Association of structural variation with cardiometabolic traits in Finns

Lei Chen,[1,2,3] Haley J. Abel,[1,2] Indraniel Das,[1] David E. Larson,[1,4] Liron Ganel,[1,2] Krishna L. Kanchi,[1] Allison A. Regier,[1,2] Erica P. Young,[1,5] Chul Joo Kang,[1] Alexandra J. Scott,[1,2] Colby Chiang,[1,2] Xinxin Wang,[1,2,3] Shuangjia Lu,[3] Ryan Christ,[1] Susan K. Service,[6] Charleston W.K. Chiang,[7,8] Aki S. Havulinna,[9,10] Johanna Kuusisto,[11,12] Michael Boehnke,[13] Markku Laakso,[11,12] Aarno Palotie,[9,14,15] Samuli Ripatti,[9,15,16] Nelson B. Freimer,[6] Adam E. Locke,[1,2] Nathan O. Stitziel,[1,2,4,*] and Ira M. Hall[1,2,3,*]

Summary

The contribution of genome structural variation (SV) to quantitative traits associated with cardiometabolic diseases remains largely unknown. Here, we present the results of a study examining genetic association between SVs and cardiometabolic traits in the Finnish population. 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 2 loci at which SVs have strong phenotypic effects: (1) a deletion of the *ALB* promoter that is greatly enriched in the Finnish population and causes decreased serum albumin level in carriers ($p = 1.47 \times 10^{-54}$) and is also associated with increased levels of total cholesterol ($p = 1.22 \times 10^{-28}$) and 14 additional cholesterol-related traits, and (2) a multi-allelic copy number variant (CNV) at *PDPR* that is strongly associated with pyruvate ($p = 4.81 \times 10^{-21}$) and alanine ($p = 6.14 \times 10^{-12}$) 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 ($p = 6.24 \times 10^{-10}$), which was also found to be strongly associated with increased glycoprotein level ($p = 3.53 \times 10^{-35}$). Our study confirms that integrating SVs in trait-mapping studies will expand our knowledge of genetic factors underlying disease risk.

Introduction

Common human diseases affecting the cardiovascular and endocrine systems are known to be associated with a variety of quantitative risk factors including various measures of cholesterol, metabolites, insulin, glucose, blood pressure, and obesity.[1,2] Understanding the genetic basis of these and other quantitative traits can shed light on the etiology, prevention, diagnosis, and treatment of disease. Family- and population-based studies have shown significant heritability for many cardiometabolic traits,[3–6] and prior genome-wide association studies (GWASs) have identified hundreds of associated loci.[7–9] However, most prior trait-mapping studies have focused on common variants ascertained by genotyping arrays or rare coding variants measured by exome sequencing, leaving out the contribution of larger and more complex forms of genome variation.

Of particular interest is the contribution of genome structural variation (SV), which encompasses diverse variant types larger than 50 base pairs (bp) in size, including copy number variants (CNVs), mobile element insertions (MEIs), inversions, and complex rearrangements. Although rare and *de novo* SVs have long been recognized to cause various rare and sporadic human disorders[10,11] and somatic SVs play a central role in cancer biology,[12] the extent to which SVs contribute more generally to common diseases and other complex traits in humans is less clear. Early genome-wide studies[13–15] failed to identify SVs associated with common diseases, but these

[1]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA; [2]Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA; [3]Department of Genetics, Yale University School of Medicine, New Haven, CT 06510, USA; [4]Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA; [5]Cardiovascular Division, Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA; [6]Center for Neurobehavioral Genetics, Jane and Terry Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA 90095, USA; [7]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA; [8]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA; [9]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki 00014, Finland; [10]Finnish Institute for Health and Welfare (THL), Helsinki 00271, Finland; [11]Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland, Kuopio 70210, Finland; [12]Department of Medicine, Kuopio University Hospital, Kuopio 70210, Finland; [13]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA; [14]Analytical and Translational Genetics Unit (ATGU), Psychiatric & Neurodevelopmental Genetics Unit, Departments of Psychiatry and Neurology, Massachusetts General Hospital, Boston, MA 02114, USA; [15]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [16]Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki 00014, Finland
*Correspondence: nstitziel@wustl.edu (N.O.S.), ira.hall@yale.edu (I.M.H.)
https://doi.org/10.1016/j.ajhg.2021.03.008.

were limited by the use of low-resolution array platforms, which only capture extremely large CNVs (>100 kb or similar), and by modest sample size. Several later studies performed targeted analysis of known SVs combined with larger-scale GWAS data,[16–18] 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.[19–21] Fine mapping of expression quantitative trait loci (eQTLs) using deep whole-genome sequencing (WGS) data has indicated that SVs are the causal variant at 3.5%–6.8% of eQTLs, and that causal SVs have larger effect sizes than causal single nucleotide variants (SNVs) and indels and are often not well-tagged by flanking SNVs.[22,23] This suggests that direct assessment of SVs in WGS-based complex trait association studies has the potential to reveal causative variants not found through other approaches.

Here, we have performed a SV association study using deep (>20×) WGS data from 4,030 individuals from Finland with extensive cardiometabolic trait measurements and extended these results to a larger set of 15,205 individuals with whole-exome sequencing (WES) and single nucleotide polymorphism (SNP) genotype data. Compared to prior work, our study benefits from (1) comprehensive SV ascertainment due to the use of deep WGS data and complementary SV detection methods, (2) deeply phenotyped individuals with existing SNP array and exome sequence data, and (3) the unique history of the Finnish population, which was shaped by multiple population bottlenecks and rapid population expansions, leading to an enrichment of some otherwise rare and low-frequency variants that can be detected by trait association at relatively modest sample sizes.[24–26] By testing for associations between structural variants and cardiometabolic traits, we identified 15 genome-wide significant loci, nine of which remained significant after multiple testing correction for the number of phenotypes, including a Finnish-enriched *ALB* promoter deletion associated with multiple traits, and a multi-allelic CNV affecting *PDPR* that is associated with pyruvate levels.

Material and methods

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.[26]

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. So, 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[22]). We estimated the minimum required sample size as 375 through an analytical approach implemented in Genetic Power Calculator.[27] 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 = 1×10^{-6}.

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 >20× coverage on the Illumina HiSeq X and NovaSeq platforms with paired-end 150 bp reads.

WGS data were aligned to the GRCh38 reference genome using BWA-MEM and processed using the functional equivalence pipeline.[28] An SV callset based on breakpoint mapping was generated using our recently published workflow[29] using the same methods as in our recent study of 17,795 human genomes.[30] Briefly, we ran LUMPY (v.0.2.13),[31] CNVnator (v.0.3.3),[32] and svtyper (v.0.1.4)[33] 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 (v.0.3.2).[29] 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.[29,30] According to our definition of SV, we filtered variants smaller than 50 bp. 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.[30] 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[34] 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 (>1 kb) within relatively unique genomic regions. Thus, in parallel we used a custom cohort-level CNV detection pipeline based on CNVnator[32] to detect smaller and more repetitive CNVs (see below).

We adapted the original GenomeSTRiP pipeline (v.2.00.1774) 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 "multi-allelic 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 S3), 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 artifacts, we excluded 612 LUMPY SVs, 740 GenomeSTRiP SVs, and 1,098 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 S1.

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 ($R_{max}^2 \geq 0.5$, flanking window size = 1 Mb). We calculated the overlap fraction between SV callsets by bedtools[35] intersect (v.2.23.0) 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,[36] using the genotype correlation matrix as input. The space clustering was evaluated by running bedtools cluster with -d (max distance) specified as 10 bp.

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 bi-allelic 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[37] implemented in EPACTS (v.3.2.9) 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 multi-allelic CNV, one single variant test was performed between the phenotype and the copy number value of the interval.

We applied matSpDlite[36] to estimate the equivalent number of independent tests. The genome-wide significance threshold was set at 1.89×10^{-6} after Bonferroni correction at level $\alpha = 0.05$ over 26,495 independent genetic variables, and the experiment-wide significance threshold was set as 3.32×10^{-8} to further correct for the 57 independent phenotypic variables also estimated using matSpDlite.[36]

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 S1). 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 bi-allelic SVs.

We separated the WES alignment data into two batches: the first composed of 10,379 samples sequenced with 100 bp paired-end reads and the second composed of 9,937 samples sequenced with 125 bp paired-end reads. For samples in each batch, we calculated the per-sample per-exon coverage by GATK[38] DepthOfCoverage (v.3.3-0) and adopted the data processing steps from the XHMM (v.1.0) pipeline[39] 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 batch 1 and 9,864 samples left in batch 2. We calculated the correlation between SV genotypes from WGS data and the normalized read-depth z-scores of exons intersected or nearby (<5 kb) 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.[26] We later did a meta-analysis under a fixed effect model using METASOFT (v.2.0.1)[40] 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 bi-allelic 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 (v.5.1)[41] 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 (2,310 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 bi-allelic 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.

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 catalog (see web resources) with the same set of keywords used in a previous study[26] (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 versus the SNP callset.

For loci containing multiple genotype-correlated SVs associated with a trait, we lumped the variants together using bedtools merge[35] 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[42] 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 ~1,000 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 3). 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 2) 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 (v.0.2)[43] 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 2.

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.

Structural variation detection and genotyping

We identified 120,793 SVs by LUMPY,[31] 111,141 CNVs by GenomeSTRiP[34] (GS), and 92,862 CNVs by our customized pipeline based on CNVnator.[32] 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 S9). Figures 1 and S2 provide an overview of the high-confidence callset, including the size distribution, composition of bi-allelic versus multi-allelic SVs, and frequency distributions. The SV size and frequency distributions are consistent with those in previous studies:[22,30,44,45] most called SVs are relatively small (<10 kb), bi-allelic, 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 (Figures 1 and S2).

Based on comparison with a set of SNP array intensity data (see material and 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 ($R^2 \geq 0.5$, see material and methods) was 56.8%, consistent with our prior GTEx study that used similar methods[22] and was evaluated extensively in the context of eQTL mapping. We also compared our callset to the high-quality SV callsets from 1000 Genomes

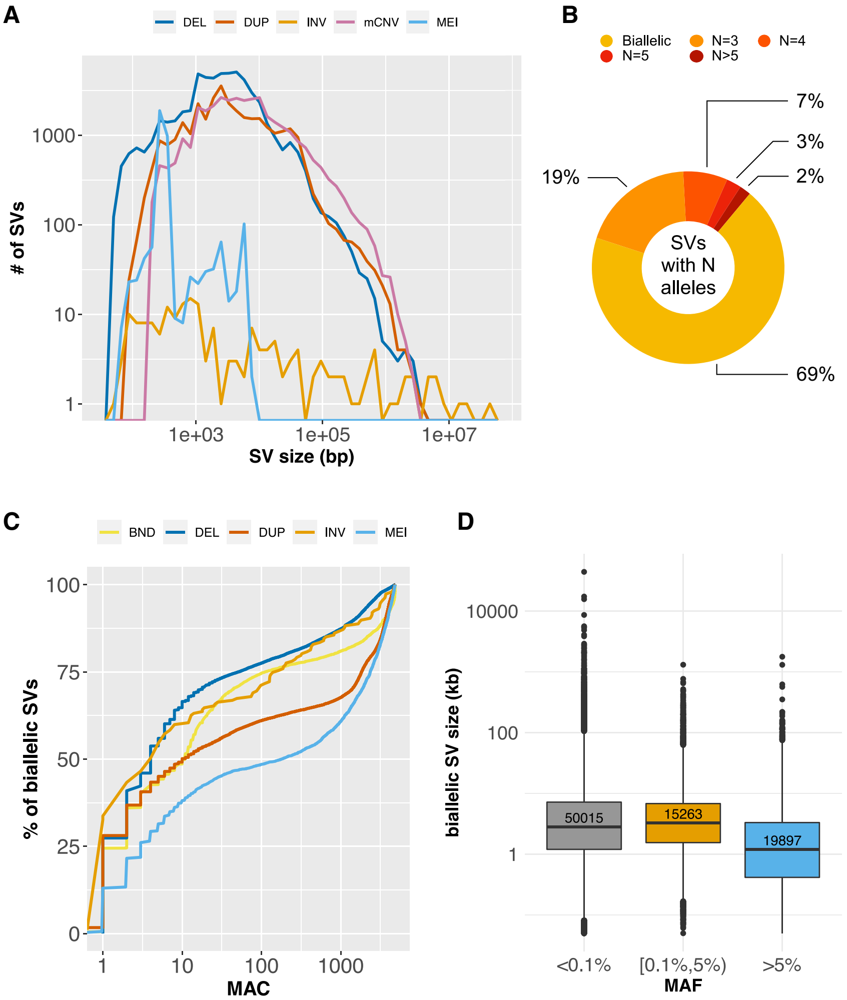

Figure 1. Overview of the high-confidence SV callset
(A) SV size distribution (log10 scale, bp) by variant type. BNDs are not included due to the ambiguous definition of variant boundaries.
(B) Proportion of bi-allelic SVs and multi-allelic CNVs, where N is defined by the number of copy number groups (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 bi-allelic 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.

(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 shows the above metrics stratified by pipelines. We estimated the genotype redundancy in total and stratified by pipelines (Table S2). 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 10 bp and measured the size of the clusters (Table S3).

Our CNVnator pipeline was the major source of redundancy and fragmentation since it detects CNVs with higher resolution—as small as 100 bp—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 S4**). 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.

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[37] in the 4,030 individuals with WGS data. We defined the genome-wide significance threshold as 1.89×10^{-6} and the experiment-wide significance threshold as 3.32×10^{-8} (see material and methods**)**. Nine associations of six loci passed genome-wide significance threshold (Table S5); six were still significant after adjusting for the equivalent number of independent phenotypes (Table 2, 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 material and methods).[26] We developed a strategy to genotype coding CNVs from WES data using read-depth information from XHMM,[39] and we 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 material and

Table 1. 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%

Quality control metrics of the SV callsets including all variants, high-confidence variants, and high-confidence common variants (defined by $\geq$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 material and 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 $r^2 \geq 0.5$) with any SNP in the 1 Mb flanking regions.
[a]CNVs only

methods). The estimated imputation accuracy of SVs corresponded well to their LD with nearby SNPs, as expected (Figure S5). 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% to 99.8% (Table S4). Overall, we were able to accurately genotype 2,053 of 4,864 candidate SVs using exome ($n = 392$) and/or array genotype ($n = 1,705$) data. 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 2).

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 2 shows the summary statistics of the lead SVs for their top traits (see also Table S5 for pre-merged summary statistics).

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

The strongest signal in the combined study was a 4 kb deletion immediately upstream of *ALB*, affecting the promoter region (Figure 2). This variant was 16-fold enriched in the Finnish population compared to non-Finnish Europeans from 1KG (MAF: 1.6% versus 0.1%) and was associated with 16 traits at genome-wide significance (Table S5, Figure S6). The top two associations were with serum albumin ($p = 1.47 \times 10^{-54}$, beta = 0.91) and total cholesterol ($p = 1.22 \times 10^{-28}$, beta = −0.49), and these are independent signals based on conditional analyses (Table S8). The cholesterol signal appears to explain the remaining 14 trait associations, all of which are highly correlated (Figure S6). 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[43] including the deletion variant and the 100 most significant SNPs on chr4:67–79 Mb (see material and 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 ($p = 0.023$, $n = 4,014$). 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 ($p = 6.52 \times 10^{-13}$, $n = 3,117$). 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 2). Thus, we hypothesize that the promoter deletion directly affects serum albumin by altering *ALB* expression and is associated with total cholesterol through its genetic correlation with other underlying causal variant(s) in the same LD block.

Prior studies[48–51] have reported five albumin-associated SNPs and two cholesterol-associated SNPs in this region. In our conditional analyses including all intrachromosomal GWAS hits,[46] the SV-albumin association remained genome-wide significant (Table 2) while the SV-cholesterol association was diminished (conditioned $p = 0.004$). To investigate the relationship between our signal and each of

Table 2. 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 frequency	p combined
deletion	*ALB*	albumin	4	3.49E−21	1.05E−10	0.91	IMP	Y	0.03	1.47E−54[a]
deletion	*HP*	glycoprotein	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	N/A	−0.12	N/A	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	N/A	−0.16	N/A	Y	0.57	1.10E−06
mCNV	SegDup cluster	glutamine	9	1.25E−06	N/A	−0.79	N/A	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	N/A	−0.39	N/A	Y	0.01	1.41E−06

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 1.89×10^{-6} and the experiment-wide significance threshold was 3.32×10^{-8} (see Table S2 and material 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,[46] using WGS data only (see material and methods)

[a] Experiment-wide significant

the seven previous GWAS SNPs, we tested the SV for association while conditioning on the reported SNPs one at a time (Table S6) and ran the association tests on those SNPs with the SV as covariate (Table S7). 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 S7, web resources); the top association was with statin medication use ($p = 6.5 \times 10^{-69}$). The second set of signals appeared in the "Endocrine, nutritional and metabolic diseases" category, led by disorders of lipoprotein metabolism and other lipidemias ($p = 1.4 \times 10^{-11}$), pure hypercholesterolemia ($p = 3.0 \times 10^{-11}$), and metabolic disorders ($p = 1.8 \times 10^{-7}$). 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.

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 ($p = 4.81 \times 10^{-21}$, beta $= -0.72$) and alanine ($p = 6.14 \times 10^{-12}$, beta $= -0.53$) levels in the serum. We reconstructed the copy number profile of this locus from short-read WGS data (see material and methods) and confirmed that the 13 correlated variant calls correspond to a single ~250 kb multi-allelic CNV (CNV1 in Figure 3) 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 3). 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 S8).

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

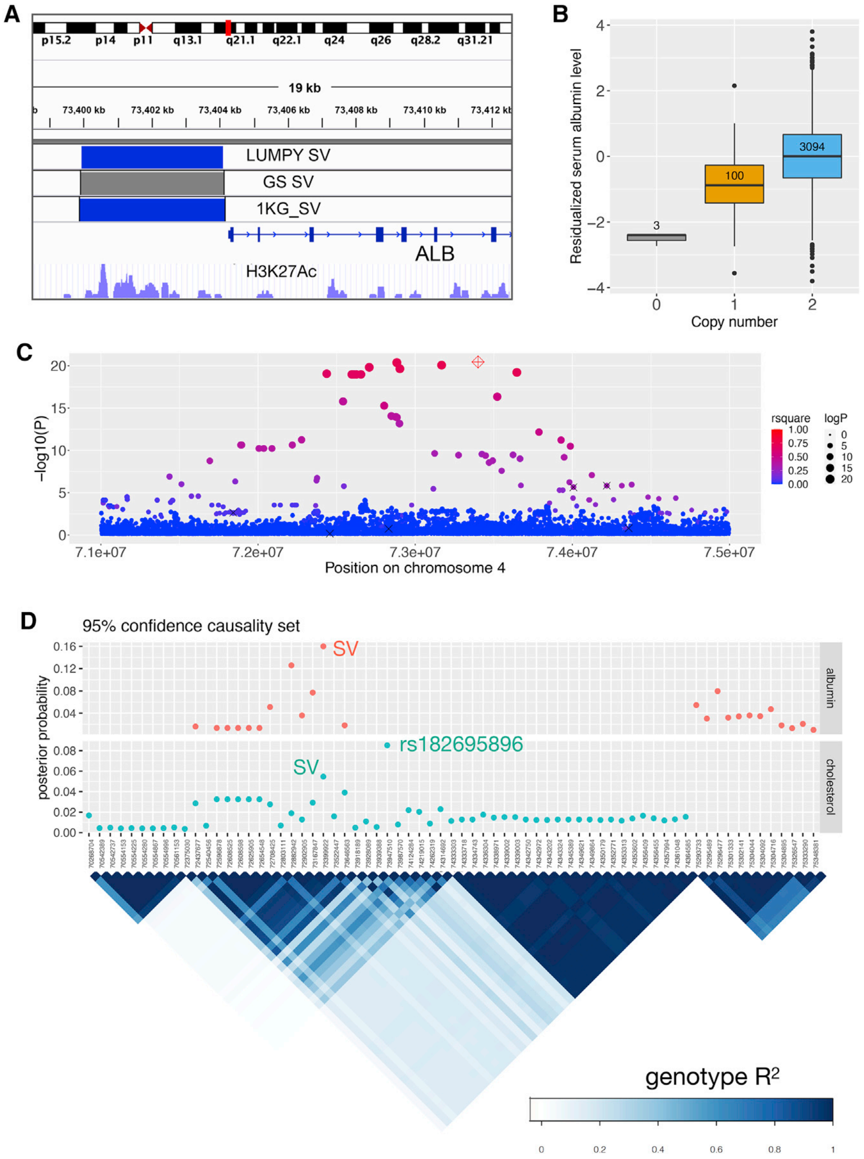

Figure 2. The *ALB* promotor deletion associated with serum albumin level and cholesterol traits

(A) The genomic location of the chr4 deletion, with coordinates detected from LUMPY, GenomeSTRiP, and 1KG. The H3K27Ac track is from the ENCODE[47] 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). 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.

(C) Local Manhattan plot of albumin association signals on chr4:71–75 Mb, 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 gray). 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.

distal to *PDPR*, as well as more divergent copies located ~55 Mb 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.[53,54,55] There are both duplication and deletion alleles of *PDPR*, 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[26]), 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, 50× 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 3) 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.[22] 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[56] created from long-read sequencing data (Figure S8). 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 3). These data suggest that CNV2 reflects highly variable structural alleles of *LOC283922* located 4 Mb away from *PDPR*, and thus it is not surprising that this CNV does not affect pyruvate levels.

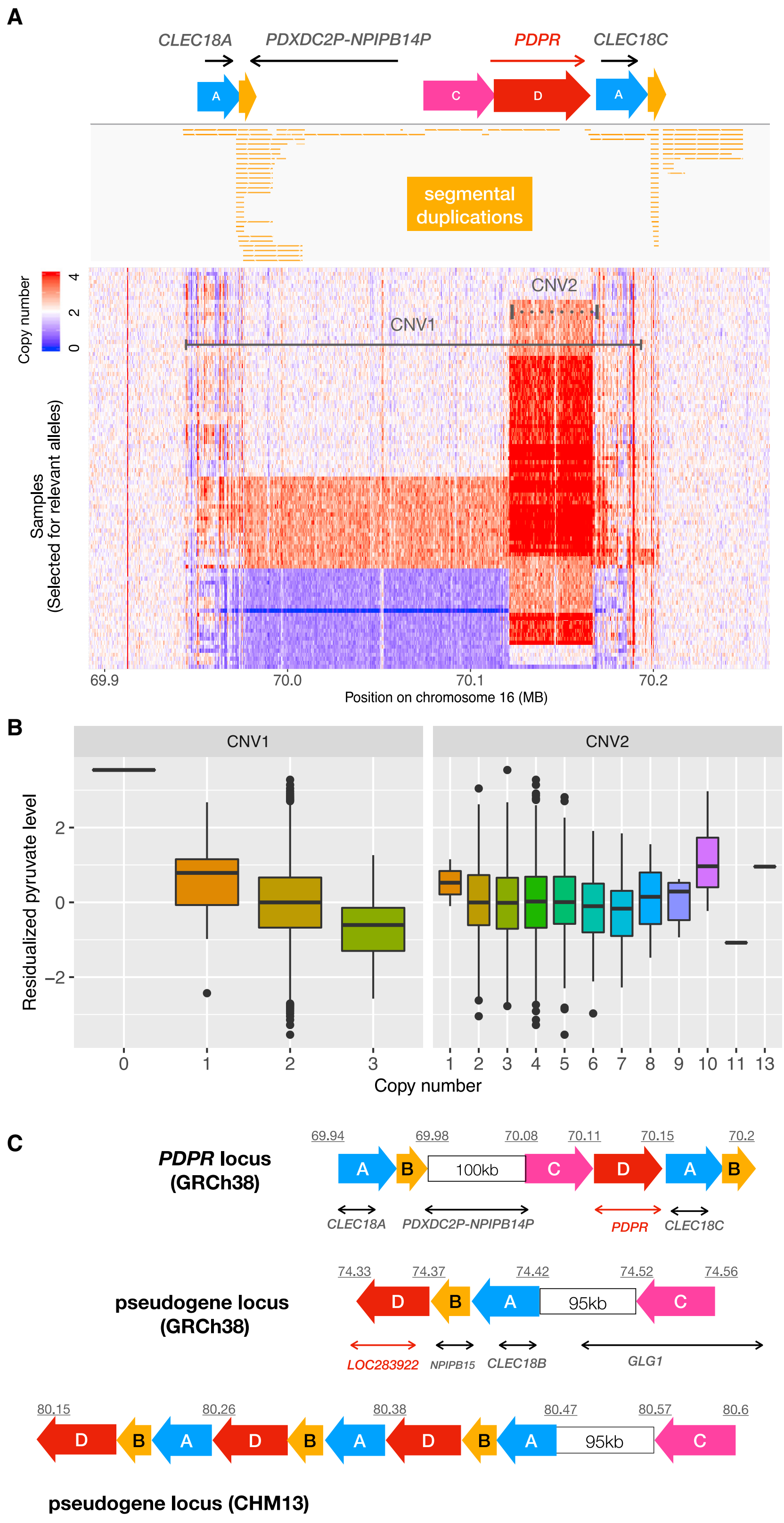

Figure 3. The multi-allelic CNV at the *PDPR* locus affecting pyruvate and alanine

(A) The *PDPR* locus showing (from top to bottom) genes, duplicated genomic segments based on dotplot analysis (see Figure S8), segmental duplication annotations from the UCSC table browser,[52] and copy number profiles for 100 samples comprising 51 carriers and 49 non-carriers for CNV1. Copy number is shown in 500 bp 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 ($p = 9.41 \times 10^{-11}$) 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 S8. The segment B corresponds to LCR16a, the core element shared by many duplicons sparsely distributed on chromosome 16.[53]

Additional trait-association signals

We confirmed a previously reported association between the recurrent *HP* deletion and decreased total serum cholesterol levels.[16] In our data, this same deletion was strongly associated with serum glycoprotein acetyls quantified by NMR ($p = 3.53 \times 10^{-35}$), and conditional analysis showed that the two associations were independent (Table S8). 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 900 kDa—more than twice the size of the dimers (86 kDa)[57]—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 1 Mb of previously published GWAS loci for the same traits. All SVs were well tagged by SNPs ($R^2 > 0.9$) 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 2). 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 $p = 4.17 \times 10^{-10}$). This variant was a 4 kb 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, $p = 1.44 \times 10^{-11}$) within 100 kb of and partially tagging the SV ($R^2 = 0.61$), 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 SNPs[46] (Table 2), 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 traits.[58]

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 2, Table S5). 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 S9), 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)[59] 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 ($p = 1.00 \times 10^{-10}$) and related traits and are the topic of a separate study.[60]

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 2). 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.

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 nine novel (i.e., not present in existing GWAS databases) and six 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 Finns.[26] 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 eQTLs[22] and selective constraint,[30] 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 multi-allelic 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 allowed us to determine 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 important discovery that variable abundance of neutrophils versus platelets in peripheral blood is strongly associated with insulin, fat mass, and related metabolic traits (as described in detail elsewhere[60]).

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.

Data and code availability

WES and phenotype data for METSIM and FINRISK are available through dbGaP (accessions phs000752 and phs000756). METSIM WGS data have been submitted to AnVIL (dbGaP accessions phs0001579). Genomic and phenotypic data for the FINRISK cohort can be obtained through THL Biobank, the Finnish Institute for Health and Welfare, Finland. Structural variant site frequency information is available in dbVAR (accession nstd204). Summary statistics are available on GitHub (see web resources). Code is available upon request.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.03.008.

Acknowledgments

We thank D. Ray from Johns Hopkins University for her comments to the manuscript. This work was funded by an NHGRI CCDG award to I.M.H. and N.O.S. (UM1 HG008853) and DK U01 DK062370, the NHGRI large-scale sequencing grant (grant number 5U54HG003079), the Sigrid Jusélius Foundation (to S.R.), the University of Helsinki HiLIFE Fellow grants 2017-2020 (to S.R.), the Academy of Finland Center of Excellence in Complex Disease Genetics (grant number 312062 to S.R., grant number 312074 to S.R. and A.P.), the Academy of Finland (grant number 285380 to S.R.), the National Heart, Lung, and Blood Institute (grant number T32HL007081 to E.P.Y.), and the National Center for Advancing Translational Sciences (grant number UL1TR002345 to E.P.Y.). 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. A.S.H. was supported by the Academy of Finland (grant no. 321356). L.C. was supported by the McDonnell International Scholars Academy Fellowship. A.J.S. was supported by the Mr. and Mrs. Spencer T. Olin Fellowship for Women in Graduate Study.

Declaration of interests

N.O.S. has received research funding from Regeneron Pharmaceuticals unrelated to this study. The rest of the authors declare no competing interests.

Received: December 12, 2020
Accepted: March 3, 2021
Published: April 1, 2021

Web resources

AnVIL, https://anvilproject.org/data

dbGaP, https://www.ncbi.nlm.nih.gov/gap

dbVar, https://www.ncbi.nlm.nih.gov/dbvar/

EBI GWAS Catalogue (2019-11-21 version), https://www.ebi.ac.uk/gwas/docs/file-downloads

Efficient and parallelizable association container toolbox (EPACTS), https://genome.sph.umich.edu/wiki/EPACTS

FinnGen project PheWeb, http://r4.finngen.fi/about

LiftOver from UCSC Genome Browser, https://genome.ucsc.edu/cgi-bin/hgLiftOver

THL Biobank, the Finnish Institute for Health and Welfare, Finland, https://thl.fi/en/web/thl-biobank

The summary statistics of all the tested SVs and traits are available through GitHub, https://github.com/hall-lab/FinnSV_paper_1220

References

1. Ortega, F.B., Lavie, C.J., and Blair, S.N. (2016). Obesity and Cardiovascular Disease. Circ. Res. *118*, 1752–1770.
2. Francula-Zaninovic, S., and Nola, I.A. (2018). Management of measurable variable cardiovascular disease' risk factors. Curr. Cardiol. Rev. *14*, 153–163.
3. Kolifarhood, G., Daneshpour, M., Hadaegh, F., Sabour, S., Mozafar Saadati, H., Akbar Haghdoust, A., Akbarzadeh, M., Sedaghati-Khayat, B., and Khosravi, N. (2019). Heritability of blood pressure traits in diverse populations: a systematic review and meta-analysis. J. Hum. Hypertens. *33*, 775–785.
4. Kim, Y., Lee, Y., Lee, S., Kim, N.H., Lim, J., Kim, Y.J., Oh, J.H., Min, H., Lee, M., Seo, H.-J., et al. (2015). On the Estimation of Heritability with Family-Based and Population-Based Samples. BioMed Res. Int. *2015*, 671349.
5. Campbell Am, L.V. (2017). Genetics of obesity. Aust. Fam. Physician *46*, 456–459.
6. Hagenbeek, F.A., Pool, R., van Dongen, J., Draisma, H.H.M., Jan Hottenga, J., Willemsen, G., Abdellaoui, A., Fedko, I.O.,

den Braber, A., Visser, P.J., et al.; BBMRI Metabolomics Consortium (2020). Heritability estimates for 361 blood metabolites across 40 genome-wide association studies. Nat. Commun. *11*, 39.

7. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.
8. Fall, T., and Ingelsson, E. (2014). Genome-wide association studies of obesity and metabolic syndrome. Mol. Cell. Endocrinol. *382*, 740–757.
9. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.
10. Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. Annu. Rev. Med. *61*, 437–455.
11. Posey, J.E., O'Donnell-Luria, A.H., Chong, J.X., Harel, T., Jhangiani, S.N., Coban Akdemir, Z.H., Buyske, S., Pehlivan, D., Carvalho, C.M.B., Baxter, S., et al.; Centers for Mendelian Genomics (2019). Insights into genetics, human biology and disease gleaned from family based genomic studies. Genet. Med. *21*, 798–812.
12. Tubio, J.M.C. (2015). Somatic structural variation and cancer. Brief. Funct. Genomics *14*, 339–351.
13. Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., Giannoulatou, E., et al.; Wellcome Trust Case Control Consortium (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature *464*, 713–720.
14. McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I.W., Maller, J.B., Kirby, A., et al. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. *40*, 1166–1174.
15. Kathiresan, S., Voight, B.F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P.M., Anand, S., Engert, J.C., Samani, N.J., Schunkert, H., et al.; Myocardial Infarction Genetics Consortium; and Wellcome Trust Case Control Consortium (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat. Genet. *41*, 334–341.
16. Boettger, L.M., Salem, R.M., Handsaker, R.E., Peloso, G.M., Kathiresan, S., Hirschhorn, J.N., and McCarroll, S.A. (2016). Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat. Genet. *48*, 359–366.
17. Usher, C.L., Handsaker, R.E., Esko, T., Tuke, M.A., Weedon, M.N., Hastie, A.R., Cao, H., Moon, J.E., Kashin, S., Fuchsberger, C., et al. (2015). Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. Nat. Genet. *47*, 921–925.
18. Zekavat, S.M., Ruotsalainen, S., Handsaker, R.E., Alver, M., Bloom, J., Poterba, T., Seed, C., Ernst, J., Chaffin, M., Engreitz, J., et al. (2018). Deep coverage whole genome sequences and plasma lipoprotein(a) in individuals of European and African ancestries. Nat. Commun. *9*, 1–14.
19. Macé, A., Tuke, M.A., Deelen, P., Kristiansson, K., Mattsson, H., Nõukas, M., Sapkota, Y., Schick, U., Porcu, E., Rüeger, S., et al. (2017). CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. Nat. Commun. *8*, 744.
20. Aguirre, M., Rivas, M.A., and Priest, J. (2019). Phenome-wide Burden of Copy-Number Variation in the UK Biobank. Am. J. Hum. Genet. *105*, 373–383.
21. Li, Y.R., Glessner, J.T., Coe, B.P., Li, J., Mohebnasab, M., Chang, X., Connolly, J., Kao, C., Wei, Z., Bradfield, J., et al. (2020). Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. Nat. Commun. *11*, 255.
22. Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T., Damani, F.N., Ganel, L., Montgomery, S.B., et al.; GTEx Consortium (2017). The impact of structural variation on human gene expression. Nat. Genet. *49*, 692–699.
23. Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J., et al.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; and Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz (2017). The impact of rare variation on gene expression across tissues. Nature *550*, 239–243.
24. Lim, E.T., Würtz, P., Havulinna, A.S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., Lappalainen, T., et al.; Sequencing Initiative Suomi (SISu) Project (2014). Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet. *10*, e1004494.
25. Davis, J.P., Huyghe, J.R., Locke, A.E., Jackson, A.U., Sim, X., Stringham, H.M., Teslovich, T.M., Welch, R.P., Fuchsberger, C., Narisu, N., 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 Genet. *13*, e1007079.
26. Locke, A.E., Steinberg, K.M., Chiang, C.W.K., Service, S.K., Havulinna, A.S., Stell, L., Pirinen, M., Abel, H.J., Chiang, C.C., Fulton, R.S., et al.; FinnGen Project (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. Nature *572*, 323–328.
27. Purcell, S., Cherny, S.S., and Sham, P.C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics *19*, 149–150.
28. Regier, A.A., Farjoun, Y., Larson, D.E., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.-J., Kher, M., Banks, E., Ames, D.C., et al. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat. Commun. *9*, 4038.
29. Larson, D.E., Abel, H.J., Chiang, C., Badve, A., Das, I., Eldred, J.M., Layer, R.M., and Hall, I.M. (2019). svtools: population-scale analysis of structural variation. Bioinformatics *35*, 4782–4787.

30. Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., et al.; NHGRI Centers for Common Disease Genomics (2020). Mapping and characterization of structural variation in 17,795 human genomes. Nature *583*, 83–89.
31. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.
32. Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974–984.
33. Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R., and Hall, I.M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat. Methods *12*, 966–968.
34. Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., Kashin, S., Boettger, L.M., and McCarroll, S.A. (2015). Large multiallelic copy number variations in humans. Nat. Genet. *47*, 296–303.
35. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.
36. Li, J., and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity *95*, 221–227.
37. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354.
38. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, 1–33.
39. Fromer, M., and Purcell, S.M. (2014). Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. Curr. Protoc. Hum. Genet. *81*, 1–21.
40. Han, B., and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am. J. Hum. Genet. *88*, 586–598.
41. Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. *103*, 338–348.
42. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.
43. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508.
44. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., et al.; 1000 Genomes Project Consortium (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.
45. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al.; Genome Aggregation Database Production Team; and Genome Aggregation Database Consortium (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451.
46. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47* (D1), D1005–D1012.
47. ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.
48. Inouye, M., Ripatti, S., Kettunen, J., Lyytikäinen, L.-P., Oksala, N., Laurila, P.-P., Kangas, A.J., Soininen, P., Savolainen, M.J., Viikari, J., et al. (2012). Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. PLoS Genet. *8*, e1002907.
49. Kettunen, J., Tukiainen, T., Sarin, A.-P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.-P., Kangas, A.J., Soininen, P., Würtz, P., Silander, K., et al. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat. Genet. *44*, 269–276.
50. Kettunen, J., Demirkan, A., Würtz, P., Draisma, H.H.M., Haller, T., Rawal, R., Vaarhorst, A., Kangas, A.J., Lyytikäinen, L.-P., Pirinen, M., et al. (2016). Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nat. Commun. *7*, 11122.
51. Surakka, I., Horikoshi, M., Mägi, R., Sarin, A.-P., Mahajan, A., Lagou, V., Marullo, L., Ferreira, T., Miraglio, B., Timonen, S., et al.; ENGAGE Consortium (2015). The impact of low-frequency and rare variants on lipid levels. Nat. Genet. *47*, 589–597.
52. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al.; University of California Santa Cruz (2003). The UCSC Genome Browser Database. Nucleic Acids Res. *31*, 51–54.
53. Jiang, Z., Tang, H., Ventura, M., Cardone, M.F., Marques-Bonet, T., She, X., Pevzner, P.A., and Eichler, E.E. (2007). Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat. Genet. *39*, 1361–1368.
54. Johnson, M.E., Cheng, Z., Morrison, V.A., Scherer, S., Ventura, M., Gibbs, R.A., Green, E.D., Eichler, E.E.; and National Institute of Health Intramural Sequencing Center Comparative Sequencing Program (2006). Recurrent duplication-driven transposition of DNA during hominoid evolution. Proc. Natl. Acad. Sci. USA *103*, 17626–17631.
55. Cantsilieris, S., Sunkin, S.M., Johnson, M.E., Anaclerio, F., Huddleston, J., Baker, C., Dougherty, M.L., Underwood, J.G., Sulovari, A., Hsieh, P., et al. (2020). An evolutionary driver of interspersed segmental duplications in primates. Genome Biol. *21*, 202.
56. Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature *585*, 79–84.
57. Sadrzadeh, S.M.H., and Bozorgmehr, J. (2004). Haptoglobin phenotypes in health and disorders. Am. J. Clin. Pathol. *121* (*Suppl*), S97–S104.
58. Lin, H., Lee, E., Hestir, K., Leo, C., Huang, M., Bosch, E., Halenbeck, R., Wu, G., Zhou, A., Behrens, D., et al. (2008). Discovery

of a cytokine and its receptor by functional screening of the extracellular proteome. Science *320*, 807–811.

59. Lopez, J.V., Yuhki, N., Masuda, R., Modi, W., and O'Brien, S.J. (1994). Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J. Mol. Evol. *39*, 174–190.

60. Ganel, L., Chen, L., Christ, R., Vangipurapu, J., Young, E., Das, I., Kanchi, K., Larson, D., Regier, A., Abel, H., et al. (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, 2020.10.23.20218586.