Supplementary Materials and Figures
Deep Sequencing of 10,000 Human Genomes

Material and Methods

Supplementary references

Supplementary Figures

Material and Methods

1. Participants and Sample Preparation

Participants were representative of major human populations and ancestries (Supplementary Fig S1). The study population was not ascertained for a specific health status: 3,940 individuals were presumed to be healthy adults, 5,656 showed signs of common disorders (cardiovascular, respiratory, metabolic syndrome, neurodegenerative disorders and aging), 664 were diagnosed with neurodevelopmental and rare disorders, and 285 were predisposed to cancer (germinal) and immune disorders. New and existing IRB-approved consent forms for participation in research and collection of biological specimens and other data used in this publication were reviewed and confirmed to be appropriate for use.

Blood specimens were collected into 4 mL EDTA Anti-Coagulant Vacutainer tubes and stored at 2-8°C for a maximum of 5 days. Genomic DNA extraction was carried out using a Chemagic DNA Blood400 kit following manufacturer's recommendations. DNA was eluted in 50uL Elution Buffer (EB, Qiagen) and stored at 4°C until used.

Double-stranded DNA was quantified with a Quant-iT fluorescence assay (Life Technologies). The genomic DNA was normalized and sheared with a Covaris LE220 instrument. Next Generation Sequencing (NGS) library preparation was carried out using the TruSeq Nano DNA HT kit (Illumina Inc), essentially following manufacturer's recommendations. Individual DNA libraries were characterized in regards to size and concentration using a LabChip DX One Touch (Perkin Elmer) and Quant-iT (Life Technologies), respectively. Libraries were normalized to 2-3.5nM and stored at -20°C until used.

2. Clustering and Sequencing

Normalized DNA libraries were combined into 6-sample pools and clustered on cBot cluster stations following the manufacturer's recommendations. Two different versions of the Clustering/SBS kits were used, v1 and v2, corresponding to the original (March 2014) and its replacement configuration (October 2015). It is worth noting that the current version, v2, includes a revised version of the original clustering protocol, requiring an upfront DNA denaturation step and a longer clustering chemistry. All flowcells were sequenced on the Illumina HiSeqX sequencer utilizing a 150 base paired-end single index read format.

3. Read mapping and genotyping

Base call (BCL) files were used to map reads to a human reference sequence (hg38 build) using ISIS Analysis Software (v. 2.5.26.13; Illumina) *(1)*. The hg38 reference sequence was modified by masking the pseudoautosomal region of chrY. The ISIS Isaac Aligner (v. 1.14.02.06) identified and marked duplicate reads, and these were removed from downstream analysis. The resulting bam files were characterized using Picard (v. 1.113-1.131), and input to the ISIS Isaac Variant Caller (v. 2.0.17). The Isaac Variant Caller was used with default settings, and yielded genomic VCF files (gVCF). For computation of accuracy, single nucleotide variants with a "PASS" flag were compared to GIAB (v. 2.19; ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv2.19). The data for the GiaB high confidence region are derived from 11 technologies: BioNano Genomics, Complete

Genomics paired-end and Long Fragment Read, Ion Proton, Oxford Nanopore, Pacific Biosciences, SOLiD, 10X Genomics GemCodeTM WGS, and Illumina paired-end, mate-pair, and synthetic long reads.

4. Ancestry admixture

The 1000 Genome Project (26 populations) *(2)* and Human Genome Diversity Project (HGDP, 52 populations) *(3)* were used as a reference panel for ancestry admixture analysis. Admixed populations including African Americans or Latin Americans from the 1000 Genome Project were excluded from the reference panel. Shared dbSNP rsIDs were used to merge genotype data from HGDP and the 1000 Genome Project. SNPs with discordant forward strand alleles between genome build 36 and 38 were removed to avoid assembly inconsistency. SNPs with discordant forward strand alleles between genome build 36 and 38 were removed to avoid assembly inconsistency. This process resulted in 3,444 samples with 636,698 SNPs, of which 116,990 were then pruned due to linkage disequilibrium using PLINK (1.9), leaving 519,708 SNPs. Allele frequencies were calculated for each population, and the top 3,000 most informative SNPs for each population (ranked based on the absolute Z-score for each allele's frequency against the whole panel of populations) were extracted. The resultant collection of 57,214 unique, ancestry-informative SNPs was used for ancestry admixture analysis using ADMIXTURE (1.23) *(4)*. Also, from the set of 519,708 SNPs, an additional set of 107,570 unique most variable SNPs (alt allele frequencies close to 0.5), were selected to supplement the above ancestry informative SNPs for PCA (performed using PLINK (1.9) *(5)*.

5. Kinship analysis

The relatedness of individuals was carried out by first extracting the genotypes for 162,997 autosomal SNVs of each sample. The samples were then merged and a bed file was generated using PLINK (1.9). The program KING (1.4) *(6)* was used to determine relatedness of the samples. Unrelated samples were identified using the default kinship coefficient cutoff of 0.0884. From this analysis it was determined there were 8,096 unrelated samples.

6. Assessment of human sample contamination

We used verifyBamID (7) to control for sample mixtures. At the conservative cutoff of 3%, we identify a 0.83% of contamination for samples processed in their entirety in our laboratory (i.e., from blood DNA extraction onwards), compared with 1.9% when considering all samples (i.e., from pre-extracted DNA).

7. False discovery rate

We estimated the false discovery rate (FDR) of our sequencing pipeline using 200 replicates of NA12878. Within the GiaB high confidence regions, we compared genotype calls to those reported by GiaB. We calculated the FDR at variant sites using the formula *FDR = FP/(TP+FP).* We also estimated the genome-wide FDR by accounting for missingness. We counted the number of sites where genotypes could not be called reliably (i.e. no-PASS calls) and denoted that as missingness (Fig. 1c). If we use 90% reproducibility as the filtering criteria, those positions with high (>=90%) reproducibility in NA12878 within GiaB high confidence region would be regarded as "positives", while those with low reproducibility would be regarded as

"negatives". We used the same FDR formula as above but with different definitions of "false positives" and "true positives". For those sites within GiaB high confidence region sequenced with high reproducibility, they would be regarded as "false positives" if there was a wrong call or the call did not pass the filter (missingness). We regarded those genotype calls that were different from GiaB as "wrong calls" (i.e. false positive calls and false negative calls mentioned in Fig. 1c). Those sites with consistent calls would be regarded as "true positives".

8. NA12878 quality metrics

We used reproducibility metrics to define regions within GiaB with high (≥90%) versus low (<90%) reproducibility at each position. The reproducibility metrics include the concordance in calls and missingness (defined in this work as a measure of low quality calls). A precise assessment of missingness is achieved by using a genomic variant call format file (gVCF) that informs every position in the genome regardless of whether a variant was identified or not. A total of 2,157 Mb (97.3%) of the GiaB high confidence region could be sequenced with high reproducibility, while 59 Mb (2.7%) were classified as less reliable (Table S1). False positive, false negative and missingness rates were considerably lower in the GiaB region sequenced with high reproducibility. At high reproducibility sites, the false discovery rate is very low (FDR = 0.0008). Other relevant metrics include a precision of 0.999, recall of 0.994 and F-measure of 0.996. If we use 90% reproducibility as the filtering criteria, the genome-wide false discovery rate is 0.0025. Other relevant metrics are a genome-wide precision of 0.998, recall of 0.980 and F-measure of 0.989

Table S1: Whole genome sequencing quality metrics. False positive calls are concentrated in the region of GiaB that has <90% reproducibility of base calling. False negative calls are more evenly represented across GiaB, missingness (no-PASS) represents the bulk of error.

	High (>=90%) reproducibility region of GiaB	Low (<90%) reproducibility region of GiaB
NA12878 positions*	2,156,734,782	58,938,283
False positive call	2,387 (0.0001%)	3,613 (0.0061%)
False negative call	16,773 (0.0008%)	78,186 (0.1327%)
Missingness	5,323,480 (0.2468%)	15,523,130 (26.3379%)

9. Extent of sequencing at the level of an individual genome

Many genome sequencing projects calculate quality statistics on a composite of all genomes sequenced, regardless of the depth or quality of an individual genome's sequence. Conversely, we chose to sequence an individual genome many times to assess the quality of our sequencing capabilities and identify regions of the genome for which we could consistently make high confidence calls. As such, our work specifically presents the confident genome calls ("extended confidence regions") for a single individual benchmarked against the complete sequence (Table S2). This difference between population genome level statistics and individual genome level statistics is significant as we move forward toward the use of an individual's genome information in the clinic.

Table S2: Individual level whole genome sequencing. Estimates are derived from the sequencing of 100 NA12878 replicates using v2 chemistry. Percentage 1 is the number of bases divided by the total length of autosomes and chrX. Percentage 2 is the number of bases divided by the total number of callable bases on autosomes and chrX (I.e. Not "N").

	Number of bases	Percentage1 (num of bases/A)	Percentage2 (no. of bases/(A-B))
Total chromosomal length of autosomes and chrX (A) ("reference genome"; hg38)	3,031,042,417		
Total number of "N" base on autosomes and chrX (B) ("inaccessible genome in reference genome")	130,962,786		
Total number of callable bases (A-B) ("accessible genome in reference genome")	2,900,079,631		
Average number of PASS position per NA12878 replicate (on autosomes + chrX) ("individual accessible genome")	2,750,001,288	90.73%	94.83%
Total length of extended confidence region (on autosomes + chrX) ("ECR")	2,583,500,276	85.23%	89.08%

10. Structural and copy number variation in NA12878

To understand the performance of structural and copy number variation analysis using Illumina short read technology, we studied precision, recall and reproducibility in the set of 200 NA12878 sequences. For short indels, we compared our calls to those reported in the newest release of GIAB (NIST v3.2.2). The average precision and recall rates achieved by ISIS Isaac Variant Caller (v.2.0.17) are 97.80% and 86.32% respectively, but with unsatisfactory reproducibility (Table S3). For SV, we compared the performance of 7 software; for deletion: Pindel (8), DELLY (9), GenomeSTRiP (10), BreakDancer (11), LUMPY (12), MatchClip2 (13), Manta (14); for insertion: Pindel (8), Manta (14). We used the list of SV from Pendleton et al. (15) as the reference set, and Manta performed the best among the 7 software and was used in the analysis of the 200 NA12878 replicas. For CNV, we compared the performance of 5 software: cn.mops (16), CNVnator (17), GenomeSTRiP (10), MatchClip2 (13), Canvas (18). We used the list of CNV from Conrad et al. (19) array data as reference set, and Canvas performed the best among the 5 software. The performance for SV and CNV is presented in Table S3. Overall, the results of analyses were deemed unsatisfactory for clinical use.

Table S3: Performance of SV and CNV calling in 200 runs of NA12878.

	Type	**Precision**	**Recall**	**Average percentage of calls with reproducibility >= 90% per sample**	**Reference set**	**Software**
Small indels (1-50 bp)	--	97.80%	86.32%	78.98% *	NIST GIAB v3.2.2	ISIS Isaac Variant Caller

						(v.2.0.17)
Small indels (1-50 bp)	--	97.21%	72.75%	78.50% *	NIST GIAB v2.19	ISIS Isaac Variant Caller (v.2.0.17)
Structural variations (>50 bp)	Deletion**	60.62%	35.86%	48.03%	Pendleton et al. 2015 (15)	MANTA v0.29.4
	Insertion***	76.53%	10.83%	29.36%	Pendleton et al. 2015 (15)	MANTA v0.29.4
Copy number variations	LOSS**	71.05%	23.94%	52.49%	Conrad et al. 2010 conrad (19)	CANVAS v1.3.5
	GAIN**	3.75%	4.55%	43.65%	Conrad et al. 2010 (19)	CANVAS v1.3.5

* Restricted to calls within GIAB HC region
** required >50% overlap
*** any overlap with reference calls within +/-20 bp of call

11. Construction of extended confidence region

We defined an extended confidence region (ECR) that includes the high confidence GiaB regions and the highly reproducible regions extending beyond the boundaries of GiaB. The set of high-reproducibility regions on autosomes and the X-chromosome was established based on analysis of NA12878 replicates. Sites with ≥90% agreement/concordance in genotype calls among NA12878 replicates were included in the high-reproducibility set; the rest were regarded as low reproducibility.

Suppl. Fig. S2 illustrates the noise we observed outside of the GiaB regions, both in terms of spurious variant calls and of apparent conservation. Of 3,088 Mb of sequence (autosomal, X- and Y-chromosomes), the overlap of GiaB high confidence and highly reproducible regions represented 69.8% of the analyzed positions. The non-GiaB regions with high variant call reproducibility covered an additional 14.1% of the genome.

We used a different strategy for defining high-reproducibility regions on the Y-chromosome because NA12878 is a female. Instead, we used genotype calls from 100 males among the 10,545 samples. Sites with ≥90% genotype calls passing quality control thresholds and without any heterozygous calls were included in the high-reproducibility set. Sites with variants calls on the Y-chromosome in 100 female samples were excluded from the high-reproducibility set (Suppl. Fig. S3). In addition, centromeric regions and known segmental duplication regions were excluded from the high-reproducibility set. They were obtained from the UCSC genome browser
(hg38, http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/cytoBand.txt.gz; http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/genomicSuperDups.txt.gz).

Illumina short read sequencing excluded 2.7% of GiaB high confidence region. Basically, these regions, after excluding segmental duplications and centromere sequences represent 35,6Mb. Approximately one third (34.3%) of the total is called reproducibly (but below 90%) in 80-89% of samples, while 7.3% of the total (2.6 Mb) is never called in any of the samples. The excluded sequence is highly distributed in small segments of less than 2020 bp. The regions are also

enriched in indel calls. Only 0.48% of the region is protein-coding. Almost all (89.14%) is annotated as repetitive sequence.

The definition of ECR allowed for more high confidence calls than those identified in GiaB (Table S4).

Table S4. Contribution of the extended confidence region to variant identification in 10,545 genomes.

		Variant Sites*	Var. sites in the ECR (%)	Var. sites in GiaB (%)
		165,007,222	146,693,004 (89%)	125,513,963 (76%)
Annotation	intergenic_region	62,486,344	54,347,350 (87%)	46,770,589 (75%)
	intron_variant	58,868,383	54,328,376 (92%)	46,931,777 (80%)
	non_coding_exon_variant	559,889	418,532 (75%)	348,897 (62%)
	3_prime_UTR_variant	1,900,062	1,738,715 (92%)	1,427,930 (75%)
	5_prime_UTR_variant	486,584	424,817 (87%)	328,399 (67%)
	upstream_gene_variant	22,449,371	19,511,283 (87%)	16,271,660 (72%)
	downstream_gene_variant	16,221,758	14,091,715 (87%)	11,953,456 (74%)
	TF_binding_site_variant	58,852	48,664 (83%)	37,261 (63%)
Variant Effect	splice_acceptor_variant	13,793	11,687 (85%)	9,347 (68%)
	splice_donor_variant	19,106	16643 (87%)	13481 (71%)
	missense_variant	1,168,296	1,055,738 (90%)	853,764 (73%)
	synonymous_variant	649,059	590,469 (91%)	476,943 (73%)
	start_lost	3,781	3,344 (88%)	2,734 (72%)
	stop_gained	30,292	26,925 (89%)	21,691 (72%)
	stop_lost	1,767	1,559 (88%)	1,245 (70%)
Pathogenicity	HGMD-DM	8,611	8,198 (95%)	6,831 (79%)
	ClinVar Pathogenic	3,390	3,191 (94%)	2,628 (78%)

* The total number of SNVs observed at these positions is 170,113,857, including multi-allelic positions. ECR: Extended confidence region

12. Reproducibility on mitochondrial DNA

We assessed the reproducibility of sequencing calls of the mitochondrial DNA using NA12878 replicates. For the 16,569 sites on mtDNA, 99.95% could be sequenced with high (>=90%) reproducibility. However, of these positions, 13,693 fell within known segmental duplication regions, according to the list of known segmental duplications obtained from the UCSC genome browser. Given this large overlap and the heteroplasmic nature of mtDNA, we excluded mtDNA from the ECR.

13. Annotation

ClinVar (VCF v4.0; fileDate=20150916) *(20)* pathogenic variants (i.e. CLINSIG=5) and HGMD (2015-R2) *(21)* disease-causing mutations (DM) were used for annotating clinically relevant variants. To generate the metaprofiles of the pathogenic variants, we used the SNVs from

ClinVar and HGMD databases. SnpEff *(22)* was used for genomic annotation and predicting effects of SNVs sites observed in the 10,545 samples (shown in Main Text Table 1). Exomic regions for protein-coding genes were extracted from GENCODE (Release 23) *(23)*.

14. Distribution of SNVs per element

The annotation file for GENCODE v23 was used as the initial annotation of genomics elements including intergenic, protein coding and RNA coding elements. Genomic positions that did not overlap any annotation in the file were defined as "intergenic". "Intronic lncRNAs" that were tagged with "sense_intronic" in either the gene_type and/or the transcript_type fields were marked as "intronic lncRNAs", while "protein coding", "lincRNA", "snoRNA" and "miRNA" in either the gene_type and/or the transcript_type were tagged with their respective names. The annotation "constitutive exons" correspond to regions that are consistently exonic in all isoforms of a gene; "alternative exons", to the regions that are exonic in at least 1 isoform, but not all; "constitutive introns", to the regions consistently intronic in all isoforms; and "alternative introns", to the regions that are consistently intronic, but are not present in all isoforms. When several exons/introns from different isoforms arising from the same gene were overlapping, only the minimum overlap of all isoforms was considered. The "origin of replication regions" (oriC) were obtained from a public HeLa ChIPseq dataset (accession number: GSM922790). Regulatory elements were obtained from the Ensembl Regulatory Build *(24)* for consistency and independence of the cell types. SNV presence was assessed at genomic positions overlapping the ECR. Only elements with at least 90% overlap with those regions were used, except for the intergenic regions, where 100% overlap was required. Values are summarized in Table S5.

In order to compare the SNV distribution across different element types, given that their size distribution varies significantly, we concatenated all elements from the same type (ex: all intergenic regions) and then reported the overall number of SNVs per kb in the total element (Figure 2A). To assess the range of variation, we chopped the newly concatenated element into 1kb windows and extracted the number of SNVs for each 1kb window (the results are shown in supplementary Figure S4).

Table S5: Overall summary of SNVs in the various genomic elements and regions.

Count= total number of the element. Length= overall sum of the ECR sizes of the elements. SNVs Total= sum of mapped SNVs in ECR. SNVs Mean= SNVs Total/Length.

Element	Count	Length	SNVs Total	SNVs Mean
oriC	11517	9433486	540262	0.0573
intergenic	30414	1142686603	65593232	0.0574
protCod (AE)	200494	33429295	1805563	0.0540
protCod (CI)	21442	110966273	6033109	0.0544
protCod (CE)	27699	8832708	488347	0.0553
protCod (AI)	146788	629489052	34918651	0.0555
intronic lncRNA (AI)	53	446218	24981	0.0560

lincRNA (AI)	4126	40803762	2290503	0.0561
intronic lncRNA (CI)	590	2972912	168752	0.0568
lincRNA (CI)	8881	98817557	5627274	0.0569
intronic lncRNA (CE)	1399	708859	41130	0.0580
miRNA	3283	295819	17184	0.0581
lincRNA (CE)	14557	5262785	314381	0.0598
intronic lncRNA (AE)	112	21762	1301	0.0598
lincRNA (AE)	8107	1722275	103570	0.0601
snoRNA	868	99090	6043	0.0610
Enhancer	119624	64671180	3525956	0.0545
Promoter Flanking	78183	146107431	8162054	0.0559
CTCF	101615	54526297	3074159	0.0564
Open Chromatin	61911	30901017	1813806	0.0587
TFBS	21804	11194969	664475	0.0594
Promoter	12621	24385606	1529677	0.0627

15. Creation of metaprofiles

A schematic description of the generation of metaprofiles is presented as Suppl. Fig S9. To assess SNV metaprofiles, GENCODE v23 was used as the starting gene annotation for the six genomic landmarks in protein coding genes (transcription start site, TSS; start codon; splice donor site, SD; splice acceptor site, SA; stop codon; polyadenylation site, pA). TSS and pA were defined as the first or last nucleotide of a transcript, respectively, start and stop codons were tagged with the same nomenclature in the annotation file, SA and SD were defined as the first nucleotide of the exon and one nucleotide after the last nucleotide of the exon, respectively. For SA/SD, exons were excluded if they were annotated as the first exon in an isoform or the last exon, respectively. To have a clean set of elements that could be aligned reliably, several filters were applied. The genomic position(s) of the genomic landmark itself had to be in the ECR. In addition, the exons selected for SA/SD metaprofiles had to have the same start or end coordinate, respectively, in all isoforms where they were present. Finally, the flanking introns of the exons selected for SA/SD had to end/start with the consensus sequence AG or GT, respectively. Redundant annotations at a given position were removed so as to avoid overcounting. To build metaprofiles, the SNV presence and frequency information was extracted for each element that passed these filters, along with every nucleotide 100 bps up and downstream of the genomic landmark. To avoid confounding effects from different genomic landmarks or low confidence regions, only positions that were at least 5 bp from another annotated genomic landmark and in the ECR were used. To compare variability across all six genomic landmarks tested, for each of the 200 positions surrounding the six genomic landmarks, the percent of elements with SNVs (number of elements with a SNV / the number of assessable elements at a given position) was divided by the mean percentage obtained across the 6 genomic landmarks (1200 bp), so that the mean normalized score across the six genomic landmarks would be equal to 1. For each of the 1200 bp the percent of SNVs present at an allelic frequency higher than 1 in a 1000 alleles was extracted.

Transmembrane domain amino acid coordinates were obtained from Uniprot (http://www.uniprot.org/) and were mapped back to genomic coordinates using UCSC knownGene table (https://genome.ucsc.edu/cgi-bin/hgTables). The elements and assessable positions were filtered with the same criteria as used for protein coding genes. The percentage of elements with SNVs was divided by the mean percentage obtained across the 6 protein coding genomic landmark for easier comparison.

Transcription factor binding sites (TFBS) genomic coordinates were obtained from Jaspar (http://jaspar.genereg.net/). The elements and assessable positions were filtered with the same criteria as used for protein coding genes. The percentage of elements with SNVs was divided by the mean percentage obtained across the 6 protein coding genomic landmark for easier comparison.

16. Identification of non-reference sequences

Sequencing read pairs with one or both reads not mapped to the hg38 reference assembly were extracted and annotated as "unmapped reads" if they showed both high sequencing quality and non-repetitiveness. Samples with more than 10% unmapped reads were excluded from analysis. Unmapped reads were then assembled using SOAPdenovo2 (v2.04) *(25)* with kmer size 91 for each sample. Assembled contigs longer than 200bp were mapped against the hg38 assembly to remove contigs that can map to the reference with >90% identity on >30% length. The remaining contigs were then mapped to the hg38 regions that were masked as repeat in UCSC goldenPath *(26)* using BLASTN *(27)* without low complexity filtering to remove contigs that contains >20% repeat sequences. The contigs passing the above filtering steps were clustered into a non-redundant set using CD-Hit (v4.6) *(28)* with 90% global identity threshold. In order to filter out contigs that are of non-human origin, we compared the non-redundant contig set against the NCBI protein database nr using DIAMOND (v0.7.9) *(29)*, and against NCBI nucleotide database nt using DNA aligner SASS (v0.3.2, Unpublished, part of the winning solution to the DTRA Algorithm Challenge) *(30)*. Contigs with non-mammal matches were considered contamination and removed from analysis. Contigs that did not match to nt or nr were included if their GC content was between 30 and 50% and their dinucleotide bias was less than 15% *(31)*. The remaining contigs were considered non-reference contigs. Those contigs were compared against the alt and patch sequences in GRCh38.p5, and contigs that mapped to the alt or patches were considered positive controls. The remaining set of contigs were classified as putative novel human sequences.

Neanderthal *(32)* and Denisovan *(33)* sequence data were downloaded from http://cdna.eva.mpg.de. Sequencing reads were mapped to hg38 assembly using BWA. The unmapped Neanderthal and Denisovan reads were then mapped to the Human non-reference contigs to calculate the overlap of novel human contigs with archaic genomes.

17. Data access

We have provided FDA with 325 vcf files for the NA12878 replicates that we sequenced. We also provided them with 6 pairs of fastq files corresponding to 3 of the replicates. This is

basically the raw data that can be run on any bioinformatics pipeline for testing and comparison purposes. Access to these replicates can be provided by PrecisionFDA on their cloud based platform upon request to the FDA.

HLI is providing access to the data in aggregate form through a public browser (link provided at publication). HLI supports the effort of making genetic data broadly available to further scientific research, but believes that controls over access will help ensure protection of the privacy of those individuals who have agreed to have their genomic sequencing data placed in the HLI database, as well as that of their family members.

References

1. Raczy C*, et al.* (2013) Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* 29(16):2041-2043.
2. Genomes Project C*, et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56-65.
3. Cavalli-Sforza LL (2007) Human evolution and its relevance for genetic epidemiology. *Annu Rev Genomics Hum Genet* 8:1-15.
4. Alexander DH, Novembre J, & Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome res* 19(9):1655-1664.
5. Chang CC*, et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7.
6. Manichaikul A*, et al.* (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867-2873.
7. Jun G*, et al.* (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* 91(5):839-848.
8. Ye K, Schulz MH, Long Q, Apweiler R, & Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25(21):2865-2871.
9. Rausch T*, et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28(18):i333-i339.
10. Handsaker RE*, et al.* (2015) Large multiallelic copy number variations in humans. *Nat Genet* 47(3):296-303.
11. Chen K*, et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 6(9):677-681.
12. Layer RM, Chiang C, Quinlan AR, & Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* 15(6):R84.
13. Wu Y, Tian L, Pirastu M, Stambolian D, & Li H (2013) MATCHCLIP: locate precise breakpoints for copy number variation using CIGAR string by matching soft clipped reads. *Front Genet* 4:157.
14. Chen X*, et al.* (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 32(8):1220-1222.

15. Pendleton M*, et al.* (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* 12(8):780-786.
16. Klambauer G*, et al.* (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic acids res* 40(9):e69.
17. Abyzov A, Urban AE, Snyder M, & Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome res* 21(6):974-984.
18. Roller E, Ivakhno S, Lee S, Royce T, & Tanner S (2016) Canvas: versatile and scalable detection of copy number variants. *Bioinformatics* 32(15):2375-2377.
19. Conrad DF*, et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464(7289):704-712.
20. Landrum MJ*, et al.* (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids res* 42(Database issue):D980-985.
21. Stenson PD*, et al.* (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 133(1):1-9.
22. Cingolani P*, et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6(2):80-92.
23. Harrow J*, et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome res* 22(9):1760-1774.
24. Zerbino DR, Wilder SP, Johnson N, Juettemann T, & Flicek PR (2015) The ensembl regulatory build. *Genome Biol* 16:56.
25. Luo R*, et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1(1):18.
26. Rosenbloom KR*, et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic acids res* 43(Database issue):D670-681.
27. Camacho C*, et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
28. Fu L, Niu B, Zhu Z, Wu S, & Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150-3152.
29. Buchfink B, Xie C, & Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12(1):59-60.
30. Servick K (2013) Bioinformatics. Top contenders blast Pentagon's new bioterror detection prize. *Science* 341(6145):449.
31. Gentles AJ & Karlin S (2001) Genome-scale compositional comparisons in eukaryotes. *Genome res* 11(4):540-546.
32. Prufer K*, et al.* (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505(7481):43-49.
33. Meyer M*, et al.* (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338(6104):222-226.
34. Poznik GD*, et al.* (2013) Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* 341(6145):562-565.

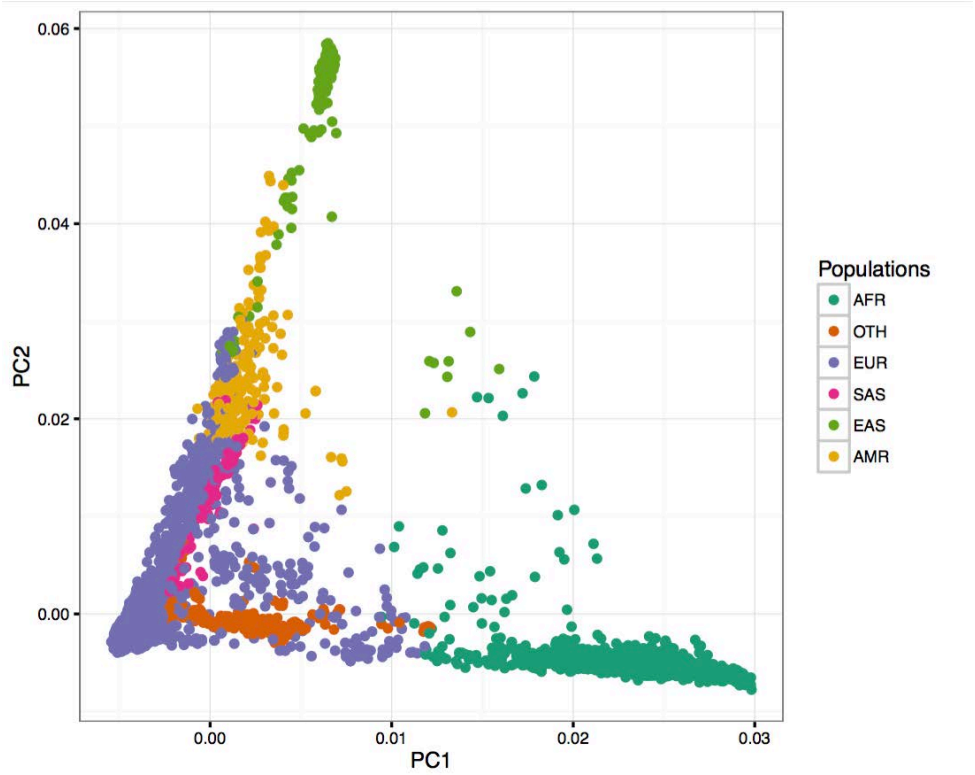
35. Rao SS*, et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665-1680.
36. Bartha I*, et al.* (2015) The Characteristics of Heterozygous Protein Truncating Variants in the Human Genome. *PLoS Comput Biol* 11(12):e1004647.

Supplementary Figures

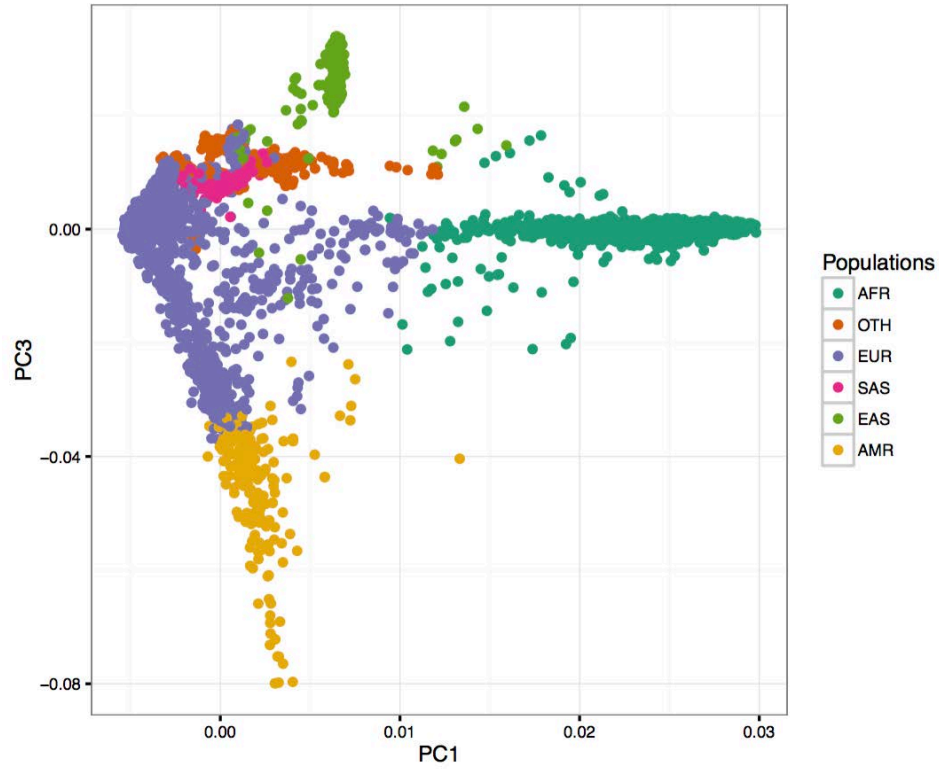
Suppl. Fig S1. Principal component analysis of study populations.
The 1000 Genome Project and Human Genome Diversity Project were used as a reference panel for ancestry admixture analysis. PCA analysis was performed using PLINK (1.9) on 162,997 ancestry informative markers. **(A)** PC1 and PC2. **(B)** PC1 and PC3. **(C)** PC3 and PC4. Genomes are colored, based on the largest admixture ancestry. The super-populations described by the 1000 Genome Project are: EUR= European, AFR= African, SAS= South Asian, EAS= East Asian, AMR= Native American, OTH= others including Siberian, Middle Easterner and Oceania; numbers are shown in the table below. **(D)** Because of the evidence in the PCA of extensive admixture, we alternatively assigned individuals to five superpopulations as described by The 1000 Genomes Project, or to an admixed population group (ADMIX, grey) on the basis of genetic ancestry. Ancestry admixture were performed using ADMIXTURE (1.23).

	Individuals in human super-populations	Numbers when considering admixture patterns	Unrelated individuals
EUR	8283	7310	5543
AFR	1397	1098	955
SAS	135	98	65
EAS	218	178	152
AMR	199	17	12
ADMIX	-	1636	1293
OTH	313	208	76

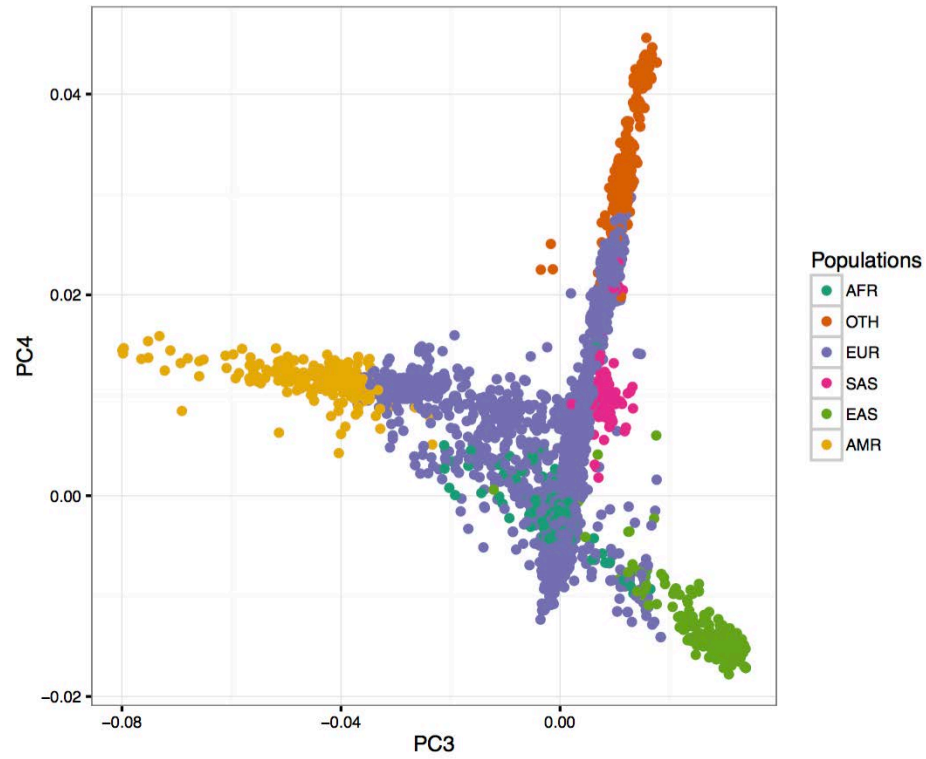
A


0.06
0.04
0.02
0.00
PC2
0.00
0.01
0.02
0.03
PC1
Populations
AFR
OTH
EUR
SAS
EAS
AMR

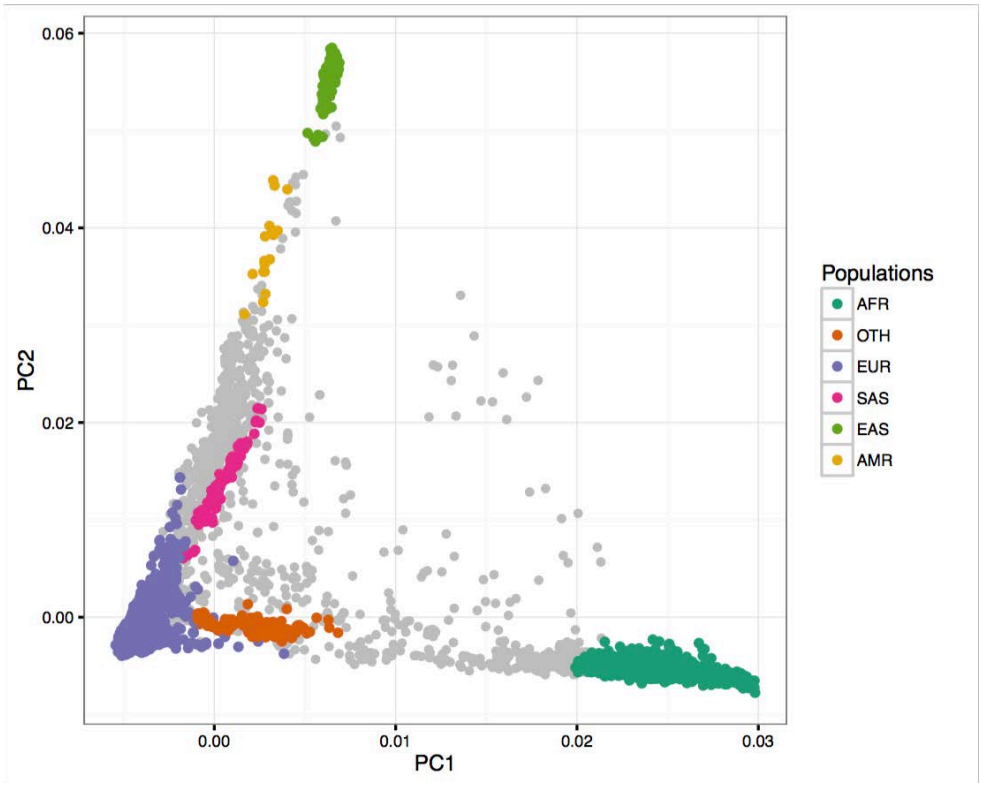
B


0.00
−0.04
−0.08
PC3
0.00
0.01
0.02
0.03
PC1
Populations
AFR
OTH
EUR
SAS
EAS
AMR

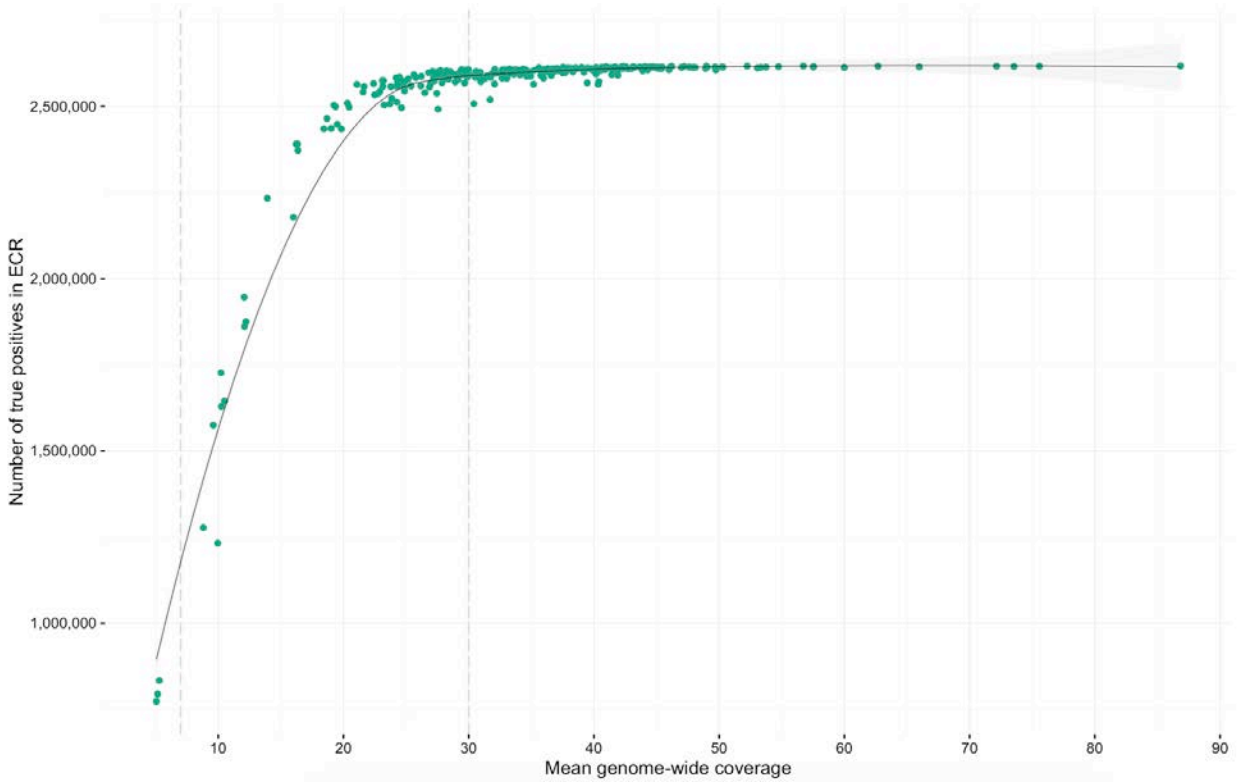
C


0.04
0.02
0.00
−0.02
PC4
−0.08
−0.04
0.00
PC3
Populations
AFR
OTH
EUR
SAS
EAS
AMR

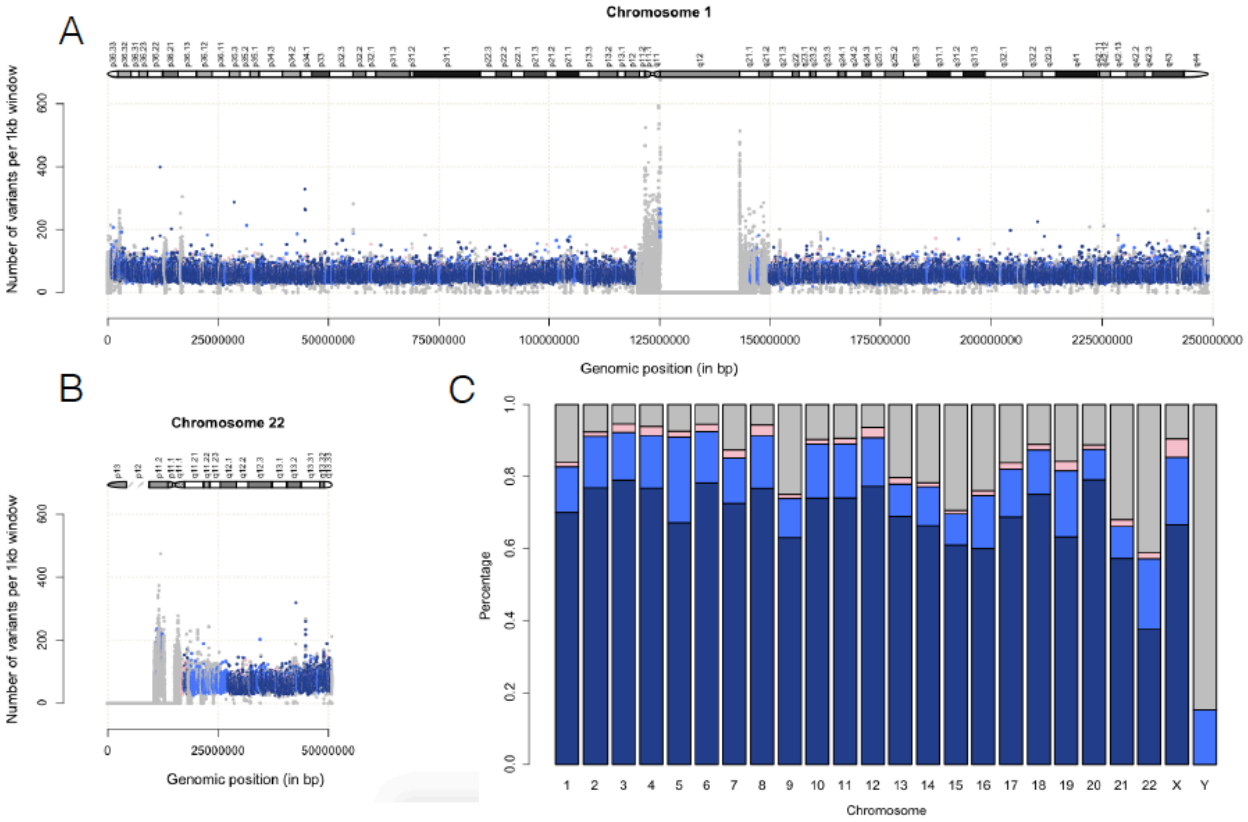
D


0.06
0.04
0.02
0.00
PC2
0.00
0.01
0.02
0.03
PC1
Populations
AFR
OTH
EUR
SAS
EAS
AMR

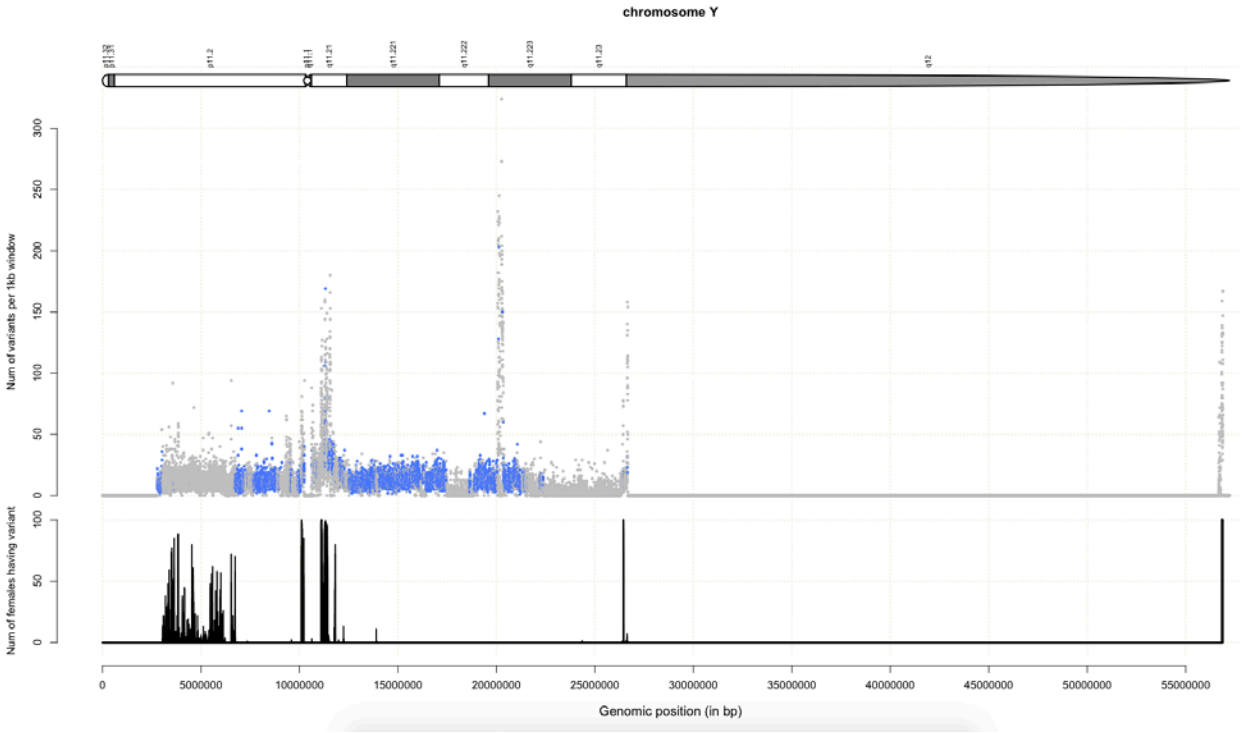
Suppl. Fig S2. Impact of sequencing depth on variant calling.
Depicted is the number of true positive SNVs detected in NA12878 replicates with different mean genome-wide coverages. Only SNV calls in both ECR and GiaB high confidence regions were considered (total number of true positives = 2,618,794). Genome-wide coverages of 7X and 30X are indicated by the vertical grey dash lines. For a genome sequenced at mean 30X, around 99% of the SNVs are detected. However, for a genome sequenced at mean 7X coverage, less than half of the true positive SNVs are detected.



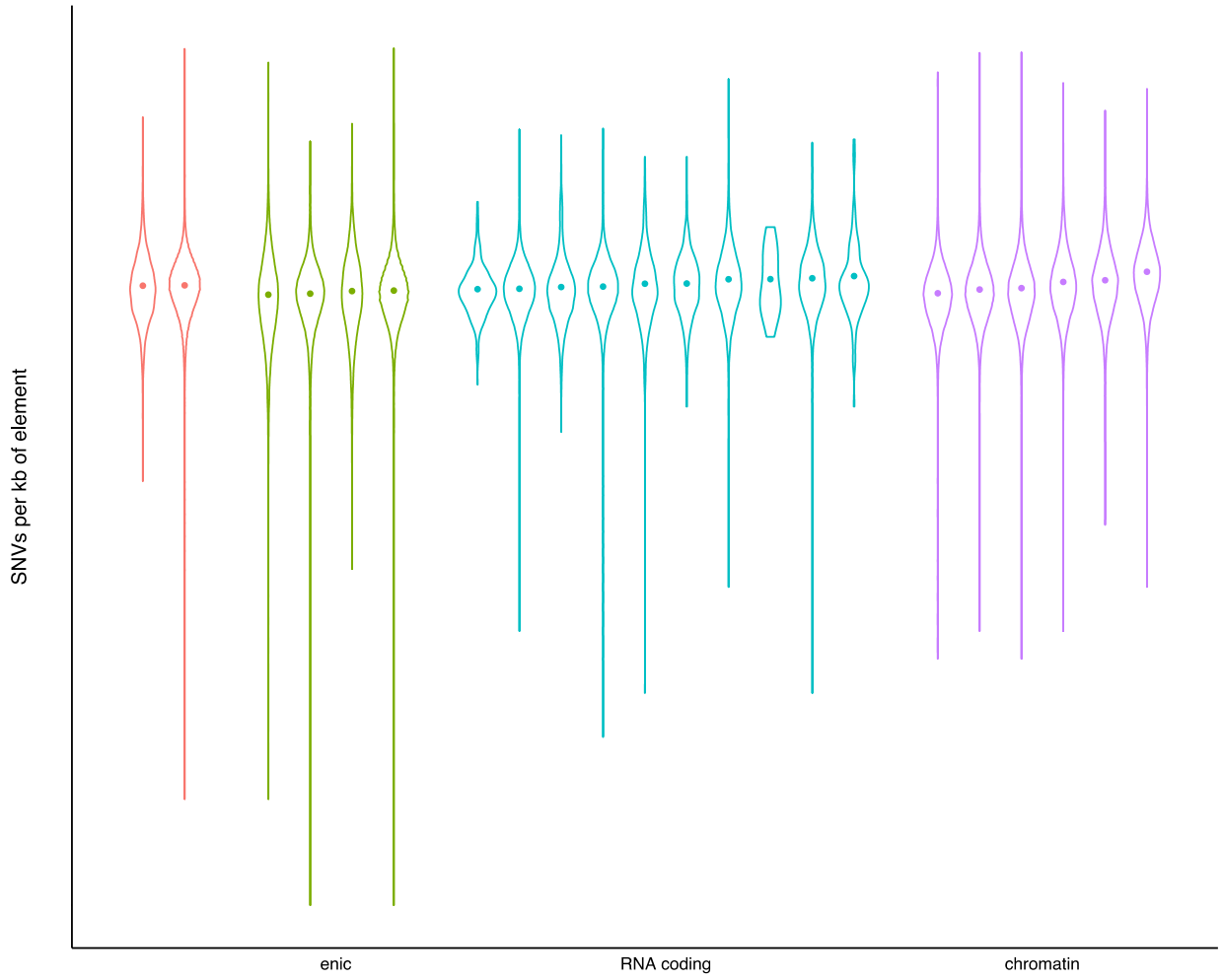
Suppl. Fig. S3. Sequence reliability and rates of variation in 10,545 genomes.
Genome view of two representative autosomal chromosomes; **(A)** Chr.1 as the longest and **(B)** Chr. 22 with the lowest proportion of sequenceable bases with the technology used. Each datapoint represents a 1kb window; the Y axis represents the number of SNVs per 1kb; dark blue are high confidence windows (the overlap of GiaB high confidence regions and regions with >=90% reproducibility in NA12878 replicates); light blue are extended confidence windows outside of GiaB; pink are GiaB only (low reproducibility with current technology); grey dots are regions outside of GiaB and extended confidence regions. **(C)** Summary statistics for all the chromosomes, using the same color coding as in previous panels.



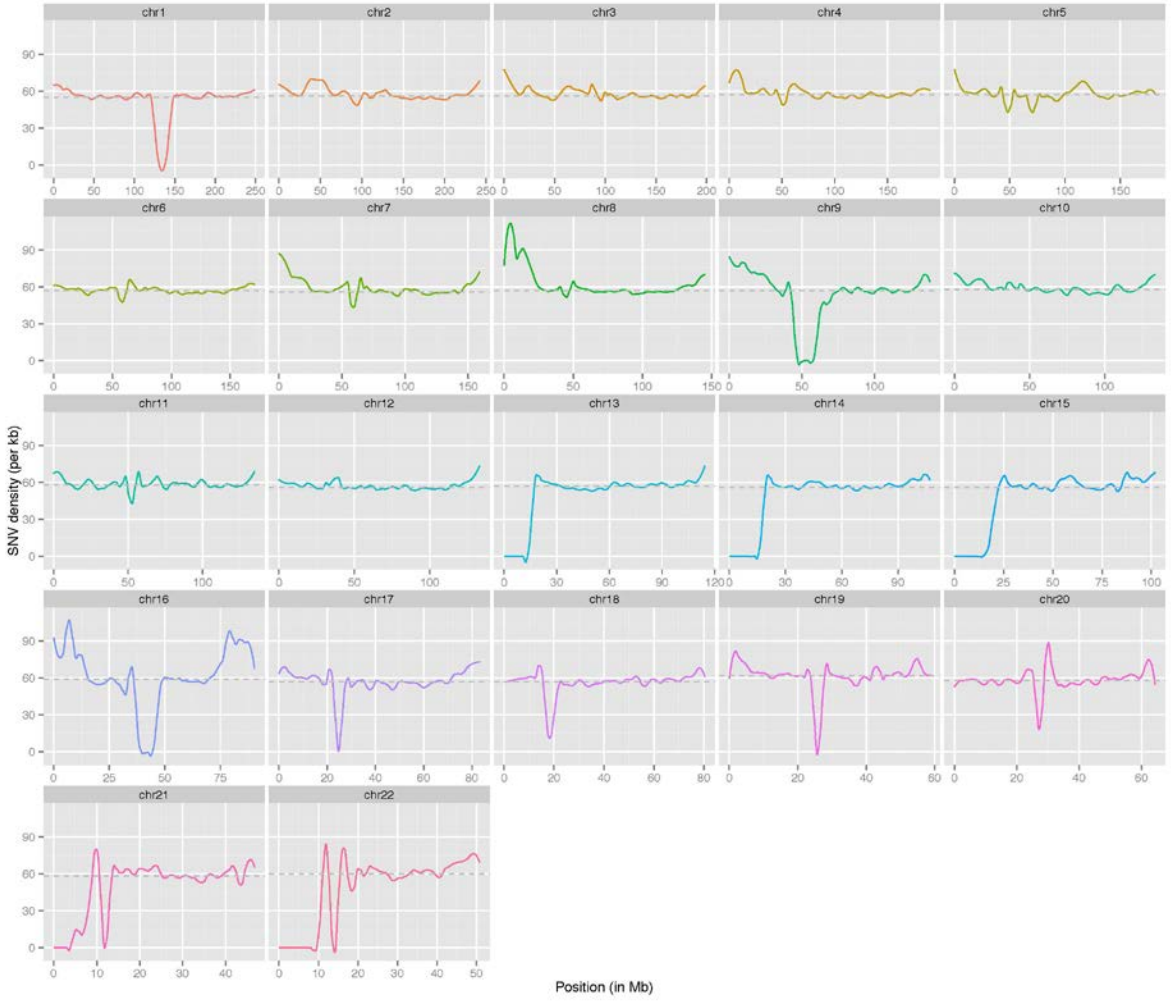
Suppl. Fig S4. Sequence reliability and rates of variation on the Y-chromosome.
Upper panel, Genome view of the Y-chromosome. Each dot represents a 1kb window; the Y axis represents the number of SNVs per 1kb; light blue are regions sequenced with high reproducibility; grey dots are regions sequenced with low reproducibility. **Lower panel**, number of females having variant at each position. We compared the ECR on Y chromosome to the list of euchromatin regions identified by Poznik et al.(34). Their list of euchromatin regions (converted from hg19 to hg38, 10Mb in total) contains 8.6Mb that are outside of segmental duplications. The ECR covers 89.7% (7.7Mb) of those regions.



Suppl. Fig S5. Single nucleotide variant distribution in the coding and non-coding genome. Distribution of SNVs in selected genomic elements (genomic, protein coding, RNA coding and regulatory elements. The distribution of each 1kb window from the total size concatenated element are represented in a violin plot. Elements are reported in the same order than in the main text Figure 2A, and the dot in the violin correspond to the SNVs per kb value computed from the total size concatenated element.



Suppl. Fig S6. Single nucleotide density across the chromosomes.
Smoothened SNV density across the genome (colored line). The line (dotted grey line) represents the median SNV density per 1kb window. Telomere regions are known for inflated rates of polymorphism. Deep valleys reflect regions that are not amenable to sequencing. Chromosomes 8 and 16 display reliable long regions of hypervariability that were the object of detailed analyses (Suppl. Fig. S7).

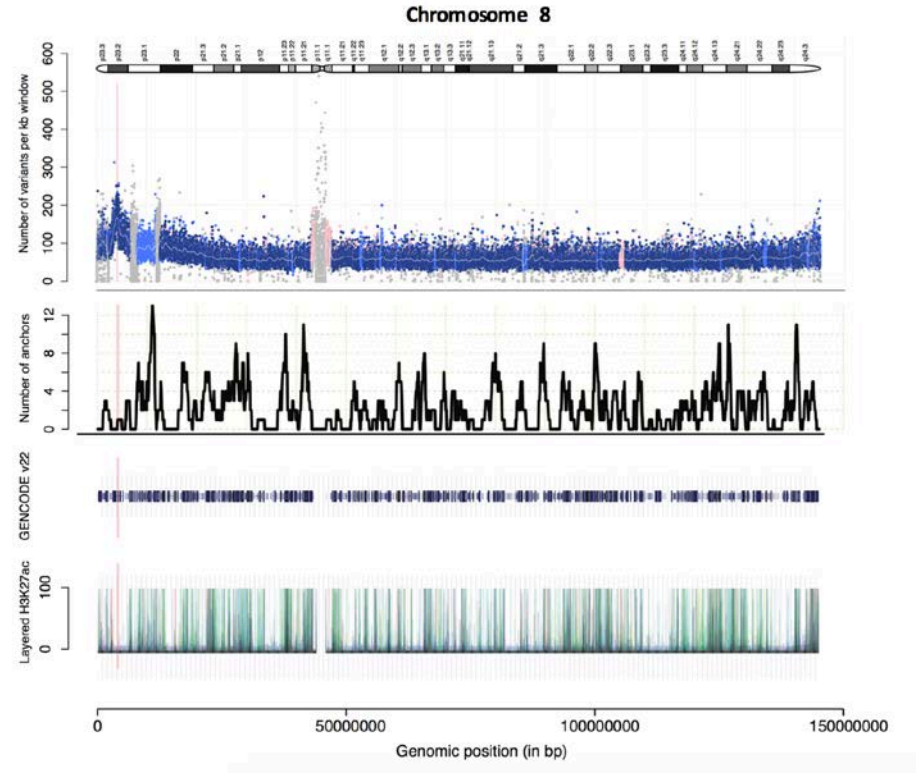


Suppl. Fig S7. Genetic hypervariability in regions lacking topological domains.
We identified 3 hypervariable megabase-long regions on autosomes based on the SNV density: one region on 8p23.2 spanning 1.3 Mb (**A**), and two regions spanning 1.45 Mb to 1.68 Mb on chromosome 16 (**B**). These hypervariable regions were identified based on the locally weighted LOESS fitting of the SNV density on each chromosome (coordinates in Table below). We used 4-Mb overlapping windows to identify a peak. A site would be regarded as peak if it has most SNVs within the 4Mb window (spanning 2Mb upstream and 2Mb downstream of the site of interest) and the number of SNVs is 3 standard deviations more than the mean of the autosomal SNV density (i.e. 98.15). We then defined the hypervariable regions by walking upstream and downstream away from the peak until the difference in number of SNV between adjacent sites was less than the median difference across the chromosome.
The density of multiple histone marks within these hypervariable regions, in particular the enhancer associated histone marks (H3K4me1, H3K4me2, H3K4me3, H3K27me3 and H3K27ac), are all depleted in the hypervariable regions (**C-E**). The coding/gene/exon densities were also significantly reduced in the identified hypervariable regions. The gene content of these regions is shown in the Table below. Both the exomic and the intronic regions present an elevated density of variants; therefore, the hypervariability is not just the reflection of a long stretch of gene-poor sequence.
We used the recently published chromatin loop anchors by Rao el 2014 *(35)* to determine the distribution of the SNVs across the genome topological domains. We observed depletion in anchors within the hypervariable regions, suggesting lack of chromatin loops in these regions. Because enhancers are typically located on the boundaries of loops, the depletion in loop anchors from the Hi-C data agrees with the observed decline in the enhancers from the ENCODE ChIP-seq data in the hypervariable regions.
Repeats, structural variations (including deletion, duplication and inversion) and segmental duplications are not enriched in these hypervariable regions based on data from RepeatMasker, Database of Genomic Variants and UCSC genome browser. We also confirmed that reads could be uniquely mapped to more than 99.8% of these regions based on simulation.

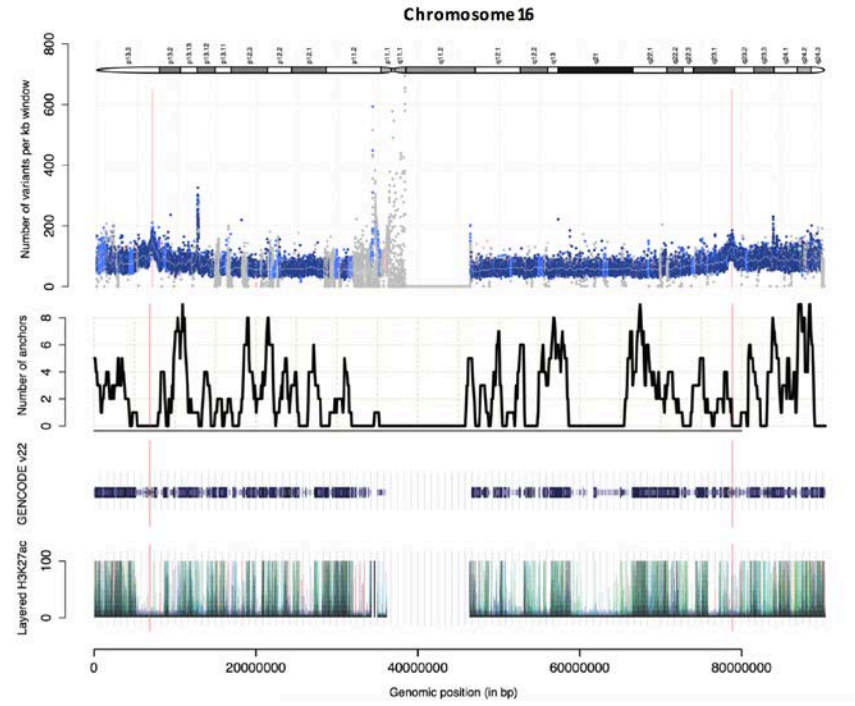
Coordinates (in hg38) and total length of uniquely mapped regions within the hypervariable regions

Chr	Start	End	Length	Total length of uniquely mapped region	Density per kb of exome content	Density per kb of intron content	Gene content
8	3,200,000	4,500,000	1,300,000	1,297,600 (99.82%)	94.50	153.19	*CSMD1*
16	6,400,000	7,850,000	1,450,000	1,448,967 (99.93%)	89.50	120.94	*RBFOX1*
16	77,900,000	79,583,500	1,683,500	1,682,850 (99.96%)	112.49	112.19	*WWOX, VAT1L, CLEC3A*

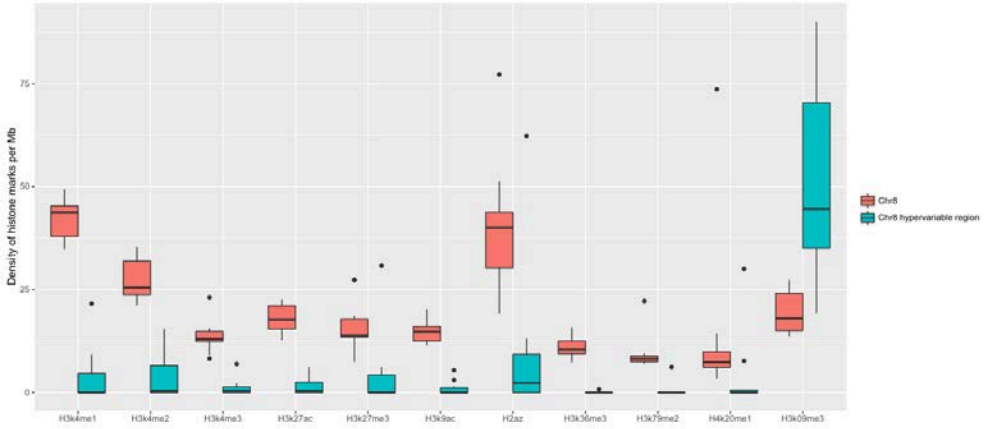
A


Chromosome 8
Number of variants per kb window
Number of anchors
GENCODE v22
Layered H3K27ac
0
50000000
100000000
150000000
Genomic position (in bp)

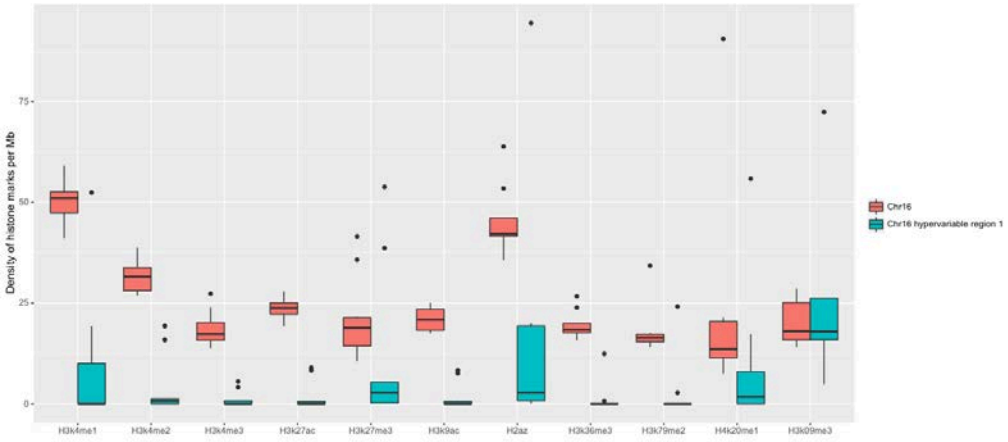
B


Chromosome 16
Number of variants per kb window
Number of anchors
GENCODE v22
Layered H3K27ac
0
20000000
40000000
60000000
80000000
Genomic position (in bp)

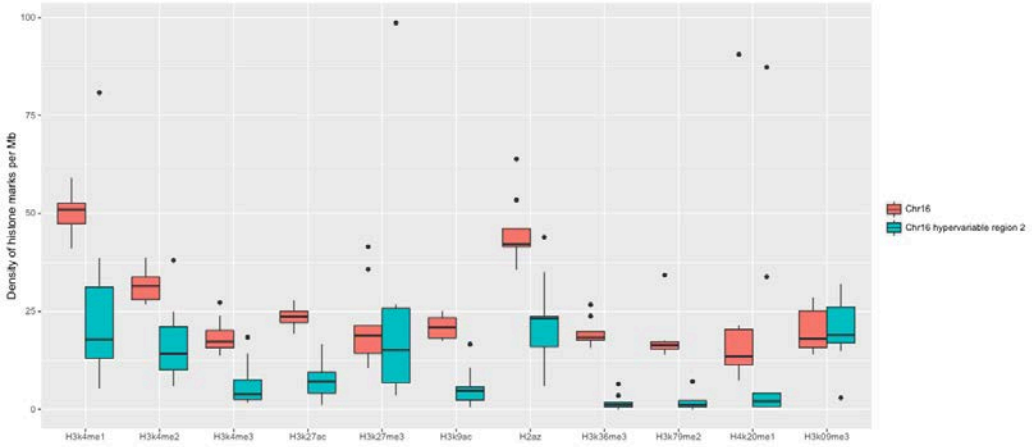
C (Chr. 8)



D (Chr. 16 First region)

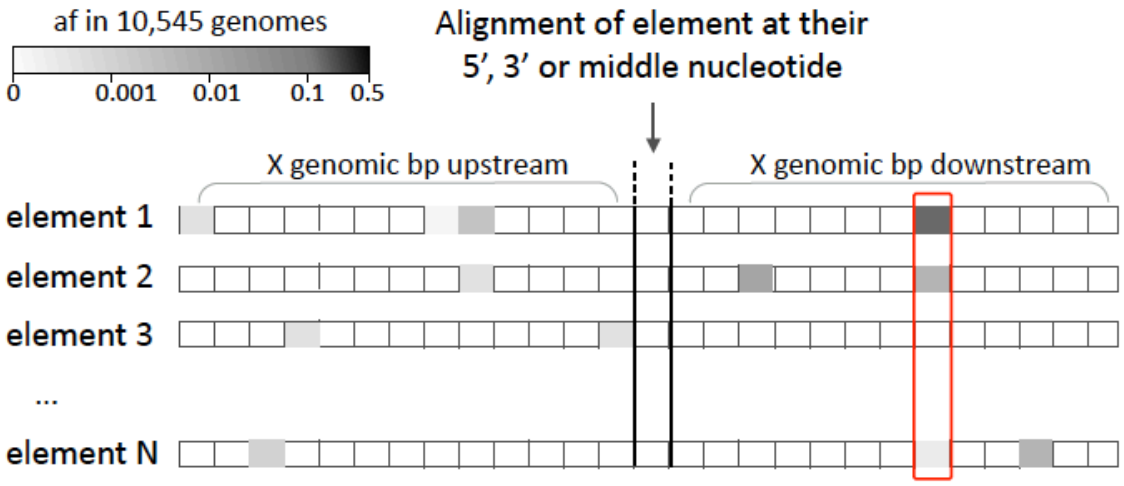


E (Chr. 16 Second region)



Suppl. Fig S8. Generation of metaprofiles

Elements sharing a common function, structure or sequence are aligned at their 5', 3' or middle nucleotide. The presence of variation and allelic frequency is recorded for each position surrounding the aligned position. Coloring of each square (position) reflects the allelic frequency of variation at that site. The matrix is analysed vertically: the count score is the fraction of positions with af > 0 (example in the red box : 3/N). The count score is further divided by the mean count score obtained across protein coding surroundings. The frequency score is the fraction of SNV with af > 0.001 (example in the red box : 2/3). The tolerance score is the product of both scores. Af, allelic frequency.



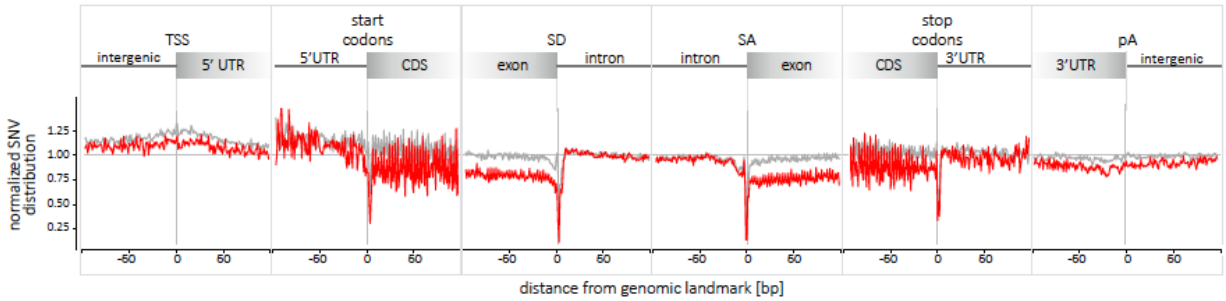
Metaprofile : vertical summary of each position at the elements

count* score:* fraction of nucleotides with af>0 (example in the red box: 3/N)

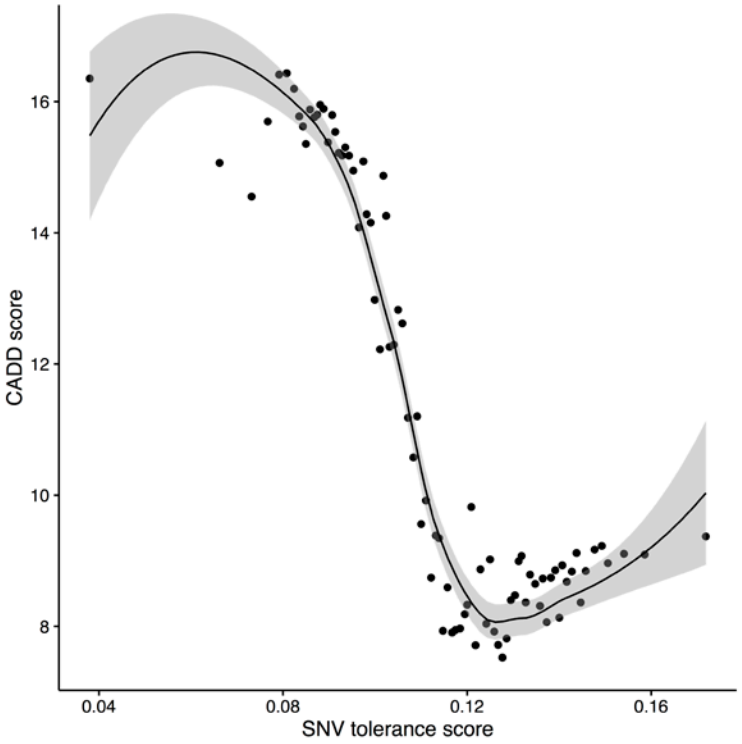
frequency* score:* fraction of SNV with af>0.001 (example in the red box: 2/3)

Tolerance* score *: product of count & frequency scores

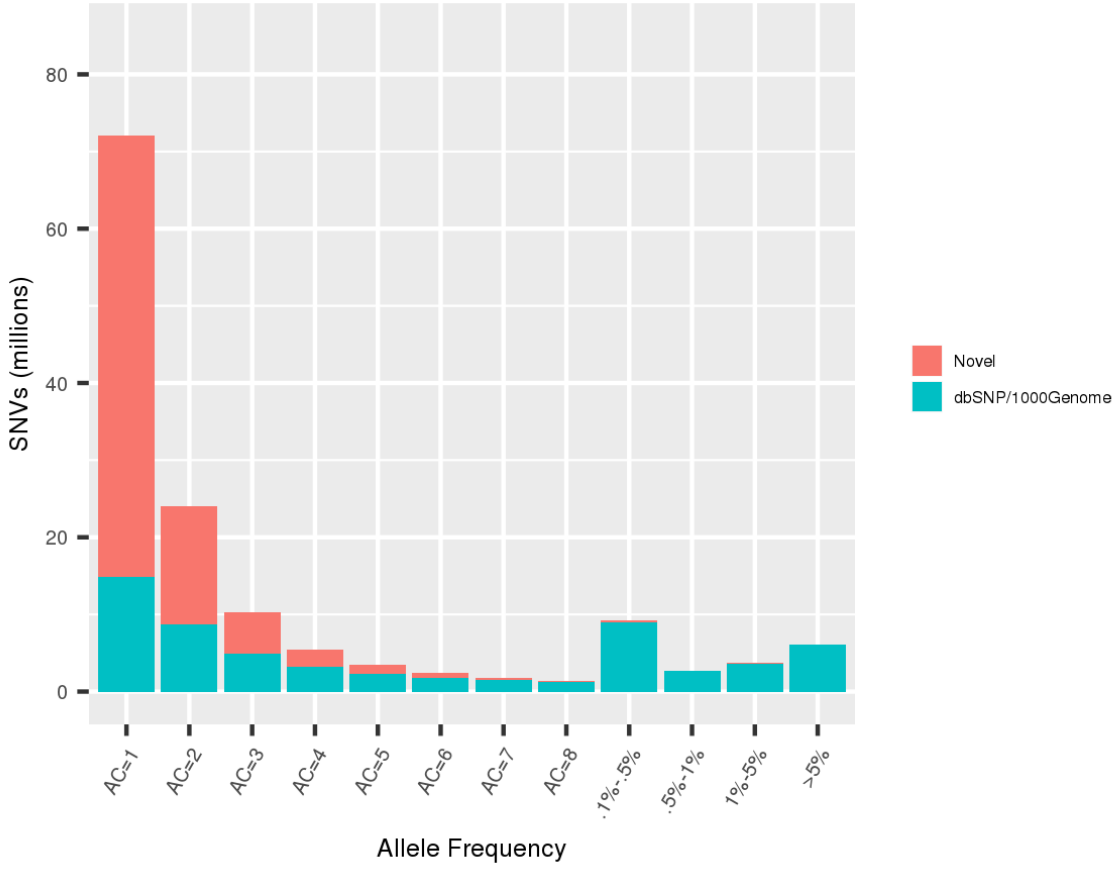
Suppl. Fig S9. Metaprofile of essential genes.
Essential genes, as defined by Bartha et al. (36), exhibit a different metaprofile pattern across the coding region. The figure depicts the transition between introns and exons. The y axis describes the enrichment/depletion of SNVs occurrence per position. In red, the metaprofile of essential genes (n=2,999, essentiality≥0.9); in grey, the metaprofile of the remaining genes with available score (n=13,163, essentiality <0.9). The x axis represents the distance from the genomic landmark.



Suppl. Fig S10. Relationship of a metaprofile tolerance score with CADD score.
Represented on X axis are the mean TS values for the coding region (+ 10 bp of intergenic or intronic boundaries), each dot represents the mean of 10 positions. The Y axis presents the mean CADD score for each bin. The LOESS curve fitting is represented by the solid line; the shaded area indicates the 95% confidence interval. CADD uses annotation from Ensembl Variant Effect Predictor, extensive information from UCSC genome browser tracks (GERP, phastCons, and phyloP; functional genomic data, transcript information and protein-level scores like Grantham, SIFT, and PolyPhen) to make functional predictions.



Suppl. Fig. S11. Distribution of allele frequencies for 150 million variants. Variants solely identified in the present study are shown in red. Variants that are also reported in dbSNP (version: human_9606_b144_GRCh38p2) and the latest (phase 3) 1000 Genome Project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/ALL.chr*.phase3_shapeit2_mvncall_integrated_v3plus_nounphased.rsID.genotypes.GRCh38_dbSNP_no_SVs.vcf.gz) are presented in blue. AC= allele counts. While 79.5% (57 of 72 million) of unique variants (AC=1) have not been reported in the past, only 0.99% (60,158 of 9.8 million) of common variants (allele frequency greater that 1%) are not represented in dbSNP and 1000 Genomes Project. This corresponds to a negligible average of 5.7 "novel" common variants per individual in the study.



Suppl. Fig S12. GC and dinucleotide content of unmapped reads.
The pattern of nucleotide-level usage is a taxonomic characteristic. GC and dinucleotide content was examined in the 4,876 unique human, or human-like contigs assembled from 2,435,202 bp of non-redundant sequence. The plots depict the distribution in GC content and dinucleotide bias of those sequences. The dark background represents the distribution of hg38-mapped sequences. The colored foreground density plot depicts unmapped contigs with confirmed human identity (green, 1,891,745 bp of non-redundant sequences mapped to known human sequences in GenBank), primate identity (light brown, 180,760 bp mapped to primate sequences in the NCBI), and the comparable distribution of 1,173,584 bp of contigs that do not have a known match in databases (purple, referred to as "human-like"). The figure plots the GC and dinucleotide pattern distribution for eukaryota, prokaryota and viruses as comparison.