Automatic detection of complex structural genome variation across world populations

Bo Zhou[1,2]*, Joseph G. Arthur[3,4]*, Hanmin Guo[1-3]*, Christopher R. Hughes[1,2*], Taeyoung Kim[5], Yiling Huang[1,2], Reenal Pattni[1,2], HoJoon Lee[6], Hanlee P. Ji[6], Giltae Song[5,7], Dean Palejev[8], Xiang Zhu[3,9,10], Wing H. Wong[3,11#], Alexander E. Urban[1,2#]

[1]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California 94305, USA
[2]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA
[3]Department of Statistics, Stanford University, Stanford, California 94305, USA
[4]Current affiliation: 10X Genomics, Pleasanton, California 94566, USA
[5]School of Computer Science and Engineering, College of Information and Biomedical Engineering, Pusan National University, Busan 46241, South Korea
[6]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, California 94305, USA
[7]Center for Artificial Intelligence Research, Pusan National University, Busan 46241, South Korea
[8]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia 1113, Bulgaria
[9]Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, USA
[10]Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, Pennsylvania 16802, USA
[11]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, California 94305, USA

* These authors contributed equally.

 # Corresponding authors:

Alexander E. Urban, Ph.D
aeurban@stanford.edu

Wing H. Wong, Ph.D
whwong@stanford.edu

Running title: Complex structural variation in human genomes

Keywords: structural variation (SV), genome associations

ABSTRACT

Complex structural variants (cxSVs), e.g. inversions with flanking deletions or interspersed inverted duplications, are part of human genetic diversity but their characteristics are not well delineated. Because their structures are difficult to resolve, cxSVs have been largely excluded from genome analysis and population-scale association studies. To permit large-scale detection of cxSVs from paired-end whole-genome sequencing, we developed Automated Reconstruction of Complex Variants (ARC-SV) using a novel probabilistic algorithm and a machine learning approach that leverages the new Human Pangenome Reference Consortium diploid assemblies. Using ARC-SV, we resolved, across 4,262 human genomes spanning all continental super-populations, 8,493 cxSVs belonging to 12 subclasses. Some cxSVs with population-specific signatures are shared with Neanderthals. Overall cxSVs are significantly enriched in regions prone to recombination and germline *de novo* mutations. Many cxSVs mark phenotypic hotspots (each significantly associated with ≥ 20 traits) identified in genome-wide association studies (GWAS), and 46.4% of all significant GWAS-SNPs catalogued to date reside within ±125 kb of at least one cxSV locus. Common SNPs near cxSVs show significant trait heritability enrichment. Genomic regions affected by cxSVs are enriched for bivalent chromatin states. Rare cxSVs are enriched in neural genes and loci undergoing rapid or accelerated evolution and recently evolved *cis*-regulatory regions for human corticogenesis. We also identified 41 fixed loci where divergence from our most recent common ancestor is via localized cxSV. Our method and analysis framework allow for the accurate, efficient, and automatic identification of cxSVs for future population-scale studies of human disease and genome biology.

INTRODUCTION

Structural variation (SV) analysis though whole-genome sequencing (WGS) is essential to the study of genome biology and disease etiology (Lappalainen et al. 2019). SVs rearrange genome sequences ≥ 50 bp and, though individually rarer than single-nucleotide variants (SNVs) and indels (The 1000 Genomes Project Consortium et al. 2012), account for a much greater portion of the DNA sequence variation between individuals (Pang et al. 2010). Due to their wide size spectrum and complexity, SVs can affect genome function through numerous mechanisms (Stankiewicz and Lupski 2010) and inordinately (by effect size) impact heritable gene expression and disease phenotypes (Stranger et al. 2007; Weischenfeldt et al. 2013; Chiang et al. 2017).

While SV detection methods based on long-read sequencing technologies are rapidly developing (Logsdon et al. 2020), large cohort and population-scale studies (Wonkam 2021; Gaziano et al. 2016; Saunders et al. 2019) continue to rely on paired-end short-read sequencing as it remains the most mature, well-supported, standardized, and cost-effective way to gather genome information (Abel et al. 2020; Collins et al. 2020). To date, standard genome SV analysis focuses almost exclusively on simple types of SVs identified by a single breakpoint, *i.e.*, by a pair of genomic locations adjacent in the sample but not the reference, such as deletions, tandem duplications, inversions, and/or translocations (Guan and Sung 2016; Kosugi et al. 2019). Detection of these simple-SV breakpoints is theoretically straightforward but difficult in practice (Escaramís et al. 2015; Cameron et al. 2019), and investigators typically apply multiple methods together with heuristic filters (Sudmant et al. 2015; Parikh et al. 2016; Mohiyuddin et al. 2015; Chaisson et al. 2019; Zhou et al. 2018b). However, SVs with complexity beyond the scope of most detection algorithms are abundant in human genomes and have been observed in a variety of phenotypic contexts including healthy, somatic, and cancer tissue (Quinlan and Hall 2012; Sudmant et al. 2015; Collins et al. 2017; Sekar et al. 2020; Collins et al. 2020; Byrska-Bishop et al. 2022; Fujimoto et al. 2021).

Our definition of a complex SV (cxSV) is any genome rearrangement that involves multiple breakpoints not reducible to non-overlapping deletions, tandem duplications, novel insertions, or inversions. More specifically, these cxSVs are localized, with all rearranged reference segments originating from a small to medium-sized (< 2 Mb) genomic region (Table S1, Figure 1, Supplementary Figure S1), unlike chromosomal-scale rearrangements or chromothripsis (Stephens et al. 2011; Forment et al. 2012). Although, complex SVs have been shown to cause Mendelian disorders and associate with other genomic disorders (Sanchis-Juan et al. 2018; Carvalho and Lupski 2016), their roles in genome biology, disease, and evolution are still very poorly understood. This is primarily due to the technological challenges in automatically resolving their sequence structures from standard WGS data where cxSVs are either inadequately represented by a single SV or incorrectly detected as multiple overlapping SVs (Sudmant et al. 2015) (Supplementary Figure S2). While cxSVs have been studied by focused breakpoint analysis with semi-manual inspection of paired-end mapping patterns (Yalcin et al. 2012; Sekar et al. 2020; Collins et al. 2020) and long sequencing reads and mate-pairs (Sudmant et al. 2015; Collins et al. 2017; Frith and Khan 2018; Cretu Stancu et al. 2017; Sedlazeck et al. 2018; Zhou et al. 2019; Lin et al. 2022), computationally efficient and accurate methods for their automatic detection and assembly from standard WGS data that are especially suitable for large-cohort or population-scale datasets have been challenging to develop (Collins et al. 2020; Abel et al. 2020; Byrska-Bishop et al. 2022)

Here, we have developed the novel method, Automated Reconstruction of Complex SVs (ARC-SV), which uses a generative statistical model and a machine learning framework based on diploid genome assemblies (Liao et al. 2022) to accurately detect and automatically sequence assemble an unprecedented spectrum of localized complex structural rearrangements from standard short-read, paired-end WGS data (Figure 1, Table S1, Supplementary Methods). This data-driven, generative approach eliminates any *a priori* assumptions of the types of complex SV structures to be observed in the sample genome. Our

method efficiently integrates the automatic detection and assembly of cxSVs to standard WGS analysis. To catalogue and to decipher the characteristics of complex structural variation in natural human genetic diversity, we then applied ARC-SV to 4,262 human genomes that span across all 8 continental super-populations across the world (Table 1). We then investigated the characteristics of their distribution across the genome, their proximity to GWAS signatures, the chromatin states of affected regions across various tissues, their enrichment in genes and regulatory regions, and their occurrence in genome regions under rapid or accelerated evolution including those that diverged from our closest primate relative via cxSV, as well as cxSVs shared between modern and archaic humans.

RESULTS

ARC-SV detection of cxSVs using a probabilistic and machine learning framework

At a given stretch of the genome, there are in principle numerous possible cxSV structures (Quinlan and Hall 2012), each producing its own unique signatures of read depth, mapped read orientations, insert sizes, and split alignments (Sekar et al. 2020). The observed signature will depend on sequencing library characteristics (especially the insert size distribution) in addition to the structure and size of the SV. Several previous methods (Collins et al. 2017, 2020; Byrska-Bishop et al. 2022) attempt to classify cxSV structures by direct comparison to these mapping signatures. Here, to achieve more general detection of cxSVs, ARC-SV simultaneously detects and genotypes SVs by integrating all reads into a probabilistic model (including reads that match the reference) which can be used to score proposed candidate diploid rearrangements (Figure 1a) in contrast to other SV analysis methods that only interrogate breakpoint-supporting reads (Collins et al. 2020; Byrska-Bishop et al. 2022) (see Supplementary Methods for detailed description of probabilistic model). Briefly, given a standard paired-end read alignment dataset, candidate breakpoints are first determined from soft-clipped reads, split-read alignments, and clustering of discordant read pairs (determined from library insert-size distribution and coverage that are unique to each sequencing library)

(Supplementary Figure S3). Breakpoint intervals determined from discordant read pair clusters are refined based on a probabilistic model and merged with those determined from soft-clipped/split reads. Based on breakpoint locations, ARC-SV then defines blocks of genome segments that are rearranged in the sample genome and constructs adjacency graphs that represent reference or novel (non-reference) adjacencies between genomic segments. Candidate haplotype structures are then proposed by enumerating paths through the adjacency graph and evaluated based on paired-end genome alignment likelihood modeling (Figure 1a, Supplementary Methods). To make complex SVs structures more interpretable, we devised our own convention of describing all cxSVs using an alphabetical notation (e.g. ABCDE → ABC'B'E where B segment is duplicated in an inverted manner between C and E, C segment is inverted, and D segment is deleted (Figure 1, Table S1). In this process of detecting cxSVs, ARC-SV also automatically detects, genotypes, and sequence assembles canonical simple SVs, e.g. deletions, inversions, tandem duplications (Table S1).

To further increase the precision and accuracy of the SV calls, i.e. to identify high-confidence calls from the ARC-SV probabilistic model, we integrated a machine learning framework trained to distinguish SV calls that validate against established "ground truths" (Figure 1b-c). Here, for the "ground truth" set, we leverage 42 new state-of-the-art diploid assemblies of diverse individuals released by the Human Pangenome Reference Consortium (HPRC) (Liao et al. 2022) (Table S2). For training and testing of our machine learning models, the 42 diploid assemblies are split into training (n=31) and test (n=11) sets stratified by genome ancestry (Table S2).

For all SVs, and perhaps especially for cxSVs, it is essential that calls are validated independently. Validation of simple SV calls is traditionally conducted by comparison to an existing gold set of predictions (Zook et al. 2020) where a SV prediction is typically said to be validated if it has high overlap to one of the "true" SVs. The commonly-used criterion is that the overlap between the predicted and true SVs is at least 50% of the larger one's size — this is

referred to as 50% reciprocal overlap (Rausch et al. 2012; Hormozdiari et al. 2009; Zhou et al. 2018b). Applying the same kind of methodology to cxSV predictions is infeasible since gold-standard cxSV call sets are not available. Here, we validate SV calls from an individual by direct sequence match to the individual's diploid genome assembly ("ground truth") via first sequence assembly of detected SVs followed by sequence alignment (Figure 1b-c).

ARC-SV automatically constructs a sequence assembly for each SV call with 1kb of flanking sequences on each side. For genomes (n=42) with HPRC assemblies (Table S2), SVs called from standard paired-end WGS of the same genomes are sequence assembled and aligned to the HPRC genome assemblies. Before alignment, predicted SV sequences are modified such that only 150 bp of flanking sequences are included. This is so that flanking sequences serve only the role of anchor and have minimal effect on the overall alignment score of the prediction sequence. Using an established validation criteria of cxSV detection using long-reads (Sedlazeck et al. 2018), a cxSV call is validated if its sequence prediction of rearranged genome segments matches the "ground truth" via complete and continuous sequence alignment (Supplementary Methods). See Figure 2 for examples. Compared to commonly used validation methods such as *in silico* simulation and overlap based on SV call set and/or alignment spans, our validation method, although significantly more stringent, is better suited to train accurate machine learning models. ARC-SV cxSV calls from each HPRC genome (Table S2) based on ~30×-coverage paired-end WGS (Byrska-Bishop et al. 2022) were validated against the diploid genome assemblies and labeled based on validation outcome (Table S3). Of the initial ARC-SV cxSV calls prior to machine learning, 54% (n=4,221) of total cxSV calls across the HPRC genomes were validated and 46% (n=3,665) were not (Table S3).

Conventionally, high-confidence SV calls are typically filtered based on metrics such as the number of supporting reads or the depth of coverage in surrounding regions (Layer et al. 2014; Rausch et al. 2012; Chen et al. 2016). While this approach is interpretable and easy to communicate, it does not make full use of the data and requires the user to manually specify

thresholds, often without any recommendations. Our alignment-based validation approach (Sedlazeck et al. 2018) provides the opportunity for a more careful study of the accuracy of SV prediction.

Our machine learning model predicts SV validation given 48 features (Table S4) transformed from the default output table of the ARC-SV probabilistic model (Table S3, Supplementary Methods). From these transformed features, we then trained random forest, extreme gradient boosting, and light gradient boosting models (Figure 1b-e, Source Data). Maximizing for precision, we trained a separate model for each SV class (cxSVs, deletions, tandem duplication, and inversion). The optimal machine learning model for predicting high-confidence cxSVs is an extreme gradient boosting (XGBoost) model achieving 95.7% precision and 91.4% recall on the test set (Table S5, Figure 1e). Machine learning models for simple SVs were also trained using the same alignment-based validation approach against HPRC diploid assemblies: deletions (XGBoost, 99% precision, 95.8% recall), tandem duplications (XGBoost, 97.8% precision, 90.9% recall), and inversions (random forest, 96.2% precision, 96.2% recall) (Tables S4-5).

Additional validation of ARC-SV detection of cxSVs

To further validate our method, we also performed alignment-based validation of high-confidence cxSV calls against other recently published diploid genome assemblies as “ground truths”; these were constructed using different methods from the HPRC assemblies (Liao et al. 2022). We validated our high-confidence cxSV calls (identified via machine learning model) for NA12878 and PGP1 (both of European ancestry) against their respective diploid genomes assembled using DipAsm (Garg et al. 2021) as well as those for HG00514 (East Asian ancestry) and HG03125 (African ancestry) against genomes assembled using a combination of PacBio HiFi and Strand-Seq (Ebert et al. 2021). We observed ≥ 92.5% F1 score across all four genome validations (Table S6, Figure 1f).

Furthermore, we also validated our method for cxSV discovery in somatic tissues. It was demonstrated previously that mosaic complex SVs can arise in the developing human brain likely during cell proliferation in mid-neurogenesis (Sekar et al. 2020). Clonal populations (n=41) were obtained from single neural progenitor cells from the cerebral cortex (FR), parietal cortex (PA) and basal ganglia (BG) regions of three phenotypically normal postmortem human fetuses (Bae et al. 2018); each clone was sequenced via standard WGS to 30× coverage, and through careful manual breakpoint analysis and local assembly, 3 localized intragenic cxSVs (2 of which are within neurodevelopmental genes *SPAST* and *NAV3*) and one inter-chromosomal cxSV were experimentally validated to be mosaic, present in 4 different clones belonging to 4 different brain regions (Sekar et al. 2020). We applied ARC-SV to all 41 clones for mosaic cxSV discovery. We then compared raw cxSV calls across different clones to identify those that may be somatic or those that occur only in subsets of clones within the same tissue of origin and obtained 42 candidates for putative mosaic cxSVs (Table S7) where we then applied our trained machine learning model to filter for high-confidence cxSVs. Our model retained only the 3 localized intragenic cxSVs and in the exact same clones that were experimentally validated (Sekar et al. 2020).

In addition, we also used PCR followed by Sanger sequencing to experimentally validate cxSV calls (n=23) from two pilot genomes that were originally used for methods development: HuRef (Zhou et al. 2018a) and HepG2 (Zhou et al. 2019) (Table S8). Primers were designed to amplify the cxSV in its entirety i.e. all breakpoints in a single reaction (Table S9). For cxSVs called as heterozygous, we ensured that reference alleles (without SV) were also amplified to validate genotype. Sanger sequencing from both forward and reverse primers was performed. We observed validation rates of 90% (9/10) for HuRef and 85% (11/13) and for HepG2 (Table S9).

Comparison against other short or long read-based cxSV detection methods

Recently, two studies on human structural variation involving the Human Structural Variation Consortium and the Genome Aggregate Database Consortium were published that incorporated detection of complex SVs from standard WGS at a population-scale (Collins et al. 2020; Byrska-Bishop et al. 2022). Both studies used the same core method developed in Collins *et al* (Collins et al. 2020) that incorporates an ensemble of different tools for breakpoint analysis to enable cxSV detection. Here, since we analyzed the same 1000 Genomes Project (1000GP) samples (n=3,202) using the same GRCh38 alignment data as Byrska-Bishop et al (Byrska-Bishop et al. 2022), we benchmarked our cxSV calls against those made by Byrska-Bishop *et al* (Byrska-Bishop et al. 2022) for genomes (n=40) with "ground truths" assemblies (Liao et al. 2022) (Table S2). Following our alignment-based validation strategy, we first sequence assembled the cxSV calls from Byrska-Bishop *et al* (Byrska-Bishop et al. 2022) with 150 bp of reference sequence flanking each side (Source Data) and then aligned these sequence of assembled cxSVs directly against their respective assembled genomes (ground truths) for validation (Supplementary Method). Since we used 31 of these genomes (Table S2) for training our machine learning models (Figure 1b-c,e-f, Table S5), we first benchmarked raw or naïve ARC-SV calls (without machine learning) for this set of genomes where we observed 188 ± 17 (mean, standard deviation) cxSV calls per genome with 53% ± 2.8% validation (Table S3), compared to 135 ± 12 cxSV calls per genome with 22% ± 2.6% validation for Byrska-Bishop *et al* (Byrska-Bishop et al. 2022) (Table S10, Figure S4). For the other 9 benchmarking genomes (Table S2), we observed 99 ± 10 cxSV calls per genome with 95% ± 0.7% validation rate for ARC-SV (with machine learning) (Table S3), compared to 131 ± 9 cxSVs calls per genome with 28% ± 5.6% validation for calls from Byrska-Bishop *et al* (Byrska-Bishop et al. 2022) (Table S10, Figure S5). Of the validated cxSV calls from Byrska-Bishop *et al* across these 40 genomes, 81% ± 6.3% were also called by ARC-SV, and of these, the ARC-SV call was not validated in 3 ± 1 cases per genome (Table S10).

In theory, long reads are better suited for resolving cxSVs as they are more likely to span and phase consecutive breakpoints. We thus compared the performance of ARC-SV to that of a novel cxSV detection method based on deep learning of long-read alignments, SVision (Lin et al. 2022). We first checked whether ARC-SV can resolve the 18 (from NA12878) manually curated cxSVs of six unique structures that was used as the benchmark where all 18 were resolved by SVision (Lin et al. 2022). ARC-SV was also able to resolve 17/18 (94%) cxSVs including all internal breakpoints (Table S11). In the Lin *et al* study, 80 high-confidence cxSV calls from HG00733 were made by SVision for which 20 (55%) were selected for experimental validation where 11 validated (Lin et al. 2022). ARC-SV resolved all 11 experimentally validated events (Table S12) and of the remaining 60 SVision calls, ARC-SV resolved 31/60 (52%) in line with the experimental validation rate for Svision (Table S13). We then sequence assembled the remaining 29/60 Svision calls from HG00733 that were not called by ARC-SV and benchmarked them against the HG00733 diploid assembly (Liao et al. 2022) where 1/29 validated (Table S13), further confirming a validation rate of ~55% for high-confidence Svision calls. In comparison, ARC-SV made 94 high-confidence cxSVs for HG00733 with 94% validation (Table S3).

Complex SVs across diverse human populations

After having extensively validated ARC-SV to decipher and automatically resolve cxSVs from standard paired-end WGS data, we applied our method to 4,262 healthy human genomes from a total 134 diverse populations across all 8 continental super-populations to investigate the diversity of complex SVs resulting from natural human genetic variation (Table 1, Figure 3a, Table S14). These genomes are compiled from the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2010; Byrska-Bishop et al. 2022), the Human Genome Diversity Project (Bergström et al. 2020), and the Simons Genome Diversity Project (Mallick et al. 2016). Across all genomes, we discovered 8,493 unique cxSVs that span up to 482 kb of the genome with the vast majority spanning between 100 bp to 10 kb (Figure 3b, Table S15). To

the best of our knowledge, this is the largest and most comprehensive catalogue of human complex SVs to date. Around 75% of the discovered cxSVs can be categorized into 11 distinct subclasses (inversion deletion, duplication-deletion, deletion-flanked inversions, etc); others are too complex to be verbally described and thus categorized as “Other” (Figure 1d, 3b,c, Table S1,S15).

Approximately 16.2% (n=1,377) of these unique cxSVs have an allele frequency ≥ 0.1% and thus considered common, and 7,116 are below 0.1% allelic frequency and considered rare; 4,661 (54.9%) of cxSVs are private to single individuals (Figure 3d, Table S15). Overall, the allele frequencies of cxSVs are lower than that of simple SVs in the human population (one-sided Mann-Whitney test, p-value = 4.2e-14, Figure S6). At an individual level, individuals of AFR descent have on average the highest number (n = 101) of cxSVs compared to other super-populations (Welch’s one-sided two sample t-test, p = 2.9e-305) (Figure 3e). Overall, the majority of unique cxSVs are rare at a population level (Figure 2D), while within any given individual, >90% of the cxSVs will be common events with ~5-10 private cxSV events with a typical span size of ~6.2 kb (Table S15). 45.2% of cxSVs overlap protein coding genes, and 66.5% overlapping putative ENCODE regulatory regions (The ENCODE Project Consortium 2012). We also find that 69.4% (956/1377) of common cxSVs (AF ≥ 0.1%) fall within previously determined structural polymorphic genomic regions determined based on the 47 HPRC genomes (Lee et al. 2022).

The majority of common cxSVs show strong signatures of population specificity (Figure 3f). For example, an interspersed inverted duplication with deletion cxSV (AD’CDE) on chromosome 4 (validated in maternal assembly of HG02818, Gambian female, Figure 2a) is common in AFR, AMR, and MDE but rare elsewhere (Table S15). This cxSV resides within *MTHFD2L*, which is associated with cytokine secretion in response to viral infections and vaccines (Kennedy et al. 2012). This cxSV deletes 2 ENCODE candidate *cis*-regulatory elements (cCREs) (E2303693 and E2303696) with distal enhancer-like signatures (The

ENCODE Project Consortium 2012) and enhancer (GH04J074217) (Fishilevich et al. 2017) which are bound by key developmental transcription factors such as IKZF1, FOXA1 and FOXA2 (Hammal et al. 2022). Another example is a cxSV (ABCBDF) on chromosome 5 (validated in the maternal assembly of HG01952, Peruvian male, Figure 2b) within *LINC01933* where the allele frequency (AF) is ~2% in AFR and EAS populations but at 9.5-18% in other populations (Table S15). The deletion of "Block E" within this cxSVs deletes a cCRE (E2422272) with distal enhancer-like signature (The ENCODE Project Consortium 2012) that is bound by ESRRA which interacts with PGC-1 transcription cofactors to regulate the expression of most genes involved in cellular energy production and biogenesis (Schreiber et al. 2004; Hammal et al. 2022). Furthermore, a cxSV (ACE'G) on chromosome 14 (validated in the paternal assembly of HG02723, Gambian female, Figure 2c, Table S15) is found only in AFR (AF=1.7%) and AMR (AF=0.09%) populations and resides within *PRKCH* which encodes a transcription factor predominantly expressed in epithelial tissues and can regulate keratinocyte differentiation (Cabodi et al. 2000). The deletion of Blocks B (12.4kb) deletes 7 cCREs, inversion of Block C affects 3 cCREs, and deletions of Blocks DEF partially deletes cCRE E1719464 (The ENCODE Project Consortium 2012) and enhancer (GH14J061376) (Fishilevich et al. 2017) which are bound by WT1 and IRF4 (Hammal et al. 2022), a key regulatory factor of in the development of human immune cells (Nam and Lim 2016). WT1 exhibit complex tissue-specific and polymorphic imprinting pattern, with biallelic, and monoallelic expression from the maternal and paternal alleles in different tissues (Jinno et al. 1994). Lastly, a cxSV (AC'E) on chromosome 12 (validated in the paternal assembly of HG03540, Figure 2d) within *SLC38A4*, a tumor suppressor gene (Li et al. 2021), is common in all populations except EUR with the highest AFs in OCN, EAS, AFR, and CAS populations.

For common cxSVs in general, principal component analysis (PCA) shows distinct clusters for diverse continental super-populations and that the distance and spatial relationship between the clusters reflect and correlate with geographical distances (Figure 3g). Batch effect

between source datasets is minimal (Supplementary Figure S7). Compared to other super-populations, AFR has the most population-specific cxSVs, which is consistent with AFR having the highest degree of genetic diversity (Figure 3e-g, Table S15). Moreover, we investigated the gene associations of cxSV regions with top loadings in the top two principal components that distinguish different populations (Supplementary Figure S8). We find that genes located near the top cxSVs contributing to population-specific signatures are enriched for those linked to the distribution of body fat and blood osmolarity (Supplementary Figure S8).

Enrichment analysis of cxSVs in recombination initiation and DNM hotspots

In addition to world population, we examined how cxSVs are distributed across the genome (Figure 4a). We observed a tendency for cxSVs to be clustered toward chromosome ends (Figure 4a) where these regions have been associated with elevated local mutation rates (Nesta et al. 2021; Audano et al. 2019) and higher recombination rates (Jensen-Seaman et al. 2004). The formation of structural variation in general has often been attributed to the improper repair of double-strand DNA breaks (Carvalho and Lupski 2016). To investigate whether cxSVs are enriched in recombination hotspots, we performed permutation test and observed singificant overlap between cxSVs and genome recombination initiation double-stranded break hotspots (Pratto et al. 2014) (enrichment = 1.33, $p < 1.0e-5$) (Figure 4a).

To test whether cxSVs are enriched in regions with elevated mutation rates, we performed permutation test on the overlap between all detected cxSVs (Table S15) and germline *de novo* mutations (DNMs) from healthy controls annotated in the Gene4Denovo database (Zhao et al. 2020). As DNMs are only present in the offspring and absent in the parents, they occur within a single generation and thus are subjected to minimal selective pressure, their distribution may largely reflect the germline mutation rate (Murat et al. 2022). We observed significant overlap (enrichment = 1.11, $p = 2.0e-5$) suggesting that cxSVs have a tendency to occur in regions with elevated mutation rates (Figure 4a).

cxSVs near GWAS SNPs

To gain insight into how cxSVs may affect complex traits and diseases, we integrated cxSVs with GWAS. Most GWAS to date focus only on the associations of common SNPs. To assess the utility of cxSVs in GWAS, we first mapped common SNPs to cxSVs using the gnomAD-SV linkage disequilibrium (LD) map which contains SNVs in strong LD with gnomAD-SVs (R2 ≥ 0.8) (Collins et al. 2020) (Supplementary Figure S9). In the gnomAD-SV LD map, there are 564,676 unique pairs of gnomAD-SVs and SNVs; 506,511 (89.7%) of which have distances less than 125 kb. For AFR individuals, 215,240 (90.5%) pairs are within 125 kb; for EUR individuals, 291,271 (89.1%) pairs are within 125 kb. Informed by theses distributions of LD-distances, we mapped a common SNP in the GWAS Catalog (p-value ≤ 5e-8) (Sollis et al. 2023) to a cxSV if their distance is within 125 kb (Methods).

Among the 40,377 unique SNPs in the GWAS Catalog (henceforth "GWAS SNPs"), 18,732 (46.4%) are mapped to at least one cxSV (Table S16). For the 7,194 unique cxSV intervals (Table S15), 4,121 (57.3%) of which contain at least one mapped GWAS SNP (within ± 125 kb). Among the 2,811 unique GWAS traits, 2,025 (72.0%) have at least one SNP mapped to a cxSV (Figure 4b, Table S16). For these 2,025 traits, 1,572 (77.6%) have higher fractions of GWAS SNPs mapped to cxSVs (Figure 4b) where 29 of show significance after Bonferroni correction (one-sided Binomial p-value ≤ 0.05/2025). Many cxSVs are mapped to GWAS SNPs associated with multiple traits (median: 3; Figure 4d, Table S16). We define a cxSV as a "phenotypic hotspot" if it is mapped to GWAS SNPs associated with ≥ 20 unique traits (Figure 4d), corresponding to the top 5% of trait counts (Table S16). For example the cxSV (AC'E) on chr19:5831855-5859707 is associated with 20 unique GWAS traits (Table S16, Figure 4d) such as cancer biomarkers, heart rate variability, and N-glycan, which is related to brain development (Medina-Cano et al. 2018). The "hottest" phenotypic hotspot contains GWAS SNPs associated with 318 unique traits (Table S16, Figure 4d). The number of unique traits associated with such mapped GWAS SNPs is uncorrelated with the size of cxSV (Pearson R: 0.004; p-value: 0.75).

Altogether, our results show that cxSVs are mapped an excess of GWAS SNPs than expected by chance.

In addition to mapping GWAS SNPs to cxSVs, we also assessed if common SNPs (MAF ≥ 0.05) mapped to cxSVs could explain a much larger proportion of complex trait heritability than one would expect by the number of these SNPs (i.e., heritability enrichment (Finucane et al. 2015)) across 1,087 GWAS datasets (Zhu et al. 2022). While significant heritability enrichment (estimate: -0.006; p-value: 0.99) was not observed for meta-analysis estimates across these datasets, we obtained significant heritability enrichments in GWAS of spondylosis (estimate: 0.504; p-value: 4.31e-05) and spondylopathy (estimate: 0.296; p-value: 4.56e-05) that passed the stringent Bonferroni correction (p-value ≤ 0.05/1087), as well as enrichments in GWAS of major depressive disorder, sedentary behavior, dorsopathy, short sleep, dietary pattern, and Crohn's disease at the nominal significance level (p-value ≤ 0.01; Figure 4c, Supplementary Table S17).

cxSVs and chromatin states

Since a significantly high fraction of GWAS SNPs are mapped to cxSVs (Figure 4b) than expected by chance and GWAS SNPs are enriched in non-coding regulatory regions (Maurano et al. 2012; Zhu et al. 2021), we then investigated how cxSVs might affect regulatory regions across various tissue types by assessing their enrichment and depletion across 4 regulatory chromatin-state categories and across 127 reference epigenomes (Roadmap Epigenomics Consortium et al. 2015). We find that cxSV regions are enriched for bivalent chromatin regions that encompass bivalent/poised TSS, Flanking Bivalent TSS/Enhancer and Bivalent Enhancer while simultaneously depleted for transcriptionally active chromatin states and enhancers (Figure 5, Table S18). This pattern of enrichment and depletion mirror closely to that of a similar analysis for the fastest evolving regions in the human genome (Mangan et al. 2022).

The function of bivalent chromatin states, which contain both the H3K27me3 repressive mark and the active H3K4me3 promoter mark and/or the active H3K4me1 enhancer mark, is

proposed to keep environmentally and developmentally responsive genes in a poised state allowing for precise activation and silencing (Bernstein et al. 2006; Macrae et al. 2022). Key environmental factors that drive human adaptation to different geographical regions include infectious agents, diet, and climate (Hollox et al. 2022). Here, we find that the highest proportion of significant bivalent chromatin-state enrichments are for epithelial, digestive, and muscle reference epigenomes (Figure 5, Table S18). Since bivalent chromatin states associate with genomic regions that are active in a localized and temporal manner (Bernstein et al. 2006; Zhao et al. 2021), our results suggest that cxSVs are enriched in genomic regions that function in response to specific developmental or environmental factors.

Rare cxSVs are enriched in neural genes

We next investigated the gene associations of common and rare cxSVs. We used FUMA (Watanabe et al. 2017) to perform gene ontology (GO) enrichment analysis for rare cxSVs (< 0.1% AF) in comparison against those that are common (≥ 0.1% AF). Of those 7,116 rare cxSVs, 3,304 overlap 2,897 unique protein-coding genes (Table S15) for which we find significant enrichment for those of neural functions (Figure 6c). After multiple test correction with FDR < 0.01, rare cxSVs are significantly enriched in 94 GO terms including synapse (q = 7.8e-9), neuron projection (q = 5.9e-7), synaptic membrane (q = 5.4e-9), and postsynapse (q = 5.2e-9) in the top 15 (Table S19). In contrast, common cxSVs are significantly enriched in only 9 GO terms (Table S20) with one for GO Cellular Component (Figure 6d) and with “nervous system process” being the least significantly enriched (q = 0.007) (Table S20). Enrichment fold change is significantly larger in rare cxSVs than common ones (binomial test, p = 0.0179, Figure S10). Our results suggest that cxSVs that are rare in the population or individually private disproportionately affect genes of neural function compared to common cxSVs. This suggests that there is a significantly high tendency for cxSVs to form in neural genes and a selective pressure against these cxSVs from rising in frequency in humans.

Rare cxSVs are enriched in genomic regions that recently diverged in humans

Many studies have identified rapidly evolving sequence changes in humans that are likely adaptive (Doan et al. 2016; Whalen and Pollard 2022; Mangan et al. 2022). Intriguingly, like cxSVs, many of these rapidly evolving regions are enriched at loci with high mutation rates (Mangan et al. 2022). We examined whether cxSVs might be occurring in genomic regions that are undergoing accelerated evolution in the human lineage and might contribute to ongoing adaptation in modern human populations. First, we investigated the co-localization between cxSVs and Human Accelerated Regions (HARs) (Doan et al. 2016), which are regions of the genome that are highly conserved across vertebrates but undergo accelerated sequence changes in the human lineage following human-chimpanzee divergence (Pollard et al. 2006). We find 58 HARs across 52 unique cxSV regions (Table S21) with significant overlap enrichment for rare cxSVs (permutation test, enrichment = 1.32, $p = 0.028$) but not for common cxSVs (Figure 6e), suggesting that in general, further changes to HARs may be selected against in modern humans. This is consistent with prior findings that sequence changes within HARs can be deleterious (Doan et al. 2016; Xu et al. 2015).

While HARs are regions that reflect human-specific adaptive changes of existing functional elements (Whalen and Pollard 2022), a recent study also identified functional genomic elements where the ancestral sequences were previously neutral, called “human ancestor quickly evolved regions” (HAQERs). HAQERs are the fastest-evolving regions in the human genome that rapidly diverged in the hominin lineage due to positive selection (Mangan et al. 2022). We find that 48/1581 unique HAQERs reside within 48 unique cxSV regions (Table S22). Like HARs, we find significant enrichment for HAQERs within rare cxSVs (permutation test, enrichment = 1.81, $p = 0.00020$) but not within common cxSVs.

Given that rare cxSVs are enriched in neural genes (Figure 6c, Table S19), we also examined whether cxSVs may affect neural enhancers/promoters in the developing cerebral cortex that recently evolved to have higher activity in humans (called “human-gained”) (Reilly et al. 2015). We find that while common cxSVs are not enriched (permutation test, $p > 0.05$), rare

(AF < 0.1%) cxSVs are significantly enriched for human-gained enhancer/promoter regions (Figure 6a). Rare cxSVs are especially enriched at loci with a large number of human-gained enhancer/promoter regions, called “epigenetic gain-enriched hotspots” (permutation test, enrichment = 1.49, p = 1.3e-2) (Figure 6a). Furthermore, rare cxSVs (p = 6.7e-3) but not common cxSVs (AF ≥ 0.1%) (p = 0.33) are significantly enriched within the 152 protein-coding genes within these “epigenetic gain-enriched hotspots” (Figure 6b). These genes perform functions related to regionalization, neuron differentiation/development, and A/P pattern formation (Reilly et al. 2015). The enrichment of rare cxSVs but not common cxSVs at HARs, HAQERs, and human-gained enhancers/promoters suggest that there is a significantly high tendency for cxSVs to arise in these human-evolved regions and selective pressure against them rising in allele frequency.

However, there are a number of cxSVs common to specific populations that affect human-evolved regions and may contribute to ongoing adaptation. For example, a cxSV (chr9:13734361-13755897, ABC’B’D) that is at 0.18% AF in EAS but rare elsewhere overlaps two HARs (HAR34 and 2xHAR.19) (Figure 7a, Table S21). HAR34 was experimentally validated to be an active enhancer in facial mesenchyme and forebrain using a transgenic mouse model (Visel et al. 2007; Whalen and Pollard 2022), and 2xHAR.19 was demonstrated experimentally to be an active enhancer in neural progenitor cells (Uebbing et al. 2021). The right breakpoint of the inversion part of this cxSV breaks 2xHAR.19 (chr9:13755747-13756163), and concomitantly, the inversion also places HAR34 ~9.5 kb away from its original genomic location (Table S21). Another example is cxSV on chr3: 2263830-2271090 (ADCDE) which is polymorphic across all continental superpopulations except OCN (Table S15) and co-localizes with HAR HACNS_888. This cxSV also overlaps a CTCF binding site (ENCODE accession E2174027), and both HACNS_888 and this CTCF binding site are within *CNTN4*, is a risk gene for neuropsychiatric disorders that modulates synaptic plasticity in the hippocampus (Oguro-Ando et al. 2021).

We also identify a cxSV (ch9:110389915-110390324, ABBCDED'F) that disrupts HAQER0272 (chr9:110389548-110390242) (Table S22); both are located within a human-specific lincRNA *ENSG00000235281* with restricted testis expression (Baran et al. 2015). This cxSV has AF 0.1-0.3% in AFR, EUR, and SAS populations and is rare elsewhere (Table S15,S22). Another example is cxSV (chr5:26793060-26793640, ABCB'E), which disrupts HAQER0188 along with NANOG and SPI1 developmental transcriptional factor binding sites in the same locus (Hammal et al. 2022). This cxSV is common (AF 0.3%) only in the EUR superpopulation (Table S15,S22). Lastly, cxSV (chr1:237402804-237402905 AC'D), which is polymorphic (AF 0.3%) in the MDE population, disrupts HAQER0130 and the inversion of Block C breaks a binding site of PAX8 (Hammal et al. 2022) (Table S22), which is a thyroid developmental transcription factor (Fernández et al. 2015).

Divergence between humans and non-human primates via cxSV

ARC-SV can also be used to identify cxSVs that differ between species. To investigate how cxSVs might be involved in the genome divergence of modern humans and our closest primate relative, we obtained paired-end whole-genome sequencing datasets for bonobos and performed cxSV detection with respect to GRCh38 and alignment-based SV validation against bonobo assembly (Mao et al. 2021). Overall, we find 377 loci that diverged between bonobos and modern humans via complex SVs that appear to be fixed within humans and derived within humans relative to other primates (Table S23, Table S15, Source Data). For example, we find a bonobo cxSV (chr5:29094936-29109104, AC'E) that encompasses HAR ANC302 (chr5:29105324-29105442) (Figure 7b). This 14 kb cxSV was orthogonally validated using PacBio long-reads of the same bonobo genome and against the bonobo genome assembly GCF_013052645.1 (Mao et al. 2021) (BLAT identity 98.6%) (Kuhn et al. 2013). It spans *LINC02109* which is a human-specific RNA expressed specifically in the testis (Baran et al. 2015). As this locus is fixed in the human population, we speculate that divergence from bonobos via this cxSV allowed for the rapid accelerated evolution of this locus in the human

lineage, possibly through creating *LINC02109*. Another cxSV that diverged between bonobos and humans but is fixed within humans is within *LINC02609*, where the bonobo sequence structure is ADCDE at chr1:90794594-90794971. Intriguingly, this locus in the human genome contains a human-gained brain enhancer (chr1:90794843-90796293) (Reilly et al. 2015) and a GWAS-SNP rs4561025, which is associated with neuroticism (Cai et al. 2020) (Figure 7c). In addition, we also find a cxSV in bonobos relative to humans on chr11:92193737-92196296 (AC'E) where the cxSV overlaps HAQER0245 and is linked to GWAS-SNP rs1125472, which is associated with mathematical ability (Lee et al. 2018) (Figure 7d). This cxSV is also conserved in chimpanzees and gorillas i.e. only the human version of this locus is non-ancestral suggesting that this is a human-derived sequence that went to fixation after divergence from the most recent common ancestor (MRCA) with chimpanzees and bonobos. Overall, we find 41 of such fixed human-derived sequences where divergence with our MRCA is via a cxSV (Table S24). These findings suggest that cxSVs may play a role in the divergence of humans from other primates.

cxSVs shared with Neanderthals

We also used ARC-SV to examine whether cxSVs that are polymorphic within humans are also identified in Neanderthals. In the Neanderthal genome (Prüfer et al. 2014), we discovered 5 cxSVs that are shared with modern humans (Table S25); 4 of which are common in the human population and validated in individuals with HPRC assemblies (Liao et al. 2022) (Table S3), and 1 of which is rare. One of these validated cxSVs in the Neanderthal genome is a dispersed duplication deletion (chr3:70399918-70401539, ADCDE, Figure 7f) that overlaps the oncogene *SAMMSON*, a lincRNA that is tissue-specifically expressed in more than 90% of human melanomas (Verfaillie et al. 2015). *SAMMSON* knockdown drastically decreases melanoma cell viability irrespective of *BRAF*, *NRAS*, or *TP53* mutations, whereas exogenous *SAMMSON* increases clonogenic potential (Leucci et al. 2016). This cxSV is highly prevalent in AFR populations (5.4% AF) and common in AMR (AF 0.36%) and MDE populations (AF 0.33%)

populations but rare elsewhere (Table S15,S25). Because of its high prevalence in AFR populations, this cxSV allele may have been introduced into the Neanderthal genome by modern humans, possibly when making a failed migration from Africa to the Middle East more than 100,000 years ago (Chen et al. 2020). Another example is a cxSV allele shared only between Neanderthal and EAS populations (AF = 0.061%), suggesting the possibility of Neanderthal introgression. This shared cxSV disrupts a cCRE with proximal enhancer-like signature (E2148159) (The ENCODE Project Consortium 2012) and resides within *PRMT2*, which regulates glioblastoma (Dong et al. 2018; Liu et al. 2022) (Figure 7e, Table S25). Introgressed cxSVs from archaic humans may contribute to the adaptation and phenotypic variability seen in specific modern human populations (Huerta-Sánchez et al. 2014; Simonti et al. 2016; Vernot et al. 2016; Hollox et al. 2022).

DISCUSSION

Here, we developed an innovative probabilistic model and machine learning framework ARC-SV to detect complex structural variation accurately, comprehensively, and automatically from standard WGS data that can also be implemented efficiently at population-scale. Importantly, ARC-SV operates using a “bottoms-up” approach where it does not look for data signatures to match pre-conceived structures (Collins et al. 2020; Byrska-Bishop et al. 2022) but rather the structures detected are purely data-driven (Supplementary Methods). Typical SV detection methods rely solely on the analysis of aberrant read signatures either of a single or multiple types (Escaramís et al. 2015; Lin et al. 2015; Guan and Sung 2016; Kosugi et al. 2019). In ARC-SV, we use such techniques to propose SVs, but ultimately combines the evidence both for and against each candidate in a holistic and more statistically rigorous fashion than previous methods.

By implementing ARC-SV to analyze 4,262 human genomes across the world, we show that complex genome structural variation is a major component of natural human genetic variation and pervasive in all human populations where individual superpopulations carry their

own distinct cxSV signatures (Figure 3f-g, Table S15). By performing cxSV analysis between bonobos, the Altai Neanderthal (Prufer et al. 2014), and humans, we show that cxSVs may have contributed to divergence in the human lineage and identify cxSVs that may have entered the human population via Neanderthal introgression. We believe that future applications of ARC-SV to examine disease cohorts or differences between additional species will uncover novel aspects of disease and evolutionary biology.

cxSVs have a significantly high tendency to be found in regions with elevated *de novo* mutation rates and recombination initiation hotspots (Pratto et al. 2014; Zhao et al. 2020) (Figure 4a). Loci with elevated mutation rates likely contribute an excess of alleles for evolution (Xie et al., 2019; Mangan et al. 2022). Considering this, it is perhaps not surprising that rare cxSVs are significantly enriched in genomic regions undergoing rapid or accelerated positive selection in humans, i.e. HARs (Whalen and Pollard 2022; Doan et al. 2016) and HAQERs (Mangan et al. 2022). Although common cxSVs are not enriched for HARs and HAQERs, suggesting that most changes in these regions are likely to be deleterious, the over-representation of rare cxSVs in HARs and HAQERs suggests that these regions continue to be highly mutable and may serve as potential substrates for ongoing adaptation within humans. We have identified a number of common cxSVs overlapping HARs and HAQERs that have risen in frequency in modern humans and may underlie variation in the facial mesenchyme, neural tissues, testis, and thyroid. Intriguingly, the tissues with bivalent chromatin regions most significantly enriched for both common and rare cxSVs include epithelial, digestive, and muscle tissue (Figure 5), which parallel known adaptive changes to different environments, such as changes to climate, infectious agents, diet, and terrain (Hollox et al. 2022).

The other side of the coin for hypermutable genomic regions is of course their tendency to generate disease predisposing alleles. Although cxSVs are often omitted in current GWAS due to the limited ability to analyze SVs (Winkler et al. 2014; Torkamaneh and Belzile 2022), we find that the GWAS SNPs that most closely associate with cxSVs also associate with disease

phenotypes in tissues that mark human-specific adaptations such as spine, gastrointestinal, and neural (Figure 4d, Table S16,S17). Interestingly, we find that cxSVs that are rare in the population or are private to individuals are enriched in genes with neurodevelopmental functions (Table S19). It is very plausible that leveraging ARC-SV to identify cxSVs in disease populations will reveal new disease-associated alleles and help uncover the the molecular pathophysiology of neurodevelopmental and neuropsychiatric disorders.

However, like any reference-based method using short-read, short-insert, WGS data, ARC-SV is also subject to uncertainty and lack of long of long-range information in short-read alignments. Currently, we show that ARC-SV outperforms long read-based cxSV detection methods such as SVision (Lin et al. 2022) (Table S13). With at least three new ultra-high-throughput short-read sequencers from three different companies just released this year to compete with the Illumina short-read platform (Pennisi 2022) and to meet the demand of new initiatives to sequence millions of individuals for national genome projects and population-scale studies (Wonkam 2021; Gaziano et al. 2016; Saunders et al. 2019), there remains a vital need to develop effective analysis methods such as ARC-SV to extract new dimensions of genetic information from short-read sequencing data. In anticipation of this new era of an unprecedented scale in the generation of population-level WGS data, we also show here that ARC-SV can be readily implemented for such competing short-read platforms Element Bioscience (Arslan et al. 2022) and MGI (Shen et al. 2020) where concordance for cxSV calls with Illumina data are 92% and 80% respectively (Table S26,S27).

A key component to the development of our statistical framework is leveraging the high-quality state-of-the-art diploid genome assemblies of 42 diverse individuals by the HPRC (Liao et al. 2022; Miga and Wang 2021) for building our machine learning models to predict cxSV events at a much greater population-scale. We anticipate that the performance of these models will substantially increase when we can re-train them with the upcoming release of the full assemblies of all 350 individuals (Liao et al. 2022; Miga and Wang 2021). Our method to

automatically detect and accurately resolve cxSVs from standard paired-end WGS opens new doorways in future research, particularly for large cohorts, to address how such aspects of genome diversity can contribute to various biological traits, evolution, genome biology, and diseases such as cancer and neurodevelopmental disorders.

Methods

Datasets for population SV calling and validation

Data for population SV calling and analysis were obtained directly from https://www.internationalgenome.org. HG01783 was removed from analysis due to ambiguous ancestry curation (indicated as both AFR and EUR). HuRef and HepG2 WGS data were obtained from Zhou et al (Zhou et al. 2018a) and Zhou et al (Zhou et al. 2019) respectively. Alignment data for calling SVs to benchmarking against those that are manually curated from NA12878 were obtained from Zhou et al (Zhou et al. 2018b).

SV detection

Complex and simple (deletions, tandem duplications, and inversions) SVs were called using both probabilistic and machine learning algorithm components (see Supplementary Methods for full description of probabilistic and machine learning models). Computation was performed using the Avitohol supercomputer (Atanassov et al. 2017). SVs with (1) greater than 85% overlap to simple repeat, low complexity, satellite repeat, and simple tandem repeat annotations (Kuhn et al. 2013), (2) greater than 50% overlap with segmental duplications, (3) affecting fewer than 50 bp (see Affected Sequence Calculation in Supplementary Methods), (4) overlap with SV blacklist regions (https://cf.10xgenomics.com/supp/genome/GRCh38/sv_blacklist.bed), (5) and overlap reference gaps were filtered out. Such SV filtering criteria was also used for paired-end WGS-based SV call comparisons. Filtering was implemented using the “arcsv merge-filter” function. Additionally, only SVs with split-read or paired-end read support ≥ 1 were retained.

SV merging

For population SV calls, only high-confidence SV calls identified from machine learning prediction (Supplementary Methods) were retained. SVs at the same locus and of the same type were merged into a master list. cxSVs at the same locus (50% reciprocal base pair overlap) and of the same type were merged.

Genomic DNA

Genomic DNA sample of Craig Venter (NS12911) was purchased from the Coriell Institute. HepG2 DNA was extracted from cells using Qiagen DNeasy Blood & Tissue Kit (Cat No. 69504), and concentration was quantified using the Qubit dsDNA BR Assay Kit (Invitrogen, Waltham, MA, USA). Purity of gDNA (OD260/280 > 1.8; OD260/230 > 1.5) was verified using NanoDrop (Thermo Scientific, Waltham, MA, USA).

Experimental validation of complex SVs

PCR primers were using Primer 3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) and verified using UCSC In-Silico PCR tool (http://genome.cse.ucsc.edu/cgi-bin/hgPcr). Each complex SV was sequence assembled with 150 bp of flanking genomic sequence on each side (Supplementary Table S7), and primers were designed to bind the flanking sequences and to amplify the assembled cxSV in its entirety including all predicted breakpoints. Each target was amplified using 50 ng of genomic DNA. The amplification was performed using Phusion Green high-fidelity DNA polymerase (Thermo Fisher Scientific Inc, Waltham, MA) under the following conditions: initial denature at 98°C for 30 sec, total 30 cycles of: 98°C for 10 sec, 60°C for 30 sec, and 72°C for 1 min (or 2 min for longer PCR product), and final extension at 72°C for 10 min. The PCR products were then analyzed by agarose gel electrophoresis. PCR products were then purified using Zymoclean Gel DNA Recovery Kit (Zymo Research, Irvine, CA) and submitted for Sanger sequencing at Sequetech Corporation (Mountain View, CA).

AF calculation

Allele frequency (AF) is calculated as the number of cxSV events divided by the number of haplotypes (i.e. 2 x number of samples). AF is calculated within all samples across different

populations. Common and rare cxSVs are defined as those whose AF is larger and smaller than 0.001, respectively.

Principal component analysis (PCA)

We performed PCA for common cxSVs using R function “prcomp”. For the top two PCs, we identified the top ten cxSVs with largest positive loadings and the top ten cxSVs with largest negative loadings as the population-specific cxSVs. We extracted 63 protein coding genes whose distance to the nearest population-specific cxSV is smaller than 100KB to perform enrichment analysis (see below).

gnomAD-SV LD maps

The Supplementary Table 7 of Collins et al (Collins et al. 2020) catalogued 15,634 SVs (AF ≥ 1%) in strong LD (R2 ≥ 0.8) with at least one SNV within +/- 1Mb on the basis of 3,470 AFR and 1,883 EUR samples from the gnomAD datasets. We used these publicly available LD maps to examine LD-distance relationship between gnomAD-SVs and SNVs, which further guided our choice of window size (125 kb) for mapping common SNPs to cxSVs in the present study.

GWAS Catalog lookup

For each cxSV, we looked up GWAS SNPs within its +/- 125 kb interval in the GWAS Catalog (Sollis et al. 2023) version 1.0.2 mapped to genome assembly GRCh38.p13 and dbSNP Build 154 (https://www.ebi.ac.uk/gwas/docs/file-downloads). We only used GWAS SNPs reaching the genome-wide significant level (p ≤ 5e-8) in this analysis.

Heritability enrichment

We quantified the heritability enrichment of common SNPs (MAF ≥ 0.05) within +/- 125 kb of cxSVs across 1087 GWAS datasets (with total h2 Z-score ≥ 2) as previously described (Zhu et al. 2022). In brief, we first created the binary annotation indicating if a common SNP falls inside +/- 125 kb of any cxSV, and then used S-LDSC (Finucane et al. 2015) to estimate the standardized effect size of the binary annotation (defined as the 1-standard deviation increase of the binary annotation) conditioning on 96 baseline genomic annotations (Hujoel et al. 2019) in

each GWAS, and lastly computed a one-sided p-value testing if the standardized effect size > 0. Given a GWAS dataset, a significantly positive standardized effect size indicates a significant heritability enrichment. As the 1087 GWAS datasets were mapped to genome assembly GRCh37, we lifted over cxSVs from GRCh38 to GRCh37 using UCSC liftOver (Kuhn et al. 2013) in this analysis.

SV enrichment analysis

We merged all unique cxSVs detected into non-overlapping genomic segments (n=7194) and examined if these genomic segments are associated with elevated recombination rates. We performed a permutation test to assess the significance of enrichment. For each permutation, we randomly selected 7194 non-overlapping regions from the genome with their sizes matching the cxSV regions, and calculated permuted overlaps defined as the number of regions that overlapped recombination hotspots (Pratto et al. 2014). The permutation procedure was repeated 100000 times. The enrichment was defined as the ratio between the data overlap (number of non-overlapping cxSV regions that overlap recombination hotspots) and the mean of 100000 permuted overlaps, and the significance p-value was calculated by comparing the data overlap with 100000 permuted overlaps. We also used the permutation test to investigate the enrichment of cxSVs in HARs, HAQERs, chromatin regions from four categories and 127 cell types, as well as human-gained enhancers. To investigate whether the cxSVs are prone to elevated mutation rate, we performed a permutation test to examine the enrichment of germline DNMs in the 7194 cxSV regions. In this specific analysis, data overlap and permuted overlap were defined as the number of DNM events falling into the 7194 cxSV regions and 7194 permuted regions, respectively. The enrichment and significance p-value were calculated similarly as in the enrichment analysis for recombination hotspots. Using GENCODE V42 (Frankish et al. 2021), we extracted 3130, 2897, and 566 genes that overlap all cxSVs, rare cxSVs, and common cxSVs, respectively from 19086 autosomal protein coding genes. Here the overlapped genes are defined as those whose distance to the nearest cxSV is less than 5 KB.

Then we used hypergeometric test to examine the enrichment of these cxSV genes in the neurodevelopmental enhancer gene set as well as the gene ontology terms (Watanabe et al. 2017).

SV detection in Neanderthal and bonobo genomes

Neanderthal WGS reads were obtained from https://www.eva.mpg.de/genetics/genome-projects/neandertal (Prüfer et al. 2014). Bonobo WGS reads were obtained from SRA (accession SRR13998205). WGS reads were aligned to hg38 (GCA_000001405.15) using BWA-MEM (Li and Durbin 2009) version 0.7.15-r1140 (parameters -Y -K 100000000). PCR duplicates were marked using Picard tools version 2.4.1 (http://broadinstitute.github.io/picard/). SV calls were called by ARC-SV (see SV detection section). For Neanderthal SV calls, only those with corresponding SV calls (95% reciprocal SV coordinate overlap and matching SV type) in modern human populations were retained. All bonobo SVs were sequence assembled with 150 bp of flanking sequence tagged on both ends and validated against bonobo genome assembly (GCA_013052645.3) using BLAT version 3.5 (Kuhn et al. 2013) requiring a minimum of 97.5% sequence identity, that query and target alignment range are within 2.5%, that total target and query gap bases are no more than 2.5% of total size, and that query alignment must initiate within the first 50 bp of 5' tagged flanking sequence and continuously align/match target until at least the last 50 bp of the 3' tagged flanking sequence. Same workflow was used for comparison against chimpanzee (GCF_002880755.1) and gorilla (GCF_008122165.1) genomes.

Author contributions

A.E.U. and W.H.W. supervised the study. B.Z., J.G.A., H.G., and X.Z. wrote the manuscript. A.E.U., W.H.W., B.Z., D.P., X.Z., G.S., and H.P.J. contributed resources. J.G.A. and W.H.W. conceived and designed the statistical framework for SV detection. B.Z., T.K., and G.S. designed and implemented machine learning framework for SV detection. J.G.A. wrote the software package. B.Z. conceived and designed method validation and implementation across datasets. B.Z., Y.H., and R.P. performed experiments. B.Z., C.R.H., Y.H., D.P., and H.L.

performed SV calling, validation, and comparative analysis. H.G. performed population SV analysis. X.Z. designed and performed GWAS catalogue and heritability enrichment analysis. H.G. designed and performed all enrichment analysis.

Acknowledgments

We thank Dr. Douglas F. Levinson, Dr. Bernardo Rodriguez Martin, Dr. Heinrich Dohna, and Dr. Janet Song for manuscript proofreading and providing feedback.

B.Z. is funded by NIH grant K01MH129758 and the Stanford Maternal and Child Health Research Institute Instructor K Support Award. J.G.A. received funding from NIH training grant T32-GM096982 and NSF Graduate Fellowship DGE-114747. H.L. and H.P.J. were funded by NIH grant U01HG01096 and R01HG006137. G.S. is funded by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2010775). D.P. acknowledges the provided access to the e-infrastructure of the NCHDC – part of the Bulgarian National Roadmap on RIs, with the financial support by Grant No D01-168/28.07.2022. X.Z. is supported by Stein Fellowship from Stanford University, Institute for Computational and Data Sciences Seed Grant and Consortium on Substance Use and Addiction Seed Grant from Pennsylvania State University. A.E.U. and W.H.W. are funded by NIH grant P50HG00773506 (renewal of P50HG007735). W.H.W is also funded by NIH grant R01HG010359 and NSF grant DMS1952386.

REFERENCES

Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, Layer RM, Neale BM, Salerno WJ, Reeves C, et al. 2020. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**: 83–89. http://www.ncbi.nlm.nih.gov/pubmed/32460305.

Arslan S, Garcia FJ, Guo M, Kellinger MW, Kruglyak S, LeVieux JA, Mah AH, Wang H, Zhao J, Zhou C, et al. 2022. Sequencing by avidity enables high accuracy with low reagent consumption. *bioRxiv* 2022.11.03.514117. http://biorxiv.org/content/early/2022/11/04/2022.11.03.514117.abstract.

Atanassov E, Gurov T, Ivanovska S, Karaivanova A. 2017. Parallel Monte Carlo on Intel MIC Architecture. *Procedia Comput Sci* **108**: 1803–1810. https://linkinghub.elsevier.com/retrieve/pii/S1877050917307226.

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**: 663-675.e19.

http://www.ncbi.nlm.nih.gov/pubmed/30661756.
Bae T, Tomasini L, Mariani J, Zhou B, Roychowdhury T, Franjic D, Pletikos M, Pattni R, Chen B-J, Venturini E, et al. 2018. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* **359**: 550–555. http://www.ncbi.nlm.nih.gov/pubmed/29217587.
Baran Y, Subramaniam M, Biton A, Tukiainen T, Tsang EK, Rivas MA, Pirinen M, Gutierrez-Arcelus M, Smith KS, Kukurba KR, et al. 2015. The landscape of genomic imprinting across diverse adult human tissues. *Genome Res* **25**: 927–36. http://www.ncbi.nlm.nih.gov/pubmed/25953952.
Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, et al. 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**. http://www.ncbi.nlm.nih.gov/pubmed/32193295.
Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**: 315–326. https://linkinghub.elsevier.com/retrieve/pii/S0092867406003801.
Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**: 3426-3440.e19. http://www.ncbi.nlm.nih.gov/pubmed/36055201.
Cabodi S, Calautti E, Talora C, Kuroki T, Stein PL, Dotto GP. 2000. A PKC-eta/Fyn-dependent pathway leading to keratinocyte growth arrest and differentiation. *Mol Cell* **6**: 1121–9. http://www.ncbi.nlm.nih.gov/pubmed/11106751.
Cai N, Revez JA, Adams MJ, Andlauer TFM, Breen G, Byrne EM, Clarke T-K, Forstner AJ, Grabe HJ, Hamilton SP, et al. 2020. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat Genet* **52**: 437–447. http://www.ncbi.nlm.nih.gov/pubmed/32231276.
Cameron DL, Di Stefano L, Papenfuss AT. 2019. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun* **10**: 3240. http://www.ncbi.nlm.nih.gov/pubmed/31324872.
Carvalho CMB, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17**: 224–38. http://www.ncbi.nlm.nih.gov/pubmed/26924765.
Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1784. http://www.ncbi.nlm.nih.gov/pubmed/30992455.
Chen L, Wolf AB, Fu W, Li L, Akey JM. 2020. Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell* **180**: 677-687.e16. https://linkinghub.elsevier.com/retrieve/pii/S0092867420300593.
Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**: 1220–2. http://www.ncbi.nlm.nih.gov/pubmed/26647377.
Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, GTEx Consortium, et al. 2017. The impact of structural variation on human gene expression. *Nat Genet* **49**: 692–699. http://www.ncbi.nlm.nih.gov/pubmed/28369037.
Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera A V, Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581**: 444–451. http://www.ncbi.nlm.nih.gov/pubmed/32461652.
Collins RL, Brand H, Redin CE, Hanscom C, Antolik C, Stone MR, Glessner JT, Mason T, Pregno G, Dorrani N, et al. 2017. Defining the diverse spectrum of inversions, complex

structural variation, and chromothripsis in the morbid human genome. *Genome Biol* **18**: 36. http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1158-6.
Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G, Giachino D, Mandrile G, Espejo Valle-Inclan J, et al. 2017. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun* **8**: 1326. http://www.ncbi.nlm.nih.gov/pubmed/29109544.
Doan RN, Bae B-I, Cubelos B, Chang C, Hossain AA, Al-Saad S, Mukaddes NM, Oner O, Al-Saffar M, Balkhy S, et al. 2016. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* **167**: 341-354.e12. http://www.ncbi.nlm.nih.gov/pubmed/27667684.
Dong F, Li Q, Yang C, Huo D, Wang X, Ai C, Kong Y, Sun X, Wang W, Zhou Y, et al. 2018. PRMT2 links histone H3R8 asymmetric dimethylation to oncogenic activation and tumorigenesis of glioblastoma. *Nat Commun* **9**: 4552. https://www.nature.com/articles/s41467-018-06968-7.
Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**. http://www.ncbi.nlm.nih.gov/pubmed/33632895.
Escaramís G, Docampo E, Rabionet R. 2015. A decade of structural variants: description, history and methods to detect structural variation. *Brief Funct Genomics* **14**: 305–14. http://www.ncbi.nlm.nih.gov/pubmed/25877305.
Fernández LP, López-Márquez A, Santisteban P. 2015. Thyroid transcription factors in development, differentiation and disease. *Nat Rev Endocrinol* **11**: 29–42. http://www.ncbi.nlm.nih.gov/pubmed/25350068.
Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, Anttila V, Xu H, Zang C, Farh K, et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**: 1228–35. http://www.ncbi.nlm.nih.gov/pubmed/26414678.
Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, et al. 2017. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**. http://www.ncbi.nlm.nih.gov/pubmed/28605766.
Forment J V, Kaidi A, Jackson SP. 2012. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nat Rev Cancer* **12**: 663–70. http://www.ncbi.nlm.nih.gov/pubmed/22972457.
Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, et al. 2021. GENCODE 2021. *Nucleic Acids Res* **49**: D916–D923. http://www.ncbi.nlm.nih.gov/pubmed/33270111.
Frith MC, Khan S. 2018. A survey of localized sequence rearrangements in human DNA. *Nucleic Acids Res* **46**: 1661–1673. http://www.ncbi.nlm.nih.gov/pubmed/29272440.
Fujimoto A, Wong JH, Yoshii Y, Akiyama S, Tanaka A, Yagi H, Shigemizu D, Nakagawa H, Mizokami M, Shimada M. 2021. Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. *Genome Med* **13**: 65. http://www.ncbi.nlm.nih.gov/pubmed/33910608.
Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, Mac S, Peluso P, Hatas E, Ghurye J, et al. 2021. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol* **39**: 309–312. http://www.ncbi.nlm.nih.gov/pubmed/33288905.
Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, Whitbourne S, Deen J, Shannon C, Humphries D, et al. 2016. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**: 214–223. https://linkinghub.elsevier.com/retrieve/pii/S0895435615004448.

Guan P, Sung W-K. 2016. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods* **102**: 36–49. http://www.ncbi.nlm.nih.gov/pubmed/26845461.
Hammal F, de Langen P, Bergon A, Lopez F, Ballester B. 2022. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res* **50**: D316–D325. https://academic.oup.com/nar/article/50/D1/D316/6423925.
Hollox EJ, Zuccherato LW, Tucci S. 2022. Genome structural variation in human evolution. *Trends Genet* **38**: 45–58. http://www.ncbi.nlm.nih.gov/pubmed/34284881.
Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**: 194–7. http://www.ncbi.nlm.nih.gov/pubmed/25043035.
Hujoel MLA, Gazal S, Hormozdiari F, van de Geijn B, Price AL. 2019. Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species. *Am J Hum Genet* **104**: 611–624. https://linkinghub.elsevier.com/retrieve/pii/S0002929719300539.
Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen C-F, Thomas MA, Haussler D, Jacob HJ. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res* **14**: 528–38. http://www.ncbi.nlm.nih.gov/pubmed/15059993.
Jinno Y, Yun K, Nishiwaki K, Kubota T, Ogawa O, Reeve AE, Niikawa N. 1994. Mosaic and polymorphic imprinting of the WT1 gene in humans. *Nat Genet* **6**: 305–309. http://www.nature.com/articles/ng0394-305.
Kennedy RB, Ovsyannikova IG, Pankratz VS, Haralambieva IH, Vierkant RA, Poland GA. 2012. Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients. *Hum Genet* **131**: 1403–21. http://www.ncbi.nlm.nih.gov/pubmed/22610502.
Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* **20**: 117. http://www.ncbi.nlm.nih.gov/pubmed/31159850.
Kuhn RM, Haussler D, Kent WJ. 2013. The UCSC genome browser and associated tools. *Brief Bioinform* **14**: 144–161. https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs038.
Lappalainen T, Scott AJ, Brandt M, Hall IM. 2019. Genomic Analysis in the Age of Human Genome Sequencing. *Cell* **177**: 70–84. http://www.ncbi.nlm.nih.gov/pubmed/30901550.
Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.
Lee H, Greer SU, Pavlichin DS, Hughes CR, Zhou B, Weissman T, Ji HP. 2022. The Human Pangenome's sequence conservation reveals a landscape of polymorphic structural variations. *bioRxiv* 2022.10.06.511239. http://biorxiv.org/content/early/2022/10/13/2022.10.06.511239.abstract.
Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, Nguyen-Viet TA, Bowers P, Sidorenko J, Karlsson Linnér R, et al. 2018. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* **50**: 1112–1121. http://www.ncbi.nlm.nih.gov/pubmed/30038396.
Leucci E, Vendramin R, Spinazzi M, Laurette P, Fiers M, Wouters J, Radaelli E, Eyckerman S, Leonelli C, Vanderheyden K, et al. 2016. Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* **531**: 518–522. http://www.nature.com/articles/nature17161.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–60. http://www.ncbi.nlm.nih.gov/pubmed/19451168.
Li J, Li M, Wang T, Liu X, Zhu X, Dai Y, Zhai K, Liu Y, Lin J, Ge R, et al. 2021. SLC38A4 functions as a tumour suppressor in hepatocellular carcinoma through modulating Wnt/β-

catenin/MYC/HMGCS2 axis. *Br J Cancer* **125**: 865–876. https://www.nature.com/articles/s41416-021-01490-y.
Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2022. A Draft Human Pangenome Reference. *bioRxiv* 2022.07.09.499321. http://biorxiv.org/content/early/2022/07/09/2022.07.09.499321.abstract.
Lin J, Wang S, Audano PA, Meng D, Flores JI, Kosters W, Yang X, Jia P, Marschall T, Beck CR, et al. 2022. SVision: a deep learning approach to resolve complex structural variants. *Nat Methods* **19**: 1230–1233. http://www.ncbi.nlm.nih.gov/pubmed/36109679.
Lin K, Smit S, Bonnema G, Sanchez-Perez G, de Ridder D. 2015. Making the difference: integrating structural variation detection tools. *Brief Bioinform* **16**: 852–64. http://www.ncbi.nlm.nih.gov/pubmed/25504367.
Liu L, Lin B, Yin S, Ball LE, Delaney JR, Long DT, Gan W. 2022. Arginine methylation of BRD4 by PRMT2/4 governs transcription and DNA repair. *Sci Adv* **8**. https://www.science.org/doi/10.1126/sciadv.add8928.
Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21**: 597–614.
Macrae TA, Fothergill-Robinson J, Ramalho-Santos M. 2022. Regulation, functions and transmission of bivalent chromatin during mammalian development. *Nat Rev Mol Cell Biol*. https://www.nature.com/articles/s41580-022-00518-2.
Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**: 201–206. http://www.ncbi.nlm.nih.gov/pubmed/27654912.
Mangan RJ, Alsina FC, Mosti F, Sotelo-Fonseca JE, Snellings DA, Au EH, Carvalho J, Sathyan L, Johnson GD, Reddy TE, et al. 2022. Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell* **185**: 4587-4603.e23. http://www.ncbi.nlm.nih.gov/pubmed/36423581.
Mao Y, Catacchio CR, Hillier LW, Porubsky D, Li R, Sulovari A, Fernandes JD, Montinaro F, Gordon DS, Storer JM, et al. 2021. A high-quality bonobo genome refines the analysis of hominid evolution. *Nature* **594**: 77–81. http://www.nature.com/articles/s41586-021-03519-x.
Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**: 1190–5. http://www.ncbi.nlm.nih.gov/pubmed/22955828.
Medina-Cano D, Ucuncu E, Nguyen LS, Nicouleau M, Lipecka J, Bizot J-C, Thiel C, Foulquier F, Lefort N, Faivre-Sarrailh C, et al. 2018. High N-glycan multiplicity is critical for neuronal adhesion and sensitizes the developing cerebellum to N-glycosylation defect. *Elife* **7**. http://www.ncbi.nlm.nih.gov/pubmed/30311906.
Miga KH, Wang T. 2021. The Need for a Human Pangenome Reference Sequence. *Annu Rev Genomics Hum Genet* **22**: 81–102. http://www.ncbi.nlm.nih.gov/pubmed/33929893.
Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, Wong WH, Lam HYK. 2015. MetaSV: An accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* **31**: 2741–2744.
Murat P, Perez C, Crisp A, van Eijk P, Reed SH, Guilbaud G, Sale JE. 2022. DNA replication initiation shapes the mutational landscape and expression of the human genome. *Sci Adv* **8**: eadd3686. http://www.ncbi.nlm.nih.gov/pubmed/36351018.
Nam S, Lim J-S. 2016. Essential role of interferon regulatory factor 4 (IRF4) in immune cell development. *Arch Pharm Res* **39**: 1548–1555. http://www.ncbi.nlm.nih.gov/pubmed/27826752.
Nesta A V, Tafur D, Beck CR. 2021. Hotspots of Human Mutation. *Trends Genet* **37**: 717–729. http://www.ncbi.nlm.nih.gov/pubmed/33199048.
Oguro-Ando A, Bamford RA, Sital W, Sprengers JJ, Zuko A, Matser JM, Oppelaar H,

Sarabdjitsingh A, Joëls M, Burbach JPH, et al. 2021. Cntn4, a risk gene for neuropsychiatric disorders, modulates hippocampal synaptic plasticity and behavior. *Transl Psychiatry* **11**: 106. http://www.ncbi.nlm.nih.gov/pubmed/33542194.

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, et al. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* **11**: R52. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-5-r52.

Parikh H, Mohiyuddin M, Lam HYK, Iyer H, Chen D, Pratt M, Bartha G, Spies N, Losert W, Zook JM, et al. 2016. svclassify: a method to establish benchmark structural variant calls. *BMC Genomics* **17**: 64. http://www.ncbi.nlm.nih.gov/pubmed/26772178.

Pennisi E. 2022. Upstart DNA sequencers could be a “game changer”. *Science* **376**: 1257–1258. http://www.ncbi.nlm.nih.gov/pubmed/35709273.

Pratto F, Brick K, Khil P, Smagulova F, Petukhova G V, Camerini-Otero RD. 2014. DNA recombination. Recombination initiation maps of individual human genomes. *Science* **346**: 1256442. http://www.ncbi.nlm.nih.gov/pubmed/25395542.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**: 43–9. http://www.ncbi.nlm.nih.gov/pubmed/24352235.

Quinlan AR, Hall IM. 2012. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet* **28**: 43–53. http://linkinghub.elsevier.com/retrieve/pii/S0168952511001685.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333–i339. http://www.ncbi.nlm.nih.gov/pubmed/22962449.

Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, Sarro R, Rakic P, Noonan JP. 2015. Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**: 1155–9. http://www.ncbi.nlm.nih.gov/pubmed/25745175.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–30. http://www.ncbi.nlm.nih.gov/pubmed/25693563.

Sanchis-Juan A, Stephens J, French CE, Gleadall N, Mégy K, Penkett C, Shamardina O, Stirrups K, Delon I, Dewhurst E, et al. 2018. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med* **10**: 95. http://www.ncbi.nlm.nih.gov/pubmed/30526634.

Saunders G, Baudis M, Becker R, Beltran S, Béroud C, Birney E, Brooksbank C, Brunak S, Van den Bulcke M, Drysdale R, et al. 2019. Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat Rev Genet* **20**: 693–701. http://www.ncbi.nlm.nih.gov/pubmed/31455890.

Schreiber SN, Emter R, Hock MB, Knutti D, Cardenas J, Podvinec M, Oakeley EJ, Kralli A. 2004. The estrogen-related receptor α (ERRα) functions in PPARγ coactivator 1α (PGC-1α)-induced mitochondrial biogenesis. *Proc Natl Acad Sci* **101**: 6472–6477. https://pnas.org/doi/full/10.1073/pnas.0308686101.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**: 461–468. http://www.ncbi.nlm.nih.gov/pubmed/29713083.

Sekar S, Tomasini L, Proukakis C, Bae T, Manlove L, Jang Y, Scuderi S, Zhou B, Kalyva M, Amiri A, et al. 2020. Complex mosaic structural variations in human fetal brains. *Genome Res* **30**: 1695–1704. http://www.ncbi.nlm.nih.gov/pubmed/33122304.

Shen H, Liu P, Li Z, Chen F, Jiang H, Shi S, Xi Y, Li Q, Wang X, Zhao J, et al. 2020. Advanced Whole Genome Sequencing Using an Entirely PCR-free Massively Parallel Sequencing Workflow. *bioRxiv* 2019.12.20.885517. http://biorxiv.org/content/early/2020/11/01/2019.12.20.885517.abstract.

Simonti CN, Vernot B, Bastarache L, Bottinger E, Carrell DS, Chisholm RL, Crosslin DR, Hebbring SJ, Jarvik GP, Kullo IJ, et al. 2016. The phenotypic legacy of admixture between modern humans and Neandertals. *Science* **351**: 737–41. http://www.ncbi.nlm.nih.gov/pubmed/26912863.

Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, et al. 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**: D977–D985. http://www.ncbi.nlm.nih.gov/pubmed/30445434.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**: 437–55. http://www.ncbi.nlm.nih.gov/pubmed/20059347.

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**: 27–40. http://www.ncbi.nlm.nih.gov/pubmed/21215367.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**: 848–53. http://www.ncbi.nlm.nih.gov/pubmed/17289997.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81. http://www.nature.com/doifinder/10.1038/nature15394.

The 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–73. http://www.nature.com/doifinder/10.1038/nature09534.

The 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65. http://www.ncbi.nlm.nih.gov/pubmed/23128226.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. http://www.ncbi.nlm.nih.gov/pubmed/22955616.

Torkamaneh D, Belzile F, eds. 2022. *Genome-Wide Association Studies*. Springer US, New York, NY https://link.springer.com/10.1007/978-1-0716-2237-7.

Uebbing S, Gockley J, Reilly SK, Kocher AA, Geller E, Gandotra N, Scharfe C, Cotney J, Noonan JP. 2021. Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proc Natl Acad Sci U S A* **118**. http://www.ncbi.nlm.nih.gov/pubmed/33372131.

Verfaillie A, Imrichova H, Atak ZK, Dewaele M, Rambow F, Hulselmans G, Christiaens V, Svetlichnyy D, Luciani F, Van den Mooter L, et al. 2015. Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nat Commun* **6**: 6683. http://www.ncbi.nlm.nih.gov/pubmed/25865119.

Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, Dannemann M, Grote S, McCoy RC, Norton H, et al. 2016. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**: 235–9. http://www.ncbi.nlm.nih.gov/pubmed/26989198.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**: D88-92.

http://www.ncbi.nlm.nih.gov/pubmed/17130149.
Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. 2017. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun 2017 81* **8**: 1–11.
Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14**: 125–38. http://www.ncbi.nlm.nih.gov/pubmed/23329113.
Whalen S, Pollard KS. 2022. Enhancer Function and Evolutionary Roles of Human Accelerated Regions. *Annu Rev Genet* **56**: 423–439. http://www.ncbi.nlm.nih.gov/pubmed/36070559.
Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, Ferreira T, Fall T, Graff M, Justice AE, et al. 2014. Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**: 1192–1212. http://www.nature.com/articles/nprot.2014.071.
Wonkam A. 2021. Sequence three million genomes across Africa. *Nature* **590**: 209–211. http://www.nature.com/articles/d41586-021-00313-7.
Xu K, Schadt EE, Pollard KS, Roussos P, Dudley JT. 2015. Genomic and network patterns of schizophrenia genetic variation in human evolutionary accelerated regions. *Mol Biol Evol* **32**: 1148–60. http://www.ncbi.nlm.nih.gov/pubmed/25681384.
Yalcin B, Wong K, Bhomra A, Goodson M, Keane TM, Adams DJ, Flint J. 2012. The fine-scale architecture of structural variants in 17 mouse genomes. *Genome Biol* **13**: R18. http://www.ncbi.nlm.nih.gov/pubmed/22439878.
Zhao G, Li K, Li B, Wang Z, Fang Z, Wang X, Zhang Y, Luo T, Zhou Q, Wang L, et al. 2020. Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Res* **48**: D913–D926. http://www.ncbi.nlm.nih.gov/pubmed/31642496.
Zhao K, Kong D, Jin B, Smolke CD, Rhee SY. 2021. A novel bivalent chromatin associates with rapid induction of camalexin biosynthesis genes in response to a pathogen signal in Arabidopsis. *Elife* **10**. http://www.ncbi.nlm.nih.gov/pubmed/34523419.
Zhou B, Arthur JG, Ho SS, Pattni R, Huang Y, Wong WH, Urban AE. 2018a. Extensive and deep sequencing of the Venter/HuRef genome for developing and benchmarking genome analysis tools. *Sci Data* **5**: 180261. http://www.nature.com/articles/sdata2018261.
Zhou B, Ho SS, Greer SU, Spies N, Bell JM, Zhang X, Zhu X, Arthur JG, Byeon S, Pattni R, et al. 2019. Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2. *Nucleic Acids Res* **47**: 3846–3861. https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz169/5377472.
Zhou B, Ho SS, Zhang X, Pattni R, Haraksingh RR, Urban AE. 2018b. Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J Med Genet* **55**: 735–743. http://www.ncbi.nlm.nih.gov/pubmed/30061371.
Zhu X, Duren Z, Wong WH. 2021. Modeling regulatory network topology improves genome-wide analyses of complex human traits. *Nat Commun* **12**: 2851. http://www.ncbi.nlm.nih.gov/pubmed/33990562.
Zhu X, Ma S, Wong WH. 2022. Genetic effects of sequence-conserved enhancer-like elements on human complex traits. *bioRxiv* 2022.08.19.504589. http://biorxiv.org/content/early/2022/08/21/2022.08.19.504589.abstract.

Figure Legends

Figure 1. High-level overview of the ARC–SV method. (a) Given a set of paired-end read alignments, candidate breakpoints are determined from discordant read pairs (1 and 2) as well as split reads (3) and soft-clipped alignments. In this case, read 1 suggests the AB and CD breakpoints (orange) while reads 2 and 3 support the BC and CD breakpoints (purple). Possible novel adjacencies (dotted) are then determined e.g. Read 1 suggests an adjacency CB (orange dotted line) while reads 2 and 3 suggest the adjacency BE. Candidate haplotype structures are proposed by enumerating paths through the novel adjacencies. Diploid genotypes are scored using a likelihood model for observed alignments. In this case, the rearrangement ABCBE on one haplotype explains the discordant pairs (1 and 2) and split read (3) that did not match the reference. (b) ARC-SV calls from 31 HPRC individuals were sequence assembled and aligned against their respective HPRC diploid assemblies (Ground Truth) and (c) labeled as "validated" or "not validated". Features from these calls were then used to train a machine learning model to identify high-confidence SVs from other "unseen" genomes. (d) Examples of complex SVs that can be called by ARC-SV. (e) ROC curve of machine learning models trained to predict cxSV validation in 11 test genomes. (f) Confusion matrices for performance of XGB machine learning model in non-HPRC genomes to predict cxSV validation.

Figure 2. Examples of cxSVs validated in HPRC assemblies and with population-specific signatures. (a-d) Plots mapped genomic locations of unique k-mers (31bp) in GRCh38 (x-axis) and respective HPRC assemblies (y-axis) at loci with cxSV identification by ARC-SV. ENCCODE cCREs with distal enhancer-like signatures indicated by "enhD". GH#s represent GeneHancer annotations. Bar plots represent AF (%) in each superpopulations.

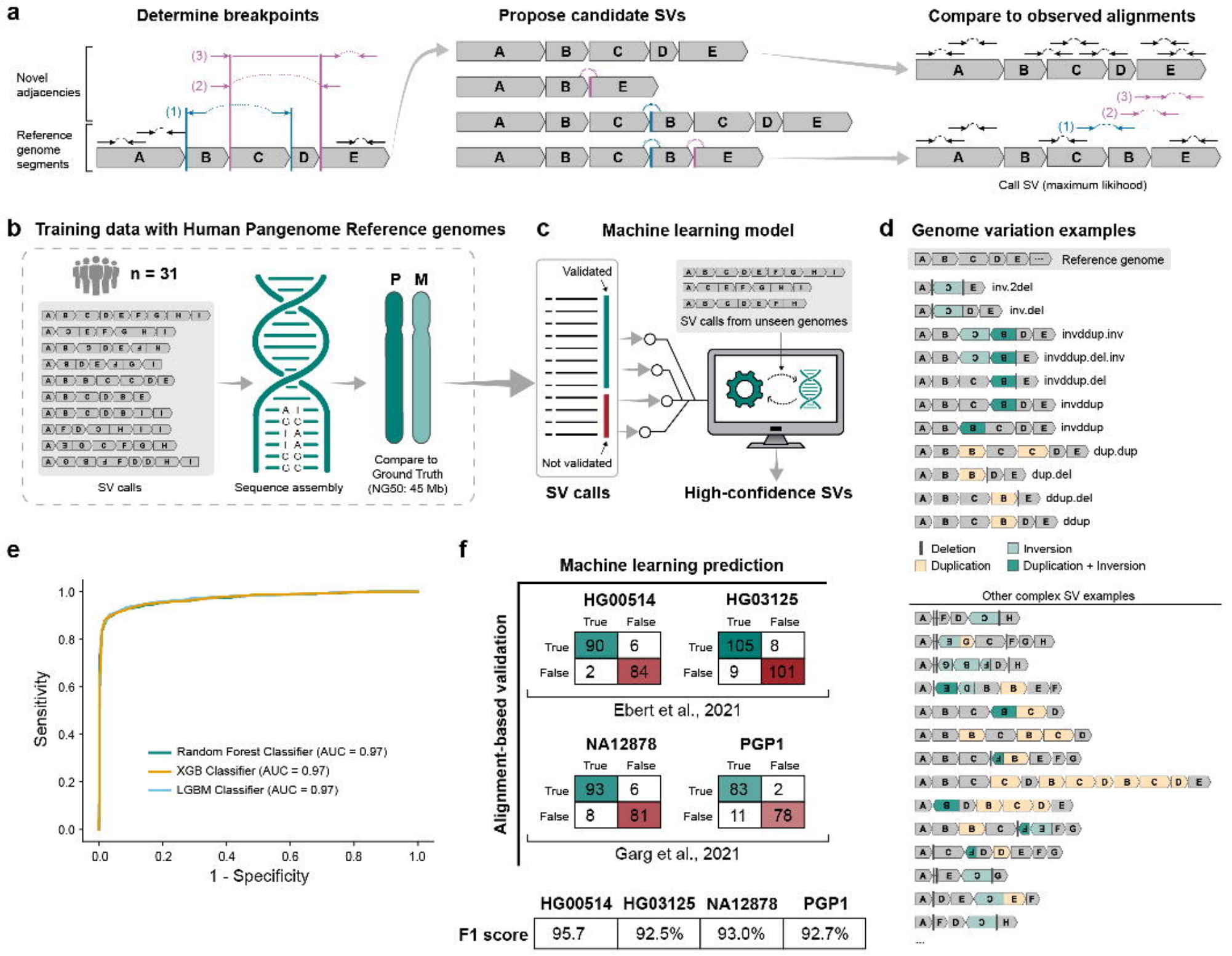
Figure 3. cxSVs across human populations. (a) Geographical distribution and ancestries of 4,262 human genomes studied. (b) Size distribution of 12 classes of cxSVs identified. (c) Total number of cxSVs by class. (d) Cumulative proportion of cxSVs against allele frequencies for different population. (e) Boxplot of number of cxSVs by superpopulation. (f) heatmap showing common cxSV across superpopulations where color scale indicate z-score of allele-frequency. (g). Scatterplot of top two principal components for common cxSVs.

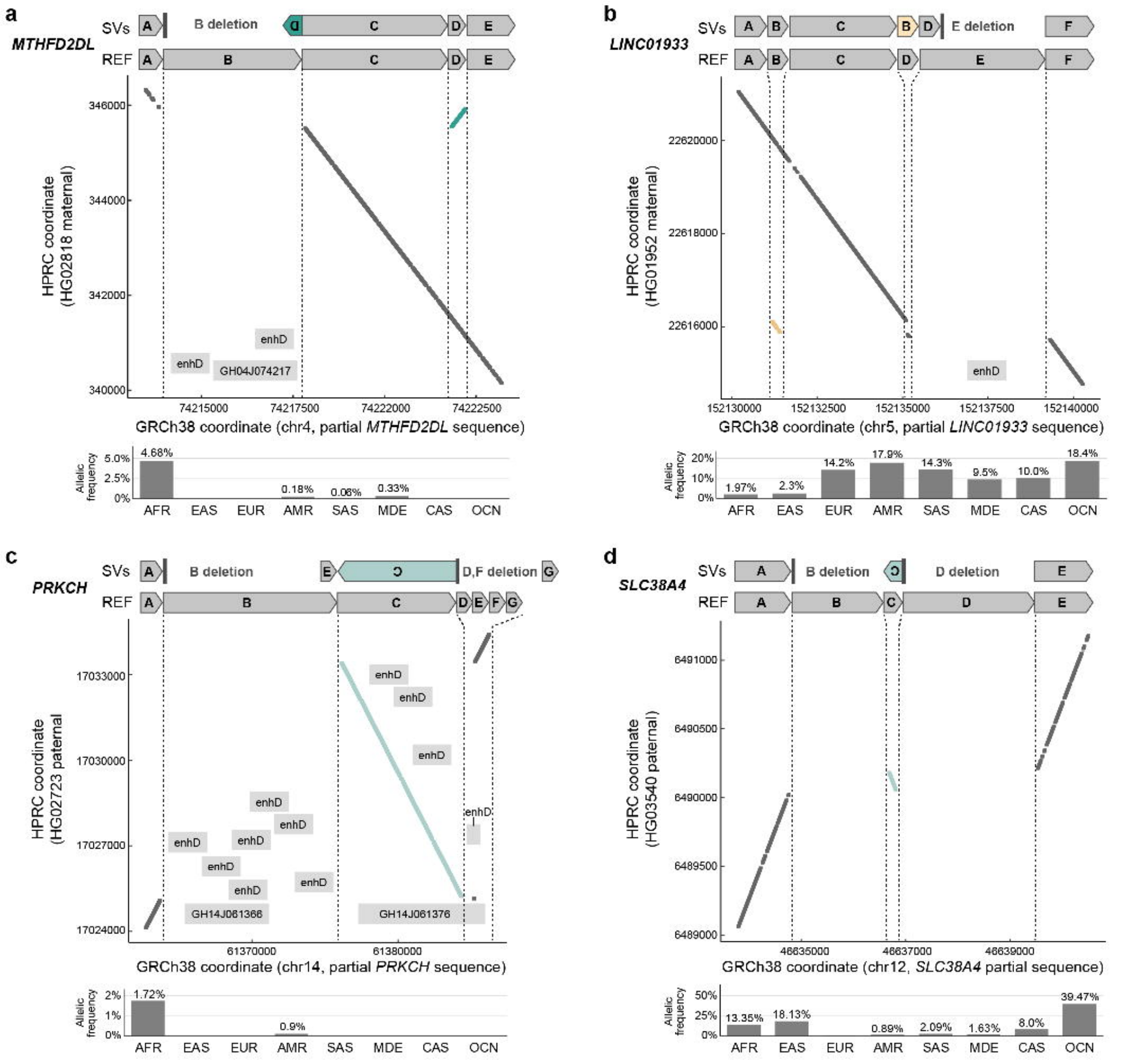
Figure 4. cxSV locations across the genome and their proximity to significant GWAS SNPs. (a) Yellow bars indicate location of cxSV along the chromosome. High density of cxSVs at sub-telomeric regions marked by red. Enrichment p-value and fold indicated by bar plot. (b) Scatter-plot of fraction of unique GWAS SNPs near cxSVs versus count of unique GWAS SNPs for trait; the blue dashed line indicates the fraction across all traits (46.4%). Color scale indicates probability density. (c) heritability enrichment of common SNPs near cxSVs; yellow line indicates nominal p-value=0.01; orange line indicates p-value=0.05 after Bonferroni correction. (d) Manhattan plot of phenotypic hotspots (each is significantly associated with ≥ 20 traits). Example of a phenotypic hotspot chr19:5831855-5859707. Each dot represents a GWAS SNP with the y-axis indicating the -log10 (p-value) from the GWAS.

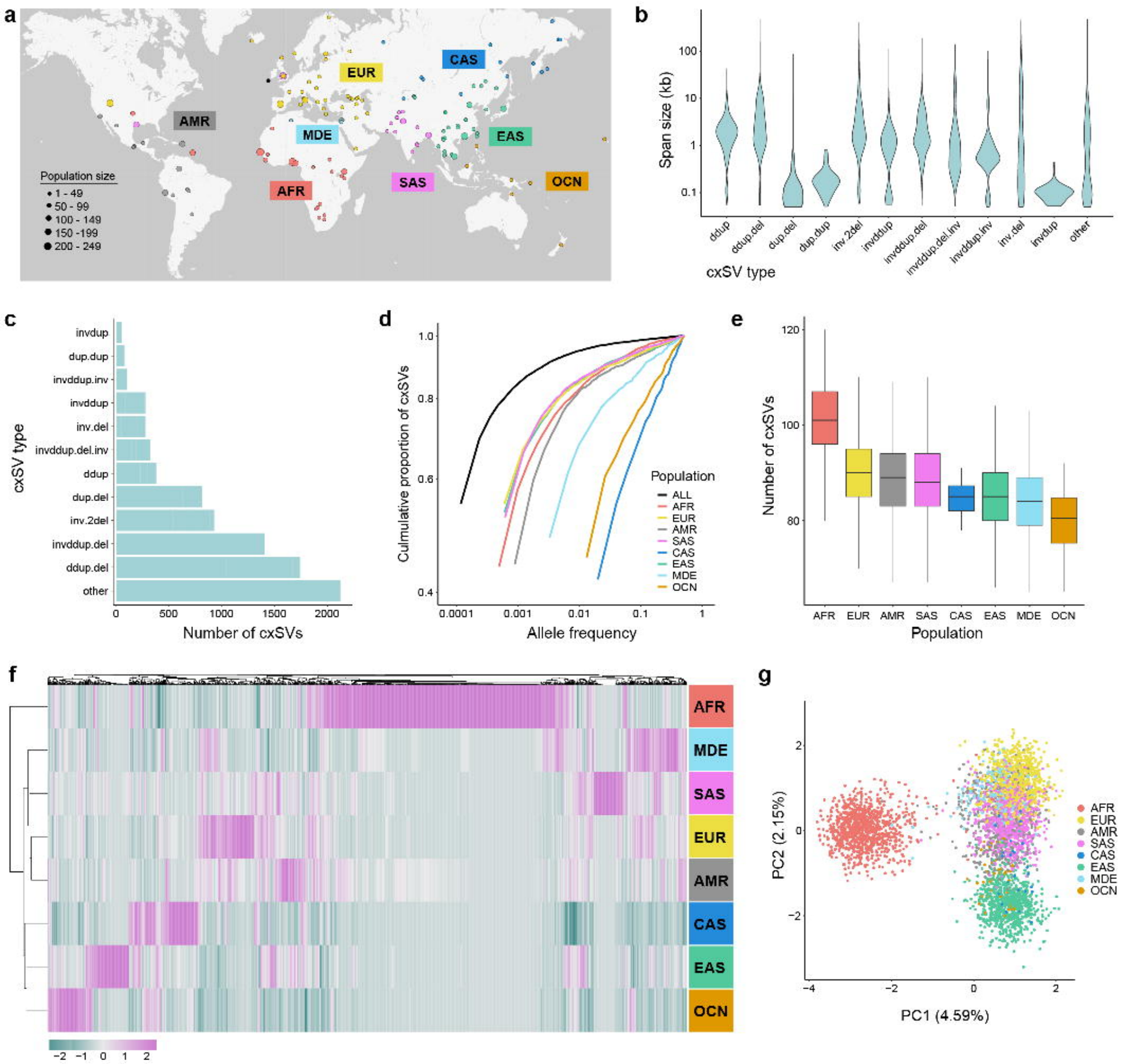
Figure 5. Enrichment and depletion of cxSVs in 127 Roadmap Reference Epigenomes across different chromatin states. Purple and green rectangles represent enrichment and depletion respectively, and deeper color represents smaller p-value. Significant enrichment or depletion (FDR <0.05) are marked by asterisks.

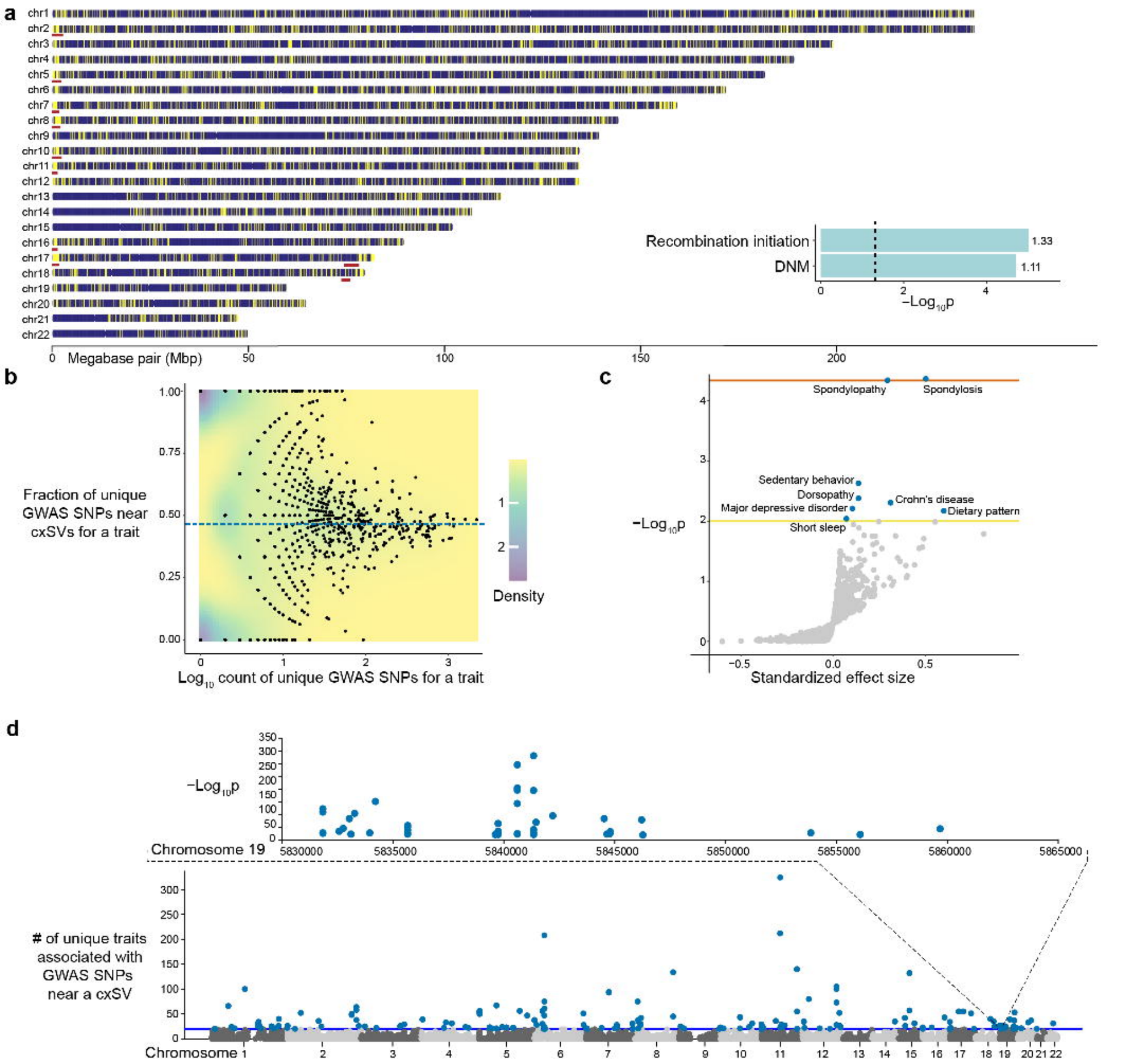
Figure 6. Enrichment of common and rare cxSVs in (a) "human-gained" enhancers/promoters, epigenetic gain-enriched hotspots, (b) protein-coding genes within epigenetic gain-enriched hotspots, (c-d) GO Cellular Component terms (with adjusted p-value), (e) HARs, and HAQERs.

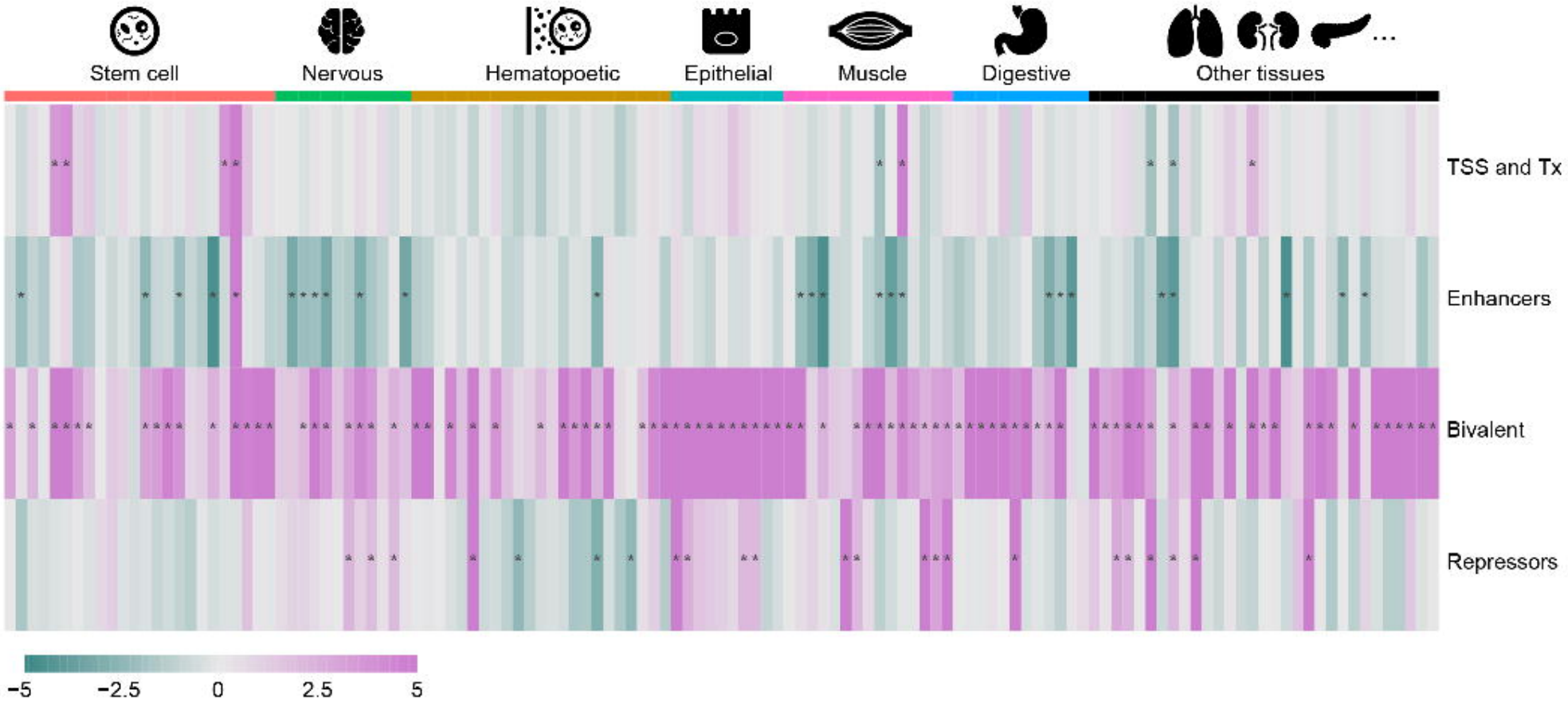
Figure 7. Examples of cxSVs at loci that recently diverged in humans. (a) cxSV identified in EAS only (0.18% AF) that affect HARs that are experimentally validated enhancers. Fixed loci in humans that diverged with bonobos and chimpanzees via a cxSV (with respect to hg38) (b) affecting HAR ANC302 within *LINC02109* and (c) within *LINC02609* that contain a human-gained brain enhancer and GWAS SNP rs4561025. (d) Fixed loci in humans that diverged with bonobos, chimpanzees, and gorillas affecting HAQER2045 and containing GWAS SNP rs1125472. (e,f) cxSVs identified in Neanderthal that are shared with humans within *PRMT2* and lincRNA *SAMMSON*. AF (%) indicated by bar-plot across superpopulations. ENCODE cCRE with proximal enhancer-like signature indicated by “enhP”.

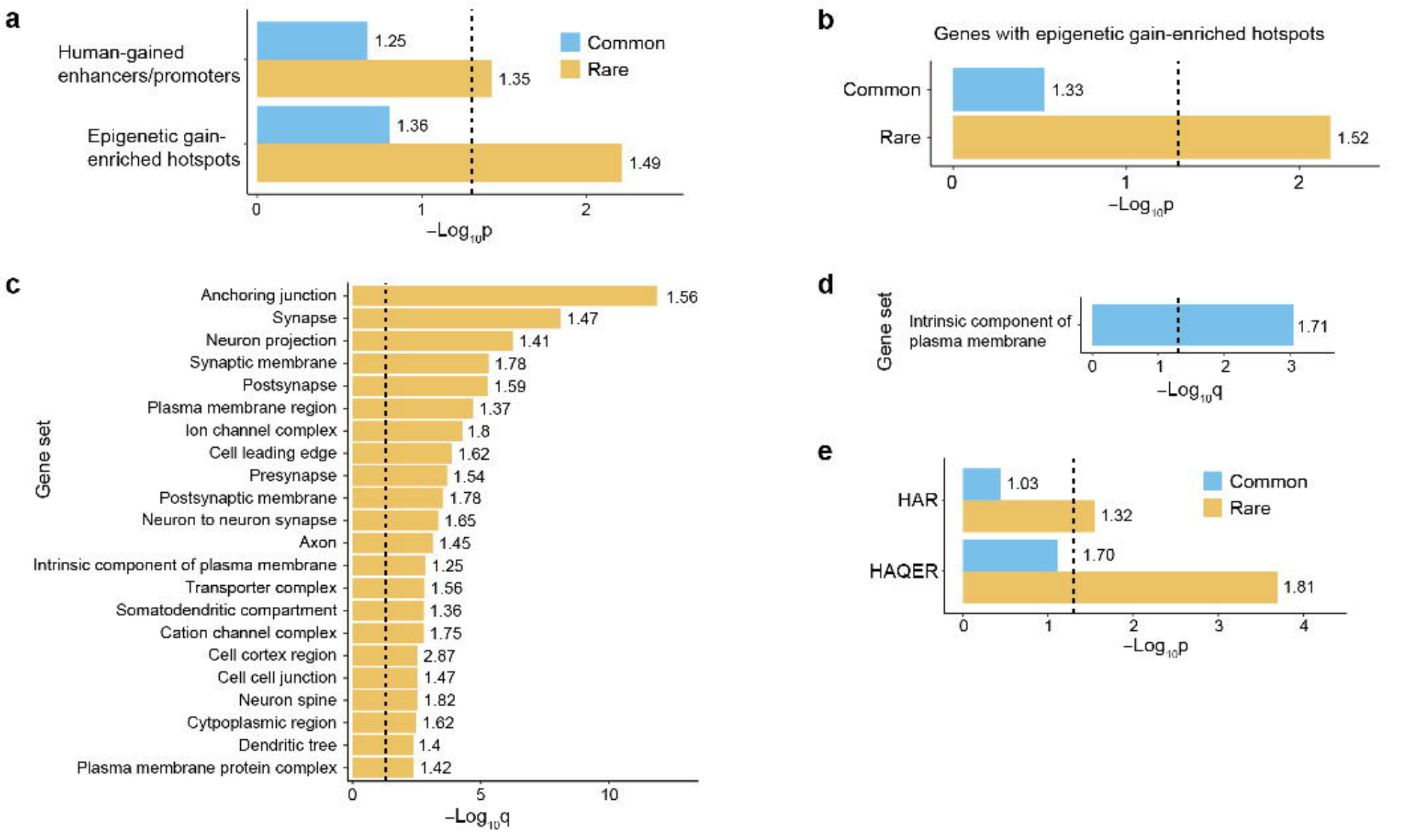

a
Determine breakpoints
Propose candidate SVs
Compare to observed alignments
Novel adjacencies
Reference genome segments
Call SV (maximum likihood)
b
Training data with Human Pangenome Reference genomes
n = 31
SV calls
Sequence assembly
P
M
Compare to Ground Truth (NG50: 45 Mb)
c
Machine learning model
Validated
Not validated
SV calls from unseen genomes
SV calls
High-confidence SVs
d
Genome variation examples
Reference genome
inv.2del
inv.del
invddup.inv
invddup.del.inv
invddup.del
invddup
invddup
dup.dup
dup.del
ddup.del
ddup
Deletion
Inversion
Duplication
Duplication + Inversion
Other complex SV examples
e
Sensitivity
1 - Specificity
Random Forest Classifier (AUC = 0.97)
XGB Classifier (AUC = 0.97)
LGBM Classifier (AUC = 0.97)
f
Machine learning prediction
Alignment-based validation
HG00514
True False
True 90 6
False 2 84
HG03125
True False
True 105 8
False 9 101
Ebert et al., 2021
NA12878
True False
True 93 6
False 8 81
PGP1
True False
True 83 2
False 11 78
Garg et al., 2021
HG00514 HG03125 NA12878 PGP1
F1 score 95.7 92.5% 93.0% 92.7%

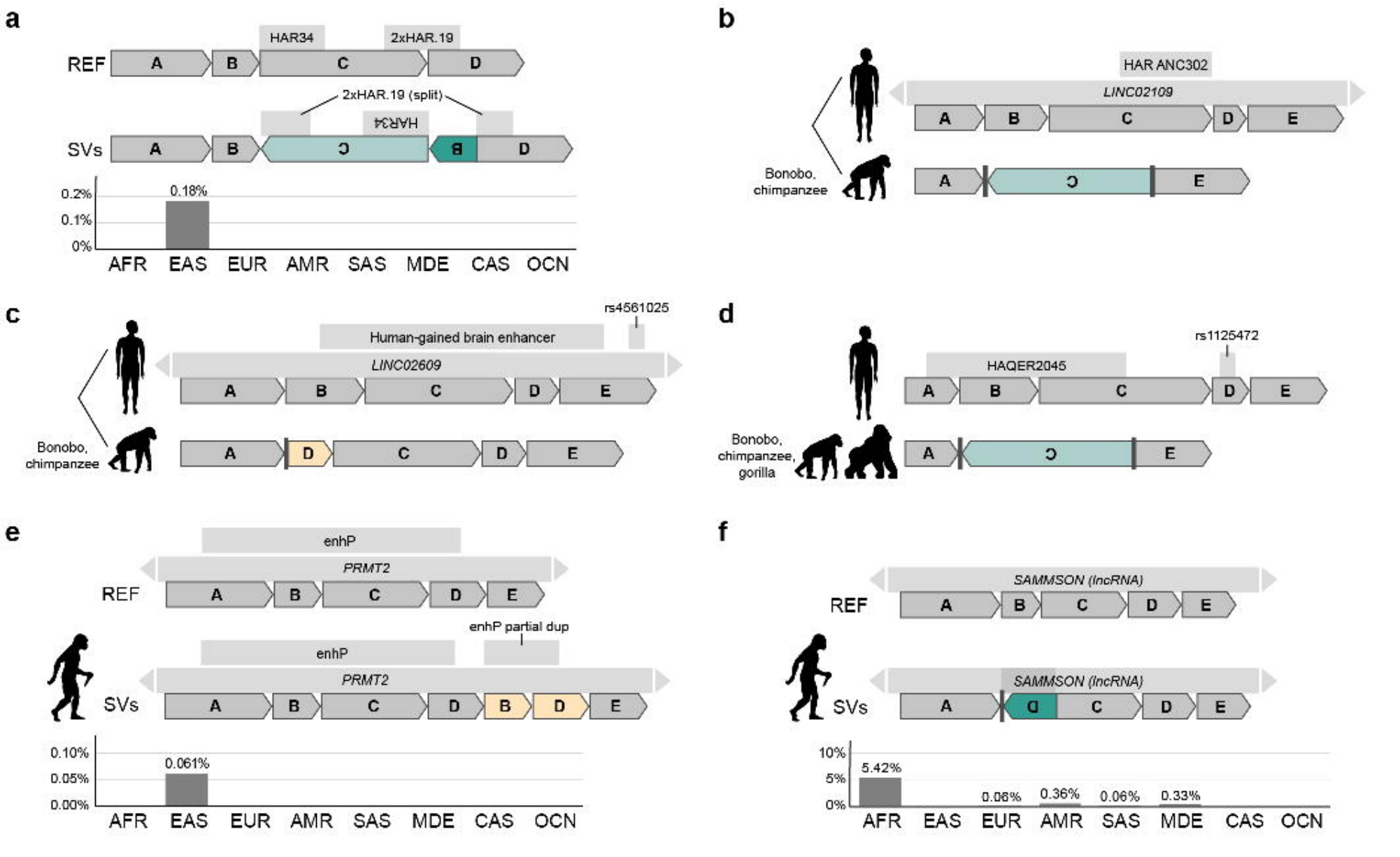

a
MTHFD2DL
SVs
A
B deletion
C
D
E
REF
A
B
C
D
E
HPRC coordinate (HG02818 maternal)
346000
344000
342000
340000
enhD
enhD
GH04J074217
74215000
74217500
74222000
74222500
GRCh38 coordinate (chr4, partial MTHFD2DL sequence)
Allelic frequency
5.0%
2.5%
0%
4.68%
0.18%
0.08%
0.33%
AFR
EAS
EUR
AMR
SAS
MDE
CAS
OCN
b
LINC01933
SVs
A
B
C
B
D
E deletion
F
REF
A
B
C
D
E
F
HPRC coordinate (HG01952 maternal)
22620000
22618000
22616000
enhD
152130000
152132500
152135000
152137500
152140000
GRCh38 coordinate (chr5, partial LINC01933 sequence)
Allelic frequency
20%
10%
0%
1.97%
2.3%
14.2%
17.9%
14.3%
9.5%
10.0%
18.4%
AFR
EAS
EUR
AMR
SAS
MDE
CAS
OCN
c
PRKCH
SVs
A
B deletion
E
C
D,F deletion
G
REF
A
B
C
D
E
F
G
HPRC coordinate (HG02723 paternal
17033000
17030000
17027000
17024000
enhD
enhD
enhD
enhD
enhD
enhD
enhD
enhD
enhD
enhD
GH14J061366
GH14J061376
61370000
61380000
GRCh38 coordinate (chr14, partial PRKCH sequence)
Allelic frequency
2%
1%
0%
1.72%
0.9%
AFR
EAS
EUR
AMR
SAS
MDE
CAS
OCN
d
SLC38A4
SVs
A
B deletion
C
D deletion
E
REF
A
B
C
D
E
HPRC coordinate (HG03540 paternal)
6491000
6490500
6490000
6489500
6489000
46635000
46637000
46639000
GRCh38 coordinate (chr12, SLC38A4 partial sequence)
Allelic frequency
50%
25%
0%
13.35%
18.13%
0.89%
2.09%
1.63%
8.0%
39.47%
AFR
EAS
EUR
AMR
SAS
MDE
CAS
OCN


a
EUR
CAS
AMR
MDE
EAS
AFR
SAS
OCN
Population size
1 - 49
50 - 99
100 - 149
150 -199
200 - 249
b
Span size (kb)
100
10
1
0.1
ddup
ddup.del
dup.del
dup.dup
inv.2del
invddup
invddup.del
invddup.del.inv
invddup.inv
inv.del
invdup
other
cxSV type
c
cxSV type
invdup
dup.dup
invddup.inv
invddup
inv.del
invddup.del.inv
ddup
dup.del
inv.2del
invddup.del
ddup.del
other
0
500
1000
1500
2000
Number of cxSVs
d
Culmulative proportion of cxSVs
1.0
0.8
0.6
0.4
0.0001
0.001
0.01
0.1
1
Allele frequency
Population
ALL
AFR
EUR
AMR
SAS
CAS
EAS
MDE
OCN
e
Number of cxSVs
120
100
80
AFR
EUR
AMR
SAS
CAS
EAS
MDE
OCN
Population
f
AFR
MDE
SAS
EUR
AMR
CAS
EAS
OCN
−2
−1
0
1
2
g
PC2 (2.15%)
2
0
−2
−4
−2
0
2
PC1 (4.59%)
AFR
EUR
AMR
SAS
CAS
EAS
MDE
OCN


a
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
0 Megabase pair (Mbp) 50 100 150 200
Recombination initiation
DNM
1.33
1.11
0 2 4
−Log10p
b
Fraction of unique GWAS SNPs near cxSVs for a trait
1.00
0.75
0.50
0.25
0.00
0 1 2 3
Log10 count of unique GWAS SNPs for a trait
1
2
Density
c
Spondylopathy
Spondylosis
Sedentary behavior
Dorsopathy
Major depressive disorder
Short sleep
Crohn's disease
Dietary pattern
−Log10p
4
3
2
1
0
−0.5 0.0 0.5
Standardized effect size
d
−Log10p
350
300
250
200
150
100
50
0
Chromosome 19
5830000 5835000 5840000 5845000 5850000 5855000 5860000 5865000
# of unique traits associated with GWAS SNPs near a cxSV
300
250
200
150
100
50
0
Chromosome 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22


Stem cell
Nervous
Hematopoetic
Epithelial
Muscle
Digestive
Other tissues
TSS and Tx
Enhancers
Bivalent
Repressors
−5
−2.5
0
2.5
5


a
Human-gained enhancers/promoters
Epigenetic gain-enriched hotspots
1.25
1.35
1.36
1.49
Common
Rare
0
1
2
−Log10p
b
Genes with epigenetic gain-enriched hotspots
Common
Rare
1.33
1.52
0
1
2
−Log10p
c
Gene set
Anchoring junction 1.56
Synapse 1.47
Neuron projection 1.41
Synaptic membrane 1.78
Postsynapse 1.59
Plasma membrane region 1.37
Ion channel complex 1.8
Cell leading edge 1.62
Presynapse 1.54
Postsynaptic membrane 1.78
Neuron to neuron synapse 1.65
Axon 1.45
Intrinsic component of plasma membrane 1.25
Transporter complex 1.56
Somatodendritic compartment 1.36
Cation channel complex 1.75
Cell cortex region 2.87
Cell cell junction 1.47
Neuron spine 1.82
Cytpoplasmic region 1.62
Dendritic tree 1.4
Plasma membrane protein complex 1.42
0
5
10
−Log10q
d
Gene set
Intrinsic component of plasma membrane
1.71
0
1
2
3
−Log10q
e
HAR
HAQER
1.03
1.32
1.70
1.81
Common
Rare
0
1
2
3
4
−Log10p


a
HAR34
2xHAR.19
REF
A
B
C
D
2xHAR.19 (split)
HAR34
SVs
A
B
C
B
D
0.2%
0.1%
0%
0.18%
AFR
EAS
EUR
AMR
SAS
MDE
CAS
OCN
b
HAR ANC302
LINC02109
A
B
C
D
E
Bonobo, chimpanzee
A
C
E
c
rs4561025
Human-gained brain enhancer
LINC02609
A
B
C
D
E
Bonobo, chimpanzee
A
D
C
D
E
d
rs1125472
HAQER2045
A
B
C
D
E
Bonobo, chimpanzee, gorilla
A
C
E
e
enhP
PRMT2
REF
A
B
C
D
E
enhP partial dup
enhP
PRMT2
SVs
A
B
C
D
B
D
E
0.10%
0.05%
0.00%
0.061%
AFR
EAS
EUR
AMR
SAS
MDE
CAS
OCN
f
SAMMSON (lncRNA)
REF
A
B
C
D
E
SAMMSON (lncRNA)
SVs
A
C
C
D
E
10%
5%
0%
5.42%
0.06%
0.36%
0.06%
0.33%
AFR
EAS
EUR
AMR
SAS
MDE
CAS
OCN