Summer 8-15-2021

Association of Structural Variation (SV) with Cardiometabolic Traits in Finns

Lei Chen

Washington University in St. Louis

Recommended Citation

Chen, Lei, "Association of Structural Variation (SV) with Cardiometabolic Traits in Finns" (2021). *Arts & Sciences Electronic Theses and Dissertations*. 2485.
https://openscholarship.wustl.edu/art_sci_etds/2485

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Human and Statistical Genetics

Dissertation Examination Committee:
Nathan O. Stitziel, Chair
Ira M. Hall, Co-Chair
Adam Locke
Nan Lin
Timothy Peterson
John Rice
Nancy Saccone

Association of Structural Variation (SV) with Cardiometabolic Traits in Finns

Arts & Sciences Graduate Students
by
Lei Chen

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2021
St. Louis, Missouri

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 1

Chapter 2

Chapter 3

Acknowledgments

Before came to the United States for a PhD, I was mentally prepared for all the miserable stories I heard over the years. However, the closer I am to the end, the more I realized how much I enjoyed this journey and how hard it is to say goodbye to the people and life here. There were so many factors associated with those joy and accomplishments, while in a metaphor of PCA, the first component would be labeled with the name of my thesis advisor, Dr. Ira Hall. Five years ago, I decided to join the lab fascinated by his cool science, and that turned out to be a decision for which I will never regret. He is not only an excellent role model as a scientist, but also a caring mentor and thoughtful leader. He always has the ability to set up challenging but realistic research goals, a superpower that gradually pulled me out of my comfort zone and accelerated my growth as a trainee scientist. I also appreciate his direct, honest and respectful communicating style, which made me feel safe and comfortable to start the frank conversations about my feelings and thoughts. More than my academic advisor, Ira is also my trusty mentor and friend for lifetime.

I also have to emphasize the contribution from my committee members, especially Nate and Adam, who have been dedicating to this project from day one, providing guidance and connecting me to resources. Thanks to Nan, Nancy and John for sharing their perspectives through the lens of statistics and thank to Tim for his valuable inputs from the view of genomics and bioinformatics.

Next I want to thank the members of Hall lab, with the first round of applause going to Haley Abel and Niel Das, who helped me overcome the most challenging part of the SV project – developing and adapting the read-depth based methods to detect CNVs. Those two geniuses also inspired me by their incredible productivity and the bandwidth of supporting multiple crucial lab

projects in parallel. Thanks to my fellow graduate students, both the “senior” batch – Sasha, Liron, Colby and Ryan Smith, and the “junior” batch – Shuangjia, Wen-Wei, and Xinxin, for all the practice talks they organized, all the innovative ideas they shared, and all the yummy snacks they put above the lab fridge. A special appreciation goes to Sasha, who took care of us like a big sister and voluntarily offered me help when I was still adapting to the new lab and the culture difference. I am so happy to see her achievements both in career and in personal life and wish her all the best in journey forward. Also, Liron, who owns a half of my acknowledgement of chapter 2. And there are also people in CCDG analysis group contributed to the fundamental work for this project: Bo, Allison, Dave, Erica, CJ, Ryan Christ, Krishna, Chad and Eddie, as well as the MGI production team. I also want to thank our collaborators in FinMetSeq Consortium, especially all those in Finland who shared the precious samples and phenotype data with us.

Many other people in WashU also helped me along the way. First is the McDonnell International Scholar Academy, who financially supported me with a five-year fellowship and provided a great platform to meet talents from all over the world. Special thanks to James and Mary Wertsch, who treated scholars like family and invited us for Thanksgiving dinner every year. I also would like to thank the HSG program of DBBS, the former and current coordinators Jeanne and Sara, the current directors John and Nate, also Pat, the former HSG director who offered me a lot of career insights from his own transition to industry. Another important person is Thi, who has been my career advisor and personal inspiration as a strong East Asian woman. A special thanks to Meghann Feely and WashU recreation center, for the plenty of fun workouts they provided before and during the pandemic that kept me active and healthy.

Then there are all my friends. Jiayang, Zhen, Jiang and Wenjun are my best friends in graduate school and also the people I missed the most in St Louis. Thanks to my rock-climbing buddies – Rachel, Ya-Lin, Wen-Wei, Yiran, and Hsun-Chia, for all the courage we gave to each other on and off the wall. Special thanks to the Morris Lab members too, for inviting Wenjun's roommate and her dog to all the fun events. Big thanks to Qing, Mingxin and Yi, my besties in other parts of the world who nurtured our long-distance friendship and served as my sounding boards for years.

Last but not the least, I would like to acknowledge my families in China, especially my parents. Growing up in a working-class family in East Asian usually means lots of pressure towards material success, while because of my parents and things they believed to be important in life, I got the privilege to pursue my own interests, happiness and a free soul.

Lei Chen

Washington University in St. Louis

August 2021

Dedicated to my 4-year-old puppy,

Shumai.

ABSTRACT OF THE DISSERTATION

Association of Structural Variation (SV) with Cardiometabolic Traits in Finns

for Arts & Sciences Graduate Students

by

Lei Chen

Doctor of Philosophy in Biology and Biomedical Sciences

Human and Statistical Genetics

Washington University in St. Louis, 2021

Professor Nathan O. Stitziel, Chair

Professor Ira M. Hall, Co-Chair

Cardiovascular diseases (CVDs) are known to be associated with a variety of quantitative risk factors such as cholesterol, metabolites, and insulin. Understanding the genetic basis of these quantitative traits can shed light on the etiology, prevention, diagnosis, and treatment of disease. However most prior trait-mapping studies have focused on single nucleotide variants (SNVs) and Indels, with the contribution of structural variation (SV) remaining unknown. In this thesis, we present the results of a study examining genetic association between SVs and cardiometabolic traits in the Finnish population. In the first chapter, we used sensitive methods to identify and genotype 129,166 high-confidence SVs from deep whole genome sequencing (WGS) data of 4,848 individuals. We tested the 64,572 common and low frequency SVs for association with 116 quantitative traits, and tested candidate associations using exome sequencing and array genotype data from an additional 15,205 individuals. We discovered 31 genome-wide significant associations at 15 loci, including two novel loci at which SVs have strong phenotypic effects: (1) a deletion of the *ALB* gene promoter that is greatly enriched in the Finnish population and causes decreased serum albumin level in carriers (p=1.47x10^{-54}), and is also associated with increased levels of total cholesterol (p=1.22x10^{-28}) and 14 additional cholesterol-related traits, and (2) a multiallelic copy number variant (CNV) at *PDPR* that is strongly associated with pyruvate (p=4.81x10^{-21}) and alanine (p=6.14x10^{-12}) levels and resides within a structurally complex genomic region that has accumulated many rearrangements over evolutionary time. We also confirmed six previously reported associations, including five led by stronger signals in single nucleotide variants (SNVs), and one linking recurrent *HP* gene deletion and cholesterol levels (p=6.24x10^{-10}), which was also found to be strongly associated with increased glycoprotein level (p=3.53x10^{-35}). The result of this chapter confirms that integrating SVs in trait-mapping studies will expand our knowledge of genetic factors underlying disease risk.

Chapter 2 and chapter 3 present two side projects derived from chapter 1: chapter 2 focused on an insulin associated chromosome 1 CNV which turned out to have indirectly measured the mitochondrial DNA copy number, of which the direct measurement showed stronger association with multiple metabolic traits. In chapter 3 we presented a pilot study of applying machine learning to genetics problems unsolvable by traditional methods. We built multi-layer neural network models to impute the highly polymorphic *AMY1* CNVs, and showed the boosted performance compared to baseline regression models as well as the best practice employed in previous publication. Both chapters proposed solutions to new questions rising from the main SV project and provided the preliminary data for other ongoing or upcoming projects in our group.

Chapter 1:

The contribution of SVs to quantitative cardiometabolic traits in Finns

The contents of this chapter were published in *American Journal of Human Genetics* in 2021: Chen L, Abel HJ, Das I, …, Stitziel NO, Hall IM. Association of structural variation with cardiometabolic traits in Finns. Am J Hum Genet. 2021 Apr 1;108(4):583-596. doi: 10.1016/j.ajhg.2021.03.008. PMID: 33798444; PMCID: PMC8059371.

1.1 Introduction

1.1.1 Studying the genetics of cardiometabolic traits

Cardiovascular diseases (CVDs) are a series of heart and blood vessel conditions, which cause 17.7 million deaths each year -- contributing 31% of the worldwide mortality(World Health Organization, 2017). Metabolic syndromes, including obesity, high insulin level, high blood pressure, fasting glucose and abnormal lipids cholesterol level are the common risk factors of CVD, coronary heart disease and type 2 diabetes(Wilson et al. 2005). Family studies and population genetic studies both suggest significant heritability of metabolic syndromes and CVDs(Hegele and Pollex 2005; Vattikuti, Guo, and Chow 2012; Pollex and Hegele 2006). Studying the genetics of related metabolic traits could help understand the etiology, prevention, diagnosis and treatment of CVD – for instance, to provide new targets for gene therapy(Wolfram and Donahue 2013). Family and population-based studies have shown significant heritability for many cardiometabolic traits(Kolifarhood et al. 2019; Kim et al. 2015; Campbell Am 2017; Hagenbeek et al. 2020), and prior genome-wide association studies (GWAS) have identified hundreds of associated loci(Willer et al. 2013; Fall and Ingelsson 2014; Visscher et al. 2017). However, limited by cost and technology, most previous genome-wide trait mapping studies have focused on common single nucleotide polymorphisms (SNPs) detectable by genotyping arrays, or SNPs and small insertion/deletion variants (indels) that are routinely assessed in genome and exome sequencing studies, leaving out the contributions of larger and more complex forms of genome variation.

1.1.2 The role of SVs in common and complex diseases – prior studies

Of particular interest is the contribution of genome structural variation (SV), which encompasses diverse variant types larger than 50bp in size, including copy number variants (CNVs), mobile element insertions (MEIs), inversions, and complex rearrangements. Although rare and de novo SVs are known to cause various rare human disorders, and somatic SVs play a central role in cancer biology, the extent to which SVs contribute more generally to common diseases and other complex traits is unclear. Early microarray-based CNV association studies from the Wellcome Trust Case Control Consortium (WTCCC) and others(Wellcome Trust Case Control Consortium et al. 2010; McCarroll et al. 2008; Myocardial Infarction Genetics Consortium et al. 2009) were largely unsuccessful in identifying new disease associated variants or genes, suggesting a minimal contribution to common disease. However, in retrospect this is perhaps unsurprising given that these studies were fairly small and underpowered relative to our current knowledge of complex trait genetic architecture and limited to a small subset of SVs – namely, large CNVs.

Recent studies from the GTEX consortium have assessed the contributions of SVs to gene expression across tissues, where SVs comprise 3.5-6.8% of eQTLs, and on average have larger effect sizes than SNP eQTLs(Chiang et al. 2017; X. Li et al. 2017). The contribution of SVs to human disease in general, and CVD in particular, has remained an open question.

For a long time our knowledge of disease-causing SVs was restricted to a few well-studied loci(Usher et al. 2015; Wu et al. 2014; Boettger et al. 2016) or large de novo or somatic CNVs that cause various genomics disorders and cancer visible at cytogenetic level with severe phenotypic effects such as Down syndrome(Jacobs 1959) or leukemia(Nowell 1962), until the development of DNA microarray technology, which enabled the high-throughput genotyping of copy number variation (CNV). In 2010, The Wellcome Trust Consortium published the first genome-wide association study specifically designed for CNVs using Comparative Genomic Hybridization (CGH) assay(Wellcome Trust Case Control Consortium et al. 2010). The study genotyped ~3,000 polymorphic CNVs and tested them with eight common diseases in ~19,000 individuals and concluded that most common CNVs were tagged by nearby SNPs and had small contributions to common traits in humans. As well-designed as it was, this study only assayed the highly polymorphic loci detected from the ~400 pilot samples, leaving behind a large number of SVs with lower frequency. Alternatively, other studies tried to utilize existing SNP array data to detect CNVs. However, since the assays were initially designed for SNPs, the probe density and the noise of intensity signals restricted the resolution and genotype quality of detected CNVs, even those well powered studies with tens of thousands of samples were only able to identify a few novel disease-associated loci(Aguirre, Rivas, and Priest 2019; Macé et al. 2017; Marshall et al. 2017). Several later studies performed targeted analysis of known SVs combined with larger-scale GWAS data(Boettger et al. 2016; Usher et al. 2015; Zekavat et al., n.d.), leading to the association of structural alleles at *HP* and *LPA* with cholesterol levels. More recent array-based CNV association studies with large sample sizes (>50,000 individuals) have revealed several genome-wide significant CNV loci for anthropometric traits and coronary disease, but these studies focused on extremely large CNVs representing <1% of the overall SV burden, leaving most SVs untested(Macé et al. 2017; Aguirre, Rivas, and Priest 2019; Y. R. Li et al. 2020).

1.1.3 Detecting SVs by sequencing data

Along with the rise of next generation sequencing (NGS) technologies, many NGS variant calling methods were invented for SVs, most of them were built on short pair-end reads data. Despite the variety of features and purposes, the short-reads SV callers can be summarized into two categories – read-depth based methods and breakpoint mapping methods. The former detects CNVs from the coverage data and the later aggregates evidence from discordant alignment patterns to identify CNVs as well as copy number neutral SVs such as inversions and translocations. Those two categories of SV calling algorithms are complementary to each other in terms of advantages and limitations. Read-depth approaches are able to cover multiallelic CNVs and repetitive regions but cannot predict the precise variant boundaries and are prone to technical confounders such as PCR-induced coverage biases. On the other hand, breakpoint mapping methods provide high resolution for predicting SV boundaries, while often fail to recognize multi-allelic SVs or the SVs within complex regions. Both approaches are preferably applied to whole genome sequencing (WGS) data, and have been utilized in large-scale projects such as 1000 Genomes Project(Sudmant et al. 2015), CCDG(Abel et al. 2020) and gnomAD(Collins et al. 2020) to extensively survey the SV landscape in human populations. However, due to the sequencing costs and analysis complexity, trait association studies have not been conducted genome-wide with adequate sample size except for a few attempts with whole exome sequencing (WES) data(Maxwell et al. 2017; Ruderfer et al. 2016), which were restricted to coding regions.

1.1.4 Identifying CVD associated SVs in Finns

The northern and eastern Finnish populations have historically been genetically isolated, mostly originated from small isolated geographic groups and rapidly expanded in recent genetic bottleneck events, during which the founder alleles were either enriched or lost under random drift. Previous studies proved that studying these populations boosted the power of detecting deleterious variants, many of which were preserved in Finnish population with detectable frequency and depleted in non-Finnish Europeans (NFE) under negative selection pressure(Lim et al. 2014; Davis et al. 2017; Locke et al. 2019).

To study the effects of SVs with higher sensitivity and across a broader spectrum of traits, we combined read-depth and breakpoint mapping methods to detect 64,572 high-confidence autosomal SVs from the WGS data of 4846 Finnish individuals, 4,030 of which also have extensively measured quantitative cardiometabolic phenotypes. To increase the power of trait mapping analysis, we genotyped 2,053 candidate SVs in an additional ~15,000 Finnish samples with WES and SNP array genotype data and tried to replicate the suggestive association signals observed in the WGS analysis. We identified 15 genome-wide significant loci associated with 31 metabolic traits, nine of which passed multiple testing correction after adjusting for the number of phenotypes. We then combined information from other types of variants, local copy number profiles, nearby genomic annotations to investigate the significant loci, and demonstrate here the interesting biology of several trait-associated SVs, including a multiallelic CNV affecting *PDPR* gene associated with pyruvate level, and a Finnish-enriched promoter deletion on *ALB* gene associated with multiple metabolic traits.

1.2 Material and Methods

1.2.1 Samples and phenotype collection

The genomic data in this study come from 10,197 METSIM participants collected from Kuopio in Eastern Finland, and 10,192 FINRISK participants collected from northeastern Finland. Both studies were approved by the Ethics Committees in Finland and all individuals contributing samples provided written informed consent. Besides collecting genotype data by SNP array and exome sequencing, both studies measured up to 254 quantitative cardiometabolic traits, among which we selected 116 traits with adequate sample sizes to maintain trait-mapping power (see below). All phenotype data were residualized for trait-specific covariates and transformed to a standard normal distribution by inverse normalization. Complete details of sample collection, genotype acquisition, and trait adjustments were described previously(Locke et al. 2019).

1.2.2 Power estimation and phenotype selection

Phenotypes with limited sample size are likely to be underpowered in trait-mapping analysis and increase the test burden if included. Thus, we selected 116 traits with large enough sample size that guaranteed 80% power to detect a hypothesized rare SV (Minor allele count (MAC) =10) with strong effect (explained 8.4% of the additive quantitative trait locus (QTL) variance, a contribution comparable to the effect of SV expression QTLs(Chiang et al. 2017)). We estimated the minimum required sample size as 375 through an analytical approach implemented in Genetic Power Calculator(Purcell, Cherny, and Sham 2003). Several other assumptions for the calculation are: 1. All samples are independent (sibship size=1); 2. The top signal is in perfect linkage disequilibrium (LD) with the causal variant; and 3. type I error rate=$1x10^{-6}$.

1.2.3 Generation of SV callsets from WGS data

For SV discovery, we used WGS data from 3,082 METSIM participants and 1,114 FINRISK participants sequenced at the McDonnell Genome Institute under the NHGRI Centers for Common Disease Genomics (CCDG) program. To increase variant detection sensitivity, we also included 779 additional Finnish participants from other cohorts and 112 multi-ethnic samples from 1000 Genomes (1KG) Project. All genomes were sequenced at >20x coverage on the Illumina HiSeq X and NovaSeq platforms with paired-end 150bp reads.

WGS data were aligned to the GRCh38 reference genome using BWA-MEM and processed using the functional equivalence pipeline(Regier et al., n.d.). An SV callset based on breakpoint mapping was generated using our recently published workflow(Larson et al. 2019) using the same methods as in our recent study of 17,795 human genomes(Abel et al. 2020). Briefly, we ran LUMPY (v0.2.13)(Layer et al. 2014), CNVnator (v0.3.3)(Abyzov et al. 2011), and svtyper (v0.1.4)(Chiang et al. 2015) to perform per-sample variant calling. After removing 22 samples that failed quality control, we merged sites discovered in all the samples and re-genotyped all sites in all samples to create a joint callset using svtools (v0.3.2)(Larson et al. 2019). Each variant was characterized as either deletion (DEL), duplication (DUP), inversion (INV), mobile element insertion (MEI), or generic rearrangement of unknown architecture (BND), based on comprehensive review of its breakpoint genotype, breakpoint coordinates, genome annotation, and read-depth evidence, as described previously(Larson et al. 2019; Abel et al. 2020). According to our definition of SV, we filtered variants smaller than 50bp. Moreover, we tuned the callset based on Mendelian error rate and flagged BNDs with mean sample quality (MSQ) score <250 and INVs with MSQ <100 as low-confidence variants. Details about this QC strategy are described elsewhere(Abel et al. 2020). For convenience, we refer to this as the "LUMPY callset".

We applied two read-depth based CNV detection methods to WGS data to detect variants that might be missed by breakpoint mapping. GenomeSTRiP(Handsaker et al. 2015) is an established tool for cohort-level CNV discovery that has proven effective in many prior studies; however, when using the recommended parameters (as we did here), detection is limited to larger CNVs (>1kb) within relatively unique genomic regions. Thus, in parallel we used a custom cohort-level CNV detection pipeline based on CNVnator(Abyzov et al. 2011) to detect smaller and more repetitive CNVs (see below).

We adapted the original GenomeSTRiP pipeline (v2.00.1774) for the large cohort of 5,087 Finnish samples: after the SVPreprocess step, samples were grouped by study cohorts and sorted by sequencing dates, then split into 54 batches with maximum size of 100. CNVs were detected within each batch by CNVDiscoveryPipeline and classified as either deletion (DEL), duplication (DUP), or mixed CNV (mCNV), with both copy number gain and loss existing in the population (referred to as "multiallelic CNV" in the text). Next, we concatenated variants from the 54 batch VCFs and re-genotyped all variants in all samples using SVGenotyper to produce a joint callset. Then we ran several GenomeSTRiP annotators (CopyNumberClassAnnotator, RedundancyAnnotator) to reclassify variants and remove redundant variant calls. During callset generation, 72 samples with abnormal read-depth profiles were excluded.

The read-depth based "CNVnator" callset was constructed using a custom pipeline that took as inputs the individual-level CNV callsets generated by CNVnator during the svtools pipeline. After removing samples with abnormal read-depth profiles, CNV calls from 4,979 samples were sorted and merged using the svtools pipeline. All merged CNV calls were re-genotyped in all samples using CNVnator. Within each connected component of overlapping CNV calls, individual variant calls were clustered based on correlation of copy-number profiles and by pairwise overlap. For each cluster, a single candidate was chosen to represent the underlying CNV. For sites with carrier frequency >0.1%, we fit the copy number distribution to a series of constrained Gaussian Mixture Models (GMMs) with varying numbers of components, and selected the site with the "best" variant representation based on a set of model metrics, including the Bayesian Information Criterion (BIC) and the distance between cluster means ("mean_sep"). For the remaining sites we selected those with the most significant copy number difference between carriers and non-carriers. With the same criteria used in GenomeSTRiP, we assigned integer copy number genotypes and CNV categories to the variants.

We used array intensity data for 2,685 METSIM samples to estimate the false discovery rate (FDR) under different filtering criteria, and to tune both CNV callsets. FDR was estimated from the Intensity Rank Sum (IRS) test statistics based on CNVs intersecting at least two SNP probes. Based on the FDR curves (**Figure 1.1**) we excluded GenomeSTRiP variants with GSCNQUAL score<2 and CNVnator DELs and DUPs with mean_sep < 0.47 or low carrier counts (DUPs<1, DELs<5, mCNVs<7).

To eliminate likely false positive calls introduced by sequencing artefacts, we excluded 612 LUMPY SVs, 740 GenomeSTRiP SVs, and 1098 CNVnator SVs that were highly enriched in any of the three sequencing year batches ($P<10^{-200}$ from Fisher's exact test). We further excluded 3 samples in the LUMPY callset, 72 samples in the GenomeSTRiP callset, and 12 samples in the CNVnator callset that carried abnormal numbers of variants (outlier samples defined by the difference of per-sample SV count from median divided by median absolute deviation (mad) larger than 10 for LUMPY/GenomeSTRiP or larger than 5 for CNVnator). Together with the samples that failed QC during variant calling, the combined list of outliers consists of 84 METSIM samples, 56 FINRISK samples, and 99 samples from other cohorts. More information about sample- and variant-level exclusions can be found in **Table 1.1**.

For each high-confidence callset, we evaluated the final FDR by using the IRS, and ran the TagVariants annotator in GenomeSTRiP to estimate the proportion of SVs in LD with nearby SNPs (R_{max}^2>=0.5, flanking window size=1Mb). We calculated the overlap fraction between SV callsets by bedtools(Quinlan and Hall 2010) intersect (v2.23.0) requiring >50% reciprocal overlap between variants. To evaluate the genotype redundancy within and between callsets, we compared the original variant counts and the equivalent number of independent genetic variables estimated by a matrix decomposition method implemented in matSpDlite(J. Li and Ji 2005), using the genotype correlation matrix as input. The space clustering was evaluated by running bedtools cluster with -d (max distance) specified as 10bp.

1.2.4 Association test with WGS data

For CNV callsets, we defined minor allele count (MAC) as the number of samples with different genotypes from the mode copy number. We kept the conventional MAC definition for the LUMPY callset since it primarily contains biallelic SVs. We set the minimum MAC threshold as 10 for variants to be included in the trait association test. We renormalized the phenotype data of the WGS samples by rank-based inverse normal transformation. We performed single-variant association tests across all renormalized metabolic traits using the EMMAX model(Kang et al. 2010) implemented in EPACTS (v3.2.9) software (see **Web Resources**). In the model, we specified the dosage-format input genotype variables as the integer copy number genotype for GenomeSTRiP variants, allele balance for LUMPY variants, and raw decimal copy number for CNVnator variants. We also incorporated in the model a kinship matrix derived from SNP data by EPACTS to account for sample relatedness and population stratification. For each multiallelic CNV, one single variant test was performed between the phenotype and the copy number value of the interval.

We applied matSpDlite(J. Li and Ji 2005) to estimate the equivalent number of independent tests. The genome-wide significance threshold was set at $1.89\text{x}10^{-6}$ after Bonferroni correction at level $\alpha = 0.05$ over 26,495 independent genetic variables, and the experiment-wide significance threshold was set as 3.32x10^{-8} to further correct for the 57 independent phenotypic variables also estimated using matSpDlite(J. Li and Ji 2005).

1.2.5 Replication using exome and array data

We attempted to replicate the association signals with a nominal $p<0.001$ in WGS analysis using genotype data for an additional 15,205 FinMetSeq participants (**Figure 1.2**). To achieve this, we employed two approaches to infer the genotypes of candidate SVs from WES and array data: WES read depth analysis for CNVs and genotype imputation for biallelic SVs.

We separated the WES alignment data into two batches: the first composed of 10,379 samples sequenced with 100bp paired-end reads and the second composed of 9,937 samples sequenced with 125bp paired-end reads. For samples in each batch, we calculated the per-sample per-exon coverage by GATK(Auwera et al. 2013) DepthOfCoverage (v3.3-0) and adopted the data processing steps from the XHMM (v1.0) pipeline(Fromer and Purcell 2014) to convert the raw coverage data into PCA-normalized read-depth z-scores. Duplicated and outlier samples were filtered simultaneously, with 9,537 samples left in batch1 and 9,864 samples left in batch2. We calculated the correlation between SV genotypes from WGS data and the normalized read-depth z-scores of exons intersected or nearby (<5kb) using samples with both WES and WGS data. Exons with $R^2<0.1$ were filtered out and the rest were passed on to validation, restricted to samples absent from the WGS analysis (n=15,205). The genetic relationship matrix used for WES replication was generated in a previous study(Locke et al. 2019). We later did a meta-analysis under a fixed effect model using METASOFT (v2.0.1)(Han and Eskin 2011) to combine the results from the two WES batches, considering the two sequencing batches were actually sampled from the same population.

We standardized the genotype representations of 2,291 biallelic candidate SVs, with copy number genotypes of duplications (CN=2,3,4) and deletions (CN=0,1,2) converted to allelic genotype format (GT=0/0, 0/1, 1/1), and extracted the SNPs and indels in the 1 Mb flanking regions of those SVs from the GATK callset generated from the same WGS data. We then phased the joint VCF with Beagle (version 5.1)(Browning, Zhou, and Browning 2018) to build a reference panel composed of 3,908 high-quality samples shared by the SV callset and the SNP callset. Then, we imputed the SV genotype in the additional 15,125 FinMetSeq samples with array genotype data by running Beagle on the genotyped SNPs. We filtered out low-imputation-quality SVs with DR2<0.3 reported by Beagle (the estimated correlation between imputed genotype and real genotype of each variant); then ran the EMMAX model on the 1,705 well-imputed SVs with the corresponding traits.

58 of the 2,053 candidate SVs had both imputed genotype and WES read-depth genotype, so we compared the imputation DR2 with exon-SV genotype R^2, then chose the measurement that was most well correlated with the WGS data. Considering the differences between directly measured WGS-based SV genotypes and predicted genotypes estimated from WES and array data, for SVs with consistent direction of effects across the discovery stage (WGS data only) and replication stage, we used Fisher’s method to combine the p-values (instead of conventional meta-analysis models that assume effect sizes across studies were sampled from the same distribution). As a sanity check for the imputation quality, we conducted leave-one-out validation for the eight genome-wide significant SVs using the reference panel only. Specifically, we took

one sample out each time as a test genome and imputed the SV genotype using the other 3,907 samples as reference and repeated the process 3,908 times to calculate the validation rate.

The array data and WES data were aligned to reference genome GRCh37 while the WGS data were aligned to reference genome GRCh38. For analysis, the coordinates were lifted over using the LiftOver utility from the UCSC GenomeBrowser (see **Web Resources**). Considering the LiftOver works less efficiently for intervals (e.g., exons) than single-base coordinates (e.g., SNPs), we chose different strategies for the WES experiment and the imputation experiment to minimize information loss. For the WES dataset, we converted the CNV coordinates from GRCh38 to GRCh37; 5,391 successfully converted (2310 intersected with exons) while 264 failed (78 intersected with exons). We dropped the CNVs that failed conversion. For the imputation experiment, we converted the coordinates of array-genotyped SNPs to GRCh38, thus all the biallelic SV candidates were kept in the replication experiment. A small number of SNPs (0.1%) dropped out during this process, which should not have big impact on the imputation considering the abundance of SNPs around each SV and the fact that this only happened to the imputed callset, not to the reference panel.

1.2.6 Candidate analysis

For genome-wide significant trait-SV associations, we collected previous GWAS signals on the same chromosome with $P<10^{-7}$ from the EBI GWAS catalogue (see **Web Resources**) with the same set of keywords used in a previous study(Locke et al. 2019) (one publication based on METSIM samples was excluded to only include findings from independent studies). We then performed conditional analysis on the original trait-SV pairs adding the GWAS hits as covariates. Conditional analyses were restricted to samples with WGS data to minimize the difference in genotype accuracy of the SV callset vs. the SNP callset.

For loci containing multiple genotype-correlated SVs associated with a trait, we lumped the variants together using bedtools merge(Quinlan and Hall 2010) and reported the coordinates of the entire region with the summary statistics of the strongest signal. To better understand these loci, we manually curated the candidates in IGV(Thorvaldsdóttir, Robinson, and Mesirov 2013) and extended the regions of interest to include surrounding genes, functional elements, previous GWAS signals and other genome annotations. We then equally split each region into ~1000 windows and used CNVnator to calculate the copy number values of those windows for 100 individuals selected to represent all genotype groups. We then plotted the window-sample copy number matrix as a heatmap with scales best presenting the locus structure (e.g. **Figure 1.10**). In addition, for SNPs in the same region, we calculated the SNP-SV genotype correlation R^2 by a linear regression model and SNP-trait p values by EMMAX, then plotted them together in a local Manhattan plot (e.g. **Figure 1.7**) using custom R scripts.

For the fine-mapping experiment of albumin, we selected the top 100 most significant SNPs on chr4:67443182-79382541 plus the *ALB* promoter deletion to calculate the pairwise genotype correlation matrix and ran CAVIAR (v0.2)(Hormozdiari et al. 2014) on those 101 variants, with the “rho” probability set at 0.95 and varying the maximum number of causal variants one to five. The same experiment was done for total cholesterol. We used the model with maximum causal variants set at two to plot the posterior probability in **Figure 1.7**.

1.3 Results

We now turn to the results of this study starting with an overview of the SV callset, followed by trait association results including the in-depth discussion of individual genome-wide significant loci.

1.3.1 Structural variation detection and genotyping

We identified 120,793 SVs by LUMPY(Layer et al. 2014), 111,141 CNVs by GenomeSTRiP(Handsaker et al. 2015) (GS), and 92,862 CNVs by our customized pipeline based on CNVnator(Abyzov et al. 2011). Considering the different genotype metrics and detection resolutions, to retain sensitivity we chose to concatenate those three callsets together and adjust for redundancy later instead of merging the variants. 129,166 high-confidence autosomal SVs passed quality control, and 64,572 passed the frequency filter for association tests **(Table 1.2**). **Figure 1.3 and Figure 1.4** provide an overview of the high-confidence callset, including the size distribution, composition of biallelic vs. multiallelic SVs, and frequency distributions. The SV size and frequency distributions are consistent with those in previous studies(Sudmant et al. 2015; Chiang et al. 2017; Abel et al. 2020; Collins et al. 2020): most called SVs are relatively small (<10kb), biallelic and rare; called MEIs exhibit the expected size distribution corresponding to Alu and L1 insertions; and allele frequency decreases with increased mean SV size, consistent with negative selection against large SVs (**Figure 1.3, Figure 1.4**).

Based on comparison with a set of SNP array intensity data (see **Methods**), we estimate an overall false discovery rate (FDR) of 4.7% for the high-confidence callset. As an indicator of true positive rate, the proportion of SV calls tagged by nearby SNPs (R^2>=0.5, see **Methods**) was 56.8%, consistent with our prior GTEx study that used similar methods(Chiang et al. 2017) and was evaluated extensively in the context of eQTL mapping. We also compared our callset to the high-quality SV callsets from 1000 Genomes (1KG) and gnomAD projects and found an overlap of 35.2%, which is reasonable considering that these studies used distinct methods and sample sets. **Table 1.5** shows the above metrics stratified by pipelines. We estimated the genotype redundancy in total and stratified by pipelines (**Table 1.3**). Overall, the "effective sample size" of independent genetic variables was 55.5% of the original variant count. Additionally, since read-depth detection methods commonly result in "fragmented" CNV calls, we estimated the fragmentation level of calls by clustering variants within 10bp and measured the size of the clusters (**Table 1.4**).

Our CNVnator pipeline was the major source of redundancy and fragmentation since it detects CNVs with higher resolution – as small as 100bp – and covers repetitive and low-complexity regions, where the coverage profile is in general much noisier than the rest of the genome. The benefit is that CNVnator detected many true CNVs missed by the two other methods. As a benchmark of the sensitivity gain, we calculated the external validation rates for SVs uniquely detected in each of our pipelines (**Figure 1.5**). 7,210 variants identified only in CNVnator overlapped with variants in 1KG and gnomAD, contributing to the 43.1% of the overall CNVnator SVs that were validated through comparison to external datasets.\

1.3.2 Association of SVs with cardiometabolic traits

We first performed single variant association tests for 64,572 high-confidence SVs (MAC≥10) and 116 quantitative traits using the EMMAX model (Kang et al. 2010) in the 4,030 individuals with WGS data. We defined the genome-wide significance threshold as 1.89x10^{-6} and the experiment-wide significance threshold as 3.32x10^{-8} (see **Methods)**. Nine associations of six loci passed genome-wide significance threshold; six were still significant after adjusting for the equivalent number of independent phenotypes (**Table 1.6,** WGS P).

We next sought to replicate these findings and to follow up on 4,855 loci with sub-threshold associations (p<0.001) via meta-analysis with larger WES (n=20,316) and array genotype datasets (n=19,033) from these same cohorts, using independent samples (n_{WES}=15,205, n_{array}=15,125) not included in the original WGS experiment (see **Methods**)(Locke et al. 2019). We developed a strategy to genotype coding CNVs from WES data using read-depth information from XHMM(Fromer and Purcell 2014), and measured copy number at the 20,058 exons intersecting with 819 candidate CNVs from WGS. We found that 281 exons from 392 CNV calls were able to recapture the copy number variability detected by WGS (at R^2>0.1). To genotype SVs using array data, we used standard imputation methods to impute 2,127 bi-allelic SVs based on the background of array-genotyped SNPs (see **Methods**). The estimated imputation accuracy of SVs corresponded well to their LD with nearby SNPs, as expected (**Figure 1.6**). To assess performance more rigorously for the eight significant SVs described below, we also performed a leave-one-out experiment, and the validation rate ranged from 93.3%-99.8% (**Table 1.7**). Overall, we were able to accurately genotype 2,053 of 4,864 candidate SVs using exome (n=392) and/or array genotype data (n=1,705). We then ran single-variant tests on those genotyped SVs with the corresponding candidate traits in the independent samples, and performed a meta-analysis to calculate a combined p-value (**Table 1.6**).

After merging fragmented SVs, we ended up with 15 independent loci associated with 31 traits at genome-wide significance, 9 of which remained significant after correction for the multiple phenotypes. **Table 1.6** shows the summary statistics of the lead SVs for their top traits (table for all the pre-merged summary statistics was too big to include in this thesis while could be found in the published paper).

1.3.3 Deletion of the *ALB* gene promoter is associated with multiple traits

The strongest signal in the combined study was a 4kb deletion immediately upstream of the *ALB* gene, affecting the promoter region (**Figure 1.7**). This variant was 16-fold enriched in the Finnish population compared to non-Finnish Europeans from 1KG (MAF: 1.6% vs. 0.1%) and was associated with 16 traits at genome-wide significance (**Figure 1.8**). The top two associations were with serum albumin ($p=1.47x10^{-54}$, beta=0.91) and total cholesterol ($p=1.22x10^{-28}$, beta=-0.49), and these are independent signals based on conditional analyses (**Table 1.10**). The cholesterol signal appears to explain the remaining 14 trait associations, all of which are highly correlated (**Figure 1.8**). This SV was well-tagged by nearby SNPs ($R^2=0.73$), and the tagging SNPs showed similar trait association patterns. To tease apart potentially indirect associations caused by LD, we performed fine-mapping analysis for serum albumin and total cholesterol with CAVIAR(Hormozdiari et al. 2014) including the deletion variant and the 100 most significant SNPs on chr4:67-79Mb (see **Methods**). The top candidate for the association with total cholesterol was a SNP (rs182695896) in moderate LD ($R^2=0.49$) with the deletion. Accounting for this SNP via conditional analysis attenuated the association between the deletion and total cholesterol (p=0.023, n=4014). The deletion was identified as the most probable causal variant for the association with albumin, and the association between the deletion and albumin remained significant after adjusting for rs182695896 ($p=6.52x10^{-13}$, n=3,117). We also observed different causality patterns for the two traits by aligning the posterior probabilities with the LD structure of the causal candidates in 95% confidence sets (**Figure 1.7**). Thus, we hypothesize that the promoter deletion directly affects serum albumin by altering *ALB* gene expression, and is associated with total cholesterol through its genetic correlation with other underlying causal variant(s) in the same LD block.

Prior studies(Inouye et al. 2012; Kettunen et al. 2012, 2016; Surakka et al. 2015) have reported five albumin associated SNPs and two cholesterol associated SNPs in this region. In our conditional analyses including all intrachromosomal GWAS hits(Buniello et al. 2019), the SV-albumin association remained genome-wide significant (**Table 1.6**) while the SV-cholesterol association was diminished (conditioned p=0.004). To investigate the relationship between our signal and each of the seven previous GWAS SNPs, we tested the SV for association while conditioning on the reported SNPs one at a time (**Table 1.8**) and ran the association tests on those SNPs with the SV as covariate (**Table 1.9**). These results suggest that the *ALB* deletion is the causal variant for three prior albumin associations (rs16850360, rs2168889, and rs1851024), is linked to one previously reported cholesterol association (rs182616603), and is independent of two prior albumin associations (rs115136538, rs184650103) and one cholesterol association (rs117087731).

We next explored the potential downstream effects of this promoter deletion in the FinnGen dataset (see Web Resources), which reports GWAS results for 1,801 disease endpoints in 135,638 individuals. We queried the top SV-tagging SNP (rs187918276, R^2=0.73) in the PheWeb browser (**Figure 1.9,** Web Resources); the top association was with statin medication use (p=6.5x10^{-69}). The second set of signals appeared in the "Endocrine, nutritional and metabolic diseases" category, led by disorders of lipoprotein metabolism and other lipidemias (p=1.4x10^{-11}), pure hypercholesterolemia (p=3.0x10^{-11}), and metabolic disorders (p=1.8x10^{-7}). These results support the medical relevance of genetic variation at this locus suggested by this and prior work; however, it is unclear whether these results are due to the *ALB* promoter deletion or the linked variants (e.g., rs182695896) associated with cholesterol.

1.3.4 A multi-allelic CNV at *PDPR* is associated with pyruvate and alanine levels

We identified a cluster of 13 highly correlated CNV calls at chr16q22.1 that were strongly associated with pyruvate ($p=4.81x10^{-21}$, beta=-0.72) and alanine ($p=6.14x10^{-12}$, beta=-0.53) levels in the serum. We reconstructed the copy number profile of this locus from short-read WGS data (see Methods) and confirmed that the 13 correlated variant calls correspond to a single ~250kb multiallelic CNV (CNV1 in **Figure 1.10**) spanning the coding sequence and 5' region of *PDPR*, a gene involved in the pyruvate metabolism pathway. *PDPR* encodes the regulatory subunit of pyruvate dehydrogenase phosphatase (PDP) which catalyzes the dephosphorylation and reactivation of pyruvate dehydrogenase complex, the catalyst of pyruvate decarboxylation. According to this mechanism, fewer copies of *PDPR* should slow down the decarboxylation reaction and lead to increased pyruvate levels, and increased copies should decrease pyruvate levels, consistent with our data (**Figure 1.10**). This CNV was also negatively associated with alanine levels, the product of pyruvate transamination, and conditional analysis suggested this association was mediated through pyruvate (**Table 1.10**).

An intriguing aspect of the *PDPR* locus is that it contains numerous segmental duplications (SDs), including highly similar local SDs scattered throughout the *PDPR* locus, additional SDs at a *PDPR* pseudogene (*LOC283922)* located 4 Mb distal to *PDPR*, as well as more divergent copies located ~55Mb away on chr16p13.11. These include LCR16a, a core element shared by many SDs on Chr16 and a well-known driver of the formation of complex segmental duplication blocks in the genomes of humans and primates(Jiang et al. 2007; Johnson et al. 2006; Cantsilieris et al. 2020). There are both duplication and deletion alleles of the *PDPR* gene, and these have indistinguishable breakpoints that correspond to LCR16a duplicons, suggesting these CNVs were caused by recurrent non-allelic homologous recombination. Similar to the *ALB* deletion described above (and many prior coding associations(Locke et al. 2019)), this CNV appears to be enriched in the Finnish population: the duplication allele was identified in 1KG with a frequency of 0.005 in non-Finnish Europeans, 50x less than the 0.025 frequency observed in our Finnish sample , and the deletion allele was not detected in 1KG. The CNV is poorly tagged by flanking SNPs (max $R^2<0.088$), making it virtually undetectable using standard GWAS methods.

In addition, a second highly polymorphic and multiallelic CNV (CNV2 in **Figure 1.10**) intersects with CNV1 and covers >90% of the gene body of *PDPR,* missing the first three exons. Notably, CNV2 did not show association with pyruvate levels in our data (p=0.6), despite being previously reported as a *cis*-eQTL for *PDPR* in multiple tissues(Chiang et al. 2017). To resolve the structure of this locus, we aligned chromosome 16 of the GRCh38 reference against itself and also against the recent high-quality CHM13 assembly(Miga et al. 2020) created from long-read sequencing data (**Figure 1.11**). Interestingly, we found that the sequence of CNV2 contains three inverted paralogs of the *LOC283922* locus (a *PDPR* pseudogene) in the CHM13 assembly, while there is only one copy of *LOC283922* in GRCh38 (**Figure 1.10**). These data suggest that CNV2 reflects highly variable structural alleles of *LOC283922* located 4Mb away from *PDPR*, and thus it is not surprising that this CNV does not affect pyruvate levels.

1.3.5 Additional trait-association signals

We confirmed a previously reported association between the recurrent HP deletion and decreased total serum cholesterol levels16. In our data, this same deletion was strongly associated with serum glycoprotein acetyls quantified by NMR (p=3.53x10-35), and conditional analysis showed that the two associations were independent (**Table 1.10**). Since Boettger et al.16 proposed a plausible mechanism for the association of HP copy number and cholesterol, here we focus on the glycoprotein association. As a serum glycoprotein, haptoglobin forms dimers in individuals with the HP1/HP1 genotype (homozygous deletion) but forms multimers in individuals carrying HP2 allele(s). The multimers can be as large as 900kDa – more than twice the size of the dimers (86kDa)55 – which could result in fewer haptoglobin molecules in HP2 carriers, and consequently fewer glycoprotein molecules overall.

We identified five trait associations involving common SVs that were within 1Mb of previously published GWAS loci for the same traits. All SVs were well-tagged by SNPs (R2>0.9) and were either intronic or upstream of genes that are functionally related to the associated phenotypes. In all five cases there were stronger SNP signals nearby, and the SV associations dropped to not more than nominal significance when conditioned on the known GWAS SNPs (**Table 1.6**). This suggests that instead of having independent effects on the phenotypes, those SVs were more likely to be in LD with the causal variants.

Additionally, we identified a low-frequency (MAF=0.01) SV associated with serum tyrosine levels (combined p=4.17x10-10). This variant was a 4kb deletion of IL34, affecting the first exon of one transcript isoform and the intronic region of the two longer isoforms. There is a stronger signal from a SNP (rs190782607, p=1.44x10-11) within 100kb of and partially tagging the SV (R2=0.61), indicating that the SV is unlikely to be the causal variant. However, the p-value of this association remained at a similar level when conditioned on known GWAS SNPs50 (**Table 1.6**), suggesting a novel signal. IL34 mediates the differentiation of monocytes and macrophages and to our knowledge has not previously been reported to be associated with amino acid traits56.

The re-discovery of known loci described above demonstrates the effectiveness of our study design. Our CNV detection pipeline also detected two associations with metabolic traits that appear to be related to blood cell-type composition rather than inherited genetic variation.

We identified three clusters of CNVs on chr7q34, chr7p14 and chr14q11.2 associated with C-reactive Protein (CRP) levels in the plasma, a biomarker for inflammation and a risk factor for heart disease (**Table 1.6**). These CNVs are large, involve subtle alterations in copy number, and correspond to T cell receptor loci, suggesting that they are likely to reflect somatic deletions due to V(D)J recombination events during T cell maturation. This hypothesis was supported by the read-depth coverage pattern (see Figure 1.12), where the measured copy number is lowest at the recombination signal sequence (RSS) used constitutively for rearrangement, and gradually increases with increasing distance to the RSS. The cause of this association is unclear but may reflect increased T-cell abundance and CRP levels due to active immune response in a subset of individuals.

Interestingly, we also indirectly measured mitochondrial (MT) genome copy number variation due to the mis-mapping of reads from mitochondrial DNA to ancient nuclear MT genome insertions (NUMT)57 on chromosomes 1 and 17, that show strong homology to segments of the MT genome. These apparent “CNVs”, which reflect MT abundance in leukocytes, were strongly associated with fasting insulin levels (p=1.00x10-10) and related traits, and are the topic of a separate study58.

We also discovered three association signals corresponding to dense clusters of fragmented CNV calls within highly repetitive and low-complexity regions including simple repeats and segmental duplications (**Table 1.6**). Interpreting patterns of variation and trait association at these loci remains challenging due to their complex and repetitive genomic architecture, and known alignment artifacts within such regions. Although we were not able to identify any technical artifacts that might explain these specific associations, they should be interpreted with caution. Further investigation of these highly repetitive loci will require improved sequencing and variant detection methods.

1.4 Discussion

We have conducted what is to our knowledge the first complex trait association study based on direct ascertainment of SV from deep WGS data. Our study leverages sensitive SV detection methods, extensive cardiometabolic quantitative trait measurements, and the unique population history of Finland. Despite the relatively modest sample size and limited power of this study, we identified 9 novel and 6 known trait associated loci. Most notably, we identified two novel loci where SVs are the likely causal variants and have strong effects on disease-relevant traits. Both SVs are ultra-rare in non-Finnish Europeans but present at elevated allele frequency in Finns – presumably due to historical population bottlenecks and expansions – which mirrors the findings from our recent study of coding variation, where many cardiometabolic trait-associated variants were enriched in Finns26. The first, a deletion of the ALB promoter, strongly decreased serum albumin levels in carriers (~1 standard deviation per copy), and also resides on a haplotype associated with cholesterol levels. This example shows that non-coding SVs can have extremely large effects, consistent with our prior results based on eQTLs22 and selective constraint30, and points to the importance of including diverse variant classes in trait association efforts . Although more work is required to understand the disease relevance of this deletion variant, we note that low levels of albumin can cause analbuminemia, which is associated with mild edema, hypotension, fatigue, lower body lipodystrophy, and hyperlipidemia.

The second, a multi-allelic CNV with both duplication and deletion alleles that affect PDPR gene dosage, has strong effects on pyruvate and alanine levels. Notably, this CNV is the product of recurrent NAHR between flanking repeats at a complex locus that has accumulated numerous segmental duplications over evolutionary time, and is not well-tagged by SNVs. This phenomenon – recurrent CNVs at segmentally duplicated loci – has been studied extensively in the context of human genomic disorders and primate genome evolution, but there are few examples for complex traits. This result underscores the importance of comprehensive variant ascertainment in WGS-based studies of common disease and other complex traits. We further note that it is unusual to observe multiallelic CNVs at a conserved metabolic gene such as PDPR; it is tempting to speculate about the role of such variation in human evolution.

Interestingly, our study also identified two novel and highly atypical trait associations that appear to be caused by variable cell type composition in the peripheral blood. Identifying these results was only possible due to our use of WGS on blood-derived DNA, combined with sensitive SV analysis methods capable of detecting sub-clonal DNA copy number differences. Our quantitative detection of subclonal T-cell receptor locus deletions formed by V(D)J recombination served as a proxy for measuring T cell abundance, and led to the novel result that CRP levels are associated with T cell abundance. We hypothesize that this association is caused by active immune response in a subset of individuals. Similarly, our quantitative detection of mitochondrial genome copy number via apparent "CNVs" at NUMT sites in the nuclear genome led to the novel and important finding that variable abundance of neutrophils vs. platelets in peripheral blood is strongly associated with insulin, fat mass, and related metabolic traits (as described in detail elsewhere58).

Taken together, these results highlight the potential role of rare, large-effect SVs in the genetics of cardiometabolic traits, and suggest that future comprehensive and well-powered WGS-based studies have the potential to contribute greatly to our understanding of common disease genetics.

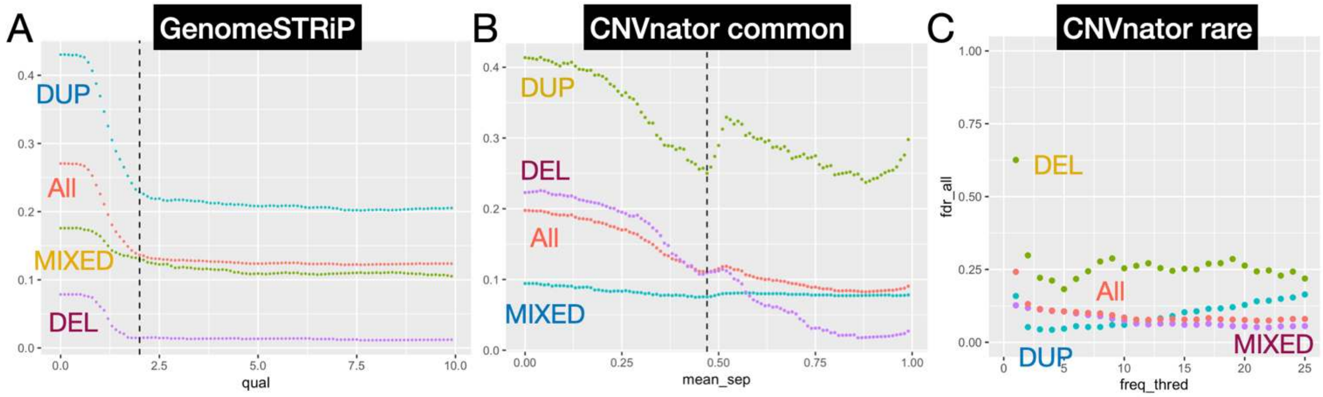


Figure 1.1. **Deciding QC filters based on FDR curves**

FDR curves under different quality thresholds for (A) GenomeSTRiP CNVs, (B) Common variants of CNVnator CNVs, and (C) Rare variants of CNVnator CNVs. The FDR was estimated from the array intensity data of METSIM samples using IntensityRankSumAnnotator from the GenomeSTRiP pipeline, among CNVs covered by at least two probes. GenomeSTRiP CNVs were filtered based on the "GSCNQUAL" score output by the software, common CNVnator CNVs were filtered by the "mean_sep" metrics from the constrained GMM model, and the rare CNVnator CNVs were filtered by carrier frequency. The results are presented for all variants as well as by different variant types, indicated by the colors shown.

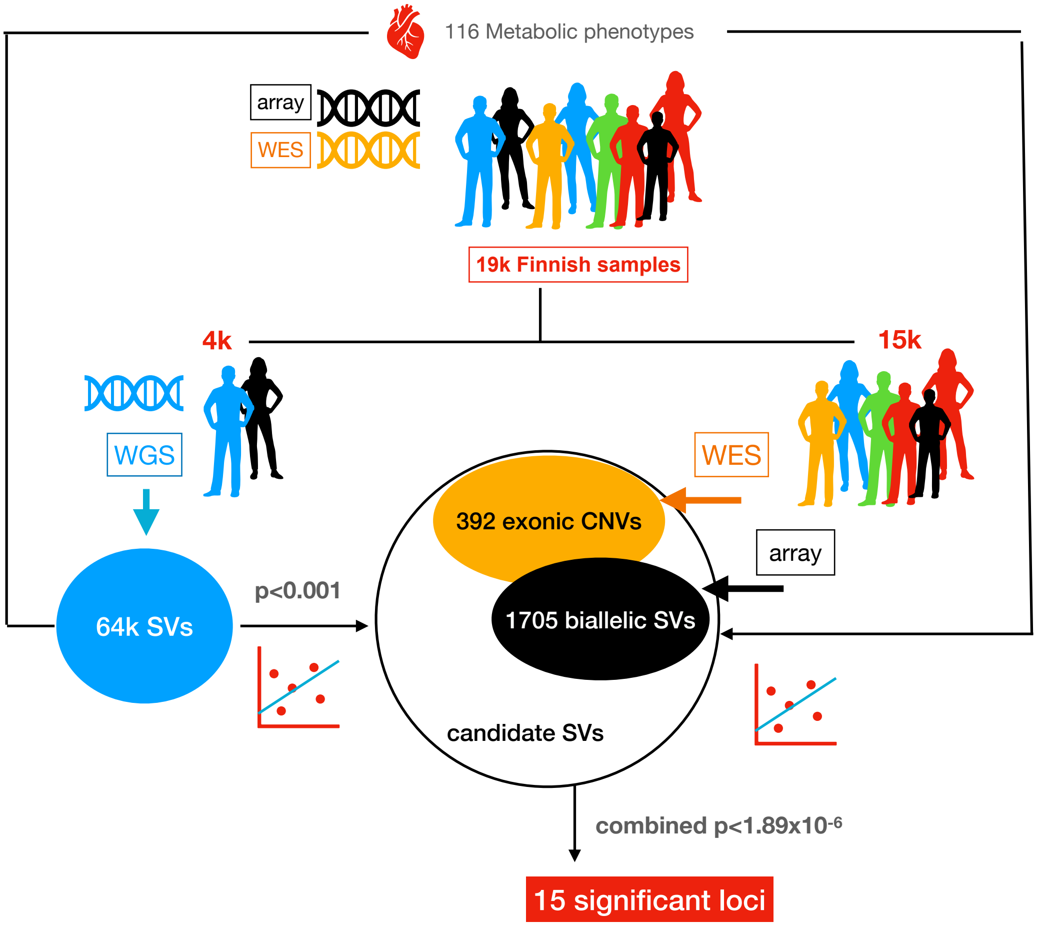


Figure 1.2. Flowchart of the overall experimental design.

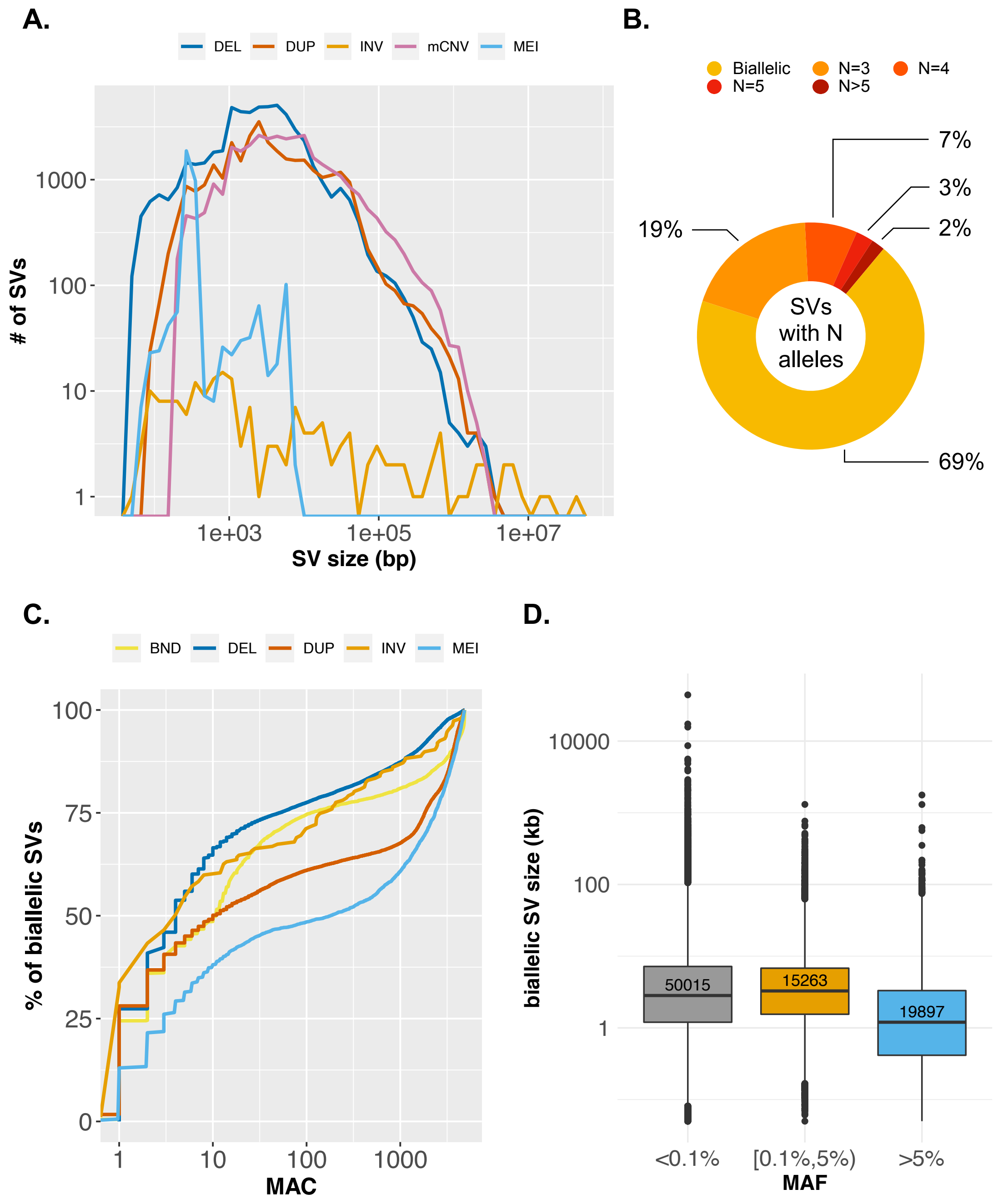


Figure 1.3. Overview of the high-confidence SV callset

(A) SV size distribution (log10 scale, bp) by variant type. BNDs are not included due to the ambiguous definition of variant boundaries. **(B)** Proportion of bi-allelic SVs and multi-allelic CNVs, where N is defined by the number of copy number groups (e.g. CN=0,1,2,3,4, etc.). **(C)**

The minor allele count distribution of all the high-confidence bi-allelic SVs stratified by variant type. **(D)** the size distribution (log10 scale) of biallelic SVs stratified by MAF groups (<0.1% - ultra-rare; 0.1%-5% - rare, >5% - common). The central line and box borders represent median, 1st and 3rd quartiles. The upper whiskers extend to the lesser extreme of the maximum and the 3rd quartile plus 1.5 times the interquartile range (IQR); the lower whiskers extend to the lesser extreme of the minimum and the 1st quartile minus 1.5 times the IQR.

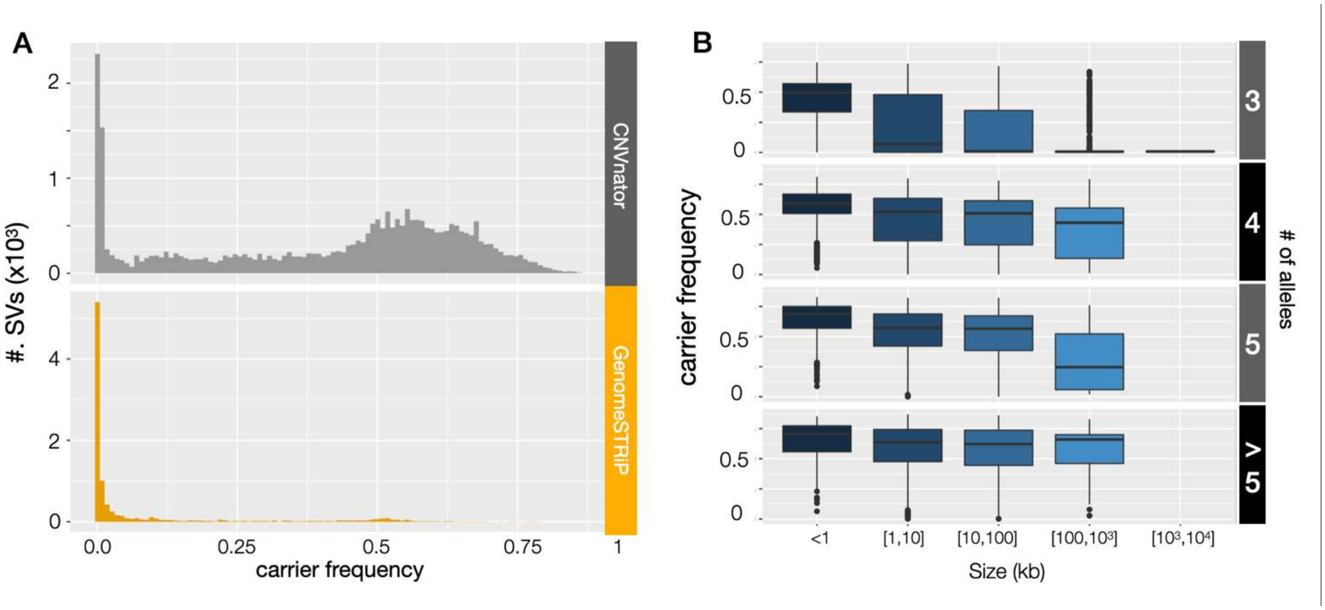
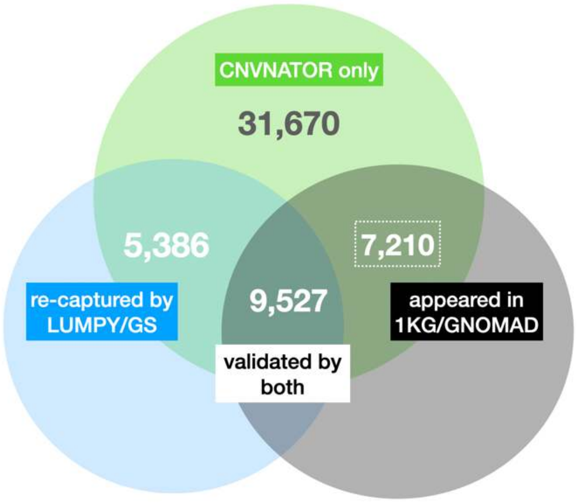
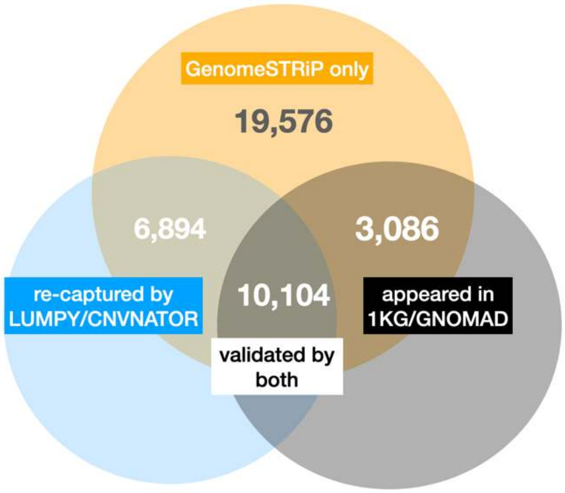


Figure 1.4. The frequency distribution of multi-allelic CNVs

(A) The carrier frequency spectrum of multi-allelic CNVs, stratified by detection methods. Note that the concentration of CNVnator variants between 0.5-0.75 were primarily caused by large segmental duplication regions near centromeres and telomeres, where the variant boundaries were challenging to define and the CNVs were detected in highly fragmented form. Such regions are often excluded from genetic analysis but were included here to maximize sensitivity. **(B)** Similar frequency distribution to (A), stratified by mCNV size groups. The central line and box borders represent median, 1st and 3rd quartiles. The upper whiskers extend to the lesser extreme of the maximum and the 3rd quartile plus 1.5 times the interquartile range (IQR); the lower whiskers extend to the lesser extreme of the minimum and the 1st quartile minus 1.5 times the IQR.


CNVNATOR only
31,670
5,386
7,210
re-captured by LUMPY/GS
9,527
appeared in 1KG/GNOMAD
validated by both


GenomeSTRiP only
19,576
6,894
3,086
re-captured by LUMPY/CNVNATOR
10,104
appeared in 1KG/GNOMAD
validated by both

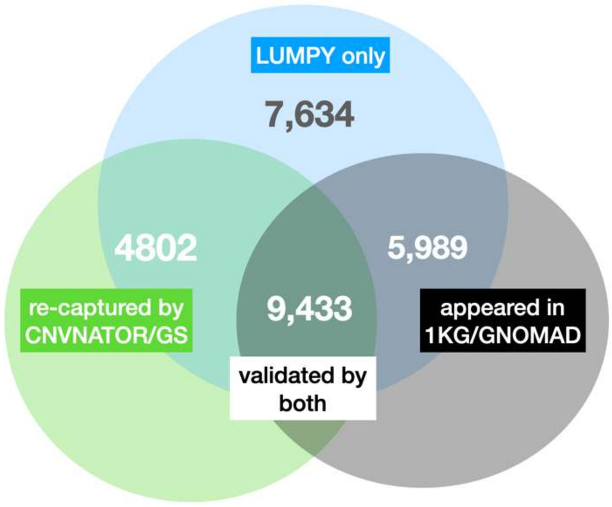

LUMPY only
7,634
4802
5,989
re-captured by CNVNATOR/GS
9,433
appeared in 1KG/GNOMAD
validated by both

Figure 1.5. Overlapping SVs among internal and external callsets

For each of the three SV detection methods used in this study, these venn diagrams show the number of CNVs that were also identified by the other two “internal” pipelines used in this study (left), and the “external” reference SV callsets from 1KG and gnomAD (right). The upper part of each diagram also shows the number of CNVs only identified by a given pipeline. Dashed rectangles were used to emphasize the number of CNVnator CNVs that were validated by external callsets but missed by the other two pipelines, showing the complementary nature of the methods used for this study. 50% reciprocal overlap was used to compare CNV calls from different callsets.

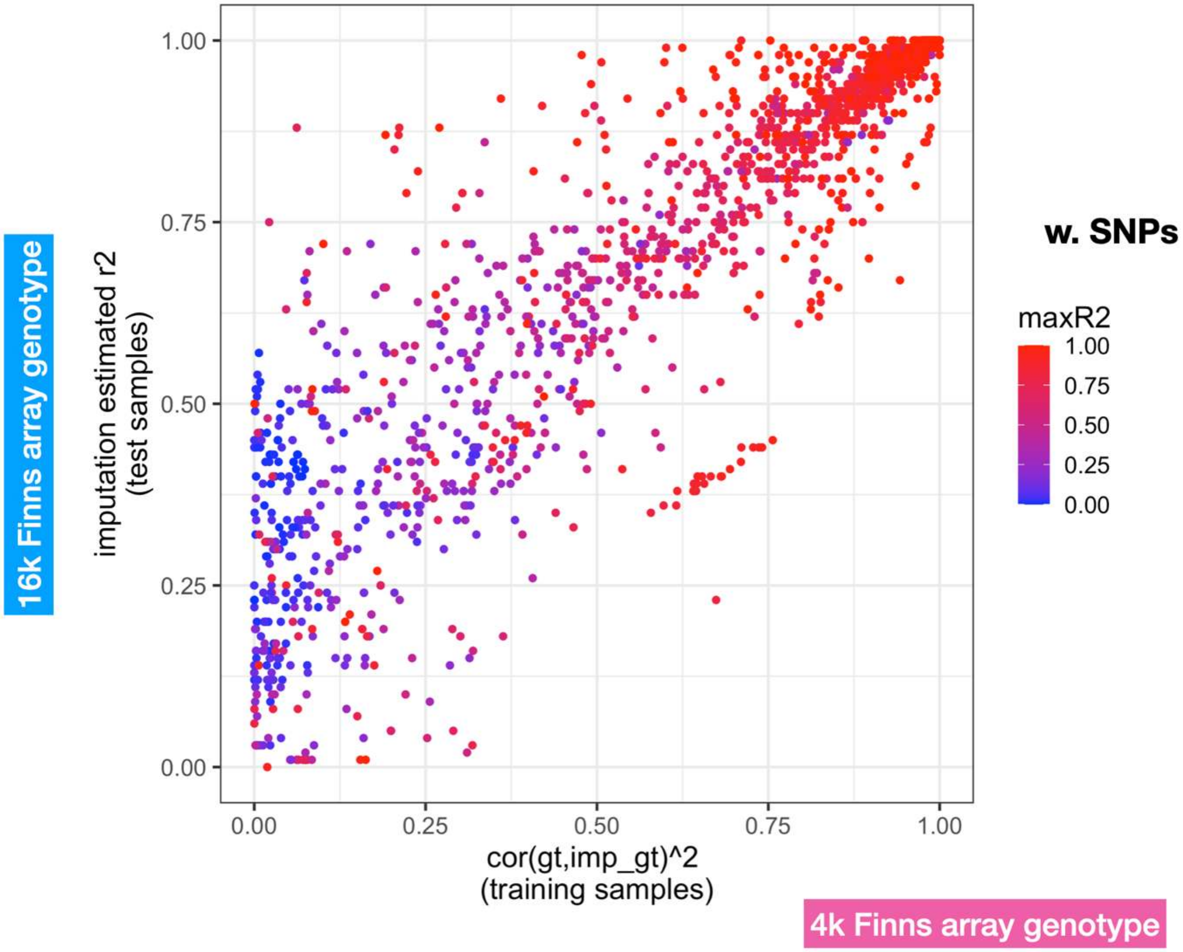


Figure 1.6. The overall evaluation of imputation quality with two metrics

The Y-axis shows the Beagle output quality score (DR2) for the ~15k tested samples, which is the estimated correlation between the imputed genotype and real genotype for each variant, and the X-axis shows the "training error" for the ~4k samples with WGS data. Training error was calculated using the WGS data as reference and array data as test input, after which the correlation of real genotype (based on WGS) and predicted genotype was calculated. The color shows how well each SV was tagged by nearby SNPs located within 1 Mb.

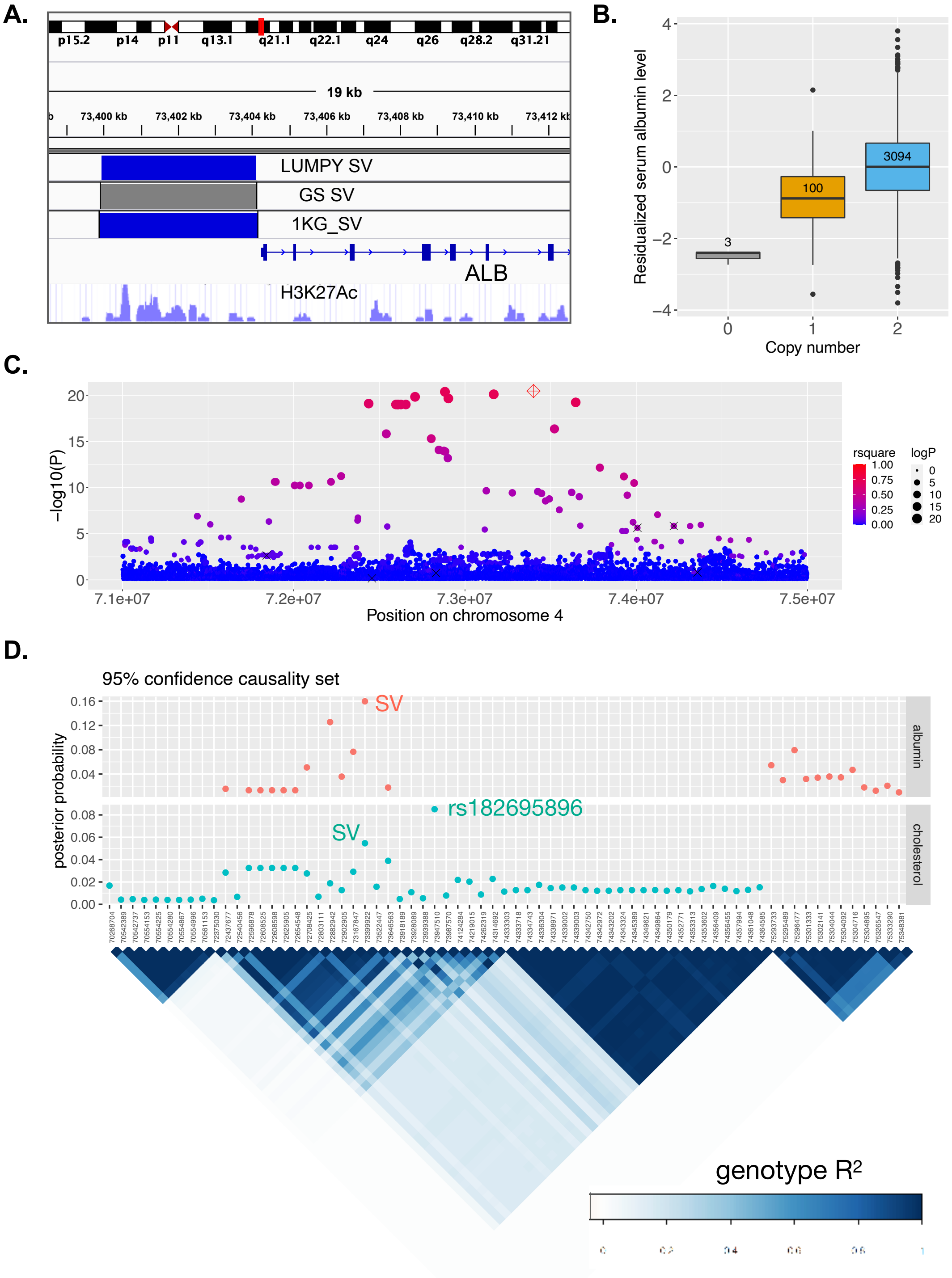

A.
p15.2
p14
p11
q13.1
q21.1
q22.1
q24
q26
q28.2
q31.21
19 kb
73,400 kb
73,402 kb
73,404 kb
73,406 kb
73,408 kb
73,410 kb
73,412 kb
LUMPY SV
GS SV
1KG_SV
ALB
H3K27Ac
B.
Residualized serum albumin level
4
2
0
−2
−4
3
100
3094
0
1
2
Copy number
C.
−log10(P)
20
15
10
5
0
7.1e+07
7.2e+07
7.3e+07
7.4e+07
7.5e+07
Position on chromosome 4
rsquare
1.00
0.75
0.50
0.25
0.00
logP
0
5
10
15
20
D.
95% confidence causality set
posterior probability
0.16
0.12
0.08
0.04
SV
albumin
0.08
0.06
0.04
0.02
0.00
rs182695896
SV
cholesterol
70268704
70542389
70542737
70554153
70554225
70554280
70554867
70554996
70561153
72375030
72437677
72540456
72596878
72608525
72608598
72625905
72654548
72708425
72803111
72882942
72902905
73167847
73399922
73522447
73646563
73918189
73928089
73939388
73947510
73987570
74124284
74219015
74262319
74314692
74333303
74333718
74334743
74336304
74338971
74339002
74339003
74342750
74342972
74343202
74343324
74345389
74349621
74349864
74350179
74352771
74353313
74353602
74356409
74356455
74357994
74361048
74364585
75293733
75295489
75296477
75301333
75302141
75304044
75304092
75304716
75304895
75326547
75333290
75348381
genotype R²
0
0.2
0.4
0.6
0.8
1

Figure 1.7. The *ALB* promotor deletion associated with serum albumin level and cholesterol traits

(A) The genomic location of the chr4 deletion, with coordinates detected from LUMPY, GenomeSTRiP and 1KG. The H3K27Ac track is from the ENCODE (ENCODE Project Consortium 2004) data obtained from the UCSC genome browser (showing the data of K562 cells). **(B)** Boxplot showing serum albumin levels stratified by genotype, with the sample size of each genotype group annotated at the center of each box. The trait value on the y-axis is the inverse normalized residual of raw measurement (residualized for age, age^2, and sex). **(C)** Local Manhattan plot of albumin association signals on chr4:71-75Mb, including the *ALB* deletion (red diamond) and SNPs with minimum allele count of 9 (filled circles). The sizes of the circles are proportional to -log10(p) and colors indicated LD (Pearson R^2) with the deletion (NA shown in grey). Six of the seven previously published GWAS signals are indicated with 'x' (the seventh was too rare in our data to be included in the test). **(D)** Fine-mapping results at the *ALB* locus for albumin and total cholesterol trait associations, using CAVIAR. The top panel shows the 95% confidence causality sets for albumin (top) and cholesterol (bottom) and posterior probability of each variant to be causal (assuming a maximum of two causal variants). The bottom panel shows the LD structure for the candidate variants, using the genotype correlation (Pearson R^2) calculated from WGS data.

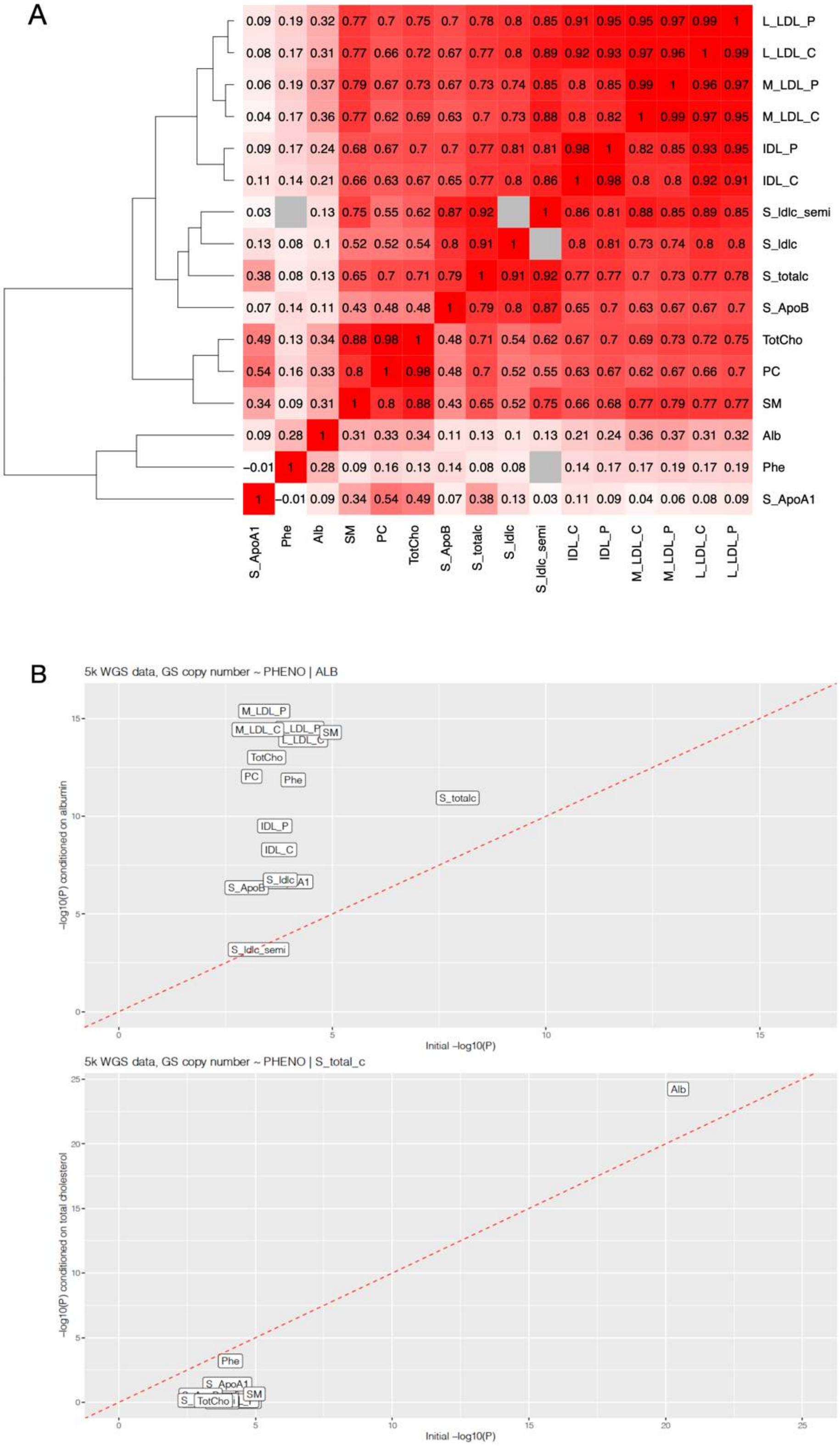
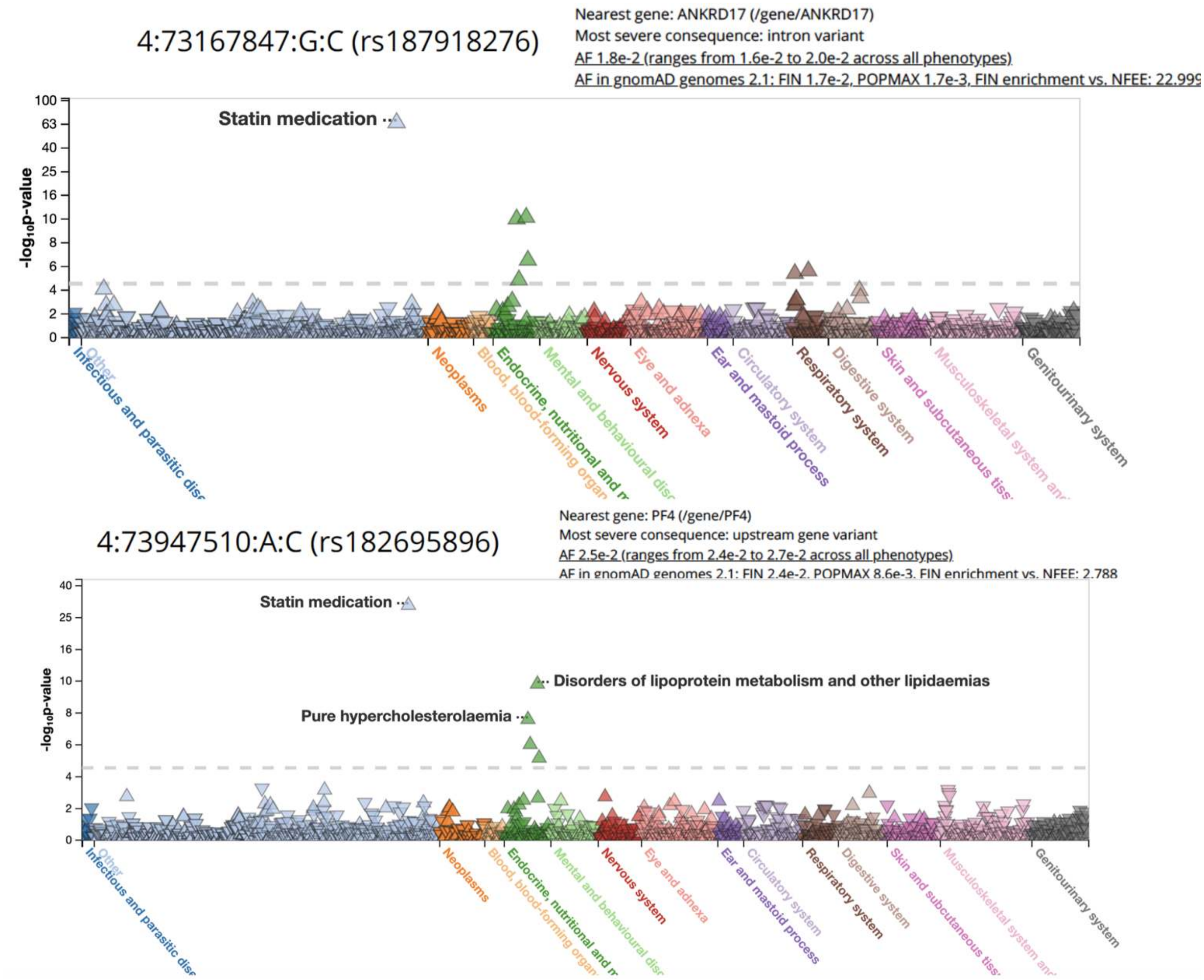


Figure 1.8. The overview of the 16 trait-association signals of *ALB* deletion

(A) The pairwise correlation (Pearson R) of the 16 traits that were significantly associated with *ALB* deletion. The cells shown in gray represent missing data, since the S_ldlc_semi trait (serum LDL cholesterol in semi-fasting samples) shared zero samples with S_ldlc (serum LDL cholesterol in fasting samples) and Phe (phenylalanine). **(B)** Comparison of the association p-value of the *ALB* deletion and the 16 traits, with (y-axis) and without (x-axis) albumin (top) and total cholesterol (bottom) as a covariate. The increases of significant level of most traits when conditioned on albumin were likely due to Berkson's paradox(Berkson 1946).



4:73167847:G:C (rs187918276)

Nearest gene: ANKRD17 (/gene/ANKRD17)
Most severe consequence: intron variant
AF 1.8e-2 (ranges from 1.6e-2 to 2.0e-2 across all phenotypes)
AF in gnomAD genomes 2.1: FIN 1.7e-2, POPMAX 1.7e-3, FIN enrichment vs. NFEE: 22.99

4:73947510:A:C (rs182695896)

Nearest gene: PF4 (/gene/PF4)
Most severe consequence: upstream gene variant
AF 2.5e-2 (ranges from 2.4e-2 to 2.7e-2 across all phenotypes)
AF in gnomAD genomes 2.1: FIN 2.4e-2, POPMAX 8.6e-3, FIN enrichment vs. NFEE: 2.788

-log10p-value
40
25
16
10
8
6
4
2
0
Statin medication
Disorders of lipoprotein metabolism and other lipidaemias
Pure hypercholesterolaemia
Other
Infectious and parasitic dise
Neoplasms
Blood, blood-forming organ
Endocrine, nutritional and m
Mental and behavioural disc
Nervous system
Eye and adnexa
Ear and mastoid process
Circulatory system
Respiratory system
Digestive system
Skin and subcutaneous tiss
Musculoskeletal system an
Genitourinary system

Figure 1.9. Potential disease endpoints of the *ALB* deletion in FinnGen dataset

Screenshots from the FinnGen PheWeb browser("PheWeb" n.d.) (Data Freeze 3) of the top tagging SNP for the *ALB* deletion (top) and for the cholesterol candidate (bottom) predicted by fine mapping with CAVIAR, showing the phenome-wide association results for each of the SNPs, colored by phenotype groups.

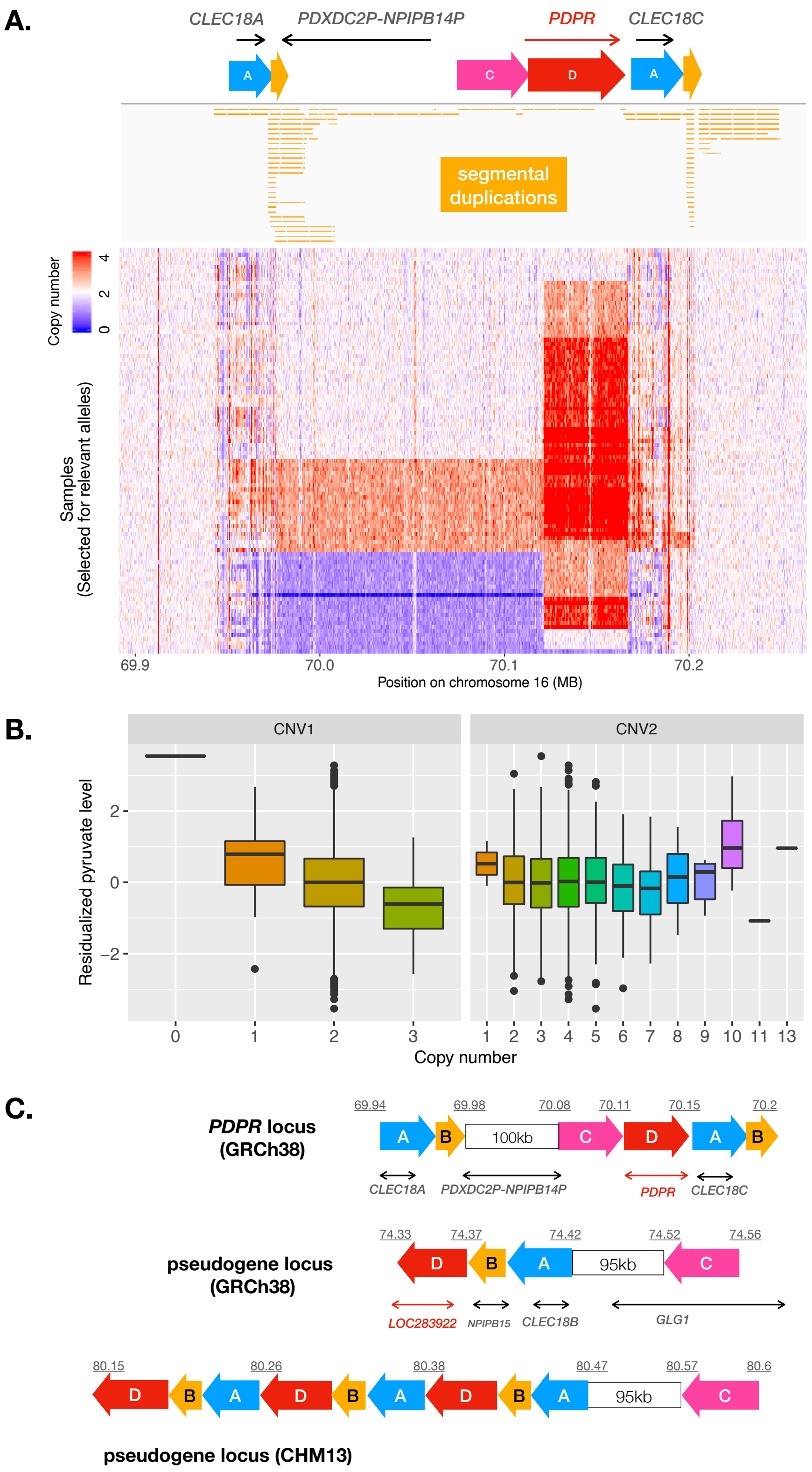

A.
CLEC18A
PDXDC2P-NPIPB14P
PDPR
CLEC18C
A
C
D
A
segmental duplications
Copy number
4
2
0
Samples
(Selected for relevant alleles)
69.9
70.0
70.1
70.2
Position on chromosome 16 (MB)
B.
CNV1
CNV2
Residualized pyruvate level
2
0
−2
0
1
2
3
1
2
3
4
5
6
7
8
9
10
11
13
Copy number
C.
PDPR locus
(GRCh38)
69.94
69.98
70.08
70.11
70.15
70.2
A
B
100kb
C
D
A
B
CLEC18A
PDXDC2P-NPIPB14P
PDPR
CLEC18C
pseudogene locus
(GRCh38)
74.33
74.37
74.42
74.52
74.56
D
B
A
95kb
C
LOC283922
NPIPB15
CLEC18B
GLG1
80.15
80.26
80.38
80.47
80.57
80.6
D
B
A
D
B
A
D
B
A
95kb
C
pseudogene locus (CHM13)

Figure 1.10. **The multi-allelic CNV at the *PDPR* locus affecting pyruvate and alanine.**

(A) The *PDPR* locus showing (from top to bottom) genes, duplicated genomic segments based on dotplot analysis (see **Figure 1.11**), segmental duplication annotations from the UCSC table browser(Karolchik et al. 2003), and copy number profiles for 100 samples comprising 51 carriers and 49 non-carriers for CNV1. Copy number is shown in 500bp windows, as determined by CNVnator, and the color saturates at four copies. The two horizontal lines indicate locations of the two CNVs (solid-CNV1, dashed-CNV2). **(B)** Pyruvate levels for 3,121 WGS samples stratified by copy number genotypes of CNV1 ($p=9.41x10^{-11}$) and CNV2 (p=0.6). **(C)** Structure of GRCh38 reference and CHM13 assembly at the *PDPR* locus (top) and its pseudogene locus (bottom two), using the same annotations as in part (A). Blocks with the same color and letter notation are highly similar DNA sequences and arrows show the direction of alignments. Diagrams were drawn based on the dot plots in **Figure 1.11**. The segment B corresponds to LCR16a, the core element shared by many duplicons sparsely distributed on chromosome 16(Jiang et al. 2007).

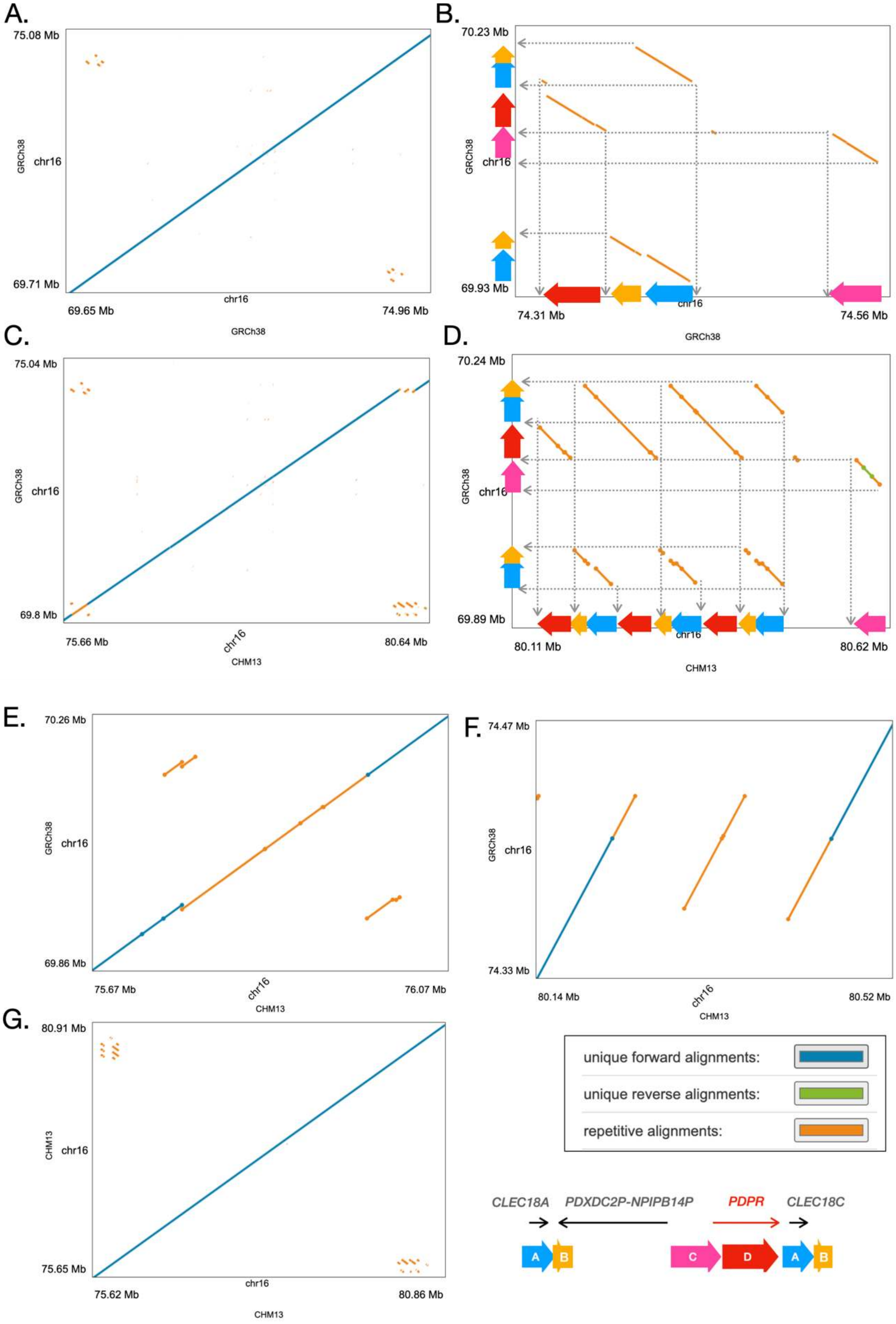

A.
75.08 Mb
GRCh38
chr16
69.71 Mb
69.65 Mb
chr16
74.96 Mb
GRCh38
B.
70.23 Mb
GRCh38
chr16
69.93 Mb
74.31 Mb
chr16
74.56 Mb
GRCh38
C.
75.04 Mb
GRCh38
chr16
69.8 Mb
75.66 Mb
chr16
80.64 Mb
CHM13
D.
70.24 Mb
GRCh38
chr16
69.89 Mb
80.11 Mb
chr16
80.62 Mb
CHM13
E.
70.26 Mb
GRCh38
chr16
69.86 Mb
75.67 Mb
chr16
76.07 Mb
CHM13
F.
74.47 Mb
GRCh38
chr16
74.33 Mb
80.14 Mb
chr16
80.52 Mb
CHM13
G.
80.91 Mb
CHM13
chr16
75.65 Mb
75.62 Mb
chr16
80.86 Mb
CHM13
unique forward alignments:
unique reverse alignments:
repetitive alignments:
CLEC18A
PDXDC2P-NPIPB14P
PDPR
CLEC18C
A
B
C
D
A
B

Figure 1.11. Aligning genome assemblies to solve the *PDPR* structure

Dot plots showing the structure of the *PDPR* and nearby pseudogene locus in both the GRCh38 and CHM13 assemblies, with repetitive alignments shown in orange and unique alignments shown in blue and green (see legend bottom right). **(A)** The *PDPR* locus in GRCh38 (y-axis) aligned to the pseudogene locus (x-axis) in GRCh38, where **(B)** shows a zoomed-in version with the diagram used for **Figure 4** using the same colors and letter. **(C)** and **(D)** show the *PDPR* locus in GRCh38 vs. the *PDPR* locus in CHM13. **(E)** and **(F)** show the pseudogene locus in GRCh38 vs. CHM13, and **(g)** shows the *PDPR* locus in CHM13 vs. itself.

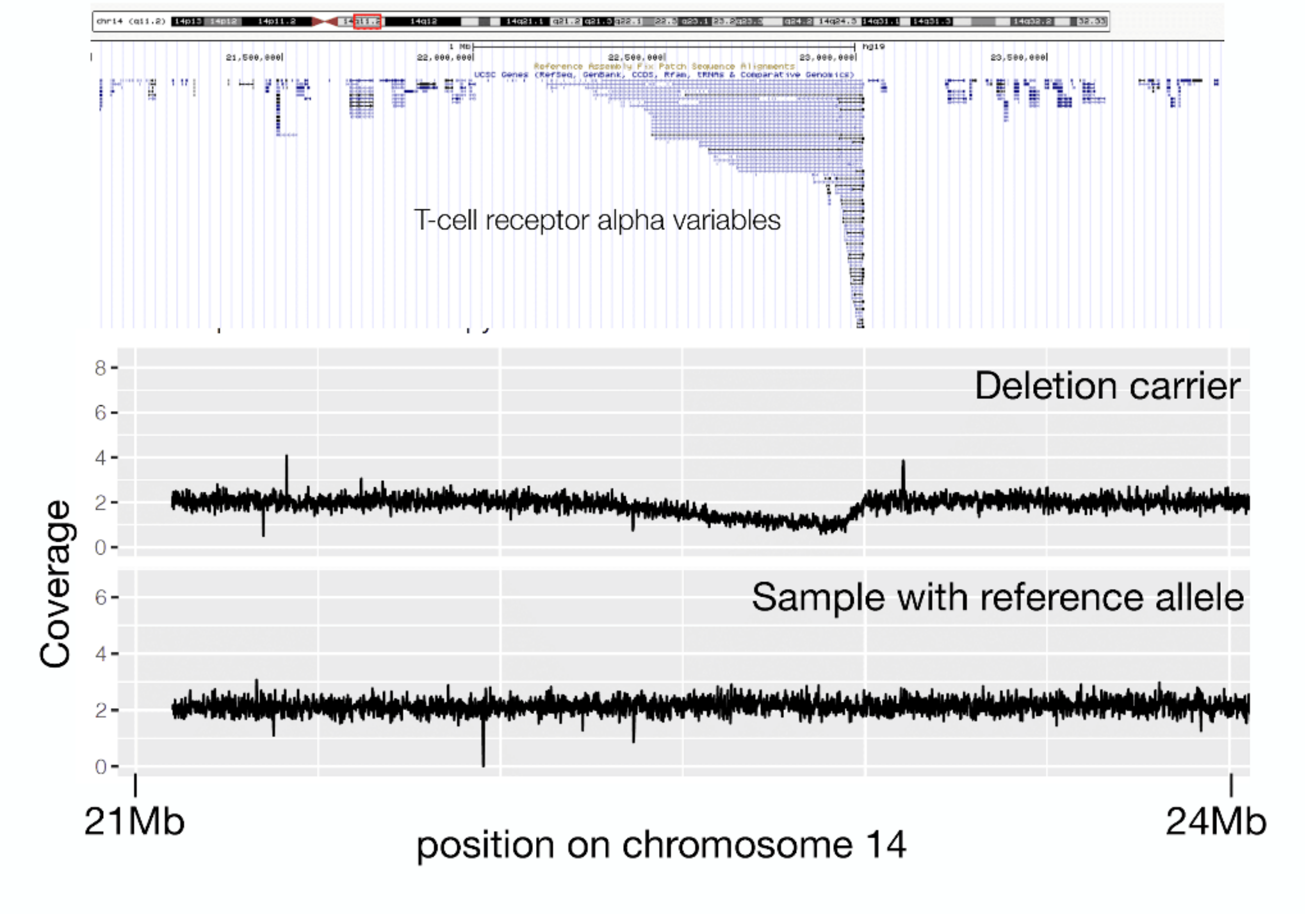
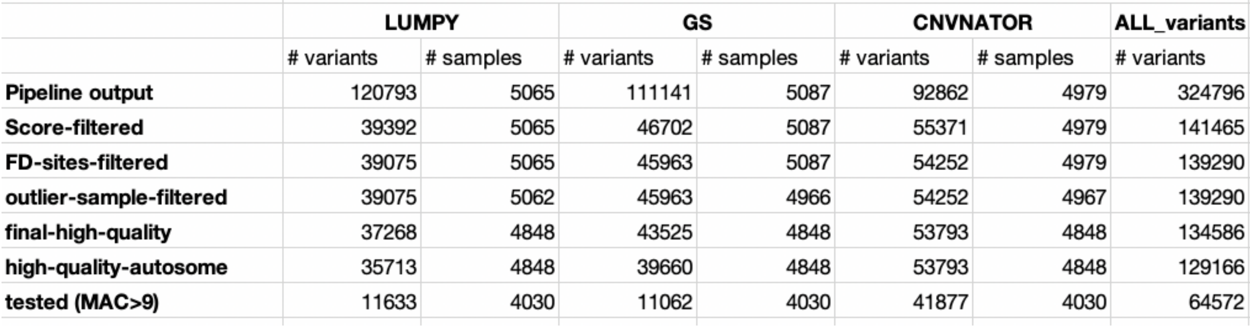


Figure 1.12. Read depth variation of the T-cell receptor genes

Read-depth coverage patterns at the chr14 T-cell receptor alpha variable region (coordinates LiftOver to GRCh37/hg19), showing one example for "deletion" carriers and one for a sample with the reference allele. The coverage values were calculated by CNVnator for 100bp windows, and the top gene track was extracted from UCSC genome browser (GRCh37/hg19).

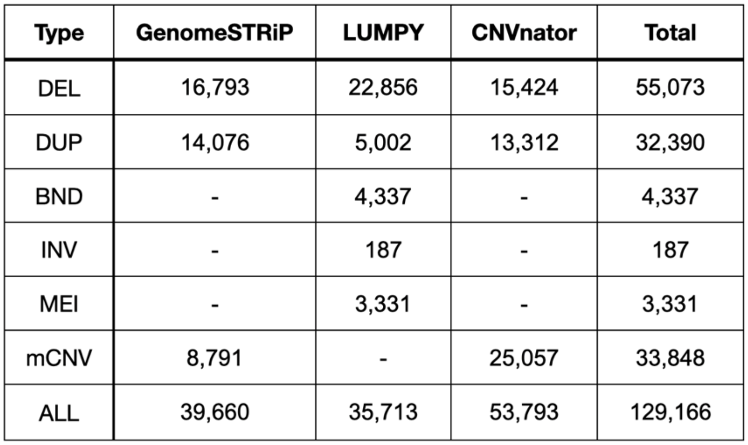
Table 1.1. Variant and sample counts in each QC step for WGS data.



	LUMPY		GS		CNVNATOR		ALL_variants
	# variants	# samples	# variants	# samples	# variants	# samples	# variants
Pipeline output	120793	5065	111141	5087	92862	4979	324796
Score-filtered	39392	5065	46702	5087	55371	4979	141465
FD-sites-filtered	39075	5065	45963	5087	54252	4979	139290
outlier-sample-filtered	39075	5062	45963	4966	54252	4967	139290
final-high-quality	37268	4848	43525	4848	53793	4848	134586
high-quality-autosome	35713	4848	39660	4848	53793	4848	129166
tested (MAC>9)	11633	4030	11062	4030	41877	4030	64572

Table 1.1. Variant and sample counts in each QC step for WGS data separated by variant calling pipelines. FD – false discovery, see **Methods** for the filtering criteria in each step.

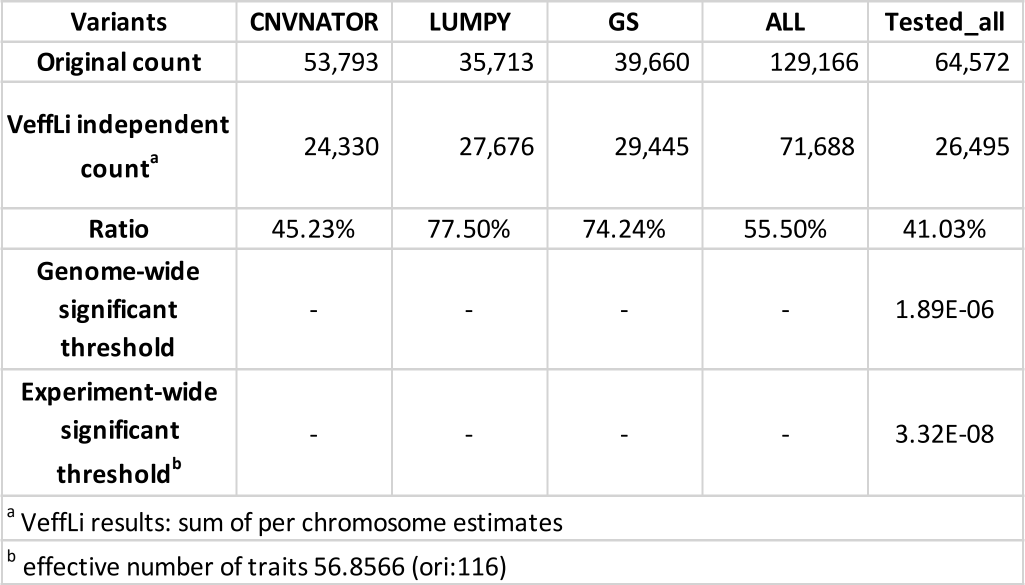
Table 1.2. High-confidence autosomal SVs count



Type	GenomeSTRiP	LUMPY	CNVnator	Total
DEL	16,793	22,856	15,424	55,073
DUP	14,076	5,002	13,312	32,390
BND	-	4,337	-	4,337
INV	-	187	-	187
MEI	-	3,331	-	3,331
mCNV	8,791	-	25,057	33,848
ALL	39,660	35,713	53,793	129,166

Table 1.2. Count of high-confidence autosomal SVs stratified by variant type and detection method including deletions (DEL), duplications (DUP), multiallelic copy number variants (mCNV), inversions (INV), mobile element insertions (MEI) and generic rearrangements of unknown architecture (BND).

Table 1.3. Genotype redundancy estimation



Variants	CNVNATOR	LUMPY	GS	ALL	Tested_all
Original count	53,793	35,713	39,660	129,166	64,572
VeffLi independent count[a]	24,330	27,676	29,445	71,688	26,495
Ratio	45.23%	77.50%	74.24%	55.50%	41.03%
Genome-wide significant threshold	-	-	-	-	1.89E-06
Experiment-wide significant threshold[b]	-	-	-	-	3.32E-08

[a] VeffLi results: sum of per chromosome estimates

[b] effective number of traits 56.8566 (ori:116)

Table 1.3. Estimation of redundant SV calls based on genotype information. Redundant variant calls identified by multiple SV detection methods are expected to have genotypes that are highly correlated. We therefore applied matSpDlite to each pipeline and to the combined callset to calculate the numbers of independent makers (VeffLi). We then applied the same method to the subset of the variants included in the trait association test and to the phenotypes to perform Bonferroni correction for the genome-wide significance threshold and experiment-wide threshold.

Table 1.4. Fragmentation level

Pipeline	#. SV	# Cluster	average cluster size	% single variant cluster	size of the largest cluster
GS	39,660	24,497	1.619	75%	96
LUMPY	35,713	23,751	1.321	90%	458[a]
LUMPY CNV	27,858	21,759	1.28	90%	149
CNVNATOR	53,793	16,962	3.171	73%	527

[a] a large inversion on chr7 with size of 44mb covered 400+ other variants

Table 1.4. Estimation of SV fragmentation based on physical clustering. Due to coverage fluctuations, CNV calls detected by read-depth analysis are often fragmented into multiple adjacent CNV calls that in fact represent a single variant. To estimate the degree of fragmentation, we clustered high-confidence autosomal CNVs within 10bp of each other and calculated the average number of SVs per cluster (average cluster size), the percentage of single variant clusters, and the maximum number of variants per cluster (size of the largest cluster).

Table 1.5. Callsets QC metrics

QC Metrics	Variants Subset	LUMPY	GS	CNVNATOR
CNV FDR [a]	all	-	27%	25%
	high confidence	0.80%	3%	9%
Counts	all	120,793	111,141	92,862
	high confidence	35,713	39,660	53,793
	common	11,633	11,062	41,877
Overlap w. 1kg [a]	all	10%	10%	11%
	high confidence	34%	21%	15%
	common	49%	34%	13%
Overlap w. gnomad [a]	all	18%	14%	25%
	high confidence	47%	27%	27%
	common	60%	40%	27%
Tagged by SNPs	high confidence	63%	62%	46%
	common	77%	65%	49%

[a] CNVs only

Table 1.5. Quality control metrics of the SV callsets including all variants, high-confidence variants, and high-confidence common variants (defined by >=10 carriers). CNV FDR was estimated by intensity rank sum test (IRS) using the SNP array data from METSIM samples. Note that LUMPY CNVs are by definition high confidence due to confirmation of independent read-depth support during variant classification steps (see **Methods**). Variant overlaps with 1KG and gnomAD were defined based on >50% reciprocal overlap. “Tagged by SNPs” was defined as SVs that are in LD (max r^2>=0.5) with any SNP in the 1Mb flanking regions.

Table 1.6. Summary statistics for all the genome-wide significant signals

SV type	Gene or annotation	Top trait	Chr	P WGS	P GWAS conditioned	BETA WGS	REP	Novel	Carrier freq.	P combined
deletion	*ALB*	Albumin	4	3.49E-21	1.05E-10	0.91	IMP	Y	0.03	1.47E-54[a]
deletion	*HP*	Glyco- -protein	16	1.38E-10	3.63E-04	-0.16	IMP	N	0.55	3.53E-35[a]
mCNV	*PDPR*	Pyruvate	16	9.41E-11	1.07E-10	-0.72	WES	Y	0.02	4.81E-21[a]
TCR	TRAV genes	CRP	14	1.30E-15	1.89E-15	1.2	WES	Y	0.36	1.51E-16[a]
deletion	*HNF1A-AS*	*CRP*	12	7.23E-04	3.60E-01	0.19	IMP	N	0.55	4E-13[a]
TCR	TRBV genes	CRP	7	3.36E-09	6.29E-09	0.84	WES	Y	0.38	2.47E-16[a]
mCNV	NUMTS	Fast insulin	1	1.00E-10	NA	-0.12	NA	Y	0	1E-10[a]
MEI	*LEPR*	CRP	1	3.94E-04	2.20E-01	0.16	IMP	N	0.51	4.5E-13[a]
deletion	*IL34*	Tyrosine	16	2.10E-04	5.45E-04	1.95	IMP	Y	0.02	4.17E-10[a]
MEI	*CDH13*	Adiponectin	16	1.24E-04	1.91E-02	-0.33	IMP	N	0.24	3.68E-08
mCNV	*AMDHD1*	Histidine	12	4.74E-04	2.72E-01	0.15	IMP	N	0.52	5.33E-07
mCNV	SegDup cluster	Fatty acid	16	1.10E-06	NA	-0.16	NA	Y	0.57	1.10E-06
mCNV	SegDup cluster	Glutamine	9	1.25E-06	NA	-0.79	NA	Y	0.43	1.25E-06
deletion	*PLTP*	Small HDL Particle	20	2.40E-04	3.81E-02	0.11	IMP	N	0.53	1.24E-06
mCNV	Simple repeats	Creatinine	4	1.41E-06	NA	-0.39	NA	Y	0.01	1.41E-06

[a] experiment-wide significant

Table 1.6. Summary statistics for 15 genome-wide significant loci with the top associated traits. Highly correlated SVs showing the same signal were manually inspected and clumped together. The genome-wide significance threshold was 1.89x10^{-6} and the experiment-wide significance threshold was 3.32x10^{-8} (see **Table 1.3** and **Methods** for details). The p value from WGS analysis and the p value from the replication experiment (IMP-imputation, WES-WES read-depth analysis, if applicable) were combined by Fisher's method and used to determine the significance level. The BETA WGS column shows the effect size in the unit of normalized trait value (e.g., for the ALB deletion, gaining one copy of the SV corresponds to 0.91 standard deviation of increased albumin level). The carrier frequency was calculated in the WGS dataset. The column of "P GWAS conditioned" shows the SV p value conditioned on all intrachromosomal GWAS SNPs from GWAS Catalog(Buniello et al. 2019), using WGS data only (see **Methods**)

Table 1.7. Leave-one-out validation for genome-wide significant SVs

VAR	FALSE	TRUE	AC_RATE
40551	17	3891	0.996
52933	113	3795	0.971
61703	55	3853	0.986
62003	7	3901	0.998
chr12_95946601_95947800	260	3648	0.933
chr16_72057601_72058200	63	3845	0.984
chr20_45906701_45907200	144	3764	0.963
CNV_chr4_73399922_73404147	9	3899	0.998

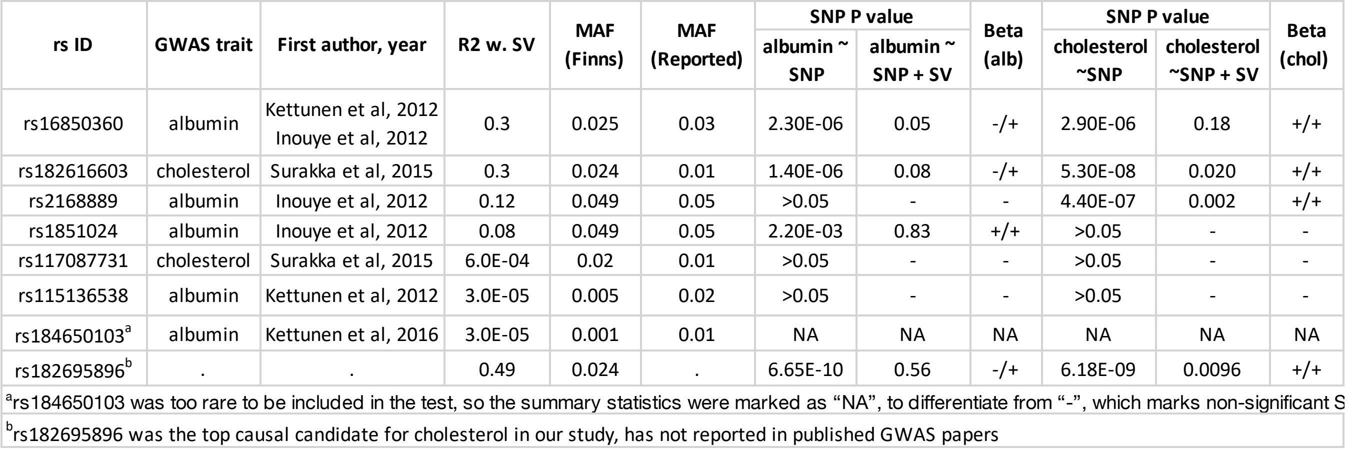
Table 1.7. The "leave-one-out" validation experiment to assess imputation quality of the eight genome-wide significant SVs. For each variant, we ran 3,908 imputation experiments and in each we used one sample as the test genome and the other samples as the reference. The accuracy rate was calculated among all 3,908 tests.

Table 1.8. Test ALB deletion conditioned on GWAS SNPs

rs ID	GWAS trait	First author, year	R2 w. SV	MAF (Finns)	MAF (Reported)	SV P value		Beta (conditional	
						albumin ~ SNP + SV	cholesterol ~ SNP + SV	albumin	cholesterol
rs16850360	albumin	Kettunen et al, 2012 Inouye et al, 2012	0.3	0.025	0.03	8.10E-18	6.00E-04	1.1	-0.39
rs182616603	cholesterol	Surakka et al, 2015	0.3	0.024	0.01	6.60E-17	4.00E-03	1.06	-0.32
rs2168889[a]	albumin	Inouye et al, 2012	0.12	0.049	0.05	6.40E-23	9.70E-05	1.05	-0.37
rs1851024	albumin	Inouye et al, 2012	0.08	0.049	0.05	2.30E-19	3.30E-08	0.91	-0.5
rs117087731	cholesterol	Surakka et al, 2015	6.00E-04	0.02	0.01	2.50E-21	1.70E-08	0.91	-0.49
rs115136538	albumin	Kettunen et al, 2012	3.00E-05	0.005	0.02	2.80E-21	1.50E-08	0.91	-0.49
rs184650103	albumin	Kettunen et al, 2016	3.00E-05	0.001	0.01	2.90E-21	1.60E-08	0.91	-0.49
rs182695896[b]	.	.	0.49	0.024	.	6.52E-13	2.33E-02	0.97	-0.27
SV ~ albumin P value = 3.49E-21, beta = 0.9107									
SV ~ cholesterol P value = 1.17E-08, beta = -0.4929									
[a] rs2168889: collider effect									
[b] rs182695896: top causal candidate for cholesterol in our study, has not reported in published GWAS papers									

Table 1.8. Association analysis between the *ALB* deletion and albumin/total cholesterol conditioned on the seven previously published GWAS SNPs and rs182695896 one at a time. None of the seven GWAS SNPs diminish the SV-albumin signal, while the first three SNPs attenuate the SV-cholesterol signal, suggesting that they might also be in LD with the underlying causal variants for cholesterol. MAF(Finns) – MAF in our data, MAF(Reported) – MAF reported in previous GWAS studies.

Table 1.9. Test the GWAS SNPs w./w.o. SV as covariate



rs ID	GWAS trait	First author, year	R2 w. SV	MAF (Finns)	MAF (Reported)	SNP P value		Beta (alb)	SNP P value		Beta (chol)
						albumin ~ SNP	albumin ~ SNP + SV		cholesterol ~SNP	cholesterol ~SNP + SV	
rs16850360	albumin	Kettunen et al, 2012 Inouye et al, 2012	0.3	0.025	0.03	2.30E-06	0.05	-/+	2.90E-06	0.18	+/+
rs182616603	cholesterol	Surakka et al, 2015	0.3	0.024	0.01	1.40E-06	0.08	-/+	5.30E-08	0.020	+/+
rs2168889	albumin	Inouye et al, 2012	0.12	0.049	0.05	>0.05	-	-	4.40E-07	0.002	+/+
rs1851024	albumin	Inouye et al, 2012	0.08	0.049	0.05	2.20E-03	0.83	+/+	>0.05	-	-
rs117087731	cholesterol	Surakka et al, 2015	6.0E-04	0.02	0.01	>0.05	-	-	>0.05	-	-
rs115136538	albumin	Kettunen et al, 2012	3.0E-05	0.005	0.02	>0.05	-	-	>0.05	-	-
rs184650103[a]	albumin	Kettunen et al, 2016	3.0E-05	0.001	0.01	NA	NA	NA	NA	NA	NA
rs182695896[b]	.	.	0.49	0.024	.	6.65E-10	0.56	-/+	6.18E-09	0.0096	+/+

[a] rs184650103 was too rare to be included in the test, so the summary statistics were marked as "NA", to differentiate from "-", which marks non-significant S

[b] rs182695896 was the top causal candidate for cholesterol in our study, has not reported in published GWAS papers

Table 1.9. The association tests between each of the seven previously published GWAS SNPs as well as rs182695896 and serum albumin/total cholesterol, with and without the *ALB* deletion as a covariate (SNPs with p-value > 0.05 were not included in the conditional analysis, with "-" in the related fields). The "Beta" column shows the direction of effects of SNPs with/without the SV in the model. rs115136538, rs184650103 and rs117087731 did not show significant association with either trait in our dataset. The other SNPs showed signals with albumin or total cholesterol which became much less significant after conditioning on SV genotype. *Note: rs184650103 was too rare to be included in the test, so the summary statistics were marked as "NA", to differentiate from "-", which marks non-significant SNPs.

Table 1.10. Conditional analysis (phenotype - phenotype)

Variant	Tested trait	Covariate trait	P WGS	P conditioned	BETA WGS	BETA conditioned	Mediator?
ALB deletion	Albumin	Total Cholesterol	3.49E-21	5.74E-25	0.9107	0.9937	N
	Total cholesterol	Albumin	1.17E-08	1.16E-11	-0.4929	-0.6558	N
PDPR mCNV	Pyruvate	Alanine	9.41E-11	6.59E-05	-0.5817	-0.4344	N
	Alanine	Pyruvate	2.93E-07	1.47E-03	-0.5744	-0.3197	Y
HP deletion	Glycoprotein	Total cholesterol	1.51E-11	2.78E-15	-0.2081	-0.1988	N
	Total cholesterol	Glycoprotein	1.01E-05	2.62E-10	0.1466	0.1604	N

Table 1.10. Conditional analysis of the three multi-trait associated variants, taking one trait as a covariate and testing the other. Additional traits were tested for the *ALB* deletion conditioned on albumin and total cholesterol, the results of which can be found in **Supplementary Figure 6b**. The covariate trait was defined as a mediator of the tested trait if the conditional p-value failed the genome-wide significance threshold ($1.89x10^{-6}$).

1.5 Author Contributions

I.M.H. and N.O.S. conceived and directed the study. L.C., H.J.A, and I.D. adapted the GenomeSTRiP pipeline to perform CNV detection at scale. H.J.A. developed the pipeline for CNV genotyping based on CNVnator. L.C. and I.D. created the GenomeSTRiP callset; L.C. and H.J.A created the CNVnator callset; D.E.L. and K.L.K created the LUMPY callset, and led data management. L.C. led all analyses related to trait association, SV genotyping using WES and array data, and investigation of candidate loci. H.J.A, D.E.L, I.D, L.G. and A.A.R. led GATK callset creation and QC for WGS data. A.P., S.R., M.L, and J.K. contributed samples and phenotypic data. All authors edited the manuscript and/or provided intellectual contributions. L.C. and I.M.H. wrote the manuscript.

1.6 Acknowledgements

We thank D. Ray from Johns Hopkins University for her comments to the manuscript. This work was funded by an NHGRI CCDG award to IMH and NOS (UM1 HG008853) and DK U01 DK062370, the NHGRI large-scale sequencing grant (grant number 5U54HG003079), the Sigrid Jusélius Foundation (to SR), the University of Helsinki HiLIFE Fellow grants 2017-2020 (to SR), the Academy of Finland Center of Excellence in Complex Disease Genetics (grant number 312062 to SR, grant number 312074 to SR and AP), the Academy of Finland (grant number 285380 to SR), the National Heart, Lung and Blood Institute (grant number T32HL007081 to EY), and the National Center for Advancing Translational Sciences (grant number UL1TR002345 to EY). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank the MGI administration and data production team, in particular R. Fulton, L. Fulton, C. Fronick, A. Wollam, S.K. Dutcher, and J. Milbrandt. The FINRISK samples used for the research were obtained from THL Biobank. We thank all study participants for their generous participation in the THL Biobank, FINRISK study, and METSIM study. ASH was supported by the Academy of Finland (grant no. 321356). LC was supported by the McDonnell International Scholars Academy Fellowship. AJS was supported by the Mr. and Mrs. Spencer T. Olin Fellowship for Women in Graduate Study.

Chapter 2:

Association between blood mtDNA content (MT-CN) and insulin related traits

The contents of this chapter were in a preprint released in 2020 and the manuscript is currently under review: Ganel, L., **Chen, L.**, Christ, R., Vangipurapu, J., Young, E., Das, I., ... & Hall, I. M. (2020). *Mitochondrial genome copy number in human blood-derived DNA is strongly associated with insulin levels and related metabolic traits and primarily reflects cell-type composition differences.* medRxiv.

2.1 Introduction

2.1.1 Insulin-associated CNVs on nuclear mitochondrial DNA segments (NUMTs)

One interesting result from the SV trait-mapping study was a CNV on chr1:628901-636500 associated with insulin level in fasting samples. This signal was first identified in a preliminary analysis with 2,063 Finnish samples (the first batch finished sequencing) using copy number window screening method (see chapter 2.2.1 for details), together with another signal on chr17 showing similar association pattern (chapter 2.3.1). Both CNVs overlapped with a type of special genomic regions called nuclear mitochondrial DNA segments (NUMTs). NUMTs evolved from ancient mitochondrial DNA fragments inserted in the eukaryotic nuclear genome, probably through non-homologous end joining (NHEJ) at double-strand breaks (DSBs) (J. V. Lopez et al. 1994; J. Lopez, Stephens, and O'Brien 1997; Rateb Dweik 2017). Because of the high similarity between NUMTs sequences and their homologous sequences on mitochondrial genome as well as several other evidence (chapter 2.3.1), it was likely that the CNVs detected on NUMTs were actually the indirect measurement of the average mtDNA copy number (MT-CN) in peripheral blood cells. To test this hypothesis, we designed an experiment to directly estimate MT-CN from the reads aligned to mitochondrial genome, and test for the association between MT-CN and metabolic traits.

2.1.2 Mitochondrial DNA copy number (MT-CN) and metabolic traits

As an important type of cellular organelle, mitochondrial provide the source of chemical energy for cells. The number of mitochondrial per cell varies across different cell types with a wild range (1~600,000), so as the mitochondrial DNA (mtDNA). There were previous studies suggesting the potential relationship between mtDNA and cardiometabolic phenotypes(Nisoli et al. 2007; Guyatt et al. 2018; X. Zhou et al. 2016; Wang and Wei 2020; Ding et al. 2015; Chen et al. 2014) . However, many studies reported conflict results about the effect direction of mitochondrial copy number to the same trait (e.g., insulin resistance), especially in the peripheral blood tissues(Lee et al. 1998; Perfield et al. 2013; Shoar et al. 2016; Song et al. 2001; Ding et al. 2015). One potential explanation is that most of the previous studies were done in clinical settings, where the sample sizes were small, and the designs were biased towards individuals with disease. Utilizing the access to large sequencing dataset collected by cohort studies from the general population together with various cardiometabolic phenotypes, we aimed to unveil the role of mtDNA copy number in metabolic syndrome with an enhanced statistical power.

2.2 Methods

2.2.1 Brief introduction of the pilot CNW experiment

In the first freeze of sequencing data, we performed a pilot experiment on 2,063 Finnish samples from FINRISK and METSIM cohorts with an association analysis on 1kb autosome variable copy number windows (vCNWs). We first split the autosomes into 1kb adjacent windows and estimated the copy number of each window using CNVnator (Abyzov et al. 2011). Since most windows were expected to be from non-variable regions, we filtered for variable copy number windows (vCNWs) via an outlier detection method proposed by Hoaglin and Iglewicz(Hoaglin and Iglewicz 1987), and tested those vCNWs for association with the quantitative metabolic phenotypes using EMMAX model. After corrected for the number of independent loci in the tests, we set up the significance threshold at 2 x 10^{-7}. The genome-wide significant signals were then investigated in a similar way as chapter 1 candidates.

2.2.2 Direct measurement of MT-CN and batch effect correction

To develop the method to directly measure the mitochondrial DNA copy number (MT-CN), we first calculated per base coverage on the mitochondrial genome from the alignment data using a combination of SAMTools(H. Li et al. 2009) (to convert CRAM format to BAM format) and BEDtools(Quinlan and Hall 2010) (the genomecov subcommand). Given the read depth distribution was fairly uniform (**Figure 2.2a**), for each individual we computed the mean coverage of mitochondrial genome divided by the mean coverage of nuclear genome to estimate the average MT-CN in peripheral blood. We then compared the direct measurement with the copy number of NUMTs (indirect measurement), to validate the first part of the hypothesis (**Figure 2.2b**). Besides Finns, we also applied this method to other population cohorts collected for CCDG project, while the rest all suffered from the hard-to-correct batch effect and were not included in the future analyses (**Figure 2.6 and chapter 2.4.1**).

In our variety of quality control analyses for the direct measurement of MT-CN, the DNA sample collection procedure turned out to be a big confounder (**Figure 2.3, chapter 2.3.2**). Specifically, among the Finns the FINRISK samples collected in the 2002 and 2007 cohorts showed strong batch effects, with significant mean-shifts compared to other cohorts. Therefore, we separately the data into three analysis batches: 1. FR92 and FR97, 2. FR02 and FR07, and 3. METSIM for the downstream analysis. For each batch the MT-CN estimates were regressed out by possible confounders (Age and Age^2) and transformed to approximately normal distribution by rank-based inverse transform method. And then the standardized MT-CN values from all batches were combined for the trait association tests.

2.2.3 Association test by direct measurement

We then applied EMMAX model to the directly measured MT-CN, with the metabolic traits as response variables. We compared the results of directly measured versus indirectly measured mtDNA content, in both pre-processed stage (raw estimates) and post-processed stage (normalized traits). We also included mitochondrial haplotype group inferred from mtSNPs as a covariate in the model, to test whether some demographic features could explain both the phenotypic and mtDNA differences.

2.2.4 Expanding the analysis to WES data

Similar to the strategy applied in chapter 1 for boosting the sample size and statistical power, we sought to expand our experiment to the 20k Finns using WES data. We first looked at the alignment data and found that the coverage of WES data on mitochondrial genome was pretty sparse (**Figure 2.5c**). So instead of using the mean coverage to estimate MT-CN, we compared several summary statistics of the WES coverage and selected the maximum to move forward, for it had the highest correlation with the WGS measurement (R^2=0.44). We then applied similar procedures to estimate the MT-CN and normalize the WES measurement, which was tested for association with quantitative traits in all ~20k samples, in samples with WGS data and in samples without WGS data (**Table 2.2**).

2.3 Results

2.3.1. CNVs on nuclear mitochondrial sequences (NUMTs) associate with insulin/fat mass traits

Among the 272,996 vCNWs tested in our preliminary analysis, there were two regions significantly associated with insulin and fat mass with similar summary statistics, one on chromosome 1 and the other on chromosome 17 (**Table 2.1**). The copy number genotype of those two regions were well correlated and far above the normal range for diploid genomes (**Figure 2.1a**). According to the annotation provided by UCSC Genome Browser, those two regions both overlapped with nuclear mitochondrial sequences (NUMTs) (one showed in **Figure 2.1b**). The vCNWs were highly similar to the homologous mitochondrial sequences, with the similarity of 98.7% and 85.4% given by BLAT(Kent 2002) (**Table 2.1**). Based on the above observations, we hypothesized that the read-depth detected NUMTs CNVs actually reflected the variation of the mitochondrial DNA copy number (MT-CN) among individuals, and this indirect measurement happened when the reads amplified from mitochondrial DNA were misaligned to the NUMTs region of nuclear genome in the upstream pipeline.

2.3.2. Direct measured MT-CN showed stronger association signals with multiple metabolic traits in Finns

We directly measured the MT-CN using the approach mentioned in chapter 2.2.2 for all the 5k Finnish samples (same set of WGS samples studied in chapter 1), and found the direct measurement well correlated with both NUMTs CNVs on chr1 and chr17 (**Figure 2.2b**, with R2 of 0.49 and 0.68, respectively). Given the decent correlation, we went ahead tested the association between MT-CN and all the 116 quantitative metabolic traits, with particular interests on insulin and fat mass, the two candidate traits showed up in the preliminary analysis using NUMTs vCNWs. For comparison we also ran CNVnator again on the same set of 5k samples to estimate the NUMTs copy number and generated the same set of summary statistics for the indirect measurements. To summarize the results, the signals of insulin and fat mass became much stronger with the directly measured WGS genotype **(Figure 2.4a**, p-value = 2.02 x 10^{-21} for insulin and p-value = 4.48 x 10^{-16} for fat mass) and conditioning on the mitochondrial haplotype groups did not change the results. Besides insulin and fat mass, we also observed many novel signals including C-reactive protein (p=9.21 x 10^{-14}), total triglycerides (p=9.21 x

10^{-14}), and HDL cholesterol (p=4.00 x 10^{-15}), suggesting MT-CN might have broader effects on multiple metabolic phenotypes.

We also used WES data to measure MT-CN and ran the association tests (chapter 2.2.4). The measurement was noisier while still able to show the same association signals in the 20k samples as well as the subsets with and without WGS data (**Table 2.2**).

To understand the connection between MT-CN and more clinically relevant phenotypes, our collaborators in METSIM project tested our direct measured MT-CN against two additional traits -- Matsuda ISI and disposition index, which measure insulin sensitivity and secretion, and both traits were significant (p=4.3x10^{-26} for insulin sensitivity, p=3.0 x 10^{-7} for insulin secretion, N=2975 for both traits). Notably, the Matsuda ISI signal was still significant when conditioned on fat mass and excluding diabetic individuals, which indicates that the association of MT-CN with insulin sensitivity was independent of fat mass, while the secretion association was likely to be linearly correlated with fat mass association.

2.4 Discussion

2.4.1 The potential and limitation of measuring MT-CN with WGS data

Compared to previous measurement of mitochondrial DNA content, such as real time PCR(Gahan et al. 2001) and quantitative PCR(Gourlain et al. 2003), our WGS-based approach is much more scalable, which gives it a big application potential given the increased availability of large WGS-sequenced cohorts in recent years. Since sample size was the common drawback of previous studies, applying our method to analyze large WGS datasets with tens of thousands of samples will shed light on the metabolic roles of MT-CN, as we explicitly proved in the preprint and briefly discussed in chapter 2.4.2.

However, this WGS-based MT-CN measurement has a few caveats as well. First, what we calculated was the average amount of mitochondrial DNA copies relative to the copies of nuclear genome in the bulk blood tissue, and the variation of this measurement could be significantly contributed by the difference in cell type composition among individuals. Similar to the variation we found on T-cell receptor genes in chapter 1, this trait was not germline mutations and it could be affected by many confounders such as immune responses. The second limitation, as illustrated in **Figure 2.3** and **2.6**, was the batch effect. As briefly mentioned in chapter 2.2.2 we performed the same analysis in all the CCDG samples with WGS data available at that time and planned on a large meta-analysis across multiple CVD cohorts from different populations. However, we found that the MT-CN distributions of samples collected in different centers separated from one another, with large differences in means as well as standard deviations. Technically we could standardize the data using the same procedure applied in Finns, while given the small sample size of most cohorts, the normalization was not likely to work very well and there was a big risk of erasing the real differences among samples when data from different standardized batches were combined. Considering the large variation in sample collection and library preparation procedures among most studies, the batch effects are likely to be common and therefore we suggest that whoever apply this method needs to be extra cautious about the quality control of the data.

2.4.2 The story beyond: genetic determinants of MT-CN and its causal relationship with metabolic traits

The strong associations with multiple metabolic traits put MT-CN in the spotlight of our research, however as mentioned previously, it was not clear whether this trait is genetically inheritable and what are the underlying causal relationships between MT-CN and CVD-related metabolic phenotypes. To address these important questions, my colleague Liron Ganel, also the first author of the preprint, conducted a series of analyses to find the genetic determinants for MT-CN, estimate the heritability of this mitochondrial trait, untangle the underlying causal relationship between MT-CN and fasting serum insulin via a modified Mendelian Randomization (MR) model, and finally interpret the association between MT-CN and metabolic syndrome in UK Biobank data with polygenic risk score (PRS) approach. Considering the length of this thesis and the free access to our manuscript, here we only briefly introduce the conclusion from his work, leaving the details to the publication. In summary, we found that genetics played a significant role in the variation of blood-derived MT-CN measured from WGS data, which turned out to be a reflection of the relative quantities of circulating immune cells in the blood. This result, together with other prior evidence (e.g., CRP as a risk factor of CVD) suggested that inflammation might play a role in metabolic syndrome.

2.5 Acknowledgement

All the downstream analyses including the MT-CN GWAS, heritability estimation and MR analyses were performed by my colleague Liron Ganel, who denoted three years of his PhD work to this project. Together with our advisor Dr. Ira Hall, he also led the writing of the manuscript, to which I only contributed a minor part. He deserves a round of the applause for years of hard work and remarkable resilience, especially after encountered lots of frustrating and confusing results. A special thanks to Ryan Christ, who was a postdoc in the lab at that time and provided significant amount of advice with his strong statistical background to help Liron ace the MR study. We would also like to thank our Finnish collaborator Jagadish Vangipurapu and Markku Laakso from University of Eastern Finland, who conducted the association tests on insulin resistance and secretion traits.

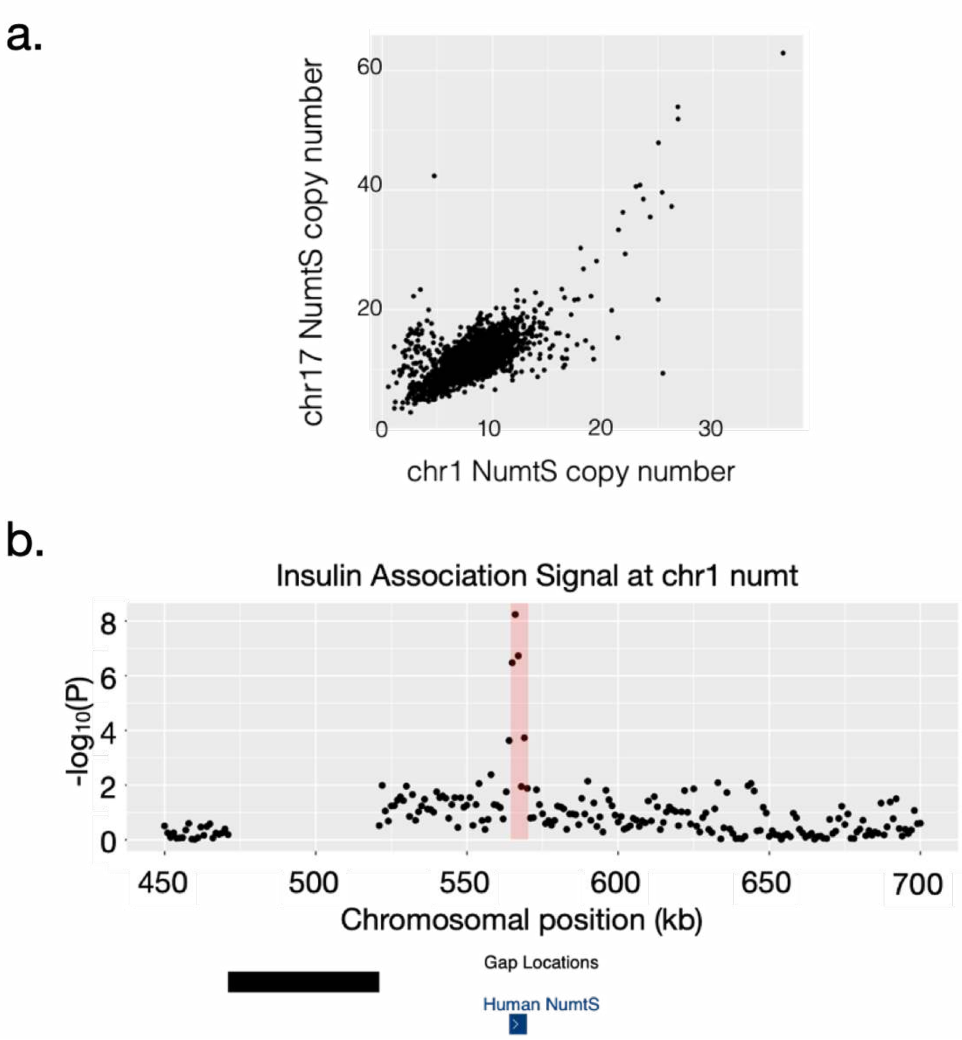


Figure 2.1 CNVs on chr1 and chr17 NUMTS regions and their association with insulin

(a) The genotype correlation between chr1 and chr17 NUMTS CNWs, with the copy number estimated by CNVnator. Each dot represents a sample.

(b) Local Manhattan plot for the chromosome 1 CNWs on region 450kb-700kb, with the NUMTS region highlighted by the pink block. The annotation tracks of Gap Locations and Human NumtS were from UCSC Genome Browser.

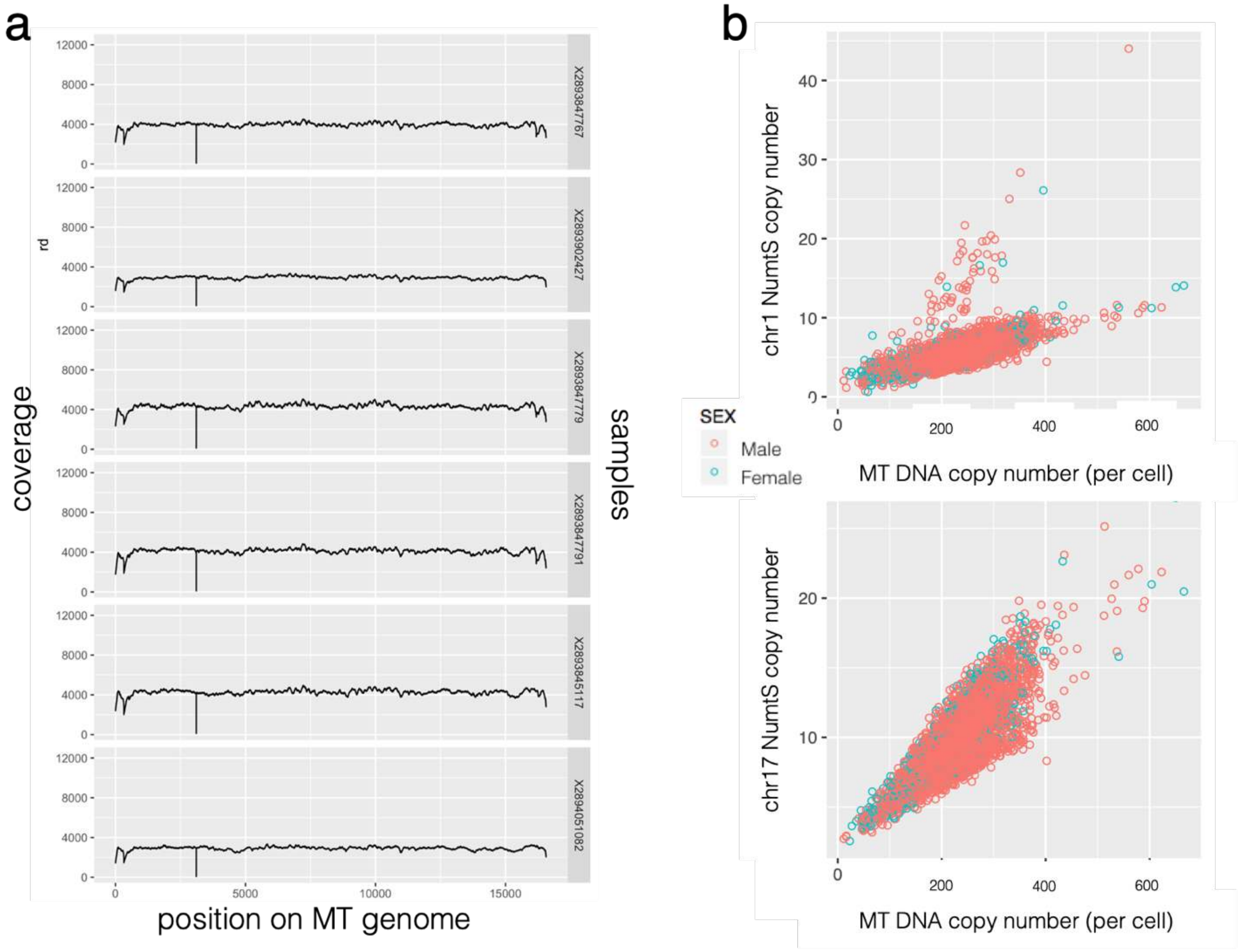


Figure 2.2 Directly measure MT-CN by aligning to mitochondrial genome

(a) Per-base coverage of the WGS reads aligned to Mitochondrial genome, showing six random samples as example.

(b) The scatterplot of directly measured MT-CN (x-axis) versus the indirect measurement from NUMTS copy number (y-axis). Each dot represents a sample, colored by sex.

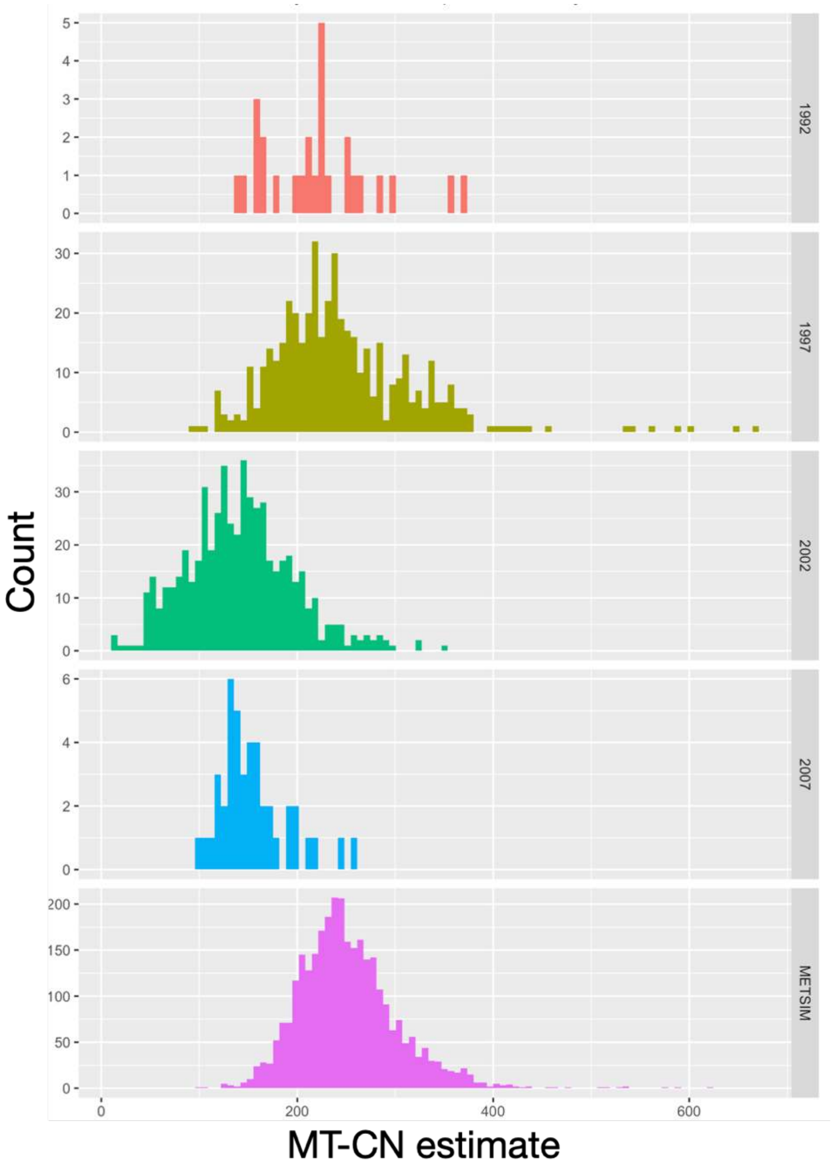


Figure 2.3 Batch effects among FINRISK cohorts

Distributions of the direct measured MT-CN from FINRISK (top four panels) and METSIM, with FINRISK samples further stratified by the year of sample collection. Batch effects mainly manifested in the 2002 and 2007 FINRISK cohorts, which is likely be caused by the change of DNA collection and storage protocol.

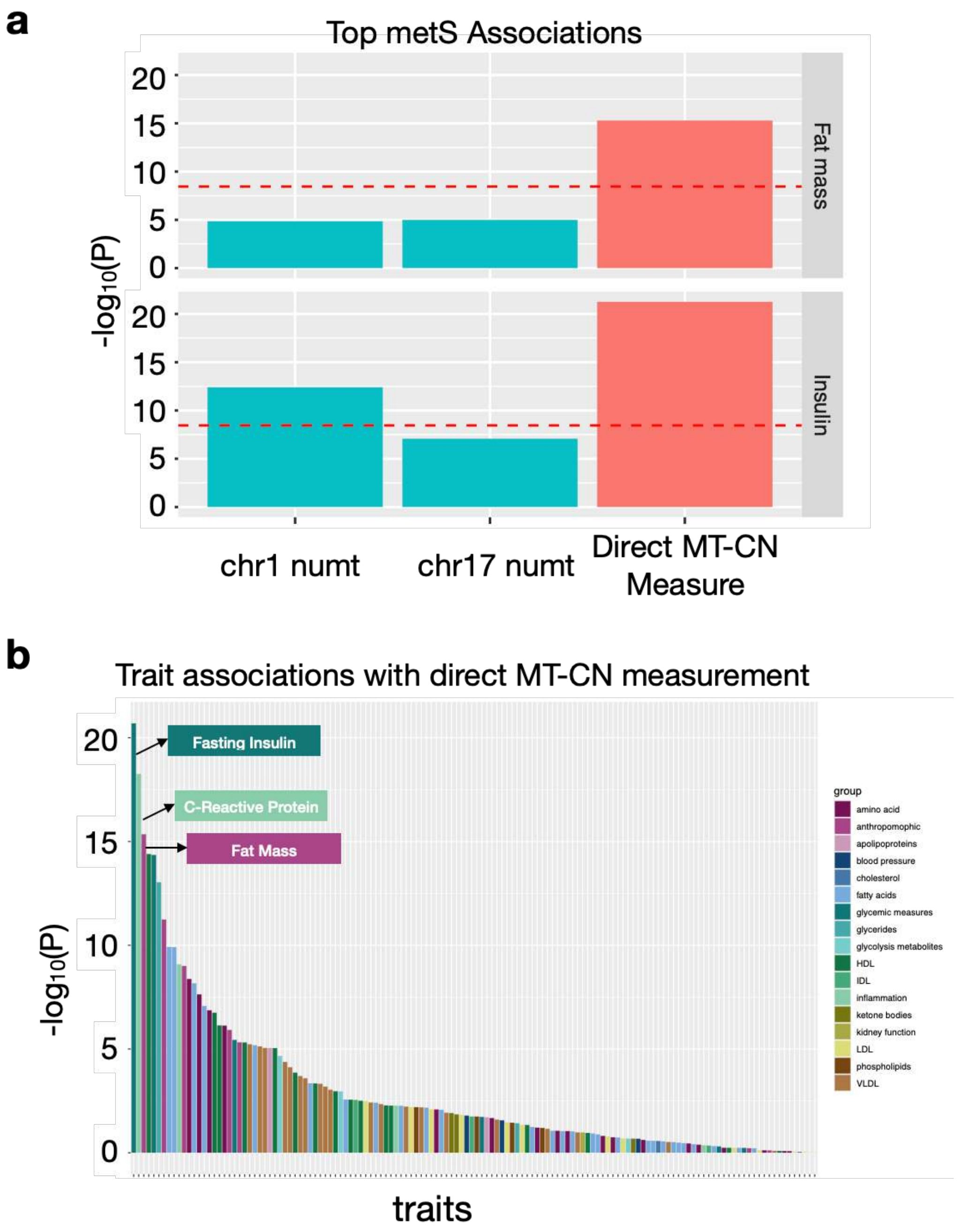


Figure 2.4 Trait association tests using directly measured MT-CN

(a) The negative log10 p-values of direct MT-CN measure (red) compared to those measured from NUMTS copy number (blue) in the association tests with insulin and fat mass, the top two candidate traits from CNW analysis.

(b) Phenom-wide negative log10 p-values between MT-CN and all the tested metabolic traits, colored by the trait groups.

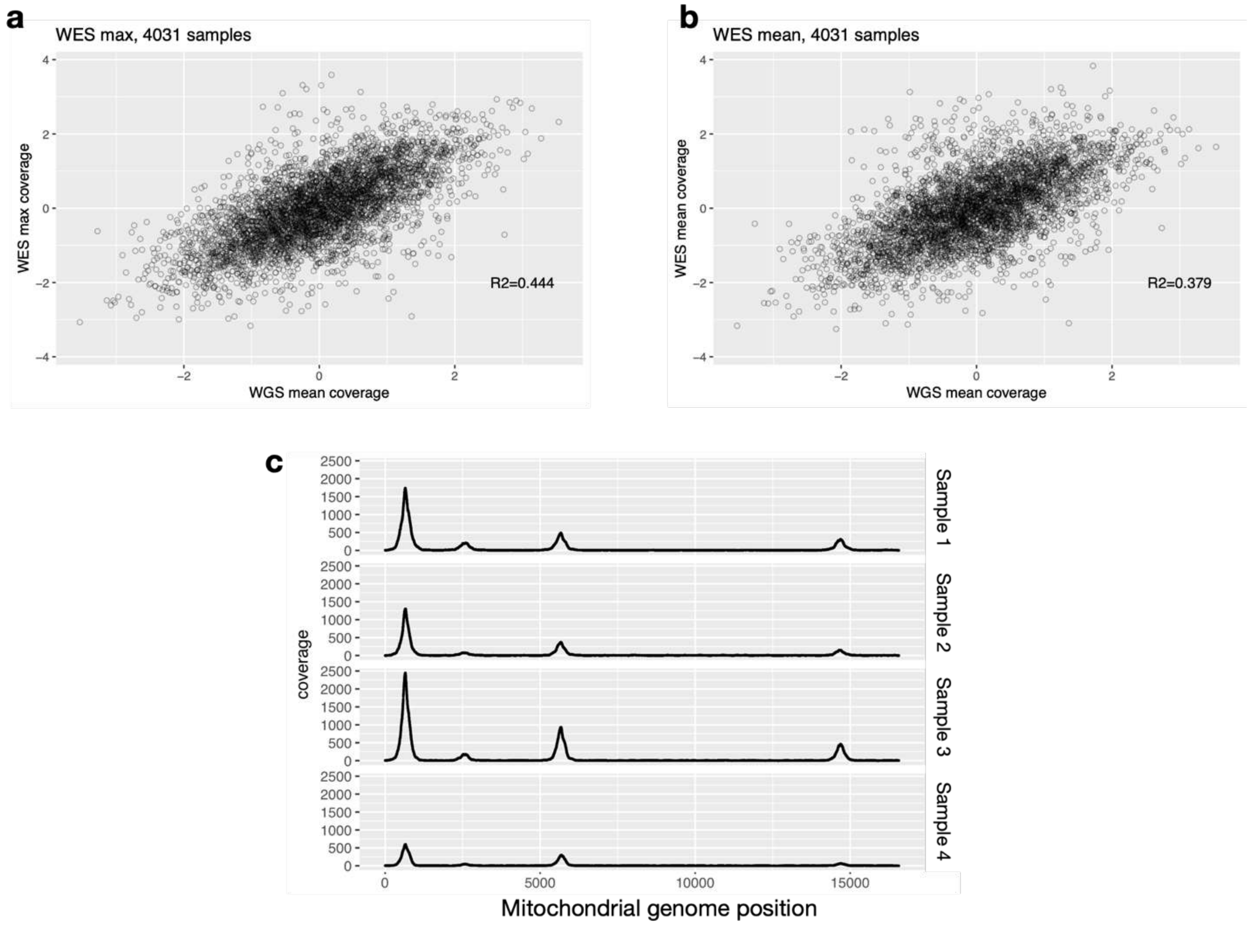


Figure 2.5 Measuring MT-CN using WES data

(a) The correlation between MT-CN measured by the mean coverage of WGS data and the max coverage of WES data using normalized MT-CN measurement.

(b) Similar to (a), but with mean coverage of WES data

(c) The per-base coverage profile for the alignment results of WES reads on mitochondrial genome, showing four samples as example.

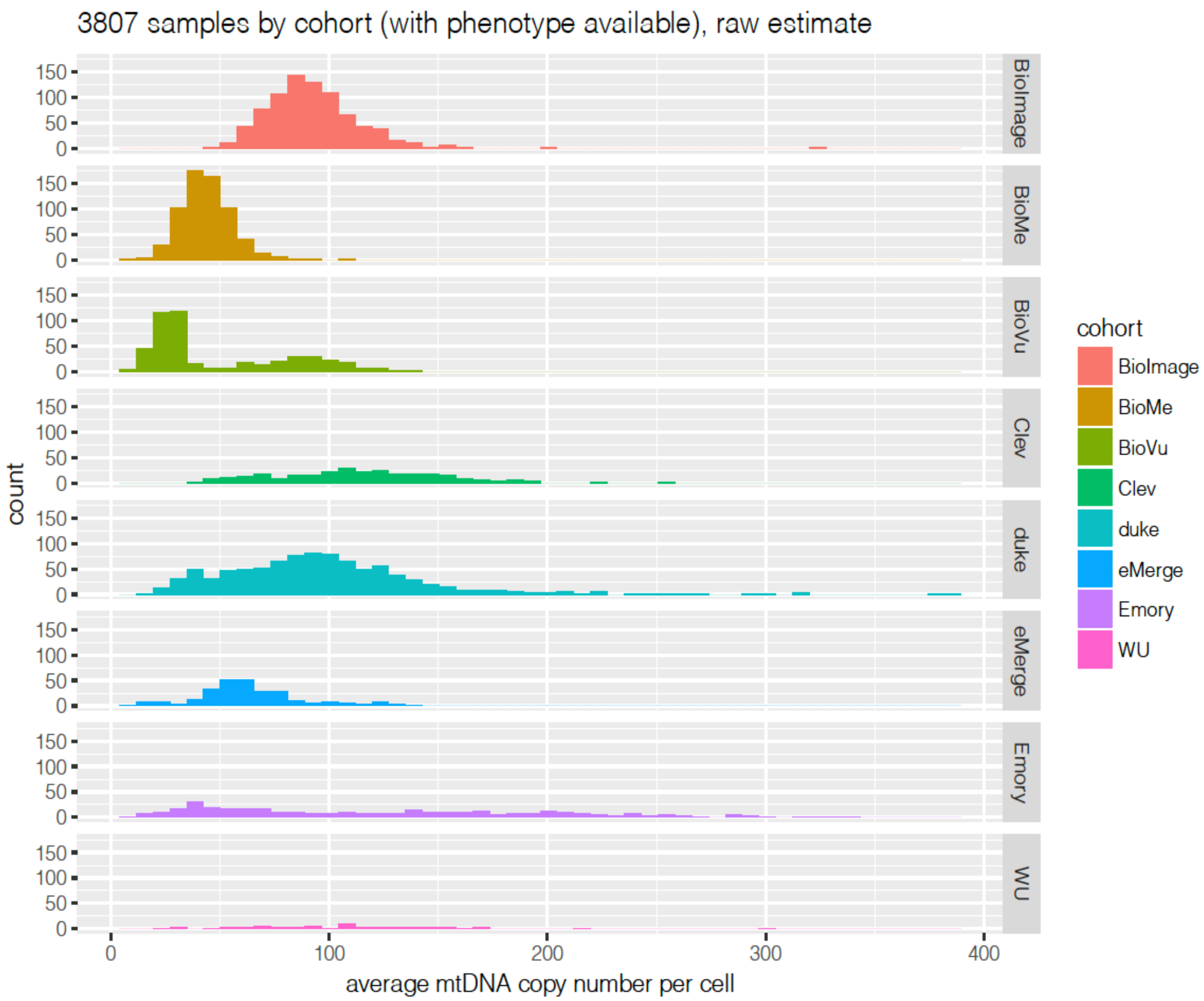


Figure 2.6 Measuring MT-CN in CCDG African American samples

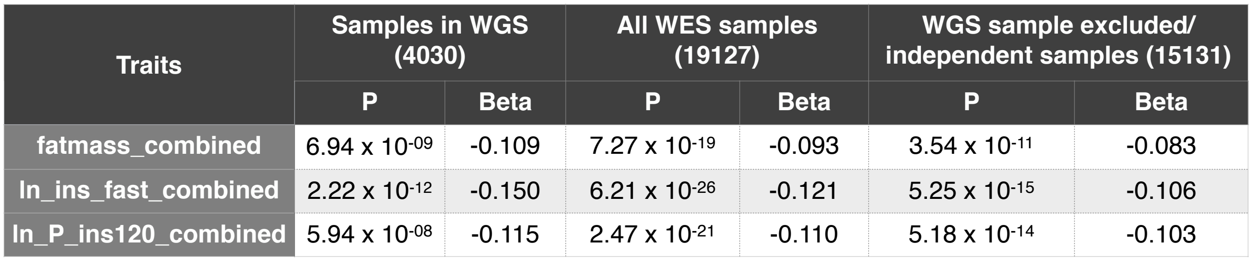
Distributions of the direct measured MT-CN in the 3807 African American samples from CCDG project, stratified and colored by the different cohorts which conducted the initial sample collection.

Table 2.1 The association signals of NUMTs CNVs

quantitative trait ~ variable copy number window genotype + kinship (~2k Finnish samples, b37)							
Nuclear coordinates (b37)	Nuclear coordinates (b38)	MT alignment (b38)	Similarity (BLAT)	Trait associated	Sample size	P	Beta
chr1:565000-566000	chr1:629620-630620	chrM:4451-5451	98.7%	ln_ins_fast_combined	1107	3.33 x 10^{-7}	-0.066
chr1:566000-567000	chr1:630620-631620	chrM:5450-6450	98.7%	fatmass_combined; ln_ins_fast_combined	1309; 1107	2.98 x 10^{-7}; 5.70 x 10^{-9}	-0.097; -0.13
chr1:567000-568000	chr1:631620-632620	chrM:6450-7449	98.5%	ln_ins_fast_combined	1107	1.86 x 10^{-7}	-0.098
chr17:22020000-22021000	chr17:22520674-22521674	chrM:15853-16569	85.4%	fatmass_combined; ln_P_CRP_combined	1309; 1964	5.46 x 10^{-9}; 4.87 x 10^{-8}	-0.058; -0.041

The information of the trait-associated NUMTS CNWs, with summary statistics from the of 2k WGS-sequenced samples (aligned to GRCh37 reference). The p-values marked by red survived the multiple testing correction for that preliminary analysis. The “Similarity” column showed the BLAT results between each of those NUMTS windows and its homologous sequence on mitochondrial genome (column on the left).

Table 2.2 WES MT Copy Number vs. Candidate Metabolic Traits



Traits	Samples in WGS (4030)		All WES samples (19127)		WGS sample excluded/ independent samples (15131)	
	P	Beta	P	Beta	P	Beta
fatmass_combined	6.94×10^{-09}	-0.109	7.27×10^{-19}	-0.093	3.54×10^{-11}	-0.083
ln_ins_fast_combined	2.22×10^{-12}	-0.150	6.21×10^{-26}	-0.121	5.25×10^{-15}	-0.106
ln_P_ins120_combined	5.94×10^{-08}	-0.115	2.47×10^{-21}	-0.110	5.18×10^{-14}	-0.103

Results of EMMAX test of association between normalized MT-CN and both fat mass and two serum insulin related traits using WES data. Association tests were performed in all samples and also separately among samples with and without WGS data.

Chapter 3:

Genotyping Recurrent SVs in Loci with Complex Rearrangement Using Machine Learning Approach

3.1 Introduction

3.1.1 The limitation of applying traditional imputation methods to complex SVs

Using imputation methods to predict the missing genotypes has become a standard procedure for most GWAS studies, especially for those conducted by microarray technologies (McCarthy et al. 2016; Das et al. 2016; Naj 2019). However, the traditional genotyping imputation methods, such as BEAGLE5 (Browning, Zhou, and Browning 2018) used in chapter 1, were primarily designed and validated for SNPs, the genotype of which usually have relatively simple representation (samples either carry zero, one or two copies of the alternative alleles). Most of the commonly used imputation software started with a haplotype-based reference panel built from either pedigree data or large cohort of unrelated samples, and then predict the missing variants using Hidden Markov Model (HMM) or similar alternatives (Marchini and Howie 2010). A very important middle step of imputation is called "phasing", where the diploid genotype measured by either sequencing or array technology was converted into genotype of two haplotypes, during which various heuristic algorithms were applied to infer whether the variant was on paternal chromosome or maternal chromosome. This phasing step, however, can become extremely complicated when dealing with complex SVs with multiple possible allelic combinations, such as CNVs with highly variable copy number genotypes. Besides the technical limitation in phasing the complex multiallelic SVs, traditional imputation is also restricted to the assumption that the variants nearby (in the sense of linear genome) are more informative than the distal variants, under the general belief of low rate of mutation and recombination. This assumption in general works well for SNPs and indels, while not for complex SVs which were usually generated by mechanisms different from smaller and simpler variants. For instance, because of the frequent rearrangements, the LD structure of some SegDup-enriched loci was likely to be disturbed where some distal variants (or a specific combination of them) might in fact provide more information than the ones close to the SV of interests. Therefore, it would be beneficial to develop novel methods to predict the missing genotypes for complex SVs.

3.1.2 The copy number polymorphism (CNP) of *AMY1* locus and its potential phenotypic effect

AMY1 genes are a cluster of protein coding gene located on chromosome 1 responsible for generating salivary amylase. The diploid copy number of this gene, which has a wild range (2 to 20) of possible values in the population, was reported to have positive association with the digestion ability of starch when compared among human populations with low- versus high-starch diet(Perry et al. 2007). And the evidence became even stronger when the comparison was done between wolves (who have only one copy per individuals) and domesticate dogs (with diploid copy number ranged from 2~30), the latter of which were under the selection pressure of adapting to a starch-rich diet (Axelsson et al. 2013). After that, the *AMY1* copy number has been broadly studied for its potential role in other disease related metabolic phenotypes such as weight, BMI, obesity, type 2 diabetes etc. (Hariharan, Mousa, and de Courten 2021; Vázquez-Moreno et al. 2020; Liu et al. 2020; O'Callaghan et al. 2019; Barber et al. 2020; Marquina et al. 2019; Pinho, Padez, and Manco 2018; Viljakainen et al. 2015). However, just like the case in chapter 2, most of the time people tried to address these questions in clinical studies with limited sample sizes, and many ended up with the conflict results. One study(Usher et al. 2015) made the efforts to evaluate the phenotypic effect of *AMY1* copy number in large GWAS. Differed from their previous published work on the *HP* recurrent deletion(Boettger et al. 2016), where they extensively studied the allelic structure and existing haplotypes to build a reference panel for imputation and tested the imputed SV in a large cohort, this group chose to use the tagging SNP (of which the R^2 was only ~0.1 with the CNV) as the proxy for *AMY1* CNV, presumably because this variant was much more difficult for phasing even after similar efforts and pedigree data were invested to solve the allelic structure of this locus. At the end the author concluded that *AMY1* copy number has nothing to do with BMI or obesity, with the argument that the power for identifying any association signal was 99% even using the partially linked SNP given their large-enough sample size. Besides strategically comprised on the genotyping accuracy, this study also had a few other caveats: 1. Due to the phenotype availability, they only tested BMI and obesity, while *AMY1* might contribute to other metabolic traits, e.g. the ones available in our study. 2. Samples used in this study were from several small cohorts (each with less than two thousand individuals) with various ethnic background and disease status, which essentially had large heterogeneity that could affect meta-analysis (e.g., one cohort might be the major drive of the conclusion). Therefore, we sought to genotype this *AMY1* CNV with a novel machine learning approach that accommodated better to this type of problems and look for potential association of this recurrent SV with all the quantitative traits in our Finnish data.

3.1.3 Neural network models and their potential application in this case

Neural network is a type of machine learning models widely used today, the development of which were inspired by the structure of biological neural networks. It usually composed of an input layer, several hidden layers and an output layer, and the hidden layers are usually designed to have different number of nodes and connecting patterns, just like neurons in the brain, to mimic the learning processes of human beings. In most popular areas of computer science, such as computer vision and nature language processing(NLP), scientists are training neural network models with hundreds of hidden layers with complex architecture constructed by enormous numbers of parameters, as known as "deep learning", empowered by the recent breakthroughs in the size of training data as well as computational efficiency and capacity (LeCun, Bengio, and

Hinton 2015). However, the application of neural network and deep learning models has been pretty limited in the field of genetics, with a narrow focus on the functional prediction of the non-coding sequences (Telenti et al. 2018; Eraslan et al. 2019; Zou et al. 2019). The limited application was likely due to several different factors, including the lack of well-labeled genetics datasets, the general difficulty in interpreting the model (both the structure and the parameters), and probably the little improvement in the performance when compared to classic and simpler models. Meanwhile, this type of models are more flexible in terms of the underlying assumptions and the data format of inputs and outputs. Also as shown in the imaging processing applications, neural nets have the unique ability of automatically selecting informative features from the high dimensional space when fed with decent amount of high-quality training data (Razzak, Naz, and Zaib 2018). Considering the advantages and disadvantages mentioned above, the genotyping experiment of *AMY1* CNV (explained in chapter 3.1.2) became a good candidate for applying neural network models to a broader human genetics problems that were hard to solve by traditional approaches. And as illustrated below, we showed that even simple models with less than five fully-connect layers (the most basic architecture) had significantly outperformed the previous best practice (tagging SNP proxy) and the baseline single-layer linear separators. This experiment is still ongoing, with efforts denoted to refine the performance of the current model. However, given how encouraging the preliminary results were, we are already in the progress of introducing the same method to other complex SV loci. In the future research, we also look forward to applying deep learning in other types of genetics problems where more and more high-quality medical and biological datasets are available.

3.2 Methods

3.2.1 Data preparation

4441 WGS-sequenced samples passed quality control for both SV callsets and GATK callset were used for training and testing the models. A subset of 406 SNPs/Indels flanking *AMY1* genes on chr1:102561901-104594600 (GRCh38) captured by both WGS data and genotype array data were extracted from the GATK callset VCF to create feature matrices. We designed two types of features using the alternative allele dosage, one from the unphased diploid genome and the other from phased haploid genome (see chapter 1.2.5). To determine the labels, we manually curated the CNVnator calls overlapped with *AMY1* and arbitrarily selected the CNV chr1:103594101-103722400, of which the genotypes could be easily defined from the copy number distribution (**Figure 3.1**). Similarly, we designed two sets of labels: 1. the continuous genotype from CNVnator output and 2. The discrete genotype from clustering the continuous values in 1. All four combinations of the features and labels were trained and tested for performances.

3.2.2 Model training and evaluation

For this preliminary analysis, all the model training and evaluations were done using tensorflow/keras(Chollet 2015), a python based deep learning API, on one laptop with 16GB memory and 6-core i9 processor. The 4441 samples were randomly separated, with 80% went into the training set and 20% went into the testing set. For models with categorical output (discrete genotype label), we used cross entropy loss function and aimed at maximizing accuracy, and mean squared error as the loss function for regression output(continuous genotype label). All models were trained with Adam optimizer for 5000 epochs with entire set of training samples. To set up the baseline performance, we first trained a simple regression model/linear separator for each group of feature and label combinations (for models with categorical outputs, an additional softmax activation function was added to convert the outputs variables). Next, we experimentally built up the neural network model starting from the simplest structure (e.g. one hidden fully connected layer with 32 nodes followed by a relu activation function) and gradually increased the complexity (adding more layers and/or nodes) until there was indication of overfitting: the improvement of the training accuracy almost saturated and the testing accuracy started to drop. Then we added regularization (L2 regularizers with factor of 0.01) and dropout layers (rate = 20%) to reduce overfitting. The final models was selected according to the training and testing performance evaluated by the correlation between the predicted genotype and WGS-measured genotype.

3.2.3 Predict AMY1 copy number in 19k Finns and association tests

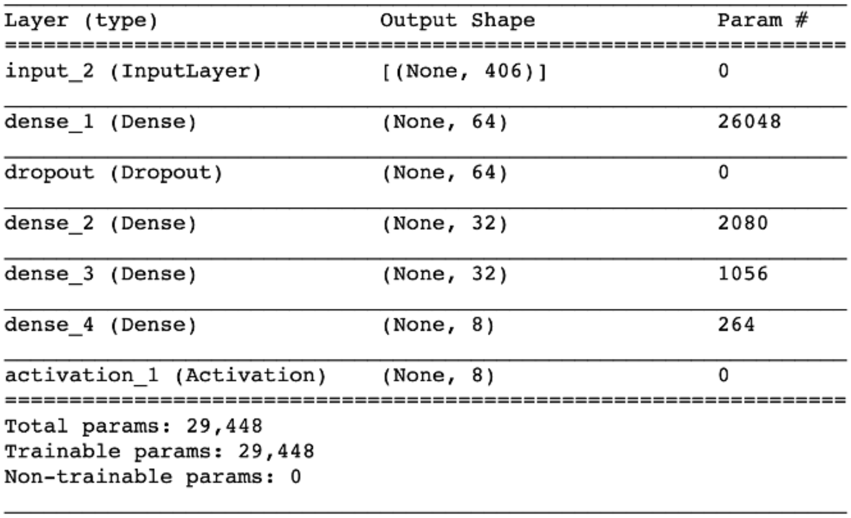
In order to better predict the "unseen" data, we first retrained the model using full set of the 4441 labeled samples with the final structure determined in chapter 3.2.2. Then the array genotype data of 21,058 Finnish samples was then input to predict the *AMY1* copy number of those individuals. We estimated the overall prediction quality from the overlapping samples between WGS and array datasets and then tested the predicted *AMY1* genotypes against 140 quantitative metabolic traits (24 additional traits added according to the same power cut-off used in chapter 1) via EMMAX model.

3.3 Results

3.3.1 Model performance

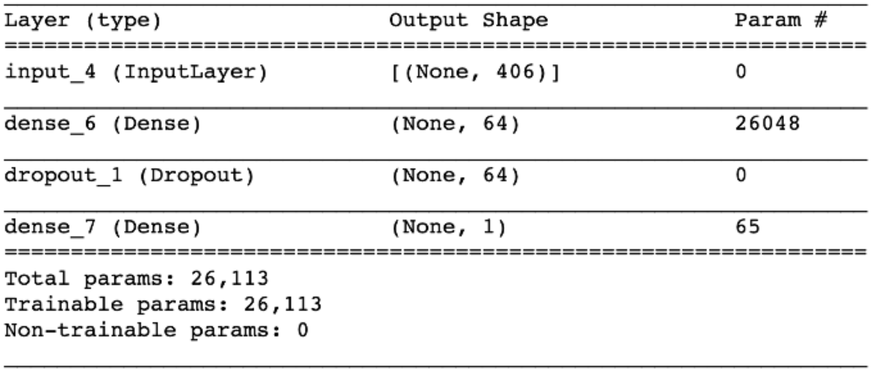
The final models for each combination are listed below:

Model #1: discrete labels ~ unphased features



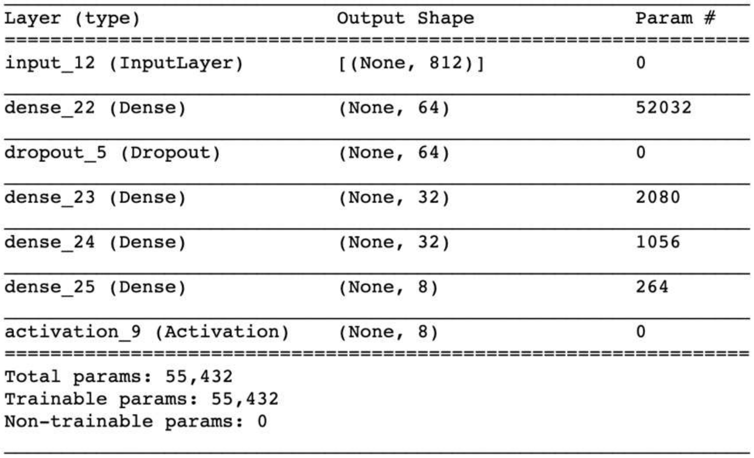
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
input_2 (InputLayer)         [(None, 406)]             0         
_________________________________________________________________
dense_1 (Dense)              (None, 64)                26048     
_________________________________________________________________
dropout (Dropout)            (None, 64)                0         
_________________________________________________________________
dense_2 (Dense)              (None, 32)                2080      
_________________________________________________________________
dense_3 (Dense)              (None, 32)                1056      
_________________________________________________________________
dense_4 (Dense)              (None, 8)                 264       
_________________________________________________________________
activation_1 (Activation)    (None, 8)                 0         
=================================================================
Total params: 29,448
Trainable params: 29,448
Non-trainable params: 0
_________________________________________________________________
```

Model #2: continuous labels ~ unphased features



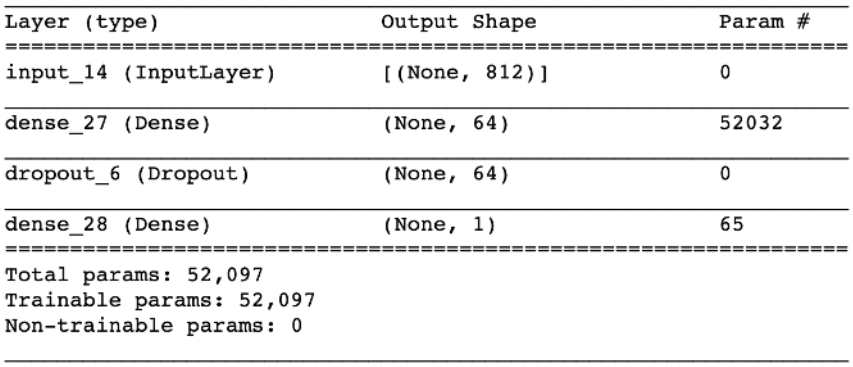
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
input_4 (InputLayer)         [(None, 406)]             0         
_________________________________________________________________
dense_6 (Dense)              (None, 64)                26048     
_________________________________________________________________
dropout_1 (Dropout)          (None, 64)                0         
_________________________________________________________________
dense_7 (Dense)              (None, 1)                 65        
=================================================================
Total params: 26,113
Trainable params: 26,113
Non-trainable params: 0
_________________________________________________________________
```

Model #3: discrete labels ~ phased features



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
input_12 (InputLayer)        [(None, 812)]             0         
_________________________________________________________________
dense_22 (Dense)             (None, 64)                52032     
_________________________________________________________________
dropout_5 (Dropout)          (None, 64)                0         
_________________________________________________________________
dense_23 (Dense)             (None, 32)                2080      
_________________________________________________________________
dense_24 (Dense)             (None, 32)                1056      
_________________________________________________________________
dense_25 (Dense)             (None, 8)                 264       
_________________________________________________________________
activation_9 (Activation)    (None, 8)                 0         
=================================================================
Total params: 55,432
Trainable params: 55,432
Non-trainable params: 0
_________________________________________________________________
```

Model #4: continuous labels ~ phased feature



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
input_14 (InputLayer)        [(None, 812)]             0         
_________________________________________________________________
dense_27 (Dense)             (None, 64)                52032     
_________________________________________________________________
dropout_6 (Dropout)          (None, 64)                0         
_________________________________________________________________
dense_28 (Dense)             (None, 1)                 65        
=================================================================
Total params: 52,097
Trainable params: 52,097
Non-trainable params: 0
_________________________________________________________________
```

And the performances of above models can be found in **Table 3.1**. To summarize: all the multi-layer neural net models outperformed the baseline models in both training data and testing data, and the performances were similar across different selections of features and labels. Given this observation and the principle of Occam's razor, we selected model #2 for the downstream analysis, as it had the simplest structure.

3.3.2 Predicted AMY1 CNP in Finns and its association with metabolic traits

We retrained the parameters of model #2 using all the labeled samples and predicted the *AMY1* copy number genotype of the ~20k samples using the updated model and the feature matrix created from SNP array data. The correlation between predicted copy number and the "ground truth" was 0.97 estimated in the ~4k overlapped samples. This approximated accuracy was presumably overestimated, since the model performance tend to drop when making predictions on the rest of the unseen data. To get a sense of the overall performance, we compared the distribution of the predicted values to the distribution of WGS-measured copy number (**Figure 3.2**). The predicted distribution also fit into a GMM, with each component centered around the same mean compared to the labeled distribution (the variant increased probably as a result of more diverse input haplotypes). We then tested for association between the predicted *AMY1* copy number and the 140 quantitative traits (**Figure 3.3**), and found five significant signals with p <0.05: serum LDL cholesterol level in semi-fasting samples (p = 0.012, beta = 0.052, n = 4602), alanine (p=0.023, beta = 0.034, n=8699), pyruvate (p=0.031, beta = 0.029, n=10743), HDL3 cholesterol (p=0.034, beta=0.028, n=10826) and glycine (p=0.042, beat = 0.031, n=8155). None of those five signals was strong enough to pass multiple testing correction for the tested phenotypes. Meanwhile, previous studies suggested a potential correlation between LDL cholesterol and acute pancreatitis(Hong et al. 2018, 2020; Ni et al. 2014), with one study observed a positive correlation between amylase level and LDL cholesterol in the blood and in the urine(Ni et al. 2014). And the observation that the signal only appeared in the semi-fasting samples was also interesting given the role of amylase in digestion activities.

3.4 Discussion

We proved in chapter 1 that SVs could have large phenotypic effect and once included in the trait-mapping studies would improve our understanding of disease genetics. However, this type of variants are also challenging to detect in large scale, since a high quality SV callset either depends on expensive and time-consuming collection of WGS data or requires highly conservative quality control for WES or array data at the price of missing out a large number of true variants. An alternative solution to that, which was already employed in our first study, is to genotype the high-confidence SVs detected from a relatively small set of WGS samples in a larger dataset with targeted sequencing (e.g., WES) or SNP array data. This approach boosted the power of our trait-mapping study, and also provided a much more “affordable” solution for routinely including SVs in GWAS studies -- simply adding those variants in the imputation panels.

However, since the classic imputation methods were only good at handling part of biallelic SVs, we still missed a fair number of complex SVs with potential phenotypic effects, such as the *AMY1* CNV. In this chapter, we delivered a new strategy for imputing the complex SVs from

SNP array data, with the application of the neural network models. And we showed that models with relatively simple structures already performed surprisingly well.

Given this proof-of-concept, the future directions of this project are:

1. Building more complex models to see if we can further improve the predicting accuracy. This could be done by either adding more hidden layers ("go deeper") or trying out other architecture such as convolutional neural network (CNN), recurrent neural network (RNN) and long short-term memory (LSTM). Of particular interests are the sequential models, given the linear nature of the DNA sequences. However, as the model becomes more sophisticated, the current size of well-labeled data for training and testing is not likely to be enough for the drastically increased parameters. As a consequence, a lot of efforts will be denoted to prevent overfitting, while the improvement on performance might be minor. Ideally we can reach a good balance of method development and application efficiency through practice and getting inputs from other machine learning experts. For example, we might find inspiration from one recently published study in which the authors also applied deep learning method to improve imputation (Kojima et al. 2020).

2. Expanding the application to other unsolved problems. A lab member Neil Zheng is currently working on experimenting the same approaches to a few pharmacogenomics loci known to be affected by recurrent SVs, including *CCL3L1*(Carpenter et al. 2012), *C4*(Szilágyi and Fust 2008), *CYP2D6*(Nofziger et al. 2020), and *DEFB4A*(Bentley et al. 2010; X.-J. Zhou et al. 2012). At the same time, he is also retraining the *AMY1* CNV model with additional samples from other ethnicity groups (e.g., including the 1000 Genome Project samples), so that we can expand the analysis to datasets like UK Biobank, where we also have the access to a lot more samples as and phenotypes. Besides genotyping SVs, we are also considering other biological applications of

deep learning technology, such as solving the structure of human genome and predicting functional and phenotypic effects from genotype data.

3.5 Acknowledgement

For the work in this chapter, I would like to acknowledge my thesis advisor Ira for supporting me to develop the initial idea, reimbursing my online training for the new skillset, and allowing me two months of the freedom to explore and implement, during which he could have me assigned to projects with higher chance for positive results (while less fun, presumably). I would also like to acknowledge Neil for continuing working on this project and hope he will carry it to the next stage and eventually find it rewarding. My appreciation also goes to Andrew Ng and other instructors who designed the Deep Learning Specialization on Coursera.org, thanks to which I quickly acquired the knowledge and technology I needed for implementing this research idea and will keep benefit from it in my future career. Finally, I want to acknowledge all the useful comments and suggestions from my committee members, especially those technical ones from Nan and Nancy.

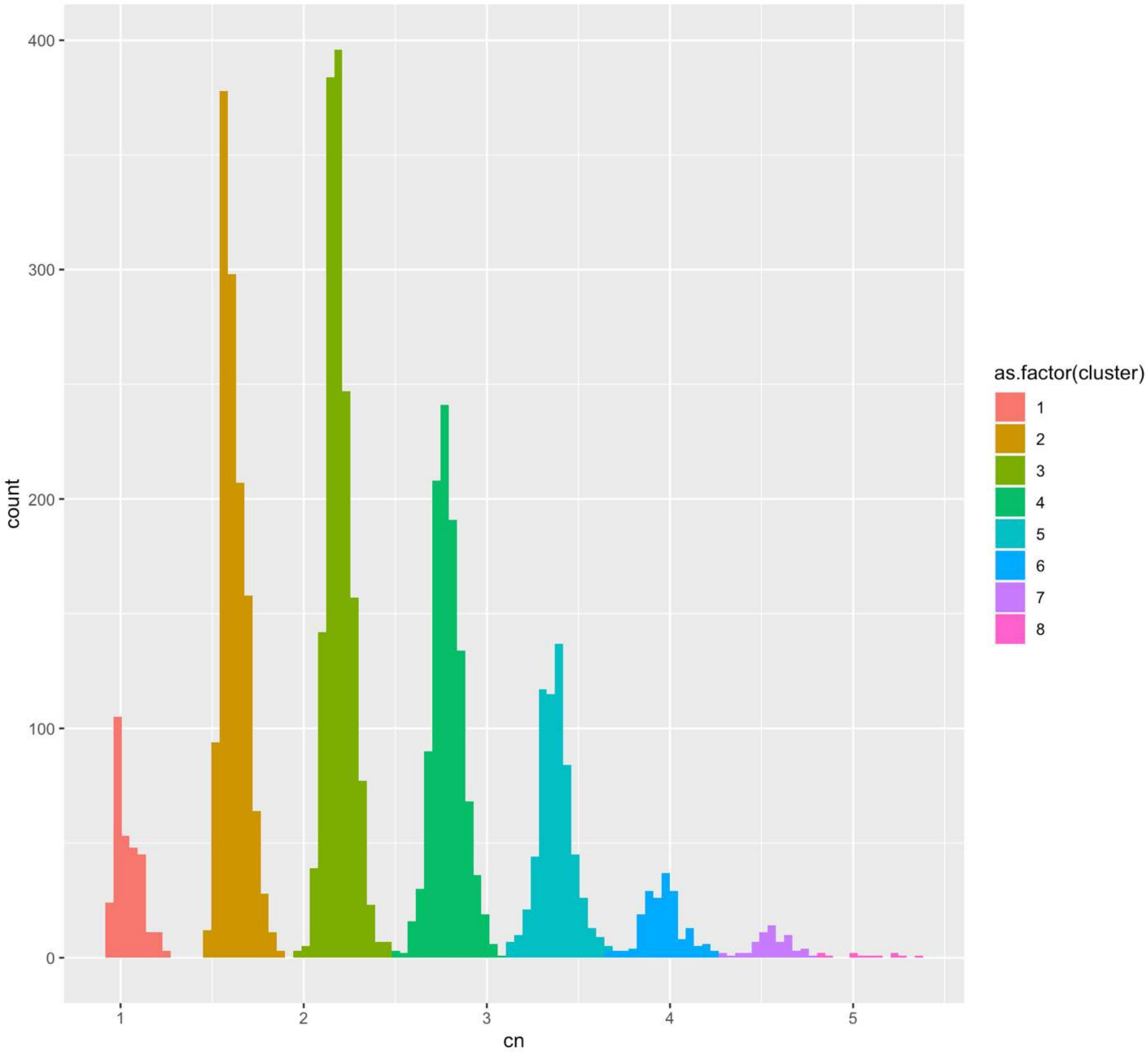


Figure 3.1. The two types of copy number labels

This figure illustrated the two types of the copy number genotype labels we used for training the models and their relationship. The histogram was plotted for the continuous copy number labels estimated by CNVnator, and the color code showed the discrete copy number labels which were the clustering results of the continuous values using k-means method (k=8).

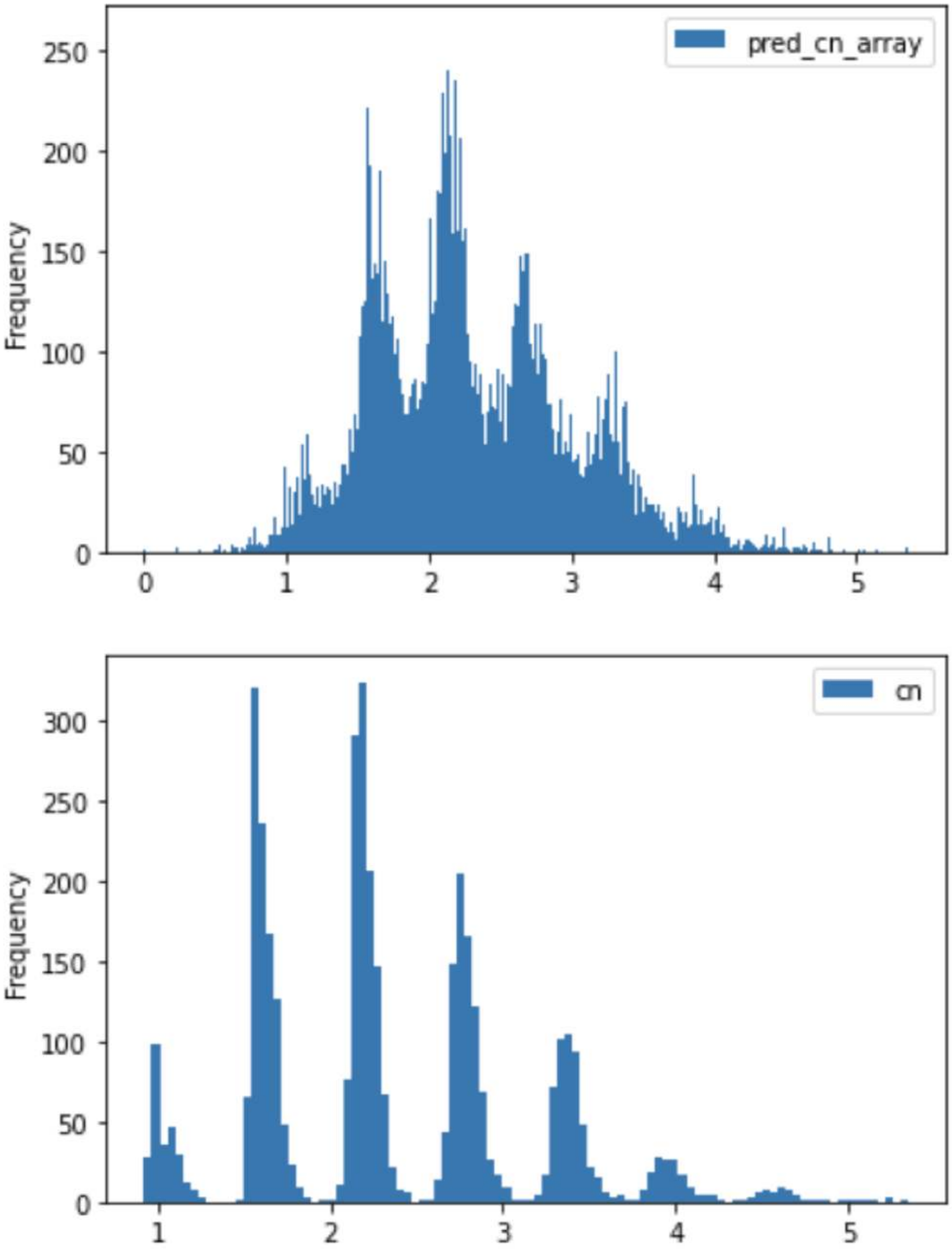


Figure 3.2. The distribution of predicted and measured *AMY1* copy number

The histogram of continuous copy number genotype predicted by the final model (top, 20k samples) and that of the directly measured genotype from WGS data (bottom, 4k samples).

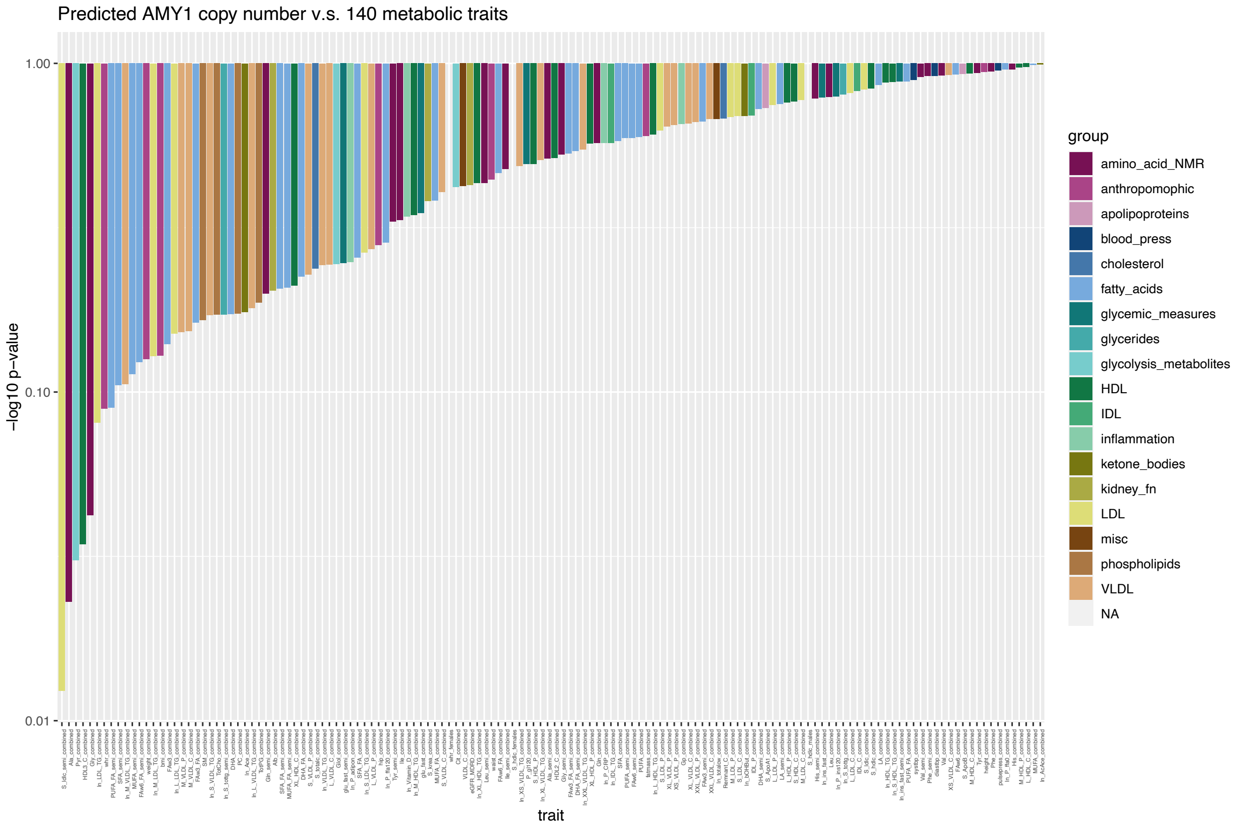
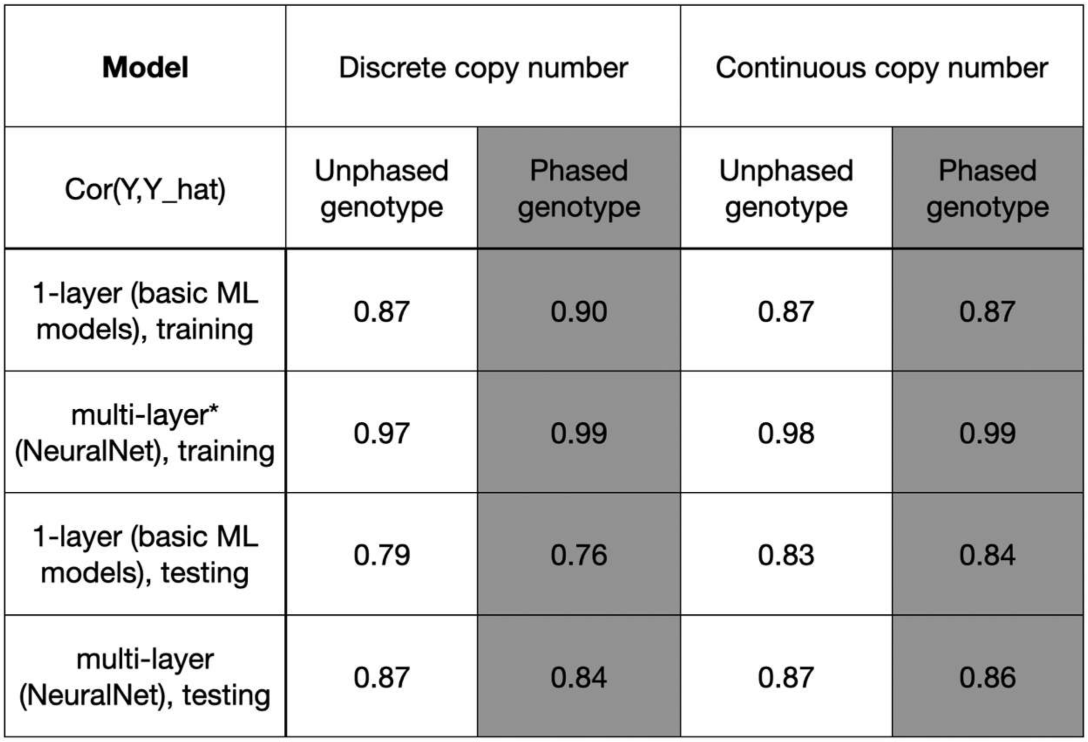


Figure 3.3. The phenome-wide association results for predicted *AMY1* copy number

The phenome-wide Manhattan plot for the predicted *AMY1* copy number variation, colors were coded by trait groups. P value on the y-axis was plotted in negative log10 scale, to better illustrate the significance level of the first five traits (p<0.05).

Table 3.1. Model performance



Model	Discrete copy number		Continuous copy number	
Cor(Y,Y_hat)	Unphased genotype	Phased genotype	Unphased genotype	Phased genotype
1-layer (basic ML models), training	0.87	0.90	0.87	0.87
multi-layer* (NeuralNet), training	0.97	0.99	0.98	0.99
1-layer (basic ML models), testing	0.79	0.76	0.83	0.84
multi-layer (NeuralNet), testing	0.87	0.84	0.87	0.86

* Different structure for the optimal categorical model and regression model

The model performance measured by the Pearson correlation coefficient of the true label (Y) and predicted label (Y_hat). Only the final selected models are presented here of which the structure of each model can be find in the chapter 3.3.1.

References/Bibliography/Works Cited

Abel, Haley J., David E. Larson, Allison A. Regier, Colby Chiang, Indraniel Das, Krishna L. Kanchi, Ryan M. Layer, et al. 2020. “Mapping and Characterization of Structural Variation in 17,795 Human Genomes.” *Nature*, May. https://doi.org/10.1038/s41586-020-2371-0.

Abyzov, Alexej, Alexander E. Urban, Michael Snyder, and Mark Gerstein. 2011. “CNVnator: An Approach to Discover, Genotype, and Characterize Typical and Atypical CNVs from Family and Population Genome Sequencing.” *Genome Research* 21 (6): 974–84. https://doi.org/10.1101/gr.114876.110.

Aguirre, Matthew, Manuel A. Rivas, and James Priest. 2019. “Phenome-Wide Burden of Copy-Number Variation in the UK Biobank.” *American Journal of Human Genetics* 105 (2): 373–83. https://doi.org/10.1016/j.ajhg.2019.07.001.

Auwera, Geraldine A., Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. “From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline.” *Current Protocols in Bioinformatics*. https://doi.org/10.1002/0471250953.bi1110s43.

Axelsson, Erik, Abhirami Ratnakumar, Maja-Louise Arendt, Khurram Maqbool, Matthew T. Webster, Michele Perloski, Olof Liberg, Jon M. Arnemo, Ake Hedhammar, and Kerstin Lindblad-Toh. 2013. “The Genomic Signature of Dog Domestication Reveals Adaptation to a Starch-Rich Diet.” *Nature* 495 (7441): 360–64. https://doi.org/10.1038/nature11837.

Barber, Thomas M., Ahsan A. Bhatti, Patrick J. D. Elder, Sarah P. Ball, Ronan Calvez, David B. Ramsden, Dan J. Cuthbertson, Andreas F. Pfeiffer, David Burnett, and Martin O. Weickert. 2020. “AMY1 Gene Copy Number Correlates with Glucose Absorption and Visceral Fat Volume, but Not with Insulin Resistance.” *The Journal of Clinical Endocrinology and Metabolism* 105 (10): e3586–96. https://doi.org/10.1210/clinem/dgaa473.

Bentley, Robert W., John Pearson, Richard B. Gearry, Murray L. Barclay, Cushla McKinney, Tony R. Merriman, and Rebecca L. Roberts. 2010. “Association of Higher DEFB4 Genomic Copy Number with Crohn’s Disease.” *The American Journal of Gastroenterology* 105 (2): 354–59. https://doi.org/10.1038/ajg.2009.582.

Berkson, Joseph. 1946. “Limitations of the Application of Fourfold Table Analysis to Hospital Data.” *Biometrics Bulletin*. https://doi.org/10.2307/3002000.

Boettger, Linda M., Rany M. Salem, Robert E. Handsaker, Gina M. Peloso, Sekar Kathiresan, Joel N. Hirschhorn, and Steven A. McCarroll. 2016. “Recurring Exon Deletions in the HP (Haptoglobin) Gene Contribute to Lower Blood Cholesterol Levels.” *Nature Genetics* 48 (4): 359–66. https://doi.org/10.1038/ng.3510.

Browning, Brian L., Ying Zhou, and Sharon R. Browning. 2018. “A One-Penny Imputed Genome from Next-Generation Reference Panels.” *American Journal of Human Genetics* 103 (3): 338–48. https://doi.org/10.1016/j.ajhg.2018.07.015.

Buniello, Annalisa, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. “The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary

Statistics 2019.” *Nucleic Acids Research* 47 (D1): D1005–12. https://doi.org/10.1093/nar/gky1120.

Campbell Am, Lesley V. 2017. “Genetics of Obesity.” *Australian Family Physician* 46 (7): 456–59. https://www.ncbi.nlm.nih.gov/pubmed/28697287.

Cantsilieris, Stuart, Susan M. Sunkin, Matthew E. Johnson, Fabio Anaclerio, John Huddleston, Carl Baker, Max L. Dougherty, et al. 2020. “An Evolutionary Driver of Interspersed Segmental Duplications in Primates.” *Genome Biology* 21 (1): 202. https://doi.org/10.1186/s13059-020-02074-4.

Carpenter, Danielle, Anna Färnert, Ingegerd Rooth, John A. L. Armour, and Marie-Anne Shaw. 2012. “CCL3L1 Copy Number and Susceptibility to Malaria.” *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 12 (5): 1147–54. https://doi.org/10.1016/j.meegid.2012.03.021.

Chen, Shuying, Xuejian Xie, Yujing Wang, Yan Gao, Xiaoli Xie, Jing Yang, and Jixian Ye. 2014. “Association between Leukocyte Mitochondrial DNA Content and Risk of Coronary Heart Disease: A Case-Control Study.” *Atherosclerosis* 237 (1): 220–26. https://doi.org/10.1016/j.atherosclerosis.2014.08.051.

Chiang, Colby, Ryan M. Layer, Gregory G. Faust, Michael R. Lindberg, David B. Rose, Erik P. Garrison, Gabor T. Marth, Aaron R. Quinlan, and Ira M. Hall. 2015. “SpeedSeq: Ultra-Fast Personal Genome Analysis and Interpretation.” *Nature Methods* 12 (10): 966–68. https://doi.org/10.1038/nmeth.3505.

Chiang, Colby, Alexandra J. Scott, Joe R. Davis, Emily K. Tsang, Xin Li, Yungil Kim, Tarik Hadzic, et al. 2017. “The Impact of Structural Variation on Human Gene Expression.” *Nature Publishing Group* 49 (5): 692–99. https://doi.org/10.1038/ng.3834.

Chollet, François. n.d. *Keras*. Github. Accessed May 21, 2021. https://github.com/keras-team/keras.

Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, et al. 2020. “A Structural Variation Reference for Medical and Population Genetics.” *Nature* 581 (7809): 444–51. https://doi.org/10.1038/s41586-020-2287-8.

Das, Sayantan, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E. Locke, Alan Kwong, Scott I. Vrieze, et al. 2016. “Next-Generation Genotype Imputation Service and Methods.” *Nature Genetics* 48 (10): 1284–87. https://doi.org/10.1038/ng.3656.

Davis, James P., Jeroen R. Huyghe, Adam E. Locke, Anne U. Jackson, Xueling Sim, Heather M. Stringham, Tanya M. Teslovich, et al. 2017. “Common, Low-Frequency, and Rare Genetic Variants Associated with Lipoprotein Subclasses and Triglyceride Measures in Finnish Men from the METSIM Study.” *PLoS Genetics* 13 (10): e1007079. https://doi.org/10.1371/journal.pgen.1007079.

Ding, Jun, Carlo Sidore, Thomas J. Butler, Mary Kate Wing, Yong Qian, Osorio Meirelles, Fabio Busonero, et al. 2015. “Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools.” *PLoS Genetics* 11 (7): e1005306. https://doi.org/10.1371/journal.pgen.1005306.

ENCODE Project Consortium. 2004. “The ENCODE (ENCyclopedia Of DNA Elements) Project.” *Science* 306 (5696): 636–40. https://doi.org/10.1126/science.1105136.

Eraslan, Gökcen, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. 2019. “Deep Learning: New Computational Modelling Techniques for Genomics.” *Nature Reviews. Genetics* 20 (7): 389–403. https://doi.org/10.1038/s41576-019-0122-6.

Fall, Tove, and Erik Ingelsson. 2014. “Genome-Wide Association Studies of Obesity and Metabolic Syndrome.” *Molecular and Cellular Endocrinology* 382 (1): 740–57. https://doi.org/10.1016/j.mce.2012.08.018.

Fromer, Menachem, and Shaun M. Purcell. 2014. “Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data.” *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]* 81 (April): 7.23.1-21. https://doi.org/10.1002/0471142905.hg0723s81.

Gahan, M. E., F. Miller, S. R. Lewin, C. L. Cherry, J. F. Hoy, A. Mijch, F. Rosenfeldt, and S. L. Wesselingh. 2001. “Quantification of Mitochondrial DNA in Peripheral Blood Mononuclear Cells and Subcutaneous Fat Using Real-Time Polymerase Chain Reaction.” *Journal of Clinical Virology: The Official Publication of the Pan American Society for Clinical Virology* 22 (3): 241–47. https://doi.org/10.1016/s1386-6532(01)00195-0.

Gourlain, K., B. Amellal, Z. Ait Arkoub, N. Dupin, C. Katlama, and V. Calvez. 2003. “Quantitative Analysis of Human Mitochondrial DNA Using a Real-Time PCR Assay.” *HIV Medicine* 4 (3): 287–92. https://doi.org/10.1046/j.1468-1293.2003.00158.x.

Guyatt, Anna L., Kimberley Burrows, Philip A. I. Guthrie, Sue Ring, Wendy McArdle, Ian N. M. Day, Raimondo Ascione, Debbie A. Lawlor, Tom R. Gaunt, and Santiago Rodriguez. 2018. “Cardiometabolic Phenotypes and Mitochondrial DNA Copy Number in Two Cohorts of UK Women.” *Mitochondrion* 39 (March): 9–19. https://doi.org/10.1016/j.mito.2017.08.007.

Hagenbeek, Fiona A., René Pool, Jenny van Dongen, Harmen H. M. Draisma, Jouke Jan Hottenga, Gonneke Willemsen, Abdel Abdellaoui, et al. 2020. “Heritability Estimates for 361 Blood Metabolites across 40 Genome-Wide Association Studies.” *Nature Communications* 11 (1): 39. https://doi.org/10.1038/s41467-019-13770-6.

Han, Buhm, and Eleazar Eskin. 2011. “Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-Wide Association Studies.” *American Journal of Human Genetics* 88 (5): 586–98. https://doi.org/10.1016/j.ajhg.2011.04.014.

Handsaker, Robert E., Vanessa Van Doren, Jennifer R. Berman, Giulio Genovese, Seva Kashin, Linda M. Boettger, and Steven A. McCarroll. 2015. “Large Multiallelic Copy Number Variations in Humans.” *Nature Genetics* 47 (3): 296–303. https://doi.org/10.1038/ng.3200.

Hariharan, Rohit, Aya Mousa, and Barbora de Courten. 2021. “Influence of AMY1A Copy Number Variations on Obesity and Other Cardiometabolic Risk Factors: A Review of the Evidence.” *Obesity Reviews: An Official Journal of the International Association for the Study of Obesity*, no. obr.13205 (January). https://doi.org/10.1111/obr.13205.

Hegele, Robert A., and Rebecca L. Pollex. 2005. “Genetic and Physiological Insights into the Metabolic Syndrome.” *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology* 289 (3): R663-9. https://doi.org/10.1152/ajpregu.00275.2005.

Hoaglin, David C., and Boris Iglewicz. 1987. “Fine-Tuning Some Resistant Rules for Outlier Labeling.” *Journal of the American Statistical Association* 82 (400): 1147. https://doi.org/10.2307/2289392.

Hong, Wandong, Vincent Zimmer, Zarrin Basharat, Maddalena Zippi, Simon Stock, Wujun Geng, Xueqin Bao, Junfeng Dong, Jingye Pan, and Mengtao Zhou. 2020. “Association of Total Cholesterol with Severe Acute Pancreatitis: A U-Shaped Relationship.” *Clinical Nutrition (Edinburgh, Scotland)* 39 (1): 250–57. https://doi.org/10.1016/j.clnu.2019.01.022.

Hong, Wandong, Vincent Zimmer, Simon Stock, Maddalena Zippi, Jones A. Q. Omoshoro-Jones, and Mengtao Zhou. 2018. "Relationship between Low-Density Lipoprotein Cholesterol and Severe Acute Pancreatitis ('the Lipid Paradox')." *Therapeutics and Clinical Risk Management* 14 (May): 981–89. https://doi.org/10.2147/tcrm.s159387.

Hormozdiari, Farhad, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. 2014. "Identifying Causal Variants at Loci with Multiple Signals of Association." *Genetics* 198 (2): 497–508. https://doi.org/10.1534/genetics.114.167908.

Inouye, Michael, Samuli Ripatti, Johannes Kettunen, Leo-Pekka Lyytikäinen, Niku Oksala, Pirkka-Pekka Laurila, Antti J. Kangas, et al. 2012. "Novel Loci for Metabolic Networks and Multi-Tissue Expression Studies Reveal Genes for Atherosclerosis." *PLoS Genetics* 8 (8): e1002907. https://doi.org/10.1371/journal.pgen.1002907.

Jacobs, P. 1959. "THE SOMATIC CHROMOSOMES IN MONGOLISM." *The Lancet*. https://doi.org/10.1016/s0140-6736(59)91892-6.

Jiang, Zhaoshi, Haixu Tang, Mario Ventura, Maria Francesca Cardone, Tomas Marques-Bonet, Xinwei She, Pavel A. Pevzner, and Evan E. Eichler. 2007. "Ancestral Reconstruction of Segmental Duplications Reveals Punctuated Cores of Human Genome Evolution." *Nature Genetics* 39 (11): 1361–68. https://doi.org/10.1038/ng.2007.9.

Johnson, Matthew E., National Institute of Health Intramural Sequencing Center Comparative Sequencing Program, Ze Cheng, V. Anne Morrison, Steven Scherer, Mario Ventura, Richard A. Gibbs, Eric D. Green, and Evan E. Eichler. 2006. "Recurrent Duplication-Driven Transposition of DNA during Hominoid Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 103 (47): 17626–31. https://doi.org/10.1073/pnas.0605426103.

Kang, Hyun Min, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-Yee Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. 2010. "Variance Component Model to Account for Sample Structure in Genome-Wide Association Studies." *Nature Genetics* 42 (4): 348–54. https://doi.org/10.1038/ng.548.

Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, et al. 2003. "The UCSC Genome Browser Database." *Nucleic Acids Research* 31 (1): 51–54. https://doi.org/10.1093/nar/gkg129.

Kent, W. J. 2002. "BLAT---the BLAST-like Alignment Tool." *Genome Research* 12 (4): 656–64. https://doi.org/10.1101/gr.229202.

Kettunen, Johannes, Ayşe Demirkan, Peter Würtz, Harmen H. M. Draisma, Toomas Haller, Rajesh Rawal, Anika Vaarhorst, et al. 2016. "Genome-Wide Study for Circulating Metabolites Identifies 62 Loci and Reveals Novel Systemic Effects of LPA." *Nature Communications* 7 (March): 11122. https://doi.org/10.1038/ncomms11122.

Kettunen, Johannes, Taru Tukiainen, Antti-Pekka Sarin, Alfredo Ortega-Alonso, Emmi Tikkanen, Leo-Pekka Lyytikäinen, Antti J. Kangas, et al. 2012. "Genome-Wide Association Study Identifies Multiple Loci Influencing Human Serum Metabolite Levels." *Nature Genetics* 44 (3): 269–76. https://doi.org/10.1038/ng.1073.

Kim, Youngdoe, Young Lee, Sungyoung Lee, Nam Hee Kim, Jeongmin Lim, Young Jin Kim, Ji Hee Oh, et al. 2015. "On the Estimation of Heritability with Family-Based and Population-Based Samples." *BioMed Research International* 2015 (August): 671349. https://doi.org/10.1155/2015/671349.

Kojima, Kaname, Shu Tadaka, Fumiki Katsuoka, Gen Tamiya, Masayuki Yamamoto, and Kengo Kinoshita. 2020. "A Genotype Imputation Method for De-Identified Haplotype Reference

Information by Using Recurrent Neural Network." *PLoS Computational Biology* 16 (10): e1008207. https://doi.org/10.1371/journal.pcbi.1008207.

Kolifarhood, Goodarz, Maryam Daneshpour, Farzad Hadaegh, Siamak Sabour, Hossein Mozafar Saadati, Ali Akbar Haghdoust, Mahdi Akbarzadeh, Bahareh Sedaghati-Khayat, and Nasim Khosravi. 2019. "Heritability of Blood Pressure Traits in Diverse Populations: A Systematic Review and Meta-Analysis." *Journal of Human Hypertension* 33 (11): 775–85. https://doi.org/10.1038/s41371-019-0253-4.

Larson, David E., Haley J. Abel, Colby Chiang, Abhijit Badve, Indraniel Das, James M. Eldred, Ryan M. Layer, and Ira M. Hall. 2019. "Svtools: Population-Scale Analysis of Structural Variation." *Bioinformatics* 35 (22): 4782–87. https://doi.org/10.1093/bioinformatics/btz492.

Layer, Ryan M., Colby Chiang, Aaron R. Quinlan, and Ira M. Hall. 2014. "LUMPY: A Probabilistic Framework for Structural Variant Discovery." *Genome Biology* 15 (6): R84. https://doi.org/10.1186/gb-2014-15-6-r84.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Lee, H. K., J. H. Song, C. S. Shin, D. J. Park, K. S. Park, K. U. Lee, and C. S. Koh. 1998. "Decreased Mitochondrial DNA Content in Peripheral Blood Precedes the Development of Non-Insulin-Dependent Diabetes Mellitus." *Diabetes Research and Clinical Practice* 42 (3): 161–67. https://doi.org/10.1016/s0168-8227(98)00110-7.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics (Oxford, England)* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, J., and L. Ji. 2005. "Adjusting Multiple Testing in Multilocus Analyses Using the Eigenvalues of a Correlation Matrix." *Heredity* 95 (3): 221–27. https://doi.org/10.1038/sj.hdy.6800717.

Li, Xin, Yungil Kim, Emily K. Tsang, Joe R. Davis, Farhan N. Damani, Colby Chiang, Gaelen T. Hess, et al. 2017. "The Impact of Rare Variation on Gene Expression across Tissues." *Nature* 550 (7675): 239–43. https://doi.org/10.1038/nature24267.

Li, Yun Rose, Joseph T. Glessner, Bradley P. Coe, Jin Li, Maede Mohebnasab, Xiao Chang, John Connolly, et al. 2020. "Rare Copy Number Variants in over 100,000 European Ancestry Subjects Reveal Multiple Disease Associations." *Nature Communications* 11 (1): 255. https://doi.org/10.1038/s41467-019-13624-1.

Lim, Elaine T., Peter Würtz, Aki S. Havulinna, Priit Palta, Taru Tukiainen, Karola Rehnström, Tõnu Esko, et al. 2014. "Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population." *PLoS Genetics* 10 (7): e1004494. https://doi.org/10.1371/journal.pgen.1004494.

Liu, Yuwei, Caren E. Smith, Laurence D. Parnell, Yu-Chi Lee, Ping An, Robert J. Straka, Hemant K. Tiwari, et al. 2020. "Salivary AMY1 Copy Number Variation Modifies Age-Related Type 2 Diabetes Risk." *Clinical Chemistry* 66 (5): 718–26. https://doi.org/10.1093/clinchem/hvaa072.

Locke, Adam E., Karyn Meltz Steinberg, Charleston W. K. Chiang, Susan K. Service, Aki S. Havulinna, Laurel Stell, Matti Pirinen, et al. 2019. "Exome Sequencing of Finnish Isolates Enhances Rare-Variant Association Power." *Nature* 572 (7769): 323–28. https://doi.org/10.1038/s41586-019-1457-z.

Lopez, J., J. C. Stephens, and S. J. O'Brien. 1997. "The Long and Short of Nuclear Mitochondrial DNA (Numt) Lineages." *Trends in Ecology & Evolution* 12 (3): 114. https://doi.org/10.1016/s0169-5347(97)84925-7.

Lopez, J. V., N. Yuhki, R. Masuda, W. Modi, and S. J. O'Brien. 1994. "Numt, a Recent Transfer and Tandem Amplification of Mitochondrial DNA to the Nuclear Genome of the Domestic Cat." *Journal of Molecular Evolution* 39 (2): 174–90. https://doi.org/10.1007/bf00163806.

Macé, Aurélien, Marcus A. Tuke, Patrick Deelen, Kati Kristiansson, Hannele Mattsson, Margit Nõukas, Yadav Sapkota, et al. 2017. "CNV-Association Meta-Analysis in 191,161 European Adults Reveals New Loci Associated with Anthropometric Traits." *Nature Communications* 8 (1): 744. https://doi.org/10.1038/s41467-017-00556-x.

Marchini, Jonathan, and Bryan Howie. 2010. "Genotype Imputation for Genome-Wide Association Studies." *Nature Reviews. Genetics* 11 (7): 499–511. https://doi.org/10.1038/nrg2796.

Marquina, Clara, Aya Mousa, Regina Belski, Harry Banaharis, Negar Naderpoor, and Barbora de Courten. 2019. "Increased Inflammation and Cardiometabolic Risk in Individuals with Low AMY1 Copy Numbers." *Journal of Clinical Medicine* 8 (3): 382. https://doi.org/10.3390/jcm8030382.

Marshall, Christian R., Daniel P. Howrigan, Daniele Merico, Bhooma Thiruvahindrapuram, Wenting Wu, Douglas S. Greer, Danny Antaki, et al. 2017. "Contribution of Copy Number Variants to Schizophrenia from a Genome-Wide Study of 41,321 Subjects." *Nature Genetics* 49 (1): 27–35. https://doi.org/10.1038/ng.3725.

Maxwell, E. K., J. S. Packer, C. O'Dushlaine, and S. E. McCarthy. 2017. "Profiling Copy Number Variation and Disease Associations from 50,726 DiscovEHR Study Exomes." *BioRxiv*. https://www.biorxiv.org/content/10.1101/119461v1.abstract.

McCarroll, Steven A., Finny G. Kuruvilla, Joshua M. Korn, Simon Cawley, James Nemesh, Alec Wysoker, Michael H. Shapero, et al. 2008. "Integrated Detection and Population-Genetic Analysis of SNPs and Copy Number Variation." *Nature Genetics* 40 (10): 1166–74. https://doi.org/10.1038/ng.238.

McCarthy, Shane, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R. Wood, Alexander Teumer, Hyun Min Kang, et al. 2016. "A Reference Panel of 64,976 Haplotypes for Genotype Imputation." *Nature Genetics* 48 (10): 1279–83. https://doi.org/10.1038/ng.3643.

Miga, Karen H., Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, et al. 2020. "Telomere-to-Telomere Assembly of a Complete Human X Chromosome." *Nature* 585 (7823): 79–84. https://doi.org/10.1038/s41586-020-2547-7.

Myocardial Infarction Genetics Consortium, Sekar Kathiresan, Benjamin F. Voight, Shaun Purcell, Kiran Musunuru, Diego Ardissino, Pier M. Mannucci, et al. 2009. "Genome-Wide Association of Early-Onset Myocardial Infarction with Single Nucleotide Polymorphisms and Copy Number Variants." *Nature Genetics* 41 (3): 334–41. https://doi.org/10.1038/ng.327.

Naj, Adam C. 2019. "Genotype Imputation in Genome-Wide Association Studies." *Et al [Current Protocols in Human Genetics]* 102 (1): e84. https://doi.org/10.1002/cphg.84.

Ni, Qingqiang, Lin Yun, Rui Xu, and Dong Shang. 2014. “Correlation between Blood Lipid Levels and Chronic Pancreatitis.” *Medicine* 93 (28): e331. https://doi.org/10.1097/md.0000000000000331.

Nisoli, Enzo, Emilio Clementi, Michele O. Carruba, and Salvador Moncada. 2007. “Defective Mitochondrial Biogenesis: A Hallmark of the High Cardiovascular Risk in the Metabolic Syndrome?” *Circulation Research* 100 (6): 795–806. https://doi.org/10.1161/01.RES.0000259591.97107.6c.

Nofziger, Charity, Amy J. Turner, Katrin Sangkuhl, Michelle Whirl-Carrillo, José A. G. Agúndez, John L. Black, Henry M. Dunnenberger, et al. 2020. “PharmVar GeneFocus: CYP2D6.” *Clinical Pharmacology and Therapeutics* 107 (1): 154–70. https://doi.org/10.1002/cpt.1643.

Nowell, C. 1962. “The Minute Chromosome (Ph1) in Chronic Granulocytic Leukemia.” *Blut Zeitschrift Für Die Gesamte Blutforschung*. https://doi.org/10.1007/bf01630378.

O’Callaghan, Nathan, Armand Valsesia, Sameer Kulkarni, Julien Marquis, Patricia Leone, Polina Mironova, Ondine Walter, et al. 2019. “Understanding Determinants of Carbohydrate Metabolism and Their Contribution to Metabolic Health; The Impact of AMY1 CNV (P21-015-19).” *Current Developments in Nutrition* 3 (Suppl 1). https://doi.org/10.1093/cdn/nzz041.P21-015-19.

Organization, World Health, and Others. 2017. “Cardiovascular Diseases-World Heart Day 2017.”

Perfield, James W., 2nd, Laura C. Ortinau, R. Taylor Pickering, Meghan L. Ruebel, Grace M. Meers, and R. Scott Rector. 2013. “Altered Hepatic Lipid Metabolism Contributes to Nonalcoholic Fatty Liver Disease in Leptin-Deficient Ob/Ob Mice.” *Journal of Obesity* 2013 (January): 296537. https://doi.org/10.1155/2013/296537.

Perry, George H., Nathaniel J. Dominy, Katrina G. Claw, Arthur S. Lee, Heike Fiegler, Richard Redon, John Werner, et al. 2007. “Diet and the Evolution of Human Amylase Gene Copy Number Variation.” *Nature Genetics* 39 (10): 1256–60. https://doi.org/10.1038/ng2123.

“PheWeb.” n.d. Accessed June 22, 2020. http://r3.finngen.fi/.

Pinho, Simão, Cristina Padez, and Licínio Manco. 2018. “High AMY1 Copy Number Protects against Obesity in Portuguese Young Adults.” *Annals of Human Biology* 45 (5): 435–39. https://doi.org/10.1080/03014460.2018.1490452.

Pollex, Rebecca L., and Robert A. Hegele. 2006. “Genetic Determinants of the Metabolic Syndrome.” *Nature Clinical Practice. Cardiovascular Medicine* 3 (9): 482–89. https://doi.org/10.1038/ncpcardio0638.

Purcell, S., S. S. Cherny, and P. C. Sham. 2003. “Genetic Power Calculator: Design of Linkage and Association Genetic Mapping Studies of Complex Traits.” *Bioinformatics* 19 (1): 149–50. https://doi.org/10.1093/bioinformatics/19.1.149.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Rateb Dweik, Nasser Eddin. 2017. “Origin of Nuclear Mitochondrial Pseudogenes (Numts).” *Journal of Phylogenetics & Evolutionary Biology* 05 (03). https://doi.org/10.4172/2329-9002.1000191.

Razzak, Muhammad Imran, Saeeda Naz, and Ahmad Zaib. 2018. “Deep Learning for Medical Image Processing: Overview, Challenges and the Future.” In *Lecture Notes in Computational Vision and Biomechanics*, 323–50. Lecture Notes in Computational

Vision and Biomechanics. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65981-7_12.

Regier, Allison A., Yossi Farjoun, David Larson, Olga Krasheninina, Hyun Min Kang, Daniel P. Howrigan, Bo-Juen Chen, et al. n.d. "Functional Equivalence of Genome Sequencing Analysis Pipelines Enables Harmonized Variant Calling across Human Genetics Projects." https://doi.org/10.1101/269316.

Ruderfer, Douglas M., Tymor Hamamsy, Monkol Lek, Konrad J. Karczewski, David Kavanagh, Kaitlin E. Samocha, Exome Aggregation Consortium, et al. 2016. "Patterns of Genic Intolerance of Rare Copy Number Variation in 59,898 Human Exomes." *Nature Genetics* 48 (10): 1107–11. https://doi.org/10.1038/ng.3638.

Shoar, Zohreh, Michael J. Goldenthal, Francesco De Luca, and Elizabeth Suarez. 2016. "Mitochondrial DNA Content and Function, Childhood Obesity, and Insulin Resistance." *Endocrine Research* 41 (1): 49–56. https://doi.org/10.3109/07435800.2015.1068797.

Song, J., J. Y. Oh, Y. A. Sung, Y. K. Pak, K. S. Park, and H. K. Lee. 2001. "Peripheral Blood Mitochondrial DNA Content Is Related to Insulin Sensitivity in Offspring of Type 2 Diabetic Patients." *Diabetes Care* 24 (5): 865–69. https://doi.org/10.2337/diacare.24.5.865.

Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. "An Integrated Map of Structural Variation in 2,504 Human Genomes." *Nature* 526 (7571): 75–81. https://doi.org/10.1038/nature15394.

Surakka, Ida, Momoko Horikoshi, Reedik Mägi, Antti-Pekka Sarin, Anubha Mahajan, Vasiliki Lagou, Letizia Marullo, et al. 2015. "The Impact of Low-Frequency and Rare Variants on Lipid Levels." *Nature Genetics* 47 (6): 589–97. https://doi.org/10.1038/ng.3300.

Szilágyi, A., and G. Fust. 2008. "Diseases Associated with the Low Copy Number of the C4B Gene Encoding C4, the Fourth Component of Complement." *Cytogenetic and Genome Research* 123 (1–4): 118–30. https://doi.org/10.1159/000184699.

Telenti, Amalio, Christoph Lippert, Pi-Chuan Chang, and Mark DePristo. 2018. "Deep Learning of Genomic Variation and Regulatory Network Data." *Human Molecular Genetics* 27 (Supplement_R1): R63–71. https://doi.org/10.1093/hmg/ddy115.

Thorvaldsdóttir, Helga, James T. Robinson, and Jill P. Mesirov. 2013. "Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration." *Briefings in Bioinformatics* 14 (2): 178–92. https://doi.org/10.1093/bib/bbs017.

Usher, Christina L., Robert E. Handsaker, Tõnu Esko, Marcus A. Tuke, Michael N. Weedon, Alex R. Hastie, Han Cao, et al. 2015. "Structural Forms of the Human Amylase Locus and Their Relationships to SNPs, Haplotypes and Obesity." *Nature Genetics* 47 (8): 921–25. https://doi.org/10.1038/ng.3340.

Vattikuti, Shashaank, Juen Guo, and Carson C. Chow. 2012. "Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits." *PLoS Genetics* 8 (3): e1002637. https://doi.org/10.1371/journal.pgen.1002637.

Vázquez-Moreno, Miguel, Aurora Mejía-Benítez, Tanmay Sharma, Jesús Peralta-Romero, Daniel Locia-Morales, Miguel Klünder-Klünder, National Obesity Network Mexico, Miguel Cruz, and David Meyre. 2020. "Association of AMY1A/AMY2A Copy Numbers and AMY1/AMY2 Serum Enzymatic Activity with Obesity in Mexican Children." *Pediatric Obesity* 15 (8): e12641. https://doi.org/10.1111/ijpo.12641.

Viljakainen, Heli, Johanna C. Andersson-Assarsson, Miriam Armenio, Minna Pekkinen, Maria Pettersson, Helena Valta, Marita Lipsanen-Nyman, Outi Mäkitie, and Anna Lindstrand. 2015. "Low Copy Number of the AMY1 Locus Is Associated with Early-Onset Female Obesity in Finland." *PloS One* 10 (7): e0131883. https://doi.org/10.1371/journal.pone.0131883.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *American Journal of Human Genetics* 101 (1): 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.

Wang, Chih-Hao, and Yau-Huei Wei. 2020. "Roles of Mitochondrial Sirtuins in Mitochondrial Function, Redox Homeostasis, Insulin Resistance and Type 2 Diabetes." *International Journal of Molecular Sciences* 21 (15): 5266. https://doi.org/10.3390/ijms21155266.

Wellcome Trust Case Control Consortium, Nick Craddock, Matthew E. Hurles, Niall Cardin, Richard D. Pearson, Vincent Plagnol, Samuel Robson, et al. 2010. "Genome-Wide Association Study of CNVs in 16,000 Cases of Eight Common Diseases and 3,000 Shared Controls." *Nature* 464 (7289): 713–20. https://doi.org/10.1038/nature08979.

Willer, Cristen J., Ellen M. Schmidt, Sebanti Sengupta, Gina M. Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, et al. 2013. "Discovery and Refinement of Loci Associated with Lipid Levels." *Nature Genetics* 45 (11): 1274–83. https://doi.org/10.1038/ng.2797.

Wilson, Peter W. F., Ralph B. D'Agostino, Helen Parise, Lisa Sullivan, and James B. Meigs. 2005. "Metabolic Syndrome as a Precursor of Cardiovascular Disease and Type 2 Diabetes Mellitus." *Circulation* 112 (20): 3066–72. https://doi.org/10.1161/CIRCULATIONAHA.105.539528.

Wolfram, Julie A., and J. Kevin Donahue. 2013. "Gene Therapy to Treat Cardiovascular Disease." *Journal of the American Heart Association* 2 (4): e000119. https://doi.org/10.1161/JAHA.113.000119.

Wu, Zhijun, Haihui Sheng, Yanjia Chen, Jing Tang, Yan Liu, Qiujing Chen, Lin Lu, and Wei Jin. 2014. "Copy Number Variation of the Lipoprotein(a) (LPA) Gene Is Associated with Coronary Artery Disease in a Southern Han Chinese Population." *International Journal of Clinical and Experimental Medicine* 7 (10): 3669–77. https://www.ncbi.nlm.nih.gov/pubmed/25419416.

Zekavat, Seyedeh M., Sanni Ruotsalainen, Robert E. Handsaker, Maris Alver, Jonathan Bloom, Tim Poterba, Cotton Seed, et al. n.d. "Deep Coverage Whole Genome Sequences and Plasma Lipoprotein(a) in Individuals of European and African Ancestries." https://doi.org/10.1101/225169.

Zhou, Xiaoshuang, Rongshan Li, Xinyan Liu, Lihua Wang, Peng Hui, Lawrence Chan, Pradip K. Saha, and Zhaoyong Hu. 2016. "ROCK1 Reduces Mitochondrial Content and Irisin Production in Muscle Suppressing Adipocyte Browning and Impairing Insulin Sensitivity." *Scientific Reports* 6 (1). https://doi.org/10.1038/srep29669.

Zhou, Xu-Jie, Fa-Juan Cheng, Ji-Cheng Lv, Huan Luo, Feng Yu, Min Chen, Ming-Hui Zhao, and Hong Zhang. 2012. "Higher DEFB4 Genomic Copy Number in SLE and ANCA-Associated Small Vasculitis." *Rheumatology (Oxford, England)* 51 (6): 992–95. https://doi.org/10.1093/rheumatology/ker419.

Zou, James, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. 2019. “A Primer on Deep Learning in Genomics.” *Nature Genetics* 51 (1): 12–18. https://doi.org/10.1038/s41588-018-0295-5.