichment for each alignment method, using GSNAP identified 1.6% more peaks overlapping a heterozygous site than BWA and predicted 32% more allelic imbalances. Further, 54% of GSNAP-identified imbalances were enriched for the reference allele compared to 60% of BWA-identified imbalance sites (**Table 2.1**). Additionally, the reference allele was enriched in 82% (23/28) of imbalances only detected when using BWA, compared to 49% (39/79) of imbalances unique to GSNAP alignments. The majority of BWA imbalances favored the reference allele in both the standard reference and the complement reference, demonstrating the presence of significant alignment bias. Together, these results demonstrate that alignment biases negatively impact accurate sequence alignment, peak calling and allelic imbalance identification.

2.2.3 AA-ALIGNER identifies sites of allelic imbalance using partial genotypes or common variant information

Complete genotypes are not available for most samples. Therefore, we evaluated how well AA-ALIGNER reproduced allelic imbalance annotations using incomplete genotype information. We separately aligned the same 50 bp CREB1 ChIP-seq reads to custom GM12878 reference genomes derived using (i) partial genotypes determined using the Human1M-Duo BeadChip array and imputed using MachAdmix [84]; and (ii) 1000 Genomes common variants (EUR, MAF>.05) to model the case of no available genotype information. Using allelic imbalances identified with complete genotype information to define true positive (TP), false positive (FP), and false negative (FN) sites, we calculated sensitivity (TP/TP+FN) and precision (TP/FP+TP), or positive predictive value.

Similar numbers of imbalances were identified using all three levels of genotype information (**Table 2.2**). Interestingly, we found that when simply including common variant alleles (no available genotypes), we detected imbalances with similar sensitivity (>73%) and precision (>75%) as with partial genotype information (**Table 2.2**). Including alleles of common variants with GSNAP significantly improved alignment performance compared to BWA with no variant information (**Table 2.3**), even though neither alignment includes any information about the sample's genotype. This improvement results from sites where including both alleles during alignment allowed for the imbalance to be detected. Of the 200 sites of imbalance detected using complete genotypes, 125 were present in the partial genotypes and 141 were common variants. Considering only these 125 and 141 sites, we find that sensitivity is 97% and 94% with 90% and 82.5% precision, respectively. In stark contrast, sensitivity of detection is 33% (partial) and 34% (common) with 45% and 47% precision at other predicted heterozygous sites, defined as sites with 5 or more reads containing each allele.

We considered whether poor performance at predicted heterozygous sites was due to either (i) incorrect identification of homozygous sites as heterozygous using sequencing data [76]; or (ii) incorrect classification of balanced heterozygous sites as imbalanced due to alignment biases. By comparing the complete genotypes from genomic sequencing to imbalances at sites predicted to be heterozygous in the sequence data, we found that of the sites incorrectly predicted to be imbalanced, 58% (18 of 31) using partial genotypes and 83% (19 of 23) using common variants were not heterozygous. When using complete genotype information, AA-ALIGNER does not report imbalances at predicted heterozygous sites. Of the imbalanced sites, 61% (11/18) using partial genotypes and 42% (8/19) using common variants were also imbalanced when using complete genotypes, underscoring the difficulty in using short reads to detect imbalances at predicted heterozygous sites. We incorrectly detected imbalance at 13 sites using partial genotypes and 4 sites using common variants because an increase or decrease in aligned reads containing one allele now caused the site to pass the significance threshold for imbalance.

We tested whether a more stringent binomial p-value threshold than 0.01 would improve performance, by reducing errors resulting from condition (ii). As expected, a stricter threshold reduced the number of imbalances detected, but it also decreased sensitivity and precision (**Table 2.4**), especially at predicted heterozygous sites. Additionally, we found at predicted heterozygous sites the p-values of false positive imbalance sites were more significant than the p-values of true positives sites when using partial genotypes (Mann-Whitney U *P*=.003) and common variants (Mann-Whitney U *P*=.03; **Figure 2.2**). These data suggest that errors in imbalance detection result more commonly from incorrect prediction of heterozygous sites than falsely calling imbalances at true heterozygous sites.

In addition to a binomial test, other statistical methods of detecting allelic imbalance have been used to measure the significance of allelic imbalance [62,74,81]. For example, a beta-binomial test is commonly used to correct for inaccurate imbalance detection caused by over dispersion of the data. Using a beta-binomial test (*P*<.01) for the 50bp pair CREB1 ChIP-seq data reduced the number of sites of allelic imbalance identified by 82-83% using complete, partial or no genotype information (**Table 2.4**). Overall sensitivity and precision of imbalance detection using partial or no genotypes declined to ~50%. Sensitivity and precision remained higher at imputed heterozygous sites (partial genotype alignment) and common variants (no genotype alignment) than predicted and uncommon variants as before. This reduction in the sensitivity and precision of imbalance detection is similar to the reduction seen when using a stricter binomial p-value threshold and is likely related to the increased p-values of false positive sites reported above.

We also considered whether common variants could be annotated more accurately than rare variants due simply to how sequences were aligned to these sites. Using BWA alignments that did not include any variant information, we predicted heterozygous sites and allelic imbalances as above. If we separate these predictions into those sites that are and are not common variants, we find that the sensitivity and precision are significantly higher for common variants (**Table 2.3)**, although still lower than when both alleles were included in the alignment.

2.2.4 Second alignment provides only modest improvement in sensitivity and precision for incomplete genotypes

Previously, Ni et al. [57] described a strategy for detecting allelic imbalance that first identifies heterozygous sites using an initial alignment without variant information, and then performs a second, allele-aware alignment including the predicted variants. We tested whether a similar second alignment would boost the sensitivity and precision of allelic imbalance identification at predicted heterozygous sites. Before the second alignment, the customized reference was updated to ensure that one allele was present at each heterozygous site predicted in the initial alignment, and non-reference alleles were added to the separate variant file. Reads were then re-aligned using this updated variant file and reference, and filtered as before.

Considering the CREB1 data with partial genotype information, this second alignment identified 11 additional correct sites of allelic imbalance while eliminating 6 incorrect sites, increasing the sensitivity to 47% and precision to 58% at predicted heterozygous sites (**Table 2.2**). When using common alleles, two additional correct imbalances were found and one incorrect site eliminated, with little change in sensitivity and precision. While a second, allele-aware alignment increases accuracy at predicted heterozygous sites, these modest gains, still accompanied by a high rate of false discovery, require an additional alignment. For all other analyses, we report imbalances detected after a single alignment.

2.2.5 Shorter read length and lower sequencing depth reduce the number of imbalance predictions but not precision or sensitivity

Most existing ChIP-seq datasets, such as from ENCODE, contain sequence reads shorter than 50 bp. We investigated how read length affects the ability of AA-ALIGNER to identify sites of allelic imbalance by trimming the 3' end of each 50 bp CREB1 ChIP-seq sequence to create 35 bp and 20 bp reads and then aligned these as before. Trimming reduced the overall number of sequenced bases considered by 30% and 60%, respectively. The total number of aligned reads decreased by 3.7% in the 35 bp alignment and 16.7% in the 20 bp alignment, further reducing total base coverage. The number of reads overlapping heterozygous sites decreased by 31.3% and 61.9%, respectively (**Figure 2.4A**), which led to an even greater reduction in number of identified allelic imbalances for 35 bp (106 imbalances; 47.0% reduction) and 20 bp (26 imbalances; 86.6% reduction) reads (**Table 2.2, Figure 2.3B**).

To determine whether reduced allelic imbalance detection was simply due to lower overall base coverage, we randomly sampled 70% and 40% of the 50 bp reads to match total base coverage levels for the above experiments using 35 bp and 20 bp reads. We found that the number of reads aligned to heterozygous sites decreased, as did imbalances identified, at the same rate as with the shorter reads (**Figure 2.3C**). Thus, reducing base coverage had a proportionate effect on allelic imbalance identification compared to reduction in mapping to heterozygous sites. In our original analysis using all 50 bp reads, we noted 22.5% of sites passed the threshold for the minimum number of reads required for each allele to be tested for imbalance by three reads or less (**Figure 2.3D**). As base coverage is reduced, a disproportionate number of these sites then fall below that threshold (N=5).

As expected, the overall number of predicted imbalance sites also decreased with base coverage when using complete genotypes. Compared to the imbalances detected with complete genotypes for each read length, the sensitivity of imbalance calls using partial genotypes or common variants remained greater than 69% and the precision greater than 75%. These data demonstrate that AA-ALIGNER maintains high detection accuracy using partial genotypes or common variants compared to complete genotypes with reduced base coverage.

2.2.6 Number of imbalances identified varies across factors and assays

To ensure that the results from the CREB1 dataset were representative of results from other experiments, we used AA-ALIGNER to predict allelic imbalance in twelve additional transcription factor ChIP-seq datasets and one DNase-seq dataset generated in the same GM12878 cell line. ChIP-seq datasets contained between 14 and 48 million aligned reads, and most reads were 36 bp in length. Overall, we found that for all alignments, imbalance predictions were accurately replicated using incomplete genotypes at sites where both alleles were used in the alignment. Imbalances at new heterozygous sites were again very poorly predicted (**Table 2.5**).

Although the precision of imbalance detection using partial genotypes and common variants was high across datasets, the number of imbalances detected varied greatly (minimum=0, maximum=291, median 19). Read length and sequencing depth influence the ability of AA-ALIGNER to identify sites of imbalance (**Figure 2.3**). We found, though, that measurements related to these characteristics (**Figure 2.4A-C**) were not highly correlated with the number of imbalances detected in these ChIP-seq datasets (0.43 ≥ Pearson R^2 ≥ 0.51). These low correlations suggest that other factors, such as the number of transcription factor binding sites (TFBS) across the genome and their overall genomic coverage also influenced imbalance detection. Alone, TFBS genomic coverage (**Figure 2.4D**) showed low correlation with the number of imbalances detected (Pearson R^2=.35), but measurements that considered sequencing depth, read length and genomic coverage together (**Figure 2.4E-G**) were highly correlated with the number of imbalances detected (0.78 ≥ Pearson R^2 ≥ 0.91). These correlations suggest that the dispersion of sequence signal across the genome needs to be considered in addition to read length and sequencing depth when evaluating the potential of AA-ALIGNER to identify allelic imbalances. While there was a positive correlation between sequencing depth and signal dispersion in ChIP-seq data, the DNase-seq data, had greater sequencing depth (aligned reads) and signal dispersion (genomic coverage), but fewer sites of allelic imbalance identified than some of the ChIP-seq data. These results suggest that sequencing depth and signal dispersion influence imbalance in DNase-seq data differently and that the correlations observed in the ChIP-seq data do not extend to DNase-seq (**Table 2.5**).

2.2.7 Allowing additional alignment mismatches increases sensitivity but decreases precision

Parameters for the different steps of allelic imbalance identification vary across reported methods and can significantly affect results. Increasing allowed alignment mismatches helps overcome missing genotypes, inaccuracies in the reference genome, and errors in the sequence reads, but also results in increased erroneous sequence alignment, particularly when aligning shorter reads. We examined how this parameter affected the performance of AA-ALIGNER with limited genotype information. The 50 bp CREB1 data was processed with complete genotypes, partial genotypes, and common variant information allowing 0, 1, 2 or 3 alignment mismatches. With complete genotype information, the number of imbalances increased only slightly with greater mismatches (<4%; **Table 2.2**).

When using partial genotypes or common variants, aligning with zero mismatches reduced the number of incorrectly aligned reads compared with our default of one mismatch, but at the cost of eliminating reads containing the non-reference allele at heterozygous sites not included during alignment. This led to increased overall precision of imbalance identification, but with significant loss of sensitivity as novel variants could not be predicted (**Table 2**). Of note, the precision of imbalance detection at known variants using zero mismatches was lower than when allowing one mismatch. Allowing two or three mismatches increased the number of imbalance sites identified using incomplete genotypes by more than 29% (**Table 2.2**). The precision at variants included in the alignment did not change, but was greatly reduced at predicted variants, indicating less stringent mismatch thresholds increase the number of misaligned reads resulting in spurious predictions of heterozygous sites and allelic imbalance at these sites. We also tested whether requiring one of the mismatches to be located at the predicted heterozygous site increased sensitivity and precision compared to allowing mismatches at any site and found that results were similar in both cases (data not shown).

2.2.8 Requiring a minimum number of reads containing each allele increases precision at predicted heterozygous sites

To balance sensitivity and precision with incomplete genotype information, we examined the impact of changing the minimum aligned read threshold for each allele required to test for imbalanced sites. Using the 50 bp CREB1 data, we found that as the required number of aligned reads increased from 2 to 10, the number of detected imbalances decreased using any level of genotype information, as expected, with small fluctuations in the overall sensitivity of imbalance identification using incomplete genotypes (**Table 2.2**). At thresholds of 15 and 20 reads per allele, the sensitivity of detection increased at predicted heterozygous sites, boosting the overall sensitivity at these thresholds. When considering imbalances at variants with both alleles included in the alignment, precision only varied slightly, but it increased at predicted sites with higher thresholds. While for most analyses we have required at least five reads per allele, these findings suggests that for known heterozygous sites, using a lower threshold will increase the number of identified sites without compromising precision.

2.2.9 Requiring higher imputation quality does not significantly improve imbalance identification

For each variant on the genotyping array, imputation quality (Rsq) reflects confidence in imputation of that variant within the population of genotyped individuals. As the imputation quality of a variant site increases, our confidence in the accuracy of the genotype assigned in GM12878 also increases. Poorly-imputed variants incorrectly identified as heterozygous in GM12878 and included during alignment can lower the precision of imbalance detection using partial genotype information. Using imputation quality thresholds from 0.3 to 0.9 as a requirement of inclusion during alignment, we tested the influence of stricter thresholds on imbalance precision and sensitivity using partial genotypes. When using a higher threshold of 0.9, some variants with a quality between 0.3 and 0.9 were still predicted to be heterozygous, increasing the precision of imbalance detection at predicted sites, but overall using a threshold of 0.9 reduced the number of false positive sites by 7 compared to 0.3 while decreasing the number of true positive sites by the same amount, resulting in a small increase in precision and decrease in sensitivity (**Table 2.2**).

2.2.10 Allelic differences in CREB1 binding experimentally supported at inflammatory bowel disease-associated loci and other predicted sites

The above analyses assume that imbalances detected using complete genotypes are the most accurate for comparing the effects of reduced information and parameter settings, but they do not address the functional accuracy of the imbalance prediction. Of special interest are sites previously shown to be associated with disease, especially a disease for which the GM12878 lymphoblastoid cell line is relevant. We identified 238 heterozygous sites in GM12878 that are in linkage disequilibrium (1000 Genomes EUR; $r^2 \geq .8$) with one of 218 index SNPs reported for a genome wide association study (GWAS, $P<1.0 \times 10^{-5}$)[31]. AA-ALIGNER predicted allelic imbalances ($P<0.01$) in CREB1 binding in GM12878 at five of these disease-associated loci (**Figure 2.5A**). Two of the sites, rs2382818 (**Figure 2.5B**) and rs713875 (**Figure 2.5C**), are at loci associated with inflammatory bowel disease susceptibility [86–88]. CREB family proteins have previously reported links to inflammation [89], B-cell lymphocytes [90], and inflammatory bowel disease [90].

At rs2382818, 27 reads containing the T allele and 6 reads containing the A allele were aligned using complete genotype, partial genotype, and common variant information (binomial $P=3.2 \times 10^{-4}$; **Figure 2.5B**, **bottom panel**). The T allele of rs2382818 most often segregates with the disease risk allele A of rs2382817 [32]. Electrophoretic mobility shift assays (EMSAs) using purified CREB1, conducted in the absence of chromatin and other nuclear proteins, can experimentally test for differential binding of CREB1 to a specific DNA sequence. Multiple, independently performed EMSAs supported allelic differences in binding at rs2382818 (**Figure**

2D). A second heterozygous site is located 2 bp downstream of rs2382818. Allowing only a single mismatch during alignment prevents reads from aligning if both alleles are not considered. At this site, a peak and an allelic imbalance were only detected when using GSNAP alignments, but not BWA (**Figure 2.5B**), demonstrating the importance of using allele-aware alignments in annotating disease-associated variants. This locus has been annotated as an enhancer based on ENCODE histone modification data [91] and linked with the expression of nearby genes (*SLC11A1, USP37, PNKD*, and *ZNF142*) [92]. We used MEME-ChIP [93] to identify a CREB1 binding motif from the 10,000 strongest ChIP-seq peaks and searched for the presence of this motif at rs2382818 using FIMO (e < $1.0x10^{-5}$) [40], but we were unable to detect the CREB1 motif at this site.

At rs713875 (*MTMR3* locus), 30 reads containing the Crohn's disease risk C allele [34] and 9 reads containing the G allele were aligned using any level of genotype information (binomial P=$1.1x10^{-3}$; **Figure 2.5C**). Allelic differences in CREB1 binding were again supported by EMSA (**Figure 2.5D**). In this example, the imbalance was detected even when only one allele was used in the alignment. The variant rs713875 is contained within a DNaseI hypersensitive site and is predicted to function as an enhancer [91]. Correlation between DNaseI hypersensitivity and gene expression levels suggests that this locus may regulate nearby genes *LIF* and *TBC1D10A*, pseudogene *CTA-85E5.7*, and non-coding RNA *RP3-438O4.4* [51]. Of these, leukemia inhibitory factor (*LIF*) is an IL-6 cytokine believed to have both inflammatory and anti-inflammatory roles [95]. As with rs2382818, we were unable to detect a CREB1 binding motif at this site. For both rs713875 and rs2382818, further study would be required to show whether allelic differences in CREB1 binding alter transcription and affect inflammatory bowel disease.

We tested for allelic differences in CREB1 binding at seven additional sites that contain a CREB1 binding motif and were predicted to be imbalanced by AA-ALIGNER. These seven included rs1107479, which has been associated with mean platelet volume [96] and age-related macular degeneration [97]. Using EMSA, we detected evidence of allelic differences in protein binding in the same direction as our predicted imbalance at 4 of the 7 sites (**Figure 2D**), for a total of 6 of 9 supported imbalances. Surprisingly, at rs1695359, we consistently detected increased protein binding for the allele predicted by our imbalance analysis to have decreased binding. Of the 6 EMSA-supported sites, only 3 were predicted to have allelic differences based on the FIMO-calculated motif score (difference>5). Of the 3 imbalance sites that were not supported by EMSA, only one (rs1695359) had a significant difference in motif binding score, and the allele with the stronger motif score demonstrated increased binding in the EMSA result, rather than the allele predicted to be enriched by imbalance detection. For comparison, we used EMSA to test 5 additional CREB1 binding locations with a heterozygous variant that fell within a CREB1 binding motif, but were not predicted as sites of allelic imbalance (*P*>.3). We found evidence of allelic differences in protein binding at two of these sites (**Figure 2.6**). For these two sites, a CREB1 motif was only predicted when the allele with stronger protein binding was present.

These data provide strong supporting evidence of allelic differences in protein binding at 6 of the 9 predicted imbalanced sites and suggest that the sequence-specific binding preferences of CREB1 influence binding at these sites. It is unclear whether the remaining three sites not supported by EMSA indicate errors in AA-ALIGNER imbalance detection, or whether these show limitations of EMSA in detecting *in vivo* allelic differences in protein binding that are dependent on chromatin context or the presence of other nuclear proteins. Likewise, it is unclear whether AA-ALIGNER failed to detect allelic imbalance at two sites with allelic differences in protein binding based on EMSA, or whether chromatin and/or other proteins compensate for reduced sequence specificity *in vivo* resulting in similar binding regardless of allele present. Overall, these EMSA results provide evidence supporting allelic differences in protein binding at individual imbalance sites detected by AA-ALIGNER.

2.3 Discussion

In this study, we have demonstrated the ability of AA-ALIGNER to remove mapping biases and to identify allelic imbalance with high sensitivity and precision when using partial or no prior genotype information compared to using complete genotype information. Thoroughly testing allelic imbalance detection using three levels of genotype information provides a clear picture of the accuracy of AA-ALIGNER when using limited genotypes compared to complete genotypes.

This is the first in-depth study of allelic imbalance detection in ChIP-seq and DNase-seq data that empirically tested the effects of key aspects of these analyses including genotype availability, read length, alignment parameters, imputation parameters, and requirements for predicting heterozygous sites. Our results indicate that including any amount of genotype information, or both alleles at common variants, significantly increases accuracy of imbalance detection compared to predictions when complete genotypes are known. We clearly show that predicting heterozygous variants with these short read data is highly inaccurate, leading to false positive rates of imbalance detection greater than 50%. We used a simple metric to predict heterozygous sites, and so one could argue that more sophisticated prediction methods could improve performance. A recent study examining the accuracy of genotyping with short reads from genomic sequencing found that removing sites with strong allelic imbalance, the very sites we are trying to identify, increased genotype accuracy [76]. That study highlighted the difficulty of identifying heterozygous sites from ChIP-seq and DNase-seq data, especially at imbalanced sites. Taken together with our data we strongly suggest that predicted genotypes should be further validated before embarking on functional analyses.

Predicting heterozygous sites in genome sequencing data is an active area of research, and many studies have demonstrated the difficulty of calling variants in sequencing data [76,81,98]. In addition to the GM12878 genotype annotation used in this study, other generally more conservative annotations exist. We found that most predicted imbalances were at common variants, and even when all common variants were included in alignments in the case of no genotypes, the true heterozygous variants and imbalances could be predicted well at these common variant sites. In contrast, the accuracy of imbalance detection at predicted heterozygous sites corresponding to rare variants is poor, even when these predicted heterozygous sites were included in a second alignment. Inaccurate imbalance detection can be caused by either i) incorrectly predicted heterozygous sites in the sequencing data (false positives) or ii) correctly predicted heterozygous sites in the sequencing data that were incorrectly annotated in the complete genotype (false negatives). Requiring more evidence to predict heterozygous sites increased the accuracy of imbalance detection, suggesting that false positives in heterozygous site predictions contributed to inaccurate imbalance detection. These incorrect predictions may be partly due to sequencing errors, but as some are still present at high minimum read thresholds, errors in sequence mapping likely contribute to false positives. The inclusion of incorrectly annotated heterozygous sites or absence of true heterozygous sites during sequence alignment can cause erroneous read mappings to highly similar genomic regions leading to incorrect heterozygous site identification.

Interestingly, many imbalances at sites not annotated as heterozygous in the complete genotype would have been considered imbalanced in the complete genotype alignment using our criteria. This suggests that errors may exist in the complete genotype data leading to false negative imbalance predictions. Further study is needed, but these data suggest that both false positives and false negatives contribute to decreased detection accuracy at predicted variants. Thus, AA-ALIGNER outputs three sets of detected imbalance sites: i) a complete set of all imbalances identified; ii) imbalances at known or common heterozygous variants (higher confidence); and iii) imbalances at predicted rare variants (lower confidence).

We showed that simply including both alleles for common variants resulted in annotations nearly as accurate as those generated from imputed genotypes. Including information about rare variants may further increase sensitivity of imbalance detection. We only considered imputed genotypes and common variants separately, but carefully combining information from these sources may perform better than either individually and is an area of future research.

Other tested parameters demonstrated the trade-off between sensitivity and precision based on their settings, but in most cases these parameters had little effect other than to change the number of predicted imbalanced sites. Nevertheless, these results can be used to guide the analysis of new data, and AA-ALIGNER allows for the easy specification of these parameters. For example, it may be prudent to apply different criteria when evaluating variant sites with known genotypes or that are common variants compared to those predicted to be heterozygous based solely on the short read data. For most of our results, we required a minimum of five reads per allele when testing for imbalances to prevent erroneous testing of homozygous variants. When strong evidence exists for heterozygosity, though, this requirement may be loosened or eliminated, allowing for greater sensitivity in identifying more extreme imbalances. While it is prudent to require a minimum read threshold of reads to detect imbalances at predicted heterozygous sites, this threshold precludes the identification of complete imbalance at known heterozygous sites where only one allele is present, such as imprinted loci. When using known heterozygous sites, AA-ALIGNER users have the option to detect complete imbalance at these sites.

The lack of a comprehensive catalog of experimentally validated sites with functional allelic differences limits our ability to evaluate allelic imbalance predictions. Our study used results obtained from complete genotypes, the best-case scenario for imbalance detection, as the standard for evaluating analyses with partial genotypes and common variants. We experimentally tested for allelic differences in CREB1 binding using EMSA at nine sites with predicted allelic imbalance and five sites with no predicted imbalance. In general, EMSA results matched predicted differences in FIMO-calculated motif scores based on the presence of each of the two alleles, though we note that we were able to detect allelic imbalance and observe differential protein binding at three sites without predicted allelic differences in motif scores. EMSAs were performed in the absence of chromatin context and other nuclear proteins, and so are limited to detecting differences in the sequence binding specificity of a protein. Despite this limitation, we detected allelic differences in CREB1 binding at 6 of 9 predicted imbalanced sites providing strong supporting evidence of allelic differences in protein binding. Further testing is required to understand the cases when EMSA results do not support predicted allelic imbalances. For example, it is unknown whether any of the 3 sites not supported by EMSA were falsely detected as imbalanced by AA-ALIGNER, or whether they failed to validate because of the limitations inherent to EMSA. Likewise, further study is needed to determine whether the two sites that AA-ALIGNER did not predict as imbalanced but that EMSA showed allelic differences in protein binding are due to limitations in AA-ALIGNER or EMSA. These results highlight the need for better experimental assays to validate allelic imbalances, and underscore the difficulty of creating comprehensive catalogs of sites with experimental evidence of differences in protein binding.

The most appropriate statistical test and significance threshold for determining imbalanced sites is not known. While the binomial test is commonly used, other statistical methods such as a beta-binomial [62,99], and Bayesian frameworks [25, 46] have been shown to accurately detect allelic imbalance. For our analyses, we used the more optimistic binomial test and determined significance using an uncorrected p-value threshold of 0.01. Our data indicate that stricter p-value thresholds do not significantly affect the sensitivity and precision of predictions using incomplete genotypes when compared to complete genotype annotations. Incorrectly predicted heterozygous sites often had very small p-values (25% at $P<10^{-7}$), thus stricter p-values will not eliminate these false positives. Likewise, using beta-binomial p-values to correct for over dispersion and setting the same uncorrected p-value cut-off greatly reduced our power to detect allelic imbalance. Using the beta-binomial p-value, imbalance detection accuracy and precision remain significantly higher for imputed and common variants than for predicted rare variants. Our experimental EMSA results were strongest overall for sites with lower p-values, although we did show evidence for altered binding at rs713875 (binomial *P*=1.0x10^{-3}) and rs2382818 (*P* =3.2x10^{-4}) but not rs72694799 (*P* =2.6x10^{-5}) (**Figure 2.5D**). Sites with less statistically significant changes in allelic data may be biologically inconsequential, or the functional effects may simply be weaker but still biologically significant. Until a larger set of experimentally supported sites exists, we cannot determine which statistical test and p-value threshold best identifies biologically relevant imbalance sites. AA-ALIGNER was designed to be modular allowing for allowing for the incorporation of alternative methods for variant identification and tests for significance of imbalances.

Copy number variants (CNVs), which can have significant impacts on disease [100], can cause one allele to overrepresented in the genomic DNA leading to biologically inconsequential imbalances in read data. Prior CNV information for the sequenced sample can be used to preclude imbalance detection within CNVs. Alternatively, sequence data from non-ChIP genomic input or other control experiments, when sequenced with sufficient read depth in the same sample, could be used to estimate an expected proportion of aligned reads per allele and to adjust for copy number variation within the binomial test. These control sequences could also correct for other biases that cause incorrect allelic imbalance detection in both the control and ChIP-seq data. Like genotype information, CNV data is not available for most samples. At this time, AA-ALIGNER does not specifically incorporate known CNV data, although known CNVs can easily be included as “blacklisted” regions and filtered post-alignment. Alternatively, presence of CNVs could be experimentally tested for at AA-ALIGNER predicted imbalance sites

2.4 Conclusions

Allelic imbalance analyses in quantitative sequence data from functional genomic experiments such as ChIP-seq and DNase-seq data is a powerful way to identify effects of genetic variation on gene regulation and uncover molecular mechanisms responsible for GWAS loci in non-coding genomic regions. Reference mapping biases at heterozygous sites and a lack of genotype information for sequenced samples greatly hinder allelic imbalance detection in most public *-seq data. Our analyses demonstrate that the AA-ALIGNER pipeline overcomes mapping biases and accurately identifies a majority of imbalance sites using only partial or no genotype information compared to complete genotype information. Additionally, we provide valuable insight into how experimental and methodological design factors effect imbalance detection.

With AA-ALIGNER, we were able to detect allelic imbalance in ChIP-seq data for a single transcription factor from a single cell line and provide supporting experimental evidence of differential protein binding at a small subset of imbalanced sites. These sites with experimental evidence included variants at two inflammatory bowel disease-associated loci. We demonstrated that mapping biases at one of these two sites prevented detection of both signal enrichment and allelic imbalance using standard analytical techniques. Existing knowledge of B-lymphocytes, regulatory regions and nearby genes suggest a plausible role for these imbalanced sites in inflammatory bowel disease pathogenesis, highlighting the utility of imbalance detection in annotating disease-associated loci. Replicating this analysis in additional cell lines and for additional factors should continue to uncover allelic imbalance at numerous other GWAS loci, providing powerful insight into likely genetic effects on regulation.

2.5 Methods

2.5.1 Genotype Data

Genomic sequencing-based variants calls for GM12878 were generated by the Broad Institute. Illumina Human-1MDuo BeadChip array genotype data generated by the HusdonAlpha Institute of Biotechnology for GM12878 and 52 other ENCODE samples were obtained from the UCSC genome browser [41]. Autosomal genotypes for all 53 samples were imputed using MaCH-Admix [84] with default parameter settings and the reference panel from the 1000 Genomes

Project Phase I version 3 (2012-03-14 release). Chromosome X genotype data in the 53 samples were pre-phased using MaCH [101] with options --states 500 and --rounds 400 and then imputed using minimac [102] with options --state 10 and --rounds 10. Post-imputation filtering of variants according to Rsq was performed as previously reported [103].

Common alleles (MAF > 0.05) used to derive the initial custom reference genome were based on 1000 Genomes Phase I version 3 EUR population [104].

2.5.2 Custom reference creation

The initial European-specific reference genome was created by replacing alleles in the hg19 reference sequence with the major allele for all common variants (MAF>.05) from the 1000 Genomes EUR population. The GM12878 custom reference was created by further modifying this initial custom reference by replacing non-reference homozygous variants with the new allele, based on information from either the full genotype or partial genotype.

2.5.3 Quantitative sequence data processing

Sequence fastq files (**Table 2.6**) were downloaded from the UCSC Genome browser ENCODE Project [41]. Sequences from each replicate were filtered with fastx_trimmer using options `-f 1 -l 50 –Q 33` and fastq_quality_filter using options `-Q 33 –p 90 –q 20 –l *N*` where *N* is the length of the reads in that dataset.

Standard GSNAP alignments were performed using the following options: --sampling=1, --terminal-threshold=10, -n 1, --query-unk-mismatch=1, --genome-unk-mismatch=1, --trim-mismatch-score=0, -t 7, and -A sam. The k-mer size parameter was set based on read length: -k=15 (50bp); -k=11 (35bp); -k=10 (20bp) with —basesize set to k-mer size. As we increased the number of mismatches allowed during alignment to *m*, we changed the option –m to *m* and –i to *m*+1 to disallow indels during alignment. The directory containing the GSNAP reference genome was specified with –D the genome name with –d. Alternate alleles at variant sites based on partial genotype information or common variants were included in alignments with the –v option. BWA alignments were performed using the bwa aln command with options –n 1, –o 0, and –e 0 and bwa samse with option –n 4. When doing a second alignment, the customized reference was updated, if necessary, to contain one of the alleles at predicted heterozygous sites from the first alignment, sequences were aligned, and the alignments were filtered as before.

Reads aligned to more than one genomic location or overlapping the ENCODE blacklist regions [41] were filtered. Potential PCR artifacts were removed using MarkDuplicates (Picard suite) with options REMOVE_DUPLICATES=TRUE, VALIDATION_STRINGENCY=LENIENT, USE_THREADING=TRUE.

To investigate the effects of reference mapping biases on peak calling, peaks were called using SPP within an Irreproducible Discovery Rate (IDR) analysis [105] as outlined by the ENCODE Consortium [40,106]. Overlaps were determined between the 10,000 peaks with the strongest signal and heterozygous sites identified by genomic sequencing (complete genotypes).

2.5.4 Identifying allelic imbalance

Only sequence bases with a Phred33 base quality score greater than 30 were considered for predicting heterozygous sites or allelic imbalances. To account for mappability differences in alignments based on which of the two alleles was present, the heterozygous base in each sequence read was changed to the alternate allele and re-aligned to the genome. Only reads aligning uniquely regardless of the allele present were used to detect allelic imbalance. Significance was assessed with a binomial probability, $b(a_{1;}\, n,\, 0.5)$, where a_1 represents the number of reads containing allele1 and n is the total number of reads at the heterozygous site and an uncorrected p-value threshold of 0.01. To calculate beta-binomial p-values, we first estimated parameter α of the beta distribution using reference allele proportions across all sites.

A Z-statistic for each tested site was calculated the following equation: $\frac{\hat{\mathrm{P}}-0.5}{\sqrt{\frac{2\alpha+N}{4N(2\alpha+1)}}}$, where $\hat{P}$ is the proportion of reads containing the reference allele and N is the total number of reads at the site.

2.5.5 Electrophoretic mobility shift assays

For each heterozygous variant examined, two sets of complementary 21-mer, biotin-labeled oligonucleotides centered on the CREB1 motif and containing one allele were synthesized by Integrated DNA Technologies. Each set was annealed to create two double-stranded probes for each variant (**Table 2.7**). EMSAs were performed according to the protocol included with the LightShift Chemiluminescent EMSA Kit (Thermo Scientific). Briefly, each reaction containing 1x binding buffer, 1 µg poly(dIdC), and 200 ng of purified CREB1 protein (CreativeBiomart CREB1-26H) was incubated for 15 minutes before adding biotin-labeled probes in a total reaction volume of 20 µl and incubating for another 25 minutes. Reactions were electrophoresed on 6% DNA retardation gels (Life Technologies) in 0.5X TBE buffer (Lonza), transferred to nylon membranes (Thermo Scientific), UV cross-linked and detected with chemilluminescence (Thermo Scientific).

Table 2.1 Allele-aware alignments with complete genotypes (GSNAP) vs no genotype information (BWA)

	GSNAP			BWA		
	Standard	**Complement**[a]	**Difference**[b]	**Standard**	**Complement**[a]	**Difference**[b]
Reads mapped uniquely	33,599,679	33,599,721	120	33,543,808	33,547,947	344,942
Reads at heterozygous sites	1,295,901	1,295,914	120	1,197,696	1,186,891	344,942
Reference allele	675,394	620,517	-	677,697	640,978	-
Non-reference allele	620,507	675,397	-	519,999	545,913	-
Peaks at heterozygous sites[c]	1,618	1,618	0	1,593	1,614	87
Allelic Imbalance Sites Identified[d]	200	200	0	151	147	56
Reference allele	108	92	-	91	82	-
Non-reference allele	92	108	-	60	65	-

[a] Alignment reference contained the non-reference allele of heterozygous sites used to create the standard reference [b] Differs in mapping or detection between alignments to standard and complement references [c] Out of 10,000 peaks with strongest signal [d] binomial p-value<.01

Table 2.2 Allelic imbalance detection accuracy in alignments using partial or no genotypes compared to complete genotypes

	Complete[a]			Partial Genotype[b] Imbalances									No Genotype[c] Imbalances								
	Total	Partial	None	Total			Known variants			Predicted variants			Total			Known variants			Predicted variants		
Factor/Assay (Condition)	N_t	N_p	N_n	N_t	Sens	Prec	N_p	Sens	Prec	N_t-N_p	Sens	Prec	N_t	Sens	Prec	N_n	Sens	Prec	N_t-N_n	Sens	Prec
CREB1 (50 bp)	200	125	141	190	73.0	76.8	134	96.8	90.3	56	33.3	44.6	203	76.0	74.9	160	93.6	82.5	43	33.9	46.5
CREB1 (35 bp)	106	70	81	104	73.6	75.0	74	97.1	91.9	30	27.8	33.3	107	77.4	76.6	87	92.6	86.2	20	28.0	35.0
CREB1 (20 bp)	26	16	16	24	69.2	75.0	17	100.0	94.1	7	20.0	28.6	22	69.2	81.8	17	100.0	94.1	5	20.0	40.0
CTCF (35 bp)	267	187	192	300	83.1	74.0	198	98.4	92.9	102	47.5	37.3	298	85.0	76.2	210	97.9	89.5	88	52.0	44.3
DNase (20 bp)	104	43	47	138	51.0	38.4	42	97.7	100.0	96	18.0	11.5	144	51.9	37.5	55	97.9	83.6	89	14.0	9.0
CREB1 (2 alns)[d]	200	125	141	195	78.5	80.5	135	97.6	90.4	60	46.7	58.3	204	77.0	75.5	156	92.2	83.3	48	40.7	50.0
Mismatches allowed																					
CREB1 (0 mm)	199	122	138	137	58.8	85.4	137	95.9	85.4	0	-	-	160	63.3	78.8	160	91.3	78.8	0	-	-
CREB1 (1m m)[e]	200	125	141	190	73.0	76.8	134	96.8	90.3	56	33.3	44.6	203	76.0	74.9	160	93.6	82.5	43	33.9	46.5
CREB1 (2 mm)	199	124	137	245	80.4	65.3	133	97.6	91.0	112	52.0	34.8	251	81.4	64.5	159	96.4	83.0	92	48.4	32.6
CREB1 (3 mm)	213	123	143	301	79.2	53.2	132	98.4	90.9	169	50.0	23.7	313	81.7	52.7	161	96.4	83.9	152	47.6	19.7
Minimum reads/allele																					
CREB1 (2 reads)	301	178	199	486	73.4	45.5	187	97.2	92.5	299	39.0	16.1	515	75.4	44.1	228	95.0	82.9	287	37.3	13.2
CREB1 (3 reads)	261	156	173	267	70.1	68.8	162	94.9	91.4	105	33.3	33.3	289	72.8	65.7	191	92.5	83.8	98	34.1	30.6
CREB1 (4 reads)	230	142	159	218	71.4	76.0	148	95.1	91.2	70	33.7	42.6	235	74.8	73.2	175	92.5	84.0	60	35.2	41.7
CREB1 (5 reads)[e]	200	125	141	190	73.0	76.8	134	96.8	90.3	56	33.3	44.6	203	76.0	74.9	160	93.6	82.5	43	33.9	46.5
CREB1 (6 reads)	198	122	136	174	70.7	80.5	130	96.7	90.8	44	28.9	50.0	188	73.7	77.7	153	93.4	83.0	35	30.6	54.3
CREB1 (7 reads)	173	109	123	154	72.8	81.8	116	97.2	91.4	38	31.2	52.6	167	75.7	78.4	138	92.7	82.6	29	34.0	58.6
CREB1 (8 reads)	157	100	111	141	72.0	80.1	107	97.0	90.7	34	28.1	47.1	148	75.2	79.7	124	92.8	83.1	24	32.6	62.5
CREB1 (9 reads)	144	91	101	130	72.2	80.0	98	96.7	89.8	32	30.2	50.0	140	75.7	77.9	115	93.1	81.7	25	34.9	60.0
CREB1 (10 reads)	124	80	88	117	74.2	78.6	88	96.2	87.5	29	34.1	51.7	125	76.6	76.0	102	92.0	79.4	23	38.9	60.9
CREB1 (15 reads)	88	60	66	82	77.3	82.9	66	96.7	87.9	16	35.7	62.5	88	80.7	80.7	76	92.4	80.3	12	45.5	83.3
CREB1 (20 reads)	63	47	52	64	84.1	82.8	53	97.9	86.8	11	43.8	63.6	67	88.9	83.6	60	96.2	83.3	7	54.5	85.7
Imputation Rsq threshold																					
CREB1 (Rsq>.3) [e]	200	125	-	190	73.0	76.8	134	96.8	90.3	56	33.3	44.6	-	-	-	-	-	-	-	-	-
CREB1 (Rsq>.4)	200	122	-	190	72.5	76.3	133	97.5	89.5	57	33.3	45.6	-	-	-	-	-	-	-	-	-
CREB1 (Rsq>.5)	200	121	-	187	72.5	77.5	129	98.3	92.2	58	32.9	44.8	-	-	-	-	-	-	-	-	-
CREB1 (Rsq>.6)	200	118	-	186	72.5	78.0	124	98.3	93.5	62	35.4	46.8	-	-	-	-	-	-	-	-	-
CREB1 (Rsq>.7)	200	117	-	185	72.0	77.8	123	98.3	93.5	62	34.9	46.8	-	-	-	-	-	-	-	-	-
CREB1 (Rsq>.8)	200	104	-	182	70.5	77.5	111	99.0	92.8	71	39.6	53.5	-	-	-	-	-	-	-	-	-
CREB1 (Rsq>.9)	200	96	-	176	69.5	79.0	99	99.0	96.0	77	42.3	57.1	-	-	-	-	-	-	-	-	-

[a] Complete genotype alignments use sequencing-based genotypes [b] Partial genotype alignments use array-based genotypes and imputation [c] No genotypes alignments use common variants (MAF>.05) from 1000 Genomes EUR [d] Imbalances called after a second alignment using refined genotypes; known variants are variants included in the first alignment [e] Condition used by default by AA-ALIGNER; N_t total imbalance count, N_p imbalances at imputated heterozygotes, N_n imbalances at common variants, Sens, percent sensitivity, Prec, percent precision

Table 2.3 Precision of imbalance detection in non-allele-aware alignments

	GSNAP			**BWA**								
	Total	MAF≥.05	MAF<.05	Total			MAF≥.05			MAF<.05		
Factor/Assay (Condition)	N_t	N_n	N_t-N_n	N_t	Sens	Prec	N_n	Sens	Prec	N	Sens	Prec
CREB1 (50 bp)	200	141	59	161	61.5	76.4	124	74.5	84.7	37	30.5	48.6
CREB1 (35 bp)	106	81	25	98	64.2	69.4	78	75.3	78.2	20	28.0	35.0
CREB1 (20 bp)	26	16	10	22	50.0	59.1	16	68.8	68.8	6	20.0	33.3
CTCF (35 bp)	267	192	75	306	74.9	65.4	218	84.4	74.3	88	50.7	43.2
Dnase (20 bp)	104	47	57	116	24.0	21.6	25	34.0	64.0	91	15.8	9.9
CREB1 (2 alns)[a]	200	141	59	193	64.5	66.8	147	76.6	73.5	46	35.6	45.7

Imbalance calls using BWA and no genotype information compared to GSNAP alignments using complete genotypes; [a]Heterozyogous sites identified in first alignment are used to create two haplotype references for second round of alignments. Nt total imbalance count Nn imbalances at common variants

Table 2.4 P-value threshold influences allelic imbalance detection using partial genotypes and common variants

	Complete[a]			**Partial Genotype[b] Imbalances**									**No Genotype[c] Imbalances**								
Imbalance	Total	Partial	None	Total			Known variants			Predicted variants			Total			Known variants			Predicted variants		
***P*<.01**	N_t	N_p	N_n	N_t	Sens	Prec	N_p	Sens	Prec	N_t-N_p	Sens	Prec	N_t	Sens	Prec	N_n	Sens	Prec	N_t-N_n	Sens	Prec
CREB1 (50 bp)	200	125	141	190	73.0	76.8	134	96.8	90.3	56	33.3	44.6	203	76.0	74.9	160	93.6	82.5	43	33.9	46.5
CREB1 (35 bp)	106	70	81	104	73.6	75.0	74	97.1	91.9	30	27.8	33.3	107	77.4	76.6	87	92.6	86.2	20	28.0	35.0
CREB1 (20 bp)	26	16	16	24	69.2	75.0	17	100.0	94.1	7	20.0	28.6	22	69.2	81.8	17	100.0	94.1	5	20.0	40.0
CTCF (35 bp)	267	187	192	300	83.1	74.0	198	98.4	92.9	102	47.5	37.3	298	85.0	76.2	210	97.9	89.5	88	52.0	44.3
DNase (20 bp)	104	43	47	138	51.0	38.4	42	97.7	100.0	96	18.0	11.5	144	51.9	37.5	55	97.9	83.6	89	14.0	9.0
***P*<.001**																					
CREB1 (50 bp)	114	67	74	108	65.8	69.4	69	95.5	92.8	39	23.4	28.2	114	69.3	69.3	85	94.6	82.4	29	22.5	31.0
CREB1 (35 bp)	60	39	42	61	70.0	68.9	40	94.9	92.5	21	23.8	23.8	62	71.7	69.4	47	92.9	83.0	15	22.2	26.7
CREB1 (20 bp)	17	10	10	14	64.7	78.6	10	100.0	100.0	4	14.3	25.0	12	64.7	91.7	10	100.0	100.0	2	14.3	50.0
CTCF (35 bp)	140	89	93	166	80.0	67.5	95	98.9	92.6	71	47.1	33.8	162	81.4	70.4	103	97.8	88.3	59	48.9	39.0
DNase (20 bp)	43	8	11	62	32.6	22.6	8	100.0	100.0	54	17.1	11.1	67	37.2	23.9	14	100.0	78.6	53	15.6	9.4
***P*<1.0x10^{-4}**																					
CREB1 (50 bp)	75	45	48	77	66.7	64.9	46	93.3	91.3	31	26.7	25.8	79	68.0	64.6	56	91.7	78.6	23	25.9	30.4
CREB1 (35 bp)	39	24	27	33	61.5	72.7	22	91.7	100.0	11	13.3	18.2	37	66.7	70.3	26	88.9	92.3	11	16.7	18.2
CREB1 (20 bp)	4	3	3	3	75.0	100.0	3	100.0	100.0	0	-	-	3	75.0	100.0	3	100.0	100.0	0	-	-
CTCF (35 bp)	89	50	54	108	76.4	63.0	54	98.0	90.7	54	48.7	35.2	104	79.8	68.3	59	69.3	88.1	45	54.3	42.2
DNase (20 bp)	21	1	2	38	19.0	10.5	1	100.0	100.0	37	15.0	8.1	38	19.0	10.5	2	100.0	100.0	36	10.5	5.6
***P*<1.0x10^{-5}** (Bonferroni correction: alpha=.05 and N=5000)																					
CREB1 (50 bp)	52	30	32	50	59.6	62.0	32	90.0	84.4	18	18.2	22.2	52	61.5	61.5	37	87.5	75.7	15	20.0	26.7
CREB1 (35 bp)	28	19	19	24	64.3	75.0	17	89.5	100.0	11	13.3	18.2	26	64.3	69.2	19	89.5	89.5	7	11.1	14.3
CREB1 (20 bp)	2	1	1	1	50.0	100.0	1	100.0	100.0	0	-	-	1	50.0	100.0	1	100.0	100.0	0	-	-
CTCF (35 bp)	63	34	36	74	74.6	63.5	36	97.1	91.7	38	48.3	36.8	71	76.2	67.6	39	94.4	87.2	32	51.9	43.8
DNase (20 bp)	16	1	1	32	18.8	9.4	1	100.0	100.0	31	13.3	6.5	33	18.8	9.1	1	100.0	100.0	32	13.3	6.2
***P*<1.0x10^{-6}** (Bonferroni correction: alpha=.005 and N=5000)																					
CREB1 (50 bp)	44	24	25	42	59.1	61.9	27	91.7	81.5	15	20.0	26.7	44	59.1	59.1	31	88	71	13	21.1	30.8
CREB1 (35 bp)	19	12	12	18	63.2	66.7	11	91.7	100.0	7	14.3	14.3	20	63.2	60	13	91.7	84.6	7	14.3	14.3
CREB1 (20 bp)	2	1	1	1	50.0	100.0	1	100.0	100.0	0	-	-	1	50	100	1	100	100	0	-	-
CTCF (35 bp)	43	21	24	57	72.1	54.4	23	100.0	91.3	34	45.5	29.4	55	76.7	60	27	95.8	85.2	28	52.6	35.7
DNase (20 bp)	12	1	1	24	16.7	8.3	1	100.0	100.0	23	9.1	4.3	24	16.7	8.3	1	100	100	23	9.1	4.3
Beta-binomial *P<0.01*																					
CREB1 (50 bp)	36	16	17	35	50.0	51.4	18	93.8	83.3	17	15.0	17.6	36	50	50	22	88.2	68.2	14	15.8	21.4

[a]Complete genotype alignments use sequencing-based genotypes [b]Partial genotype alignments use array-based genotypes and imputation [c]No genotypes alignments use common variants (MAF>.05) from 1000 Genome EUR; N_t total imbalance count, N_p imbalances at imputed heterozygotes, N_n imbalances at common variants, Sens, percent sensitivity, Prec, percent precision

Table 2.5 Precision of allelic imbalance detection extends to other transcription factor ChIP-seq data

	Complete[a]			**Partial Genotype[b] Imbalances**									**No Genotype[c] Imbalances**								
	Total	Partial	None	Total			Known variants			Predicted variants			Total			Known variants			Predicted variants		
Factor/Assay (Condition)	N_t	N_p	N_n	N_t	Sens	Prec	N_p	Sens	Prec	N_t-N_p	Sens	Prec	N_t	Sens	Prec	N_n	Sens	Prec	N_t-N_p	Sens	Prec
CREB1 (50 bp)	200	125	141	190	73.0	76.8	134	96.8	90.3	56	33.3	44.6	203	76.0	74.9	160	93.6	82.5	43	33.9	46.5
CREB1 (35 bp)	106	70	81	104	73.6	75.0	74	97.1	91.9	30	27.8	33.3	107	77.4	76.6	87	92.6	86.2	20	28.0	35.0
CREB1 (20 bp)	26	16	16	24	69.2	75.0	17	100.0	94.1	7	20.0	28.6	22	69.2	81.8	17	100.0	94.1	5	20.0	40.0
CTCF (35 bp)	267	187	192	300	83.1	74.0	198	98.4	92.9	102	47.5	37.3	298	85.0	76.2	210	97.9	89.5	88	52.0	44.3
Dnase (20 bp)	104	43	47	138	51.0	38.4	42	97.7	100.0	96	18.0	11.5	144	51.9	37.5	55	97.9	83.6	89	14.0	9.0
Other Factors																					
EBF1 (35 bp)	291	233	248	283	86.6	89.0	233	98.3	98.3	50	39.7	46.0	296	90.7	89.2	251	97.2	96.0	45	53.5	51.1
ZNF143 (35 bp)	90	48	53	103	70.0	61.2	55	100.0	87.3	48	35.7	31.2	107	73.3	61.7	61	98.1	85.2	46	37.8	30.4
ELF1 (35bp)	52	29	29	51	63.5	64.7	31	96.6	90.3	20	21.7	25.0	51	65.4	66.7	35	96.6	80.0	16	26,1	37.5
STAT5A (35 bp)	35	9	13	32	42.9	46.9	9	100.0	100.0	23	23.1	26.1	30	42.9	50.0	12	92.3	100.0	18	13.6	16.7
BCL3 (35 bp)	19	1	1	7	15.8	42.9	1	100.0	100.0	6	11.1	33.3	7	15.8	42.9	1	100.0	100.0	6	11.1	33.2
PAX5 (35 bp)	16	10	10	12	62.5	83.3	10	100.0	100.0	2	0.0	0.0	12	62.5	83.3	10	100.0	100.0	2	0.0	0.0
POL2 (35 bp)	20	8	8	12	45.0	75.0	8	100.0	100.0	1	8.3	25.0	12	45.0	75.0	8	100.0	100.0	4	8.3	25.0
C/EBPb (35 bp)	8	8	8	0	-	-	0	-	-	0	-	-	2	0.0	0.0	1	0.0	0.0	1	0.0	0.0
ZBTB33 (35 bp)	7	3	3	5	24.9	60.0	3	100.0	100.0	2	0.0	0.0	5	42.9	60.0	3	100.0	100.0	2	0.0	0.0
P300 (35 bp)	0	0	0	1	-	0.0	0	-	-	1	-	0.0	1	-	0.0	0	-	0.0	1	-	0.0
NFkB (28 bp)	0	0	0	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-

[a]Complete genotype alignments use sequencing-based genotypes [b]Partial genotype alignments use array-based genotypes and imputation [c]No genotypes alignments use common variants (MAF>.05) from 1000 Genomes EUR; N_t total imbalance count, N_p imbalances at imputed heterozygotes, N_n imbalances at common variants, Sens, percent sensitivity, Prec, percent precision

Table 2.6 Alignment and base coverage statistics for ChIP-seq transcription factor and Dnase-seq data

			Read Counts		Base Counts				
ENCODE Dataset	**Factor**	**Bp**	Total	Heterozygous[a]	Aligned[b]	1X Coverage[c]	10X Coverage[d]	Hets[e]	Imbalances[f]
wgEncodeHaibTfbsGm12878Elf1sc631V0416101	DNase	20	156,758,018	2,758,908	3,135,160,360	1,405,991,528	21,124,395	8445	106
wgEncodeHaibTfbsGm12878Creb1sc240V0422111	CREB1	50	33,599,679	1,295,901	1,679,983,950	970,842,159	10,446,092	5282	200
wgEncodeSydhTfbsGm12878Ctcfsc15914c20Std	CTCF	36	27,756,145	791,146	999,221,220	564,722,839	8,555,802	5402	267
wgEncodeSydhTfbsGm12878Ebf1sc137065Std	EBF1	36	41,942,339	1,204,579	1,509,924,204	953,631,250	7,860,098	3981	291
wgEncodeSydhTfbsGm12878Znf143166181apStd	ZNF143	36	34,120,761	1,005,238	1,228,347,396	840,169,377	5,288,296	3287	90
wgEncodeHaibTfbsGm12878Elf1sc631V0416101	ELF1	36	24,915,146	779,055	896,945,256	565,177,535	5,090,994	2803	52
wgEncodeHaibTfbsGm12878Stat5asc74442V0422111	STAT5a	36	47,942,522	1,372,232	1,725,930,792	1,076,897,191	3,390,070	1311	35
wgEncodeHaibTfbsGm12878Bcl3V0416101	BCL3	36	21,628,124	645,670	778,612,464	588,947,638	1,540,684	890	19
wgEncodeHaibTfbsGm12878Pax5c20	PAX5	36	19,630,594	596,047	706,701,384	540,278,004	1,263,550	448	16
wgEncodeOpenChromChipGm12878Pol2	POL2	36	14,908,730	417,043	536,714,280	418,436,494	966,477	325	12
wgEncodeHaibTfbsGm12878Cebpbsc150V0422111	CEBPb	36	40,620,992	1,137,812	1,462,355,712	1,029,621,494	783,682	331	8
wgEncodeSydhTfbsGm12878NfkbTnfaIggrab	NFkB	28	19,683,073	432,673	551,126,044	461,637,264	454,451	7	0
wgEncodeHaibTfbsGm12878Zbtb33	ZBTB33	36	21,413,742	647,263	770,894,712	628,127,307	166,212	122	7
wgEncodeHaibTfbsGm12878P300	P300	36	15,865,233	477,718	571,148,388	488,303,953	138,015	51	0

[a]Reads at heterozygous sites from complete genotypes [b]Reads aligned using complete genotypes [c]Genomic bases with at least 1 read [d]Genomic bases covered by at least 10 reads [e]Predicted heterozygous sites with at least 5 reads containing each allele [f]SItes with imbalance, binomial *P*<.01; Bp, base pairs

Table 2.7 EMSA probes for experimental validation

Figure 2.5 Probes

	rsID	Position	Allele 1 Probe	Allele 2 Probe
A	rs2382818	chr2:219155907	5'-ACCTCTCTGAAGGGCTCATTT	5'-ACCTCTCTGATGTGCTCATTT
B	rs28712309	chr4:120375837	5'-GCCATTGTGACGTCACGGAAG	5'-GCCATTGTGGCGTCACGGAAG
C	rs72694709	chr4:170533776	5'-TGGCGTGTGACGTCAGCGCGT	5'-TGGCGTATGACGTCAGCGCGT
D	rs28711909	chr4:185769900	5'-CCACTTATGACGTAGCTTTTG	5'-CCACTTATGACATAGCTTTTG
E	rs1107479	chr12:57030685	5-TGCCAAGGACGTCACAGGCAG	5'-TGCCAAGGACGTCATAGGCAG
F	rs73177939	chr12:104351332	5'-CCTCCTGTGACCTCTTAAGAG	5'-CCTCCTGTGACGTCTTAAGAG
G	rs12953558	chr18:76829069	5'-GATCGGTGACGTCATCGGGCC	5'-CATCGGTGACGTAATCGGTCC
H	rs12624512	chr20:62482272	5'-TGGGGAAGACCTCACATAGGC	5'-TGGGGAAGACGTCACATAGGC
I	rs713875	chr22:30592486	5'-GGTGGCCAGCGTCAGCGTTTG	5'-GGTGGCCAGGGTCAGCGTTTG
Figure 2.6 Probes				
A	rs72807213	chr5:175816018	5-GAGGACAGTGATGTCGGAGGG	5'- GAGGACAGTGATATCGGAGGG
B	rs9388486	chr6:126661154	5'-GCTCTCAATGACGTCAGGTAT	5'- GCTCTCAATGACGCCAGGTAT
C	rs274035	chr7:23450198	5'-ACCTCTAGTGATGTAAAGTCT	5'-ACCTCTAATGATGTAAAGTCT
D	rs12145434	chr12:121454342	5'-CTTCATGACGTCACGTGAGAG	5'-CTTCATGACGTCACGAGAGAG
E	rs55811458	chr22:30592486	5'-GCAACTGGTGACATCATGAGA	5'-GCAACTAGTGACATCATGAGA

All probes are labeled with biotin on the 5' end (Integrated DNA technologies- /5Biosq/ tag)

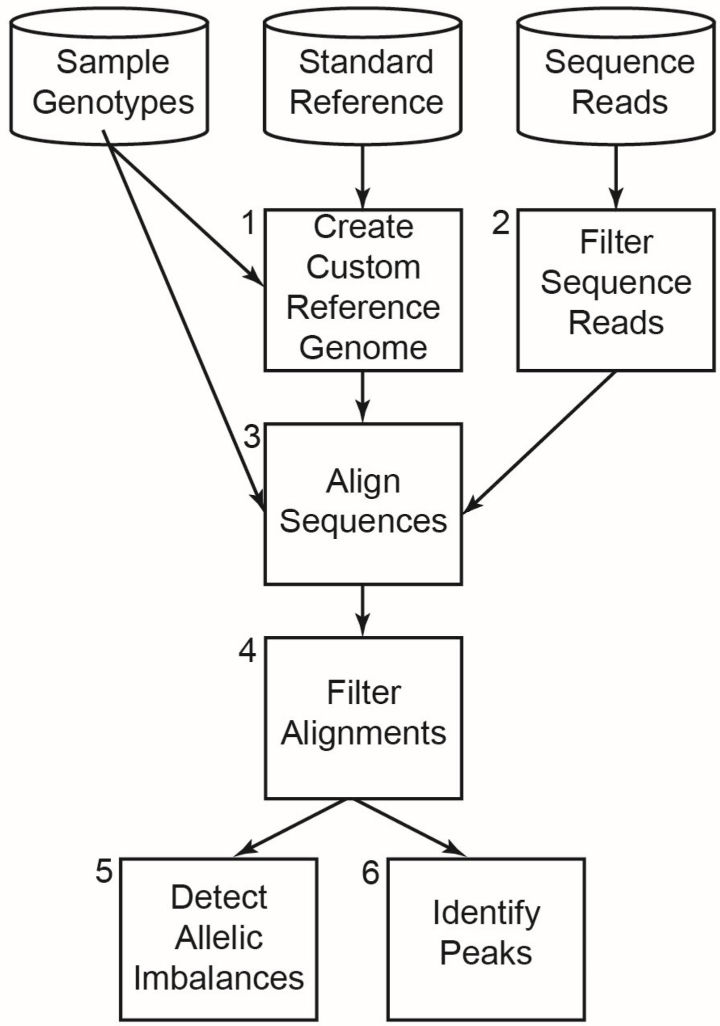


Figure 2.1 Overview of AA-ALIGNER. Sample genotypes or common variants are used to create a custom reference genome (1). Sequence reads are filtered to remove low quality reads (2) and aligned to the custom reference using GSNAP including alternate alleles (3). Alignments are filtered further to increase alignment quality (4) and used to detect sites of allelic imbalance (5, binomial test) and identify peaks (6). Allelic imbalance is tested at heterozygous sites included in the customized reference genome and at predicted heterozygous sites, identified based on a minimum number of mapped reads containing each of two alleles. If desired, predicted heterozygous sites can be used to update the custom reference and be included in a second alignment repeating steps 3-6.

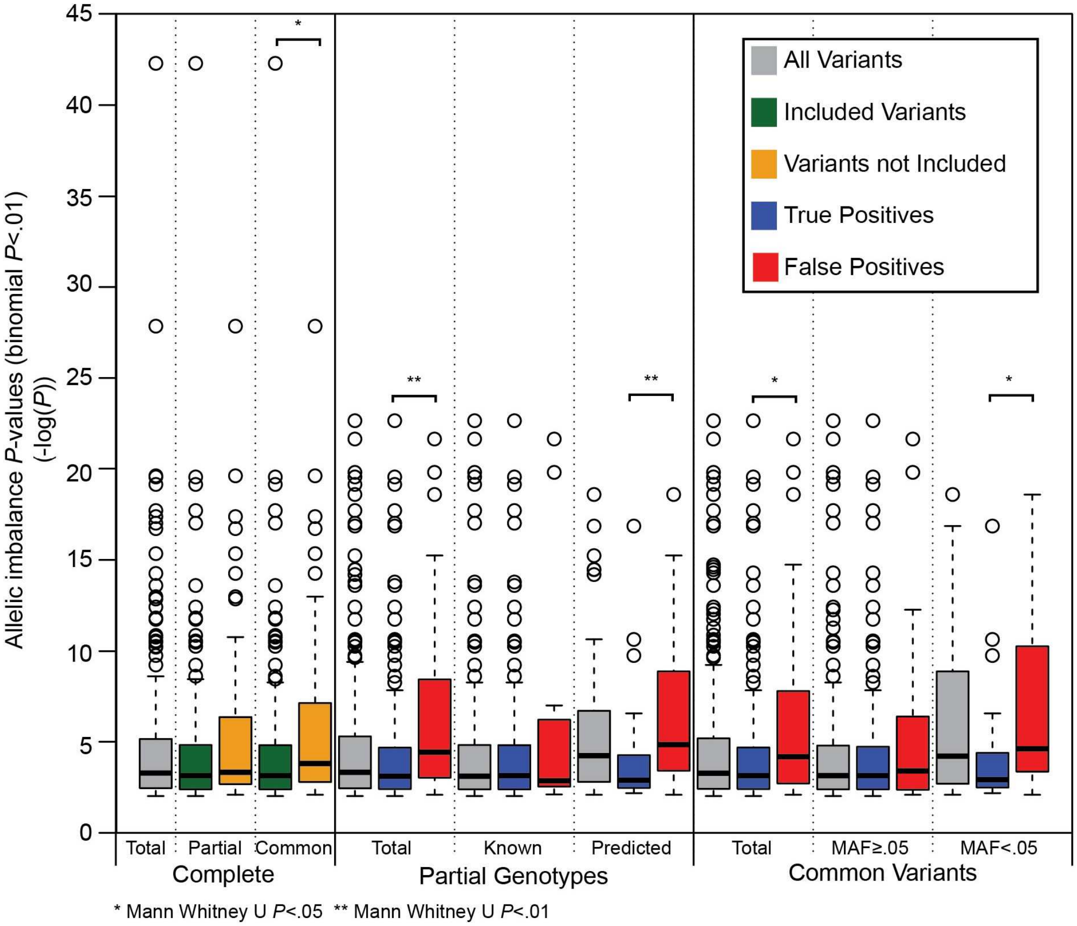
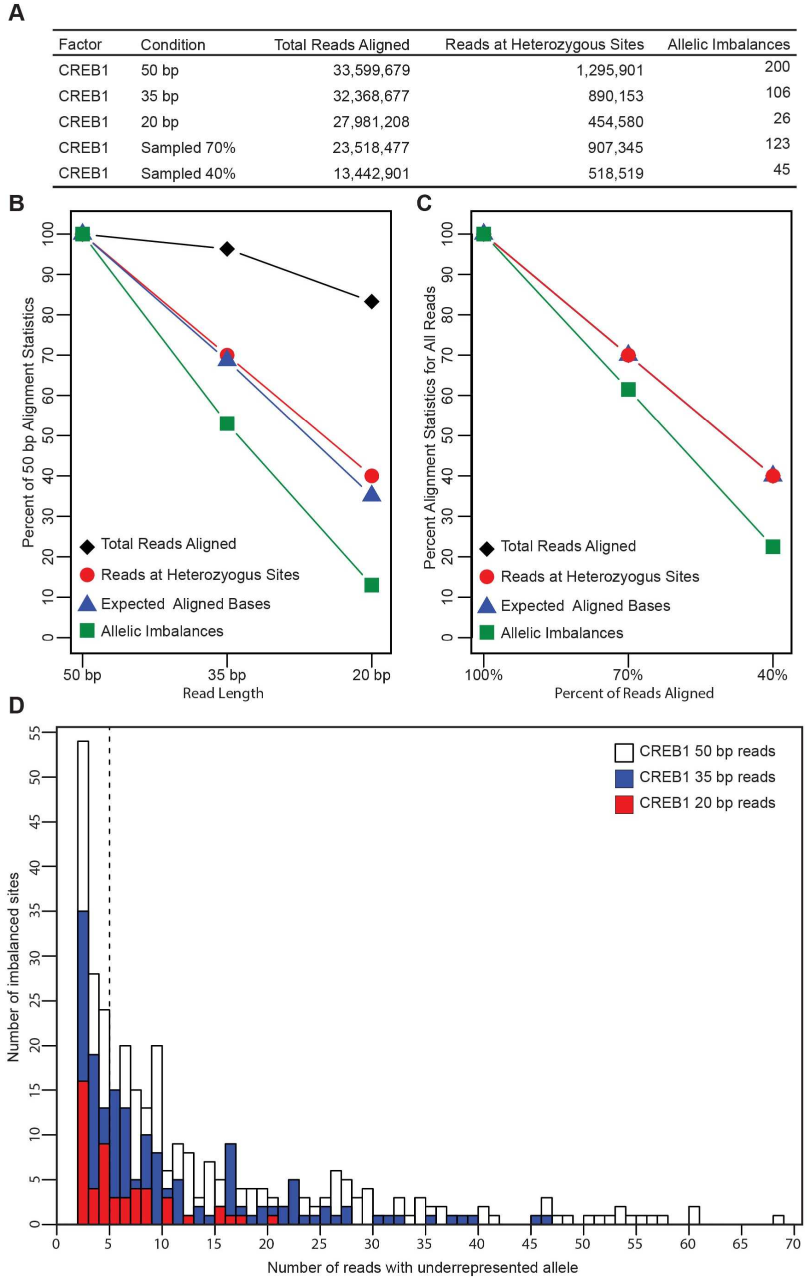


Figure 2.2 False positive imbalance sites have more significant p-values. Boxplot of *P*-values at sites of allelic imbalance using complete (left), and partial (middle) genotypes and common variants (right). For the complete genotype alignment, *P*-values are further subdivided by inclusion in partial genotypes and common variants. For partial genotypes and common variant alignments, all *P*-values are displayed in addition to being divided into inclusion during alignment and predicted from the alignment, and then further divided into whether or not the sites was predicted using complete genotypes (false positive vs true positive). Subdivided groups with significant differences in *P*-value distributions (Mann Whitney U test; *P*<.01 and *P<.001*) are indicated.



A

Factor	Condition	Total Reads Aligned	Reads at Heterozygous Sites	Allelic Imbalances
CREB1	50 bp	33,599,679	1,295,901	200
CREB1	35 bp	32,368,677	890,153	106
CREB1	20 bp	27,981,208	454,580	26
CREB1	Sampled 70%	23,518,477	907,345	123
CREB1	Sampled 40%	13,442,901	518,519	45

Figure 2.3 Read length influences sequence alignment and allelic imbalance detection at heterozygous sites. (A) Alignment statistics for alignments of CREB1 ChIP-seq data, and when using different (B) read lengths and (C) sequence depths, plotted as percent of the 50 bp statistics. (D) Histogram of the number of reads containing the underrepresented allele at sites with significant imbalance(binomial *P*<.01, uncorrected) with 2 or more reads containing each allele. Vertical dashed line indicates the minimum of 5 or more reads containing each allele.

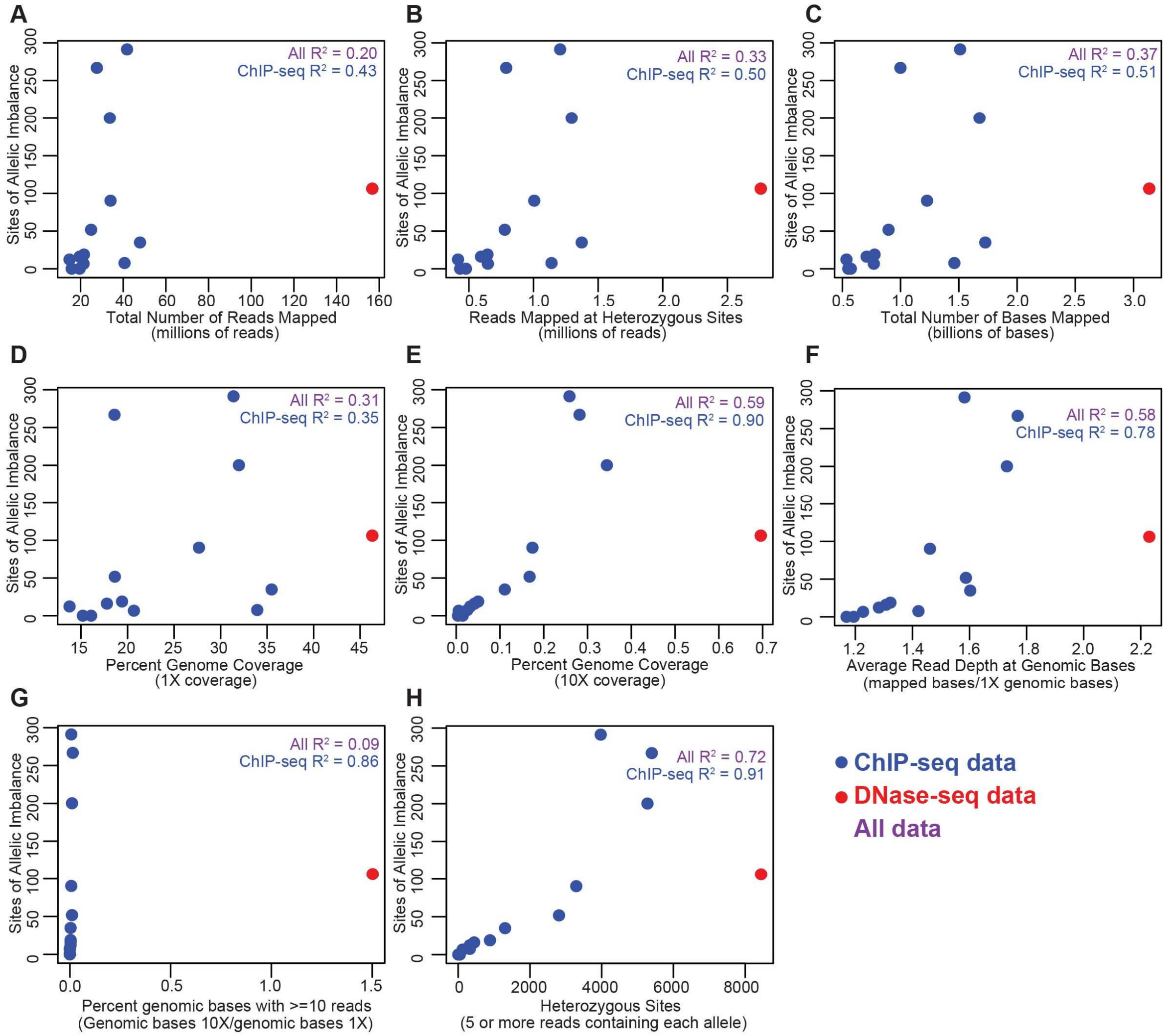
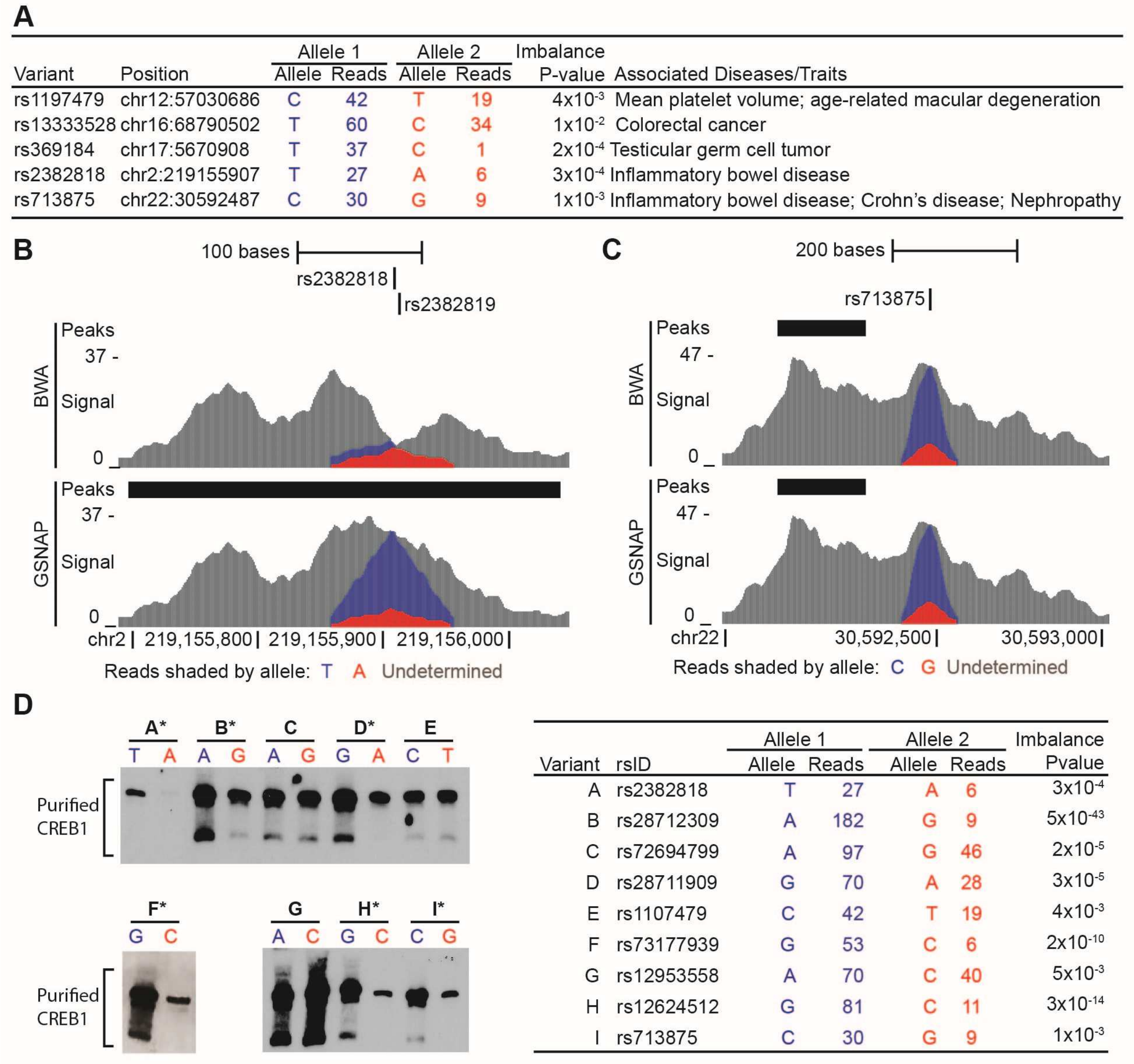


Figure 2.4 Correlation of alignment statistics and number of imbalances detected. The number of sites of allelic imbalance in thirteen ChIP-seq and one DNase-seq dataset compared to the (A) total number of reads aligned; (B) number of reads aligned to heterozygous sites; (C) total number of bases aligned (read length x total reads aligned); the percent of genome with greater than (D) 1X and (E) 10X coverage; (F) the number of heterozygous sites identified; (G) the average read depth at bases with 1X or more coverage; and (H) the ratio of sites with 10X coverage to 1X coverage. Pearson correlation R^2 values for each statistic and number of allelic imbalances are displayed for all data and only ChIP-seq data.

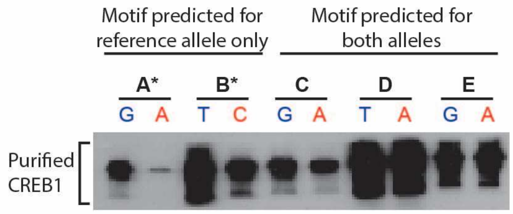


		Allele 1		Allele 2		Imbalance	
Variant	Position	Allele	Reads	Allele	Reads	P-value	Associated Diseases/Traits
rs1197479	chr12:57030686	C	42	T	19	$4x10^{-3}$	Mean platelet volume; age-related macular degeneration
rs13333528	chr16:68790502	T	60	C	34	$1x10^{-2}$	Colorectal cancer
rs369184	chr17:5670908	T	37	C	1	$2x10^{-4}$	Testicular germ cell tumor
rs2382818	chr2:219155907	T	27	A	6	$3x10^{-4}$	Inflammatory bowel disease
rs713875	chr22:30592487	C	30	G	9	$1x10^{-3}$	Inflammatory bowel disease; Crohn's disease; Nephropathy

		Allele 1		Allele 2		Imbalance
Variant	rsID	Allele	Reads	Allele	Reads	Pvalue
A	rs2382818	T	27	A	6	$3x10^{-4}$
B	rs28712309	A	182	G	9	$5x10^{-43}$
C	rs72694799	A	97	G	46	$2x10^{-5}$
D	rs28711909	G	70	A	28	$3x10^{-5}$
E	rs1107479	C	42	T	19	$4x10^{-3}$
F	rs73177939	G	53	C	6	$2x10^{-10}$
G	rs12953558	A	70	C	40	$5x10^{-3}$
H	rs12624512	G	81	C	11	$3x10^{-14}$
I	rs713875	C	30	G	9	$1x10^{-3}$

Figure 2.5 Validation of allelic imbalance detected at GWAS loci and other predicted sites

Figure 2.5 Validation of allelic imbalance detected at GWAS loci and other predicted sites. (A) We detected significant allelic imbalance (binomial P<0.01) in CREB1 ChIP-seq sequence reads at variants at five disease- and trait-associated loci. (B) At rs2382818, sequence reads that failed to align when only single alleles were considered (top) were correctly aligned in an allele-aware alignment (bottom). The increase in aligned reads allowed for the detection of a CREB1 peak (black box) and allelic imbalance at the variant for which more reads were aligned containing the T allele than the A allele were aligned. Total sequence signal is displayed and reads are shaded based which allele they contain. (C) We detected a significantly greater proportion of reads containing the C allele of rs713875 than the G allele. (D) EMSA using purified CREB1 and labeled probes containing each allele at nine sites of allelic imbalance to test for allelic differences in binding. Alleles colored blue are predicted to bind CREB1 more strongly than alleles colored red. Allelic differences in protein binding consistent with these predictions were observed for starred (*) variants. Only CREB1-bound probe is shown. Similar results were observed in a replicate experiment.


Motif predicted for reference allele only
Motif predicted for both alleles
A*
B*
C
D
E
G
A
T
C
G
A
T
A
G
A
Purified CREB1

		Allele 1		Allele 2		Imbalance
Variant	rsID	Allele	Reads	Allele	Reads	P-value
A	rs72807213	G	24	A	17	0.35
B	rs9388486	T	36	C	33	0.81
C	rs274035	G	19	A	17	0.87
D	rs12145434	T	78	A	82	0.81
E	rs55811458	G	42	A	53	0.30

Figure 2.6 Allelic differences in binding at sites without predicted allelic imbalance. EMSA using purified CREB1 and labeled probes containing each allele at five sites with reads mapping but not significant allelic imbalance to test for allelic differences in binding. Alleles colored blue are the reference allele and alleles colored red are the other allele. Allelic differences in protein binding were detected at two sites (starred) without predicted imbalance but with predicted allelic differences in the presence of CREB1 motif. Only CREB1-bound probe is shown.

CHAPTER 3: ALLELIC IMBALANCE DETECTION IN QUANTITATIVE SEQUENCE DATA PREDICTS GENETIC EFFECTS ON PROTEIN BINDING AT LOCI ASSOCIATED WITH CARDIOMETABOLIC TRAITS AND DISEASES

3.1 Background

Complex, cardiometabolic diseases such as coronary artery disease (CAD) and type 2 diabetes (T2D), present a major health concern worldwide. Genome-wide association studies (GWAS) have identified hundreds of genetic loci associated with these diseases[3,4] and other cardiometabolic risk factors[5,7–12,14–16,107]. Around 93% of the variants located at GWAS loci are located in non-coding regions[19] and likely influence the transcription of one or more nearby genes. Indeed, at some of these loci, experimental studies have identified specific genetic variants with allelic differences in DNA-protein interactions, enhancer activity, and/or transcript levels of nearby genes[22]. For example, rs12740374, a variant associated with low-density lipoprotein cholesterol (LDL-C) plasma levels has been show to influence *SORT1* expression through allelic differences in C/EBPB protein binding[23]. Likewise, the T2D risk allele of rs11603334 has been shown to disrupt binding of PAX family proteins and increase transcription of the *ARAP1* promoter[31] and rs11257655, located in another T2D-associated locus near *CDC123/CAMK1D*, demonstrates allelic differences in FOXA1 and FOXA2 binding and enhancer activity[27].

Quantitative sequence data, such as ChIP-seq and DNase-seq, can identify regions actively regulating gene transcription and, more specifically, the location of DNA-protein binding sites across the genome[108]. Finding overlap between these regions and GWAS variants identifies the variants likely influencing the regulation of nearby genes. While multiple variants at GWAS loci can be located in regulatory regions, not all of these variants demonstrate allelic differences in protein binding and enhancer activity. For example, two variants near *ARAP1*, rs11603334 and rs1552224, are located in regions with regulatory evidence, but only rs11603334 has experimental evidence of allelic differences in protein binding and enhancer activity[31]. Quantitative sequence data can also detect specific variants with allelic differences in protein binding that potentially influence the transcript levels of nearby genes and/or cardiometabolic phenotypes. Similar numbers of sequence reads originating from heterozygous sites, are expected to contain each allele. Enrichment of one allele in quantitative sequence data, or allelic imbalance, at heterozygous sites can indicate allelic differences in protein binding or chromatin accessibility. We have shown in Chapter 2 that allelic imbalance detection in ChIP-seq data identifies sites with experimental evidence of allelic differences in protein binding, including sites located at GWAS loci.

Two major limitations of allelic imbalance detection are reference mapping biases introduced during sequence alignment[72], and the necessity of sample genotype data to identify and correct for these biases to accurately identify sites of allelic imbalance. We have previously described, our pipeline, AA-ALIGNER[109], that uses an allele-aware aligner, GSNAP[70], to remove mapping biases and identify sites of allelic imbalance. We have shown that AA-ALIGNER can accurately identify allelic imbalance using limited or no genotype information (**Chapter 2**). Additionally, we have shown that imbalance prediction accuracy is much greater at established heterozygous sites in the reference genome, than sites predicted to be heterozygous in the ChIP-seq or DNase-seq data. Likewise, when heterozygous sites are not known, imbalances are more accurately predicted at common variants (MAF>0.05) than rare variants.

Using AA-ALIGNER, we have identified sites of allelic imbalance in samples from three tissues with documented relevance to cardiometabolic phenotypes: liver[110], pancreatic islets[111], and adipose[112]. We report correlations between data characteristics and the number of imbalances detected and describe biological insights gained from imbalance detection. Using linkage disequilibrium (LD), we report allelic imbalances that are located at expression quantitative trait loci (eQTLs) identified in liver[113,114], adipose[115], and islet[116] samples and/or cardiometabolic phenotype-associated loci[21]. Finally, we describe experimental evidence of allelic differences in protein binding and/or enhancer activity for variants associated with four cardiometabolic phenotype, highlighting the utility of using the allelic imbalance sites to understand gene transcription regulation at these and other loci.

3.2 Results

3.2.1 Allelic imbalance detection in quantitative sequence data from cardiometabolic-relevant tissues

We predicted allelic differences in protein binding by detecting allelic imbalance in quantitative sequence data. Using AA-ALIGNER, we aligned sequence reads from and detected allelic imbalance in 117 publicly available ChIP-seq and DNase-seq datasets from a liver cell line (HepG2), primary human pancreatic islets from 12 individuals, and two adipose cell lines (hASC and SGBS). Together, these ChIP-seq experiments captured the binding sites of 72 proteins in one or more samples. After combining replicate datasets into a single experiment, i.e. ChIP-seq datasets for the same protein from the same sample, we tested for allelic imbalance at heterozygous sites in a total of 90 ChIP-seq experiments and 1 DNase-seq experiment. (**Table 3.1**) We detected significant allelic imbalance at 22,414 heterozygous sites (see **Appendix 1** for subset) across the genome (uncorrected binomial P<.01), with evidence of allelic imbalance detected in more than one experiment at 6,338 (28%) of these sites (**Table 3.1**). At a majority of these sites (98%), imbalance was detected for multiple transcription factors and/or DNase hypersensitivity in a single tissue. A small percentage of sites (2%) were imbalanced for the same or different experiments in multiple tissues.

3.2.2 Combining replicate datasets with low imbalance concordance increased the number of sites of allelic imbalance identified

For increased power to detect imbalances, we combined replicated datasets into one experiment before detecting imbalance. To examine concordance between replicate datasets, we compared the sites of allelic imbalance detected after combining datasets to sites identified in each individual dataset (**Figure 3.1**). Combining replicates identified, on average, 52% more sites of imbalance, suggesting that allelic imbalance detection is influenced directly by sequencing depth. After combining replicates, 19% of imbalances originally identified in individual replicates were no longer significant. When looking across individual replicates only a small percentage (median 14%) of imbalance sites was detected in more than one replicate. This low concordance between replicates could be influenced by sequencing depth of individual replicates, but is also likely influenced in part by variation in experimental protocols used to generate each replicate.

3.2.3 Percent of genome with 8X coverage is highly correlated with the number of imbalances detected

Focusing on datasets from HepG2, we further investigated the relationship between the number of allelic imbalance sites detected and characteristics of the sequence data, such as sequencing depth. We found a modest correlation between sequencing depth and the number of imbalances detected, as suggested by the combined replicate results ($0.53 \leq$ Pearson $r^2 \leq 0.59$; **Figure 3.2A-C**). When examining genomic coverage, we found little correlation between the percent of the genome with at least one read mapped, (1X coverage) and imbalance detection (Pearson r^2=0.33; **Figure 3.2D**), but a very high correlation between the percent of the genome with 8 or more reads mapped (8X coverage) and imbalance detection (Pearson r^2 =0.91; **Figure 3.2E**). Sites with 8 or more reads mapping have the minimum signal intensity required to detect significant imbalance at heterozygous sites (**Figure 3.2F**). Metrics characterizing combinations of sequencing depth, 1X coverage and/or 8X coverage ($0.70 \leq$ Pearson $r^2 \leq 0.76$; **Figure 3.2G-H**) were more correlated with imbalance detection than sequencing depth alone, but less correlated than 8X coverage alone. Correlation values were similar when looking at all datasets and only ChIP-seq datasets, suggesting the DNase-seq and ChIP-seq data are influenced similarly by these characteristics.

3.2.4 Changing the allele present in the reference does not change the alignment

We expect that in the absence of reference mapping bias, around 50% of reads at heterozygous site should contain the reference allele. Likewise, we expect that a similar number of imbalanced sites to have reference allele enrichment as non-reference allele enrichment. Large proportions of reads containing the reference allele suggest a greater number of sites with reference allele enrichment and could indicate the presence of false positives introduced by reference mapping biases. We looked for evidence of this in our data by testing for a correlation between the proportion of reads containing the reference allele at heterozygous sites and the number of imbalances detected. We found little correlation between the number of imbalances detected and the number of reads containing the reference allele (Pearson r^2=0.27; **Figure 3.2I**), although one dataset had a higher proportion of reads containing the reference allele than the others (**Figure 3.2I, green diamond**). Over 54% of these CEBP/B ChIP-seq reads that mapped to heterozygous sites contained the reference allele, compared to the median of 50% over all datasets.

Increased mapping of reads containing the reference allele could be the result of i) reference mapping biases or ii) other unknown experimental or biological factors. To rule out reference mapping biases, we realigned the data to a complement reference genome in which the base at each heterozygous site is changed from the reference to the non-reference allele.

For this analysis, we only changed the reference at sites identified as heterozygous in HepG2 using a genotyping array and imputation to the 1000 Genomes reference panel (see **Methods**). For comparison, we similarly aligned two other ChIP-seq datasets with 51% of reads containing the reference allele. For all three datasets, mapping reads to the complement genome produced the same alignment and had no effect on imbalance detection, indicating that AA-Aligner had successfully removed reference mapping biases (**Table 3.2**) and that other experimental and biological factors are likely responsible for the enrichment of reads containing the reference allele.

3.2.5 The major allele is commonly enriched at sites of allelic imbalance

Like reads mapping to heterozygous sites, we observed greater number of allelic imbalance sites (55.4%) with reference allele enrichment than non-reference allele enrichment in the HepG2 experiments. At 74% of these imbalanced sites the reference allele is also, the major, or more common, allele (allele frequency >0.5; 1000 Genome EUR) and we hypothesized that the increased reference allele enrichment is a function of major allele enrichment. Comparing the number of sites with major allele enrichment to a binomial distribution we found that major allele is significantly enriched (55.6% of sites) at more sites than expected by chance (P=1.8x10^{-97}) .

3.2.6 Allelic imbalance sites are in the same region as predicted binding motifs for imbalanced proteins

We have detected allelic imbalance at sites identified as heterozygous using genotyping arrays and imputation and sites predicted to be heterozygous because sequence reads from ChIP-seq and DNase-seq contain more than one allele. These heterozygous sites identified only in the sequence data may be real variants that not correctly identified by imputation because they are rare or unique to the sequenced individual. We have shown previously that imbalance detection is more accurate at the heterozygous sites identified using imputation than these sites predicted from the sequence data. As such, we created a high confidence set of imbalanced sites containing only allelic imbalances found in HepG2 data at heterozygous sites identified by imputation. Using protein binding motifs identified using ChIP-seq data from the ENCODE project[117], we searched for evidence of direct protein binding to these heterozygous sites. Using FIMO[94] and available motifs, we identified the genomic locations of protein binding motifs ($P<1.0\text{x}10^{-4}$) for a subset (50/61) of all the proteins imbalanced in HepG2. Of sites with evidence of imbalance for these proteins, 80% were within 500 bp of and 15% were located within a predicted binding site for an imbalanced protein (**Figure 3.3A**), suggesting that allelic imbalance is often detected for proteins binding proximal to but not at the heterozygous sites.

3.2.7 Presence of allelic imbalance in one protein of established protein-protein pairs is associated with presences of imbalance in the second protein

We combined these high-confidence imbalance sites with heterozygous sites that have reads mapping but no suggestive evidence of allelic imbalance (8 or more reads mapped per experiment and imbalance $P>0.5$) to create a curated set of heterozygous sites with mapped reads. Of the 75,941 heterozygous sites in our curated set, 19,523 have significant imbalance ($P\leq0.01$) in at least one experiment, and the remaining 56,418 have no evidence of imbalance ($P\geq0.5$). We used this curated set to investigate the relationship between the imbalance statuses of proteins with reads mapping to the same site.

We hypothesize that because of interactions between proteins, the imbalance status (imbalanced or not imbalanced) of proteins with at least 8 reads mapping to heterozygous sites and a binding motif at those sites (motif protein) can influence the imbalance status of other proteins with reads 8 or more reads to the same site. We tested this hypothesis by identifying all heterozygous sites from our curated list that have evidence of a motif protein. For each site, we used the imbalance status of this motif protein and each of the other mapped proteins to classify each protein pair at that site into one of four categories—(i) both proteins were imbalanced; (ii) only the motif protein was imbalanced; (iii) only the other protein was imbalanced; and (iv) neither protein was imbalanced. Using a Fisher's exact test we found a statistical association (Bonferroni corrected for 1,892 tests; P<2.6x10^{-5}) between the imbalance status of the two proteins in 15 protein-protein pairs (**Figure 3.3B**).

In our test, two of the three motif protein-other protein pairs with the most significant associations involved both CTCF and Rad21 (CTCF-Rad21 P=2.8x10^{-9}, Rad21-CTCF P=2.6x10^{-8}), suggesting that when CTCF and RAD21 co-localize to the same region the imbalance status of one protein influences the other. This observation is supported by work demonstrating that CTCF recruits the Rad21-containing cohesin complex to DNase hypersensitivity sites[118]. Although not reaching our threshold for significance, we also detected suggestive evidence that a relationship exists between CTCF and Smc3 (P=4.8x10^{-4}), another cohesin subunit.

3.2.8 Allelic imbalance in DNase-seq data coincides with allelic imbalance in ChIP-seq data at a subset of allelic imbalance sites

With our curated set of heterozygous sites, we next evaluated our ability to identify significant imbalance in ChIP-seq data using DNase-seq data. While reads from one or more ChIP-seq experiments mapped to 85% of sites with DNase-seq reads, DNase-seq reads only mapped to 11% of sites with ChIP-seq reads mapping (**Figure 3.3C**). Focusing only on sites with allelic imbalance, imbalance was detected in one or more ChIP-seq experiments at 71% of sites with allelic imbalance in DNase-seq reads, but imbalance was detected in DNase-seq reads at only 6% of sites with imbalance in ChIP-seq experiments (**Figure 3.4D**). These data suggest that DNase-seq data frequently identifies sites of allelic imbalance in ChIP-seq data, does not comprehensively identify allelic imbalance in protein binding across the genome. These results are likely a function of the sequencing depth and broad signal dispersion of DNase-seq data, and we expect that increased sequencing depth would increase the ability of DNase-seq data to detect additional sites that are also imbalanced in the ChIP-seq data.

3.2.9 Allelic imbalances at published eQTL loci

We searched for evidence of allelic imbalance at eQTL loci identified in in liver[113,114], adipose[115], and pancreatic islet[116] samples. We considered an imbalanced site to be located at an eQTL locus if it was in linkage disequilibrium (LD; r^2>=.7; 1000 Genomes EUR) with any of the 33,959 reported eQTL variants (P<=1x10^{-5}). Of the 22,414 sites with allelic imbalance, 167 were located at loci with evidence of an eQTL, suggesting that differential protein binding may influence gene transcription at these loci (**Table 3.3**, **Appendix 2**).

We identified two sites of allelic imbalance at eQTL loci that also have evidence of an association with cardiometabolic diseases. At rs12091564, we detected enrichment of the C allele over the T allele in multiple factors in HepG2 (CREB1, MYBL2, NR2F2, and TBP). This variant also has reported associations with *NOTCH2NL* transcription in islets[116] (P=2.7x10^{-6}) and coronary artery disease risk[119] (risk allele C; P=2.0x10^{-7}). At rs13356762, we detected enrichment of the G allele over the T allele in TAF1 reads from HepG2 cells. This variant is also associated with *C5orf35* in islets[116] (P=2.5x10^{-8}) and type 2 diabetes risk[119] (risk allele A; P=4.0x10^{-6}). At this same locus, rs185220 (G allele enrichment over A allele) is also associated with *C5orf35* expression in islets but fell just below our LD threshold (r^2 = 0.68) with the type 2 diabetes-associated variant (**Table 3.3, Appendix 2, Appendix 3**).

3.2.10 Allelic imbalances at cardiometabolic phenotype-associated loci

We expanded our search outside of eQTL loci and looked for evidence of allelic imbalance at loci associated with cardiometabolic traits and disease (**Appendix 4**). We detected allelic imbalance at 199 of the 66,513 variants located at cardiometabolic genome-wide association loci reported in the NHGRI catalog (LD r^2>=.7; genome-wide significance P<=5x10^{-8}) (**Appendix 3**). Of these sites, 5 were located at coronary artery disease-associated loci, 11 at type 2 diabetes loci, and 47 at loci associated with lipid levels (HDL-C, LDL-C, triglycerides, and total cholesterol) (**Table 3.4**, **Table 3.5**). Of note, we detected allelic imbalance at two sites, rs6713419 and rs10184004, located near *COBLL1* and *GRB14*, and in LD with variants associated with both type 2 diabetes risk and triglyceride levels. Evidence of allelic imbalance at this and other GWAS loci suggest that differential protein binding at imbalanced sites is contributing to cardiometabolic disease risk and trait measurements.

3.2.11 Experimental conformation of allelic differences in protein binding and enhancer activity at imbalanced sites

Four of the imbalanced sites located at GWAS loci also have experimental evidence demonstrating allelic differences in protein binding and/or enhancer activity (**Figure 3.4A**). At rs4969182 near *PGS1*, we detected enrichment of the A allele in data for 6 proteins including FOXA1 and FOXA2. Experimentally, this same allele shows increased binding of FOXA1 and FOXA2 to the A allele, as well as increased enhancer activity[33]. Likewise at rs4846913, near *GALNT2*, we predicted allelic imbalance in 5 datasets and the enriched allele demonstrated increased protein binding to C/EBPb and enhancer activity[120]. The enriched allele at rs62102718 near *PEPD*, also has experimentally validated allelic differences in protein binding[121].

Finally, we tested rs6813195 for allelic differences in enhancer activity in MIN6, mouse insulinoma, cells using a dual luciferase assay and observed increase enhancer activity for the allele predicted to have increased binding of FOXA2 in human islets (**Figure 3.4B**). Together these experimental data highlight the utility of using allelic imbalance detection to predict allelic differences in protein binding and transcriptional activity at cardiometabolic phenotype-associated loci.

3.3 Discussion

Allelic imbalance detection in quantitative sequence data is a powerful tool for understanding genetic effects on the regulation of gene transcription. We used AA-ALIGNER to detect allelic imbalance in ChIP-seq and DNase-seq data generated in cell lines and primary cells from tissues playing a role in cardiometabolic phenotypes. Imbalance detection in these samples has provided not only biological insights into the regulation of gene transcription at specific cardiometabolic GWAS loci, but also more general insights into protein binding at imbalanced sites.

We found evidence of allelic imbalance at hundreds of loci associated with cardiometabolic traits and diseases. While these imbalanced sites may be located at GWAS loci by chance, it is likely that many of them are playing an active role in regulating the transcription of nearby genes and influencing the associated phenotype. For example, LD data suggests that at two variants near *GRB14*, rs6713419 and rs10184004, the alleles predicted to have increased MAFK binding are on the same haplotype as the alleles associated with both increased triglyceride levels and type 2 diabetes risk. This effect is likely mediated by changes in gene transcription, and *GBR14,* which binds to the insulin receptor and negatively regulates insulin signaling[122], is a strong candidate target. Differential protein binding could influence *GRB14* transcription and ultimately insulin signaling, although experimental validation is needed to confirm differential protein binding and *GRB14* transcription. Allelic differences in enhancer activity have been experimentally observed, however, at three other sites with allelic imbalance, and we are confident that future experimental testing will produce similar evidence for additional imbalanced sites. As the *GRB14* locus demonstrates, predicted imbalances can provide a starting hypothesis for these experiments and expedite experimental exploration of gene transcription regulation at GWAS loci.

In addition to GWAS loci, we also found allelic imbalance at sites associated with gene expression. One variant with allelic imbalance, rs12091564 is associated with allelic differences in *NOTCH2NL* transcription in islets as well as coronary artery disease risk. The Notch signaling plays a role in cardiovascular disease[123], making it plausible that differential regulations of *NOTCH2NL* by rs12091564 influences coronary artery disease risk. Two other imbalance sites, rs13356762 and rs185220 are associated with *C5orf35* expression and T2D. This gene encodes *SETD9* and although it is unclear what role this protein might play in T2D risk, our allelic imbalance results provide a candidate variant to test for differences in regulatory activity. We have additionally identified allelic imbalance at eQTLs outside of GWAS loci that may not be immediately applicable in understanding the genetic effects on cardiometabolic phenotypes, but could be important for understanding genetic effects on gene transcription in general.

In addition to providing candidate regulatory variants for experimental study, our analyses have provided us with some insights into the mechanics of protein binding at sites of allelic imbalance. First, we observed enrichment of reads containing the major allele at more imbalanced sites than expected by chance, suggesting that variants promoting increased protein binding may be evolutionarily favored, or conversely, variants disrupting binding disfavored. Second, we have used allelic imbalance to perform a preliminary exploration of the binding relationship of proteins co-localized to the same heterozygous site. We found evidence of an association between the presence of imbalance in CTCF and cohesin subunits Rad21 and SMC. This finding is supported by other work demonstrating CTCF, Rad21, and SMC co-localization in HepG2 cells[124] and a direct interaction between CTCF and Rad21[118] . While our analysis offers preliminary evidence of direct binding relationships between proteins, it is important to note that it may be limited by many factors such as accuracy of binding motif locations, sequencing depth, and ChIP-seq data availability.

Data availability is one of the greatest limiting factors of imbalance detection. Our analyses were particularly limited by the small number of ChIP-seq samples generated in pancreatic islets and adipose tissue. It is likely that in these tissues we were unable to detect allelic imbalance at many sites influencing gene transcription at cardiometabolic GWAS loci. We were limited further because allelic imbalance detection can only be done at heterozygous sites. Even with the abundance of data from a liver cell line, we failed to detect allelic imbalance at sites with documented allelic differences in protein binding in liver because these sites are homozygous in HepG2 cells[23,27,30]. Despite this limitation, allelic imbalance detection is very useful even in only a single dataset. Analyzing quantitative sequence data from more than one individuals would help to overcome this limitation, but analyzing large numbers of ChIP-seq datasets in multiple individuals can be resource prohibitive.

DNase-seq data can identify the binding sites of many transcription factors in a single assay and is an attractive option for identifying protein binding sites in a population of individuals[125]. While we detected allelic imbalance in ChIP-seq data at a majority of sites imbalanced in DNase-seq data, we only predicted a small fraction of sites imbalanced in ChIP-seq data using DNase-seq data. DNase-seq has a more disperse signal than most ChIP-seq data and requires a much deeper sequencing depth to achieve the same signal intensity found in ChIP-seq data with fewer reads. As signal intensity was highly correlated with imbalance detection, it is likely that with greater sequencing depth DNase-seq data would be able to identify a greater proportion of ChIP-seq imbalances. Protein binding to DNA creates a localized site of protection from DNase-seq reads, or footprints[125]. Reduced DNase-seq read coverage in footprints further limits imbalance detection in DNase-seq data at heterozygous sites directly bound by protein. Additionally, the number of cells required to generate adequate sequencing depth with DNase-seq can be prohibitive when using a limited number of primary cells. ATAC-seq, similar to DNase-seq, requires fewer cells and may reduce this limitation, but further study is needed to assess the efficiency and accuracy of allelic imbalance in that data.

We have limited our analyses to ChIP-seq and DNase-seq data generated in a single liver cell line, pancreatic islet samples from 12 individuals, and two adipose cell lines. Additional protein ChIP-seq and DNase-seq data exists for other liver and pancreatic cell lines as well as primary cells and samples from these tissues. Additionally, RNA-seq, FAIRE-seq and histone modification ChIP-seq are also available for samples from these and other samples related to cardiometabolic phenotypes. As we expand allelic imbalance identification into this additional data, we expect to find additional evidence of allelic imbalance at cardiometabolic phenotype-associated loci and gain further insight into transcriptional activity at these loci.

3.4 Conclusion

In conclusion, we have identified allelic imbalance in ChIP-seq and DNase-seq data at thousands of genomic sites. These imbalances, which predict allelic differences in protein binding were identified in cultures of cell lines or primary cells from liver, pancreas, and adipose tissues and can provide direct insight into gene transcription regulation at cardiometabolic phenotype-associated loci. The hundreds of allelic imbalance sites we have identified at eQTL and/or GWAS loci may influence gene transcription at these loci and are prime candidates for future experimental analyses. We have documented allelic imbalance at sites for which allelic differences in protein binding and/or enhancer activity have been experimentally observed, including novel evidence of allelic differences in enhancer activity of at rs6813195. This study provides not only thousands of candidate regulatory sites with predicted allelic differences in protein binding, but also examples of the biological insights into transcription that can be gained from examining these sites.

3.5 Methods

3.5.1 Genotype imputation

Genotypes for HepG2, generated by the HudsonAlpha Institute of Biotechnology using the Illumina Human-1MDuo BeadChip array, were download from the UCSC genome browser[41]. We imputed autosomal genotypes for HepG2 and 52 other samples from the ENCODE project using MaCH-Admix[84] with default parameter settings and the reference panel from the 1000 Genomes Project Phase I version 3 (2012-03-14 release). Chromosome X genotype data was first pre-phased using MaCH[101] with options --states 500 and --rounds 400 and then imputed using minimac[102] with options --state 10 and --rounds 10. Imputation quality Rsq was used to filter variants post imputation as previously reported[103].

3.5.2 Sequence mapping and imbalance detection in quantitative sequence data

ChIP-seq and DNase-seq data from the ENCODE consortium[40] was downloaded from the UCSC genome browser[41] and data from other sources was obtained from the short read archive (SRA)[69] . Sequence reads were filtered and aligned, the alignments processed, and allelic imbalance detected using AA-ALIGNER as previously described, but with a few modifications. Before alignment we filtered sequences using TagDust[126] and the following parameters `-q –f 0.001 -s` to remove known adapter sequences. Rather than using MarkDuplicates (Picard Tools[127]) to reduce PCR artifacts, we only align the 5 reads with highest quality when a datasets contains more than 5 reads with same sequence.

Sequences from primary islets samples, hASC, and SGBS were aligned to a major allele reference sequence created by changing bases of hg19 at common variants to match the major allele. (1000 Genomes EUR MAF>.05). Data from HepG2 cells were aligned to a HepG2-specific reference genome created by changing each base in the major allele reference to match the allele with the highest predicted dosage at imputed variants.

3.5.3 Identifying motif occurrences

We searched for occurrences of transcription factor binding motifs previously identified using ChIP-seq data[117] within the regions ±500 bp of sites of allelic imbalance. We looked for motif occurrences in two sets of sequences, one containing the reference allele and the other containing the non-reference allele, using the motif scanner FIMO[94] with the parameters `--max-strand –max-stored-scores 1000000 –no-qvalue`.

3.5.4 Cell Culture

Mouse derived insulinoma MIN6 cells were cultured in DMEM (Sigma), supplemented with 10% FBS, 1 mM sodium pyruvate, 0.1 mM β-mercaptoethanol and maintained at 37°C with 5% $CO_{2.}$

3.5.5 Generation of luciferase reporter constructs, transient DNA transfection and luciferase reporter assays

To test the allele specific transcriptional activity, we PCR-amplified a 201 bp fragment (chr4: 153520425-153520625) surrounding the SNP rs6813195 (Forward primer: GGGAGAGGAAGCAAGTAAACAAG, reverse primer: CAGGCAATCTGTCCACCTC) from DNA of individuals homozygous for both major and minor alleles. Restriction sites for KpnI and XhoI were added to primers during amplification, and the resulting PCR products were digested with KpnI and XhoI and cloned in both orientations into the multiple cloning sites of the minimal promoter-containing firefly luciferase reporter vector pGL4.23 (Promega, Madison, WI). Fragments are designated as ‘forward’ or ‘reverse’ based on their orientation with respect to the genome. Three to six independent clones for each allele for each' orientation were isolated,

Approximately 200,000 MIN6 cells per well were seeded in 24-well plates. At 80% confluency, cells were co-transfected with luciferase constructs (250 ng per well) and *Renilla* control reporter vector (phRL-TK, Promega) (80 ng per well) using Lipofectamine LTX (Invitrogen). Transfected cells were incubated at 37°C with 5% CO_2 for 48 hours then lysed with passive lysis buffer (Promega), and luciferase activity was measured using the Dual-luciferase assay system (Promega). To control for transfection efficiency, raw values for firefly luciferase activity were divided by raw *Renilla* luciferase activity values, and fold change was calculated as normalized luciferase values divided by pGL4.23 minimal promoter empty vector control values. Data are reported as the fold change in mean (± SE) relative luciferase activity per allele. A two-sided *t*-test was used to compare luciferase activity between alleles.

Table 3.1 Summary of Detected Allelic Imbalances

		Samples by Tissue Type		
	All Samples	**Liver**	**Pancreatic Islets**	**Adipose**
Samples	**15**	**1**	**12**	**2**
Total Experiments	**91**	**70**	**17**	**4**
ChIP-seq datasets	90	69	17	4
DNase-seq datasets	1	1	0	0
Sites with allelic imbalance[a]	**22,414**	**21,308**	**1,074**	**75**
Sites with imbalance in >= 1 experiment[a]	**6,338**	**6,292**	**157**	**21**

A sample is a cell line or primary cell isolated from a single individual. DNase-seq datasets from the same sample or ChIP-seq datasets for the same protein from the same sample are represented as a single experiment. [a]For some sites allelic imbalance was detected in samples from more than one tissue classification

Table 3.2 Allele-aware alignments with complete genotypes (GSNAP) vs no genotype information (BWA)

	C/EBPB		CREB1		TAF1	
	Standard	Complement[a]	Standard	Complement[a]	Standard	Complement[a]
Reads mapped uniquely	34,806,038	34,806,038	48,939,432	48,939,435	23,565,876	23,565,876
Reads at heterozygous sites	833,781	833,781	1,613,584	1,613,587	1,197,696	1,186,891
Reference allele	457,433	376,348	822,922	790,665	312,267	288,331
Non-reference allele	376,348	457,433	790,662	822,922	288,331	312,267
Allelic Imbalance Sites Identified[b]	778	778	1204	1204	310	310
Reference allele	418	360	695	509	187	123
Non-reference allele	360	418	509	695	123	187

[a]Alignment reference contained the non-reference allele of heterozygous sites identified by imputation [b]Imbalances at known heterozygous sites (binomial p-value<.01)

Table 3.3 Site of allelic imbalance with published evidence of eQTL (P<1.0x10^{-10})

Imbalanced		Enriched			
Variant	**Position**	**Allele**	**Imbalanced Protein(s)**	**Other Allele**	**eQTL Genes**
rs1494813	chr1:45957290	C	(L) MXI1	T	(I) *CCDC163P*[116]
rs12127787	chr1:89458761	C	(L) CREB1,USF1[b], USF2[b], YY1	T	(I) *GBP3*[116]
rs61844237	chr1:245133662	G	(L) DNASE	C	(I) *EFCAB2*[116]
rs1554612	chr2:48827497	C	(L) CTCF,FOXA2,RAD21	T	(A) *STON1*[115]
rs2070063	chr2:64862055	A	(L) MAFK	G	(A) *SERTAD2*[115]
rs2364723	chr2:178126546	C	(L) MBD4	G	(L) *NFE2L2*[114]
rs9841194	chr3:125635739	T	(L) POL2	C	(I) *LOC100125556*[116]
rs7661077	chr4:7219889	C	(L) C/EBPB	T	(L) *SORCS2*[114]
rs10030238	chr4:141808805	A	(L) HNF4G	G	(A) *RNF150*[115]
rs2227426	chr4:155493171	G	(L) POL2	A	(L) C9orf66[114]
rs3195676	chr5:34008100	C	(L) BHLHE40, MAX,TAF1	T	(A) *AMACR*[115]
rs185220	chr5:56205357	G	(L) DNASE	A	(I) *C5orf35*[116]
rs3132555[a]	chr6:31082910	G	(L) RAD21	C	(I) *CDSN*[116]
rs3094209[a]	chr6:31089982	G	(L) C/EBPB	A	(I) *CDSN*[116]
rs9271092	chr6:32576296	A	(L) RAD21	G	(I) *HLA-DRB1*[116]
rs9271093	chr6:32576341	G	(L) CTCF	A	(I) *HLA-DRB1*, *HLA-DRA*, *HLA-DRB5*[116]
rs9271094	chr6:32576347	G	(L) CTCF	C	(I) *HLA-DRB1*[116]
rs9271096	chr6:32576426	A	(L) CTCF	G	(I) *HLA-DRB1*[116]
rs539298	chr6:160770360	G	(L) BHLHE40,C/EBPB,CREB1,DNASE,ELF1, HNF4A, HNF4G, JUND,MAX,MYBL2,NFIC, NR2F2,P300, POL2,RAD21	A	(L) *SLC22A3*[113]
rs8200	chr7:75696606	G	(L) POL2	C	(L) *AKs022137*[114]
rs6985299	chr8:71613079	T	(L) MAX	C	(I) *XK92*[116]
rs11985375	chr8:71613472	G	(L) CREB1,MAX	A	(I) *XK92*[116]
rs17141322	chr10:17604700	A	(I) PDX1	C	(I) *ST8SIA6*[116]
rs9787897	chr11:74659302	T	(L) FOXA2	A	(I) *XRRA1*[116]
rs567956	chr11:74659779	C	(L) POL2	T	(I) *XRRA1*[116]
rs2165163	chr11:74660143	C	(L) DNASE,MAX	G	(I) *XRRA1*[116]
rs933462	chr12:9103665	G	(L) HDAC2	T	(A) *KLRG1*[115]
rs9925556	chr16:2880105	T	(L) FOXA2,DNASE	C	(L) *ZG16B*[114]
rs1981760	chr16:50723074	C	(L) EZH2	T	(L) *CARD15*[114]
rs16949649	chr17:49230308	T	(L) C/EBPB	C	(A) *NME1*[115]
rs2598414	chr17:74067099	C	(L) BHLHE40	T	(A) *SRP68*[115]
rs2376585	chr17:76417883	T	(L) CEBPA,CTCF,DNASE,FOXA1, FOXA1,MAX,NR2F2, ZBTB33,ZEB1	C	(A) *DNAH17*[115]
rs8101689	chr19:30185697	A	(L) HDAC2	G	(A) *C19orf12*[115]
rs1343703	chr19:49955155	G	(I) FOXA2 (L) ARID3A[b]	C	(L) *NOP17*[114]
rs562954	chr20:48092076	G	(L) CEPBP	A	(L) *AK055386*,*KCNB1*[114]
rs4828057	chrX:100006043	A	(I) MAFB	C	(A) *SYTL4*[115]

Variants with allelic imbalance and a reported association with gene expression (P<1x10^{-10}) are shown. [a]Variant is in perfect LD (r^2=1) with the eQTL variant rs3130981 [b]Protein is enriched for the other allele rather than enriched allele. Allelic imbalance and eQTLs were identified in (A) adipose tissue, (I) pancreatic islets, or (L) liver tissues

Table 3.4 Sites of allelic imbalance located at cardiometabolic disease-associated loci

		Allelic Imbalance				Disease Association				
Locus	Imbalanced Variant	Position	Enriched Allele	Imbalanced Protein(s)	Other Allele	Disease	Reported Variant[a]	LD r^2[b]	Coupling[c]	Risk Allele
MIA3	rs4846770	chr1:222795569	G	(L)CEBPD	C	CAD	rs17465637[128,129]	0.95	G,C	C
HCG27	rs6921948	chr6:31171257	A	(I)FOXA2	C	CAD	rs3869109[130]	0.73	C,G	G
YP17A1,CNNM2,NT5C2	chr10:104692633	chr10:104692633	A	(L)CEBPA	C	CAD	rs12413409[128,131,132]	0.94	C,G	G
	chr10:104952499	chr10:104952499	C	(L)DNASE	T	CAD	rs12413409[128,131,132]	0.94	C,G	G
ADAMTS7	rs11856536	chr15:79094325	A	(L)HDAC2	G	CAD	rs3825807[128]	0.98	A,A	A
RBMS1,ITGB8	rs6706545	chr2:161181478	A	(L)C/EBPB, FOXA2	T	T2D	rs7593730[133]	0.98	A,C	C
	rs10929982	chr2:161236277	T	(L)C/EBPB	C	T2D	rs7593730	0.77	T,C	C
COBLL1,GRB14	rs6713419	chr2:165508300	T	(L)MAFK	C	T2D	rs3923113[134,135]	0.97	T,A	A
	rs10184004	chr2:165508389	C	(L)MAFK	T	T2D	rs3923113	0.85	C,A	A
TMEM154	rs6813195	chr4:153520475	C	(I)FOXA2	T	T2D	rs6813195[135]	-	-	C
CDC123,CAMK1D	rs34428576	chr10:12281111	A	(L)CEBPA,C/EBPB,CEBPD,CREB1,DNASE, FOXA1, HDAC2,HNF4A, JUND, MAX, NFIC,NR2F2,P300,RAD21, ZBTB7A	G	T2D	rs12779790[136]	0.72	A,G	G
HHEX	rs4933736	chr10:94471595	T	(L)FOXA2	C	T2D	rs5015480[136–140]	0.74	C,C	C
FITM2,RHDML,HNF4A	rs4812816	chr20:42930872	C	(L)MAZ	A	T2D	rs6017317[141]	0.74	A,G	G
	rs6065723	chr20:42956922	C	(L)MAFF,MAFK	T	T2D	rs6017317	0.77	T,G	G
SL30A8	rs35859536	chr8:118191475	C	(I)PDX1	T	T2D	rs3802177[135,139]	0.99	C,G	G
GLIS3	rs57884925	chr9:4285119	C	(L)MAFK	G	T2D	rs7041847[135,141]	0.93	G,A	A

[a]For associations with more than one reported variant the variant in highest linkage disequilibrium with the imbalance site is shown. [b] Linkage Disequilibrium r^2 was calculated for EUR samples in 1000 Genomes Phase1v3 [c]Allele coupling calculated in 1000 Genomes EUR samples written as: imbalanced variant allele, reported variant allele. (I) for islets and (L) for liver indicates the tissue containing the imbalance. CAD, coronary artery disease; T2D, type 2 diabetes

Table 3.5 Sites of allelic imbalance located at lipid trait-associated loci

			Allelic Imbalance			Trait Association				
Locus	Imbalanced Variant	Position	Enriched Allele	Imbalanced Factor(s)	Other Allele	Trait	Reported Variant[a]	LD r^2[b]	Allele Coupling[c]	Inc. Allele[d]
LDLRAP1, TMEM57	rs9438904	chr1:25756860	T	(L)YY1	C	LDL-C	rs12027135[107,142]	0.98	C,T	T
						TC	rs12027135[107,142]	0.98	C,T	T
ANGPTL1	rs17361251	chr1:178520577	A	(L)CEBPA,C/EBPB,CEBPD,FOXA1,FOXA2,HNF4A,HNF4G ,MAX,NFIC,NR2F2,P300,SP1,ZBTB7A	C	HDL-C	rs4650994[107]	1.00	C,G	G
	rs17276513	chr1:178520604	A	(L)CEBPA,C/EBPB,DNASE,FOXA1,FOXA2,HDAC2,HNF4A, HNF4G,MAX,NR2F2,SP1,P300,ZBTB7A	T	HDL-C	rs4650994	0.99	T,G	G
	rs17276527	chr1:178520680	A	(L)CREB1,DNASE,FOXA1,FOXA2,HDAC2,HNF4A,HNF4G, NFIC,MAX,P300,SP1,ZEB1	G	HDL-C	rs4650994	1.00	G,G	G
GALNT2	rs4846913	chr1:230294715	A	(I)MAFB (L)CEBPA,C/EBPB,CEBPD,NR2F2	C	HDL-C	rs4846914[107,142–144]	1.00	A,A	A
						TG	rs4846914[107,142,143]	1.00	C,G	G
IRF2BP2	rs526936	chr1:234852204	A	(I)FOXA2 (L)POL2[e]	G	LDL-C	rs514230[107,142]	0.92	A,T	T
						TC	rs514230[107,142]	0.92	A,T	T
	rs556107	chr1:234853059	C	(L)HEY1	T	LDL-C	rs514230	0.93	T,T	T
						TC	rs514230	0.93	T,T	T
APOB	rs1367117	chr2:21263900	G	(L)MAX,HEY1	A	LDL-C	rs1367117[107,142]	-	-	A
						TC	rs1367117[107,142]	-	-	A
APOB	rs312983	chr2:21378580	A	(L)FOXA1	C	LDL-C	rs562338[145,146]	0.72	C,G	G
	rs312984	chr2:21378778	C	(L)ARID3A,FOXA1,FOXA2,HNF4A,MAX,NFIC, RAD21,ZEB1	T	LDL-C	rs562338	0.73	T,G	G
	rs312985	chr2:21378805	A	(L)CREB1,FOXA1,FOXA2,HDAC2,HNF4A,HNF4G,MYBL2, NRSF,P300,SP1,ZEB1	G	LDL-C	rs562338	0.73	G,G	G
	rs1652418	chr2:21388456	T	(L)MAZ,SMC3,RAD21	C	LDL-C	rs562338	0.72	C,G	G
	rs544039	chr2:21398985	C	(L)CTCF,RAD21	A	LDL-C	rs562338	0.71	A,G	G
GCKR	rs1260326	chr2:27730940	C	(L)CTCF	T	TC	rs1260326[107,142,143,147,148]	-	-	T
						TG	rs1260326[107,142]	-	-	T
	rs780095	chr2:27741105	G	(L)FOXA1	A	TC	rs1260326	0.81	A,T	T
						TG	rs1260333	0.98	A,T	T
	rs780094	chr2:27741237	C	(L)C/EBPB,FOXA2,MAFK*,MAX,NR2F2,NRSF,ZEB1	T	TC	rs1260326	0.91	T,T	T
						TG	rs780094[143,146,149,150]	-	-	T
EHBP1	rs2136737	chr2:62969310	G	(L)FOXA2,HNF4A*	C	LDL-C	rs2710642[107]	0.77	G,A	A
	rs1553832	chr2:63013515	G	(L)BHLHE40	C	LDL-C	rs2710642	0.80	C,A	A
	rs56373728	chr2:63095792	G	(L)POL2	A	LDL-C	rs2710642	0.93	A,A	A
	rs2710642	chr2:63149557	G	(L)POL2	A	LDL-C	rs2710642	-	-	A

[a]For associations with more than one reported variant the variant in highest linkage disequilibrium with the imbalance site is shls 2-own· [c] Linkage Disequilibrium r^2 was calculated for EUR samples in 1000 Genomes Phase1v3 [b]Allele coupling calculated in 1000 Genomes EUR samples written as Imbalanced variant allele, Reported variant allele. [c]The increasing allele is associated with higher trait levels. [d]Inc. allele is associated with increase in trait measurement. [d]Indicates factor is enriched for other allele rather than enriched allele. (I) for islets and (L) for liver indicates the tissue containing the imbalance. HDL-C, high density lipoprotien cholesterol; LDL-C, low density lipoprotein cholesterol; TG, triglycerides; TC, total cholesterol

Table 3.5 cont'd. Sites of allelic imbalance located at lipid trait-associated loci

s			Allelic Imbalance			Trait Association				
	Imbalanced		Enriched		Other					Inc.
Locus	Variant	Position	Allele	Imbalanced Factor(s)	Allele	Trait	Reported Variant[a]	LD r^{2b}	Allele Coupling[c]	Allele[d]
COBLL1, GRB4	rs6713419	chr2:165508300	T	(L)MAFK	C	TG	rs10195252[142]	0.87	T,T	T
	rs10184004	chr2:165508389	C	(L)MAFK	T	TG	rs10195252	0.99	C,T	T
FN1	rs1250240	chr2:216295014	G	(L)POL2	A	LDL-C	rs1250229[107]	0.84	G,C	C
	rs1250241	chr2:216295312	A	(L)POL2	T	LDL-C	rs1250229	0.84	A,C	C
	rs1250244	chr2:216297796	C	(L)FOXA2,POL2	G	LDL-C	rs1250229	0.86	C,C	C
	rs1250258	chr2:216300185	T	(L)NFIC,POL2	C	LDL-C	rs1250229	0.93	T,C	C
	rs1250259	chr2:216300482	A	(L)NR2F2	T	LDL-C	rs1250229	0.94	A,C	C
GSK3B	rs6800622	chr3:119580678	C	(L)MAFK	A	HDL-C	rs6805251[107]	0.98	A,T	T
PP1R3B	rs6984305	chr8:9178268	T	(L)CTCF	A	HDL-C	rs9987289[107,142]	0.80	T,G	G
						LDL-C	rs9987289[107,142]	0.80	T,G	G
						TC	rs9987289[107,142]	0.80	G,T	G
ABCA1	rs4149269	chr9:107647121	G	(L)POL2	A	HDL-C	rs4149268	0.98	A,C	C
ZNF259,APOA1, APOC3,APOA4, APOA5,BUD13	rs180351	chr11:116607641	T	(L)CTCF	C	TG	rs603446[151]	0.91	C,C	C
UBASH3B	rs11218752	chr11:122552600	C	(L)CTCF	T	HDL-C	rs7941030[107,142]	0.81	T,C	C
						TC	rs7941030[107,142]	0.81	T,C	C
CETP	rs12720926	chr16:56998918	A	(L)DNASE	G	HDL-C	rs1532624[150,152]	0.94	G,A	A
DPEP3	rs7199443	chr16:67841129	G	(L)MAX	T	HDL-C	rs255049[153]	0.72	G,G	G
	rs7196789	chr16:67927124	C	(L)YY1	T	HDL-C	rs255049	0.77	T,G	G
	rs1134760	chr16:67964203	C	(L)POL2	T	HDL-C	rs255049	0.81	C,G	G
	rs20549	chr16:67969930	G	(L)POL2	A	HDL-C	rs255049	0.81	G,G	G
	rs1109166	chr16:67977382	C	(L)CREB1, HNF4A,HNF4G,FOXA1, FOXA2,NR2F2	T	HDL-C	rs255049	0.84	C,G	G
MPP3	rs17742347	chr17:41846468	C	(L)POL2	T	TG	rs8077889[107]	0.90	T,C	C
	rs17674998	chr17:41879544	A	(L)ZBTB33	G	TG	rs8077889	0.99	G,C	C
	rs9901676	chr17:41911818	T	(L)EZH2	C	TG	rs8077889	0.92	C,C	C
PGS1	rs4969182	chr17:76393030	T	(L)C/EBPB,FOXA1,FOXA2,MAX,MYBL2,NR2F2	C	HDL-C	rs4129767[107,142]	0.96	C,A	A
	rs4969183	chr17:76393372	A	(L)BHLHE40	G	HDL-C	rs4129767	0.96	G,A	A
INSR	rs10410204	chr19:7224350	C	(L)FOXA1,FOXA2	T	TG	rs7248104[107]	0.98	T,G	G
	rs7248104	chr19:7224431	A	(L)FOXA1,FOXA2,MAX,NR2F2,YY1	G	TG	rs7248104	-	-	G
PEPD	rs62102718	chr19:33891013	A	(L)HNF4G	T	HDL-C	rs731839[107]	0.76	A,A	A
						TG	rs731839[107]	0.76	A,A	A
SPTLC3	rs1321940	chr20:12959885	G	(L)FOXA1	A	LDL-C	rs364585[107]	0.99	G,G	G

[a]For associations with more than one reported variant the variant in highest linkage disequilibrium with the imbalance site is shown. [c] Linkage Disequilibrium r^2 was calculated for EUR samples in 1000 Genomes Phase1v3 [b]Allele coupling calculated in 1000 Genomes EUR samples written as Imbalanced variant allele, Reported variant allele. [c]The increasing allele is associated with higher trait levels. [d]Inc. allele is associated with increase in trait measurement. [d]Indicates factor is enriched for other allele rather than enriched allele. (I) for islets and (L) for liver indicates the tissue containing the imbalance. HDL-C, high density lipoprotien cholesterol; LDL-C, low density lipoprotein cholesterol; TG, triglycerides; TC, total cholesterol

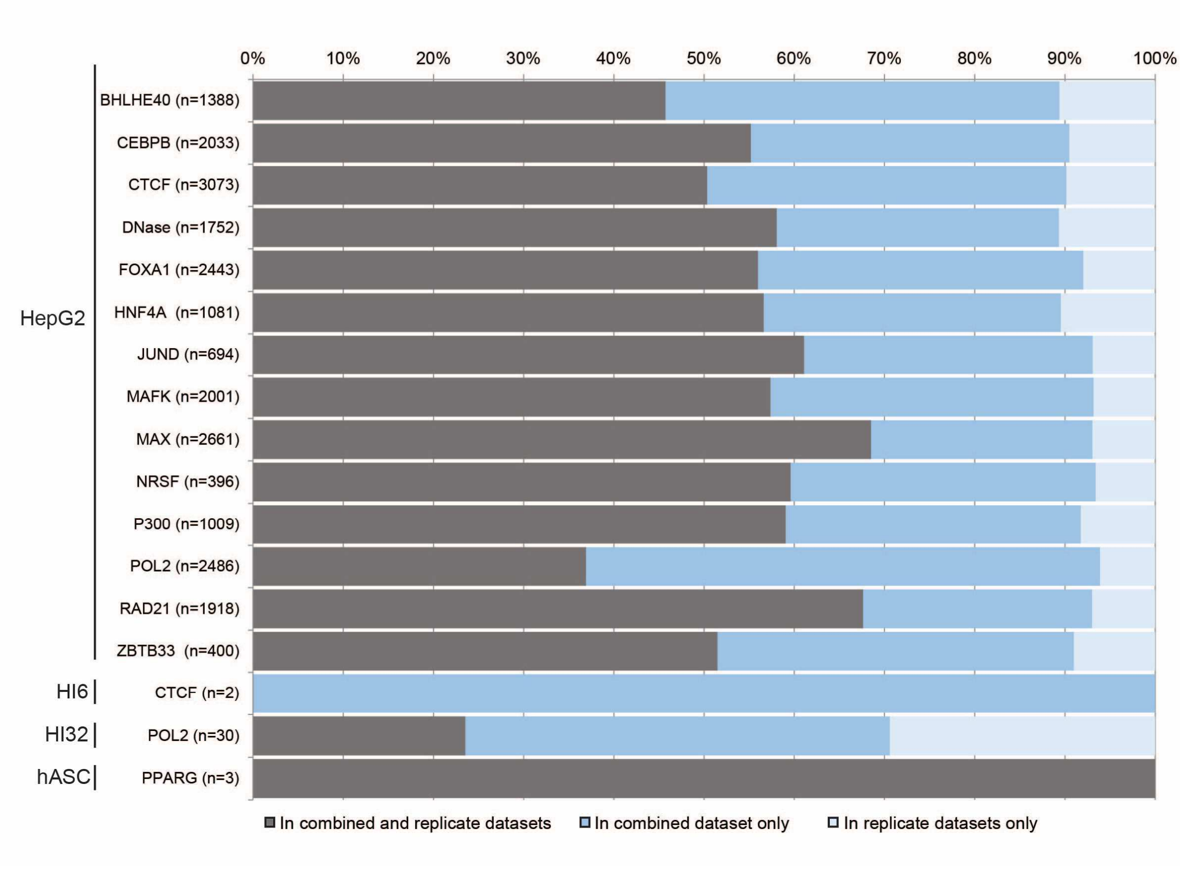


Figure 3.1 Concordance between replicate datasets. Concordance between allelic imbalance detection in combined and individual replicate ChIP-seq datasets for the same protein from the same sample or DNase-seq datasets from the same sample. n represents the number of sites of allelic imbalance in the individual and replicate datasets.

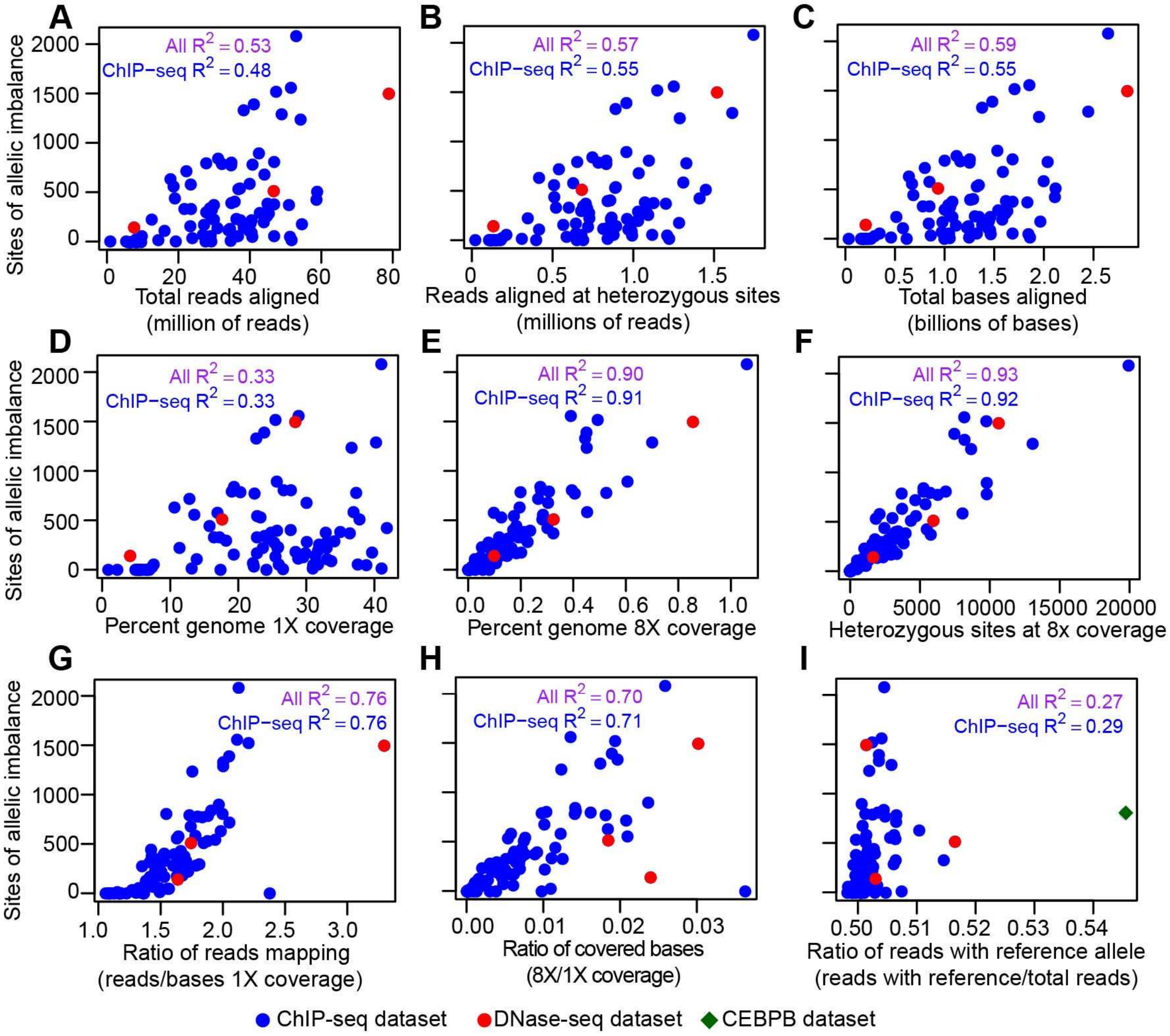


Figure 3.2 Percent of genome with 8X coverage is correlated with imbalance detection. . Measures of sequencing depth (A-C), genomic coverage (D-F), and combinations of measurements from both groups (G-I) are plotted vs the number of sites of allelic imbalance detected. Each circle represents on dataset. Pearson correlation R^2 values when considering only ChIP-seq datasets and all datasets are displayed.

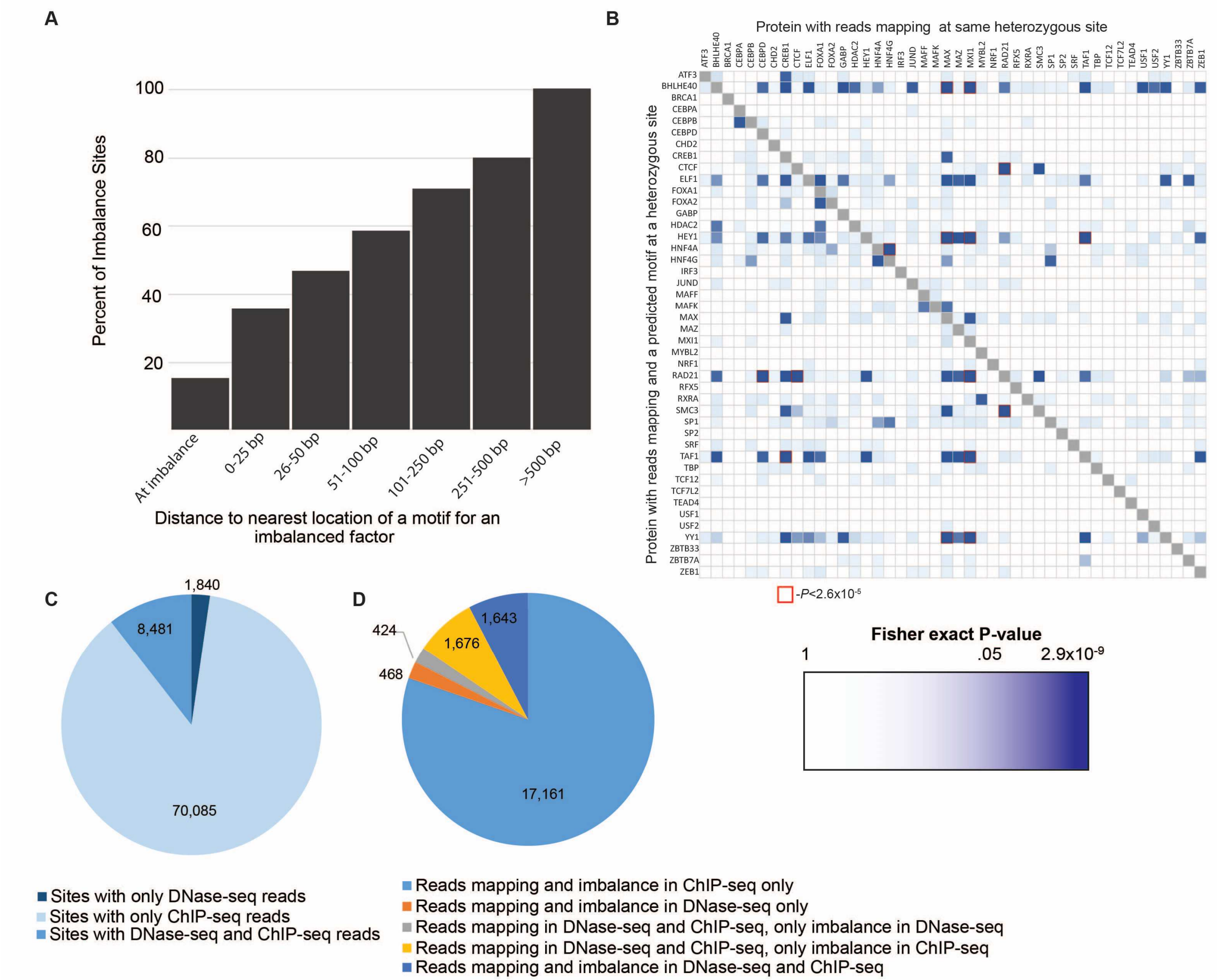


Figure 3.3 Co-occupancy of binding motifs and reads from DNAse-seq and ChIP-seq experiments at allelic imbalance sites.

Figure 3.3 Co-occupancy of binding motifs and reads from DNAse-seq and ChIP-seq experiments at allelic imbalance sites. (A) Transcription factor binding motifs were located near allelic imbalance sites and the distance to the nearest motif of an imbalance factor was calculated. (B) Comparing the imbalance status (imbalanced or not) of proteins with reads mapping to and a binding motif overlapping heterozygous sites (row) and other factors with reads mapping to the same site (column), we tested for association between the imbalance status of the two proteins. Cells are shaded by Fisher exact test P-values and pairs reaching significance at a Bonferroni correct P-value ($P<2.6x10^{-5}$) are outlined in red. Co-occurrence of DNase-seq and ChIP-seq reads at (C) heterozygous sites or (D) imbalanced sites.

A

		Allelic Imbalance		EMSA		Luciferase	
Locus	SNP	All.	Protein(s)	Allele	Protein	Allele	Disease/Trait
PGS1	rs4969182	T	(L) CEBPB,FOXA1,FOXA2,MAX,MYBL2,NR2F2	T ↑	FOXA[33]	T ↑	HDL-C,TG
GALNT2	rs4846913	A	(I) MAFB (L) CEBPA,CEBPB,CEBPD,NR2F2	A ↑	CEBPB[120]	A ↑	HDL-C, TG
PEPD	rs62102718	A	(L) HNF4G	A ↑[121]	-	-	HDL-C, TG
TMEM154	rs6813195	C	(I) FOXA2	-	-	C ↑	T2D

All, enriched allele; HDL-C, high density lipoprotein cholesterol; TG, triglycerides; T2D, Type-2 diabetes

B

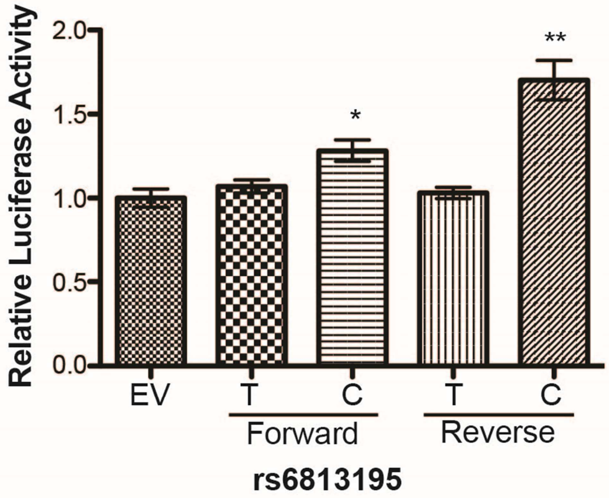


Figure 3.4. Experimental evidence of allelic differences in protein binding and transcriptional activity at sites of allelic imbalance.

Figure 3.4 Experimental evidence of allelic differences in protein binding and transcriptional activity at sites of allelic imbalance. (A) At cardiometabolic GWAS loci, we detected allelic imbalance at four sites for which the enriched allele has been observed to have increased binding in an electrophoretic mobility shift assay (EMSA) and/or increased enhancer activity in a dual luciferase assay. Proteins with evidence of a supershift are listed. (B) The C allele of rs6813195 displayed increased enhancer activity over the T allele in both the forward and reverse orientations in MIN6 cells. Error bars represent the standard error of mean across 3-6 biological replicates per allele in each orientaion. *P<.05 **P<.001

CHAPTER 4: CONCLUSIONS

4.1 Introduction

In this work, I have described the detection of allelic imbalances in quantitative sequence data and demonstrated how these analyses provide insights about the regulation of gene transcription, particularly at cardiometabolic phenotype-associated loci. Here I provide a brief summary highlighting important findings and discuss the impact of these findings on cardiometabolic disease risk factor research, as well as research focused on other complex phenotypes. I discuss the advantages and limitations of allelic imbalance detection, and the role of imbalance detection in population studies utilizing quantitative sequence data.

4.2 Overview of findings

Chapter 2 describes our allele-aware alignment pipeline, AA-ALIGNER, and evaluated the impact of common analytical and experimental decisions on pipeline performance. Most importantly, we demonstrated that under a variety of conditions and in multiple datasets, AA-ALIGNER removes mapping biases and accurately identifies allelic imbalance even when limited or no sample genotypes are available. We also reported that allelic imbalance detection is much more accurate at sites identified as heterozygous during imputation than at sites predicted to be heterozygous solely from the sequence data. Surprisingly, this increased accuracy was observed when we predicted heterozygous sites at common variants compared to uncommon or rare variants. The accuracy of imbalance detection at imputed heterozygous sites and common sites remained high across a variety of parameters such as read length, number of mismatches allowed, number of reads required to be predicted as heterozygous, and imputation quality Rsq threshold. These results allowed us to expand, with confidence, allelic imbalance detection in samples directly relevant to cardiometabolic phenotypes, but lacking complete genotype information.

Chapter 3 discusses imbalance detection in samples from a liver cell line, primary pancreatic islet cells, and two adipose cell lines. In 91 experiments from these samples, we detected over 22,000 sites of allelic imbalance, and 29% of these sites have evidence of imbalance in multiple experiments. Only considering sites with evidence of imbalance for a protein with binding motifs available, we found evidence of a binding motif within 500 bp of the imbalanced site at 80% of sites, suggesting that these imbalanced factors are binding at or near sites of allelic imbalance. A subset of sites of allelic imbalance are located at an eQTL locus, a cardiometabolic phenotype-associated locus, or in a few cases both, suggesting that differential protein binding at these sites may be influencing gene transcription and the studied phenotype. We document experimental data that confirms these connections at a handful of sites, supporting the hypothesis that imbalanced sites at other loci are strong candidates for experimental testing.

4.3 Immediate impact on genetic studies of cardiometabolic phenotypes

Our findings have immediate implications on current research on the genetic factors underlying cardiometabolic diseases and traits. Primarily, our imbalance detections have identified over 300 sites of allelic imbalance at GWAS loci, effectively identifying the same number of candidate variants to test experimentally for allelic differences in protein binding, enhancer activity, and gene transcription. At the *HHEX* locus, previously studied in our lab, initial efforts to identify candidate variants based on their location in open chromatin regions and transcription factor ChIP-seq peaks failed to identify variants with allelic differences in protein binding and enhancer activity. One variant, rs4933736, was not selected as a candidate in this initial survey, but has evidence of allelic imbalance in our data and is now a top candidate at this locus.

Second, we have identified over 160 sites of imbalance with a reported association to gene expression, but not cardiometabolic phenotypes. Though an association has not been reported in the GWAS catalog, some of the genes at these eQTLs may play a role in cardiometabolic phenotypes, making these sites of allelic imbalance good candidates for experimental testing. For example, rs2364723 which is associated with *NFE2L2* transcript levels has also been reported to be associated with triglyceride levels[154] in a study not included in the GWAS catalog. While it is not feasible to experimentally test over 400 candidate variants at eQTL or GWAS loci in the short term, allelic imbalance detection has reduce the number of candidate variants compared to the thousands of variants in LD at cardiometabolic GWAS loci.

Chapter 3 describes allelic imbalance in only a subset of samples (liver, pancreatic islets, and adipose) that are related to cardiovascular disease. Quantitative sequence data has also been generated in other tissues such as skeletal muscle, brain and heart with known connections to cardiometabolic traits and diseases. Identifying allelic imbalance in these other samples would immediately provide additional candidate variants and further insight into gene transcription at GWAS loci.

4.4 Long-term impact on researching genetic effects on cardiometabolic phenotypes

While these sites of allelic imbalance provide an immediate contribution to research studying previously identified cardiometabolic GWAS loci, they will continue to provide new contributions in the future. As novel eQTL and GWAS loci are discovered, sites of allelic imbalance offer compelling candidate variants to test for experimental testing and have the potential to greatly expedite experimental characterization of these loci. Additionally, as more examples of predicted allelic imbalance are discovered at sites demonstrating experimentally observed allelic differences in regulatory activity, allelic imbalance detection may begin to play a more prominent role in genetic studies. I anticipate that greater effort will be placed in generating quantitative sequence data in additional samples with relevance to cardiometabolic traits and diseases. One further use of this strategy would apply to quantitative sequence data generated from tissue samples from individuals with genotype and phenotype data. Integrating these data can identify genotype-phenotype associations that are mediated by gene regulatory regions, an idea discussed briefly near the end of this chapter.

4.5 Studying complex, non-cardiometabolic phenotypes

We have focused our analyses on cardiometabolic traits and disease, but sites of allelic imbalance also play an important role in studying other phenotypes. As shown by the experimental data at Crohn's disease loci presented in **Chapter 2**, allelic imbalance sites occur at GWAS loci for and in tissues relevant to other traits and diseases. A wealth of data exists for samples relevant to other complex traits and disease. In particular, quantitative sequence data has been generated in bulk for the lymphoblastoid cell line (LCL) GM12878, and data has also been generated for more than 70 other LCLs. As we have shown, allelic imbalance sites detected in data generated from these samples are particularly useful in studying diseases related to the immune system and inflammation.

4.6 Advantages of allelic imbalance detection

A common method of regulatory variant prediction is to identify variants overlapping peaks from one or more quantitative sequence datasets. While powerful in identifying variants in regulatory regions this method does not identify which of these variants demonstrate allelic differences in regulatory activity. Allelic imbalance detection can predicts allelic differences in regulatory activity, specifically protein-DNA binding, providing strong evidence of allelic effects on regulation. Allelic imbalance detection directly interrogates genetic effects on quantitative sequence data providing strong evidence of potential allelic differences in regulatory activity than the simple overlap analyses.

A second, common method of identifying variants with allelic differences in transcriptional regulation is to identify eQTLs. These analyses identify groups of variant in LD with each other that are associated with gene expression, but in most cases cannot distinguish which variant(s) in the group are responsible for the association. Together with allelic imbalance detection, which has the advantage of predicting which variants have evidence of allelic differences in protein binding, these analyses provide evidence of the effects of individual variants on gene transcription regulation. Imbalance detection in ChIP-seq or DNase-seq data cannot directly detect allelic differences in gene regulation. However, if non-coding variants and coding variants are in linkage disequilibrium with each other, allelic imbalance can be detected in ChIP-seq or DNase-seq and RNA-seq from the same sample to infer a direct relationship between imbalances in regulatory regions and the transcript levels of genes. This combined analysis can be powerful, especially in the absence of other eQTL evidence, to detect allelic differences in gene transcription regulation in as little as two quantitative sequence datasets from a single sample. An advantage of this approach is that allelic imbalance detection is able to detect subtle differences between alleles that eQTL analyses do not have the power to detect due to variability between samples.

4.7 Limitations of allelic imbalance detection

Along with these advantages, there are multiple limitations in allelic imbalance detection analyses. While allelic imbalance can indicate allelic effects on gene regulatory regions, they do not necessarily indicate allelic differences in gene transcript levels. Compensatory mechanisms can correct for imbalances and maintain gene transcript levels. Cell homeostasis limits the application of allelic imbalance detection in identifying genetic variants influencing complex diseases and traits through the regulation of gene transcription.

Another current limitation is the availability of sufficient quantitative sequence data. While many datasets exist for some samples such as GM12878 and HepG2, other samples, particular primary cells and tissues, have quantitative sequence data available for very few transcription factors and histone modifications. This limitation is especially apparent in **Chapter 3**, because very little data exists for pancreatic islets and adipose tissue likely due to limited availability of human pancreatic islet cells and difficulty performing some quantitative sequencing assays on adipose cells. Additionally, the choice of protein or histone modifications examined in ChIP-seq experiments varies widely between samples and depends in large part on the preferences of investigators generating the data.

The largest limitation of imbalance detection is directly connected to its advantages. Allelic imbalance can only be detected at heterozygous sites. While allelic imbalance can predict differences in DNA-protein binding using data from a single individual, the analysis is limited to finding imbalance at heterozygous sites in that individual. In **Chapter 3**, we used predominately data from HepG2 to study cardiometabolic disease and traits due to the large number of transcription factors studied in that one cell line. As a result, we failed to detect imbalances at many variants that have published evidence of allelic differences in protein binding because they are homozygous in HepG2. To overcome this limitation, quantitative sequence data must be generated from multiple people with different genetic backgrounds. As the sample size increases a greater proportion of variants are able to be tested for imbalance, but at some point the population becomes more akin to that used for eQTL analyses and allelic imbalance detection is not as advantageous as at low sample sizes[55,56].

Overall, allelic imbalance detection is limited by the number of experiments required to be performed in individuals with diverse genetic backgrounds. Perfectly comprehensive imbalance detection requires the generation of ChIP-seq data for each individual transcription factor in samples of the same tissue from a population large enough to find at least one individual that is heterozygous for each variant. Studying multiple tissues further increases the number of ChIP-seq experiments required, and generating data on this scale is currently not feasible. DNase-seq and ATAC-seq can identify allelic differences in binding of multiple proteins reducing the number of ChIP-seq experiments requires. Our data, however, suggests that the sequence depth of currently available DNase-seq experiments is not deep enough for comprehensive detection of imbalances in ChIP-seq data. Further study is needed to determine if deeper sequencing of DNase-seq data can truly allow for more comprehensive imbalance detection and limit the number of ChIP-seq experiments required.

4.8 Allelic imbalance and population studies of quantitative sequence data

Like phenotype data or gene expression data, associations can be detected between genotypes and non-RNA-seq quantitative sequence data if the data is available for enough individuals in a study population. For example, associations between genotype and DNase hypersensitivity (ds-QTL) [55], histone modification[56], and Pol2 binding[56] have been reported using expression and genotype data from LCLs. Additionally, joint ds-QTL and eQTL analyses in these samples have identified genetic variants that simultaneously affect DNase hypersensitivity and gene expression [55]. There is no phenotype information for the LCL samples, preventing any testing for phenotypic associations. In studies like this, allele-aware alignment of the sequence data is critical, because reference mapping biases could introduce artificial associations or obscure real associations. AA-ALIGNER is particularly well suited for confidently removing these mapping biases when using imputed genotypes for study participants.

Until quantitative sequence data exists for a large number of individuals in study populations, allelic imbalance detection will play an important role in understanding the genetic effects on gene regulation. Even in the presence of these large populations, allelic imbalance detection will still have the power to interrogate genetic effects on transcription regulation at the level of individual samples and detect subtle changes not able to be detected by association testing. Additionally, quantitative sequence data at heterozygous sites found in multiple individuals can be combined to increase overall sequencing depth and power to detect significant imbalance.

In conclusion, allelic imbalance detection in quantitative sequence data is a powerful tool for studying genetic effects on the regulation of gene transcription at cardiometabolic phenotype-associated loci. Using allelic imbalance detection, we have identified many candidate variants to be tested experimentally for allelic differences in protein binding, enchancer activity, and gene transcription regulation. As more quantitative sequence data is generated and analyzed, imbalance detection will continue to play an important role in identifying additional variants likely influencing gene transcription at cardiometabolic and other complex trait- and disease-associated loci.

APPENDIX 1: Allelic imbalance sites referenced in text

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs9438904	chr1:25756860	HepG2	YY1	C	8	T	0	0.00781	imputed
rs3795688	chr1:26560481	HepG2	CTCF	A	83	C	52	0.00956	imputed
rs1494813	chr1:45957290	HepG2	MXI1	C	8	T	0	0.00781	imputed
rs9793263	chr1:46722389	HepG2	MAFK	G	34	A	2	1.94E-08	imputed
rs3862273	chr1:48251342	HepG2	YY1	T	9	G	0	0.00391	imputed
rs17125090	chr1:63988904	HepG2	MXI1	A	25	G	7	0.0021	imputed
rs17125090	chr1:63988904	HepG2	MAX	A	90	G	28	8.91E-09	imputed
rs2301054	chr1:64107028	HepG2	HDAC2	A	8	G	0	0.00781	imputed
rs2301054	chr1:64107028	HepG2	HNF4A	A	31	G	9	0.00068	imputed
rs12021623	chr1:66153586	HepG2	HNF4G	C	19	A	5	0.00661	imputed
rs942849	chr1:84427499	HepG2	JUND	G	9	A	0	0.00391	imputed
rs11161503	chr1:85462582	HepG2	EZH2	C	8	G	0	0.00781	imputed
rs11161505	chr1:85462665	HepG2	EZH2	T	8	G	0	0.00781	imputed
rs2268667	chr1:85793746	HepG2	FOXA1	G	12	A	1	0.00342	imputed
rs2177461	chr1:85861976	HepG2	ZBTB33	G	9	C	0	0.00391	imputed
rs12127787	chr1:89458761	HepG2	USF1	T	102	C	6	1.25E-23	imputed
rs12127787	chr1:89458761	HepG2	USF2	T	23	C	1	2.98E-06	imputed
rs12127787	chr1:89458761	HepG2	CREB1	C	33	T	12	0.00246	imputed
rs12127787	chr1:89458761	HepG2	YY1	C	30	T	11	0.00432	imputed
rs10858091	chr1:109935578	HepG2	SIN3AK20	C	11	T	1	0.00635	imputed
rs10858091	chr1:109935578	HepG2	TCF12	C	8	T	0	0.00781	imputed
rs10858091	chr1:109935578	HepG2	HEY1	C	24	T	6	0.00143	imputed
rs2140924	chr1:109935775	HepG2	CREB1	A	19	C	4	0.0026	imputed
rs573491	chr1:110026891	HepG2	TAF1	T	31	G	9	0.00068	imputed
rs2781553	chr1:110026989	HepG2	YY1	G	11	T	0	0.00098	imputed
rs839605	chr1:120217524	HepG2	CTCF	C	68	A	1	2.37E-19	imputed
rs639761	chr1:120217558	HepG2	CTCF	G	131	A	2	1.64E-36	imputed
rs639761	chr1:120217558	HepG2	CEBPA	G	8	A	0	0.00781	imputed
rs639761	chr1:120217558	HepG2	CEBPB	G	9	A	0	0.00391	imputed
rs639761	chr1:120217558	HepG2	CREB1	G	8	A	0	0.00781	imputed
rs639761	chr1:120217558	HepG2	RAD21	G	13	A	1	0.00183	imputed
rs640195	chr1:120217650	HepG2	CTCF	T	172	A	2	1.27E-48	imputed
rs640195	chr1:120217650	HepG2	ARID3A	T	10	A	0	0.00195	imputed
rs640195	chr1:120217650	HepG2	CEBPB	T	8	A	0	0.00781	imputed
rs640195	chr1:120217650	HepG2	CREB1	T	8	A	0	0.00781	imputed
rs12091564	chr1:145395604	HepG2	TBP	C	303	T	145	6.60E-14	imputed
rs12091564	chr1:145395604	HepG2	MYBL2	C	14	T	1	0.00098	imputed
rs12091564	chr1:145395604	HepG2	CREB1	C	16	T	3	0.00443	imputed
rs12091564	chr1:145395604	HepG2	NR2F2	C	17	T	3	0.00258	imputed
rs6677420	chr1:150946490	HepG2	MAX	A	9	G	0	0.00391	imputed
rs6674171	chr1:154491683	HepG2	MAFF	A	12	G	1	0.00342	imputed
rs6674171	chr1:154491683	HepG2	MAFK	A	33	G	8	0.00011	imputed
rs2236869	chr1:169535196	HepG2	FOXA1	G	17	T	4	0.0072	imputed
rs2281007	chr1:171111351	HepG2	CEBPA	G	14	A	2	0.00418	imputed
rs17361251	chr1:178520577	HepG2	FOXA2	A	34	C	0	1.16E-10	imputed
rs17361251	chr1:178520577	HepG2	CEBPD	A	8	C	0	0.00781	imputed
rs17361251	chr1:178520577	HepG2	CEBPB	A	31	C	0	9.31E-10	imputed
rs17361251	chr1:178520577	HepG2	MAX	A	14	C	1	0.00098	imputed
rs17361251	chr1:178520577	HepG2	ZBTB7A	A	11	C	1	0.00635	imputed
rs17361251	chr1:178520577	HepG2	FOXA1	A	46	C	0	2.84E-14	imputed
rs17361251	chr1:178520577	HepG2	NR2F2	A	11	C	0	0.00098	imputed
rs17361251	chr1:178520577	HepG2	SP1	A	14	C	0	0.00012	imputed
rs17361251	chr1:178520577	HepG2	NFIC	A	8	C	0	0.00781	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs17361251	chr1:178520577	HepG2	HNF4G	A	33	C	0	2.33E-10	imputed
rs17361251	chr1:178520577	HepG2	CEBPA	A	22	C	0	4.77E-07	imputed
rs17361251	chr1:178520577	HepG2	P300	A	26	C	1	4.17E-07	imputed
rs17361251	chr1:178520577	HepG2	HNF4A	A	77	C	0	1.32E-23	imputed
rs17276513	chr1:178520604	HepG2	FOXA2	A	44	T	0	1.14E-13	imputed
rs17276513	chr1:178520604	HepG2	CEBPB	A	31	T	0	9.31E-10	imputed
rs17276513	chr1:178520604	HepG2	MAX	A	13	T	0	0.00024	imputed
rs17276513	chr1:178520604	HepG2	DNASE	A	13	T	0	0.00024	imputed
rs17276513	chr1:178520604	HepG2	ZBTB7A	A	8	T	0	0.00781	imputed
rs17276513	chr1:178520604	HepG2	FOXA1	A	42	T	0	4.55E-13	imputed
rs17276513	chr1:178520604	HepG2	NR2F2	A	11	T	0	0.00098	imputed
rs17276513	chr1:178520604	HepG2	SP1	A	16	T	0	3.05E-05	imputed
rs17276513	chr1:178520604	HepG2	HDAC2	A	12	T	0	0.00049	imputed
rs17276513	chr1:178520604	HepG2	CEBPA	A	9	T	0	0.00391	imputed
rs17276513	chr1:178520604	HepG2	P300	A	26	T	1	4.17E-07	imputed
rs17276513	chr1:178520604	HepG2	HNF4G	A	21	T	0	9.54E-07	imputed
rs17276513	chr1:178520604	HepG2	HNF4A	A	81	T	0	8.27E-25	imputed
rs17276527	chr1:178520680	HepG2	FOXA2	A	53	G	0	2.22E-16	imputed
\rs17276527	chr1:178520680	HepG2	MAX	A	12	G	0	0.00049	imputed
rs17276527	chr1:178520680	HepG2	DNASE	A	13	G	0	0.00024	imputed
rs17276527	chr1:178520680	HepG2	FOXA1	A	61	G	1	2.73E-17	imputed
rs17276527	chr1:178520680	HepG2	ZEB1	A	12	G	0	0.00049	imputed
rs17276527	chr1:178520680	HepG2	SP1	A	12	G	0	0.00049	imputed
rs17276527	chr1:178520680	HepG2	NFIC	A	9	G	0	0.00391	imputed
rs17276527	chr1:178520680	HepG2	HDAC2	A	9	G	0	0.00391	imputed
rs17276527	chr1:178520680	HepG2	P300	A	23	G	0	2.38E-07	imputed
rs17276527	chr1:178520680	HepG2	HNF4G	A	9	G	0	0.00391	imputed
rs17276527	chr1:178520680	HepG2	CREB1	A	12	G	0	0.00049	imputed
rs17276527	chr1:178520680	HepG2	HNF4A	A	59	G	0	3.47E-18	imputed
rs2488400	chr1:197702225	HepG2	FOXA2	G	9	C	0	0.00391	imputed
rs2201601	chr1:213031448	HepG2	YY1	G	27	C	9	0.00393	imputed
rs11120067	chr1:213094557	HepG2	CTCF	A	22	G	4	0.00053	imputed
rs9970073	chr1:214156165	HepG2	MAX	A	57	G	1	4.09E-16	imputed
rs340879	chr1:214156514	HepG2	RAD21	T	11	C	0	0.00098	imputed
rs4846770	chr1:222795569	HepG2	CEBPD	G	10	C	0	0.00195	imputed
rs12077115	chr1:225924711	HepG2	CEBPA	A	9	G	0	0.00391	imputed
rs12040438	chr1:229718258	HepG2	BHLHE40	T	9	C	0	0.00391	imputed
rs4846913	chr1:230294715	HepG2	CEBPD	A	11	C	1	0.00635	imputed
rs4846913	chr1:230294715	HepG2	CEBPB	A	57	C	21	5.57E-05	imputed
rs4846913	chr1:230294715	HepG2	NR2F2	A	35	C	15	0.0066	imputed
rs4846913	chr1:230294715	HepG2	CEBPA	A	17	C	2	0.00073	imputed
rs4846913	chr1:230294715	HI81	MAFB	C	21	A	5	0.00249	common
rs526936	chr1:234852204	HI32	FOXA2	A	38	G	17	0.00646	common
rs526936	chr1:234852204	HepG2	POL2	G	48	A	22	0.00255	imputed
rs556107	chr1:234853059	HepG2	HEY1	C	22	T	6	0.00372	imputed
rs2066381	chr1:240995427	HI32	PDX1	A	24	G	6	0.00143	common
rs61844237	chr1:245133662	HepG2	DNASE	G	80	C	43	0.00108	imputed
rs2291426	chr1:245134114	HepG2	CREB1	A	23	C	6	0.00232	imputed
rs4020081	chr1:245209045	HepG2	FOXA1	C	104	T	62	0.00139	imputed
rs4020081	chr1:245209045	HepG2	P300	C	15	T	2	0.00235	imputed
rs4020081	chr1:245209045	HepG2	CREB1	C	28	T	3	4.65E-06	imputed
rs4020082	chr1:245209134	HepG2	FOXA1	A	58	G	25	0.00038	imputed
rs6759670	chr2:950291	HepG2	RAD21	C	18	A	1	7.63E-05	imputed
rs34122754	chr2:9884076	HepG2	CTCF	C	8	G	0	0.00781	imputed
rs4669449	chr2:9884205	HepG2	EZH2	G	8	A	0	0.00781	imputed
rs4669888	chr2:12980514	HepG2	CEBPB	G	14	A	1	0.00098	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs4669888	chr2:12980514	HepG2	CEBPA	G	8	A	0	0.00781	imputed
rs633808	chr2:20957592	HepG2	TCF12	G	8	T	0	0.00781	imputed
rs1367117	chr2:21263900	HepG2	MAX	G	8	A	0	0.00781	imputed
rs1367117	chr2:21263900	HepG2	HEY1	G	29	A	10	0.00338	imputed
rs312983	chr2:21378580	HepG2	FOXA1	A	22	C	3	0.00016	imputed
rs312984	chr2:21378778	HepG2	FOXA2	C	62	T	33	0.00383	imputed
rs312984	chr2:21378778	HepG2	MAX	C	9	T	0	0.00391	imputed
rs312984	chr2:21378778	HepG2	RAD21	C	12	T	0	0.00049	imputed
rs312984	chr2:21378778	HepG2	FOXA1	C	174	T	71	3.72E-11	imputed
rs312984	chr2:21378778	HepG2	ZEB1	C	15	T	3	0.00754	imputed
rs312984	chr2:21378778	HepG2	NFIC	C	24	T	6	0.00143	imputed
rs312984	chr2:21378778	HepG2	ARID3A	C	15	T	1	0.00052	imputed
rs312984	chr2:21378778	HepG2	HNF4A	C	31	T	6	4.13E-05	imputed
rs312985	chr2:21378805	HepG2	NRSF	A	10	G	0	0.00195	imputed
rs312985	chr2:21378805	HepG2	FOXA2	A	94	G	42	9.67E-06	imputed
rs312985	chr2:21378805	HepG2	MYBL2	A	20	G	6	0.00936	imputed
rs312985	chr2:21378805	HepG2	FOXA1	A	224	G	99	2.80E-12	imputed
rs312985	chr2:21378805	HepG2	ZEB1	A	15	G	2	0.00235	imputed
rs312985	chr2:21378805	HepG2	SP1	A	29	G	3	2.56E-06	imputed
rs312985	chr2:21378805	HepG2	HDAC2	A	23	G	7	0.00522	imputed
rs312985	chr2:21378805	HepG2	P300	A	42	G	20	0.00715	imputed
rs312985	chr2:21378805	HepG2	HNF4G	A	20	G	3	0.00049	imputed
rs312985	chr2:21378805	HepG2	CREB1	A	15	G	3	0.00754	imputed
rs312985	chr2:21378805	HepG2	HNF4A	A	44	G	10	3.39E-06	imputed
rs1652418	chr2:21388456	HepG2	MAZ	T	8	C	0	0.00781	imputed
rs1652418	chr2:21388456	HepG2	RAD21	T	60	C	28	0.00085	imputed
rs1652418	chr2:21388456	HepG2	SMC3	T	15	C	2	0.00235	imputed
rs386643898	chr2:21398985	HepG2	CTCF	C	8	A	0	0.00781	imputed
rs386643898	chr2:21398985	HepG2	RAD21	C	21	A	2	6.60E-05	imputed
rs7572949	chr2:24162018	HepG2	CEBPB	T	12	C	0	0.00049	imputed
rs36101491	chr2:24387532	HepG2	CTCF	T	368	C	197	5.73E-13	imputed
rs36101491	chr2:24387532	HepG2	RAD21	T	89	C	47	0.0004	imputed
rs36101491	chr2:24387532	HepG2	ZBTB7A	T	10	C	0	0.00195	imputed
rs17046192	chr2:24461334	HepG2	ZBTB7A	A	8	G	0	0.00781	imputed
rs17046192	chr2:24461334	HepG2	FOXA1	A	40	G	15	0.00102	imputed
rs11676939	chr2:24479057	HepG2	CEBPB	C	11	T	1	0.00635	imputed
rs72803210	chr2:24615710	HepG2	MAFK	A	18	G	0	7.63E-06	imputed
rs77421503	chr2:24625676	HepG2	MAX	C	10	G	0	0.00195	imputed
rs10460551	chr2:24627074	HepG2	HDAC2	C	8	T	0	0.00781	imputed
rs7580081	chr2:25097072	HepG2	CEBPB	C	18	G	1	7.63E-05	imputed
rs11684202	chr2:25887558	HepG2	RAD21	G	9	A	0	0.00391	imputed
rs2011616	chr2:27302561	HepG2	HDAC2	G	9	A	0	0.00391	imputed
rs2580759	chr2:27432500	HepG2	CTCF	G	11	T	1	0.00635	imputed
rs11608	chr2:27435374	HepG2	YY1	G	42	A	16	0.00086	imputed
rs1141313	chr2:27460968	HepG2	POL2	G	9	A	0	0.00391	imputed
rs1260326	chr2:27730940	HepG2	CTCF	C	9	T	0	0.00391	imputed
rs780095	chr2:27741105	HepG2	FOXA1	G	25	A	6	0.00088	imputed
rs780094	chr2:27741237	HepG2	NRSF	C	8	T	0	0.00781	imputed
rs780094	chr2:27741237	HepG2	FOXA2	C	24	T	5	0.00055	imputed
rs780094	chr2:27741237	HepG2	CEBPB	C	9	T	0	0.00391	imputed
rs780094	chr2:27741237	HepG2	MAX	C	16	T	2	0.00131	imputed
rs780094	chr2:27741237	HepG2	FOXA1	C	31	T	10	0.00145	imputed
rs780094	chr2:27741237	HepG2	ZEB1	C	12	T	1	0.00342	imputed
rs780094	chr2:27741237	HepG2	NR2F2	C	24	T	4	0.00018	imputed
rs780094	chr2:27741237	HepG2	MAFK	T	22	C	6	0.00372	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs3749147	chr2:27851918	HepG2	MXI1	G	34	A	13	0.00309	imputed
rs1919128	chr2:27801759	HepG2	BHLHE40	A	8	G	0	0.00781	imputed
rs3749147	chr2:27851918	HepG2	TAF1	G	94	A	38	1.19E-06	imputed
rs3749147	chr2:27851918	HepG2	CEBPD	G	10	A	0	0.00195	imputed
rs3749147	chr2:27851918	HepG2	MAX	G	59	A	32	0.00611	imputed
rs3749147	chr2:27851918	HepG2	ELF1	G	28	A	6	0.0002	imputed
rs3749147	chr2:27851918	HepG2	DNASE	G	35	A	10	0.00025	imputed
rs3749147	chr2:27851918	HepG2	ZBTB7A	G	13	A	1	0.00183	imputed
rs3749147	chr2:27851918	HepG2	NR2F2	G	18	A	3	0.00149	imputed
rs3749147	chr2:27851918	HepG2	YY1	G	133	A	86	0.00182	imputed
rs3749147	chr2:27851918	HepG2	POL2	G	71	A	32	0.00015	imputed
rs3749147	chr2:27851918	HepG2	GABP	G	58	A	15	4.09E-07	imputed
rs162330	chr2:38319496	HepG2	CEBPB	C	14	A	2	0.00418	imputed
rs2881327	chr2:46641955	HepG2	MAX	G	17	A	3	0.00258	imputed
rs2881327	chr2:46641955	HepG2	FOXA1	G	25	A	5	0.00032	imputed
rs2881327	chr2:46641955	HepG2	HNF4G	G	8	A	0	0.00781	imputed
rs1554612	chr2:48827497	HepG2	CTCF	C	67	T	23	3.80E-06	imputed
rs1554612	chr2:48827497	HepG2	FOXA2	C	11	T	0	0.00098	imputed
rs1554612	chr2:48827497	HepG2	RAD21	C	42	T	20	0.00715	imputed
rs72800719	chr2:54280195	HepG2	CTCF	G	33	C	14	0.00794	imputed
rs10192403	chr2:54313507	HepG2	FOXA1	C	8	T	0	0.00781	imputed
rs62165172	chr2:55736237	HepG2	FOXA2	G	10	A	0	0.00195	imputed
rs782599	chr2:55847423	HepG2	CTCF	T	31	C	11	0.00289	imputed
rs782599	chr2:55847423	HepG2	FOXA1	T	8	C	0	0.00781	imputed
rs2136737	chr2:62969310	HepG2	FOXA2	G	12	C	1	0.00342	imputed
rs2136737	chr2:62969310	HepG2	HNF4A	C	8	G	0	0.00781	imputed
rs1553832	chr2:63013515	HepG2	BHLHE40	G	12	C	1	0.00342	imputed
rs368327833	chr2:63095792	HepG2	POL2	G	9	A	0	0.00391	imputed
rs2710642	chr2:63149557	HepG2	POL2	G	11	A	1	0.00635	imputed
rs2070063	chr2:64862055	HepG2	MAFK	A	32	G	13	0.00661	imputed
rs35125132	chr2:65041562	HepG2	EZH2	T	8	G	0	0.00781	imputed
rs11890701	chr2:70360457	HI32	PDX1	T	19	A	3	0.00086	common
rs7582417	chr2:70367861	HepG2	POL2	A	14	G	0	0.00012	imputed
rs386647116	chr2:70368391	HepG2	CEBPB	A	9	G	0	0.00391	imputed
rs11692018	chr2:70369154	HepG2	POL2	A	33	C	12	0.00246	imputed
rs11692018	chr2:70369154	HepG2	NFIC	A	12	C	0	0.00049	imputed
rs12713688	chr2:70369503	HepG2	MXI1	G	17	C	3	0.00258	imputed
rs12713688	chr2:70369503	HepG2	MAX	G	46	C	22	0.0049	imputed
rs12713688	chr2:70369503	HepG2	HEY1	G	17	C	4	0.0072	imputed
rs4338986	chr2:70376574	HepG2	HEY1	C	18	A	4	0.00434	imputed
rs10469966	chr2:73752368	HepG2	CEBPB	G	9	A	0	0.00391	imputed
rs62150376	chr2:83295262	HepG2	FOXA1	T	31	C	5	1.29E-05	imputed
rs62150376	chr2:83295262	HepG2	HDAC2	T	14	C	1	0.00098	imputed
rs62150376	chr2:83295262	HepG2	HNF4G	T	11	C	1	0.00635	imputed
rs2241883	chr2:88424066	HepG2	POL2	T	35	C	8	4.19E-05	imputed
rs13431601	chr2:102873713	HepG2	EZH2	G	8	A	0	0.00781	imputed
rs2276561	chr2:113956371	HepG2	MAX	G	16	C	2	0.00131	imputed
rs2305133	chr2:113956821	HepG2	CTCF	C	129	G	76	0.00026	imputed
rs2305133	chr2:113956821	HepG2	BHLHE40	C	9	G	0	0.00391	imputed
rs2305133	chr2:113956821	HepG2	RAD21	C	52	G	21	0.00037	imputed
rs931472	chr2:113969948	HepG2	DNASE	C	21	T	2	6.60E-05	imputed
rs1049137	chr2:113975110	HepG2	POL2	G	8	A	0	0.00781	imputed
rs2289897	chr2:113977454	HepG2	CEBPB	A	8	G	0	0.00781	imputed
rs4849176	chr2:113977936	HepG2	POL2	C	16	T	3	0.00443	imputed
rs4849178	chr2:113982608	HepG2	FOXA1	A	8	G	0	0.00781	imputed
rs2166421	chr2:113990242	HepG2	FOXA1	C	26	T	4	5.95E-05	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs2166421	chr2:113990242	HepG2	HNF4A	C	11	T	1	0.00635	imputed
rs7421852	chr2:113990261	HepG2	FOXA1	A	20	G	6	0.00936	imputed
rs10206269	chr2:113990393	HepG2	HNF4A	A	12	C	1	0.00342	imputed
rs4849181	chr2:113991970	HepG2	CTCF	G	47	A	16	0.00012	imputed
rs7564909	chr2:129504503	HepG2	CEBPB	A	10	T	0	0.00195	imputed
rs6758916	chr2:129531820	HepG2	MAFK	T	19	A	4	0.0026	imputed
rs1010241	chr2:135154312	HepG2	BHLHE40	A	8	C	0	0.00781	imputed
rs62168897	chr2:135717997	HepG2	POL2	C	13	T	2	0.00739	imputed
rs34272267	chr2:150536025	HepG2	CEBPB	C	42	G	9	3.39E-06	imputed
rs6706545	chr2:161181478	HepG2	FOXA2	A	16	T	1	0.00027	imputed
rs6706545	chr2:161181478	HepG2	CEBPB	A	9	T	0	0.00391	imputed
rs10929982	chr2:161236277	HepG2	CEBPB	T	14	C	0	0.00012	imputed
rs73029563	chr2:165008166	HepG2	CEBPB	C	8	G	0	0.00781	imputed
rs13004226	chr2:165080678	HepG2	MAFK	C	77	G	37	0.00023	imputed
rs6713419	chr2:165508300	HepG2	MAFK	T	8	C	0	0.00781	imputed
rs10184004	chr2:165508389	HepG2	MAFK	C	25	T	1	8.05E-07	imputed
rs6754950	chr2:170630370	HepG2	ZBTB33	A	13	G	0	0.00024	imputed
rs7579463	chr2:171253722	HepG2	CTCF	A	44	C	13	4.71E-05	imputed
rs7579463	chr2:171253722	HepG2	RAD21	A	43	C	20	0.00515	imputed
rs2364723	chr2:178126546	HepG2	MBD4	C	8	G	0	0.00781	imputed
rs13427277	chr2:188075497	HepG2	CTCF	A	8	G	0	0.00781	imputed
rs840601	chr2:188159887	HepG2	TCF7L2	A	9	G	0	0.00391	imputed
rs696092	chr2:188210214	HepG2	MAFK	A	42	G	17	0.00155	imputed
rs1355521	chr2:188307747	HepG2	CEBPB	G	13	A	0	0.00024	imputed
rs10201618	chr2:188326949	HepG2	USF1	C	27	A	6	0.00032	imputed
rs8176547	chr2:188340349	HepG2	POL2	G	12	T	1	0.00342	imputed
rs8176547	chr2:188340349	HepG2	MAFK	G	27	T	6	0.00032	imputed
rs8176546	chr2:188340396	HepG2	POL2	T	16	C	2	0.00131	imputed
rs8176546	chr2:188340396	HepG2	MAFK	T	32	C	10	0.00094	imputed
rs938929	chr2:198780860	HepG2	RAD21	G	9	A	0	0.00391	imputed
rs12991600	chr2:202337236	HepG2	MAFK	G	19	A	2	0.00022	imputed
rs6745050	chr2:204691538	HepG2	MAFK	T	48	C	19	0.00052	imputed
rs6747951	chr2:206829310	HepG2	MAX	G	17	C	4	0.0072	imputed
rs6747951	chr2:206829310	HepG2	FOXA1	G	11	C	1	0.00635	imputed
rs1250240	chr2:216295014	HepG2	POL2	G	25	A	9	0.00904	imputed
rs1250241	chr2:216295312	HepG2	POL2	A	35	T	14	0.0038	imputed
rs1250244	chr2:216297796	HepG2	FOXA2	C	11	G	0	0.00098	imputed
rs1250244	chr2:216297796	HepG2	POL2	C	68	G	23	2.52E-06	imputed
rs1250258	chr2:216300185	HepG2	POL2	T	47	C	23	0.00558	imputed
rs1250258	chr2:216300185	HepG2	NFIC	T	12	C	1	0.00342	imputed
rs1250259	chr2:216300482	HepG2	NR2F2	A	8	T	0	0.00781	imputed
rs10171839	chr2:219051314	HepG2	MAX	G	11	A	1	0.00635	imputed
rs13062	chr2:219260651	HepG2	CTCF	A	15	C	3	0.00754	imputed
rs13062	chr2:219260651	HepG2	MAX	A	14	C	2	0.00418	imputed
rs13062	chr2:219260651	HepG2	ZEB1	A	11	C	1	0.00635	imputed
rs13423632	chr2:232079116	HepG2	MAX	C	21	T	5	0.00249	imputed
rs16827879	chr2:232092301	HepG2	CTCF	T	64	C	34	0.00319	imputed
rs16827879	chr2:232092301	HepG2	RAD21	T	89	C	55	0.00577	imputed
rs4477910	chr2:234643737	HepG2	FOXA2	T	27	A	0	1.49E-08	imputed
rs4477910	chr2:234643737	HepG2	FOXA1	T	30	A	0	1.86E-09	imputed
rs4477910	chr2:234643737	HepG2	NR2F2	T	8	A	0	0.00781	imputed
rs4477910	chr2:234643737	HepG2	HNF4A	T	9	A	0	0.00391	imputed
rs77438791	chr2:239035642	HepG2	MBD4	G	8	A	0	0.00781	imputed
rs4401206	chr2:241796905	HepG2	MXI1	G	35	A	15	0.0066	imputed
rs4401206	chr2:241796905	HepG2	USF1	G	76	A	32	2.76E-05	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs4401206	chr2:241796905	HepG2	CEBPD	G	8	A	0	0.00781	imputed
rs4401206	chr2:241796905	HepG2	MAX	G	213	A	120	3.91E-07	imputed
rs4401206	chr2:241796905	HepG2	DNASE	G	66	A	29	0.00019	imputed
rs4401206	chr2:241796905	HepG2	RAD21	G	12	A	1	0.00342	imputed
rs4401206	chr2:241796905	HepG2	YY1	G	17	A	3	0.00258	imputed
rs4401206	chr2:241796905	HepG2	P300	G	26	A	9	0.00599	imputed
rs4401206	chr2:241796905	HepG2	HNF4G	G	25	A	4	0.0001	imputed
rs4401206	chr2:241796905	HepG2	CREB1	G	24	A	5	0.00055	imputed
rs4401206	chr2:241796905	HepG2	HNF4A	G	54	A	27	0.0036	imputed
rs10933641	chr2:241813788	HepG2	MAX	C	57	T	27	0.0014	imputed
rs10933641	chr2:241813788	HepG2	DNASE	C	35	T	15	0.0066	imputed
rs10933641	chr2:241813788	HepG2	POL2	C	35	T	8	4.19E-05	imputed
rs10206101	chr2:241835543	HepG2	ZEB1	C	60	A	13	2.31E-08	imputed
rs10206101	chr2:241835543	HepG2	JUND	C	9	A	0	0.00391	imputed
rs10206101	chr2:241835543	HepG2	HNF4G	C	8	A	0	0.00781	imputed
rs10206101	chr2:241835543	HepG2	FOSL2	C	11	A	1	0.00635	imputed
rs10933517	chr2:241836338	HepG2	DNASE	C	16	T	1	0.00027	imputed
rs10933517	chr2:241836338	HepG2	RAD21	C	9	T	0	0.00391	imputed
rs4675858	chr2:241840558	HepG2	CEBPB	A	8	G	0	0.00781	imputed
rs4417704	chr2:241846573	HepG2	CTCF	G	20	A	3	0.00049	imputed
rs4417704	chr2:241846573	HepG2	MAX	G	8	A	0	0.00781	imputed
rs4417704	chr2:241846573	HepG2	RAD21	G	12	A	1	0.00342	imputed
rs62186584	chr2:241853621	HepG2	CREB1	C	15	T	2	0.00235	imputed
rs9653611	chr2:243006956	HepG2	EZH2	G	11	C	0	0.00098	imputed
rs60533128	chr2:243028088	HepG2	P300	G	9	A	0	0.00391	imputed
rs58906257	chr2:243028248	HepG2	TBP	C	18	T	2	0.0004	imputed
rs58906257	chr2:243028248	HepG2	MAX	C	30	T	4	6.16E-06	imputed
rs58906257	chr2:243028248	HepG2	DNASE	C	20	T	3	0.00049	imputed
rs58906257	chr2:243028248	HepG2	POL2	C	95	T	12	3.49E-17	imputed
rs58906257	chr2:243028248	HepG2	SP1	C	8	T	0	0.00781	imputed
rs58906257	chr2:243028248	HepG2	HEY1	C	28	T	8	0.00119	imputed
rs58906257	chr2:243028248	HepG2	HDAC2	C	25	T	5	0.00032	imputed
rs58906257	chr2:243028248	HepG2	HNF4A	C	10	T	0	0.00195	imputed
rs58906257	chr2:243028248	HepG2	MAFK	C	14	T	1	0.00098	imputed
rs57603292	chr2:243028537	HepG2	FOXA1	A	9	G	0	0.00391	imputed
rs59191623	chr2:243028595	HepG2	RAD21	A	10	G	0	0.00195	imputed
rs990284	chr3:104972	HepG2	FOXA2	A	18	G	2	0.0004	imputed
rs990284	chr3:104972	HepG2	FOXA1	A	39	G	11	9.02E-05	imputed
rs990284	chr3:104972	HepG2	P300	A	11	G	1	0.00635	imputed
rs7640929	chr3:27513944	HepG2	CEBPA	A	10	C	0	0.00195	imputed
rs75016701	chr3:42091501	HepG2	FOXA1	G	28	A	9	0.00256	imputed
rs78607708	chr3:42096955	HepG2	FOXA2	G	50	A	15	1.57E-05	imputed
rs78607708	chr3:42096955	HepG2	CEBPB	G	60	A	26	0.00032	imputed
rs78607708	chr3:42096955	HepG2	FOXA1	G	53	A	15	4.12E-06	imputed
rs78607708	chr3:42096955	HepG2	CEBPA	G	42	A	13	0.00011	imputed
rs78607708	chr3:42096955	HepG2	P300	G	23	A	4	0.00031	imputed
rs3774750	chr3:50208406	HepG2	MAX	G	13	C	2	0.00739	imputed
rs2233474	chr3:50388607	HepG2	DNASE	C	30	A	11	0.00432	imputed
rs7639267	chr3:52568805	HepG2	HEY1	T	8	G	0	0.00781	imputed
rs1108842	chr3:52720080	HepG2	SIN3AK20	C	25	A	7	0.0021	imputed
rs1108842	chr3:52720080	HepG2	TAF1	C	75	A	39	0.00096	imputed
rs1108842	chr3:52720080	HepG2	CREB1	C	91	A	58	0.00853	imputed
rs2710323	chr3:52815905	HepG2	FOXA2	C	12	T	1	0.00342	imputed
rs2710323	chr3:52815905	HepG2	BHLHE40	C	11	T	0	0.00098	imputed
rs66815886	chr3:64703394	HepG2	BHLHE40	T	8	G	0	0.00781	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs6800622	chr3:119580678	HepG2	MAFK	C	16	A	2	0.00131	imputed
rs2976742	chr3:125417021	HepG2	CTCF	C	23	T	0	2.38E-07	imputed
rs12695470	chr3:125635718	HepG2	DNASE	A	8	C	0	0.00781	imputed
rs12695470	chr3:125635718	HepG2	POL2	A	9	C	0	0.00391	imputed
rs9841194	chr3:125635739	HepG2	POL2	T	13	C	2	0.00739	imputed
rs12497980	chr3:125636117	HepG2	MXI1	G	10	A	0	0.00195	imputed
rs12497980	chr3:125636117	HepG2	MAX	G	49	A	15	2.44E-05	imputed
rs4377449	chr3:125642330	HepG2	USF1	G	22	A	0	4.77E-07	imputed
rs9826071	chr3:125648165	HepG2	CTCF	T	95	C	53	0.0007	imputed
rs9826071	chr3:125648165	HepG2	RAD21	T	24	C	7	0.00333	imputed
rs17523380	chr3:125802874	HepG2	RAD21	T	8	C	0	0.00781	imputed
rs2939820	chr3:128127643	HepG2	CTCF	G	25	A	3	2.74E-05	imputed
rs4683799	chr3:139210258	HI87	NKX2_2	G	24	C	4	0.00018	common
rs11714980	chr3:167452991	HepG2	ZBTB7A	T	9	C	0	0.00391	imputed
rs58575091	chr3:186545319	HepG2	CHD2	T	15	C	2	0.00235	imputed
rs4689909	chr4:4643276	HepG2	HEY1	G	8	A	0	0.00781	imputed
rs7661077	chr4:7219889	HepG2	CEBPB	C	15	T	3	0.00754	imputed
rs11932616	chr4:26063055	HepG2	CTCF	C	59	T	1	1.06E-16	imputed
rs11932616	chr4:26063055	HepG2	RAD21	C	20	T	0	1.91E-06	imputed
rs78578320	chr4:68566689	HepG2	DNASE	G	20	A	5	0.00408	imputed
rs28653581	chr4:68567025	HepG2	DNASE	G	89	T	49	0.00084	imputed
rs4075927	chr4:79575058	HepG2	USF1	G	24	A	3	4.92E-05	imputed
rs4075927	chr4:79575058	HepG2	BHLHE40	G	8	A	0	0.00781	imputed
rs45499402	chr4:89043634	HepG2	FOXA1	G	24	C	5	0.00055	imputed
rs6841731	chr4:89228928	HepG2	MAFK	A	21	G	2	6.60E-05	imputed
rs2869930	chr4:89242372	HepG2	FOXA2	G	14	C	1	0.00098	imputed
rs2869930	chr4:89242372	HepG2	FOXA1	G	22	C	3	0.00016	imputed
rs77826206	chr4:95613564	HepG2	MAFK	T	8	C	0	0.00781	imputed
rs10030238	chr4:141808805	HepG2	HNF4G	A	11	G	1	0.00635	imputed
rs6813195	chr4:153520475	HI101	FOXA2	C	33	T	13	0.00453	common
rs2227426	chr4:155493171	HepG2	POL2	G	12	A	1	0.00342	imputed
rs6846466	chr4:166424428	HI101	FOXA2	T	16	C	3	0.00443	common
rs28641985	chr4:189376705	HepG2	SRF	A	52	G	0	4.44E-16	imputed
rs28641985	chr4:189376705	HepG2	FOXA2	A	20	G	0	1.91E-06	imputed
rs28641985	chr4:189376705	HepG2	ELF1	A	29	G	0	3.73E-09	imputed
rs28641985	chr4:189376705	HepG2	FOXA1	A	13	G	1	0.00183	imputed
rs28641985	chr4:189376705	HepG2	POL2	A	29	G	1	5.77E-08	imputed
rs28641985	chr4:189376705	HepG2	HEY1	A	13	G	0	0.00024	imputed
rs28641985	chr4:189376705	HepG2	CREB1	A	8	G	0	0.00781	imputed
rs28641985	chr4:189376705	HepG2	MAFK	A	8	G	0	0.00781	imputed
rs31490	chr5:1344458	HepG2	CEBPB	A	8	G	0	0.00781	imputed
rs835158	chr5:14873254	HepG2	CEBPB	G	8	C	0	0.00781	imputed
rs10941891	chr5:21391789	HepG2	CTCF	G	38	C	0	7.28E-12	imputed
rs10941891	chr5:21391789	HepG2	RAD21	G	24	C	0	1.19E-07	imputed
rs10941891	chr5:21391789	HepG2	SMC3	G	9	C	0	0.00391	imputed
rs3195676	chr5:34008100	HepG2	TAF1	C	46	T	23	0.00762	imputed
rs3195676	chr5:34008100	HepG2	BHLHE40	C	11	T	1	0.00635	imputed
rs3195676	chr5:34008100	HepG2	MAX	C	27	T	10	0.00763	imputed
rs13356762	chr5:56110992	HepG2	TAF1	G	20	T	6	0.00936	imputed
rs2548663	chr5:56172778	HepG2	CEBPB	G	8	A	0	0.00781	imputed
rs185220	chr5:56205357	HepG2	DNASE	G	41	A	7	6.24E-07	imputed
rs252923	chr5:56205662	HepG2	YY1	G	9	T	0	0.00391	imputed
rs33321	chr5:56206073	HepG2	HEY1	G	17	T	4	0.0072	imputed
rs16876512	chr5:78407261	HepG2	TCF12	T	8	C	0	0.00781	imputed
rs1643635	chr5:79928216	HepG2	MBD4	G	9	A	0	0.00391	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs10168	chr5:79950403	HepG2	MAX	C	9	T	0	0.00391	imputed
rs10168	chr5:79950403	HepG2	DNASE	C	42	T	8	1.16E-06	imputed
rs386689494	chr5:79961675	HepG2	MAFK	C	8	T	0	0.00781	imputed
rs1677645	chr5:79964419	HepG2	CEBPB	C	14	T	2	0.00418	imputed
rs226198	chr5:81573992	HepG2	USF1	C	10	T	0	0.00195	imputed
rs226198	chr5:81573992	HepG2	MAX	C	79	T	33	1.64E-05	imputed
rs4869419	chr5:92567539	HepG2	SRF	A	8	G	0	0.00781	imputed
rs10070197	chr5:95241868	HepG2	CEBPB	C	14	T	2	0.00418	imputed
rs1458017	chr5:95251346	HepG2	HEY1	A	11	G	1	0.00635	imputed
rs261973	chr5:95869427	HepG2	MAFK	T	11	C	1	0.00635	imputed
rs730870	chr5:125874993	HepG2	FOXA1	G	135	A	64	5.34E-07	imputed
rs192231	chr5:140248539	HepG2	CTCF	C	42	T	8	1.16E-06	imputed
rs192231	chr5:140248539	HepG2	RAD21	C	38	T	5	2.50E-07	imputed
rs13160685	chr5:149822331	HepG2	CEBPB	A	26	G	9	0.00599	imputed
rs1039438	chr5:156476770	HepG2	P300	G	22	A	7	0.00813	imputed
rs9328078	chr6:1979522	HepG2	CEBPB	C	8	A	0	0.00781	imputed
rs12195826	chr6:2565752	HepG2	FOXA2	G	24	A	6	0.00143	imputed
rs7739320	chr6:3054146	HepG2	DNASE	C	78	T	24	7.68E-08	imputed
rs12203636	chr6:3064249	HepG2	MYBL2	A	11	G	0	0.00098	imputed
rs12203636	chr6:3064249	HepG2	MAX	A	24	G	4	0.00018	imputed
rs12203636	chr6:3064249	HepG2	ZEB1	A	15	G	3	0.00754	imputed
rs12203636	chr6:3064249	HepG2	POL2	A	61	G	17	5.66E-07	imputed
rs12203636	chr6:3064249	HepG2	HEY1	A	37	G	11	0.00022	imputed
rs12196777	chr6:3064523	HepG2	DNASE	C	48	T	11	1.24E-06	imputed
rs12665605	chr6:3067003	HepG2	POL2	G	13	A	2	0.00739	imputed
rs12663589	chr6:3069057	HepG2	POL2	C	34	T	13	0.00309	imputed
rs9504915	chr6:6749069	HepG2	YY1	A	13	G	2	0.00739	imputed
rs9393818	chr6:10967918	HepG2	BHLHE40	T	8	G	0	0.00781	imputed
rs2295602	chr6:11005842	HepG2	BHLHE40	T	8	C	0	0.00781	imputed
rs3798713	chr6:11008622	HepG2	FOXA1	G	56	C	12	6.21E-08	imputed
rs3798713	chr6:11008622	HepG2	HNF4A	G	8	C	0	0.00781	imputed
rs953413	chr6:11012859	HepG2	FOXA2	G	58	A	23	0.00013	imputed
rs953413	chr6:11012859	HepG2	FOXA1	G	115	A	48	1.59E-07	imputed
rs953413	chr6:11012859	HepG2	NR2F2	G	26	A	5	0.00019	imputed
rs953413	chr6:11012859	HepG2	P300	G	39	A	18	0.00751	imputed
rs56190003	chr6:11088533	HepG2	HEY1	T	17	C	2	0.00073	imputed
rs13362715	chr6:11088630	HepG2	POL2	C	13	T	0	0.00024	imputed
rs13362715	chr6:11088630	HepG2	HEY1	C	14	T	0	0.00012	imputed
rs9379687	chr6:24721787	HepG2	MXI1	C	24	A	5	0.00055	imputed
rs9379687	chr6:24721787	HepG2	BHLHE40	C	26	A	9	0.00599	imputed
rs9379687	chr6:24721787	HepG2	MAX	C	95	A	49	0.00016	imputed
rs9379687	chr6:24721787	HepG2	HEY1	C	29	A	8	0.00075	imputed
rs9379687	chr6:24721787	HepG2	CREB1	C	18	A	4	0.00434	imputed
rs1165176	chr6:25830298	HepG2	FOXA1	A	28	G	5	6.62E-05	imputed
rs1165183	chr6:25836380	HepG2	POL2	G	19	A	5	0.00661	imputed
rs198853	chr6:26104096	HepG2	TBP	C	71	T	20	7.25E-08	imputed
rs198853	chr6:26104096	HepG2	SRF	C	9	T	0	0.00391	imputed
rs198853	chr6:26104096	HepG2	RXRA	C	19	T	5	0.00661	imputed
rs198853	chr6:26104096	HepG2	MYBL2	C	70	T	36	0.00124	imputed
rs198853	chr6:26104096	HepG2	POL2	C	165	T	96	2.31E-05	imputed
rs198853	chr6:26104096	HepG2	HEY1	C	98	T	52	0.00022	imputed
rs9380049	chr6:28048535	HepG2	HEY1	A	37	G	10	9.85E-05	imputed
rs9380049	chr6:28048535	HepG2	HDAC2	A	8	G	0	0.00781	imputed
rs9380049	chr6:28048535	HepG2	CREB1	A	24	G	6	0.00143	imputed
rs9380050	chr6:28048538	HepG2	HEY1	A	38	G	11	0.00014	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs9380050	chr6:28048538	HepG2	HDAC2	A	9	G	0	0.00391	imputed
rs9380050	chr6:28048538	HepG2	HNF4G	A	8	G	0	0.00781	imputed
rs9380050	chr6:28048538	HepG2	CREB1	A	22	G	5	0.00151	imputed
rs2281588	chr6:28072602	HepG2	CTCF	G	45	A	14	6.53E-05	imputed
rs2281588	chr6:28072602	HepG2	GABP	G	77	A	39	0.00053	imputed
rs2281588	chr6:28072602	HepG2	CREB1	G	24	A	8	0.007	imputed
rs17711801	chr6:28092307	HepG2	SP2	C	43	G	18	0.00187	imputed
rs17711801	chr6:28092307	HepG2	GABP	C	37	G	17	0.00907	imputed
rs9380056	chr6:28104476	HepG2	ZEB1	C	10	T	0	0.00195	imputed
rs9380056	chr6:28104476	HepG2	POL2	C	34	T	14	0.00552	imputed
rs9380057	chr6:28104634	HepG2	MAX	G	45	T	10	2.06E-06	imputed
rs9380057	chr6:28104634	HepG2	ELF1	G	33	T	13	0.00453	imputed
rs9380057	chr6:28104634	HepG2	DNASE	G	29	T	8	0.00075	imputed
rs9380057	chr6:28104634	HepG2	YY1	G	23	T	5	0.00091	imputed
rs9380057	chr6:28104634	HepG2	POL2	G	53	T	20	0.00014	imputed
rs9380057	chr6:28104634	HepG2	GABP	G	197	T	54	2.66E-20	imputed
rs9380057	chr6:28104634	HepG2	HEY1	G	33	T	10	0.00061	imputed
rs9380057	chr6:28104634	HepG2	CREB1	G	40	T	14	0.00054	imputed
rs9380057	chr6:28104634	HepG2	HNF4A	G	19	T	3	0.00086	imputed
rs9357065	chr6:28129580	HepG2	SIN3AK20	T	10	C	0	0.00195	imputed
rs9357065	chr6:28129580	HepG2	TAF1	T	31	C	2	1.31E-07	imputed
rs9357065	chr6:28129580	HepG2	MBD4	T	9	C	0	0.00391	imputed
rs9357065	chr6:28129580	HepG2	MAX	T	46	C	1	6.82E-13	imputed
rs9357065	chr6:28129580	HepG2	DNASE	T	47	C	1	3.48E-13	imputed
rs9357065	chr6:28129580	HepG2	RAD21	T	10	C	0	0.00195	imputed
rs9357065	chr6:28129580	HepG2	ZEB1	T	13	C	0	0.00024	imputed
rs9357065	chr6:28129580	HepG2	YY1	T	14	C	0	0.00012	imputed
rs9357065	chr6:28129580	HepG2	POL2	T	128	C	2	1.25E-35	imputed
rs9357065	chr6:28129580	HepG2	EZH2	T	9	C	0	0.00391	imputed
rs9357065	chr6:28129580	HepG2	SRF	T	21	C	2	6.60E-05	imputed
rs9357065	chr6:28129580	HepG2	CEBPD	T	9	C	0	0.00391	imputed
rs9357065	chr6:28129580	HepG2	CEBPB	T	15	C	0	6.10E-05	imputed
rs9357065	chr6:28129580	HepG2	ELF1	T	28	C	0	7.45E-09	imputed
rs9357065	chr6:28129580	HepG2	ZBTB7A	T	9	C	0	0.00391	imputed
rs9357065	chr6:28129580	HepG2	NR2F2	T	12	C	1	0.00342	imputed
rs9357065	chr6:28129580	HepG2	GABP	T	160	C	1	1.11E-46	imputed
rs9357065	chr6:28129580	HepG2	HEY1	T	74	C	0	1.06E-22	imputed
rs13201769	chr6:30756066	HepG2	CTCF	A	8	G	0	0.00781	imputed
rs13201769	chr6:30756066	HepG2	NR2F2	A	13	G	2	0.00739	imputed
rs13201769	chr6:30756066	HepG2	SP1	A	8	G	0	0.00781	imputed
rs3132555	chr6:31082910	HepG2	RAD21	G	13	C	1	0.00183	imputed
rs1042149	chr6:31082960	HepG2	CTCF	G	20	A	4	0.00154	imputed
rs386579266	chr6:31089982	HepG2	CEBPB	G	8	A	0	0.00781	imputed
rs6921948	chr6:31171257	HI101	FOXA2	A	27	C	3	8.43E-06	common
rs813115	chr6:31620020	HepG2	DNASE	G	35	A	11	0.00054	imputed
rs813115	chr6:31620020	HepG2	NRSF	G	12	A	1	0.00342	imputed
rs813115	chr6:31620020	HepG2	GABP	G	16	A	3	0.00443	imputed
rs4348358	chr6:32399092	HepG2	MAFK	G	75	A	42	0.00293	predicted
rs9268606	chr6:32400070	HepG2	CTCF	G	11	A	0	0.00098	imputed
rs9271092	chr6:32576296	HepG2	RAD21	A	8	G	0	0.00781	imputed
rs9271093	chr6:32576341	HepG2	CTCF	G	11	A	0	0.00098	imputed
rs9271094	chr6:32576347	HepG2	CTCF	G	15	C	0	6.10E-05	imputed
rs9271096	chr6:32576426	HepG2	CTCF	A	9	G	0	0.00391	imputed
rs17843603	chr6:32620241	HepG2	BHLHE40	G	12	A	0	0.00049	imputed
rs1063349	chr6:32627906	HepG2	HNF4A	T	15	C	0	6.10E-05	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs386699568	chr6:32627923	HepG2	HNF4A	A	16	G	0	3.05E-05	imputed
rs9274535	chr6:32634620	HepG2	RFX5	T	174	C	42	2.77E-20	predicted
rs35721478	chr6:33725129	HepG2	FOXA2	T	8	C	0	0.00781	imputed
rs35721478	chr6:33725129	HepG2	FOXA1	T	19	C	0	3.81E-06	imputed
rs206936	chr6:34302869	HepG2	POL2	A	8	G	0	0.00781	imputed
rs6912971	chr6:34355658	HepG2	MAFK	G	9	C	0	0.00391	imputed
rs9469779	chr6:34393845	HepG2	MXI1	A	26	G	8	0.00294	imputed
rs9469779	chr6:34393845	HepG2	TAF1	A	87	G	51	0.00275	imputed
rs9469779	chr6:34393845	HepG2	MAX	A	66	G	27	6.47E-05	imputed
rs9469779	chr6:34393845	HepG2	POL2	A	126	G	49	5.24E-09	imputed
rs9469779	chr6:34393845	HepG2	ELF1	A	78	G	42	0.0013	imputed
rs9469779	chr6:34393845	HepG2	HEY1	A	96	G	43	8.12E-06	imputed
rs7757900	chr6:34398879	HepG2	CEBPD	T	8	C	0	0.00781	imputed
rs9368813	chr6:34399814	HepG2	BHLHE40	C	8	T	0	0.00781	imputed
rs12192544	chr6:46620252	HepG2	EZH2	C	9	G	0	0.00391	imputed
rs283080	chr6:118606000	HepG2	RAD21	A	12	C	1	0.00342	imputed
rs7770081	chr6:133089569	HepG2	CTCF	T	99	G	44	4.89E-06	imputed
rs12211701	chr6:133119757	HepG2	DNASE	G	17	C	2	0.00073	imputed
rs9493446	chr6:133125643	HepG2	MAFK	T	22	C	5	0.00151	imputed
rs9493450	chr6:133135807	HepG2	POL2	T	263	C	117	5.00E-14	imputed
rs9493450	chr6:133135807	HepG2	HEY1	T	164	C	65	4.72E-11	imputed
rs6937795	chr6:137291281	HI88	NKX2_2	A	25	C	9	0.00904	common
rs6937795	chr6:137291281	HI32	PDX1	C	19	A	4	0.0026	common
rs6917676	chr6:137291296	HI88	NKX2_2	T	16	G	3	0.00443	common
rs6917676	chr6:137291296	HI32	PDX1	G	21	T	6	0.00592	common
rs11155000	chr6:139099401	HepG2	TCF7L2	C	9	T	0	0.00391	imputed
rs539298	chr6:160770360	HepG2	MAX	G	23	A	1	2.98E-06	imputed
rs539298	chr6:160770360	HepG2	DNASE	G	14	A	0	0.00012	imputed
rs539298	chr6:160770360	HepG2	RAD21	G	12	A	1	0.00342	imputed
rs539298	chr6:160770360	HepG2	POL2	G	24	A	1	1.55E-06	imputed
rs539298	chr6:160770360	HepG2	HNF4G	G	34	A	3	1.23E-07	imputed
rs539298	chr6:160770360	HepG2	MYBL2	G	9	A	0	0.00391	imputed
rs539298	chr6:160770360	HepG2	BHLHE40	G	15	A	0	6.10E-05	imputed
rs539298	chr6:160770360	HepG2	CEBPB	G	16	A	0	3.05E-05	imputed
rs539298	chr6:160770360	HepG2	ELF1	G	22	A	1	5.72E-06	imputed
rs539298	chr6:160770360	HepG2	NR2F2	G	44	A	2	3.08E-11	imputed
rs539298	chr6:160770360	HepG2	NFIC	G	8	A	0	0.00781	imputed
rs539298	chr6:160770360	HepG2	JUND	G	8	A	0	0.00781	imputed
rs539298	chr6:160770360	HepG2	P300	G	32	A	3	4.18E-07	imputed
rs539298	chr6:160770360	HepG2	CREB1	G	20	A	3	0.00049	imputed
rs539298	chr6:160770360	HepG2	HNF4A	G	41	A	1	1.96E-11	imputed
rs9505962	chr6:169726667	HepG2	RAD21	T	69	C	41	0.00972	imputed
rs73069540	chr7:26904770	HepG2	HDAC2	T	8	C	0	0.00781	imputed
rs10281169	chr7:36922825	HepG2	CTCF	A	25	G	2	5.65E-06	imputed
rs10281169	chr7:36922825	HepG2	RAD21	A	11	G	1	0.00635	imputed
rs8200	chr7:75696606	HepG2	POL2	G	17	C	3	0.00258	imputed
rs10953284	chr7:77169782	HepG2	DNASE	G	17	C	0	1.53E-05	imputed
rs10953284	chr7:77169782	HepG2	HNF4G	G	14	C	0	0.00012	imputed
rs10953284	chr7:77169782	HepG2	CEBPB	G	16	C	1	0.00027	imputed
rs10953284	chr7:77169782	HepG2	HNF4A	G	26	C	1	4.17E-07	imputed
rs705379	chr7:94953895	HepG2	DNASE	G	10	A	0	0.00195	imputed
rs776745	chr7:99291337	HepG2	FOXA1	G	32	T	12	0.00366	imputed
rs6962760	chr7:141465363	HI32	FOXA2	T	19	C	4	0.0026	common
rs11783893	chr8:2100976	HepG2	CTCF	C	8	G	0	0.00781	imputed
rs35382339	chr8:8559255	HepG2	CTCF	A	21	G	2	6.60E-05	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs35382339	chr8:8559255	HepG2	DNASE	G	17	A	4	0.0072	imputed
rs6984305	chr8:9178268	HepG2	CTCF	T	46	A	16	0.00018	imputed
rs11777082	chr8:39797703	HepG2	BHLHE40	A	8	G	0	0.00781	imputed
rs2279128	chr8:71581559	HepG2	BHLHE40	T	30	G	10	0.00222	imputed
rs6985299	chr8:71613079	HepG2	MAX	T	20	C	6	0.00936	imputed
rs11985375	chr8:71613472	HepG2	MAX	G	35	A	7	1.51E-05	imputed
rs11985375	chr8:71613472	HepG2	CREB1	G	23	A	6	0.00232	imputed
rs10810299	chr9:14964272	HepG2	POL2	A	10	C	0	0.00195	imputed
rs34428576	chr10:12281111	HepG2	MAX	G	12	A	0	0.00049	imputed
rs34428576	chr10:12281111	HepG2	DNASE	G	8	A	0	0.00781	imputed
rs34428576	chr10:12281111	HepG2	RAD21	G	11	A	0	0.00098	imputed
rs34428576	chr10:12281111	HepG2	CEBPD	G	18	A	0	7.63E-06	imputed
rs34428576	chr10:12281111	HepG2	CEBPB	G	120	A	0	1.50E-36	imputed
rs34428576	chr10:12281111	HepG2	ZBTB7A	G	14	A	0	0.00012	imputed
rs34428576	chr10:12281111	HepG2	FOXA1	G	14	A	0	0.00012	imputed
rs34428576	chr10:12281111	HepG2	NR2F2	G	11	A	0	0.00098	imputed
rs34428576	chr10:12281111	HepG2	NFIC	G	8	A	0	0.00781	imputed
rs34428576	chr10:12281111	HepG2	JUND	G	11	A	0	0.00098	imputed
rs34428576	chr10:12281111	HepG2	HDAC2	G	14	A	0	0.00012	imputed
rs34428576	chr10:12281111	HepG2	CEBPA	G	57	A	0	1.39E-17	imputed
rs34428576	chr10:12281111	HepG2	P300	G	15	A	0	6.10E-05	imputed
rs34428576	chr10:12281111	HepG2	CREB1	G	13	A	0	0.00024	imputed
rs34428576	chr10:12281111	HepG2	HNF4A	G	12	A	0	0.00049	imputed
rs1414395	chr10:13334136	HepG2	CEBPB	T	14	G	2	0.00418	imputed
rs4747275	chr10:16552472	HepG2	MAFK	G	17	A	2	0.00073	imputed
rs17141322	chr10:17604700	HI32	PDX1	A	19	C	4	0.0026	common
rs16916563	chr10:63507642	HepG2	FOXA2	G	11	A	0	0.00098	imputed
rs4933736	chr10:94471595	HepG2	FOXA2	T	8	C	0	0.00781	imputed
rs11190179	chr10:101365313	HepG2	CEBPB	G	17	A	0	1.53E-05	imputed
rs2295776	chr10:102295629	HepG2	BHLHE40	G	8	T	0	0.00781	imputed
rs2495758	chr10:102321900	HepG2	CTCF	C	275	G	3	1.47E-77	imputed
rs2495758	chr10:102321900	HepG2	MAX	C	28	G	0	7.45E-09	imputed
rs2495758	chr10:102321900	HepG2	DNASE	C	14	G	1	0.00098	imputed
rs2495758	chr10:102321900	HepG2	RAD21	C	153	G	0	1.75E-46	imputed
rs2495758	chr10:102321900	HepG2	YY1	C	15	G	0	6.10E-05	imputed
rs2495758	chr10:102321900	HepG2	CHD2	C	8	G	0	0.00781	imputed
rs2495758	chr10:102321900	HepG2	ELF1	C	25	G	0	5.96E-08	imputed
rs2495758	chr10:102321900	HepG2	CREB1	C	22	G	2	3.59E-05	imputed
rs112699822	chr10:104692633	HepG2	CEBPA	A	8	C	0	0.00781	imputed
rs1926032	chr10:104829469	HepG2	CTCF	C	17	T	3	0.00258	imputed
rs10510007	chr10:116636721	HepG2	CEBPB	G	11	A	0	0.00098	imputed
rs10835531	chr11:1516110	HI101	FOXA2	A	20	G	6	0.00936	common
rs538954	chr11:65756808	HepG2	ZBTB7A	C	8	T	0	0.00781	imputed
rs12288023	chr11:67421341	HepG2	CEBPB	T	20	C	0	1.91E-06	imputed
rs613128	chr11:68638058	HepG2	MAX	G	11	T	0	0.00098	imputed
rs514833	chr11:68657734	HepG2	CTCF	C	32	T	0	4.66E-10	imputed
rs514833	chr11:68657734	HepG2	DNASE	C	10	T	0	0.00195	imputed
rs514833	chr11:68657734	HepG2	RAD21	C	23	T	0	2.38E-07	imputed
rs514833	chr11:68657734	HepG2	ZBTB33	C	16	T	0	3.05E-05	imputed
rs629426	chr11:68671104	HepG2	YY1	G	23	A	6	0.00232	imputed
rs9787897	chr11:74659302	HepG2	FOXA2	T	9	A	0	0.00391	imputed
rs567956	chr11:74659779	HepG2	POL2	C	9	T	0	0.00391	imputed
rs2165163	chr11:74660143	HepG2	MAX	C	69	G	40	0.00704	imputed
rs2165163	chr11:74660143	HepG2	DNASE	C	28	G	8	0.00119	imputed
rs4944968	chr11:74724276	HepG2	DNASE	C	12	G	1	0.00342	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs3781884	chr11:113217364	HepG2	CTCF	G	17	A	0	1.53E-05	imputed
rs180351	chr11:116607641	HepG2	CTCF	T	82	C	15	2.38E-12	imputed
rs7109649	chr11:116943544	HepG2	FOXA2	T	27	C	6	0.00032	imputed
rs7109649	chr11:116943544	HI87	NKX2_2	T	21	C	6	0.00592	common
rs11218752	chr11:122552600	HepG2	CTCF	C	11	T	1	0.00635	imputed
rs57246949	chr11:129476625	HepG2	CEBPB	G	10	A	0	0.00195	imputed
rs2534721	chr12:6580144	HepG2	DNASE	C	89	A	56	0.00766	imputed
rs933462	chr12:9103665	HepG2	HDAC2	G	8	T	0	0.00781	imputed
rs2417257	chr12:13444717	HepG2	CTCF	C	8	T	0	0.00781	imputed
rs2239182	chr12:48255411	HepG2	FOXA1	C	15	T	2	0.00235	imputed
rs4509811	chr12:58335142	HepG2	CEBPD	A	8	G	0	0.00781	imputed
rs11612569	chr12:101800809	HepG2	CEBPB	A	10	G	0	0.00195	imputed
rs869916	chr12:103244013	HepG2	HNF4A	T	23	G	5	0.00091	imputed
rs4764939	chr12:103522952	HepG2	HNF4A	C	8	T	0	0.00781	imputed
rs12828810	chr12:121152017	HepG2	CTCF	G	15	T	2	0.00235	imputed
rs7139079	chr12:121415293	HepG2	POL2	A	9	G	0	0.00391	imputed
rs2258287	chr12:121454313	HepG2	USF1	A	38	C	3	1.05E-08	imputed
rs2258287	chr12:121454313	HepG2	POL2	A	31	C	7	0.00012	imputed
rs1154513	chr12:122391963	HI32	FOXA2	A	15	G	3	0.00754	common
rs12864047	chr13:74796108	HepG2	HNF4G	C	11	T	1	0.00635	imputed
rs9302064	chr13:95966851	HepG2	FOXA1	A	48	C	12	3.18E-06	imputed
rs9302064	chr13:95966851	HepG2	JUND	A	17	C	4	0.0072	imputed
rs1010461	chr14:21153788	HepG2	POL2	C	16	A	3	0.00443	imputed
rs17109371	chr14:25429892	HepG2	ATF3	C	10	T	0	0.00195	imputed
rs10138510	chr14:25430134	HepG2	ZBTB33	G	8	A	0	0.00781	imputed
rs76138569	chr14:25512157	HepG2	MAFK	C	53	T	27	0.00487	imputed
rs1769591	chr14:34378886	HepG2	FOXA1	G	96	A	59	0.00369	imputed
rs11624787	chr14:53288450	HepG2	FOXA2	C	11	G	0	0.00098	imputed
rs11624787	chr14:53288450	HepG2	CEBPB	C	9	G	0	0.00391	imputed
rs11624787	chr14:53288450	HepG2	FOXA1	C	22	G	0	4.77E-07	imputed
rs11624787	chr14:53288450	HepG2	P300	C	12	G	0	0.00049	imputed
rs11624787	chr14:53288450	HepG2	HNF4A	C	11	G	0	0.00098	imputed
rs17090719	chr14:94846661	HepG2	MBD4	T	9	C	0	0.00391	imputed
rs17090719	chr14:94846661	HepG2	POL2	T	79	C	0	3.31E-24	imputed
rs17090719	chr14:94846661	HepG2	HEY1	T	13	C	0	0.00024	imputed
rs2034652	chr15:40802768	HepG2	CEBPB	A	8	G	0	0.00781	imputed
rs8036737	chr15:40874256	HepG2	MAFK	G	8	A	0	0.00781	imputed
rs8042519	chr15:45996341	HepG2	CTCF	C	14	G	0	0.00012	imputed
rs11857380	chr15:58712203	HepG2	FOXA1	G	15	T	0	6.10E-05	imputed
rs7178540	chr15:62380132	HI87	MAFB	G	22	A	7	0.00813	common
rs11854147	chr15:75052771	HepG2	FOXA1	T	77	C	36	0.00014	imputed
rs11072502	chr15:75052820	HepG2	FOXA1	G	82	A	34	9.69E-06	imputed
rs11072506	chr15:75052994	HepG2	POL2	A	176	G	113	0.00025	imputed
rs11072506	chr15:75052994	HepG2	HNF4A	A	8	G	0	0.00781	imputed
rs11857695	chr15:75165751	HepG2	JUND	T	28	G	1	1.12E-07	imputed
rs11857695	chr15:75165751	HepG2	CREB1	T	58	G	16	9.67E-07	imputed
rs7175950	chr15:78236353	HepG2	POL2	A	11	G	0	0.00098	imputed
rs11856536	chr15:79094325	HepG2	HDAC2	A	8	G	0	0.00781	imputed
rs4932370	chr15:91404705	HI102	NKX6_1	G	24	A	8	0.007	common
rs9925556	chr16:2880105	HepG2	DNASE	T	9	C	0	0.00391	imputed
rs9925556	chr16:2880105	HepG2	FOXA1	T	21	C	1	1.10E-05	imputed
rs149597	chr16:11343942	HepG2	POL2	G	15	C	1	0.00052	imputed
rs243330	chr16:11350991	HepG2	POL2	C	18	T	2	0.0004	imputed
rs243329	chr16:11352313	HepG2	BHLHE40	A	8	T	0	0.00781	imputed
rs376374	chr16:11370616	HepG2	MAX	A	10	G	0	0.00195	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs11640295	chr16:11390728	HepG2	POL2	G	25	C	6	0.00088	imputed
rs11640295	chr16:11390728	HepG2	HEY1	G	20	C	6	0.00936	imputed
rs7189239	chr16:11402515	HepG2	MXI1	T	8	C	0	0.00781	imputed
rs7189239	chr16:11402515	HepG2	MAX	T	30	C	3	1.40E-06	imputed
rs7189239	chr16:11402515	HepG2	ZEB1	T	15	C	2	0.00235	imputed
rs7189239	chr16:11402515	HepG2	POL2	T	15	C	1	0.00052	imputed
rs7189239	chr16:11402515	HepG2	HDAC2	T	13	C	1	0.00183	imputed
rs7189239	chr16:11402515	HepG2	HNF4A	T	44	C	17	0.00073	imputed
rs8059989	chr16:12185746	HepG2	MAFK	T	26	C	9	0.00599	imputed
rs9925009	chr16:12186352	HepG2	P300	A	8	T	0	0.00781	imputed
rs4780465	chr16:12707523	HepG2	TEAD4	T	12	C	1	0.00342	imputed
rs4780465	chr16:12707523	HepG2	MBD4	T	8	C	0	0.00781	imputed
rs4780465	chr16:12707523	HepG2	DNASE	T	24	C	8	0.007	imputed
rs4780465	chr16:12707523	HepG2	POL2	T	26	C	4	5.95E-05	imputed
rs4780465	chr16:12707523	HepG2	BHLHE40	T	17	C	3	0.00258	imputed
rs4780465	chr16:12707523	HepG2	NR2F2	T	24	C	6	0.00143	imputed
rs4780465	chr16:12707523	HepG2	P300	T	35	C	15	0.0066	imputed
rs4780465	chr16:12707523	HepG2	HNF4A	T	29	C	7	0.00031	imputed
rs1560104	chr16:12708208	HepG2	MAX	C	14	T	0	0.00012	imputed
rs1560104	chr16:12708208	HepG2	POL2	C	28	T	9	0.00256	imputed
rs1560104	chr16:12708208	HepG2	CEBPD	C	9	T	0	0.00391	imputed
rs11075256	chr16:15187912	HepG2	MXI1	G	18	C	4	0.00434	imputed
rs11075256	chr16:15187912	HepG2	TAF1	G	13	C	2	0.00739	imputed
rs11075256	chr16:15187912	HepG2	MAX	G	50	C	19	0.00024	imputed
rs11075256	chr16:15187912	HepG2	POL2	G	60	C	24	0.00011	imputed
rs11075256	chr16:15187912	HepG2	SP2	G	9	C	0	0.00391	imputed
rs11075256	chr16:15187912	HepG2	HEY1	G	17	C	3	0.00258	imputed
rs7194098	chr16:20464350	HepG2	POL2	C	10	G	0	0.00195	imputed
rs7194098	chr16:20464350	HepG2	SRF	C	9	G	0	0.00391	imputed
rs9937581	chr16:20473903	HepG2	MAFK	G	56	A	0	2.78E-17	imputed
rs9937581	chr16:20473903	HepG2	MAFF	G	18	A	1	7.63E-05	imputed
rs1394678	chr16:20491058	HepG2	CTCF	C	8	T	0	0.00781	imputed
rs62032983	chr16:23653343	HepG2	ELF1	T	12	C	0	0.00049	imputed
rs62032983	chr16:23653343	HepG2	CREB1	T	8	C	0	0.00781	imputed
rs181203	chr16:28512371	HepG2	EZH2	A	8	C	0	0.00781	imputed
rs62034319	chr16:28532188	HepG2	HDAC2	T	11	G	1	0.00635	imputed
rs2106480	chr16:28537971	HepG2	DNASE	T	35	C	15	0.0066	imputed
rs2106480	chr16:28537971	HepG2	CEBPB	T	44	C	20	0.00369	imputed
rs2106480	chr16:28537971	HepG2	FOXA1	T	45	C	19	0.00156	imputed
rs2106480	chr16:28537971	HepG2	P300	T	44	C	18	0.0013	imputed
rs2106480	chr16:28537971	HepG2	HNF4A	T	52	C	27	0.00655	imputed
rs62034351	chr16:28565489	HepG2	DNASE	G	62	A	25	9.06E-05	imputed
rs62034351	chr16:28565489	HepG2	POL2	G	80	A	40	0.00033	imputed
rs62034351	chr16:28565489	HepG2	HEY1	G	57	A	13	1.03E-07	imputed
rs7191618	chr16:28565667	HepG2	MAX	C	55	G	29	0.00604	imputed
rs7191618	chr16:28565667	HepG2	DNASE	C	50	G	21	0.00077	imputed
rs7191618	chr16:28565667	HepG2	RAD21	C	13	G	1	0.00183	imputed
rs743590	chr16:28608230	HepG2	MAX	G	18	A	3	0.00149	imputed
rs743590	chr16:28608230	HepG2	YY1	G	13	A	0	0.00024	imputed
rs743590	chr16:28608230	HepG2	POL2	G	25	A	7	0.0021	imputed
rs62031562	chr16:28609329	HepG2	POL2	A	21	T	4	0.00091	imputed
rs7187776	chr16:28857645	HepG2	HNF4G	A	8	G	0	0.00781	imputed
rs7187776	chr16:28857645	HepG2	HEY1	A	25	G	6	0.00088	imputed
rs7187776	chr16:28857645	HepG2	CREB1	A	104	G	47	4.03E-06	imputed
rs62037367	chr16:28874547	HepG2	CREB1	C	27	G	5	0.00011	imputed
rs7198606	chr16:28875122	HepG2	SRF	T	8	G	0	0.00781	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs7198606	chr16:28875122	HepG2	SP2	T	59	G	27	0.00073	imputed
rs7198606	chr16:28875122	HepG2	CREB1	T	31	G	10	0.00145	imputed
rs11864750	chr16:28875204	HepG2	NR2F2	A	8	T	0	0.00781	imputed
rs4072402	chr16:28937259	HepG2	CTCF	C	10	T	0	0.00195	imputed
rs2303222	chr16:31085470	HepG2	POL2	C	29	T	10	0.00338	imputed
rs1981760	chr16:50723074	HepG2	EZH2	C	9	T	0	0.00391	imputed
rs12720926	chr16:56998918	HepG2	DNASE	A	29	G	9	0.00166	imputed
rs7199443	chr16:67841129	HepG2	MAX	G	21	T	1	1.10E-05	imputed
rs7196789	chr16:67927124	HepG2	YY1	T	71	C	38	0.00203	imputed
rs1134760	chr16:67964203	HepG2	POL2	C	11	T	1	0.00635	imputed
rs20549	chr16:67969930	HepG2	POL2	G	62	A	31	0.00171	imputed
rs1109166	chr16:67977382	HepG2	FOXA2	C	63	T	29	0.00051	imputed
rs1109166	chr16:67977382	HepG2	HNF4G	C	41	T	19	0.00622	imputed
rs1109166	chr16:67977382	HepG2	FOXA1	C	86	T	47	0.00091	imputed
rs1109166	chr16:67977382	HepG2	NR2F2	C	101	T	44	2.49E-06	imputed
rs1109166	chr16:67977382	HepG2	CREB1	C	26	T	7	0.00132	imputed
rs1109166	chr16:67977382	HepG2	HNF4A	C	78	T	41	0.00089	imputed
rs35223604	chr16:70052345	HepG2	CTCF	C	25	G	8	0.00455	imputed
rs4985407	chr16:70285901	HepG2	CTCF	G	78	A	39	0.0004	imputed
rs4985407	chr16:70285901	HepG2	SIN3AK20	G	17	A	1	0.00014	imputed
rs4985407	chr16:70285901	HepG2	MAX	G	116	A	31	9.39E-13	imputed
rs4985407	chr16:70285901	HepG2	DNASE	G	75	A	22	6.07E-08	imputed
rs4985407	chr16:70285901	HepG2	YY1	G	49	A	16	5.08E-05	imputed
rs4985407	chr16:70285901	HepG2	POL2	G	54	A	24	0.0009	imputed
rs4985407	chr16:70285901	HepG2	CHD2	G	12	A	1	0.00342	imputed
rs4985407	chr16:70285901	HepG2	BHLHE40	G	15	A	3	0.00754	imputed
rs4985407	chr16:70285901	HepG2	ELF1	G	31	A	8	0.00029	imputed
rs4985407	chr16:70285901	HepG2	GABP	G	40	A	19	0.00864	imputed
rs4985407	chr16:70285901	HepG2	CREB1	G	52	A	19	0.00011	imputed
rs8052763	chr16:75251659	HepG2	HNF4A	C	8	G	0	0.00781	imputed
rs12923290	chr16:78079956	HepG2	YY1	G	15	A	2	0.00235	imputed
rs12923290	chr16:78079956	HepG2	NRSF	G	434	A	290	9.74E-08	imputed
rs12921945	chr16:78080133	HepG2	YY1	T	16	G	3	0.00443	imputed
rs12921945	chr16:78080133	HepG2	NRSF	T	322	G	182	4.64E-10	imputed
rs12923626	chr16:78080139	HepG2	YY1	C	16	A	3	0.00443	imputed
rs12923626	chr16:78080139	HepG2	NRSF	C	288	A	143	2.52E-12	imputed
rs386792435	chr16:78080141	HepG2	YY1	G	16	T	3	0.00443	imputed
rs386792435	chr16:78080141	HepG2	NRSF	G	280	T	130	9.91E-14	imputed
rs12923218	chr16:78080274	HepG2	NRSF	G	27	C	6	0.00032	imputed
rs4888731	chr16:78080418	HepG2	NRSF	G	11	A	1	0.00635	imputed
rs12448415	chr16:87871096	HepG2	POL2	G	15	A	2	0.00235	imputed
rs12931876	chr16:87874182	HepG2	CTCF	T	13	C	1	0.00183	imputed
rs34508683	chr16:87876375	HepG2	POL2	C	15	T	3	0.00754	imputed
rs34508683	chr16:87876375	HepG2	EZH2	C	8	T	0	0.00781	imputed
rs28609922	chr16:87876631	HepG2	MAX	A	45	C	13	3.01E-05	imputed
rs28609922	chr16:87876631	HepG2	BHLHE40	A	23	C	6	0.00232	imputed
rs28609922	chr16:87876631	HepG2	P300	A	8	C	0	0.00781	imputed
rs71391360	chr16:87877313	HepG2	POL2	G	13	C	1	0.00183	imputed
rs4843270	chr16:87878072	HepG2	CTCF	A	14	C	2	0.00418	imputed
rs4843270	chr16:87878072	HepG2	POL2	A	18	C	0	7.63E-06	imputed
rs386793901	chr16:87878076	HepG2	POL2	A	15	G	0	6.10E-05	imputed
rs4843718	chr16:87878476	HepG2	CREB1	A	10	G	0	0.00195	imputed
rs35459492	chr16:87878883	HepG2	POL2	G	11	C	1	0.00635	imputed
rs876985	chr16:87882124	HepG2	P300	C	17	T	4	0.0072	imputed
rs747687	chr17:775334	HepG2	ZBTB33	G	8	C	0	0.00781	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs78601373	chr17:776345	HepG2	MAFK	C	25	T	0	5.96E-08	imputed
rs78601373	chr17:776345	HepG2	MAFF	C	13	T	0	0.00024	imputed
rs507506	chr17:7118322	HepG2	RAD21	G	28	A	11	0.00948	imputed
rs2017365	chr17:7122624	HepG2	MAX	G	14	A	1	0.00098	imputed
rs62066869	chr17:20059342	HepG2	DNASE	A	54	C	26	0.00232	imputed
rs2526479	chr17:20130221	HepG2	CTCF	G	140	T	93	0.00251	imputed
rs2526479	chr17:20130221	HepG2	MAX	G	13	T	0	0.00024	imputed
rs1453559	chr17:38020419	HepG2	MXI1	C	8	T	0	0.00781	imputed
rs1453559	chr17:38020419	HepG2	DNASE	C	9	T	0	0.00391	imputed
rs1453559	chr17:38020419	HepG2	BHLHE40	C	9	T	0	0.00391	imputed
rs12936231	chr17:38029120	HepG2	CTCF	G	35	C	0	5.82E-11	imputed
rs12936231	chr17:38029120	HepG2	RAD21	G	41	C	0	9.09E-13	imputed
rs12936231	chr17:38029120	HepG2	ARID3A	G	14	C	1	0.00098	imputed
rs386796961	chr17:38072247	HepG2	POL2	G	8	T	0	0.00781	imputed
rs7224129	chr17:38075426	HepG2	POL2	G	17	A	4	0.0072	imputed
rs4065275	chr17:38080865	HepG2	CTCF	G	43	A	2	5.89E-11	imputed
rs4065275	chr17:38080865	HepG2	DNASE	A	17	G	3	0.00258	imputed
rs4065275	chr17:38080865	HepG2	RAD21	G	15	A	2	0.00235	imputed
rs8076131	chr17:38080912	HepG2	CTCF	A	28	G	2	8.68E-07	imputed
rs8076131	chr17:38080912	HepG2	RAD21	A	24	G	5	0.00055	imputed
rs8076131	chr17:38080912	HepG2	HEY1	G	9	A	0	0.00391	imputed
ss56891470	chr17:41438468	HI32	POL2	T	73	C	34	0.00021	predicted
rs12938996	chr17:41438674	HI32	POL2	A	81	G	50	0.00851	predicted
rs17742347	chr17:41846468	HepG2	POL2	C	19	T	4	0.0026	imputed
rs17674998	chr17:41879544	HepG2	ZBTB33	A	8	G	0	0.00781	imputed
rs9901676	chr17:41911818	HepG2	EZH2	T	9	C	0	0.00391	imputed
rs12948653	chr17:46259254	HepG2	CTCF	A	24	C	8	0.007	imputed
rs16949649	chr17:49230308	HepG2	CEBPB	T	8	C	0	0.00781	imputed
rs6503905	chr17:57287454	HepG2	ELF1	A	15	G	1	0.00052	imputed
rs6503905	chr17:57287454	HepG2	CREB1	A	21	G	6	0.00592	imputed
rs8076760	chr17:61920497	HepG2	MAX	T	32	C	6	2.43E-05	imputed
rs8076760	chr17:61920497	HepG2	TCF12	T	9	C	0	0.00391	imputed
rs8076760	chr17:61920497	HepG2	CHD2	T	113	C	52	2.33E-06	imputed
rs8076760	chr17:61920497	HepG2	GABP	T	91	C	44	6.43E-05	imputed
rs6808	chr17:62400575	HepG2	CREB1	C	8	G	0	0.00781	imputed
rs12936766	chr17:62408949	HepG2	CEBPB	G	19	C	3	0.00086	imputed
rs4968721	chr17:62409586	HepG2	FOXA2	G	17	C	1	0.00014	imputed
rs4968721	chr17:62409586	HepG2	FOXA1	G	26	C	8	0.00294	imputed
rs4968721	chr17:62409586	HepG2	JUND	G	25	C	4	0.0001	imputed
rs4968721	chr17:62409586	HepG2	P300	G	17	C	3	0.00258	imputed
rs4366742	chr17:64212242	HepG2	POL2	T	21	C	5	0.00249	imputed
rs8064837	chr17:64242703	HepG2	SRF	G	8	C	0	0.00781	imputed
rs2598414	chr17:74067099	HepG2	BHLHE40	C	9	T	0	0.00391	imputed
rs4969182	chr17:76393030	HepG2	FOXA2	T	30	C	4	6.16E-06	imputed
rs4969182	chr17:76393030	HepG2	MAX	T	13	C	2	0.00739	imputed
rs4969182	chr17:76393030	HepG2	MYBL2	T	10	C	0	0.00195	imputed
rs4969182	chr17:76393030	HepG2	CEBPB	T	12	C	1	0.00342	imputed
rs4969182	chr17:76393030	HepG2	FOXA1	T	71	C	1	3.09E-20	imputed
rs4969182	chr17:76393030	HepG2	NR2F2	T	11	C	0	0.00098	imputed
rs4969183	chr17:76393372	HepG2	BHLHE40	A	9	G	0	0.00391	imputed
rs2376585	chr17:76417883	HepG2	CTCF	T	12	C	0	0.00049	imputed
rs2376585	chr17:76417883	HepG2	FOXA2	T	97	C	36	1.19E-07	imputed
rs2376585	chr17:76417883	HepG2	MAX	T	34	C	15	0.0094	imputed
rs2376585	chr17:76417883	HepG2	DNASE	T	68	C	20	2.77E-07	imputed
rs2376585	chr17:76417883	HepG2	ZEB1	T	18	C	3	0.00149	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs2376585	chr17:76417883	HepG2	FOXA1	T	211	C	101	4.52E-10	imputed
rs2376585	chr17:76417883	HepG2	NR2F2	T	31	C	12	0.0054	imputed
rs2376585	chr17:76417883	HepG2	CEBPA	T	44	C	15	0.0002	imputed
rs2376585	chr17:76417883	HepG2	ZBTB33	T	17	C	1	0.00014	imputed
rs62078747	chr17:80055206	HepG2	MXI1	C	10	G	0	0.00195	imputed
rs62078747	chr17:80055206	HepG2	MAX	C	29	G	9	0.00166	imputed
rs59251877	chr17:80056498	HepG2	MAX	G	65	A	30	0.00042	imputed
rs59251877	chr17:80056498	HepG2	JUND	G	8	A	0	0.00781	imputed
rs7225637	chr17:80059758	HepG2	TAF1	G	8	A	0	0.00781	imputed
rs7225637	chr17:80059758	HepG2	ELF1	G	42	A	15	0.00046	imputed
rs11658040	chr17:80059891	HepG2	MAX	C	30	T	11	0.00432	imputed
rs11658040	chr17:80059891	HepG2	DNASE	C	33	T	14	0.00794	imputed
rs11658040	chr17:80059891	HepG2	HNF4A	C	17	T	0	1.53E-05	imputed
rs9908277	chr17:80060829	HepG2	MXI1	T	41	C	16	0.00126	imputed
rs9908277	chr17:80060829	HepG2	MAX	T	197	C	107	2.74E-07	imputed
rs9894129	chr17:80075700	HepG2	MAX	A	25	G	4	0.0001	imputed
rs9894129	chr17:80075700	HepG2	CEBPB	A	29	G	11	0.00643	imputed
rs9894129	chr17:80075700	HepG2	HDAC2	A	12	G	1	0.00342	imputed
rs9894129	chr17:80075700	HepG2	HNF4A	A	22	G	3	0.00016	imputed
rs9916649	chr17:80075739	HepG2	FOXA2	G	8	A	0	0.00781	imputed
rs9916649	chr17:80075739	HepG2	MAX	G	32	A	10	0.00094	imputed
rs9916649	chr17:80075739	HepG2	RXRA	G	19	A	4	0.0026	imputed
rs9916649	chr17:80075739	HepG2	BHLHE40	G	14	A	0	0.00012	imputed
rs9916649	chr17:80075739	HepG2	SP1	G	27	A	9	0.00393	imputed
rs9916649	chr17:80075739	HepG2	CREB1	G	17	A	3	0.00258	imputed
rs7218075	chr17:80076808	HepG2	FOXA2	C	11	G	1	0.00635	imputed
rs7218075	chr17:80076808	HepG2	POL2	C	8	G	0	0.00781	imputed
rs62079996	chr17:80076862	HepG2	FOXA1	A	20	G	4	0.00154	imputed
rs6502065	chr17:80095642	HepG2	CEBPB	C	22	T	2	3.59E-05	imputed
rs182498	chr18:21079363	HepG2	RAD21	C	8	T	0	0.00781	imputed
rs4800162	chr18:21117419	HepG2	FOXA1	G	13	T	0	0.00024	imputed
rs12607673	chr18:50906636	HepG2	CTCF	T	46	C	18	0.00062	imputed
rs12607674	chr18:50906642	HepG2	CTCF	T	38	C	15	0.00219	imputed
rs34589926	chr18:50906676	HepG2	CTCF	G	29	T	6	0.00012	imputed
rs7256735	chr19:2169121	HI88	NKX2_2	T	29	G	8	0.00075	common
rs10410204	chr19:7224350	HepG2	FOXA2	C	21	T	0	9.54E-07	imputed
rs10410204	chr19:7224350	HepG2	FOXA1	C	48	T	1	1.78E-13	imputed
rs7248104	chr19:7224431	HepG2	FOXA2	A	89	G	3	5.24E-23	imputed
rs7248104	chr19:7224431	HepG2	MAX	A	38	G	2	1.49E-09	imputed
rs7248104	chr19:7224431	HepG2	YY1	A	12	G	0	0.00049	imputed
rs7248104	chr19:7224431	HepG2	FOXA1	A	113	G	3	6.27E-30	imputed
rs7248104	chr19:7224431	HepG2	NR2F2	A	9	G	0	0.00391	imputed
rs7259455	chr19:11253310	HepG2	DNASE	C	61	T	13	1.39E-08	imputed
rs2303696	chr19:18548884	HepG2	DNASE	C	8	T	0	0.00781	imputed
rs8103622	chr19:18572834	HepG2	CTCF	C	118	T	63	5.29E-05	imputed
rs8101689	chr19:30185697	HepG2	HDAC2	A	9	G	0	0.00391	imputed
rs62102718	chr19:33891013	HepG2	HNF4G	A	15	T	1	0.00052	imputed
rs55792845	chr19:37498685	HI87	NKX2_2	C	27	T	3	8.43E-06	common
rs296368	chr19:48372298	HepG2	RAD21	T	71	C	42	0.00815	imputed
rs1343703	chr19:49955155	HI32	FOXA2	G	29	C	8	0.00075	common
rs1343703	chr19:49955155	HepG2	ARID3A	C	31	G	12	0.0054	imputed
rs39714	chr19:54693682	HI87	NKX2_2	C	20	G	6	0.00936	common
rs36624	chr19:54693868	HI87	NKX2_2	C	34	T	13	0.00309	common
rs34541537	chr19:58912737	HepG2	CREB1	C	55	T	18	1.69E-05	imputed
rs35138622	chr19:58912788	HepG2	CREB1	C	88	A	26	4.59E-09	imputed

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

				Enriched		Other			Variant
Variant	Position	Sample	Protein/Assay	Allele	Reads	Allele	Reads	P-value[a]	Source[b]
rs35117909	chr19:58912906	HepG2	CREB1	G	37	A	15	0.00319	imputed
rs11878203	chr19:58920050	HepG2	DNASE	G	42	A	13	0.00011	imputed
rs11697620	chr20:3150165	HepG2	CTCF	C	21	T	1	1.10E-05	imputed
rs11697620	chr20:3150165	HepG2	MAX	C	12	T	1	0.00342	imputed
rs1321940	chr20:12959885	HepG2	FOXA1	G	67	A	37	0.00423	imputed
rs13042787	chr20:17436571	HepG2	BHLHE40	T	8	C	0	0.00781	imputed
rs6111720	chr20:17868623	HepG2	CEBPD	T	8	C	0	0.00781	imputed
rs6060266	chr20:33733078	HI34	CTCF	C	42	T	16	0.00086	common
rs4812816	chr20:42930872	HepG2	MAZ	C	8	A	0	0.00781	imputed
rs6065723	chr20:42956922	HepG2	MAFK	C	24	T	5	0.00055	imputed
rs6065723	chr20:42956922	HepG2	MAFF	C	8	T	0	0.00781	imputed
rs6073534	chr20:43365504	HepG2	USF1	C	25	T	7	0.0021	imputed
rs6073538	chr20:43379264	HepG2	DNASE	C	8	A	0	0.00781	imputed
rs386814838	chr20:48092076	HepG2	CEBPB	G	12	A	1	0.00342	imputed
rs6068599	chr20:52259618	HepG2	TBP	C	13	T	0	0.00024	imputed
rs6068599	chr20:52259618	HepG2	CTCF	C	9	T	0	0.00391	imputed
rs6068599	chr20:52259618	HepG2	MBD4	C	12	T	0	0.00049	imputed
rs6068599	chr20:52259618	HepG2	MAX	C	15	T	0	6.10E-05	imputed
rs6068599	chr20:52259618	HepG2	RAD21	C	8	T	0	0.00781	imputed
rs6068599	chr20:52259618	HepG2	ZEB1	C	13	T	0	0.00024	imputed
rs6068599	chr20:52259618	HepG2	POL2	C	32	T	0	4.66E-10	imputed
rs6068599	chr20:52259618	HepG2	ARID3A	C	13	T	0	0.00024	imputed
rs6068599	chr20:52259618	HepG2	MAFK	C	8	T	0	0.00781	imputed
rs6068599	chr20:52259618	HepG2	MYBL2	C	11	T	0	0.00098	imputed
rs6068599	chr20:52259618	HepG2	NR2F2	C	10	T	0	0.00195	imputed
rs6068599	chr20:52259618	HepG2	FOXA1	C	10	T	0	0.00195	imputed
rs6068599	chr20:52259618	HepG2	HEY1	C	19	T	0	3.81E-06	imputed
rs6068599	chr20:52259618	HepG2	NFIC	C	9	T	0	0.00391	imputed
rs6068599	chr20:52259618	HepG2	HDAC2	C	10	T	0	0.00195	imputed
rs6068599	chr20:52259618	HepG2	P300	C	23	T	0	2.38E-07	imputed
rs6068599	chr20:52259618	HepG2	CREB1	C	8	T	0	0.00781	imputed
rs2041278	chr20:52268995	HepG2	FOXA2	G	27	T	10	0.00763	imputed
rs2832245	chr21:30554634	HI32	FOXA2	C	30	T	7	0.00019	common
rs6000200	chr22:36612258	HI32	PDX1	C	31	G	12	0.0054	common
rs4828057	chrX:100006043	HI81	MAFB	A	20	C	5	0.00408	common

Only imbalances sites at eQTL or cardiometabolic GWAS loci are shown. [a]binomial P-value [b]Heterozygous sites were imputed, common variants (MAF>.05 1000G EUR) or predicted from the sequence data

APPENDIX 2: Allelic imbalance at eQTL loci (P<1.0x10^{-5})

		Enr.		Other	
Variant	Position	Allele	Imbalanced Protein/Assay	Allele	Gene(s)
rs1494813	chr1:45957290	C	(L) MXI1	T	(I) CCDC163P
rs9793263	chr1:46722389	G	(L) MAFK	A	(A) C1orf190
rs17125090	chr1:63988904	A	(L) MAX,MXI1	G	(I) ITGB3BP
rs11161503	chr1:85462582	C	(L) EZH2	G	(I) MCOLN3
rs11161505	chr1:85462665	T	(L) EZH2	G	(I) MCOLN3
rs12127787	chr1:89458761	T	(L) CREB1*,USF1,USF2,YY1*	C	(I) GBP3
rs10858091	chr1:109935578	C	(L) HEY1,SIN3AK20,TCF12	T	(I) MYBPHL,SYPL2
rs2140924	chr1:109935775	A	(L) CREB1	C	(I) MYBPHL,SYPL2
rs573491	chr1:110026891	T	(L) TAF1	G	(I) SYPL2
rs2781553	chr1:110026989	G	(L) YY1	T	(I) SYPL2
rs12091564	chr1:145395604	C	(L) CREB1,MYBL2,NR2F2,TBP	T	(I) NOTCH2NL
rs2201601	chr1:213031448	G	(L) YY1	C	(I) NCRNA00292
rs11120067	chr1:213094557	A	(L) CTCF	G	(I) NCRNA00292
rs61844237	chr1:245133662	G	(L) DNASE	C	(I) EFCAB2
rs2291426	chr1:245134114	A	(L) CREB1	C	(I) EFCAB2
rs4020081	chr1:245209045	C	(L) CREB1,FOXA1,P300	T	(I) EFCAB2
rs4020082	chr1:245209134	A	(L) FOXA1	G	(I) EFCAB2
rs6759670	chr2:950291	C	(L) RAD21	A	(I) LOC339822
rs34122754	chr2:9884076	C	(L) CTCF	G	(I) GRHL1
rs4669449	chr2:9884205	G	(L) EZH2	A	(I) GRHL1
rs633808	chr2:20957592	G	(L) TCF12	T	(I) C2orf43
rs36101491	chr2:24387532	T	(L) CTCF,RAD21,ZBTB7A	C	(I) C2orf84
rs17046192	chr2:24461334	A	(L) FOXA1,ZBTB7A	G	(I) C2orf84
rs11676939	chr2:24479057	C	(L) CEBPB	T	(I) C2orf84
rs72803210	chr2:24615710	A	(L) MAFK	G	(I) C2orf84
rs77421503	chr2:24625676	C	(L) MAX	G	(I) C2orf84
rs10460551	chr2:24627074	C	(L) HDAC2	T	(I) C2orf84
rs2011616	chr2:27302561	G	(L) HDAC2	A	(A) KHK
rs162330	chr2:38319496	C	(L) CEBPB	A	(L) CYP1B1
rs1554612	chr2:48827497	C	(L) CTCF,FOXA2,RAD21	T	(A) STON1 (L) STON1
rs2070063	chr2:64862055	A	(L) MAFK	G	(A) SERTAD2
rs2241883	chr2:88424066	T	(L) POL2	C	(L) ARHGAP5,FABP1,TIGD2
rs2276561	chr2:113956371	G	(L) MAX	C	(I) LOC654433
rs2305133	chr2:113956821	C	(L) BHLHE40,CTCF,RAD21	G	(I) LOC654433
rs931472	chr2:113969948	C	(L) DNASE	T	(I) LOC654433
rs1049137	chr2:113975110	G	(L) POL2	A	(I) LOC654433

Tissues are designated by (L) liver[113,114], (I) islets[116], (A) adipose[115]. *indicates sequence reads are enriched for other allele rather than enriched allele (Enr. Allele)

Variant	Position	Enr. Allele	Imbalanced Protein/Assay	Other Allele	Gene(s)
rs2289897	chr2:113977454	A	(L) CEBPB	G	(I) LOC654433
rs4849176	chr2:113977936	C	(L) POL2	T	(I) LOC654433
rs4849178	chr2:113982608	A	(L) FOXA1	G	(I) LOC654433
rs2166421	chr2:113990242	C	(L) FOXA1,HNF4A	T	(I) LOC654433
rs7421852	chr2:113990261	A	(L) FOXA1	G	(I) LOC654433,PAX8
rs10206269	chr2:113990393	A	(L) HNF4A	C	(I) LOC654433,PAX8
rs4849181	chr2:113991970	G	(L) CTCF	A	(I) LOC654433,PAX8
rs2364723	chr2:178126546	C	(L) MBD4	G	(L) NFE2L2
rs938929	chr2:198780860	G	(L) RAD21	A	(I) PLCL1
rs12991600	chr2:202337236	G	(L) MAFK	A	(A) TRAK2
rs10171839	chr2:219051314	G	(L) MAX	A	(A) ARPC2
rs13062	chr2:219260651	A	(L) CTCF,MAX,ZEB1	C	(A) SLC11A1
rs13423632	chr2:232079116	C	(L) MAX	T	(I) HTR2B
rs16827879	chr2:232092301	T	(L) CTCF,RAD21	C	(I) HTR2B
rs77438791	chr2:239035642	G	(L) MBD4	A	(I) ESPNL
rs10206101	chr2:241835543	C	(L) FOSL2,HNF4G,JUND,ZEB1	A	(I) C2orf54
rs10933517	chr2:241836338	C	(L) DNASE,RAD21	T	(I) C2orf54
rs990284	chr3:104972	A	(L) FOXA1,FOXA2,P300	G	(L) CHL1
rs3774750	chr3:50208406	G	(L) MAX	C	(A) MST1R
rs2233474	chr3:50388607	C	(L) DNASE	A	(A) CYB561D2
rs2976742	chr3:125417021	C	(L) CTCF	T	(I) LOC100125556
rs12695470	chr3:125635718	A	(L) DNASE,POL2	C	(I) LOC100125556
rs9841194	chr3:125635739	T	(L) POL2	C	(I) LOC100125556
rs12497980	chr3:125636117	G	(L) MAX,MXI1	A	(I) LOC100125556
rs4377449	chr3:125642330	G	(L) USF1	A	(I) LOC100125556
rs9826071	chr3:125648165	T	(L) CTCF,RAD21	C	(I) LOC100125556
rs17523380	chr3:125802874	T	(L) RAD21	C	(I) LOC100125556
rs2939820	chr3:128127643	G	(L) CTCF	A	(L) HSS00171311
rs7661077	chr4:7219889	C	(L) CEBPB	T	(L) SORCS2
rs78578320	chr4:68566689	G	(L) DNASE	A	(I) GNRHR
rs28653581	chr4:68567025	G	(L) DNASE	T	(I) GNRHR
rs10030238	chr4:141808805	A	(L) HNF4G	G	(A) RNF150
rs2227426	chr4:155493171	G	(L) POL2	A	(L) C9orf66
rs6846466	chr4:166424428	T	(I) FOXA2	C	(I) MIR578
rs28641985	chr4:189376705	A	(L) CREB1,ELF1,FOXA1,FOXA2,HEY1,MAFK,POL2,SRF	G	(I) LOC401164
rs10941891	chr5:21391789	G	(L) CTCF,RAD21,SMC3	C	(I) GUSBP1
rs3195676	chr5:34008100	C	(L) BHLHE40,MAX,TAF1	T	(A) AMACR
rs13356762	chr5:56110992	G	(L) TAF1	T	(I) C5orf35
rs2548663	chr5:56172778	G	(L) CEBPB	A	(I) C5orf35
rs185220	chr5:56205357	G	(L) DNASE	A	(I) C5orf35

Tissues are designated by (L) liver[113,114], (I) islets[116], (A) adipose[115]. *indicates sequence reads are enriched for other allele rather than enriched allele (Enr. Allele)

Variant	Position	Enr. Allele	Imbalanced Protein/Assay	Other Allele	Gene(s)
rs252923	chr5:56205662	G	(L) YY1	T	(I) C5orf35
rs33321	chr5:56206073	G	(L) HEY1	T	(I) C5orf35
rs1643635	chr5:79928216	G	(L) MBD4	A	(I) DHFR
rs10168	chr5:79950403	C	(L) DNASE,MAX	T	(I) DHFR
rs386689494	chr5:79961675	C	(L) MAFK	T	(I) DHFR
rs1677645	chr5:79964419	C	(L) CEBPB	T	(I) DHFR
rs226198	chr5:81573992	C	(L) MAX,USF1	T	(I) ATP6AP1L
rs730870	chr5:125874993	G	(L) FOXA1	A	(L) HSS00124116
rs192231	chr5:140248539	C	(L) CTCF,RAD21	T	(I) VTRNA1-2
rs7739320	chr6:3054146	C	(L) DNASE	T	(I) LOC401233
rs12203636	chr6:3064249	A	(L) HEY1,MAX,MYBL2,POL2,ZEB1	G	(I) LOC401233
rs12196777	chr6:3064523	C	(L) DNASE	T	(I) LOC401233
rs12665605	chr6:3067003	G	(L) POL2	A	(I) LOC401233
rs12663589	chr6:3069057	C	(L) POL2	T	(I) LOC401233
rs9379687	chr6:24721787	C	(L) BHLHE40,CREB1,HEY1,MAX,MXI1	A	(A) FAM65B
rs198853	chr6:26104096	C	(L) HEY1,MYBL2,POL2,RXRA,SRF,TBP	T	(L) HFE
rs13201769	chr6:30756066	A	(L) CTCF,NR2F2,SP1	G	(A) IER3
rs3132555	chr6:31082910	G	(L) RAD21	C	(I) CDSN
rs1042149	chr6:31082960	G	(L) CTCF	A	(I) PSORS1C1
rs386579266	chr6:31089982	G	(L) CEBPB	A	(I) CDSN
rs9271092	chr6:32576296	A	(L) RAD21	G	(I) HLA-DQA1,HLA-DRB1
rs9271093	chr6:32576341	G	(L) CTCF	A	(I) HLA-DQB1,HLA-DRA, HLA-DRB1, HLA-DRB5
rs9271094	chr6:32576347	G	(L) CTCF	C	(I) HLA-DQA1,HLA-DRB1
rs9271096	chr6:32576426	A	(L) CTCF	G	(I) HLA-DQA1,HLA-DRB1
rs17843603	chr6:32620241	G	(L) BHLHE40	A	(I) HLA-DQA1,HLA-DQB1, HLA-DRB1
rs1063349	chr6:32627906	T	(L) HNF4A	C	(I) HLA-DQA1,HLA-DQB1, HLA-DRB1
rs386699568	chr6:32627923	A	(L) HNF4A	G	(I) HLA-DQA1,HLA-DQB1, HLA-DRB1
rs9274535	chr6:32634620	T	(L) RFX5	C	(I) HLA-DQB1,HLA-DRA, HLA-DRB1, HLA-DRB5
rs12192544	chr6:46620252	C	(L) EZH2	G	(L) SLC25A27
rs9493450	chr6:133135807	T	(L) HEY1,POL2	C	(A) SNORD100
rs11155000	chr6:139099401	C	(L) TCF7L2	T	(A) CCDC28A
rs539298	chr6:160770360	G	(L) BHLHE40,CEBPB,CREB1,DNASE,ELF1,HNF4A,HNF4G,JUND,MAX,MYBL2,NFIC, NR2F2,P300,POL2,RAD21	A	(L) SLC22A3
rs8200	chr7:75696606	G	(L) POL2	C	(L) AK022137
rs10953284	chr7:77169782	G	(L) CEBPB,DNASE,HNF4A,HNF4G	C	(L) PTPN12
rs776745	chr7:99291337	G	(L) FOXA1	T	(I) CYP3A5
rs6962760	chr7:141465363	T	(I) FOXA2	C	(I) FLJ40852
rs11783893	chr8:2100976	C	(L) CTCF	G	(I) MYOM2
rs2279128	chr8:71581559	T	(L) BHLHE40	G	(I) XKR9

Tissues are designated by (L) liver[113,114], (I) islets[116], (A) adipose[115]. *indicates sequence reads are enriched for other allele rather than enriched allele (Enr. Allele)

Variant	Position	Enr. Allele	Imbalanced Protein/Assay	Other Allele	Gene(s)
rs6985299	chr8:71613079	T	(L) MAX	C	(I) XKR9
rs11985375	chr8:71613472	G	(L) CREB1,MAX	A	(I) XKR9
rs10810299	chr9:14964272	A	(L) POL2	C	(I) LOC389705
rs1414395	chr10:13334136	T	(L) CEBPB	G	(A) PHYH
rs4747275	chr10:16552472	G	(L) MAFK	A	(I) C1QL3
rs17141322	chr10:17604700	A	(I) PDX1	C	(I) ST8SIA6
rs10510007	chr10:116636721	G	(L) CEBPB	A	(A) KIAA1600
rs10835531	chr11:1516110	A	(I) FOXA2	G	(I) LOC338651
rs538954	chr11:65756808	C	(L) ZBTB7A	T	(L) NM_032325,NM_174952
ds613128	chr11:68638058	G	(L) MAX	T	(I) MRPL21
rs514833	chr11:68657734	C	(L) CTCF,DNASE,RAD21,ZBTB33	T	(I) MRPL21
rs629426	chr11:68671104	G	(L) YY1	A	(I) MRPL21
rs9787897	chr11:74659302	T	(L) FOXA2	A	(I) XRRA1
rs567956	chr11:74659779	C	(L) POL2	T	(I) XRRA1
rs2165163	chr11:74660143	C	(L) DNASE,MAX	G	(I) XRRA1
rs4944968	chr11:74724276	C	(L) DNASE	G	(I) XRRA1
rs3781884	chr11:113217364	G	(L) CTCF	A	(I) TTC12
rs2534721	chr12:6580144	C	(L) DNASE	A	(I) TAPBPL
rs933462	chr12:9103665	G	(L) HDAC2	T	(A) KLRG1
rs2239182	chr12:48255411	C	(L) FOXA1	T	(L) CASKIN2
rs4509811	chr12:58335142	A	(L) CEBPD	G	(I) XRCC6BP1
rs1010461	chr14:21153788	C	(L) POL2	A	(A) RNASE4
rs1769591	chr14:34378886	G	(L) FOXA1	A	(L) EGLN3
rs11624787	chr14:53288450	C	(L) CEBPB,FOXA1,FOXA2,HNF4A,P300	G	(L) STYX
rs2034652	chr15:40802768	A	(L) CEBPB	G	(I) C15orf57,MRPL42P5
rs8036737	chr15:40874256	G	(L) MAFK	A	(I) C15orf57,MRPL42P5
rs7175950	chr15:78236353	A	(L) POL2	G	(I) LOC645752
rs9925556	chr16:2880105	T	(L) DNASE,FOXA1	C	(L) NM_145252
rs11075256	chr16:15187912	G	(L) HEY1,MAX,MXI1,POL2,SP2,TAF1	C	(L) RRN3
rs62032983	chr16:23653343	T	(L) CREB1,ELF1	C	(I) DCTN5
rs2303222	chr16:31085470	C	(L) POL2	T	(L) VKORC1
rs1981760	chr16:50723074	C	(L) EZH2	T	(L) CARD15
rs35223604	chr16:70052345	C	(L) CTCF	G	(I) EXOSC6
rs4985407	chr16:70285901	G	(L) BHLHE40,CHD2,CREB1,CTCF,DNASE, ELF1,GABP,MAX,POL2,SIN3AK20,YY1	A	(I) EXOSC6
rs876985	chr16:87882124	C	(L) P300	T	(A) SLC7A5
rs5689147	chr17:41438468	T	(I) POL2	C	(I) NBR2
rs12938996	chr17:41438674	A	(I) POL2	G	(I) NBR2
rs12948653	chr17:46259254	A	(L) CTCF	C	(A) HOXB5
rs16949649	chr17:49230308	T	(L) CEBPB	C	(A) NME1 (L) NME1
rs8076760	chr17:61920497	T	(L) CHD2,GABP,MAX,TCF12	C	(I) FTSJ3

Tissues are designated by (L) liver[113,114], (I) islets[116], (A) adipose[115]. *indicates sequence reads are enriched for other allele rather than enriched allele (Enr. Allele)

Variant	Position	Enr. Allele	Imbalanced Protein/Assay	Other Allele	Gene(s)
rs2598414	chr17:74067099	C	(L) BHLHE40	T	(A) SRP68
rs2376585	chr17:76417883	T	(L) CEBPA,CTCF,DNASE,FOXA1,FOXA2, MAX,NR2F2,ZBTB33,ZEB1	C	(I) DNAH17
rs8101689	chr19:30185697	A	(L) HDAC2	G	(I) C19orf12
rs1343703	chr19:49955155	C	(L) ARID3A (I) FOXA2*	G	(L) NOP17
rs34541537	chr19:58912737	C	(L) CREB1	T	(I) ZNF584
rs35138622	chr19:58912788	C	(L) CREB1	A	(I) ZNF584
rs35117909	chr19:58912906	G	(L) CREB1	A	(I) ZNF584
rs11878203	chr19:58920050	G	(L) DNASE	A	(I) ZNF584
rs11697620	chr20:3150165	C	(L) CTCF,MAX	T	(A) ProSAPiP1
rs6111720	chr20:17868623	T	(L) CEBPD	C	(L) ADRBK1,AP4M1,ATP2B1, CAPN10, CORO1B,GIPC1,LZTR1,PIAS3, SLC39A13,TAF6,WHSC2,WIZ
rs386814838	chr20:48092076	G	(L) CEBPB	A	(L) AK055386,Contig29707,KCNB1
rs2832245	chr21:30554634	C	(I) FOXA2	T	(I) RWDD2B
rs6000200	chr22:36612258	C	(I) PDX1	G	(A) APOL2
rs4828057	chrX:100006043	A	(I) MAFB	C	(A) SYTL4

Tissues are designated by (L) liver[113,114], (I) islets[116], (A) adipose[115]. *indicates sequence reads are enriched for other allele rather than enriched allele (Enr. Allele)

APPENDIX 3: Allelic imbalance at cardiometabolic trait and disease GWAS loci

Locus	Variant	Position	Enr. Allele	Imbalanced Proteins	Other Allele	Trait	Reported Variant	LD $(r^2)^a$	Coupling	Risk. Allele
TMEM57, LDLRAP1	rs9438904	chr1:25756860	C	(L) YY1	T	Total Cholesterol	rs12027135	0.98	T,A	A
						LDL cholesterol	rs12027135	0.98	T,A	A
LEPR	rs12021623	chr1:66153586	C	(L) HNF4G	A	C-reactive protein	rs4420065	0.90	C,T	C
DDAH1	rs2268667	chr1:85793746	G	(L) FOXA1	A	Serum dimethylarginine levels	rs1884139	0.90	G,G	T
	rs2177461	chr1:85861976	G	(L) ZBTB33	C	Serum dimethylarginine levels	rs1884139	0.97	G,G	T
PHGDH	rs839605	chr1:120217524	C	(L) CTCF	A	Blood metabolite levels	rs1163251	0.98	C,C	T
	rs639761	chr1:120217558	G	(L) CEBPA,CEBPB,CREB1,CTCF,RAD21	A	Blood metabolite levels	rs1163251	0.97	G,C	T
	rs640195	chr1:120217650	T	(L) ARID3A,CEBPB,CREB1,CTCF,RAD21	A	Blood metabolite levels	rs1163251	0.98	T,C	T
	rs483180	chr1:120267505	G	(L) HDAC2	C	Metabolite levels	rs478093	0.99	C,G	G
						Metabolic traits	rs477992	0.99	C,G	A
F5, SELP	rs2236869	chr1:169535196	G	(L) FOXA1	T	Activated partial thromboplastin time	rs6028	0.90	T,T	C
FMO3	rs2281007	chr1:171111351	G	(L) CEBPA	A	Blood metabolite levels	rs7061710	0.79	G,G	C
ANGPTL1	rs17361251	chr1:178520577	A	(L) CEBPA,CEBPB,CEBPD,FOXA1,FOXA2,HNF4A, HNF4G,MAX,NFIC,NR2F2,P300,SP1,ZBTB7A	C	HDL cholesterol	rs4650994	1.00	A,A	G
	rs17276513	chr1:178520604	A	(L) CEBPA,CEBPB,DNASE,FOXA1,FOXA2,HDAC2, HNF4A,HNF4G,MAX,NR2F2,P300,SP1,ZBTB7A	T	HDL cholesterol	rs4650994	0.99	A,A	G
	rs17276527	chr1:178520680	A	(L) CREB1,DNASE,FOXA1,FOXA2,HDAC2,HNF4A, HNF4G,MAX,NFIC,P300,SP1,ZEB1	G	HDL cholesterol	rs4650994	1.00	A,A	G
PROX1	rs9970073	chr1:214156165	A	(L) MAX	G	Fasting glucose-related traits (interaction with BMI)	rs340874	0.73	G,C	-
						Fasting glucose-related traits	rs340874	0.73	G,C	C
	rs340879	chr1:214156514	T	(L) RAD21	C	Fasting glucose-related traits (interaction with BMI)	rs340874	0.77	C,C	-
						Fasting glucose-related traits	rs340874	0.77	C,C	C
MIA3	rs4846770	chr1:222795569	G	(L) CEBPD	C	Myocardial infarction (early onset)	rs17465637	0.95	G,C	C
						Coronary heart disease	rs17465637	0.95	G,C	C
GALNT2	rs4846913	chr1:230294715	A	(L) CEBPA,CEBPB,CEBPD,NR2F2 (I) MAFB	C	HDL cholesterol	rs2144300	1.00	A,T	T
						Triglycerides	rs4846914	1.00	A,A	G
						Metabolite levels	rs10127775	1.00	A,T	-
IRF2BP2, TOMM20	rs526936	chr1:234852204	A	(L) POL2 (I) FOXA2	G	Total Cholesterol	rs514230	0.92	A,T	A
						LDL cholesterol	rs514230	0.92	A,T	A
	rs556107	chr1:234853059	C	(L) HEY1	T	Total Cholesterol	rs514230	0.93	T,T	A
						LDL cholesterol	rs514230	0.93	T,T	A
TRIB2	rs4669888	chr2:12980514	G	(L) CEBPA,CEBPB	A	Pericardial fat	rs10198628	0.98	A,A	-

* indicates other allele is enriched rather than enriched allele(Enr. Allele), ^{a}R2 and coupling (imbalanced allele, reported allele) calculated in 1000 Genome EUR. Risk allele increases risk or trait measurement based on reported effect allele in NHGRI GWAS catalog. - indicates no effect allele reported in catalog. Only reported variant in highest LD shown

Locus	Variant	Position	Enr. Allele	Imbalanced Proteins	Other Allele	Trait	Reported Variant	LD (r^2)^a	Coupling	Risk. Allele
APOB	rs1367117	chr2:21263900	G	(L) HEY1,MAX	A	Total Cholesterol	rs1367117	1.00	-	A
						LDL cholesterol	rs1367117	1.00	-	A
						Lipid metabolism phenotypes	rs1367117	1.00	-	-
	rs312983	chr2:21378580	A	(L) FOXA1	C	LDL cholesterol	rs562338	0.72	C,G	G
						Lipid metabolism phenotypes	rs312985	0.99	C,G	-
	rs312984	chr2:21378778	C	(L) ARID3A,FOXA1,FOXA2,HNF4A,MAX,NFIC, RAD2,ZEB1	T	LDL cholesterol	rs562338	0.73	T,G	G
						Lipid metabolism phenotypes	rs312985	1.00	T,G	-
	rs312985	chr2:21378805	A	(L) CREB1,FOXA1,FOXA2,HDAC2,HNF4A,HNF4G, MYBL2,NRSF,P300,SP1,ZEB1	G	LDL cholesterol	rs562338	0.73	G,G	G
						Lipid metabolism phenotypes	rs312985	1.00	-	-
	rs1652418	chr2:21388456	T	(L) MAZ,RAD21,SMC3	C	LDL cholesterol	rs515135	0.72	C,C	T
						Lipid metabolism phenotypes	rs312985	0.99	C,G	-
	rs544039	chr2:21398985	C	(L) CTCF,RAD21	A	LDL cholesterol	rs515135	0.71	A,C	T
						Lipid metabolism phenotypes	rs506585	1.00	A,A	-
ADCY3, POMC	rs7580081	chr2:25097072	C	(L) CEBPB	G	Body mass index	rs6545814	0.94	G,A	G
						Obesity	rs10182181	0.73	G,A	G
SLC5A6	rs2580759	chr2:27432500	G	(L) CTCF	T	Blood metabolite levels	rs1395	0.71	T,A	A
	rs11608	chr2:27435374	G	(L) YY1	A	Blood metabolite levels	rs1395	0.88	A,A	A
	rs1141313	chr2:27460968	G	(L) POL2	A	Blood metabolite levels	rs1395	0.91	A,A	A
GCKR	rs1260326	chr2:27730940	C	(L) CTCF	T	Triglycerides-Blood Pressure (TG-BP)	rs780093	0.91	T,T	A
						Metabolic syndrome	rs780094	0.91	T,T	A
						Waist Circumference - Triglycerides (WC-TG)	rs780093	0.91	T,T	A
						Palmitoleic acid (16:1n-7) plasma levels	rs780093	0.91	T,T	-
						Fasting glucose-related traits (interaction with BMI)	rs780094	0.91	T,T	-
						Serum albumin level	rs1260326	1.00	-	T
						Two-hour glucose challenge	rs1260326	1.00	-	T
						Hypertriglyceridemia	rs1260326	1.00	-	T
						Lipid metabolism phenotypes	rs1260326	1.00	-	-
						Fasting insulin-related traits	rs780094	0.91	T,T	C
						Blood metabolite levels	rs1260326	1.00	-	T
						Metabolite levels	rs1260326	1.00	-	-
						Non-albumin protein levels	rs1260326	1.00	-	C
						Metabolic traits	rs1260326	1.00	-	A
						Fasting insulin-related traits (interaction with BMI)	rs780094	0.91	T,T	-
						C-reactive protein	rs1260326	1.00	-	T
						Triglycerides	rs1260326	1.00	-	T

* indicates other allele is enriched rather than enriched allele (Enr. Allele), ^{a}R2 and coupling (imbalanced allele, reported allele) calculated in 1000 Genome EUR. Risk allele increases risk or trait measurement based on reported effect allele in NHGRI GWAS catalog. - indicates no effect allele reported in catalog. Only reported variant in highest LD shown

Locus	Variant	Position	Enr. Allele	Imbalanced Proteins	Other Allele	Trait	Reported Variant	LD (r^2)[a]	Coupling	Risk. Allele
GCKR (cont'd)	rs1260326 (cont'd)	chr2:27730940	C	(L) CTCF	T	Blood metabolite ratios	rs1260326	1.00	-	T
						Total Cholesterol	rs1260326	1.00	-	T
						Fasting glucose-related traits	rs780094	0.91	T,T	C
	rs780095	chr2:27741105	G	(L) FOXA1	A	Triglycerides-Blood Pressure (TG-BP)	rs780093	0.88	A,T	A
						Metabolic syndrome	rs780094	0.86	A,T	A
						Waist Circumference - Triglycerides (WC-TG)	rs780093	0.88	A,T	A
						Palmitoleic acid (16:1n-7) plasma levels	rs780093	0.88	A,T	-
						Fasting glucose-related traits (interaction with BMI)	rs780094	0.86	A,T	-
						Serum albumin level	rs1260326	0.81	A,T	T
						Two-hour glucose challenge	rs1260326	0.81	A,T	T
						Hypertriglyceridemia	rs1260326	0.81	A,T	T
						Lipid metabolism phenotypes	rs1260326	0.81	A,T	-
						Fasting insulin-related traits	rs780094	0.86	A,T	C
						Non-albumin protein levels	rs1260326	0.81	A,T	C
						Metabolite levels	rs1260326	0.81	A,T	-
						Blood metabolite levels	rs1260326	0.81	A,T	T
						Fasting insulin-related traits (interaction with BMI)	rs780094	0.86	A,T	?
						Metabolic traits	rs780094	0.86	A,T	T
						C-reactive protein	rs780094	0.86	A,T	A
						Triglycerides	rs1260333	0.98	A,A	C
						Blood metabolite ratios	rs1260326	0.81	A,T	T
						Total Cholesterol	rs1260326	0.81	A,T	T
						Fasting glucose-related traits	rs780094	0.86	A,T	C
	rs780094	chr2:27741237	C	(L) CEBPB,FOXA1,FOXA2,MAFK,MAX,NR2F2, NRSF,ZEB1	T	C-reactive protein	rs780094	1.00	-	A
						Metabolic syndrome	rs780094	1.00	-	A
						Metabolic traits	rs780094	1.00	-	T
						Fasting insulin-related traits (interaction with BMI)	rs780094	1.00	-	-
						Fasting glucose-related traits (interaction with BMI)	rs780094	1.00	-	-
						Fasting insulin-related traits	rs780094	1.00	-	C
						Fasting glucose-related traits	rs780094	1.00	-	C
						Triglycerides	rs780094	1.00	-	T
						Waist Circumference - Triglycerides (WC-TG)	rs780093	0.98	T,T	A
						Triglycerides-Blood Pressure (TG-BP)	rs780093	0.98	T,T	A

* indicates other allele is enriched rather than enriched allele(Enr. Allele), [a]R2 and coupling (imbalanced allele, reported allele) calculated in 1000 Genome EUR. Risk allele increases risk or trait measurement based on reported effect allele in NHGRI GWAS catalog. - indicates no effect allele reported in catalog. Only reported variant in highest LD shown

			Enr.		Other		Reported	LD		Risk.
Locus	Variant	Position	Allele	Imbalanced Proteins	Allele	Trait	Variant	(r^2)^a	Coupling	Allele
GCKR (cont'd)	rs780094 (cont'd)	chr2:27741237	C	(L) CEBPB,FOXA1,FOXA2,MAFK,MAX,NR2F2, NRSF,ZEB1	T	Palmitoleic acid (16:1n-7) plasma levels	rs780093	0.98	T,T	-
						Two-hour glucose challenge	rs1260326	0.91	T,T	T
						Hypertriglyceridemia	rs1260326	0.91	T,T	T
						Lipid metabolism phenotypes	rs1260326	0.91	T,T	?
						Non-albumin protein levels	rs1260326	0.91	T,T	C
						Metabolite levels	rs1260326	0.91	T,T	-
						Blood metabolite levels	rs1260326	0.91	T,T	T
						Blood metabolite ratios	rs1260326	0.91	T,T	T
						Total Cholesterol	rs1260326	0.91	T,T	T
						Serum albumin level	rs1260326	0.91	T,T	T
C2orf16	rs1919128	chr2:27801759	A	(L) BHLHE40	G	Waist Circumference - Triglycerides (WC-TG)	rs1919128	1.00	-	A
CCDC121	rs3749147	chr2:27851918	G	(L) CEBPD,DNASE,ELF1,GABP,MAX,MXI1,NR2F2, POL2,TAF1,YY1,ZBTB7A	A	Waist Circumference - Triglycerides (WC-TG)	rs3749147	1.00	-	C
EHBP1	rs2136737	chr2:62969310	C	(L) FOXA2,HNF4A	G	LDL cholesterol	rs2710642	0.77	G,A	G
	rs1553832	chr2:63013515	G	(L) BHLHE40	C	LDL cholesterol	rs2710642	0.80	C,A	G
	rs56373728	chr2:63095792	G	(L) POL2	A	LDL cholesterol	rs2710642	0.93	A,A	G
	rs2710642	chr2:63149557	G	(L) POL2	A	LDL cholesterol	rs2710642	1.00	-	G
ALMS1, NAT8, TPRKB, DUSP11	rs10469966	chr2:73752368	G	(L) CEBPB	A	Metabolite levels	rs9309473	0.89	G,A	G
						Blood metabolite levels	rs10469966	1.00	-	A
						Metabolic traits	rs13391552	0.80	G,G	A
Intergenic	rs62150376	chr2:83295262	T	(L) FOXA1,HDAC2,HNF4G	C	Hypertension	rs10496289	0.93	C,C	-
ACMSD	rs62168897	chr2:135717997	C	(L) POL2	T	Blood metabolite levels	rs6430553	0.74	C,C	T
FIGN	rs73029563	chr2:165008166	C	(L) CEBPB	G	Blood pressure	rs1446468	0.90	C,T	T
GRB14, COBLL1	rs6713419	chr2:165508300	T	(L) MAFK	C	type 2 diabetes	rs3923113	0.97	C,C	A
						Triglycerides	rs10195252	0.87	C,C	C
	rs10184004	chr2:165508389	C	(L) MAFK	T	type 2 diabetes	rs3923113	0.85	T,C	A
						Triglycerides	rs10195252	0.99	T,C	C
CTLA4	rs6745050	chr2:204691538	T	(L) MAFK	C	Type 1 diabetes	rs3087243	0.88	C,G	A
						Type 1 diabetes autoantibodies	rs3087243	0.88	C,G	A
UGT1A	rs4477910	chr2:234643737	T	(L) FOXA1,FOXA2,HNF4A,NR2F2	A	Blood metabolite levels	rs887829	0.74	A,C	T
						Metabolite levels	rs887829	0.74	A,C	T
						Metabolic traits	rs887829	0.74	A,C	T
AGXT	rs4401206	chr2:241796905	G	(L) CEBPD,CREB1,DNASE,HNF4A,HNF4G,MAX, MXI1,P300,RAD21,USF1,YY1	A	Blood metabolite levels	rs4675874	0.83	G,A	A
	rs10933641	chr2:241813788	C	(L) DNASE,MAX,POL2	T	Blood metabolite levels	rs4675874	0.86	C,A	A
	rs4675858	chr2:241840558	A	(L) CEBPB	G	Blood metabolite levels	rs4675874	0.98	A,A	A
	rs4417704	chr2:241846573	G	(L) CTCF,MAX,RAD21	A	Blood metabolite levels	rs4675874	0.75	G,A	A
	rs62186584	chr2:241853621	C	(L) CREB1	T	Blood metabolite levels	rs4675874	0.72	C,A	A

* indicates other allele is enriched rather than enriched allele(Enr. Allele), ^{a}R2 and coupling (imbalanced allele, reported allele) calculated in 1000 Genome EUR. Risk allele increases risk or trait measurement based on reported effect allele in NHGRI GWAS catalog. - indicates no effect allele reported in catalog. Only reported variant in highest LD shown

Locus	Variant	Position	Enr. Allele	Imbalanced Proteins	Other Allele	Trait	Reported Variant	LD (r^2)[a]	Coupling	Risk. Allele
SLC4A7	rs7640929	chr3:27513944	A	(L) CEBPA	C	Diastolic blood pressure	rs13082711	0.89	C,T	T
						Blood pressure	rs13082711	0.89	C,T	T
GNL3	rs7639267	chr3:52568805	T	(L) HEY1	G	Adiponectin levels	rs2590838	0.70	G,G	G
	rs1108842	chr3:52720080	C	(L) CREB1,SIN3AK20,TAF1	A	Adiponectin levels	rs1108842	1.00	-	C
	rs2710323	chr3:52815905	C	(L) BHLHE40,FOXA2	T	Adiponectin levels	rs1108842	0.80	T,A	C
GSK3B	rs6800622	chr3:119580678	C	(L) MAFK	A	HDL cholesterol	rs6805251	0.98	C,C	T
ADIPOQ	rs58575091	chr3:186545319	T	(L) CHD2	C	Adiponectin levels	rs266717	0.80	T,C	C
STX18, MSX1	rs4689909	chr4:4643276	G	(L) HEY1	A	Congenital heart disease	rs870142	0.98	G,C	A
ANXA3	rs4075927	chr4:79575058	G	(L) BHLHE40,USF1	A	Non-albumin protein levels	rs10007186	0.95	G,T	
ABCG2	rs45499402	chr4:89043634	G	(L) FOXA1	C	Serum uric acid levels	rs2231142	1.00	G,G	T
PPM1K	rs6841731	chr4:89228928	A	(L) MAFK	G	Blood metabolite levels	rs1440581	0.89	G,T	T
						Metabolite levels	rs1440581	0.89	G,T	?
	rs2869930	chr4:89242372	G	(L) FOXA1,FOXA2	C	Blood metabolite levels	rs10022462	1.00	G,C	T
TMEM154	rs6813195	chr4:153520475	C	(I) FOXA2	T	type 2 diabetes	rs6813195	1.00	-	C
TERT	rs31490	chr5:1344458	A	(L) CEBPB	G	Serum prostate-specific antigen levels	rs401681	0.89	A,T	C
ANKH	rs835158	chr5:14873254	G	(L) CEBPB	C	Blood metabolite levels	rs835154	0.76	C,G	A
BHMT, BHMT2	rs16876512	chr5:78407261	T	(L) TCF12	C	Blood metabolite levels	rs16876394	0.86	C,T	T
						Metabolite levels	rs17823642	0.83	C,C	C
ELL2	rs10070197	chr5:95241868	C	(L) CEBPB	T	Serum total protein level	rs3777200	0.99	T,C	T
	rs1458017	chr5:95251346	A	(L) HEY1	G	Serum total protein level	rs3777200	0.81	G,C	T
ELOVL2	rs2295602	chr6:11005842	T	(L) BHLHE40	C	Blood metabolite levels	rs4713169	0.87	T,G	C
	rs3798713	chr6:11008622	G	(L) FOXA1,HNF4A	C	Blood metabolite levels	rs4713169	0.83	G,G	C
	rs953413	chr6:11012859	G	(L) FOXA1,FOXA2,NR2F2,P300	A	Blood metabolite levels	rs4713169	0.88	G,G	C
	rs56190003	chr6:11088533	T	(L) HEY1	C	Blood metabolite levels	rs4713169	0.94	T,G	C
	rs13362715	chr6:11088630	C	(L) HEY1,POL2	T	Metabolite levels	rs3798722	0.75	C,A	-
						Metabolic traits	rs9393903	0.79	C,G	A
						Blood metabolite ratios	rs9393915	0.85	C,C	T
SLC17A3	rs1165176	chr6:25830298	A	(L) FOXA1	G	Cardiovascular disease risk factors	rs11754288	0.78	G,G	A
						Blood metabolite levels	rs2762353	0.86	G,G	A
						Blood metabolite ratios	rs1185567	0.88	G,G	A
	rs1165183	chr6:25836380	G	(L) POL2	A	Cardiovascular disease risk factors	rs11754288	0.78	G,G	A
						Blood metabolite levels	rs2762353	0.86	G,G	A
						Blood metabolite ratios	rs1185567	0.87	G,G	A
HCG27, HLA-C	rs6921948	chr6:31171257	A	(I) FOXA2	C	Coronary heart disease	rs3869109	0.73	A,A	G
BAT2, BAT5	rs813115	chr6:31620020	G	(L) DNASE,GABP,NRSF	A	Diastolic blood pressure	rs805303	1.00	A,G	G
						Systolic blood pressure	rs805303	1.00	A,G	G
						Hypertension	rs805303	1.00	A,G	G

* indicates other allele is enriched rather than enriched allele(Enr. Allele), [a]R2 and coupling (imbalanced allele, reported allele) calculated in 1000 Genome EUR. Risk allele increases risk or trait measurement based on reported effect allele in NHGRI GWAS catalog. - indicates no effect allele reported in catalog. Only reported variant in highest LD shown

			Enr.		Other		Reported	LD		Risk.
Locus	Variant	Position	Allele	Imbalanced Proteins	Allele	Trait	Variant	(r^2)^a	Coupling	Allele
MHC	rs4348358	chr6:32399092	G	(L) MAFK	A	Type 1 diabetes	rs9268645	1.00	G,C	-
	rs9268606	chr6:32400070	G	(L) CTCF	A	Type 1 diabetes	rs9268645	1.00	G,C	-
SLC35F1, C6orf204, PLN	rs283080	chr6:118606000	A	(L) RAD21	C	Cardiac structure and function	rs89107	0.79	C,G	G
Intergenic	rs73069540	chr7:26904770	T	(L) HDAC2	C	Type 1 diabetes	rs7804356	0.95	C,T	-
PON-1	rs705379	chr7:94953895	G	(L) DNASE	A	Paraoxonase activity	rs854572	0.84	G,C	
PPP1R3B	rs6984305	chr8:9178268	T	(L) CTCF	A	Fasting glucose-related traits (interaction with BMI)	rs4841132	0.80	T,G	-
						HDL cholesterol	rs9987289	0.80	T,G	A
						Metabolite levels	rs4841132	0.80	T,G	-
						HDL Cholesterol - Triglycerides (HDLC-TG)	rs9987289	0.80	T,G	A
						Fasting insulin-related traits (interaction with BMI)	rs4841132	0.80	T,G	-
						C-reactive protein	rs9987289	0.80	T,G	A
						Total Cholesterol	rs9987289	0.80	T,G	-
						LDL cholesterol	rs9987289	0.80	T,G	T
IDO1	rs11777082	chr8:39797703	A	(L) BHLHE40	G	Blood metabolite levels	rs2160860	0.88	G,A	A
CDC123, CAMK1D	rs34428576	chr10:12281111	G	(L) CEBPA,CEBPB,CEBPD,CREB1,DNASE, FOXA1,HDAC2,HNF4A,JUND,MAX,NFIC,NR2F2,P300, RAD21,ZBTB7A	A	type 2 diabetes	rs12779790	0.72	G,A	G
c10orf107, TMEM26, RTKN2, RHOBTB1, ARID5B	rs16916563	chr10:63507642	G	(L) FOXA2	A	Diastolic blood pressure	rs1530440	0.92	G,C	T
HHEX	rs4933736	chr10:94471595	T	(L) FOXA2	C	type 2 diabetes	rs1111875	0.74	C,C	G
HIF1AN, SEC31B, NDUFB8, WNT8B, SCD	rs2295776	chr10:102295629	G	(L) BHLHE40	T	Palmitoleic acid (16:1n-7) plasma levels	rs11190604	0.96	G,A	-
	rs2495758	chr10:102321900	C	(L) CHD2,CREB1,CTCF,DNASE,ELF1,MAX,RAD21,YY1	G	Palmitoleic acid (16:1n-7) plasma levels	rs11190604	0.98	C,A	-
CYP17A1, AS3MT, CNNM2, NT5C2	-	chr10:104692633	A	(L) CEBPA	C	Blood pressure	rs11191548	0.83	C,T	T
						Coronary heart disease	rs12413409	0.94	C,G	-
						Systolic blood pressure	rs11191548	0.83	C,T	T
	rs1926032	chr10:104829469	C	(L) CTCF	T	Blood pressure	rs11191548	0.76	C,T	T
						Systolic blood pressure	rs11191548	0.76	C,T	T
ACY3	rs12288023	chr11:67421341	T	(L) CEBPB	C	Serum metabolite levels	rs12288023	1.00	-	C

* indicates other allele is enriched rather than enriched allele(Enr. Allele), ^{a}R2 and coupling (imbalanced allele, reported allele) calculated in 1000 Genome EUR. Risk allele increases risk or trait measurement based on reported effect allele in NHGRI GWAS catalog. - indicates no effect allele reported in catalog. Only reported variant in highest LD shown

Locus	Variant	Position	Enr. Allele	Imbalanced Proteins	Other Allele	Trait	Reported Variant	LD (r^2)^a	Coupling	Risk. Allele
ZNF259, APOA1, APOC3, APOA4, APOA5, BUD13	rs180351	chr11:116607641	T	(L) CTCF	C	Triglycerides	rs603446	0.91	T,T	T
UBASH3B	rs11218752	chr11:122552600	C	(L) CTCF	T	HDL cholesterol	rs7941030	0.81	C,T	G
						Total Cholesterol	rs7941030	0.81	C,T	G
PAH	rs869916	chr12:103244013	T	(L) HNF4A	G	Blood metabolite levels	rs1498694	1.00	G,A	A
ASCL1	rs4764939	chr12:103522952	C	(L) HNF4A	T	C-reactive protein	rs10745954	0.87	C,A	A
ACADS	rs12828810	chr12:121152017	G	(L) CTCF	T	Metabolite levels	rs2014355	0.91	T,T	T
						Blood metabolite levels	rs2066938	1.00	T,A	A
						Metabolic traits	rs2066938	1.00	T,A	G
HNF1A	rs7139079	chr12:121415293	A	(L) POL2	G	C-reactive protein	rs7310409	0.82	A,G	A
OASL	rs2258287	chr12:121454313	A	(L) POL2,USF1	C	Cardiovascular disease risk factors	rs3213545	0.72	C,G	A
WDR66	rs1154513	chr12:122391963	A	(I) FOXA2	G	Blood metabolite levels	rs493519	0.94	G,C	T
Intergenic	rs12864047	chr13:74796108	C	(L) HNF4G	T	Sudden cardiac arrest	rs12429889	0.78	T,T	-
ABCC4	rs9302064	chr13:95966851	A	(L) FOXA1,JUND	C	Blood metabolite levels	rs9302065	0.99	C,G	A
SERPINA1	rs17090719	chr14:94846661	T	(L) HEY1,MBD4,POL2	C	Metabolite levels	rs1303	0.75	T,T	-
CSK, ULK3	rs11857695	chr15:75165751	T	(L) CREB1,JUND	G	Diastolic blood pressure	rs6495122	0.70	T,C	A
ADAMTS7	rs11856536	chr15:79094325	A	(L) HDAC2	G	Coronary heart disease	rs3825807	0.98	A,A	A
FURIN, FES	rs4932370	chr15:91404705	G	(I) NKX6	A	Diastolic blood pressure	rs2521501	0.71	A,T	T
						Systolic blood pressure	rs2521501	0.71	A,T	T
ACSM5, ACSM2A	rs7194098	chr16:20464350	C	(L) POL2,SRF	G	Blood metabolite levels	rs11647589	0.89	C,A	A
	rs9937581	chr16:20473903	G	(L) MAFF,MAFK	A	Blood metabolite levels	rs1394678	0.80	G,C	T
	rs1394678	chr16:20491058	C	(L) CTCF	T	Blood metabolite levels	rs1394678	1.00	-	T
IL27	rs181203	chr16:28512371	A	(L) EZH2	C	Type 1 diabetes	rs4788084	0.73	A,C	G
	rs62034319	chr16:28532188	T	(L) HDAC2	G	Type 1 diabetes	rs4788084	0.99	T,C	G
	rs2106480	chr16:28537971	T	(L) CEBPB,DNASE,FOXA1,HNF4A,P300	C	Type 1 diabetes	rs4788084	1.00	T,C	G
	rs62034351	chr16:28565489	G	(L) DNASE,HEY1,POL2	A	Type 1 diabetes	rs4788084	0.79	G,C	G
	rs7191618	chr16:28565667	C	(L) DNASE,MAX,RAD21	G	Type 1 diabetes	rs4788084	0.89	C,C	G
	rs743590	chr16:28608230	G	(L) MAX,POL2,YY1	A	Type 1 diabetes	rs4788084	0.74	G,C	G
	rs62031562	chr16:28609329	A	(L) POL2	T	Type 1 diabetes	rs4788084	0.71	A,C	G
SH2B1	rs7187776	chr16:28857645	A	(L) CREB1,HEY1,HNF4G	G	Body mass index	rs7498665	0.71	A,A	G
						Obesity	rs7498665	0.71	A,A	G
	rs62037367	chr16:28874547	C	(L) CREB1	G	Body mass index	rs7498665	1.00	C,A	G
						Obesity	rs7498665	1.00	C,A	G
	rs7198606	chr16:28875122	T	(L) CREB1,SP2,SRF	G	Body mass index	rs7498665	1.00	T,A	G
						Obesity	rs7498665	1.00	T,A	G
	rs11864750	chr16:28875204	A	(L) NR2F2	T	Body mass index	rs7498665	0.99	A,A	G
						Obesity	rs7498665	0.99	A,A	G

* indicates other allele is enriched rather than enriched allele(Enr. Allele), ^{a}R2 and coupling (imbalanced allele, reported allele) calculated in 1000 Genome EUR. Risk allele increases risk or trait measurement based on reported effect allele in NHGRI GWAS catalog. - indicates no effect allele reported in catalog. Only reported variant in highest LD shown

Locus	Variant	Position	Enr. Allele	Imbalanced Proteins	Other Allele	Trait	Reported Variant	LD $(r^2)^a$	Coupling	Risk. Allele
RABEP2, SH2B1	rs4072402	chr16:28937259	C	(L) CTCF	T	Obesity	rs7184597	0.85	C,C	T
CETP	rs12720926	chr16:56998918	A	(L) DNASE	G	Lipid metabolism phenotypes	rs1532624	0.94	A,C	-
						HDL cholesterol	rs1532624	0.94	A,C	C
						Blood metabolite levels	rs1800775	0.75	A,C	A
						Triglycerides	rs1800775	0.75	A,C	C
						Lipid traits	rs1800775	0.75	A,C	A
EDC4, LCAT	rs7199443	chr16:67841129	G	(L) MAX	T	Metabolic syndrome	rs8060686	0.88	T,T	T
	rs7196789	chr16:67927124	T	(L) YY1	C	Metabolic syndrome	rs8060686	0.97	C,T	T
						HDL cholesterol	rs2271293	0.71	C,G	A
	rs1134760	chr16:67964203	C	(L) POL2	T	Metabolic syndrome	rs8060686	0.95	T,T	T
	rs20549	chr16:67969930	G	(L) POL2	A	Metabolic syndrome	rs8060686	0.95	A,T	T
	rs1109166	chr16:67977382	C	(L) CREB1,FOXA1,FOXA2,HNF4A,HNF4G,NR2F2	T	Metabolic syndrome	rs8060686	0.92	T,T	T
Intergenic	rs8052763	chr16:75251659	C	(L) HNF4A	G	Type 1 diabetes	rs7202877	0.74	C,T	G
SLC7A5	rs12448415	chr16:87871096	G	(L) POL2	A	Blood metabolite levels	rs8051149	0.73	G,A	A
	rs12931876	chr16:87874182	T	(L) CTCF	C	Blood metabolite levels	rs8051149	0.78	C,A	A
	rs34508683	chr16:87876375	C	(L) EZH2,POL2	T	Blood metabolite levels	rs8051149	0.99	C,A	A
	rs28609922	chr16:87876631	A	(L) BHLHE40,MAX,P300	C	Blood metabolite levels	rs8051149	0.99	A,A	A
	rs56722741	chr16:87877313	G	(L) POL2	C	Blood metabolite levels	rs8051149	0.99	G,A	A
	rs4843270	chr16:87878072	A	(L) CTCF,POL2	C	Blood metabolite levels	rs8051149	0.99	A,A	A
	rs4843715	chr16:87878076	A	(L) POL2	G	Blood metabolite levels	rs8051149	1.00	A,A	A
	rs4843718	chr16:87878476	A	(L) CREB1	G	Blood metabolite levels	rs8051149	0.99	A,A	A
	rs35459492	chr16:87878883	G	(L) POL2	C	Blood metabolite levels	rs8051149	1.00	G,A	A
ORMDL3	rs1453559	chr17:38020419	C	(L) BHLHE40,DNASE,MXI1	T	Type 1 diabetes	rs2290400	0.85	T,T	C
	rs12936231	chr17:38029120	G	(L) ARID3A,CTCF,RAD21	C	Type 1 diabetes	rs2290400	0.87	C,T	C
	rs1031460	chr17:38072247	G	(L) POL2	T	Type 1 diabetes	rs2290400	0.86	G,T	C
	rs7224129	chr17:38075426	G	(L) POL2	A	Type 1 diabetes	rs2290400	0.92	A,T	C
	rs4065275	chr17:38080865	G	(L) CTCF,DNASE,RAD21	A	Type 1 diabetes	rs2290400	0.86	G,T	C
	rs8076131	chr17:38080912	A	(L) CTCF,HEY1,RAD21	G	Type 1 diabetes	rs2290400	0.73	A,T	C
MPP3	rs17742347	chr17:41846468	C	(L) POL2	T	Triglycerides	rs8077889	0.90	C,A	C
	rs17674998	chr17:41879544	A	(L) ZBTB33	G	Triglycerides	rs8077889	0.99	A,A	C
	rs9901676	chr17:41911818	T	(L) EZH2	C	Triglycerides	rs8077889	0.92	T,A	C
C17orf71	rs6503905	chr17:57287454	A	(L) CREB1,ELF1	G	Circulating myeloperoxidase levels	rs6503905	1.00	-	A
PGS1	rs4969182	chr17:76393030	T	(L) CEBPB,FOXA1,FOXA2,MAX,MYBL2,NR2F2	C	HDL cholesterol	rs4129767	0.96	C,A	G
	rs4969183	chr17:76393372	A	(L) BHLHE40	G	HDL cholesterol	rs4129767	0.96	G,A	G

* indicates other allele is enriched rather than enriched allele(Enr. Allele), ^{a}R2 and coupling (imbalanced allele, reported allele) calculated in 1000 Genome EUR. Risk allele increases risk or trait measurement based on reported effect allele in NHGRI GWAS catalog. - indicates no effect allele reported in catalog. Only reported variant in highest LD shown

Locus	Variant	Position	Enr. Allele	Imbalanced Proteins	Other Allele	Trait	Reported Variant	LD (r^2)[a]	Coupling	Risk. Allele
CCDC57	rs62078747	chr17:80055206	C	(L) MAX,MXI1	G	Blood metabolite levels	rs4625783	0.78	G,C	T
	rs59251877	chr17:80056498	G	(L) JUND,MAX	A	Blood metabolite levels	rs4625783	0.76	A,C	T
	rs7225637	chr17:80059758	G	(L) ELF1,TAF1	A	Blood metabolite levels	rs4625783	0.82	A,C	T
	rs11658040	chr17:80059891	C	(L) DNASE,HNF4A,MAX	T	Blood metabolite levels	rs4625783	0.82	T,C	T
	rs9908277	chr17:80060829	T	(L) MAX,MXI1	C	Blood metabolite levels	rs4625783	0.83	C,C	T
	rs9894129	chr17:80075700	A	(L) CEBPB,HDAC2,HNF4A,MAX	G	Blood metabolite levels	rs4625783	0.94	G,C	T
	rs9916649	chr17:80075739	G	(L) BHLHE40,CREB1,FOXA2,MAX,RXRA,SP1	A	Blood metabolite levels	rs4625783	0.94	A,C	T
	rs7218075	chr17:80076808	C	(L) FOXA2,POL2	G	Blood metabolite levels	rs4625783	0.93	G,C	T
	rs62079996	chr17:80076862	A	(L) FOXA1	G	Blood metabolite levels	rs4625783	0.92	G,C	T
	rs6502065	chr17:80095642	C	(L) CEBPB	T	Blood metabolite levels	rs4625783	0.98	T,C	T
AP3D1, DOT1L, SF3A2	rs7256735	chr19:2169121	T	(I) NKX2	G	Myocardial infarction	rs3803915	0.95	T,C	C
INSR	rs10410204	chr19:7224350	C	(L) FOXA1,FOXA2	T	Triglycerides	rs7248104	0.98	C,A	A
	rs7248104	chr19:7224431	A	(L) FOXA1,FOXA2,MAX,NR2F2,YY1	G	Triglycerides	rs7248104	1.00	-	A
ISYNA1	rs2303696	chr19:18548884	C	(L) DNASE	T	Blood metabolite levels	rs4808136	0.80	T,G	A
	rs8103622	chr19:18572834	C	(L) CTCF	T	Blood metabolite levels	rs4808136	0.92	C,G	A
PEPD	rs62102718	chr19:33891013	A	(L) HNF4G	T	Adiponectin levels	rs731839	0.76	A,A	G
						HDL cholesterol	rs731839	0.76	A,A	G
						Triglycerides	rs731839	0.76	A,A	G
SULT2A1	rs296368	chr19:48372298	T	(L) RAD21	C	Blood metabolite levels	rs182420	0.84	C,C	T
SPTLC3	rs1321940	chr20:12959885	G	(L) FOXA1	A	Blood metabolite levels	rs4814176	0.99	G,C	T
						Blood metabolite ratios	rs4814176	0.99	G,C	T
						LDL cholesterol	rs364585	0.99	G,G	A
EDEM2	rs6060266	chr20:33733078	C	(I) CTCF	T	Protein C levels	rs6120849	0.96	T,C	T
FITM2, R3HDML, HNF4A	rs4812816	chr20:42930872	C	(L) MAZ	A	Type 2 diabetes	rs6017317	0.74	C,T	G
	rs6065723	chr20:42956922	C	(L) MAFF,MAFK	T	Type 2 diabetes	rs6017317	0.77	C,T	G

* indicates other allele is enriched rather than enriched allele(Enr. Allele), [a]R2 and coupling (imbalanced allele, reported allele) calculated in 1000 Genome EUR. Risk allele increases risk or trait measurement based on reported effect allele in NHGRI GWAS catalog. - indicates no effect allele reported in catalog. Only reported variant in highest LD shown

APPENDIX 4: Cardiometabolic diseases and traits from GWAS catalog

Activated partial thromboplastin time
Adiponectin levels
Adiposity
Anthropometric traits
Apolipoprotein Levels
&beta;2-Glycoprotein I (&beta;2-GPI) plasma levels
Blood metabolite levels
Blood metabolite ratios
Blood pressure
Blood pressure (age interaction)
Blood pressure measurement (cold pressor test)
Blood pressure measurement (high sodium and potassium intervention)
Blood pressure measurement (high sodium intervention)
Blood pressure measurement (low sodium intervention)
Blood pressure (response to antihypertensive medication)
Blood pressure variability
Body mass index
Body mass index and cholesterol (psychopharmacological treatment)
Body mass index and fat mass
Body mass (lean)
Cardiac hypertrophy
Cardiac muscle measurement
Cardiac structure and function
Cardiac Troponin-T levels
Cardiovascular disease risk factors
Cardiovascular event reduction in the elderly at risk for vascular disease (statin therapy interaction)
Cardiovascular heart disease in diabetics
Cholesterol
Cholesterol and Triglycerides
Total Cholesterol
Coronary artery calcification
Coronary artery calcification (smoking interaction)
Coronary artery disease
Coronary artery disease or ischemic stroke
Coronary artery disease or large artery stroke
Coronary heart disease
Coronary heart disease in familial hypercholesterolemia
Coronary restenosis
Circulating myeloperoxidase levels (plasma)
Circulating myeloperoxidase levels (serum)
Congenital heart disease
C-reactive protein
C-reactive protein levels
Diabetes (gestational)
Diabetes (incident)
Diabetes related insulin traits
Diastolic blood pressure
Dilated cardiomyopathy
Fasting glucose-related traits
Fasting glucose-related traits (interaction with BMI)
Fasting insulin (interaction)
Fasting insulin-related traits
Fasting insulin-related traits (interaction with BMI)
Fasting plasma glucose
Fat body mass
Glycemic control in type 1 diabetes (HbA1c)
Glycemic traits
HDL cholesterol
HDL Cholesterol - Triglycerides (HDLC-TG)
Heart failure
Hypertension
Hypertension (young onset)
Hypertriglyceridemia
Hypertrophic cardiomyopathy
Insulin-related traits
Insulin resistance/response
LDL cholesterol
LDL (oxidized)
Lipid metabolism phenotypes
Lipid traits
Lipopolysaccharide induced cytokine levels
Lipoprotein diameter
Lp (a) levels
Metabolic syndrome
Metabolic syndrome (bivariate traits)
Metabolic traits
Metabolite levels
Myocardial infarction
Myocardial infarction (early onset)
Natriuretic peptide levels
NHDL cholesterol
Non-albumin protein levels
Obesity
Obesity and blood pressure
Obesity (early onset extreme)
Obesity (extreme)
Obesity-related traits
Palmitic acid (16:0) plasma levels
Palmitoleic acid (16:1n-7) plasma levels
Paraoxonase activity
Pericardial fat
Peripartum cardiomyopathy
Peripheral artery disease
Plasma cystastin c levels in acute coronary syndrome
Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid)
Plasma omega-6 polyunsaturated fatty acid leve;(linoleic acid)
Plasma omega-6 polyunsaturated fatty acid levels (adrenic acid)
Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid)
Plasma omega-6 polyunsaturated fatty acid levels (linoleic acid)
Proinsulin levels
Protein C levels
Soluble leptin receptor levels
Serum albumin level
Serum dimethylarginine levels (symmetric)
Serum metabolite levels
Serum prostate-specific antigen levels

Serum protein levels (sST2)
Serum total protein level
Serum uric acid levels
Subcutaneous adipose tissue
Sudden cardiac arrest
Systolic blood pressure
Triglycerides
Triglycerides-Blood Pressure (TG-BP)
Two-hour glucose challenge
Type 1 diabetes
Type 1 diabetes autoantibodies
Type 2 diabetes
Type 2 diabetes and 6 quantitative traits
Type 2 diabetes and gout
Type 2 diabetes and other traits
Type 2 diabetes (dietary heme iron intake interaction)
Type 2 diabetes nephropathy
Type 2 diabetes (young onset) and obesity
Vascular constriction
Visceral adipose tissue/subcutaneous adipose tissue ratio
Visceral fat
Waist circumference
Waist circumference and related phenotypes
Waist circumference (sex interaction)
Waist Circumference - Triglycerides (WC-TG)
Waist-hip ratio
Waist-to-hip circumference ratio (interaction)
Waist to hip ratio (sex interaction)

REFERENCES

1. World Health Organization *Global status report on noncommunicable diseases 2014*. 298 (2014).at <http://apps.who.int/iris/bitstream/10665/148114/1/9789241564854_eng.pdf>

2. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *The New England journal of medicine* **363**, 166–76 (2010).

3. McPherson, R. From Genome-Wide Association Studies to Functional Genomics: New Insights Into Cardiovascular Disease. *The Canadian journal of cardiology* **29**, 29–23 (2012).

4. Björkegren, J. L. M., Kovacic, J. C., Dudley, J. T. & Schadt, E. E. Genome-Wide Significant Loci: How Important Are They?: Systems Genetics to Understand Heritability of Coronary Artery Disease and Other Common Complex Disorders. *Journal of the American College of Cardiology* **65**, 830–845 (2015).

5. Willer, C. J. & Mohlke, K. L. Finding genes and variants for lipid levels after genome-wide association analysis. *Current opinion in lipidology* **23**, 98–103 (2012).

6. Willer, C. J. *et al.* Discovery and Refinement of Loci Associated with Lipid Levels. *Nature genetics* (2013).

7. Dastani, Z. *et al.* Novel Loci for Adiponectin Levels and Their Influence on Type 2 Diabetes and Metabolic Traits: A Multi-Ethnic Meta-Analysis of 45,891 Individuals. *PLoS Genetics* **8**, e1002607 (2012).

8. Wu, Y. *et al.* Genome-wide association study for adiponectin levels in Filipino women identifies CDH13 and a novel uncommon haplotype at KNG1-ADIPOQ. *Human molecular genetics* **19**, 4955–64 (2010).

9. Heid, I. M. *et al.* Clear detection of ADIPOQ locus as the major gene for plasma adiponectin: results of genome-wide association analyses including 4659 European individuals. *Atherosclerosis* **208**, 412–20 (2010).

10. Richards, J. B. *et al.* A genome-wide association study reveals variants in ARL15 that influence adiponectin levels. *PLoS genetics* **5**, e1000768 (2009).

11. Jee, S. H. *et al.* Adiponectin concentrations: a genome-wide association study. *American journal of human genetics* **87**, 545–52 (2010).

12. Chung, C.-M. *et al.* A genome-wide association study reveals a quantitative trait locus of adiponectin on CDH13 that predicts cardiometabolic outcomes. *Diabetes* **60**, 2417–23 (2011).

13. Warren, L. L. *et al.* Deep Resequencing Unveils Genetic Architecture of ADIPOQ and Identifies a Novel Low-Frequency Variant Strongly Associated With Adiponectin Variation. *Diabetes* **61**, 1297–301 (2012).

14. Fall, T. & Ingelsson, E. Genome-wide association studies of obesity and metabolic syndrome. *Molecular and cellular endocrinology* **382**, 740–57 (2014).

15. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).

16. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

17. Huyghe, J. R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nature genetics* **advance on**, (2012).

18. Huang, Q. Genetic Study of Complex Diseases in the Post-GWAS Era. *Journal of Genetics and Genomics* **42**, 87–98 (2015).

19. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.)* **337**, 1190–5 (2012).

20. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nature reviews. Genetics* **10**, 184–94 (2009).

21. Hindorff, L. *et al.* A Catalog of Published Genome-Wide Association Studies. at <http://www.genome.gov/gwastudies>

22. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *American journal of human genetics* **93**, 779–97 (2013).

23. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–9 (2010).

24. Carlsson, P. & Mahlapuu, M. Forkhead transcription factors: key players in development and metabolism. *Developmental biology* **250**, 1–23 (2002).

25. Tsompana, M. & Buck, M. J. Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin* **7**, 33 (2014).

26. Holwerda, S. & De Laat, W. Chromatin loops, gene positioning, and gene expression. *Frontiers in genetics* **3**, 217 (2012).

27. Lecompte, S. *et al.* Genetic and molecular insights into the role of PROX1 in glucose metabolism. *Diabetes* **62**, 1738–45 (2013).

28. Fogarty, M. P., Xiao, R., Prokunina-Olsson, L., Scott, L. J. & Mohlke, K. L. Allelic expression imbalance at high-density lipoprotein cholesterol locus MMAB-MVK. *Human molecular genetics* **19**, 1921–9 (2010).

29. Fogarty, M. P., Panhuis, T. M., Vadlamudi, S., Buchkovich, M. L. & Mohlke, K. L. Allele-Specific Transcriptional Activity at Type 2 Diabetes-Associated Single Nucleotide

Polymorphisms in Regions of Pancreatic Islet Open Chromatin at the JAZF1 Locus. *Diabetes* **62**, 1–7 (2013).

30. Fogarty, M. P., Cannon, M. E., Vadlamudi, S., Gaulton, K. J. & Mohlke, K. L. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. *PLoS genetics* **10**, e1004633 (2014).

31. Kulzer, J. R. *et al.* A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *American journal of human genetics* **94**, 186–97 (2014).

32. Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets. *Nature genetics* **42**, 255–9 (2010).

33. Lo, K. S., Vadlamudi, S., Fogarty, M. P., Mohlke, K. L. & Lettre, G. Strategies to fine-map genetic associations with lipid levels by combining epigenomic annotations and liver-specific transcription profiles. *Genomics* **104**, 105–12 (2014).

34. Bahassi, E. M. & Stambrook, P. J. Next-generation sequencing technologies: breaking the sound barrier of human genetics. *Mutagenesis* **29**, 303–10 (2014).

35. McVean, G. A. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

36. Furey, T. S. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics* **13**, 840–52 (2012).

37. Giresi, P. G. & Lieb, J. D. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods (San Diego, Calif.)* **48**, 233–9 (2009).

38. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols* **2010**, pdb.prot5384 (2010).

39. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213–8 (2013).

40. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

41. University of California, S. C. UCSC Genome Browser ENCODE downloads. at <ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/>

42. ENCODE Consortium ENCODE Project. at <https://www.encodeproject.org/>

43. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

44. Roadmap Epigenomics Project Human Epigenome Atlas. at <http://www.genboree.org/epigenomeatlas/index.rhtml>

45. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic acids research* **39**, D19–21 (2011).

46. Leinonen, R. *et al.* Improvements to services at the European Nucleotide Archive. *Nucleic Acids Research* **38**, D39–D45 (2009).

47. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome research* **22**, 1748–59 (2012).

48. Karczewski, K. J., Snyder, M., Altman, R. B. & Tatonetti, N. P. Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association. *PLoS genetics* **10**, e1004122 (2014).

49. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods* **9**, 215–6 (2012).

50. Paul, D. S. *et al.* Maps of open chromatin highlight cell type-restricted patterns of regulatory sequence variation at hematological trait loci. *Genome research* **23**, 1130–41 (2013).

51. Sheffield, N. C. *et al.* Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome research* **23**, 777–88 (2013).

52. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research* **40**, D930–4 (2012).

53. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome research* **22**, 1790–7 (2012).

54. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164 (2010).

55. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–4 (2012).

56. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science (New York, N.Y.)* **342**, 747–9 (2013).

57. Ni, Y., Hall, A. W., Battenhouse, A. & Iyer, V. R. Simultaneous SNP identification and assessment of allele-specific bias from ChIP-seq data. *BMC genetics* **13**, 46 (2012).

58. Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome research* **22**, 860–9 (2012).

59. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science (New York, N.Y.)* **328**, 235–9 (2010).

60. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology* **7**, 522 (2011).

61. Lalonde, E. *et al.* RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome research* **21**, 545–54 (2011).

62. Harvey, C. T. *et al.* QuASAR: Quantitative Allele Specific Analysis of Reads. *Bioinformatics (Oxford, England)* (2014).doi:10.1093/bioinformatics/btu802

63. Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome research* **21**, 1728–37 (2011).

64. Satya, R. V., Zavaljevski, N. & Reifman, J. A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic acids research* **40**, e127 (2012).

65. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–8 (2010).

66. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **26**, 589–95 (2010).

67. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**, 1851–8 (2008).

68. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).

69. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21 (2013).

70. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)* **26**, 873–81 (2010).

71. Schbath, S. *et al.* Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *Journal of computational biology : a journal of computational molecular cell biology* **19**, 796–813 (2012).

72. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics (Oxford, England)* **25**, 3207–12 (2009).

73. Birney, E., Lieb, J. D., Furey, T. S., Crawford, G. E. & Iyer, V. R. Allele-specific and heritable chromatin signatures in humans. *Human molecular genetics* **19**, R204–9 (2010).

74. Skelly, D. a, Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome research* **21**, 1728–37 (2011).

75. Van der Auwera, G. A. *et al. Current Protocols in Bioinformatics*. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **11**, 11.10.1–11.10.33 (John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2002).

76. Wall, J. D. *et al.* Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Research* **24**, 1734–1739 (2014).

77. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods* **4**, 651–7 (2007).

78. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science (New York, N.Y.)* **328**, 232–5 (2010).

79. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science (New York, N.Y.)* **342**, 750–2 (2013).

80. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science (New York, N.Y.)* **342**, 744–7 (2013).

81. León-Novelo, L. G., McIntyre, L. M., Fear, J. M. & Graze, R. M. A flexible Bayesian method for detecting allelic imbalance in RNA-seq data. *BMC genomics* **15**, 920 (2014).

82. Kharchenko, P. V, Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology* **26**, 1351–9 (2008).

83. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491–8 (2011).

84. Liu, E. Y., Li, M., Wang, W. & Li, Y. MaCH-admix: genotype imputation for admixed populations. *Genetic epidemiology* **37**, 25–37 (2013).

85. Hindorff, L. A. *et al.* A Catalog of Published Genome-wide Association Studies. at <http://www.genome.gov/gwastudies>

86. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–24 (2012).

87. Imielinski, M. *et al.* Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nature genetics* **41**, 1335–40 (2009).

88. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics* **42**, 1118–25 (2010).

89. Kotla, S., Singh, N. K., Heckle, M. R., Tigyi, G. J. & Rao, G. N. The transcription factor CREB enhances interleukin-17A production and inflammation in a mouse model of atherosclerosis. *Science signaling* **6**, ra83 (2013).

90. Wen, A. Y., Sakamoto, K. M. & Miller, L. S. The role of the transcription factor CREB in immune function. *Journal of immunology (Baltimore, Md. : 1950)* **185**, 6413–9 (2010).

91. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–9 (2011).

92. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–61 (2014).

93. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics (Oxford, England)* **27**, 1696–7 (2011).

94. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)* **27**, 1017–8 (2011).

95. Knight, D. Leukaemia inhibitory factor (LIF): a cytokine of emerging importance in chronic airway inflammation. *Pulmonary pharmacology & therapeutics* **14**, 169–76 (2001).

96. Gieger, C. *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**, 201–8 (2011).

97. Chen, W. *et al.* Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 7401–6 (2010).

98. Heap, G. A. *et al.* Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Human molecular genetics* **19**, 122–34 (2010).

99. Sun, W. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* **68**, 1–11 (2012).

100. Almal, S. H. & Padh, H. Implications of gene copy-number variation in health and diseases. *Journal of human genetics* **57**, 6–13 (2012).

101. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology* **34**, 816–34 (2010).

102. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics* **44**, 955–9 (2012).

103. Liu, E. Y. *et al.* Genotype imputation of Metabochip SNPs using a study-specific reference panel of ~4,000 haplotypes in African Americans from the Women's Health Initiative. *Genetic epidemiology* **36**, 107–17 (2012).

104. 1000 Genomes Project Consortium 1000 Genomes phase I version 3 data. at <ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/>

105. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics* **5**, 1752–1779 (2011).

106. Kundaje, A. ENCODE: TF ChIP-seq peak calling using the Irreproducibility Discovery Rate (IDR) framework. at <https://sites.google.com/site/anshulkundaje/projects/idr>

107. Willer, C. J. *et al.* Discovery and Refinement of Loci Associated with Blood Lipid Levels. *Nature genetics*

108. Bryzgalov, L. O. *et al.* Detection of regulatory SNPs in human genome using ChIP-seq ENCODE data. *PloS one* **8**, e78833 (2013).

109. Furey, T. Allele-Aware ALignments for the Investigation of GeNetic Effects on Regulation (AA-ALIGNER). at <http://fureylab.web.unc.edu/software/AA_ALIGNER/>

110. Wiernsperger, N. Hepatic function and the cardiometabolic syndrome. *Diabetes, metabolic syndrome and obesity : targets and therapy* **6**, 379–88 (2013).

111. Rorsman, P. & Braun, M. Regulation of insulin secretion in human pancreatic islets. *Annual review of physiology* **75**, 155–79 (2013).

112. Lafontan, M. Adipose tissue and adipocyte dysregulation. *Diabetes & metabolism* **40**, 16–28 (2014).

113. Innocenti, F. *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS genetics* **7**, e1002078 (2011).

114. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS biology* **6**, e107 (2008).

115. Nica, A. C. *et al.* The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. *PLoS Genetics* **7**, e1002003 (2011).

116. Fadista, J. *et al.* Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 13924–9 (2014).

117. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research* **42**, 2976–2987 (2013).

118. Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422–33 (2008).

119. Slavin, T. P., Feng, T., Schnell, A., Zhu, X. & Elston, R. C. Two-marker association tests yield new disease associations for coronary artery disease and hypertension. *Human genetics* **130**, 725–33 (2011).

120. Roman, T. *et al.* Functional regulatory variants at the GALNT2 human high-density lipoprotein cholesterol locus. *in preparation*

121. Davis, J. P. personal communication.

122. Nouaille, S. *et al.* Interaction between the insulin receptor and Grb14: a dynamic study in living cells using BRET. *Biochemical pharmacology* **72**, 1355–66 (2006).

123. Rusanescu, G., Weissleder, R. & Aikawa, E. Notch signaling in cardiovascular disease and calcification. *Current cardiology reviews* **4**, 148–56 (2008).

124. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research* **22**, 1798–812 (2012).

125. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods* **6**, 283–9 (2009).

126. Lassmann, T., Hayashizaki, Y. & Daub, C. O. TagDust--a program to eliminate artifacts from next generation sequencing data. *Bioinformatics (Oxford, England)* **25**, 2839–40 (2009).

127. Picard Tools. at <http://picard.sourceforge.net.>

128. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics* **43**, 333–8 (2011).

129. Palm, W. & De Lange, T. How shelterin protects mammalian telomeres. *Annual review of genetics* **42**, 301–34 (2008).

130. Davies, R. W. *et al.* A genome-wide association study for coronary artery disease identifies a novel susceptibility locus in the major histocompatibility complex. *Circulation. Cardiovascular genetics* **5**, 217–25 (2012).

131. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nature genetics* **43**, 339–44 (2011).

132. Dichgans, M. *et al.* Shared Genetic Susceptibility to Ischemic Stroke and Coronary Artery Disease: A Genome-Wide Analysis of Common Variants. *Stroke* **45**, 24–36 (2013).

133. Qi, L. *et al.* Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Human molecular genetics* **19**, 2706–15 (2010).

134. Kooner, J. S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nature genetics* **43**, 984–9 (2011).

135. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics* **46**, 234–44 (2014).

136. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics* **40**, 638–45 (2008).

137. Perry, J. R. B. *et al.* Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. *PLoS genetics* **8**, e1002741 (2012).

138. Shu, X. O. *et al.* Identification of New Genetic Risk Variants for Type 2 Diabetes. *PLoS Genetics* **6**, e1001127 (2010).

139. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics* **42**, 579–89 (2010).

140. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science (New York, N.Y.)* **316**, 1336–41 (2007).

141. Cho, Y. S. *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nature genetics* **44**, 67–72 (2012).

142. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–13 (2010).

143. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genetics* **40**, 189–197 (2008).

144. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature genetics* **41**, 56–65 (2009).

145. Sandhu, M. S. *et al.* LDL-cholesterol concentrations: a genome-wide association study. *The Lancet* **371**, 483–491 (2008).

146. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics* **40**, 161–9 (2008).

147. Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nature genetics* **42**, 210–5 (2010).

148. Zhou, L. *et al.* A genome wide association study identifies common variants associated with lipid levels in the Chinese population. *PloS one* **8**, e82420 (2013).

149. Coram, M. A. *et al.* Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. *American journal of human genetics* **92**, 904–16 (2013).

150. Aulchenko, Y. S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature genetics* **41**, 47–55 (2009).

151. Kim, Y. J. *et al.* Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nature genetics* **43**, 990–5 (2011).

152. Weissglas-Volkov, D. *et al.* Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci. *Journal of medical genetics* **50**, 298–308 (2013).

153. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics* **41**, 35–46 (2009).

154. Figarska, S. M., Vonk, J. M. & Boezen, H. M. NFE2L2 polymorphisms, mortality, and metabolism in the general population. *Physiological genomics* **46**, 411–7 (2014).