Open access • Posted Content • DOI:10.1101/563866

Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program

Daniel Taliun, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson ...+191 more authors

Institutions: University of Michigan, University of Maryland, Baltimore, University of Washington, University of California, Berkeley ...+57 more institutions

Published on: 06 Mar 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Reference genome, Whole genome sequencing and Genome

2021-02-10

Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program

Daniel Taliun
University of Michigan

Et al.

Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program

https://doi.org/10.1038/s41586-021-03205-y

Received: 6 March 2019

Accepted: 7 January 2021

Published online: 10 February 2021

Open access

Check for updates

A list of authors and their affiliations appears at the end of the paper.

The Trans-Omics for Precision Medicine (TOPMed) programme seeks to elucidate the genetic architecture and biology of heart, lung, blood and sleep disorders, with the ultimate goal of improving diagnosis, treatment and prevention of these diseases. The initial phases of the programme focused on whole-genome sequencing of individuals with rich phenotypic data and diverse backgrounds. Here we describe the TOPMed goals and design as well as the available resources and early insights obtained from the sequence data. The resources include a variant browser, a genotype imputation server, and genomic and phenotypic data that are available through dbGaP (Database of Genotypes and Phenotypes)[1]. In the first 53,831 TOPMed samples, we detected more than 400 million single-nucleotide and insertion or deletion variants after alignment with the reference genome. Additional previously undescribed variants were detected through assembly of unmapped reads and customized analysis in highly variable loci. Among the more than 400 million detected variants, 97% have frequencies of less than 1% and 46% are singletons that are present in only one individual (53% among unrelated individuals). These rare variants provide insights into mutational processes and recent human evolutionary history. The extensive catalogue of genetic variation in TOPMed studies provides unique opportunities for exploring the contributions of rare and noncoding sequence variants to phenotypic variation. Furthermore, combining TOPMed haplotypes with modern imputation methods improves the power and reach of genome-wide association studies to include variants down to a frequency of approximately 0.01%.

Advancing DNA-sequencing technologies and decreasing costs are enabling researchers to explore human genetic variation at an unprecedented scale[2,3]. For these advances to improve our understanding of human health, they must be deployed in well-phenotyped human samples and used to build resources such as variation catalogues[3,4], control collections[5,6] and imputation reference panels[7–9]. Here we describe high-coverage whole-genome sequencing (WGS) analyses of the first 53,831 TOPMed samples (Box 1 and Extended Data Tables 1, 2); additional data are being made available as quality control, variant calling and dbGaP curation are completed (altogether more than 130,000 TOPMed samples are now available in dbGaP).

A key goal of the TOPMed programme is to understand risk factors for heart, lung, blood and sleep disorders by adding WGS and other 'omics' data to existing studies with deep phenotyping (Supplementary Information 1.1 and Supplementary Fig. 1). The programme currently consists of more than 80 participating studies, around 1,000 investigators and more than 30 working groups (https://www.nhlbiwgs.org/working-groups-public). TOPMed participants are ethnically and ancestrally diverse (Extended Data Fig. 1, Supplementary Information 1.1.4 and Supplementary Fig. 2). Through a combination of race and ethnicity information (from participant questionnaires and/or study inclusion criteria), we classified study participants into 'population groups', which varied in composition according to the goals of each analysis. In some analyses, these groups were further refined using genetic ancestry (see Methods and Supplementary Information for details).

Our study extends previous efforts by identifying and characterizing the rare variants that comprise the majority of human genomic variation[7,10–12]. Rare variants represent recent and potentially deleterious changes that can affect protein function, gene expression or other biologically important elements[11,13,14].

TOPMed WGS quality assessment

WGS of the TOPMed samples was performed over multiple studies, years and sequencing centres. To minimize batch effects, we standardized laboratory methods, mapped and processed sequence data centrally using a single pipeline, and performed variant calling and genotyping jointly across all samples (see Methods). We annotated each variant site with multiple sequence quality metrics and trained machine learning filters to identify and exclude inconsistencies that are revealed when the same individual was sequenced repeatedly. Available WGS data were processed periodically to produce genotype data 'freezes'. The 53,831 samples described here are drawn from TOPMed freeze 5.

Stringent variant and sample quality filters were applied and the resulting genotype call sets were evaluated in several ways (Supplementary Information 1.2.2, 1.3, 1.4). First, we compared genotypes for samples sequenced in duplicate (the mean alternative allele concordance was 0.9995 for single-nucleotide variants (SNVs) and 0.9930 for insertions or deletions (indels)). Second, we compared genotypes to those from previous whole-exome sequencing datasets (protein-coding regions from GENCODE[15]; 80% of variants were found with both approaches and overlapping variant calls had a concordance of 0.9993 for SNVs and 0.9974 for indels) (Supplementary Tables 1–3). Third, we compared genotypes to those obtained using alternative

Box 1

TOPMed participant consents and data access

The TOPMed programme comprises more than 80 participating studies, of which 32 are represented in the 53,831 whole genomes described here. TOPMed has leveraged existing studies with deep phenotyping and longitudinal follow-up data and with varied informed consent procedures and options. Consent groups range from broad 'general research use' and 'health, medical and biomedical' categories to disease-specific categories for heart, lung, blood and/or sleep disorders. Many studies have further consent modifiers, such as limiting use to not-for-profit organizations or requiring documentation of local IRB approval. Participant consents guide the appropriate use of data by TOPMed investigators as well; therefore, the set of study-consent groups used varies across different analyses reported in this paper (Extended Data Table 3).

TOPMed data have been deposited in dbGaP and access is adjudicated by a staff committee of the National Institutes of Health. The committee verifies that applications are consistent with data use limitations and consent groups for each sample. Study investigators have no role in the decision, except in a small subset of studies that require a letter of collaboration. A summary of currently available data and any use restrictions is available at https://www.ncbi.nlm.nih.gov/gap/advanced_search/?TERM=topmed.

Although TOPMed studies have separate dbGaP accessions, formats are standardized to facilitate combining data, with all variants from the joint genotype call set included in the variant call format (VCF) files, unique sample identifiers across all of TOPMed and sample attributes with TOPMed-specific variables. Notably, cross-study analyses require the identification of a set of compatible study-consent groups. In addition to genotype calls, CRAM files with aligned sequence reads are also available, hosted in commercial clouds and with access managed by dbGaP. The dbGaP accession numbers for all TOPMed studies referenced in this paper are listed in Extended Data Tables 2, 3.

The TOPMed imputation reference panel is available to users for imputation into their own samples via an imputation server. The server performs imputation into these samples, while the reference panel data themselves are not exposed to the user because they derive from multiple studies with variable consent types and other data use limitations (Extended Data Table 3).

informatics tools (compared to GATK v.4.1.3, TOPMed has lower Mendelian inconsistency rates and minimizes batch effects) (Supplementary Table 4). These reproducibility estimates indicate the high quality of the genotype calls and effectiveness of machine-learning-based quality filters.

Batch effects were evaluated by (1) comparing distributions of genetic principal components among sequencing centres, which are very similar between European American and African American individuals (Supplementary Figs. 3–5); (2) comparing alternative allele concordance between duplicates among centres, which is high (the largest difference being 4×10^{-4}), and the patterns of between- versus within-centre differences, which indicate random errors rather than systematic centre differences (Supplementary Figs. 6–8); and (3) performing tests of association between variants and batches, which show a very small fraction of variants with genome-wide significance (0.004%, Supplementary Figs. 9,10) (Supplementary Information 1.2). We conclude that batch effects appear to be minor, thus enabling multi-study association testing.

410 million genetic variants in 53,831 samples

A total of 7.0×10^{15} bases of DNA-sequencing data were generated, consisting of an average of 129.6×10^{9} bases of sequence distributed across 864.2 million paired reads (each 100–151 base pairs (bp) long) per individual. For a typical individual, 99.65% of the bases in the reference genome were covered, to a mean read depth of 38.2×.

Sequence analysis identified 410,323,831 genetic variants (381,343,078 SNVs and 28,980,753 indels), corresponding to an average of one variant per 7 bp (Extended Data Table 4). Overall, 78.7% of these variants had not been described in dbSNP build 149; TOPMed variants now account for the majority of variants in dbSNP. Among all variant alleles, 46.0% were singletons, observed once across all 53,831 participants. Among 40,722 unrelated participants (see Methods), the proportion of singleton variants was higher at 53.1% (Table 1). Down-sampling analyses show that the proportion of singletons increases until around 15,000 unrelated individuals are sequenced and then decreases very gradually (Supplementary Fig. 11). The fraction of singletons in each region or class of sites closely tracks functional constraints. For example, among all 4,651,453 protein-coding variants in unrelated individuals, the proportion of singletons was the highest for the 104,704 frameshift variants (68.4%), high among the 97,217 putative splice and truncation variants (62.1%), intermediate among the 2,965,093 nonsynonymous variants (55.6%) and lowest among the 1,435,058 synonymous variants (49.8%). Beyond protein-coding sequences, we found increased proportions of singletons in promoters (55.0%), 5′ untranslated regions (54.7%), regions of open chromatin (53.4%) and 3′ untranslated regions (53.3%); we found lower proportions of singletons in intergenic regions (53.0%) (Supplementary Table 5). Although putative transcription factor binding sites initially appeared to show fewer singletons (52.7%) than the remainder of the genome (53.1%), this pattern did not hold when we analysed highly mutable CpG sites separately. In fact, transcription factor binding sites were enriched for singletons in both CpG sites and non-CpG sites, an example of Simpson's paradox[16].

We identified an average of 3.78 million variants in each genome. Among these, an average of 30,207 (0.8%) were novel and 3,510 (0.1%) were singletons. Among all variants, we observed 3.17 million nonsynonymous and 1.53 million synonymous variants (a 2.1:1 ratio), but individual genomes contained similar numbers of nonsynonymous and synonymous variants (11,743 nonsynonymous and 11,768 synonymous, on average) (Extended Data Table 4). The difference can be explained if more than half of the nonsynonymous variants are removed from the population by natural selection before they become common.

Putative loss-of-function variants

A notable class of variants is the 228,966 putative loss-of-function (pLOF) variants that we observed in 18,493 (95.0%) GENCODE[15] genes (Extended Data Table 5 and Supplementary Fig. 12). This class includes the highest proportion of singletons among all of the variant classes that we examined. An average individual carried 2.5 unique pLOF variants. We identified more pLOF variants per individual than in previous surveys based on exome sequencing—an increase that was mainly driven by the identification of additional frameshift variants (Supplementary Table 6) and by a more uniform and complete coverage of protein-coding regions (Supplementary Figs. 13, 14).

We searched for gene sets with fewer rare pLOF variants than expected based on gene size. The gene sets with strong functional constraint included genes that encode DNA- and RNA-binding proteins, spliceosomal complexes, translation initiation machinery and

Table 1 | Number of variants in 40,722 unrelated individuals in TOPMed

	All unrelated individuals (n = 40,722)		Per individual			
	Total	Singletons (%)	Average	5th percentile	Median	95th percentile
Total variants	**384,127,954**	**203,994,740 (53)**	**3,748,599**	**3,516,166**	**3,563,978**	**4,359,661**
SNVs	357,043,141	189,429,596 (53)	3,553,423	3,335,442	3,380,462	4,125,740
Indels	27,084,813	14,565,144 (54)	195,176	180,616	183,503	233,928
Novel variants	**298,373,330**	**191,557,469 (64)**	**29,202**	**20,312**	**24,106**	**44,336**
SNVs	275,141,134	177,410,620 (64)	25,027	17,520	20,975	36,861
Indels	23,232,196	14,146,849 (61)	4,175	2,747	3,145	7,359
Coding variation	**4,651,453**	**2,523,257 (54)**	**23,909**	**22,158**	**22,557**	**27,716**
Synonymous	1,435,058	715,254 (50)	11,651	10,841	11,056	13,678
Nonsynonymous	2,965,093	1,648,672 (56)	11,384	10,632	10,856	13,221
Stop/essential splice	97,217	60,347 (62)	474	425	454	566
Frameshift	104,704	71,577 (68)	132	112	127	165
In-frame	51,997	29,110 (56)	102	85	99	128

Novel variants are taken as variants that were not present in dbSNP build 149, the most recent dbSNP version without TOPMed submissions.

RNA splicing and processing proteins (Supplementary Table 7). Genes associated with human disease in COSMIC[17] (31% depletion), the GWAS catalogue[18] (around 8% depletion), OMIM[19] (4% depletion) and ClinVar[20] (4% depletion) all contained fewer rare pLOF variants than expected (each comparison $P < 10^{-4}$).

The distribution of genetic variation

We examined the distribution of variant sites across the genome by counting variants across ordered 1-megabase (Mb) concatenations of contiguous sequence with a similar conservation level (indicated by combined annotation-dependent depletion (CADD score[21]), and in segments categorized by coding versus noncoding status (Fig. 1 and Extended Data Fig. 2). As expected, the vast majority of human genomic variation is rare (minor allele frequency (MAF) < 0.5%)[10,11] and located in putatively neutral, noncoding regions of the genome (Fig. 1). Although coding regions have lower average levels of both common (MAF ≥ 0.5%) and rare variation, we identified some ultra-conserved noncoding regions with even lower levels of genetic variation[22] (Fig. 1 and Supplementary Fig. 15).

Segments with notably high or low levels of variation do exist. For example, one region on chromosome 8p (GRC 38 positions 1,000,001–7,000,000 bp) has the highest overall levels of variation (Extended Data Fig. 2). This is consistent with previous findings, as this region has been shown to have one of the highest mutation rates across the human genome[23].

Although levels of common and rare variation within segments are significantly correlated ($R^2 = 0.462$, $P \leq 2 \times 10^{-16}$) (Supplementary Fig. 16), there are outliers. For example, segments overlapping the major histocompatibility complex (MHC) have the highest levels of common variation but no notable increase in levels of rare variation, consistent with balancing selection[24–26]. A detailed examination of the MHC shows peaks of increased variation and nucleotide diversity consistent with assembly-based analyses of the region[27] (Supplementary Fig. 17). Segments with a high proportion of coding bases feature a strong depletion in the number of common variants but only a modest depletion in rare variants (Supplementary Fig. 18).

Insights into mutation processes

A hallmark of human genetic variation is that SNVs tend to cluster together throughout the genome[3,28]. Such patterns of clustering contain important information about demographic history[29], signals of natural selection[30] and processes that generate mutations[31]. To dissect the spatial clustering of SNVs, we analysed a collection of 50,264,223 singleton SNVs ascertained in a subset of 3,000 unrelated individuals selected to have low levels of genetically estimated admixture—1,000 each of African, East Asian and European ancestry[32] (see Methods).

In these data, we observed that 1.9% of singletons in a given individual occur at distances of less than 100 bp apart[33,34] (Supplementary Figs. 19, 20). In coalescent simulations (see Methods), only 0.16% of the simulated singletons within an individual were less than 100 bp apart (Supplementary Figs. 19, 20). Although demographic history contributes to singleton clustering (Supplementary Information 1.6), population genetic processes alone do not fully account for the observed clustering patterns, particularly for the most closely spaced singletons. To better understand the latent factors that contribute to the observed clustering, we modelled the inter-singleton distance distribution as a mixture of exponential processes (see Methods). The best-fitting version of this model consisted of four mixture components (Fig. 2).

Component 1 represents singletons that occurred an average of around 2–8 bp apart and accounted for approximately 1.5% of singletons in each sample. These singletons are substantially enriched for A>T and C>A transversions (Extended Data Fig. 3a), consistent with the signatures of trans-lesion synthesis that causes multiple non-independent point mutations within very short spans[35]. The density of component 1 singletons is also associated with CpG island density (Supplementary Fig. 21). Component 2 represents singletons occurring 500–5,000 bp apart, accounting for around 12–24% of singletons. These singletons are enriched for C>G transversions and show prominent subtelomeric concentrations on chromosomes 8p, 9p, 16p and 16q[36,37] (Extended Data Fig. 3 and Supplementary Fig. 22), consistent with the recently described maternally derived C>G mutation clusters[36,37]. The exact mechanism that underlies this distinctive clustering pattern is unknown, but may involve either hypermutability of single-stranded DNA intermediates during the repair of double-stranded breaks[36,37] or transcription-associated mutagenesis, with increased damage on the non-transcribed strand[38]. Our results are compatible with both these mechanisms: component 2 singletons are enriched near regions of H3K4 trimethylation, a mark associated with double-stranded break response[39], and depleted in exon-dense regions (Supplementary Fig. 21). Component 3 singletons (occurring approximately 30–50 kilobases (kb) apart) accounted for around 43–49% of all singletons, and component 4 singletons (occurring approximately 125–170 kb apart) accounted for around 31–37% of all singletons. These latter components

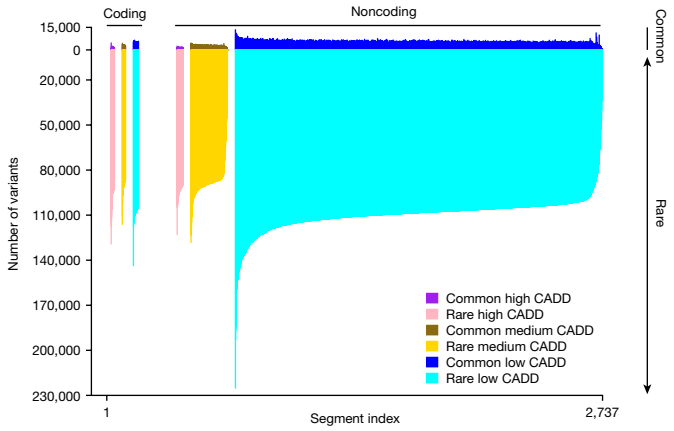


Fig. 1 | Distribution of genetic variants across the genome. Common (allele frequency ≥ 0.5%) and rare (allele frequency < 0.5%) variant counts are shown above and below the *x* axis, respectively, within 1-Mb concatenated segments (see Methods). Segments are stratified by CADD functionality score, and sorted based on their number of rare variants according to the functionality category. There were 22 high CADD, 22 medium CADD and 34 low CADD coding segments, and 40 high CADD, 238 medium CADD and 2,381 low CADD noncoding segments. Noncoding regions of the genome with low CADD scores (<10, reflecting lower predicted function) have the largest levels of common and rare variation (noncoding plot region, dark and light blue, respectively), followed by low CADD coding regions (coding plot region, dark and light blue, respectively). Overall, the vast majority of human genomic variation comprises rare variation.

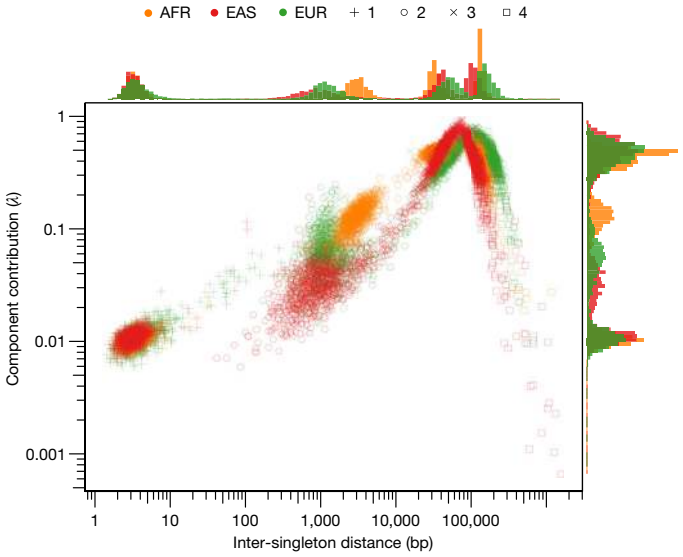


Fig. 2 | Characteristics of singleton clustering patterns. Parameter estimates for exponential mixture models of singleton density. Each point represents one of the four components in one of the 3,000 individuals in the sample, coloured according to the genetically inferred population of that individual. The rate parameters of each component are shown across the *x* axis, and the lambda parameters (that is, the proportion that the component contributes to the mixture) are shown on the *y* axis (on a log–log scale). Histograms show the distribution of the lambda and rate parameters for each component. AFR, African ancestry; EAS, East Asian ancestry; EUR, European ancestry.

have nearly identical mutational spectra (Extended Data Fig. 3a) and are distributed about uniformly in the genome.

Beyond SNVs and indels

To evaluate the potential of our data to generate even more comprehensive variation datasets, we developed and applied a method based on de novo assembly of unmapped and mismapped read pairs, enabling us to assemble sequences that are present in a sample but absent, or improperly represented, in the reference. As the majority of non-reference human sequence is present in the assembled genomes of other primates[40,41], we leveraged available hominid references (see Methods) to specifically discover retained ancestral sequences that have been deleted in some human lineages, including on the reference haplotype.

In total, we placed 1,017 ancestral sequences, of which we were able to fully resolve 713, ranging in length from 100 bp to 39 kb (N50 = 1,183), and accounting for a total of 528,233 bp (Fig. 3a). We partially resolved 304 events, for which we assembled part of the ancestral sequence but could place only one breakpoint on the reference sequence (see Supplementary Information 1.7). Out of all 1,017 events, 551 (54.18%) occur within GENCODE v.29[15] genes (a proportion that is not significantly different from 54.80% of the current reference genome GRCh38 that is within genes). The assembled sequences contain repetitive motifs at a significantly higher rate than the genome as a whole (58.2% versus 50.1%) (Supplementary Tables 8–10). There is a strong overrepresentation of simple and low complexity sequences both in the reference breakpoints and within the bodies of the non-reference sequences, which could be indicative of the instability of these motifs and/or errors in the reference.

Considering only fully resolved events with genotyping rates above 95% (n = 541), we identified between 232 kb and 418 kb of retained ancestral sequence per diploid individual. Allele frequencies of assembled retained sequences are greater than those observed for SNVs and indels, with 76.7% of the assembled sequences present at allele frequency of more than 5% and only 12% of assembled sequences with allele frequency of less than 0.5% (Supplementary Fig. 23). This could reflect difficulty in assembling rare haplotypes. Consistent with observations for SNVs and indels, individuals of African ancestry had, on average, more non-reference alleles (Fig. 3b, Supplementary Fig. 24 and Supplementary Table 11). The overwhelming majority of assembled events are shared by multiple continental groups. We found 58 genic (5 of which are exonic) and 48 intergenic sequences present in a homozygous state in all individuals in the cohort, suggesting that the reference sequence may be incomplete at particular loci, directly affecting the annotation of common forms of genes, such as *UBE2QL1*, *FOXO6* and *FURIN* (Supplementary Fig. 25).

Comparing our findings to two previous short-read studies on different smaller datasets[40,41], 356 sequences (251 kb) are unique to our call set. Additionally, we resolved the length and both breakpoints for 94 events (104 kb) for which only one breakpoint had been reported (Fig. 3c). Further investigation of the overlap with insertions called using long reads on 15 genomes[42], showed that—with a single exception—all previously described events with an allele frequency of more than 12% could be confirmed (Supplementary Fig. 26).

Variation in *CYP2D6*

A complementary approach to de novo genome assembly is to develop approaches that combine multiple types of information—including previously observed haplotype variation, SNVs, indels, copy number and homology information—to identify and classify haplotypes in interesting regions of the genome. One such region is around the *CYP2D6* gene, which encodes an enzyme that metabolizes approximately 25% of prescription drugs and the activity of which varies substantially among individuals[43–45]. More than 150 *CYP2D6* haplotypes have been described, some involving a gene conversion with its nearby non-functional but highly similar paralogue *CYP2D7*.

We performed *CYP2D6* haplotype analysis for all 53,831 TOPMed individuals[43,46]. We called a total of 99 alleles (66 known and 33 novel)

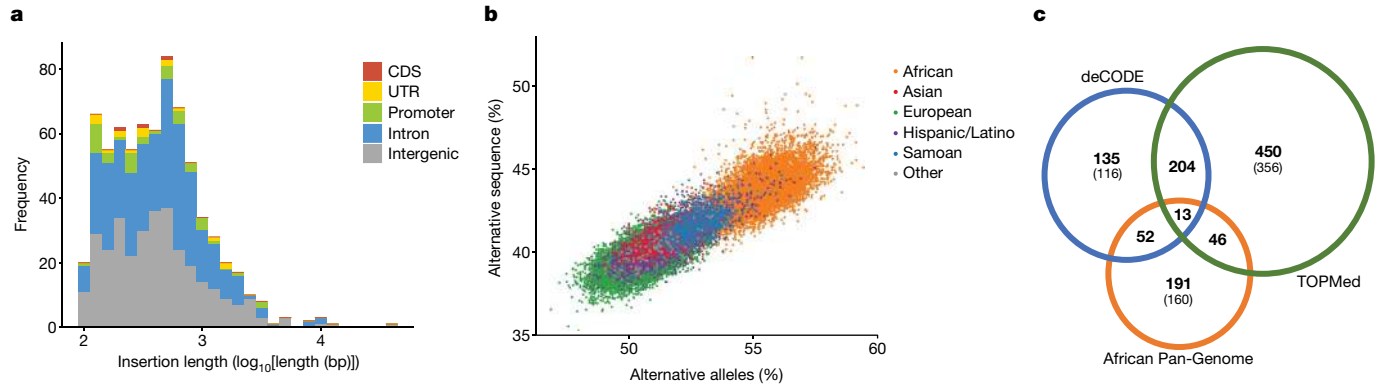


Fig. 3 | Retained non-reference ancestral sequences discovered from unmapped reads. a, Length distribution of fully resolved ancestral sequences, coloured by overlap with GENCODE v.29 genic features. **b**, Percentage of non-reference (alternative) alleles compared with the percentage of non-reference sequence identified per individual, coloured by population group. **c**, Venn diagram showing the positional concordance with insertions identified using short-read data from two previous studies[40,41]. The number of sequences specific to each study and that have not been partially resolved in the other studies is given between brackets.

representing increased function, decreased function and loss of function (Supplementary Table 12). Nineteen of the known alleles and all of the novel alleles are defined by structural variants, including complex *CYP2D6-CYP2D7* hybrids and extensive copy number variation, which ranged from zero to eight gene copies (Supplementary Figs. 27, 28).

Heterozygosity and rare variant sharing

The TOPMed variation data also present an opportunity to examine expectations about rare variation, and to specifically investigate which studies show distinct patterns of variation that might be expected to provide unique insights. To do this, we grouped TOPMed participants by study and by population group, and calculated genetically determined ancestry components, heterozygosity, number of singletons and rare variant sharing (Fig. 4, Supplementary Table 13 and Supplementary Data 1).

As expected, African American and Caribbean population groups have the greatest heterozygosity[7,47], followed by Hispanic/Latino, European American, Amish, East Asian and Samoan groups. This is consistent with a gradual loss of heterozygosity tracking the recent African origin of modern humans and subsequent migration from Africa to the rest of the globe[47,48]. The Asian population groups have among the lowest heterozygosity in our sample (even lower than the Amish, a European ancestry founder population with notably low heterozygosity[49,50]), but also the greatest singleton counts (in contrast to the Amish, who have the lowest; see Supplementary Information 1.8).

Using rare variation, we are also able to distinguish fine-scale patterns of population structure (Fig. 4, Supplementary Fig. 29 and Supplementary Information 1.9). Broadly, we observe sharing between population groups with shared continental ancestry (whether African, European, Asian or American). Nevertheless, additional patterns emerge. The Amish are unique among the included studies: they exhibit little rare variant sharing with outside groups and also the greatest rare variant sharing within the study—consistent with a marked founder effect. Furthermore, we observe an approximately 4× greater rare variant sharing between African American and Caribbean population groups than between European American population groups, even after correcting for sample size differences (Supplementary Fig. 30).

Haplotype sharing

A corollary to rare variant sharing is rare haplotype sharing through segments inherited from a recent common ancestor (Supplementary Figs. 31, 32). The distribution of identical-by-descent segments enables estimates of effective population sizes over the past 300 generations (Extended Data Fig. 4 and Supplementary Fig. 33). The Amish study shows the greatest average levels of within-study identical-by-descent sharing, consistent with a founder event 14 generations ago[50,51]. The demographic histories are broadly similar between population groups, with the exception of the Amish, who experienced a more extreme bottleneck when moving from Europe to America, and Samoan individuals, who have had a smaller effective population size than the East Asian populations from which they separated around 5,000 years ago[52–54]. Both non-Amish European ancestry and African ancestry populations appear to have experienced a bottleneck around 5–10 generations ago, consistent with moving to America, whether through colonization or forced migration (82% of TOPMed participants are US residents).

Large samples alleviate the effects of linkage

The relative numbers of singletons, doubletons and other very rare variants can be used to infer human demographic history[11,55,56]. Although much of demographic inference in past studies focused on fourfold degenerate synonymous sites in protein sequences, these sites evolve under the influence of strong selection at nearby protein-coding sites[57,58], which can affect the inferred timing and magnitude of population size changes[59]. WGS enables us to access intergenic regions of the genome that are minimally affected by selection. We measured how the site frequency spectrum and demographic inference changed as a function of sample size and an index of selection at linked sites (McVicker's *B* statistic[60]) using TOPMed individuals whose genomes suggested mostly European ancestry and low admixture. Estimates of effective population size of European individuals based on the 1% of the genome with the weakest effect of selection at linked sites consistently yielded around 1.1 million individuals (Fig. 5, Supplementary Figs. 34, 35 and Supplementary Table 14).

Human adaptations

When adaptive mutations arise, they can quickly spread. This process generates distinct genomic patterns surrounding the locus, including extended regions of low-diversity haplotypes and few singletons. We scanned for evidence of very recent ongoing positive selection by taking advantage of our WGS data and large samples. We used the singleton density score[61] to search for regions where positive selection has occurred or is ongoing in three ancestry groups: European (n = 21,196), African (n = 2,117) and East Asian (n = 1,355). Broadly, each of these populations showed evidence for adaptation in immune system

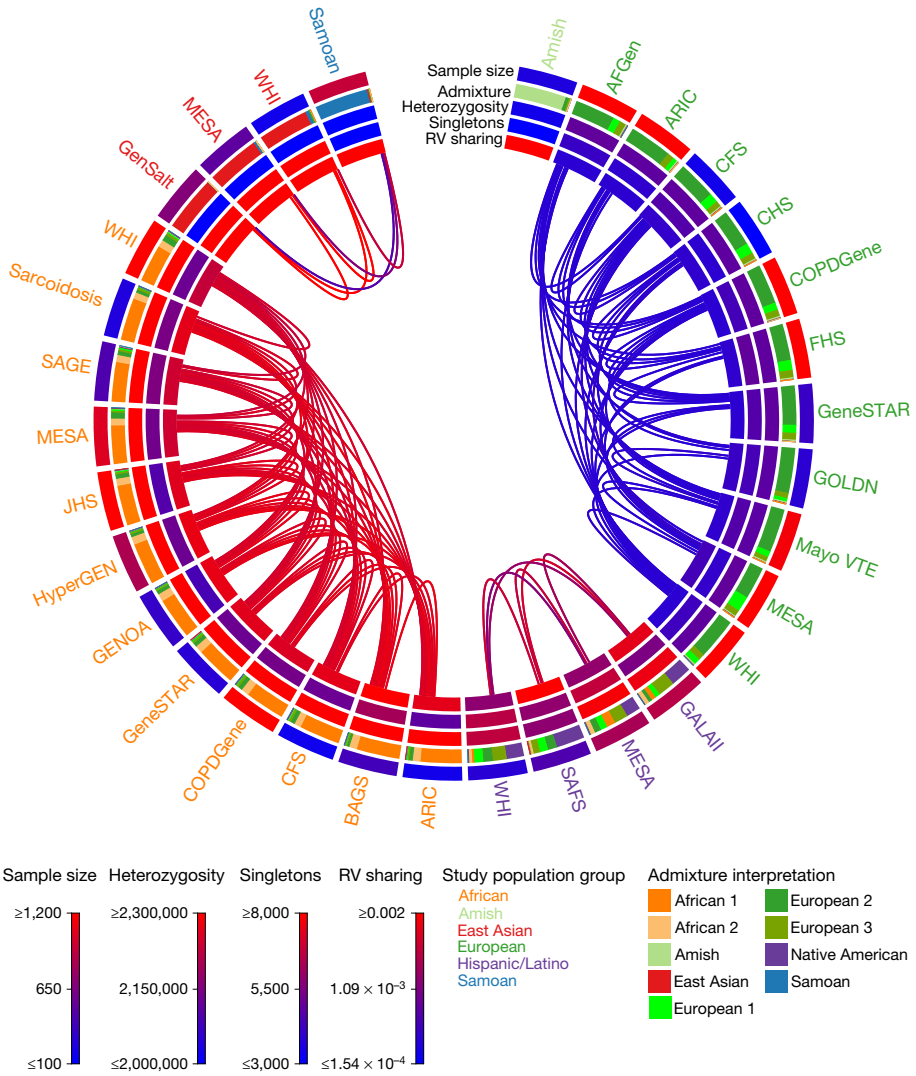


Fig. 4 | Ancestry, genetic diversity and rare-variant genetic relatedness across the TOPMed studies. Each study label is shaded based on their population group. From the outside moving in each track represents: the unrelated sample size of each study used in these calculations, average admixture values, average number of heterozygous sites in each individual's genome, average number of singleton variants in each individual's genome and the average within-study rare-variant (RV) sharing comparisons. The links depict the 75th percentile of between-study rare-variant sharing comparisons. All between-study rare-variant sharing comparisons can be found in Supplementary Fig. 29. The sample size, average heterozygosity, number of singletons, within-cohort rare-variant sharing and admixture values by TOPMed study and population group can be found in Supplementary Table 13. Study name abbreviations are defined in Extended Data Tables 1, 2 and Supplementary Table 20.

genes, albeit with a variety of different gene targets, which probably reflects historical differences in pathogen exposure.

The European population shows selection signals (Supplementary Fig. 36a) in the vicinity of *LCT* and the MHC locus, reflecting well-known signals for adaptation to lactose metabolism and immune system function[61]. We further identify a strong selection signal implicating *HERC2*, a gene that is associated with iris pigmentation[62]. The African population shows a selection signal (Supplementary Fig. 36b) at a locus situated among a cluster of antimicrobial alpha- and beta-defensin genes[63], which has an important role in innate immunity, suggesting a possible adaptive response to environmental pathogens. Other regions implicated include a locus 23 kb upstream of *NRG3*, a previously identified putative target of selection expressed in neural tissue[64,65] and the calcium sensor *STIM1*. Mutations in *STIM1* are known to cause immunodeficiency[66]. The East Asian population shows a selection signal (Supplementary Fig. 36c) at *GJA5*, a gap junction protein that forms intercellular channels to allow transport between cells, and at *PRAG1*, a pseudokinase that interacts with cytoplasmic tyrosine kinase (*CSK*), which ultimately affects antibacterial immune response[67]. Combined with a strong signal at the MHC locus, this once again suggests adaptation in immune system function. We also find evidence of positive selection at two alcohol metabolism genes at mutations known to confer protection against alcoholism: the R48H polymorphism (rs1229984) in *ADH1B*[68,69] and the E504K polymorphism (rs671) in *ALDH2*[70,71].

The TOPMed imputation resource

In addition to enabling detailed analysis of TOPMed sequenced samples, TOPMed can enhance the analysis of any genotyped samples[72]. To this end, we constructed a TOPMed-based imputation reference panel that now includes 97,256 individuals (Extended Data Table 3), including 308,107,085 SNVs and indels (Supplementary Table 15). This is, to our knowledge, the first imputation reference panel that is based exclusively on deep WGS data in diverse samples and greatly exceeds previously published alternatives[7,8]. For example, the average imputation quality r^2 for variants with a frequency of 0.001 in genomes of individuals with an African ancestry increased from around 0.17 in previous panels to 0.96 (Supplementary Fig. 37). Similar improvements were observable in all ancestries that we considered except in South Asian individuals. The minimum allele frequency at which variants could be well-imputed ($r^2 > 0.3$) decreased to around 0.002–0.003% (European or African ancestry in TOPMed). This means that 89% of the approximately 80,000 rare variants with MAF < 0.5% in an average genome of African ancestry can be recovered through genotype imputation using the TOPMed panel.

To illustrate the possibilities, we imputed TOPMed variants in array-genotyped participants of the UK Biobank[2] and compared the results to exome-sequencing data of overlapping individuals. Of the 463,182 exome-sequencing variants with MAF > 0.05% in 49,819 participants of the UK Biobank, the majority (84.86%) were also present in the

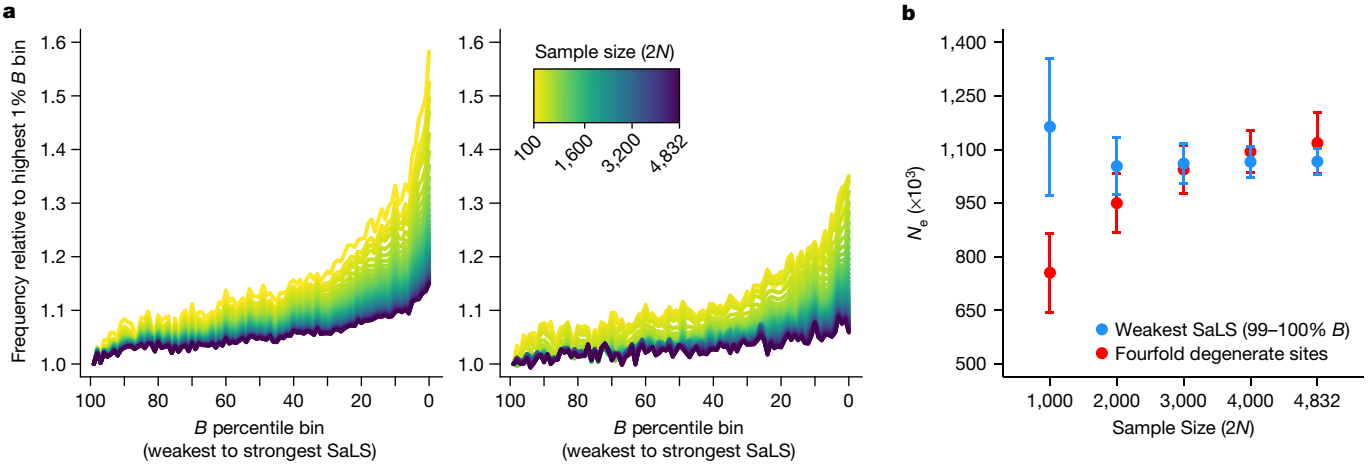


Fig. 5 | Relative increase in singletons and doubletons of the site frequency spectrum across McVicker's *B* and the population size inferred from demographic inference using various sample sizes. a, The relative increase in the singleton (left) and doubleton (right) bins of the site frequency spectrum for decreasing percentile bins of McVicker's *B* compared with the highest percentile bin of *B*. The higher percentiles of *B* indicate weaker effects of selection at linked sites (SaLS). These relative increases are plotted for different sample sizes. **b**, Each point corresponds to the population size inferred in the last generation of an exponential growth model for Europeans. Demographic inference was conducted with different sample sizes for fourfold degenerate sites (n = 4,718,653 sites) and the highest 1% *B* sites (n = 10,977,437 sites). Error bars show 95% confidence intervals (see Supplementary Table 14 for parameter values). N_e, effective population size.

TOPMed-imputed data with imputation quality >0.3. This proportion was lower (52.97%) for 3,587,193 non-singleton exome-sequencing variants with MAF ≤ 0.05%. The TOPMed-imputed genotypes were highly correlated with the exome-sequencing genotypes—the average correlation ranged from 0.73 (MAF ≤ 0.05%) to 0.98 (MAF > 25%) (Supplementary Fig. 38).

An initial association analysis of 94,081 imputed rare autosomal (allele frequency ≤ 0.5%) pLOF variants identified, among other findings, several known rare variant associations with breast cancer: a frameshift variant in *CHEK2* and a stop gain variant in *PALB2* (see Methods and Supplementary Table 16). We also found that the burden of rare pLOF variants in *BRCA2* (comprising 35 rare pLOF variants; $P = 1.6 \times 10^{-8}$; cumulative allele frequency in cases versus controls, 0.13% versus 0.05%) was increased among cases. The individually associated pLOF variants would not have been detected using previous reference panels (Supplementary Table 16). Other examples of rare variant association signals included associations with the burden of rare pLOF variants in *USH2A* and retinal dystrophies (47 rare pLOF variants; allele frequency in cases versus controls, 3% versus 0.2%), *IFT140* and kidney cyst (18 rare pLOF variants; allele frequency in cases versus controls, 0.5% versus 0.1%), and *MYOC* and glaucoma (14 rare pLOF variants; allele frequency in cases versus controls, 0.5% versus 0.1%).

Conclusion and future prospects

We show that TOPMed WGS data provide a rich resource for developing and testing methods for surveying human variation, for inference of human demography and for exploring functional constraints on the genome[73,74]. In addition to these uses, we expect that TOPMed data will improve nearly all ongoing studies of common and rare disorders by providing both a deep catalogue of variation in healthy individuals and an imputation resource that enables array-based studies to achieve a completeness that was previously attainable only through direct sequencing.

Members of the broader scientific community are using TOPMed resources through the WGS and phenotype data available on dbGaP, the BRAVO variant server and the imputation reference panel on the TOPMed imputation server. Full utilization of the programme's resources by the scientific community will require new approaches for dealing with the large size of the omics data, the diversity of the phenotypic data types and structures, and the need to share data in a manner that supports the privacy and consent preferences of participants. These issues are currently being addressed in partnership with the NHLBI BioData Catalyst[75] cloud-computing programme.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03205-y.

1. Mailman, M. D. et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).
2. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
3. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
4. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 431–443 (2020).
5. Bodea, C. A. et al. A method to exploit the structure of genetic ancestry space to enhance case–control studies. *Am. J. Hum. Genet.* **98**, 857–868 (2016).
6. Guo, M. H., Plummer, L., Chan, Y.-M., Hirschhorn, J. N. & Lippincott, M. F. Burden testing of rare variants identified through exome sequencing via publicly available control data. *Am. J. Hum. Genet.* **103**, 522–534 (2018).
7. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
8. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
9. Das, S., Abecasis, G. R. & Browning, B. L. Genotype imputation from large reference panels. *Annu. Rev. Genomics Hum. Genet.* **19**, 73–96 (2018).
10. Fu, W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
11. Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
12. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
13. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
14. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).
15. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47** (D1), D766–D773 (2019).
16. Blyth, C. R. On Simpson's paradox and the sure-thing principle. *J. Am. Stat. Assoc.* **67**, 364–366 (1972).
17. Forbes, S. A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).

18. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP–trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
19. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
20. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46** (D1), D1062–D1067 (2018).
21. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
22. Katzman, S. et al. Human genome ultraconserved elements are ultraselected. *Science* **317**, 915 (2007).
23. Nusbaum, C. et al. DNA sequence and analysis of human chromosome 8. *Nature* **439**, 331–335 (2006).
24. Piertney, S. B. & Oliver, M. K. The evolutionary ecology of the major histocompatibility complex. *Heredity* **96**, 7–21 (2006).
25. Bernatchez, L. & Landry, C. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *J. Evol. Biol.* **16**, 363–377 (2003).
26. Black, F. L. & Hedrick, P. W. Strong balancing selection at HLA loci: evidence from segregation in South Amerindian families. *Proc. Natl Acad. Sci. USA* **94**, 12452–12456 (1997).
27. Jensen, J. M. et al. Assembly and analysis of 100 full MHC haplotypes from the Danish population. *Genome Res.* **27**, 1597–1607 (2017).
28. Hellmann, I. et al. Why do human diversity levels vary at a megabase scale? *Genome Res.* **15**, 1222–1231 (2005).
29. Choudhury, A. et al. Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. *BMC Genomics* **15**, 437 (2014).
30. Torres, R., Szpiech, Z. A. & Hernandez, R. D. Human demographic history has amplified the effects of background selection across the genome. *PLoS Genet.* **14**, e1007387 (2018).
31. Carlson, J. et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat. Commun.* **9**, 3753 (2018).
32. Kessler, M. D. & O'Connor, T. D. Accurate and equitable medical genomic analysis requires an understanding of demography and its influence on sample size and ratio. *Genome Biol.* **18**, 42 (2017).
33. Harris, K. & Nielsen, R. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res.* **24**, 1445–1454 (2014).
34. Besenbacher, S. et al. Multi-nucleotide de novo mutations in humans. *PLoS Genet.* **12**, e1006315 (2016).
35. Waters, L. S. et al. Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. *Microbiol. Mol. Biol. Rev.* **73**, 134–154 (2009).
36. Jónsson, H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
37. Goldmann, J. M. et al. Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **50**, 487–492 (2018).
38. Seplyarskiy, V. B. et al. Population sequencing data reveal a compendium of mutational processes in human germline. Preprint at https://doi.org/10.1101/2020.01.10.893024 (2020).
39. Faucher, D. & Wellinger, R. J. Methylated H3K4, a transcription-associated histone modification, is involved in the DNA damage response pathway. *PLoS Genet.* **6**, e1001082 (2010).
40. Sherman, R. M. et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
41. Kehr, B. et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nat. Genet.* **49**, 588–593 (2017).
42. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675 (2019).
43. Lee, S.-B. et al. Stargazer: a software tool for calling star alleles from next-generation sequencing data using *CYP2D6* as a model. *Genet. Med.* **21**, 361–372 (2019).
44. Zhou, S.-F. Polymorphism of human cytochrome P450 2D6 and its clinical significance: part I. *Clin. Pharmacokinet.* **48**, 689–723 (2009).
45. Crews, K. R. et al. Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update. *Clin. Pharmacol. Ther.* **95**, 376–382 (2014).
46. Lee, S.-B., Wheeler, M. M., Thummel, K. E. & Nickerson, D. A. Calling star alleles with Stargazer in 28 pharmacogenes with whole genome sequences. *Clin. Pharmacol. Ther.* **106**, 1328–1337 (2019).
47. Ramachandran, S. et al. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl Acad. Sci. USA* **102**, 15942–15947 (2005).
48. Li, J. Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
49. .McKusick, V. A. *Medical Genetic Studies of the Amish: Selected Papers* (Johns Hopkins Univ. Press, 1978).
50. Beiler, K. *Fisher Family History* (Eby's Quality Publishing, 1988).
51. Lee, W.-J., Pollin, T. I., O'Connell, J. R., Agarwala, R. & Schäffer, A. A. PedHunter 2.0 and its usage to characterize the founder structure of the Old Order Amish of Lancaster County. *BMC Med. Genet.* **11**, 68 (2010).
52. Wollstein, A. et al. Demographic history of Oceania inferred from genome-wide data. *Curr. Biol.* **20**, 1983–1992 (2010).
53. Lipson, M. et al. Population turnover in remote Oceania shortly after initial settlement. *Curr. Biol.* **28**, 1157–1165 (2018).
54. Harris, D. N. et al. Evolutionary history of modern Samoans. *Proc. Natl Acad. Sci. USA* **117**, 9458–9465 (2020).
55. Gravel, S. et al. Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA* **108**, 11983–11988 (2011).
56. Gao, F. & Keinan, A. Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. *Genetics* **202**, 235–245 (2016).
57. Schrider, D. R., Shanku, A. G. & Kern, A. D. Effects of linked selective sweeps on demographic inference and model selection. *Genetics* **204**, 1207–1223 (2016).
58. Ewing, G. B. & Jensen, J. D. The consequences of not accounting for background selection in demographic inference. *Mol. Ecol.* **25**, 135–141 (2016).
59. Ragsdale, A. P., Moreau, C. & Gravel, S. Genomic inference using diffusion models and the allele frequency spectrum. *Curr. Opin. Genet. Dev.* **53**, 140–147 (2018).
60. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
61. Field, Y. et al. Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).
62. Kayser, M. et al. Three genome-wide association studies and a linkage analysis identify *HERC2* as a human iris color gene. *Am. J. Hum. Genet.* **82**, 411–423 (2008).
63. Ganz, T. & Lehrer, R. I. Defensins. *Pharmacol. Ther.* **66**, 191–205 (1995).
64. Zhang, D. et al. Neuregulin-3 (NRG3): a novel neural tissue-enriched protein that binds and activates ErbB4. *Proc. Natl Acad. Sci. USA* **94**, 9562–9567 (1997).
65. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
66. Picard, C. et al. *STIM1* mutation associated with a syndrome of immunodeficiency and autoimmunity. *N. Engl. J. Med.* **360**, 1971–1980 (2009).
67. Safari, F., Murata-Kamiya, N., Saito, Y. & Hatakeyama, M. Mammalian Pragmin regulates Src family kinases via the Glu-Pro-Ile-Tyr-Ala (EPIYA) motif that is exploited by bacterial effectors. *Proc. Natl Acad. Sci. USA* **108**, 14938–14943 (2011).
68. Jörnvall, H., Hempel, J., Vallee, B. L., Bosron, W. F. & Li, T. K. Human liver alcohol dehydrogenase: amino acid substitution in the beta 2 beta 2 Oriental isozyme explains functional properties, establishes an active site structure, and parallels mutational exchanges in the yeast enzyme. *Proc. Natl Acad. Sci. USA* **81**, 3024–3028 (1984).
69. Osier, M. et al. Linkage disequilibrium at the *ADH2* and *ADH3* loci and risk of alcoholism. *Am. J. Hum. Genet.* **64**, 1147–1157 (1999).
70. Hempel, J., Kaiser, R. & Jörnvall, H. Mitochondrial aldehyde dehydrogenase from human liver. Primary structure, differences in relation to the cytosolic enzyme, and functional correlations. *Eur. J. Biochem.* **153**, 13–28 (1985).
71. Hsu, L. C., Tani, K., Fujiyoshi, T., Kurachi, K. & Yoshida, A. Cloning of cDNAs for human aldehyde dehydrogenases 1 and 2. *Proc. Natl Acad. Sci. USA* **82**, 3771–3775 (1985).
72. Kowalski, M. H. et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* **15**, e1008500 (2019).
73. Bick, A. G. et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).
74. Li, X. et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**, 969–983 (2020).
75. BioData Catalyst Consortium. The NHLBI BioData Catalyst. Zenodo https://doi.org/10.5281/zenodo.3822858 (2020).

Daniel Taliun[1,2,216], Daniel N. Harris[3,4,5,216], Michael D. Kessler[3,4,5,216], Jedidiah Carlson[6,7,216], Zachary A. Szpiech[8,9,216], Raul Torres[10,216], Sarah A. Gagliano Taliun[1,2,216], André Corvelo[11,216], Stephanie M. Gogarten[12], Hyun Min Kang[1,2], Achilleas N. Pitsillides[13], Jonathon LeFaive[1,2], Seung-been Lee[7], Xiaowen Tian[12], Brian L. Browning[14], Sayantan Das[1,2], Anne-Katrin Emde[11], Wayne E. Clarke[11], Douglas P. Loesch[3,4,5], Amol C. Shetty[3,4,5], Thomas W. Blackwell[1,2], Albert V. Smith[1,2], Quenna Wong[12], Xiaoming Liu[15], Matthew P. Conomos[12], Dean M. Bobo[16], François Aguet[17], Christine Albert[18], Alvaro Alonso[19], Kristin G. Ardlie[17], Dan E. Arking[20], Stella Aslibekyan[21], Paul L. Auer[22], John Barnard[23], R. Graham Barr[24,25], Lucas Barwick[26], Lewis C. Becker[27], Rebecca L. Beer[28], Emelia J. Benjamin[29,30,31], Lawrence F. Bielak[32], John Blangero[33,34], Michael Boehnke[1,2], Donald W. Bowden[35], Jennifer A. Brody[36,37], Esteban G. Burchard[38,39], Brian E. Cade[40,41], James F. Casella[42,43], Brandon Chalazan[44], Daniel I. Chasman[45,46], Yii-Der Ida Chen[47], Michael H. Cho[48], Seung Hoan Choi[17], Mina K. Chung[49,50,51], Clary B. Clish[52], Adolfo Correa[53,54,55], Joanne E. Curran[33,34], Brian Custer[56,57], Dawood Darbar[58], Michelle Daya[59], Mariza de Andrade[60], Dawn L. DeMeo[48], Susan K. Dutcher[61,62], Patrick T. Ellinor[63], Leslie S. Emery[12], Celeste Eng[39], Diane Fatkin[64,65,66], Tasha Fingerlin[67], Lukas Forer[68], Myriam Fornage[69], Nora Franceschini[70], Christian Fuchsberger[1,2,68,71], Stephanie M. Fullerton[72], Soren Germer[11], Mark T. Gladwin[73,74,75], Daniel J. Gottlieb[76,77], Xiuqing Guo[47], Michael E. Hall[53], Jiang He[78,79], Nancy L. Heard-Costa[31,80], Susan R. Heckbert[37,81], Marguerite R. Irvin[82], Jill M. Johnsen[36,83], Andrew D. Johnson[31,84], Robert Kaplan[85], Sharon L. R. Kardia[32], Tanika Kelly[78], Shannon Kelly[86,87,88], Eimear E. Kenny[16], Douglas P. Kiel[17,40,89,90], Robert Klemmer[1,2], Barbara A. Konkle[36,83], Charles Kooperberg[91], Anna Köttgen[92,93], Leslie A. Lange[94],

Jessica Lasky-Su[40,41,48,95], Daniel Levy[29,31,84], Xihong Lin[96], Keng-Han Lin[1,2], Chunyu Liu[13], Ruth J. F. Loos[97,98], Lori Garman[99], Robert Gerszten[100], Steven A. Lubitz[18], Kathryn L. Lunetta[13], Angel C. Y. Mak[39], Ani Manichaikul[101,102], Alisa K. Manning[40,103,104], Rasika A. Mathias[105], David D. McManus[106], Stephen T. McGarvey[107,108,109], James B. Meigs[110], Deborah A. Meyers[111], Julie L. Mikulla[28], Mollie A. Minear[28], Braxton D. Mitchell[4,5,112], Sanghamitra Mohanty[113,114], May E. Montasser[4,5], Courtney Montgomery[99], Alanna C. Morrison[115], Joanne M. Murabito[29], Andrea Natale[113], Pradeep Natarajan[40,63,116,117], Sarah C. Nelson[12], Kari E. North[70], Jeffrey R. O'Connell[4,5], Nicholette D. Palmer[35], Nathan Pankratz[118], Gina M. Peloso[13], Patricia A. Peyser[32], Jacob Pleiness[1,2], Wendy S. Post[119], Bruce M. Psaty[36,37,81,120,121], D. C. Rao[122], Susan Redline[40,41], Alexander P. Reiner[81,91], Dan Roden[123], Jerome I. Rotter[47], Ingo Ruczinski[124], Chloé Sarnowski[13], Sebastian Schoenherr[68], David A. Schwartz[125], Jeong-Sun Seo[126,127,128], Sudha Seshadri[31,129], Vivien A. Sheehan[130,131], Wayne H. Sheu[132], M. Benjamin Shoemaker[123], Nicholas L. Smith[81,121,133], Jennifer A. Smith[32,134], Nona Sotoodehnia[37], Adrienne M. Stilp[12], Weihong Tang[135], Kent D. Taylor[47], Marilyn Telen[136], Timothy A. Thornton[12], Russell P. Tracy[137], David J. Van Den Berg[138], Ramachandran S. Vasan[29,31], Karine A. Viaud-Martinez[139], Scott Vrieze[140], Daniel E. Weeks[141,142], Bruce S. Weir[12], Scott T. Weiss[40,41,48,95], Lu-Chen Weng[18], Cristen J. Willer[6,143,144], Yingze Zhang[73,74,75], Xutong Zhao[1,2], Donna K. Arnett[145], Allison E. Ashley-Koch[146], Kathleen C. Barnes[59], Eric Boerwinkle[147,148], Stacey Gabriel[17], Richard Gibbs[148], Kenneth M. Rice[12], Stephen S. Rich[101,102], Edwin K. Silverman[48], Pankaj Qasba[28], Weiniu Gan[28], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium*, George J. Papanicolaou[28], Deborah A. Nickerson[7,149,150], Sharon R. Browning[12], Michael C. Zody[11], Sebastian Zöllner[1,2,151], James G. Wilson[152], L. Adrienne Cupples[13,31]✉, Cathy C. Laurie[12]✉, Cashell E. Jaquish[28]✉, Ryan D. Hernandez[38,153,154,155,156]✉, Timothy D. O'Connor[3,4,5]✉ & Gonçalo R. Abecasis[1]✉

[1]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA. [2]Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI, USA. [3]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA. [4]Program in Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. [5]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. [6]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [7]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [8]Department of Biology, Pennsylvania State University, University Park, PA, USA. [9]Institute for Computational and Data Sciences, Pennsylvania State University, University Park, PA, USA. [10]Biomedical Sciences Graduate Program, University of California, San Francisco, San Francisco, CA, USA. [11]New York Genome Center, New York, NY, USA. [12]Department of Biostatistics, University of Washington, Seattle, WA, USA. [13]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [14]Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA, USA. [15]USF Genomics, College of Public Health, University of South Florida, Tampa, FL, USA. [16]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [17]The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [18]Massachusetts General Hospital, Boston, MA, USA. [19]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA. [20]McKusick-Nathans Institute, Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [21]University of Alabama, Birmingham, AL, USA. [22]Zilber School of Public Health, University of Wisconsin Milwaukee, Milwaukee, WI, USA. [23]Cleveland Clinic, Cleveland, OH, USA. [24]Department of Medicine, Columbia University Medical Center, New York, NY, USA. [25]Department of Epidemiology, Columbia University Medical Center, New York, NY, USA. [26]The Emmes Corporation, Rockville, MD, USA. [27]Johns Hopkins University, Baltimore, MD, USA. [28]National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA. [29]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [30]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA. [31]Framingham Heart Study, Framingham, MA, USA. [32]Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, USA. [33]Department of Human Genetics, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA. [34]South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA. [35]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [36]Department of Medicine, University of Washington, Seattle, WA, USA. [37]Cardiovascular Health Research Unit, University of Washington, Seattle, WA, USA. [38]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA. [39]Department of Medicine, University of California, San Francisco, San Francisco, CA, USA. [40]Department of Medicine, Harvard Medical School, Boston, MA, USA. [41]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [42]Department of Pediatrics, Johns Hopkins University, Baltimore, MD, USA. [43]Division of Pediatric Hematology, Johns Hopkins University, Baltimore, MD, USA. [44]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. [45]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA. [46]Harvard Medical School, Boston, MA, USA. [47]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation, Harbor-UCLA Medical Center, Torrance, CA, USA. [48]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [49]Department of Cardiovascular Medicine, Heart & Vascular Institute, Cleveland Clinic, Cleveland, OH, USA. [50]Department of Cardiovascular and Metabolic Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [51]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH, USA. [52]Metabolomics Platform, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [53]Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA. [54]Department of Pediatrics, University of Mississippi Medical Center, Jackson, MS, USA. [55]Department of Population Health Science, University of Mississippi Medical Center, Jackson, MS, USA. [56]Vitalant Research Institute, San Francisco, CA, USA. [57]Department of Laboratory Medicine, University of California, San Francisco, San Francisco, CA, USA. [58]Department of Medicine, University of Illinois at Chicago, Chicago, IL, USA. [59]Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [60]Mayo Clinic, Rochester, MN, USA. [61]McDonnell Genome Institute, Washington University, St Louis, MO, USA. [62]Department of Genetics, Washington University, St Louis, MO, USA. [63]Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [64]Molecular Cardiology Division, Victor Chang Cardiac Research Institute, Darlinghurst, New South Wales, Australia. [65]Faculty of Medicine, University of New South Wales, Kensington, New South Wales, Australia. [66]Cardiology Department, St Vincent's Hospital, Darlinghurst, New South Wales, Australia. [67]National Jewish Health, Center for Genes, Environment and Health, Denver, CO, USA. [68]Institute of Genetic Epidemiology, Department of Genetics and Pharmacology, Medical University of Innsbruck, Innsbruck, Austria. [69]Institute of Molecular Medicine, University of Texas Health Science Center at Houston, Houston, TX, USA. [70]Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA. [71]Institute for Biomedicine, Eurac Research, Bolzano, Italy. [72]Department of Bioethics & Humanities, University of Washington School of Medicine, Seattle, WA, USA. [73]Pittsburgh Heart, Lung, Blood and Vascular Medicine Institute, University of Pittsburgh, Pittsburgh, PA, USA. [74]Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [75]Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [76]VA Boston Healthcare System, Boston, MA, USA. [77]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA, USA. [78]Department of Epidemiology, Tulane University, New Orleans, LA, USA. [79]Tulane University Translational Science Institute, Tulane University, New Orleans, LA, USA. [80]Department of Neurology, Boston University School of Medicine, Boston, MA, USA. [81]Department of Epidemiology, University of Washington, Seattle, WA, USA. [82]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA. [83]Bloodworks Northwest Research Institute, Seattle, WA, USA. [84]Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Framingham, MA, USA. [85]Albert Einstein College of Medicine, New York, NY, USA. [86]Department of Epidemiology, Vitalant Research Institute, San Francisco, CA, USA. [87]Department of Pediatrics, UCSF Benioff Children's Hospital, Oakland, CA, USA. [88]Division of Pediatric Hematology, UCSF Benioff Children's Hospital, Oakland, CA, USA. [89]Hinda and Arthur Marcus Institute for Aging Research, Hebrew SeniorLife, Boston, MA, USA. [90]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [91]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [92]Department of Epidemiology, Johns Hopkins University, Baltimore, MD, USA. [93]Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany. [94]Department of Medicine, University of Colorado at Denver, Aurora, CO, USA. [95]Brigham and Women's Hospital, Boston, MA, USA. [96]Biostatistics and Statistics, Harvard University, Boston, MA, USA. [97]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [98]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [99]Department of Genes and Human Disease, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA. [100]Beth Israel Deaconess Medical Center, Boston, MA, USA. [101]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [102]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA. [103]Clinical and Translational Epidemiology Unit, Mongan Institute, Massachusetts General Hospital, Boston, MA, USA. [104]Metabolism Program, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [105]Department of Medicine, Johns Hopkins University, Baltimore, MD, USA. [106]Cardiovascular Medicine, University of Massachusetts Medical School, Worcester, MA, USA. [107]International Health Institute, Brown University, Providence, RI, USA. [108]Department of Epidemiology, Brown University, Providence, RI, USA. [109]Department of Anthropology, Brown University, Providence, RI, USA. [110]Division of General Internal Medicine, Massachusetts General Hospital, Harvard Medical School, The Broad Institute of MIT and Harvard, Boston, MA, USA. [111]University of Arizona, Tucson, AZ, USA. [112]Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD, USA. [113]Texas Cardiac Arrhythmia Institute, St David's Medical Center, Austin, TX, USA. [114]Department of Internal Medicine, Dell Medical School, Austin, TX, USA. [115]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA. [116]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [117]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [118]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA. [119]Division of Cardiology, Department of Medicine, Johns Hopkins University, Baltimore, MD, USA. [120]Department of Health Services, University of Washington, Seattle, WA, USA. [121]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA. [122]Division of Biostatistics, Washington University in St Louis, St Louis, MO, USA. [123]Vanderbilt University Medical Center, Nashville, TN, USA. [124]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [125]University of Colorado at Denver, Denver, CO, USA. [126]Precision Medicine Center, Seoul National University Bundang Hospital, Seongnam, Republic of Korea. [127]Macrogen Inc, Seoul, Republic of Korea. [128]Gong Wu Genomic Medicine Institute, Seoul National University

Bundang Hospital, Seongnam, Republic of Korea. [129]Glenn Biggs Institute for Alzheimer's and Neurodegenerative Diseases, University of Texas Health Sciences Center at San Antonio, San Antonio, TX, USA. [130]Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, USA. [131]Aflac Cancer and Blood Disorders Center, Children's Healthcare of Atlanta, Atlanta, GA, USA. [132]Taichung Veterans General Hospital Taiwan, Taichung City, Taiwan. [133]Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research and Development, Seattle, WA, USA. [134]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA. [135]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, USA. [136]Duke University, Durham, NC, USA. [137]Department of Pathology & Laboratory Medicine, University of Vermont Larner College of Medicine, Burlington, VT, USA. [138]Center for Genetic Epidemiology, Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA. [139]Illumina Laboratory Services, Illumina Inc, San Diego, CA, USA. [140]Department of Psychology, University of Minnesota, Minneapolis, MN, USA. [141]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA. [142]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA. [143]Department of Internal Medicine-Cardiology, University of Michigan, Ann Arbor, MI, USA. [144]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [145]Department of Epidemiology, University of Kentucky, Lexington, KY, USA. [146]Duke Molecular Physiology Institute, Duke University Medical Center, Durham, NC, USA. [147]University of Texas Health Science Center at Houston, Houston, TX, USA. [148]Baylor College of Medicine Human Genome Sequencing Center, Houston, TX, USA. [149]Northwest Genomics Center, Seattle, WA, USA. [150]Brotman Baty Institute, Seattle, WA, USA. [151]Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA. [152]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA. [153]Department of Human Genetics, McGill University, Montreal, Quebec, Canada. [154]Quantitative Biosciences Institute, University of California, San Francisco, San Francisco, CA, USA. [155]Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. [156]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA. [216]These authors contributed equally: Daniel Taliun, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, André Corvelo. *A list of authors and their affiliations appears in the online version of the paper. [52]e-mail: adrienne@bu.edu; cclaurie@uw.edu; jaquishc@nhlbi.nih.gov; ryan.hernandez@mcgill.ca; timothydoconnor@gmail.com; goncalo@umich.edu

NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

Namiko Abe[11], Laura Almasy[157], Seth Ament[158], Peter Anderson[159], Pramod Anugu[160], Deborah Applebaum-Bowden[161], Tim Assimes[162], Dimitrios Avramopoulos[27], Emily Barron-Casella[27], Terri Beaty[27], Gerald Beck[23], Diane Becker[27], Amber Beitelshees[158], Takis Benos[163], Marcos Bezerra[164], Joshua Bis[159], Russell Bowler[165], Ulrich Broeckel[166], Jai Broome[159], Karen Bunting[11], Carlos Bustamante[162], Erin Buth[159], Jonathan Cardwell[125], Vincent Carey[95], Cara Carty[167], Richard Casaburi[168], Peter Castaldi[95], Mark Chaffin[169], Christy Chang[158], Yi-Cheng Chang[170], Sameer Chavan[125], Bo-Juen Chen[11], Wei-Min Chen[171], Lee-Ming Chuang[170], Ren-Hua Chung[172], Suzy Comhair[23], Elaine Cornell[173], Carolyn Crandall[168], James Crapo[165], Jeffrey Curtis[174], Coleen Damcott[158], Sean David[175], Colleen Davis[159], Lisa de las Fuentes[176], Michael DeBaun[177], Ranjan Deka[178], Scott Devine[158], Qing Duan[179], Ravi Duggirala[180], Jon Peter Durda[173], Charles Eaton[181], Lynette Ekunwe[160], Adel El Boueiz[182], Serpil Erzurum[23], Charles Farber[171], Matthew Flickinger[174], Myriam Fornage[183], Chris Frazar[159], Mao Fu[158], Lucinda Fulton[176], Shanshan Gao[125], Yan Gao[160], Margery Gass[184], Bruce Gelb[16], Xiaoqi Priscilla Geng[174], Mark Geraci[185], Auyon Ghosh[95], Chris Gignoux[162], David Glahn[186], Da-Wei Gong[158], Harald Goring[187], Sharon Graw[188], Daniel Grine[125], C. Charles Gu[176], Yue Guan[158], Namrata Gupta[169], Jeff Haessler[184], Nicola L. Hawley[186], Ben Heavner[159], David Herrington[189], Craig Hersh[95], Bertha Hidalgo[21], James Hixson[183], Brian Hobbs[95], John Hokanson[125], Elliott Hong[158], Karin Hoth[190], Chao Agnes Hsiung[172], Yi-Jen Hung[191], Haley Huston[192], Chii Min Hwu[132], Rebecca Jackson[193], Deepti Jain[159], Min A. Jhun[174], Craig Johnson[159], Rich Johnston[194], Kimberly Jones[27], Sekar Kathiresan[169], Alyna Khan[159], Wonji Kim[182], Greg Kinney[125], Holly Kramer[195], Christoph Lange[196], Ethan Lange[125], Leslie Lange[125], Cecelia Laurie[159], Meryl LeBoff[95], Jiwon Lee[95], Seunggeun Shawn Lee[174], Wen-Jane Lee[132], David Levine[159], Joshua Lewis[158], Xiaohui Li[197], Yun Li[179], Henry Lin[197], Honghuang Lin[198], Keng Han Lin[174], Simin Liu[181], Yongmei Liu[136], Yu Liu[199], James Luo[28], Michael Mahaney[200], Barry Make[27], JoAnn Manson[95], Lauren Margolin[169], Lisa Martin[201], Susan Mathai[125], Susanne May[159], Patrick McArdle[158], Merry-Lynn McDonald[21], Sean McFarland[202], Daniel McGoldrick[159], Caitlin McHugh[159], Hao Mei[160], Luisa Mestroni[188], Nancy Min[160], Ryan L. Minster[163], Matt Moll[95], Arden Moscati[16], Solomon Musani[160], Stanford Mwasongwe[160], Josyf C. Mychaleckyj[171], Girish Nadkarni[16], Rakhi Naik[27], Take Naseri[203], Sergei Nekhai[204], Bonnie Neltner[125], Heather Ochs-Balcom[205], David Paik[162], James Pankow[206], Afshin Parsa[158], Juan Manuel Peralta[180], Marco Perez[162], James Perry[158], Ulrike Peters[184], Lawrence S. Phillips[194], Toni Pollin[158], Julia Powers Becker[125], Meher Preethi Boorgula[125], Michael Preuss[16], Dandi Qiao[95], Zhaohui Qin[194], Nicholas Rafaels[125], Laura Raffield[179], Laura Rasmussen-Torvik[207], Aakrosh Ratan[171], Robert Reed[158], Elizabeth Regan[165], Muagututi'a Sefuiva Reupena[208], Carolina Roselli[169], Pamela Russell[125], Sarah Ruuska[192], Kathleen Ryan[158], Ester Cerdeira Sabino[209], Danish Saleheen[210], Shabnam Salimi[158], Steven Salzberg[27], Kevin Sandow[197], Vijay G. Sankaran[211], Christopher Scheller[174], Ellen Schmidt[174], Karen Schwander[176], Frank Sciurba[163], Christine Seidman[46], Jonathan Seidman[46], Stephanie L. Sherman[194], Aniket Shetty[125], Wayne Hui-Heng Sheu[132], Brian Silver[212], Josh Smith[159], Tanja Smith[11], Sylvia Smoller[85], Beverly Snively[189], Michael Snyder[162], Tamar Sofer[95], Garrett Storm[125], Elizabeth Streeten[158], Yun Ju Sung[176], Jody Sylvia[95], Adam Szpiro[159], Carole Sztalryd[158], Hua Tang[162], Margaret Taub[27], Matthew Taylor[125], Simeon Taylor[158], Machiko Threlkeld[159], Lesley Tinker[184], David Tirschwell[159], Sarah Tishkoff[213], Hemant Tiwari[21], Catherine Tong[159], Michael Tsai[206], Dhananjay Vaidya[27], Peter VandeHaar[174], Tarik Walker[125], Robert Wallace[190], Avram Walts[125], Fei Fei Wang[159], Heming Wang[95], Karol Watson[168], Jennifer Wessel[185], Kayleen Williams[159], L. Keoki Williams[214], Carla Wilson[95], Joseph Wu[162], Huichun Xu[158], Lisa Yanek[27], Ivana Yang[125], Rongze Yang[158], Norann Zaghloul[158], Maryam Zekavat[169], Snow Xueyan Zhao[165], Wei Zhao[174], Degui Zhi[183], Xiang Zhou[174] & Xiaofeng Zhu[215]

[157]Children's Hospital of Philadelphia, Philadelphia, PA, USA. [158]University of Maryland, Baltimore, MD, USA. [159]University of Washington, Seattle, WA, USA. [160]University of Mississippi, Jackson, MS, USA. [161]National Institutes of Health, Bethesda, MD, USA. [162]Stanford University, Stanford, CA, USA. [163]University of Pittsburgh, Pittsburgh, PA, USA. [164]Fundação de Hematologia e Hemoterapia de Pernambuco–Hemope, Recife, Brazil. [165]National Jewish Health, Denver, CO, USA. [166]Medical College of Wisconsin, Milwaukee, WI, USA. [167]Washington State University, Seattle, WA, USA. [168]University of California, Los Angeles, Los Angeles, CA, USA. [169]Broad Institute, Cambridge, MA, USA. [170]National Taiwan University, Taipei, Taiwan. [171]University of Virginia, Charlottesville, VA, USA. [172]National Health Research Institute Taiwan, Zhunan Township, Taiwan. [173]University of Vermont, Burlington, VT, USA. [174]University of Michigan, Ann Arbor, MI, USA. [175]University of Chicago, Chicago, IL, USA. [176]Washington University in St Louis, St Louis, MO, USA. [177]Vanderbilt University, Nashville, TN, USA. [178]University of Cincinnati, Cincinnati, OH, USA. [179]University of North Carolina, Chapel Hill, NC, USA. [180]University of Texas Rio Grande Valley School of Medicine, Edinburg, TX, USA. [181]Brown University, Providence, RI, USA. [182]Harvard University, Boston, MA, USA. [183]University of Texas Health at Houston, Houston, TX, USA. [184]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [185]Indiana University, Indianapolis, IN, USA. [186]Yale University, New Haven, CT, USA. [187]University of Texas Rio Grande Valley School of Medicine, San Antonio, TX, USA. [188]University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [189]Wake Forest Baptist Health, Winston-Salem, NC, USA. [190]University of Iowa, Iowa City, IA, USA. [191]Tri-Service General Hospital National Defense Medical Center, Taipei, Taiwan. [192]Blood Works Northwest, Seattle, WA, USA. [193]Ohio State University Wexner Medical Center, Columbus, OH, USA. [194]Emory University, Atlanta, GA, USA. [195]Loyola University, Maywood, IL, USA. [196]Harvard School of Public Health, Boston, MA, USA. [197]Lundquist Institute, Torrance, CA, USA. [198]Boston University, Boston, MA, USA. [199]Stanford University, Palo Alto, CA, USA. [200]University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA. [201]George Washington University, Washington, DC, USA. [202]Harvard University, Cambridge, MA, USA. [203]Ministry of Health, Government of Samoa, Apia, Samoa. [204]Howard University, Washington, DC, USA. [205]University at Buffalo, Buffalo, NY, USA. [206]University of Minnesota, Minneapolis, MN, USA. [207]Northwestern University, Chicago, IL, USA. [208]Lutia I Puava Ae Mapu I Fagalele, Apia, Samoa. [209]Universidade de Sao Paulo, Sao Paulo, Brazil. [210]Columbia University, New York, NY, USA. [211]Broad Institute, Harvard University, Boston, MA, USA. [212]UMass Memorial Medical Center, Worcester, MA, USA. [213]University of Pennsylvania, Philadelphia, PA, USA. [214]Henry Ford Health System, Detroit, MI, USA. [215]Case Western Reserve University, Cleveland, OH, USA.

Methods

DNA samples

WGS for the 53,831 samples reported here was performed on samples that had previously been collected from and consented to by research participants from 33 NHLBI-funded research projects. All studies were approved by the corresponding institutional review boards (Supplementary Information 4). All sequencing was done from DNA extracted from whole blood, with the exception of 17 Framingham samples (lymphoblastoid cell lines) and HapMap samples NA12878 and NA19238 (lymphoblastoid cell lines) used periodically as sequencing controls. Cell lines were tested for mycoplasma contamination by aligning sequence data to the human genome, and authenticated by comparison with previous genetic analysis.

WGS

WGS targeting a mean depth of at least 30× (paired-end, 150-bp reads) using Illumina HiSeq X Ten instruments was carried out over several years at six sequencing centres (Supplementary Table 17). All sequencing used PCR-free library preparation kits purchased from KAPA Biosystems, equivalent to the protocol in the Illumina TruSeq PCR-Free Sample Preparation Guide (Illumina, FC-121-2001). Centre-specific details are available from the TOPMed website (https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-project-freeze-5b-phases-1-and-2). In addition, 30× coverage WGS for 1,606 samples from four contributing studies were sequenced before the start of the TOPMed sequencing project and are included in this dataset. These were sequenced at Illumina using HiSeq 2000 or 2500 instruments, have 2 × 100-bp or 2 × 125-bp paired-end reads and sometimes used PCR amplification.

Sequence data processing and variant calling

Sequence data processing was performed periodically to produce genotype data 'freezes' that included all samples available at the time. All sequences were remapped using BWA-MEM[76] to the hs38DH 1000 Genomes build 38 human genome reference including decoy sequences, following the protocol published previously[77]. Variant discovery and genotype calling was performed jointly, across TOPMed studies, for all samples in a given freeze using the GotCloud[78,79] pipeline. This procedure results in a single, multi-study genotype call set. A support vector machine quality filter for variant sites was trained using a large set of site-specific quality metrics and known variants from arrays and the 1000 Genomes Project as positive controls and variants with Mendelian inconsistencies in multiple families as negative controls (see online documentation[80] for more details). After removing all sites with a minor allele count less than 2, the genotypes with a minimal depth of more than 10× were phased using Eagle 2.4[81]. Sample-level quality control included checks for pedigree errors, discrepancies between self-reported and genetic sex, and concordance with previous genotyping array data. Any errors detected were addressed before dbGaP submission. Details regarding WGS data acquisition, processing and quality control vary among the TOPMed data freezes. Freeze-specific methods are described on the TOPMed website (https://www.nhlbiwgs.org/data-sets) and in documents included in each TOPMed accession released on dbGaP (for example, see document phd008024.1 in phs000956.v4.p1).

Access to sequence data

Copies of individual-level sequence data for each study participant are stored on both Google and Amazon clouds. Access involves an approved dbGaP data access request and is mediated by the NCBI Sequence Data Delivery Pilot mechanism. This mechanism uses fusera software[82] running on the user's cloud instance to handle authentication and authorization with dbGaP. It provides read access to sequence data for one or more TOPMed (or other) samples as .cram files (with associated .crai index files) within a fuse virtual file system mounted on the cloud computing instance. Samples are identified by 'SRR' run accession numbers assigned in the NCBI Sequence Read Archive (SRA) database and shown under each study's phs number in the SRA Run Selector (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi). The phs numbers for all TOPMed studies are readily found by searching dbGaP for the string 'TOPMed'. The fusera software is limited to running on Google or Amazon cloud instances to avoid incurring data egress charges. Fusera, samtools and other tools are also packaged in a Docker container for ease of use and are available for download from Docker Hub[83].

Sample sets

Several sample sets derived from three different WGS data freezes were used in the analyses presented here: freeze 3 (GRCh37 alignment, around 18,000 samples jointly called in 2016), freeze 5 (GRCh38 alignment, approximately 65,000 samples jointly called in 2017), and freeze 8 (GRCh38 alignment, about 140,000 samples jointly called in 2019). Extended Data Table 3 indicates which TOPMed study-consent groups were used in each of several different types of analyses described in this paper. Most analyses were performed on a set of 53,831 samples derived from freeze 5 ('General variant analyses' in Extended Data Table 3) or on a subset thereof approved for population genetic studies ('Population genetics' in Extended Data Table 3). The set of 53,831 was selected from freeze 5 using samples eligible for dbGaP sharing at the time of analysis, excluding (1) duplicate samples from the same participant; (2) one member of each monozygotic twin pair; (3) samples with questionable identity or low read depth (<98% of variant sites at depth ≥ 10×); and (4) samples with consent types inconsistent with analyses presented here. The 'unrelated' sample set consisting of 40,722 samples refers to a subset of the 53,831 samples of individuals who are unrelated with a threshold of third degree (less closely related than first cousins), identified using the PC-AiR method[84]. Exact numbers of samples used in each analysis are listed in Supplementary Table 18.

High-coverage whole-exome sequencing in BioMe study

From around 10,000 BioMe study samples present in TOPMed freeze 8, we randomly selected 1,000 samples for which whole-exome sequencing (WES) data were available. These samples were whole-exome sequenced using Illumina v4 HiSeq 2500 at an average 36.4× depth. Genetic variants were jointly called using the GATK v.3.5.0 pipeline across all 31,250 BioMe samples with WES data. A series of quality control filters, known as the Goldilocks filter, were applied before data delivery to the Charles Bronfman Institute for Personalized Medicine (IPM). First, a series of filters was applied to particular cells comprising combinations of sites and samples—that is, genotypic information for one individual at one locus. Quality scores were normalized by depth of coverage and used with depth of coverage itself to filter sites, using different thresholds for SNVs and short indels. For SNVs, cells with depth-normalized quality scores less than 3, or depth of coverage less than 7 are set to missing. For indels, cells with depth-normalized quality scores less than 5, or depth of coverage less than 10 are set to missing. Then, variant sites were filtered, such that all samples carrying variation have heterozygous (0/1) genotype calls and all samples carrying heterozygous variation fail the allele balance cut-off; these sites were removed from the dataset at this stage. The allele balance cut-off, as with the depth and quality scores used for cell filtering above, differed depending on whether the site was a SNV or an indel: SNVs require at least one sample to carry an alternative allele balance ≥ 15%, and indels require at least one sample to carry an alternative allele balance ≥ 20%. These filters resulted in the removal of 441,406 sites, leaving 8,761,478 variants in the dataset. After subsetting to 1,000 randomly selected individuals, we had 1,076,707 autosomal variants that passed quality control. We further removed variants with call rate <99% (that is, missing in more than 10 individuals), reducing the number of analysed autosomal variants to 1,044,517. The comparison results of TOPMed

WGS and BioMe WES data are described in Supplementary Information 1.3.1.

Low-coverage WGS and high-coverage WES in the Framingham Heart Study

Investigators of the Framingham Heart Study (FHS) evaluated WGS data from TOPMed in comparison with sequencing data from CHARGE Consortium WGS and WES datasets. Supplementary Table 19 provides the counts and depth of each sequencing effort. The overlap of these three groups is 430 FHS study participants, on whom we report here. We use a subset of 263 unrelated study participants to calculate the numbers of singletons and doubletons, MAF, heterozygosity and all rates, to avoid bias from the family structure. Supplementary Information 1.3.2 provides further detail on the sequencing efforts and a detailed description of the comparison results.

Identifying pLOF variants

pLOF variants were identified using Loss Of Function Transcript Effect Estimator (LOFTEE) v.0.3-beta[85] and Variant Effect Predictor (VEP) v.94[86]. The genomic coordinates of coding elements were based on GENCODE v.29[15]. Only stop-gained, frameshift and splice-site-disturbing variants annotated as high-confidence pLOF variants were used in the analysis. The pLOF variants with allele frequency > 0.5% or within regions masked due to poor accessibility were excluded from analysis (see Supplementary Information 1.5 for details).

We evaluated the enrichment and depletion of pLOF variants (allele frequency < 0.5%) in gene sets (that is, terms) from Gene Ontology (GO)[87,88]. For each gene annotated with a particular GO term, we computed the number of pLOF variants per protein-coding base pair, L, and proportion of singletons, S. We then tested for lower or higher mean L and S in a GO term using bootstrapping (1,000,000 samples) with adjustment for the gene length of the protein-coding sequence (CDS): (1) sort all genes by their CDS length in ascending order and divide them into equal-size bins (1,000 genes each); (2) count how many genes from a GO term are in each bin; (3) from each bin, sample with replacement the same number of genes and compute the average L and S; (4) count how many times sampled L and S were lower or higher than observed values. The P values were computed as the proportion of bootstrap samples that exceeded the observed values. The fold change of average L and S was computed as a ratio of observed values to the average of sampled values. We tested all 12,563 GO terms that included more than one gene. The P-value significance threshold was thus ~2×10^{-6}. The enrichment and depletion of pLOF variants in public gene databases was tested in a similar way.

Sequencing depth at protein-coding regions

We compared sequencing depth at protein-coding regions in TOPMed WGS and ExAC WES data. The ExAC WES depth at each sequenced base pair on human genome build GRCh37 was downloaded from the ExAC browser website (http://exac.broadinstitute.org). When sequencing depth summary statistics for a base pair were missing, we assumed depth <10× for this base pair. Only protein-coding genes from consensus coding sequence were analysed and the protein-coding regions (CDS) were extracted from GENCODE v.29. When analysing ExAC sequencing depth, we used GENCODE v.29 lifted to human genome build GRCh37. When comparing sequencing depth for each gene individually in TOPMed and ExAC, we used only genes present in both GRCh38 and GRCh37 versions of GENCODE v.29.

Novel genetic variants in unmapped reads

Analysis of unmapped reads was performed using 53,831 samples from TOPMed data freeze 5. From each sample, we extracted and filtered all read pairs with at least one unmapped mate and used them to discover human sequences that were absent from the reference. The pipeline included four steps: (1) per-sample de novo assembly of unmapped reads; (2) contig alignment to the *Pan paniscus*, *Pan troglodytes*, *Gorilla gorilla* and *Pongo abelii* genome references and subsequent hominid-reference-based merging and scaffolding of sequences pooled together from all samples; (3) reference placement and breakpoint calling; and (4) variant genotyping. The detailed description of each step is provided in Supplementary Information 1.7.

Identification of *CYP2D6* alleles using Stargazer's genotyping pipeline

Details of the Stargazer genotyping pipeline have been described previously[43]. In brief, SNVs and indels in *CYP2D6* were assessed from a VCF file generated using GATK-HaplotypeCaller[89]. The VCF file was phased using the program Beagle[90] and the 1000 Genomes Project haplotype reference panel. Phased SNVs and indels were then matched to star alleles. In parallel, read depth was calculated from BAM files using GATK-DepthOfCoverage[89]. Read depth was converted to copy number by performing intra-sample normalization[43]. After normalization, structural variants were assessed by testing all possible pairwise combinations of pre-defined copy number profiles against the observed copy number profile of the sample. For new SVs, breakpoints were statistically inferred using changepoint[91]. Information regarding new SVs was stored and used to identify subsequent SVs in copy number profiles. Output data included individual diplotypes, copy number plots and a VCF of SNVs and indels that were not used to define star alleles.

Genome-wide distribution of genetic variation

Contiguous segment analysis. We excluded indels and multi-allelic variants, and categorized the remaining variants as common (allele frequency ≥ 0.005) or rare (allele frequency < 0.005), and as coding or noncoding based on protein-coding exons from Ensembl 94[92]. Variant counts were analysed across 2,739 non-empty (that is, with at least one variant) contiguous 1-Mb chromosomal segments, and counts in segments at the end of chromosomes with length $L < 10^6$ bp were scaled up proportionally by the factor $10^6 \times L^{-1}$. For each segment, the coding proportion, C, was calculated as the proportion of bases overlapping protein-coding exons. The distribution of C is fairly narrow, with 80% of segments having $C \leq 0.0195$, 99% of segments have $C \leq 0.067$ and only 3 segments having $C \geq 0.10$. Owing to the significant negative correlation between C and the number of variants in a segment, and potential mapping effects, we use linear regression to adjust the variant counts per segment according to the model count = $\beta \times C + A$ + count_adj, where A is the proportion of segment bases overlapping the accessibility mask (Supplementary Information 1.5). Unless otherwise noted, we present analyses and results that use these adjusted count values.

Concatenated segment analysis. Distinct base classifications were defined by both coding and noncoding annotations (based on Ensembl 94[92]) and CADD in silico prediction scores[21] (downloaded from the CADD server for all possible SNVs). For each base, we used the maximum possible CADD score (when using the minimum CADD score, results were qualitatively the same). Bases beyond the final base with a CADD score per chromosome were excluded. This resulted in six distinct types of concatenated segments: high (CADD ≥ 20), medium (10 ≤ CADD < 20) and low (CADD < 10) CADD scores for coding and similarly for noncoding variants. Common (allele frequency ≥ 0.005) and rare (allele frequency < 0.005) variant counts were then calculated across these concatenated segments. Multi-allelic variants and those in regions masked due to accessibility were excluded. Counts in segments at the end of chromosomes were scaled up as in the contiguous analysis.

Singleton clustering analysis

Data. From the TOPMed freeze 5 dataset, we selected a subset of 1,000 unrelated individuals of African ancestry, 1,000 unrelated individuals of East Asian ancestry and 1,000 unrelated individuals of European

ancestry, with the ancestry of each individual inferred across 7 global reference populations using RFMix[93]. In each of these subsamples, we recalculated the allele counts of each SNV and extracted SNVs that were singletons within that sample, then calculated the distance to the nearest singleton (either upstream or downstream from the focal singleton) occurring within the same individual. Note that a singleton defined here is not necessarily a singleton in the entire TOPMed freeze 5 dataset. We chose to limit the size of each population subsample to $n = 1{,}000$ for three reasons: first, to ensure the different population subsamples carried roughly a similar number of singletons; second, to ensure homogeneous ancestry within each subsample so that our analysis of singleton clustering patterns was not an artefact of admixed haplotypes; third, to limit the incidence of recurrent mutations at hypermutable sites, which can alter the underlying mutational spectrum of singleton SNVs in large samples[94]. Although the TOPMed Consortium sequenced individuals from several other diverse population groups (for example, Samoan, Hispanic/Latino individuals), the majority of these individuals were of admixed ancestry and the singletons from these smaller samples reflected mutations that have accumulated over a longer period of time, so the mutation spectra and genome-wide distributions of these samples would be more susceptible to distortion by other evolutionary processes such as selection and biased gene conversion[31].

Simulations. To quantify the effects of external branch length heterogeneity on singleton clustering patterns, we used the stdpopsim library[95] to simulate variants across chromosome 1 for 2,000 European and 2,000 African haploid samples, using a previously reported demographic model[10]. Simulations were performed using a per-site, per-generation mutation rate[96] of 1.29×10^{-8}, and using recombination rates derived from the HapMap genetic map[97]. Because our aim was to compare these simulated singletons to unphased singletons observed in the TOPMed data, we randomly assigned each of the 2,000 haploid samples from each population into one of 1,000 diploid pairs, and calculated the inter-singleton distances per diploid sample, ignoring the haplotype on which each simulated singleton originated.

Mixture model parameter estimation. The distribution of singletons suggest an underlying nonhomogeneous Poisson process, where the rate of incidence varies across the genome. In other areas of research, it has been shown that the waiting times between events arising from other nonhomogeneous Poisson processes, such as volcano eruptions or extreme weather events, can be accurately modelled as a mixture of exponential distributions[98,99]. Taking a similar approach, we model the distribution of inter-singleton distances across all S_i singletons in individual i as a mixture of K exponential component distributions ($f_k(d_i;\theta_{i,k})$), given by:

$$f(d_i;\lambda_i,\ \theta_i) = \sum_{k=1}^{K} \lambda_{i,k} f_k(d_i;\theta_{i,k})$$

where $\theta_{i,1} < \theta_{i,2} < \ldots < \theta_{i,K}$ and $\lambda_{i,k} = S_{i,k}/S_i$ is the proportion of singletons arising from component k, such that $\sum_{k=1}^{K} \lambda_{i,k} = 1$.

We estimate the parameters of this mixture ($\lambda_{i,1}, \ldots, \lambda_{i,K}, \theta_{i,1}, \ldots, \theta_{i,K}$) using the expectation–maximization algorithm as implemented in the mixtools R package[100]. Code for this analysis is available for download from the GitHub repository[101]. To identify an optimal number of mixture components, we iteratively fit mixture models for increasing values of K and calculated the log-likelihood of the observed data D given the parameter estimates ($\hat{\lambda}_{i,1}, \ldots, \hat{\lambda}_{i,K}, \hat{\theta}_{i,1}, \ldots, \hat{\theta}_{i,K}$), stopping at K components if the P value of the likelihood ratio test between $K-1$ and K components was >0.01 (χ^2 test with two degrees of freedom). The goodness-of-fit plateaued at four components for the majority of individuals, so we used the four-component parameter estimates from each individual in all subsequent analyses.

Now let $k_{i,j}$ indicate which of the four processes generated singleton j in individual i. We calculated the probability of being generated by process k as:

$$p(k_{i,j} = k | d_{i,j};\ k \in \{1, \ldots, 4\}) = \frac{p(d_i, k)}{p(d_i)} = \frac{\lambda_{i,k} f_k(d_i;\theta_{i,k})}{\sum_{k=1}^{4} \lambda_{i,k} f_k(d_i;\theta_{i,k})}.$$

We then classified the process-of-origin for each singleton according to the following optimal decision rule:

$$\hat{k}_{i,j} = \arg\max_{k \in \{1,\ldots,4\}} p(k|d_{i,j}).$$

Identification of mixture component hotspots. After assigning singletons to the most likely mixture component, we pooled singletons across individuals of a given ancestry group and counted the number of occurrences in each component in non-overlapping 1-Mb windows throughout the genome. We defined hotspots as the top 5% of 1-Mb bins containing the most singletons in a component in each ancestry group.

Modelling the relationship between clustering patterns and genomic features. In each 1-Mb window, we calculated the average signal for 12 genomic features (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9ac, H3K9me3, exon density, DNase hypersensitivity, CpG island density, lamin-associated domain density and recombination rate), using the previously described source datasets[31]. For each mixture component, we then applied the following negative binomial regression model to estimate the effects of each feature on the density of that component in 1-Mb windows:

$$\log(Y_{a,k,w}) = \beta_0 + \beta_1 X_{1,w} + \ldots + \beta_{12} X_{12,w}$$

Where $Y_{a,k,w}$ is the number of singletons in ancestry subsample a of mixture component k in window w and $X_{1,w}, \ldots, X_{12,w}$ are the signals of each of the 12 genomic features in corresponding window w.

Evolutionary genetics of individuals with diverse ancestry

Rare variant sharing. In these analyses, we used 39,722 unrelated individuals that had provided consent for population genetics research. Each individual was grouped into their TOPMed study, except for individuals from the AFGen project, which were treated as one study (Extended Data Tables 1, 2). Individuals from the FHS and ARIC projects individuals, which overlapped with the AFGen project, remained in their respective studies and were not grouped into the AFGen project. Individuals for whom the population group was either missing or 'other' were removed from the analysis. We then removed all indels, multi-allelic variants and singletons from the remaining 39,168 individuals. Each study was then split by population group. We excluded studies that had fewer than 19 samples from the analysis; however all 39,168 samples were used to define singleton filtering. We used the Jaccard index[102], J, to determine the intersection of rare variants ($2 \le$ sample count ≤ 100) between two individuals divided by the union of the rare variants of that pair, where the sample count indicates the number of individuals with either a heterozygote or homozygote variant. We then determined the average J value between and within each study.

To confirm that J is not biased by sample size, we randomly sampled 500 individuals from each of two studies with European (AFGen and FHS) and African (COPDGene and JHS) population groups in TOPMed freeze 3, without replacement. We then recalculated J between and within these randomly sampled studies, considering alternative allele counts between 2 and 100 within these 2,000 individuals.

Haplotype sharing. We used the RefinedIBD program[103] to call segments of identical-by-descent (IBD) sharing of length ≥ 2 cM on the autosomes using passing SNVs with MAF > 5%. All 53,831 samples were included in this analysis, and we used genotype data phased with

Eagle2[81]. As IBD logarithm of odds (LOD) scores are often deflated in populations with strong founding bottlenecks, such as the Amish, we used a LOD score threshold of 1.0 instead of the default 3.0. To account for possible phasing and genotyping errors, we filled gaps between IBD segments for the same pair of individuals if the gap had a length of at most 0.5 cM and at most one discordant genotype. As a result of the lower LOD threshold, regions with a low variant density can have an excess of apparent IBD segments. We therefore identified regions with highly elevated levels of detected IBD using a previously described procedure[104] and removed any IBD segments that fell wholly within these regions.

We divided the data by study and by population group within each study. In the analyses of IBD sharing levels and recent effective size, we did not include studies without appropriate consent or population groups with fewer than 80 individuals within a study. We calculated the total length of IBD segments for each pair of individuals, and we averaged these totals within each population group in a study and between each pair of population-by-study groups. We also estimated recent effective population sizes for each group using IBDNe[104].

Demographic estimation under selection at linked sites. We selected 2,416 samples from the TOPMed data freeze 3 that (1) had a high percentage of European ancestry; (2) were unrelated; and (3) gave consent for population genetics research. More detailed information about ancestry estimation and filters is provided in Supplementary Information 1.10.

We performed several steps to filter the genome for high-quality neutral sites, which were based on a previously described ascertainment scheme[30] (Supplementary Information 1.10). After filtering, positions in the genome were annotated for how strongly affected they were by selection at linked sites using the background selection coefficient, McVicker's B statistic[60]. We used all sites annotated with a B value for performing general analyses. However, when performing demographic inferences, we limited our analyses to regions of the genome within the top 1% of the genome-wide distribution of B ($B \geq 0.994$). These sites correspond to regions of the genome inferred to be under the weakest amount of background selection (that is, under the weakest effects of selection at linked sites). Sites in the genome were also polarized to ancestral and derived states using ancestral annotations called with high-confidence from the GRCh37 e71 ancestral sequence. After keeping only polymorphic bi-allelic sites, we had 20,324,704 sites, of which 191,631 had $B \geq 0.994$. We also identified 91,177 fourfold degenerate synonymous sites (irrespective of B) that were polymorphic (bi-allelic) and had high-confidence ancestral and derived states.

We performed demographic inference with the moments[105] program by fitting a model of exponential growth with three parameters (N_{Eur0}, N_{Eur}, T_{Eur}) to the site-frequency spectrum. This included two free parameters: the starting time of exponential growth (T_{Eur}) and the ending population size after growth (N_{Eur}). The ancestral size parameter (that is, the population size when growth begins), N_{Eur0}, was kept constant in our model such that the relative starting size of the population was always 1. We applied the inference procedure to either fourfold degenerate sites or sites with $B \geq 0.994$. The site frequency spectrum used for inference was unfolded and based on the polarization step described above. The inference procedure was fit using sample sizes ($2N$) of 1,000, 2,000, 3,000, 4,000 and 4,832 chromosomes. To convert the scaled genetic parameters output by the inference procedure to physical units, we used the resulting theta (also inferred by moments) and a mutation rate[106] of 1.66×10^{-8} to generate corresponding effective population sizes (N_e). To convert generations to years, we assumed a generation time of 25 years. The 95% confidence intervals were generated by resampling the site frequency spectrum 1,000 times and using the Godambe information matrix to generate parameter uncertainties[107]. A more detailed description is available in Supplementary Information 1.10.

Selection. We started with 39,649 unrelated individuals selected from the TOPMed data freeze 5 for which we had consent for population genetic analyses (Extended Data Table 3). As the singleton density score (SDS) requires thousands of samples and a baseline demographic history, we subset our data by population group and limited our population analysis to those population groups for which we had well-studied demographic histories: broadly European, broadly African and broadly East Asian. To avoid potential problems introduced by admixture, we required that our samples had more than 90% inferred European, African or East Asian ancestry as inferred by a seven-way ancestry inference pipeline (Supplementary Information 1.11). This left $n = 21,196$ European samples, $n = 2,117$ African samples and $n = 1,355$ East Asian samples. We specifically excluded Amish samples from the European group as they are a unique founder population. We analysed each population separately. Only bi-allelic sites with an unambiguous ancestral state, inferred using the WGSA pipeline[108], were used. Sites near chromosome boundaries, near centromeres and in regions with poor accessibility were excluded. We used the previously published R scripts[61] to perform all demographic history simulations and SDS computations in each population. We then normalized raw SDS scores within 1% frequency bins and treated the normalized scores as Z-scores to convert them to P values as described previously[61]. Raw and normalized SDS scores are included in Supplementary Data 2.

TOPMed imputation panel

Construction. We divided each autosomal chromosome and the X chromosome into overlapping chunks (with chunk size of 1 Mb each and with 0.1 Mb overlap between consecutive chunks), and then phased each of the chunks using Eagle v.2.4[81]. We removed all singleton sites and compressed the haplotype chunks into m3vcf format[109]. Afterwards, we ligated the compressed haplotype chunks for each chromosome to generate the final reference panel.

Evaluation of imputation accuracy. For all TOPMed individuals, genetic ancestries were estimated using the top four principal components projected onto the principal component space of 938 Human Genome Diversity Project (HGDP) individuals using verifyBamID2[110]. For each TOPMed individual, we identified the 10 closest individuals from 2,504 individuals from the 1000 Genomes Project phase 3 based on Euclidean distances in the principal component space estimated by verifyBamID2. If all of the 10 closest individuals from the 1000 Genomes Project phase 3 belonged to the same super-population—among African, admixed American, East Asian, European and South Asian populations—we estimated that the TOPMed individual also belonged to that super-population. Among the 97,256 reference panel individuals, 90,339 (93%) were assigned to a super-population, with the following breakdown: African, 24,267 individuals; admixed American, 17,085 individuals; European, 47,159 individuals; East Asian, 1,184 individuals; South Asian, 644 individuals. We randomly selected 100 individuals from each super-population in the BioMe TOPMed study, and selected markers on chromosome 20 present on the Illumina HumanOmniExpress (8v1-2_A) array. The selected genotypes were phased with Eagle 2.4.1[81], using the 1000 Genomes Project phase 3 ($n = 2,504$), Haplotype Reference Consortium (HRC, $n = 32,470$) and TOPMed ($n = 96,756$) reference panels, excluding the 500 individuals from the TOPMed reference panel. The phased genotypes were imputed using Minimac4[111] from each reference panel, and the imputation accuracy was estimated as the squared correlation coefficient (r^2) between the imputed dosages and the genotypes calls from the sequence data. The allele frequencies were estimated among all TOPMed individuals estimated to belong to the same super-population, and the r^2 values were averaged across variants in each MAF category. Variants present in 100 sequenced individuals but absent from the reference panels were assumed to have $r^2 = 0$ for the purposes of computing the average r^2. The minimum MAF

to achieve $r^2 > 0.3$ was calculated from the average r^2 in each MAF category by finding the MAF that crosses $r^2 = 0.3$ using linear interpolation. The average number of rare variants (MAF < 0.5%) and the fraction of imputable rare variants ($r^2 > 0.3$) were calculated based on the number of non-reference alleles in imputed samples above and below the minimum MAF, assuming Hardy–-Weinberg equilibrium.

Imputation of the UK Biobank to the TOPMed panel and association analyses. After phasing the UK Biobank genetic data (carried out on 81 chromosomal chunks using Eagle v.2.4), the phased data were converted from GRCh37 to GRCh38 using LiftOver[112]. Imputation was performed using Minimac4[111].

We compared the correlation of genotypes between the exome-sequencing data released by the UK Biobank (following their SPB pipeline[113]) and the TOPMed-imputed genotypes. The comparison assessed 49,819 individuals and 3,052,260 autosomal variants that were found in both the exome-sequencing and TOPMed-imputed datasets (matched by chromosome, position and alleles, and with an imputation quality of at least 0.3 in the TOPMed-imputed data). We split the variants into MAF bins for which the MAF from the exome data was used to define the bins, and computed Pearson correlations averaged within each bin.

We tested single pLOF, nonsense, frameshift and essential splice-site variants[85,86] for association with 1,419 PheCodes constructed from composites of ICD-10 (International Classification of Diseases 10th revision) codes to define cases and controls. Construction of the PheCodes has been previously described[114]. We performed the association analysis in the 'white British' individuals, which resulted in 408,008 individuals after the following quality control metrics were applied: (1) samples did not withdraw consent from the UK Biobank study as of the end of 2019; (2) 'submitted gender' matches 'inferred sex'; (3) phased autosomal data available; (4) outliers for the number of missing genotypes or heterozygosity removed; (5) no putative sex chromosome aneuploidy; (6) no excess of relatives; (7) not excluded from kinship inference; and (8) in the UK Biobank defined the 'white British' ancestry subset. To perform the association analyses, we used a logistic mixed model test implemented in SAIGE[114] with birth year and the top four principal components (computed from the white British subset) as covariates. For the pLOF burden tests, for each autosomal gene with at least two rare pLOF variants ($n = 12{,}052$ genes), a burden variable was created in which dosages of rare pLOF variants were summed for each individual. This sum of dosages was tested for association with the 1,419 traits using SAIGE. The same covariates used in the single-variant tests were included. For both the single-variant and the burden tests, we used 5×10^{-8} as the genome-wide significance threshold.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

A detailed description of the TOPMed participant consents and data access is provided in Box 1. TOPMed data used in this manuscript are available through dbGaP. The dbGaP accession numbers for all TOPMed studies referenced in this paper are listed in Extended Data Tables 2, 3. A complete list of TOPMed genetic variants with summary level information used in this manuscript is available through the BRAVO variant browser (bravo.sph.umich.edu). The TOPMed imputation reference panel described in this manuscript can be used freely for imputation through the NHLBI BioData Catalyst at the TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/). DNA sequence and reference placement of assembled insertions are available in VCF format (without individual genotypes) on dbGaP under the TOPMed GSR accession phs001974.

Code availability

All code for TOPMed data quality checks and variant calling is available at https://github.com/statgen/topmed_variant_calling. Code for the WGS and WES data comparisons is available at https://github.com/statgen/sequencing_comparison. Code for modelling the singleton distance distribution is available at https://github.com/carjed/topmed_singleton_clusters. Code for identifying novel genetic variants in unmapped reads is available at https://github.com/nygenome/topmed_unmapped. Code for gene-burden association tests using rare pLOF variants is available at https://github.com/sgagliano/GeneBurden. Code for the imputed and genotype UK Biobank WES data comparisons is available at https://github.com/sgagliano/UKB_WES_vs_TOPMed_IMP.

76. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
77. Regier, A. A. et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun*. **9**, 4038 (2018).
78. Jun, G. & Kang, H. M. GotCloud. https://genome.sph.umich.edu/wiki/GotCloud (accessed 2019–2020).
79. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome Res*. **25**, 918–925 (2015).
80. Center for Statistical Genetics. statgen: topmed variant calling. GitHub https://github.com/statgen/topmed_variant_calling (2020).
81. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet*. **48**, 1443–1448 (2016).
82. The MITRE Corporation. mitre: fusera. GitHub https://github.com/mitre/fusera (2019).
83. Center for Statistical Genetics. statgen: statgen-tools. Docker Hub https://hub.docker.com/r/statgen/statgen-tools.
84. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol*. **39**, 276–293 (2015).
85. Karczewski, K. J. et al. loftee. GitHub https://github.com/konradjk/loftee (2015).
86. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol*. **17**, 122 (2016).
87. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat. Genet*. **25**, 25–29 (2000).
88. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res*. **45** (D1), D331–D338 (2017).
89. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**, 1297–1303 (2010).
90. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet*. **81**, 1084–1097 (2007).
91. Killick, R. & Eckley, I. A. changepoint: an R package for changepoint analysis. *J. Stat. Softw*. **58**, 1–19 (2014).
92. Zerbino, D. R. et al. Ensembl 2018. *Nucleic Acids Res*. **46** (D1), D754–D761 (2018).
93. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet*. **93**, 278–288 (2013).
94. Harpak, A., Bhaskar, A. & Pritchard, J. K. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. *PLoS Genet*. **12**, e1006489 (2016).
95. Adrion, J. R. et al. A community-maintained standard library of population genetic models. *eLife* **9**, e54967 (2020).
96. Tian, X., Browning, B. L. & Browning, S. R. Estimating the genome-wide mutation rate with three-way identity by descent. *Am. J. Hum. Genet*. **105**, 883–893 (2019).
97. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
98. Mendoza-Rosas, A. T. & De la Cruz-Reyna, S. A mixture of exponentials distribution for a simple and precise assessment of the volcanic hazard. *Nat. Hazards Earth Syst. Sci*. **9**, 425–431 (2009).
99. Rossi, F., Fiorentino, M. & Versace, P. Two-component extreme value distribution for flood frequency analysis. *Wat. Resour. Res*. **20**, 847–856 (1984).
100. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. S. mixtools: an R package for analyzing mixture models. *J. Stat. Softw*. **32**, 1–29 (2009).
101. Carlson, J. carjed: topmed singleton clusters. GitHub https://github.com/carjed/topmed_singleton_clusters (2020).
102. Prokopenko, D. et al. Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project. *Bioinformatics* **32**, 1366–1372 (2016).
103. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
104. Browning, S. R. & Browning, B. L. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet*. **97**, 404–418 (2015).
105. Jouganous, J., Long, W., Ragsdale, A. P. & Gravel, S. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics* **206**, 1549–1567 (2017).
106. Palamara, P. F. et al. Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *Am. J. Hum. Genet*. **97**, 775–789 (2015).

107. Coffman, A. J., Hsieh, P. H., Gravel, S. & Gutenkunst, R. N. Computationally efficient composite likelihood statistics for demographic inference. *Mol. Biol. Evol.* **33**, 591–593 (2016).
108. Liu, X. et al. WGSA: an annotation pipeline for human genome sequencing studies. *J. Med. Genet.* **53**, 111–112 (2016).
109. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
110. Zhang, F. et al. Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Res.* **30**, 185–194 (2020).
111. Center for Statistical Genetics. Minimac4. https://genome.sph.umich.edu/wiki/Minimac4 (2018).
112. Casper, J. et al. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* **46** (D1), D762–D769 (2018).
113. Van Hout, C. V. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
114. Zhou, W. et al. Efficiently controlling for case–control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

Acknowledgements WGS for the TOPMed programme was supported by the National Heart, Lung and Blood Institute (NHLBI). Specific funding sources for each study and genomic centre are provided in Supplementary Table 20. Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity quality control and general study coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN2682018000001I). We thank the studies and participants who provided biological samples and data for TOPMed. The full study-specific acknowledgments are included in Supplementary Information 2. The UK Biobank analyses were conducted using the UK Biobank Resource under application number 24460. Other acknowledgments are included in Supplementary Information 3. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the US Department of Health and Human Services.

Author contributions Supplementary Table 21 lists the analysts and senior scientists who contributed to particular sections of this paper. T.W.B., Q.W., F.A., K.G.A., P.L.A., R.G.B., R.L.B., J. Blangero, M.B., E.G.B., J.F.C., Y.-D.I.C., M.H.C., A. Correa, J.E.C., D.L.D., P.T.E., M.F., N.F., S.M.F., D.J.G., M.E.H., J.H., S.R.H., M.R.I., A.D.J., S.K., D.P.K., C.K., A.K., L.A.L., J.L.-S., D.L., C.L., K.L.L., A.M., A.K.M., R.A.M., S.T.M., J.B.M., J.L.M., M.A.M., B.D.M., M.E.M., C.M., A.C.M., J.M.M., P.N., K.E.N., N.P., G.M.P., W.S.P., B.M.P., D.C.R., S.R., A.P.R., J.I.R., I.R., C.S., S. Seshadri, V.A.S., W.H.S., N.L.S., N.S., K.D.T., T.A.T., R.S.V., S.V., D.E.W., B.S.W., S.T.W., C.J.W., D.K.A., A.E.A.-K., K.C.B., E.B., S. Gabriel, R. Gibbs, K.M.R., S.S.R., E.K.S., P.Q., W.G., G.J.P., D.A.N., S.Z., J.G.W., L.A.C., C.C.L., C.E.J., R.D.H., T.D.O. and G.R.A. contributed to the conception or design of the TOPMed programme and its operations. C.A., A.A., D.E.A., S.A., J. Barnard, L.B., L.C.B., E.J.B., L.F.B., J. Blangero, D.W.B., J.A.B., E.G.B., B.E.C., B. Chalazan, D.I.C., Y.-D.I.C., M.K.C., A. Correa, J.E.C., B. Custer, D.D., M.D., M.D.A., P.T.E., C.E., D.F., T.F., M.T.G., X.G., J.H., N.L.H.-C., S.R.H., J.M.J., R. Kaplan, S.L.R.K., T.K., S.K., E.E.K., D.P.K., R. Klemmer, B.A.K., C.K., L.A.L., J.L.-S., R.J.F.L., L.G., R. Gerszten, S.A.L., K.L.L., A.C.Y.M., R.A.M., D.D.M., S.T.M., D.A.M., B.D.M., S.M., C.M., A.N., K.E.N., J.R.O., N.D.P., P.A.P., W.S.P., B.M.P., D.C.R., S.R., D.R., J.I.R., D.A.S., S. Seshadri, V.A.S., W.H.S., M.B.S., N.L.S., J.A.S., W.T., K.D.T., M.T., R.P.T., D.J.V.D.B., R.S.V., D.E.W., S.T.W., Y.Z., D.K.A., A.E.A.-K., K.C.B., E.B., S.S.R., E.K.S., J.G.W., L.A.C. and R.D.H. provided phenotypic data and/or biosamples. F.A., K.G.A., L.C.B., J. Blangero, B.E.C., C.B.C., J.E.C., S.K.D., P.T.E., S. Germer, X.G., D.L., R.J.F.L., S.T.M., K.E.N., J.I.R., J.-S.S., K.D.T., D.J.V.D.B., R.S.V., K.A.V.-M., D.E.W., A.E.A.-K., K.C.B., E.B., S. Gabriel, R. Gibbs, G.J.P. and D.A.N. acquired WGS and/or other omics data. D.T., D.N.H., M.D.K., J.C., Z.A.S., R.T., S.A.G.T., A. Corvelo, S.M.G., H.M.K., A.N.P., J. LeFaive, S.-b.L., X.T., B.L.B., S.D., A.-K.E., W.E.C., D.P.L., A.C.S., T.W.B., A.V.S., Q.W., X. Liu, M.P.C., D.M.B., L.S.E., L.F., C.F., S. Germer, X. Lin, K.-H.L., S.C.N., J.P., S. Schoenherr, A.M.S., X.Z., E.B., D.A.N. and C.C.L. created software, processed and/or analysed WGS or other data for data summaries in this paper. D.T., D.N.H., M.D.K., J.C., Z.A.S., R.T., S.A.G.T., A. Corvelo, S.D., S. Germer, S.R.B., L.A.C., C.C.L., C.E.J., R.D.H., T.D.O. and G.R.A. drafted the manuscript and revised the paper according to co-author suggestions. All authors reviewed the manuscript, suggested revisions as needed and approved the final version. A full list of members and affiliations of the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium is available at https://www.nhlbiwgs.org/topmed-banner-authorship.

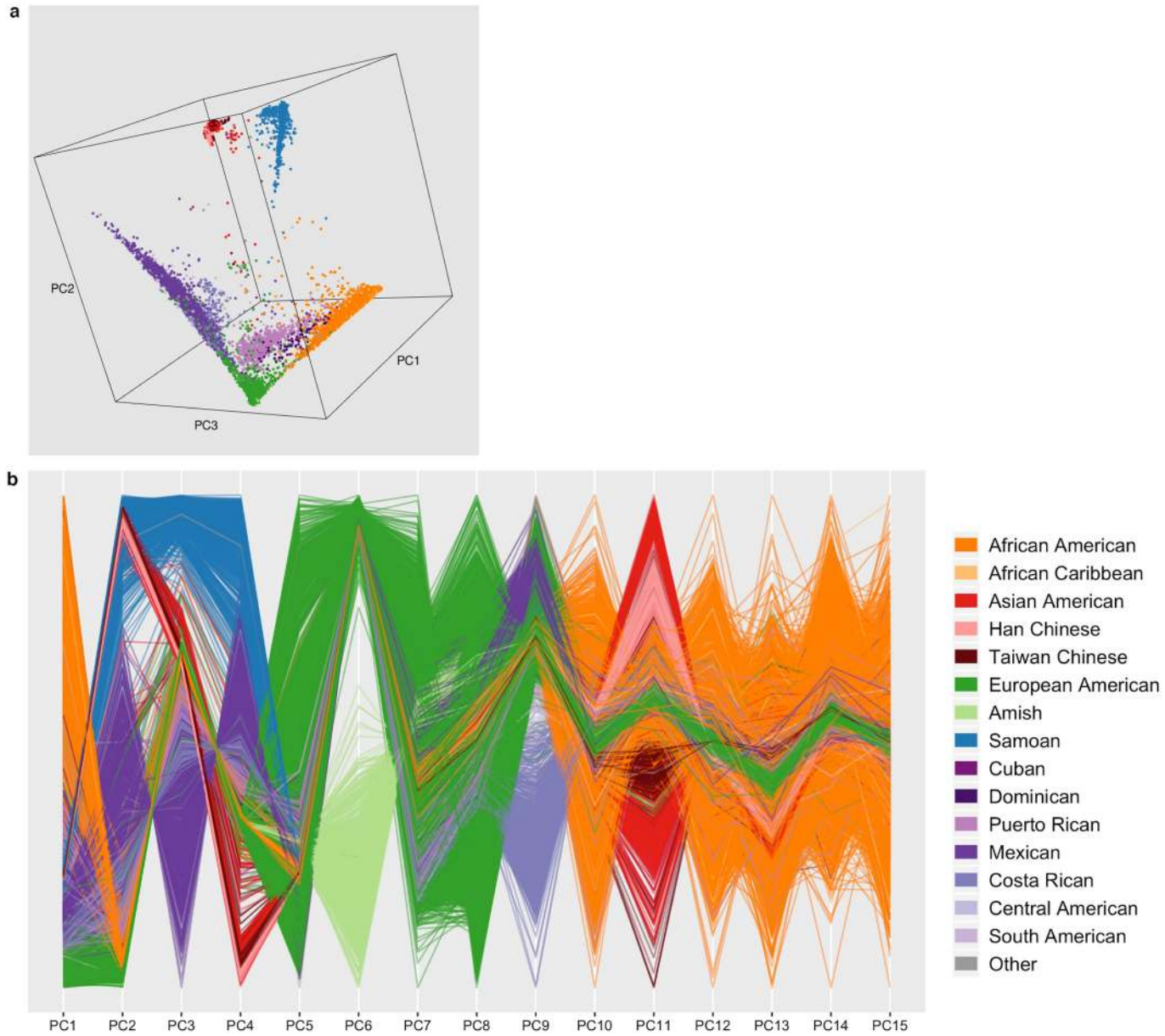
Competing interests S.D. holds equity in 23andMe. S.A. holds equity in 23andMe. R.G.B. has received funding from NIH, the COPD Foundation and Alpha1 Foundation. J.F.C. is an inventor on a patent licensed to ImmunArray. M.H.C. has received grant support from GSK. D.L.D. has received personal fees from Novartis. P.T.E. is supported by a grant from Bayer to the Broad Institute focused on the genetics and therapeutics of cardiovascular diseases. P.T.E. has also served on advisory boards or consulted for Quest Diagnostics and Novartis. M.T.G. is a co-inventor on pending patent applications and planned patents directed to the use of recombinant neuroglobin and haeme-based molecules as antidotes for CO poisoning, which have been licensed by Globin Solutions. Globin Solutions also has an option to a potential therapeutic for CO poisoning from VCU, hydroxycobalamin. M.T.G. is a shareholder, advisor and director in Globin Solutions. M.T.G. is a co-inventor on patents directed to the use of nitrite salts in cardiovascular diseases, which were previously licensed to United Therapeutics and Hope Pharmaceuticals, and are now licensed to Globin Solutions. M.T.G. is a co-investigator in a research collaboration with Bayer Pharmaceuticals to evaluate riociguate as a treatment for patients with sickle cell disease. M.T.G. has served as a consultant for Epizyme, Actelion Clinical Research, Acceleron Pharma, Catalyst Biosciences, Modus Therapeutics, Sujana Biotech and United Therapeutics Corporation. M.T.G. is on Bayer HealthCare's Heart and Vascular Disease Research Advisory Board. D.P.K. receives grants to his institution from Amgen and Radius Health, and serves on scientific advisory boards for Solarea Bio and Pfizer. K.H.L. holds equity in 23andMe. S.A.L. receives sponsored research support from Bristol Myers Squibb/Pfizer, Bayer, Boehringer Ingelheim and Fitbit, has consulted for Bristol Myers Squibb/Pfizer and Bayer, and participates in a research collaboration with IBM. D.D.M. receives research support from Bristol Myers Squibb, Care Evolution, Samsung, Apple Computer, Pfizer, Biotronik, Boehringer Ingelheim, Philips Research Institute, Flexcon, Fitbit and has consulted for Bristol Myers Squibb, Pfizer, Fitbit, Philips, Samsung Electronics, Rose Consulting, Boston Biomedical Associates and FlexCon. D.D.M. is also a member of the Operations Committee and Steering Committee for the GUARD-AF Study (NCT04126486) sponsored by Bristol Meyers Squibb and Pfizer. J.B.M. is an Academic Associate for Quest Diagnostics. For B.D.M.: the Amish Research Program receives partial support from Regeneron Pharmaceuticals. M.E.M. is an inventor on a patent that was published by the United States Patent and Trademark Office on 6 December 2018 under Publication Number US 2018-0346888, and an international patent application that was published on 13 December 2018 under Publication Number WO-2018/226560 regarding B4GALT1 Variants And Uses Thereof. P.N. reports grants from Amgen, Apple, Boston Scientific and Novartis, consulting income from Apple, Blackstone Life Sciences, Genentech and Novartis, and spousal employment at Vertex, all unrelated to the present work. B.M.P. serves on the DSMB of a clinical trial funded by the manufacturer (Zoll LifeCor) and on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. J.-S.S. serves as the chairman of Macrogen. S.T.W. is paid royalties by UpToDate. The spouse of C.J.W. works at Regeneron Pharmaceuticals. R.A.G. is an employee of Baylor College of Medicine that receives revenue from Genetic Testing. E.K.S. in the past three years received grant support from GlaxoSmithKline and Bayer. M.C.Z. owns stock in ThermoFisher and Merck. L.A.C. spends part of her time consulting for Dyslipidemia Foundation, a non-profit company, as a statistical consultant. G.R.A. is an employee of Regeneron Pharmaceuticals, he owns stock and stock options for Regeneron Pharmaceuticals.

Additional information

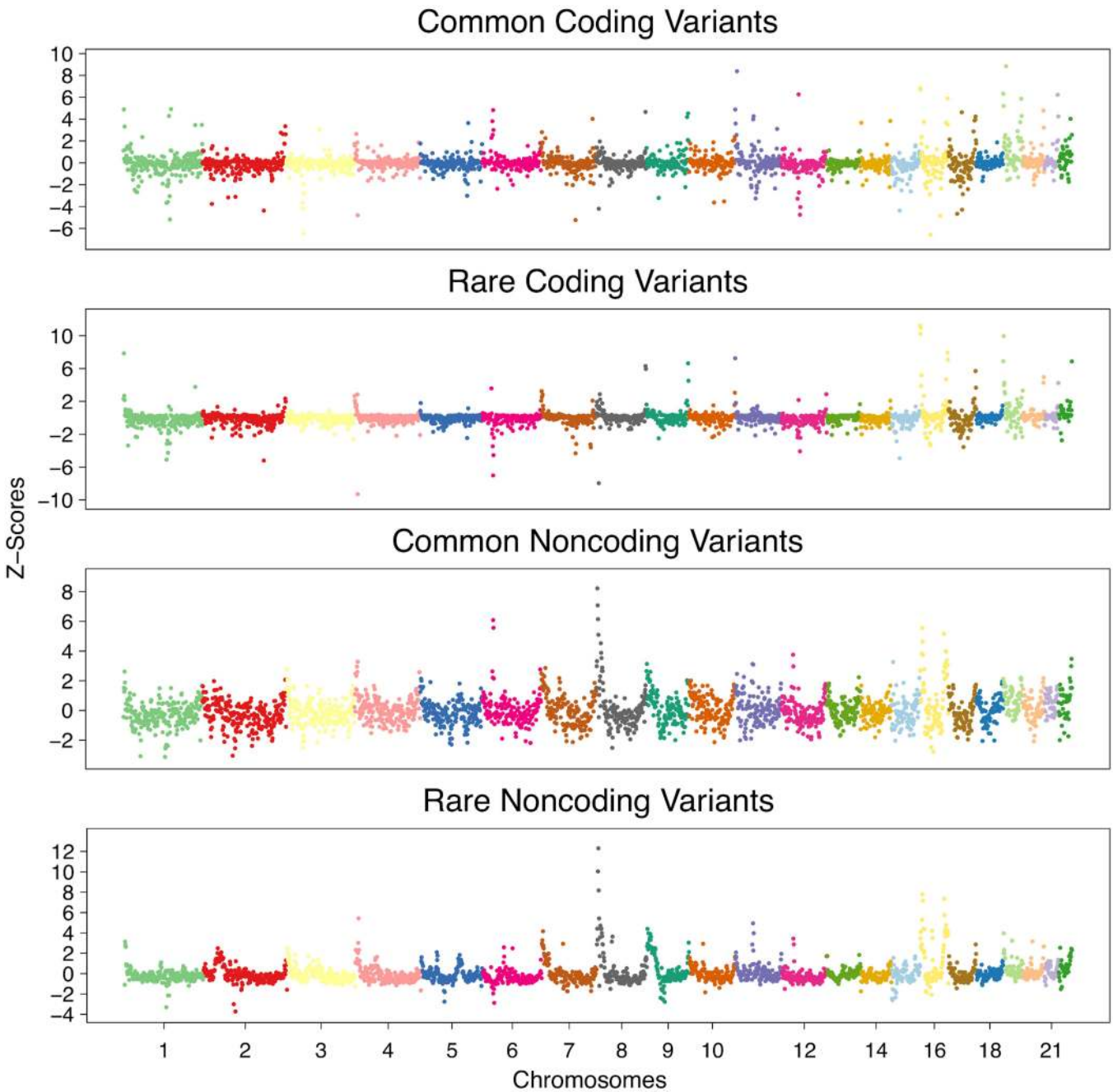
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03205-y.

Correspondence and requests for materials should be addressed to L.A.C., C.C.L., C.E.J., R.D.H., T.D.O. or G.R.A.

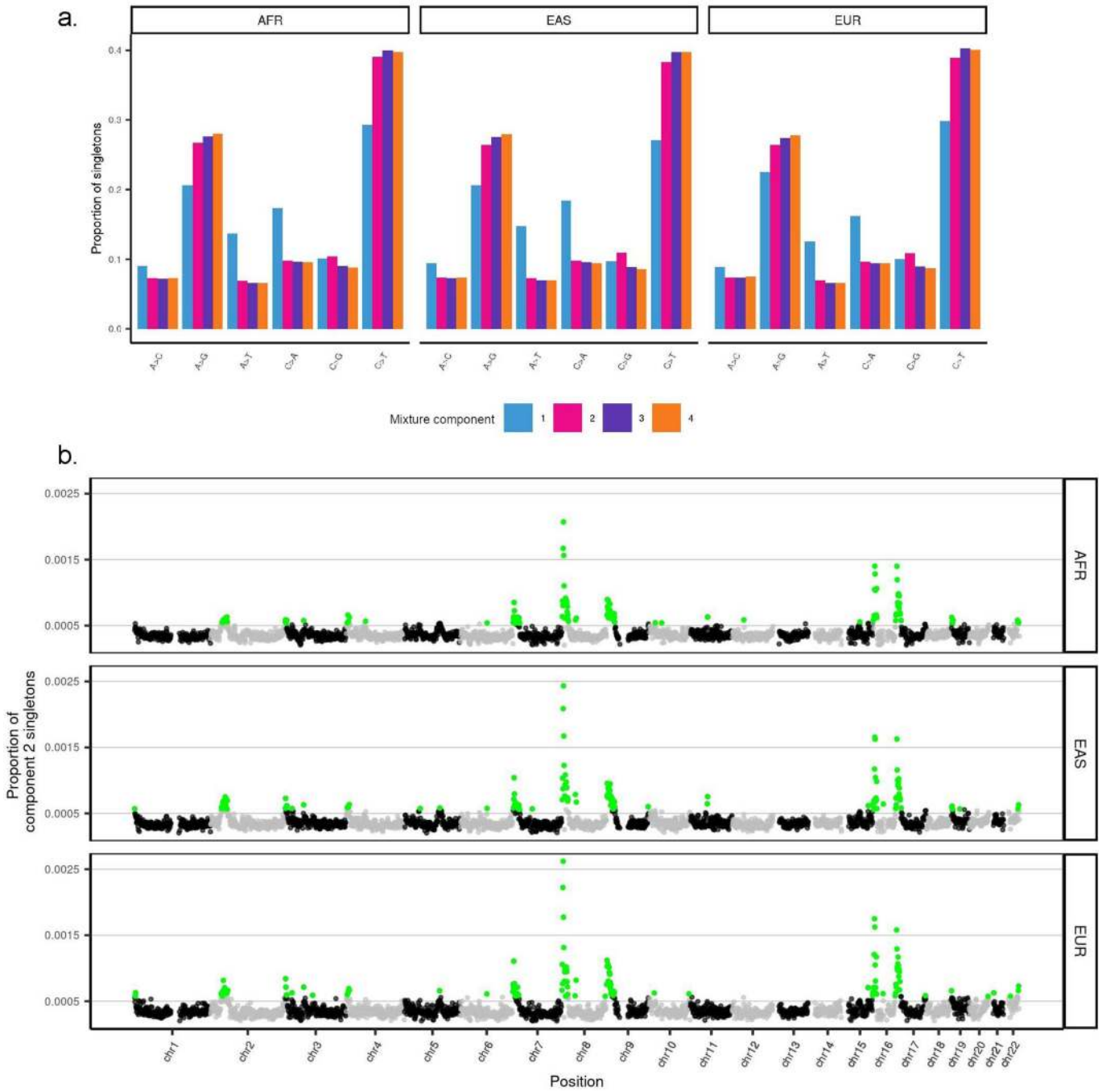
Peer review information *Nature* thanks Joshua Akey and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



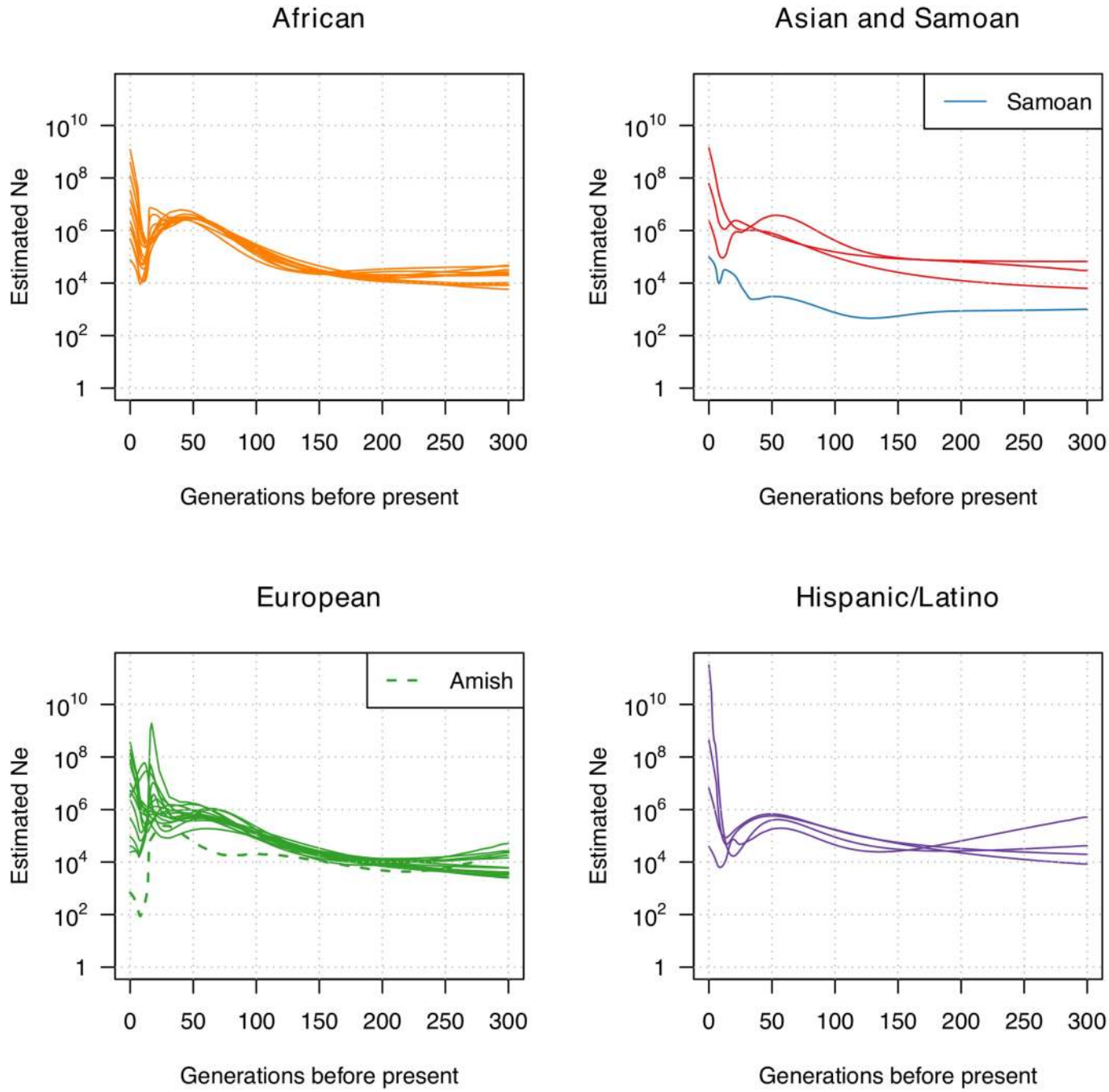
Extended Data Fig. 1 | Principal components of the genotypic data from freeze 5 pooled across studies. a, Three-dimensional plot of principal components (PC) 1, 2 and 3. **b**, Parallel coordinate plot colour-coded by categories defined according to race, ancestry and/or ethnic information provided by the study participants and/or by study investigators according to study inclusion criteria. Individuals with missing values for ancestry or ethnicity are excluded.



Extended Data Fig. 2 | Distribution of genetic variants across the genome. After filtering to focus on regions of the genome that are accessible through short-read sequencing, most contiguous 1-Mb segments show similar levels of common (5,141 ± 1,298 variants with MAF ≥ 0.5%) and rare variation (120,414 ± 19,862 variants with MAF < 0.5%). From top to bottom, panel 1 shows the levels of variation across the genome for common coding variants, panel 2 for rare coding variants, panel 3 for common noncoding variants and panel 4 for rare noncoding variants. Variation levels are represented by the *Z*-score (*X*-mean/s.d.) of the adjusted variant counts per 1-Mb contiguous segment for each variant category.



Extended Data Fig. 3 | Characteristics of singleton clustering patterns. **a**, Mutational spectra of singletons assigned to each of the four mixture components, separated by population. **b**, Density of mixture component 2 singletons in 1-Mb windows across the genome. Windows with mixture component 2 singleton counts above the 95th percentile (calculated genome-wide per population subsample) are classified as hotspots and are highlighted in green.



Extended Data Fig. 4 | Estimates of recent effective population size by population group. Each line represents the estimate from a single study, considering only individuals with an annotated population group. The included studies are the same as those in Supplementary Fig. 31. The Amish and Samoan results are individually identified due to their distinct recent population size trajectories. N_e, effective population size. The overlay view is shown in Supplementary Fig. 33.

Extended Data Table 1 | TOPMed projects and participating parent studies included in genotype data freeze 5

Project Abbreviation	Project Name	Phenotype Focus*	Participating TOPMed Parent Studies†
AA_CAC	African American Coronary Artery Calcification	CAC	DHS, GeneSTAR, GENOA, MESA
AFGen	Atrial Fibrillation Genetics Consortium	AF	ARIC, CCAF, HVH, FHS, MGH_AF, Partners, VAFAR, VU_AF, WGHS
Amish	Genetics of Cardiometabolic Health in the Amish	HLB	Amish
BAGS	Barbados Asthma Genetics Study	Asthma	BAGS
CFS	Cleveland Family Study	HLB, Sleep	CFS
COPD	Genetic Epidemiology of COPD	COPD	COPDGene, EOCOPD
CRA_CAMP	The Genetic Epidemiology of Asthma in Costa Rica and the Childhood Asthma Management Program	Asthma	CRA
FHS	Framingham Heart Study	HLB	FHS
GeneSTAR	Genetic Studies of Atherosclerosis Risk	Platelet Aggregation	GeneSTAR
GenSalt	Genetic Epidemiology Network of Salt Sensitivity	Hypertension	GenSalt
GOLDN	Genetics of Lipid Lowering Drugs and Diet Network	Lipids	GOLDN
HyperGEN_GENOA	Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy	Hypertension	GENOA, HyperGEN
JHS	Jackson Heart Study	HLB	JHS
MESA	Multi-Ethnic Study of Atherosclerosis	HLB	MESA
PGX_Asthma	Pharmacogenomics of Bronchodilator Response in Minority Children with Asthma	Asthma	GALAII, SAGE
SAFS	San Antonio Family Studies	HLB	SAFS
Sarcoidosis	Genetics of Sarcoidosis in African Americans	Sarcoidosis	Sarcoidosis
Samoan	Samoan Adiposity Study	Adiposity	Samoan
THRV	Taiwan Study of Hypertension using Rare Variants	Hypertension	THRV
VTE	Venous Thromboembolism	VTE, HLB	ARIC, CHS, HVH, Mayo_VTE, WHI
WHI	Women's Health Initiative	HLB, Stroke, VTE	WHI

See Supplementary Information 1.1.2 for definitions of TOPMed projects and parent studies. AF, atrial fibrillation; CAC, coronary artery calcification; HLB, general heart, lung and blood; VTE, venous thromboembolism. Note, some case-only collections are included. See Extended Data Table 2 for study abbreviations and additional study information.

*Primary phenotype focus for TOPMed samples.

†Some TOPMed studies participate in more than one project.

Extended Data Table 2 | Studies that contributed to the freeze-5 genotype call set

Study/Cohort Abbreviation	TOPMed Accession	TOPMed Study Name* ("NHLBI TOPMed:")	Sample Size†	Parent Study Accession	Parent Study Design
Amish	phs000956	Genetics of Cardiometabolic Health in the Amish	1,111		family/population sample
ARIC	phs001211	Trans-Omics for Precision Medicine Whole Genome Sequencing Project: ARIC	3,619	phs000280	prospective cohort
BAGS	phs001143	The Genetics and Epidemiology of Asthma in Barbados	1,022		family
CCAF	phs001189	Cleveland Clinic Atrial Fibrillation Study	360		cross-sectional case-control
CFS	phs000954	The Cleveland Family Study (WGS)	994	phs000284	family
CHS	phs001368	Cardiovascular Health Study	69	phs000287	prospective cohort
COPDGene	phs000951	Genetic Epidemiology of COPD (COPDGene) in the TOPMed Program	8,909	phs000179	case-control, longitudinal follow-up
CRA	phs000988	The Genetic Epidemiology of Asthma in Costa Rica	1,142		family
DHS	phs001412	Diabetes Heart Study African American Coronary Artery Calcification (AA CAC)	337		family/population sample
EOCOPD	phs000946	Boston Early-Onset COPD Study in the TOPMed Program	74	phs001161	family
FHS	phs000974	Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study	4,166	phs000007	prospective cohort
GALAII	phs000920	Genes-environments and Admixture in Latino Asthmatics (GALA II) Study	999	phs001180	pharmacogenomic
GeneSTAR	phs001218	GeneSTAR (Genetic Study of Atherosclerosis Risk)	1,637		family
GENOA	phs001345	Genetic Epidemiology Network of Arteriopathy (GENOA)	1,143	phs001238	family
GenSalt	phs001217	Genetic Epidemiology Network of Salt Sensitivity (GenSalt)	1,689	phs000784	family
GOLDN	phs001359	Genetics of Lipid Lowering Drugs and Diet Network (GOLDN)	899	phs000741	family
HVH	phs000993	Heart and Vascular Health Study (HVH)	625	phs001013	cross-sectional case-control
HyperGEN	phs001293	HyperGEN - Genetics of Left Ventricular (LV) Hypertrophy	1,776		cross-sectional case-control
JHS	phs000964	Jackson Heart Study	3,406	phs000286	prospective cohort
Mayo_VTE	phs001402	Whole Genome Sequencing of Venous Thromboembolism (WGS of VTE)	1,251	phs000289	cross-sectional case-control
MESA	phs001416	MESA and MESA Family AA-CAC	4,875	phs000209	prospective cohort
MGH_AF	phs001062	MGH Atrial Fibrillation Study	984	phs001001	family
Partners	phs001024	Partners HealthCare Biobank	128		cross-sectional case-control
SAFS	phs001215	San Antonio Family Heart Study (WGS)	1,508		family
SAGE	phs000921	Study of African Americans, Asthma, Genes and Environment (SAGE) Study	499		pharmacogenomic
Sarcoidosis	phs001207	African American Sarcoidosis Genetics Resource	606		family and cross-sectional
Samoan	phs000972	Genome-wide Association Study of Adiposity in Samoans	1,232	phs000914	population sample
THRV	phs001387	Rare Variants for Hypertension in Taiwan Chinese (THRV)	1,525		case families and controls
VAFAR	phs000997	The Vanderbilt AF Ablation Registry	163		cases with longitudinal follow-up
VU_AF	phs001032	The Vanderbilt Atrial Fibrillation Registry	1,110		families with longitudinal follow-up
WGHS	phs001040	Novel Risk Factors for the Development of Atrial Fibrillation in Women	115		prospective cohort
WHI	phs001237	Women's Health Initiative (WHI)	10,047	phs000200	prospective cohort

Each study has a dbGaP accession for the TOPMed sequence data and genotypes, although some also have pre-existing parent study accessions. Phenotypic data are mainly in the parent accessions, although some are in the TOPMed accessions. See also Supplementary Figs. 39 and 40 for information about the ancestral and/or ethnic and sex composition of each study. The relationships between these studies and their TOPMed project(s) are summarized in Extended Data Table 1. All of the TOPMed and parent study accessions in this table have been released on dbGaP (see https://www.ncbi.nlm.nih.gov/gap/?term=TOPMed and https://www.nhlbiwgs.org/group/project-studies?field_is_this_a_value=sub).

*Study name as it appears in dbGaP, with 'NHLBI TOPMed:' prepended.

†Approximate sample size for freeze-4 and freeze-5 releases combined.

Extended Data Table 3 | TOPMed study-consent groups used in analyses and tools

Study/Cohort Abbreviation	TOPMed Accession	Consent Group	Freeze 5 VCF						Freeze 8 VCF			Freeze 3 VCF	Freeze 5 BAM
			PCA, Kinship	General analyses	Population genetics	Selection & adaptation	BRAVO variant server	Imputation reference panel	Imputation accuracy	WGS & WES comparison		Demography	De novo assembly
Amish	phs000956	HMB-IRB-MDS	X	X	X		X	X					X
ARIC	phs001211	DS-CVD-IRB	X	X	X	X	X	X				X	X
		HMB-IRB	X	X	X	X	X	X				X	X
AustralianFamilialAF	phs001435	HMB-NPU-MDS					X	X					
BAGS	phs001143	GRU-IRB	X	X	X	X	X	X				X	X
BioMe	phs001644	HMB-NPU						X	X	X			
CARDIA	phs001612	HMB-IRB						X					
		HMB-IRB-NPU						X					
CCAF	phs001189	GRU-IRB	X	X	X	X	X					X	X
CFS	phs000954	DS-HLBS-IRB-NPU	X	X	X	X	X	X				X	X
CHS	phs001368	HMB-MDS	X	X	X	X	X	X					X
		HMB-NPU-MDS	X	X	X	X	X	X					X
COPDGene	phs000951	HMB	X	X	X	X	X	X				X	X
		DS-CS-RD	X	X									X
CRA	phs000988	DS-ASTHMA-IRB-MDS-RD	X	X			X						X
DECAF	phs001546	GRU					X						
DHS	phs001412	DS-DHD-IRB-COL-NPU	X	X			X	X					X
		HMB-IRB-COL-NPU	X	X			X	X					X
EOCOPD	phs000946	DS-CS-RD	X	X									X
FHS	phs000974	HMB-IRB-MDS	X	X	X	X	X	X				X	X
		HMB-IRB-NPU-MDS	X	X	X	X	X	X				X	X
GALAI	phs001542	DS-LD-IRB-COL						X					
GALAII	phs000920	DS-LD-IRB-COL	X	X	X		X	X				X	X
GeneSTAR	phs001218	DS-CVD-IRB-NPU-MDS	X	X	X	X	X	X					X
GENOA	phs001345	DS-ASC-RF-NPU	X	X	X	X	X						X
GenSalt	phs001217	DS-HCR-IRB	X	X	X	X	X						X
GOLDN	phs001359	DS-CVD-IRB	X	X	X	X	X	X					X
HCHS_SOL	phs001395	HMB						X					
		HMB-NPU						X					
HVH	phs000993	DS-CVD-IRB-MDS	X	X	X	X	X	X				X	X
		HMB-IRB-MDS	X	X	X	X	X	X				X	X
HyperGEN	phs001293	DS-CVD-IRB-RD	X	X	X	X	X	X					X
		GRU-IRB	X	X	X	X	X	X					X
IPF	phs001607	DS-ILD-IRB-NPU						X					
		DS-LD-IRB-NPU						X					
		DS-PFIB-IRB-NPU						X					
		DS-PUL-ILD-IRB-NPU						X					
		HMB-IRB-NPU						X					
JHS	phs000964	DS-FDO-IRB	X	X	X	X	X	X				X	X
		DS-FDO-IRB-NPU	X	X	X	X	X	X				X	X
		HMB-IRB	X	X	X	X	X	X				X	X
		HMB-IRB-NPU	X	X	X	X	X	X				X	X
LTRC	phs001662	HMB-MDS						X					
Mayo_VTE	phs001402	GRU	X	X	X	X	X						X
MESA	phs001416	HMB	X	X	X	X	X	X					X
		HMB-NPU	X	X	X	X	X	X					X
MGH_AF	phs001062	DS-AF-IRB-RD	X	X	X	X	X					X	X
		HMB-IRB	X	X	X	X	X					X	X
miRhythm	phs001434	GRU					X						
MLOF	phs001515	HMB-PUB					X	X					
OMG_SCD	phs001608	DS-SCD-IRB-PUB-COL-MDS-RD					X	X					
Partners	phs001024	HMB	X	X	X	X	X					X	X
PharmHU	phs001466	HMB					X						
REDS-III_Brazil	phs001468	GRU-IRB-PUB-COL-NPU					X						
SAFS	phs001215	DS-DHD-IRB-PUB-MDS-RD	X	X	X		X	X					X
SAGE	phs000921	DS-LD-IRB-COL	X	X	X	X	X	X				X	X
Sarcoidosis	phs001207	DS-SAR-IRB	X	X	X	X	X	X					X
Samoan	phs000972	GRU-IRB-PUB-COL-NPU-GSO	X	X	X								X
SARP	phs001446	GRU					X						
THRV	phs001387	DS-CVD-IRB-COL-NPU-RD	X				X						
VAFAR	phs000997	HMB-IRB	X	X	X	X	X	X					X
VU_AF	phs001032	GRU-IRB	X	X	X	X	X	X				X	X
walk_PHaSST	phs001514	DS-SCD-IRB-PUB-COL-NPU-MDS-RD					X	X					
		HMB-IRB-PUB-COL-NPU-MDS-GSO					X	X					
WGHS	phs001040	HMB	X	X	X	X	X						X
WHI	phs001237	HMB-IRB	X	X	X	X	X	X					X
		HMB-IRB-NPU	X	X	X	X	X	X					X

Consent group data use limitations are defined as follows: GRU, general research use; HMB, limited to health, medical and/or biomedical purposes; DS, use of the data must be related to specified disease. Consent group data use limitation modifiers include the following: IRB, requestor must provide documentation of local IRB approval; PUB, requestor agrees to make results of studies using the data available to the larger scientific community; COL, requestor must provide a letter of collaboration with the primary study investigator(s); NPU, use of the data are limited to not-for-profit organizations; MDS, use of the data includes methods development research; GSO, use of the data are limited to genetic studies only. AF, atrial fibrillation; ASC-RF, arteriosclerosis and its risk factors; CVD, cardiovascular disease; CS, chronic obstructive pulmonary disease (COPD) and smoking; DHD, diabetes and heart disease; FDO, focus disease only (in JHS, FDO is blood pressure, heart/CVD, obesity, diabetes, kidney disease, or lung disease and risk factors); HCR, high blood pressure and related cardiovascular-renal disease; HLBS, heart, lung, blood and sleep disorders; ILD, interstitial lung disease; LD, lung disease; PFIB, pulmonary fibrosis; PUL, pulmonary, interstitial lung disease; RD, related disorders; SAR, sarcoidosis; SCD, sickle cell disease.

Extended Data Table 4 | Coverage, sequencing depth and number of variants

	All Individuals		Per Individual			
	Total	Singletons (%)	Average	5th %tile	Median	95th %tile
Samples	53,831	-	-	-	-	-
Bases (Gb)	6,973,670	-	130	107	128	157
Depth (x)	-	-	38	31	38	46
Genome Covered (%)	-	-	98.5	96.2	99.2	99.9
Depth >10x	-	-	97.9	95.4	98.7	99.6
Total Variants	410,323,831	188,947,391 (46)	3,776,362	3,515,416	3,567,439	4,364,075
SNVs	381,343,078	175,419,690 (46)	3,579,423	3,334,782	3,383,710	4,129,868
Indels	28,980,753	13,527,701 (47)	196,940	180,567	183,759	234,245
Novel* Variants	323,113,479	178,243,307 (55)	30,207	20,363	26,347	44,379
SNVs	298,028,808	165,082,153 (55)	25,861	17,568	22,909	36,897
Indels	25,084,671	13,161,154 (52)	4,345	2,752	3,378	7,392
Coding Variation	4,970,331	2,334,217 (47)	23,916	22,156	22,591	27,744
Synonymous	1,525,971	656,746 (43)	11,743	10,840	11,073	13,693
Non-synonymous	3,172,551	1,527,247 (48)	11,468	10,633	10,875	13,237
Stop/Essential Splice	105,042	56,801 (54)	478	426	456	568
Frameshift	113,805	67,903 (60)	133	112	127	167
Inframe	55,806	27,118 (49)	103	85	99	129

*Variant was not present in dbSNP build 149, the most recent dbSNP version without TOPMed submissions.

Extended Data Table 5 | pLOF variants in 53,831 individuals

	All Individuals		Per Individual			
	Total	Singletons (%)	Average	5th %tile	Median	95th %tile
pLoF	228,966	58.5	209	182	202	251
Stop gained	79,766	55.6	72	60	72	87
Frameshift	100,393	60.3	92	77	90	115
Splice	48,807	59.6	44	34	43	57
pLoF (AF < 0.5%)	217,795	58.8	20.7	10	19	35
Stop gained (AF < 0.5%)	75,904	55.8	7.9	3	7	15
Frameshift (AF < 0.5%)	95,064	60.6	8.3	3	8	15
Splice (AF < 0.5%)	46,827	59.9	4.5	1	4	9

nature research

Corresponding author(s): Gonçalo Abecasis

Last updated by author(s): November 19, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Complete description of data collection and used tools/software is available at https://www.nhlbiwgs.org/data-sets. We also provide a detailed description of TOPMed program's organization and data collection in Supplementary Information 1.1.
Data analysis	All code for TOPMed data quality checks and variant calling is available at https://github.com/statgen/topmed_variant_calling. Code for the WGS and WES data comparisons is available at https://github.com/statgen/sequencing_comparison. Code for modeling singleton distance distribution is available at https://github.com/carjed/topmed_singleton_clusters. Code for identifying novel genetic variants in unmapped reads is available at https://github.com/nygenome/topmed_unmapped. Code for gene burden association tests using rare pLoF variants is available at https://github.com/sgagliano/GeneBurden. Code for the imputed and genotype UK Biobank WES data comparisons is available at https://github.com/sgagliano/UKB_WES_vs_TOPMed_IMP. All programs used are open-source software developed by the academic community and published in scientific literature. For each used program we provide a version number and/or corresponding reference in main and supplementary texts, and methods. These programs include: ADMIXTURE v1.3.0, ANNOVAR, bcftools, Eagle 2.4, EPACTS, Fusera, GATK v4, GENESIS (R package), GotCloud, IBDNe, LiftOver, LOFTEE v0.3-beta, Minimac4, RefineIBD, RFMix, samtools, SeqVarTools (R package), Stargazer, VEP v94, verifyBamID2, WGSA.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

TOPMed data used in this manuscript are available through dbGaP. The detailed description of the TOPMed participant consents and data access is provided in Box 1. The dbGaP accession numbers for all TOPMed studies referenced in this paper are listed in Extended Data Tables 2 and 3. Complete list of TOPMed genetic variants with summary level information used in this manuscript is available through the BRAVO variant browser (bravo.sph.umich.edu). TOPMed imputation reference panel described in this manuscript can be used freely for imputation through the NHLBI BioData Catalyst at TOPMed Imputation Server (imputation.biodatacatalyst.nhlbi.nih.gov). The insertion callset (no genotypes) is available on dbGaP under the TOPMed GSR accession phs001974.

The following publicly available databases/datasets were used in the analysis: 1000 Genomes Project (https://www.internationalgenome.org), CADD (https://cadd.gs.washington.edu), ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), COSMIC (https://cancer.sanger.ac.uk/cosmic), dbSNP (https://www.ncbi.nlm.nih.gov/snp/), Ensembl (https://uswest.ensembl.org/index.html), ExAC (https://gnomad.broadinstitute.org), GENCODE (https://www.gencodegenes.org), GWAS Catalog (https://www.ebi.ac.uk/gwas/), HGDP (https://www.hagsc.org/hgdp/), OMIM (https://omim.org), UK Biobank (https://www.ukbiobank.ac.uk).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample-size calculation was performed. We used all whole genome sequencing data available at the moment at the TOPMed program.
Data exclusions	Sample level QC and variant level QC for TOPMed are extensively described in the methods and supplementary text. There were no exclusions based on participants disease status.
Replication	We did not attempt to reproduce any findings in a separate dataset, because no high depth whole genome sequencing datasets of comparable size were available at the moment of writing.
Randomization	Randomization was not performed, because this is a population-based study aggregating whole-genome sequencing data from >80 different established studies with varying designs.
Blinding	Blinding was not performed, because this is a population-based study aggregating whole-genome sequencing data from >80 different established studies with varying designs.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Lymphoblastoid cell lines were used for a small number of samples.
Authentication	Compared to previous genetic analysis.
Mycoplasma contamination	Mycoplasma sequence data was aligned to human genome which should avoid any mycoplasma originating reads.
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Human research participants

Policy information about studies involving human research participants

Population characteristics	A primary goal of the TOPMed program is to improve scientific understanding of the fundamental biological processes that underlie heart, lung, blood, and sleep (HLBS) disorders. 37% of all participants come from studies with focus on heart disorders, 33% - on lung disorders, 11% - on blood disorders, 1% - on sleep disorders, and 18% - on multiple phenotypes (Supplementary Figure 1). 41% of all participants have European ancestry, 31% - African ancestry, 15% - Hispanic/Latino, 9% - Asian ancestry, 4% - others/unknown (Supplementary Figures 2 and 39). 60% of individuals are females (Supplementary Figure 40).
Recruitment	TOPMed consists of ~155k participants from >80 different studies with varying designs: prospective cohorts, case-control studies, extended family structures and population isolates. Studies were biased towards individuals with heart, lung, blood, and sleep disorders, and who are willing to participate in research. More details are available at https://www.nhlbiwgs.org.
Ethics oversight	Informed consent was obtained from all participants, and the recruiting institutions for the >80 different TOPMed studies provided ethical oversight (see Box 1 and Supplementary Information 1.1.1-1.1.2). See Supplementary Information 4 for per study ethics statements.

Note that full information on the approval of the study protocol must also be provided in the manuscript.