Similarity-based analysis of allele frequency distribution among multiple populations identifies adaptive genomic structural variants

Authors SAITOU, MARIE.[1,2,3], MASUDA, NAOKI,[4,5], GOKCUMEN, OMER.[1]

Affiliations:

[1]Dept. of Biological Sciences,

University at Buffalo, State University of New York, Buffalo, NY 14260-2900, USA

[2] Currently at the Faculty of Biosciences,

Norwegian University of Life Sciences, Universitetstunet 3, 1430 Ås, Norway.

[3] Dept. of Medicine,

The University of Chicago. Section of Genetic Medicine, 5841 S. Maryland Ave., Chicago, IL, 60637-1447, USA.

[4]Department of Mathematics,

University at Buffalo, State University of New York, Buffalo, NY 14260-2900, USA

[5]Computational and Data-Enabled Science and Engineering Program,

University at Buffalo, State University of New York, Buffalo, NY 14260-5030, USA

Correspondence:

O.G. omergokc@buffalo.edu

N.M. naokimas@buffalo.edu

Abstract

Structural variants have a considerable impact on human genomic diversity. However, their evolutionary history remains mostly unexplored. Here, we developed a new method to identify potentially adaptive structural variants based on a similarity-based analysis that incorporates genotype frequency data from 26 populations simultaneously. Using this method, we analyzed 57,629 structural variants and identified 576 structural variants that show unusual population differentiation. Of these putatively adaptive structural variants, we further showed that 24 variants are multiallelic and overlap with coding sequences, and 20 variants are significantly associated with GWAS traits. Closer inspection of the haplotypic variation associated with these putatively adaptive and functional structural variants reveals deviations from neutral expectations due to (i) population differentiation of rapidly evolving multi-allelic variants, (ii) incomplete sweeps, and (iii) recent population-specific negative selection. Overall, our study provides new methodological insights, documents hundreds of putatively adaptive variants, and introduces evolutionary models that may better explain the complex evolution of structural variants.

Introduction

Emerging technologies have recently revealed hundreds of thousands of genomic structural variants, including polymorphic duplications, deletions, inversions, and mobile transposable elements in the human genome (Hurles et al. 2008; Conrad et al. 2010; Pang et al. 2010; Mukamel et al. 2021). Unlike single nucleotide variants, each structural variant affects a continuous block in the genome and thus is more likely to result in a phenotypic effect (Hurles et al. 2008; Weischenfeldt et al. 2013; Sudmant, Rausch, et al. 2015). Several structural variants have been documented to have considerable effects on human disease and evolution (Dennis & Eichler 2016; Payer et al. 2017; Hsieh et al. 2019; Ho et al. 2020; Mukamel et al. 2021). Some of these functional variants reach >20% allele frequency in human populations, and some affect the copy number variation of entire protein-coding genes (McCarroll et al. 2005; Handsaker et al. 2015).

The poster child for adaptive structural variation in humans is the copy number variation of the amylase gene. Several studies put forward evidence for positive selection of higher amylase gene copy numbers in the human lineage, and further in high-starch-consuming human populations (Perry et al. 2007). Another striking example of potentially adaptive structural variants is the deletion of *LCE3B* and *LCE3C.* This variant is one of the leading susceptibility markers to psoriasis (de Cid et al. 2009). This deletion was shown to be retained in the human lineage since Human-Neanderthal divergence under balancing selection (Pajic et al. 2016), arguably maintaining a balance between protection against pathogens and facilitating immune-mediated disorders. Recently, a genome-wide analysis identified several Neanderthal- and Denisovan-introgressed structural variants that show strong signatures of adaptation (Hsieh et al. 2019; Yan et al. 2021). Collectively, these studies, along with others (see (Saitou & Gokcumen 2020) for a detailed review), imply that several common structural variants contribute to human phenotypic variation and may have evolved under diverse adaptive scenarios.

Despite the increasing appreciation of their role in human adaptive evolution, structural variants have not been scrutinized as much as single nucleotide variants due to technical difficulties. From a methodological perspective, structural variants are more challenging to discover and genotype due to their localization in highly repetitive sections of the genome. In addition, they are generated through complex mutational mechanisms, often involving gene conversions and unequal recombination events (Kidd et al. 2010; Handsaker et al. 2011; Lupski 2015; Carvalho & Lupski 2016; Sekar et al. 2016). As a result, structural variants are usually harbored by complex haplotypes, and they often are not tagged perfectly with flanking variants (Sudmant, Mallick, et al. 2015). The complexity of haplotypic architecture harboring structural variants complicates analyses of neutrality and integration of structural variants to genome-wide association studies. Most studies primarily interrogate single nucleotide variants and structural variants are often considered if only a “tag” single nucleotide variant can be found. For example, our work has resolved the complex evolutionary history of the common deletion of the metabolizing *GSTM1* gene (Saitou, Satta & Gokcumen 2018). Locus-specific studies showed a strong association between this deletion and bladder cancer (The *GSTM1* deletion is the risk allele, $p = 4\times10^{-11}$)

(Rothman et al. 2010). A haplotype-based analysis of the locus suggested that this deletion has formed multiple times through independent mutation events and undergone gene conversion events (Saitou, Satta, Gokcumen, et al. 2018). Thus, due to the lack of single nucleotide variants tagging this deletion, most genome-wide association studies and traditional selection scans did not include this deletion. Investigating individual haplotypes that harbor the deletion led us to identify one particular haplotype associated with the deletion that has been subject to a recent selective sweep in the East Asian populations (Saitou, Satta & Gokcumen 2018).

A second factor that complicates the evolutionary study of structural variants is that some are multiallelic (Quinlan & Hall 2012; Handsaker et al. 2015). For example, the haptoglobin locus harbors two large multiallelic and recurrent structural variants that are not tagged by any single nucleotide variant. Only after careful, locus-specific resolution of haplotypic variation were they shown to be associated with cholesterol levels (Boettger et al. 2016). Similarly, *AMY1* (Perry et al. 2007), as we noted above, and *DMBT1 (Polley et al. 2015)* loci harbor multiallelic structural variations that were associated with dietary and metabolic traits. However, even for amylase gene copy number variation, arguably the best-studied structural variant in the human genome from an evolutionary perspective, the timing and existence of putative adaptive forces remain elusive (Mathieson & Mathieson 2018). In fact, structural variants are often consciously left out from most selection scans along with segmental duplications and other repetitive regions due to the complications that we described above (e.g.**,** (Schrider & Kern 2017)). In sum, we argue that the full impact of structural variants on human evolution has not been understood and may explain some of the most exciting, yet to be described, adaptive variation in humans.

Given the complexity of haplotypes that harbor a considerable number of structural variants, measures that depend on accurate genotyping of haplotypic variation, such as allele frequency spectra (e.g., Tajima's D (Tajima 1993)) or linkage-disequilibrium/homozygosity (e.g., iHS (Voight et al. 2006), XP-EHH (Sabeti et al. 2007)) are often underpowered. Instead, direct population differentiation metrics may be the most appropriate and unbiased way to identify putatively adaptive structural variants among human populations. Population differentiation-based methods are robust to haplotype disruption due to gene conversion, recurrence, or the presence of multiple alleles. Most studies that identify adaptive structural variants have employed population-differentiation-based methods (Redon et al. 2006; Xue et al. 2008; Sudmant, Mallick, et al. 2015; Almarri et al. 2020; Bergström et al. 2020). Deviations from expected allele frequency distribution can provide information on several types of selection (positive, negative, or stabilizing), and differential selection with complex histories (selection on standing variation, recent geography-specific negative selection, oscillating selective forces such as dynamic environmental change (Key et al. 2014)). This is important because it has been shown that "classical" sweeps were rare (Hernandez et al. 2011) and selection on standing variants are likely to be the major force of human genomic adaptation (Schrider & Kern 2017), as recently shown for multiple alleles shaping skin color (Crawford et al. 2017; Martin et al. 2017).

To measure the population differentiation of genetic variants, F_{ST} statistics (Weir & Cockerham 1984) and V_{ST} statistics for copy number variants (Redon et al. 2006) are commonly used. More recent research has developed methods to compare multiple populations, primarily for admixture analysis (*e.g.*, F_3 statistics (Reich et al. 2009)). However, these methods can only compare two or three populations to each other. Recently, (Duforet-Frebourg et al. 2016) developed a PCA-based method to identify single nucleotide variants with population differentiation by analyzing 10 populations simultaneously, confirming well-known targets for positive selection, and discovered new candidate genes. Here, we developed a new, similarity-based method to identify adaptive structural variants with unusual allele frequency distribution with which one can analyze (1) multiallelic variants and (2) the distribution of genotype frequency in multiple populations collectively.

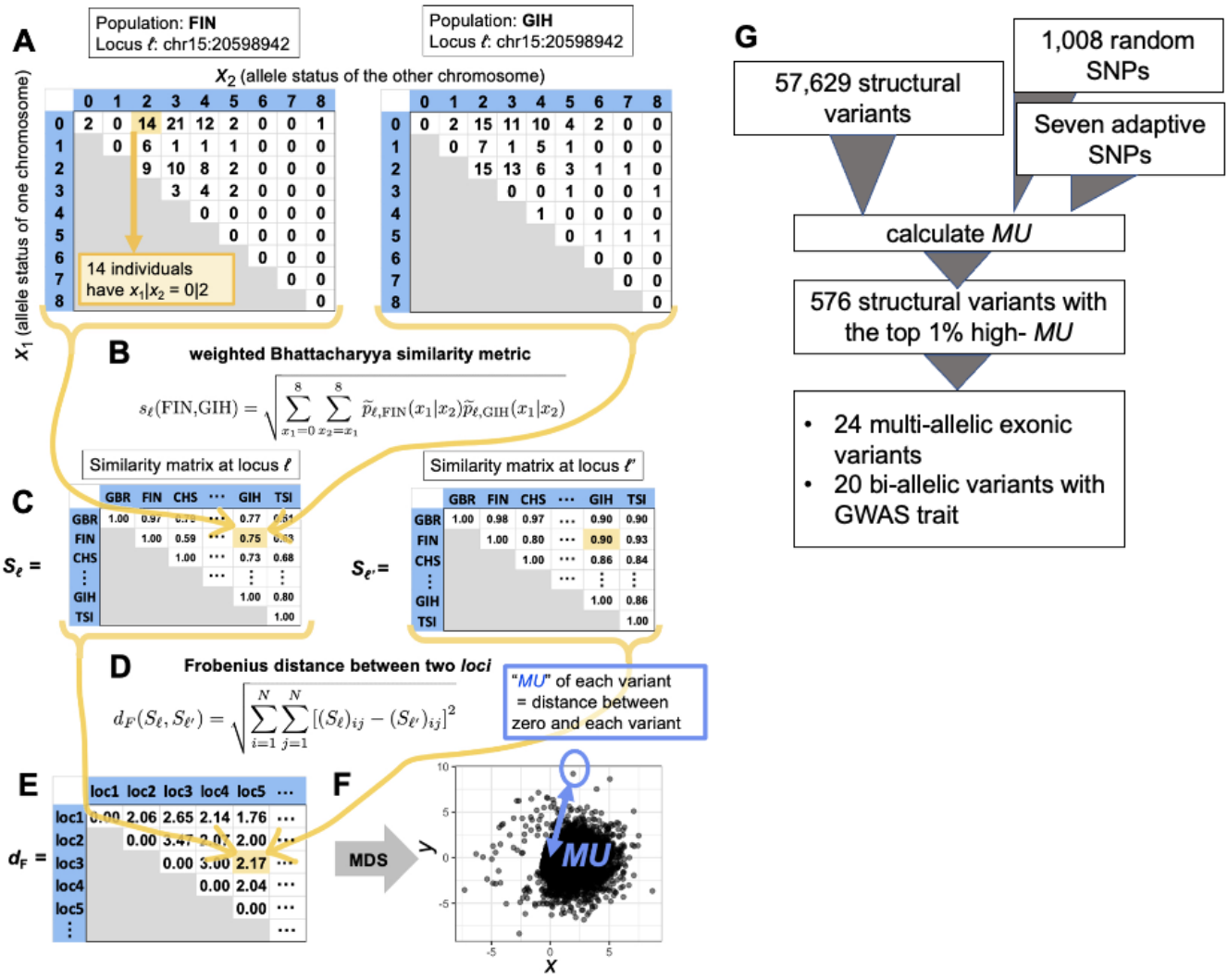


Figure 1. An overview of the calculation of "*Measure of Unusualness (MU)*"(see Methods for details). **A.** For each population and each locus, we have a 9 x 9 matrix representing the genotype. The row and column of the matrix represent one of the two chromosomes each. The cells contain the frequencies of specific genotype combinations. **B.** We calculate the similarity in the 9 x 9 matrices exemplified in panel A between each pair of populations. The similarity value ranges between 0 and 1. **C.** In this manner, for each locus, we obtain a 26 x 26 matrix representing the similarity between different pairs of populations. The diagonal entries of the similarity matrix are equal to 1 because any population is identical to itself, yielding the largest possible value of the similarity, which is 1. **D.** We calculate the distance between the 26 x 26 matrices for each pair of loci. **E.** In this manner, we obtain a distance matrix representing the distance between the different pairs of loci. **F.** We carry out the MDS to project the obtained distance matrix into the two-dimensional embedding space. Each circle represents a locus. The distance between the locus and the origin in the embedding space defines *MU* for each locus. **G.**The general results from our pipeline.

Results and Discussion

Structural Variants with Unusual Population Differentiation

Inspired by the emerging work that integrates all available population differentiation information to understand demographic and adaptive trends (Duforet-Frebourg et al. 2016), we developed a new method based on the Bhattacharyya similarity metric specifically to identify putatively adaptive outliers among structural variants ((Bhattacharyya 1943)**, Methods, fig.1A-F, fig S1**).

Briefly, we characterize each locus (ℓ) as an $N \times N$ similarity matrix S_ℓ based on the genotype frequency of the $N = 26$ populations in the 1000 Genomes Project phase 3 data set (Sudmant et al. 2015). We measure a modified Bhattacharyya similarity metric between each pair of populations based on the transformed probability distribution (for the original Bhattacharyya metric, see (Bhattacharyya 1943; Cha & Srihari 2002)). To increase the sensitivity to identify the population differentiation of variants with many alleles, we use a weighted variant of the Bhattacharyya similarity metric. Overall, we analyzed M = 58,644 variants, including 57,629 structural variants, 1,008 uniformly randomly chosen single nucleotide polymorphisms, as well as 7 single nucleotide polymorphisms (SNPs) that were reported to be under adaptive evolution (6 SNPs under positive selection and one SNP under balancing selection) (Norton et al. 2007; Kimura et al. 2009; Mou et al. 2008; Smith et al. 2009; Ding et al. 2013; Ko et al. 2013; Basu Mallick et al. 2013; Wilde et al. 2014; Wu et al. 2016; Deng & Xu 2018). We constructed a distance matrix for each of the M loci and compared the similarity matrix S_ℓ across all these loci ($\ell = 1, ..., M$). We define the distance between S_ℓ (at one locus) and $S_{\ell'}$ (at another locus) by the Frobenius norm, denoted by d_F. The $M \times M$ Frobenius distance matrix, denoted by F, tabulates the difference between each pair of loci, and its (ℓ, ℓ') entry is given by $d_F(S_\ell, S_{\ell'})$. These steps provided us with a matrix indicating how structural variants relate to each other based on their global genotype frequency distribution.

We assumed, based on previous literature (Conrad et al. 2010), that the majority of structural variants will be evolving under neutrality or near neutrality. Therefore, population differentiation should primarily be driven by genetic drift. We then reasoned that structural variants that have unusual allele frequency distribution among the 26 populations as compared to the genome-wide observations are likely to have evolved under non-neutral conditions. To visualize the relationships between structural variants based on their global allele frequency distribution, we ran a multidimensional scaling (MDS) algorithm. To empirically measure these relationships, we calculated the distance between the origin and each variant in the MDS space and defined it as "*Measure of Unusualness (MU)*", or degree of the unusual allele frequency distribution (**Methods, fig. 1F**). This measure informs on the unusualness of global population differentiation of a given structural variant, as compared to the entirety of the data set.

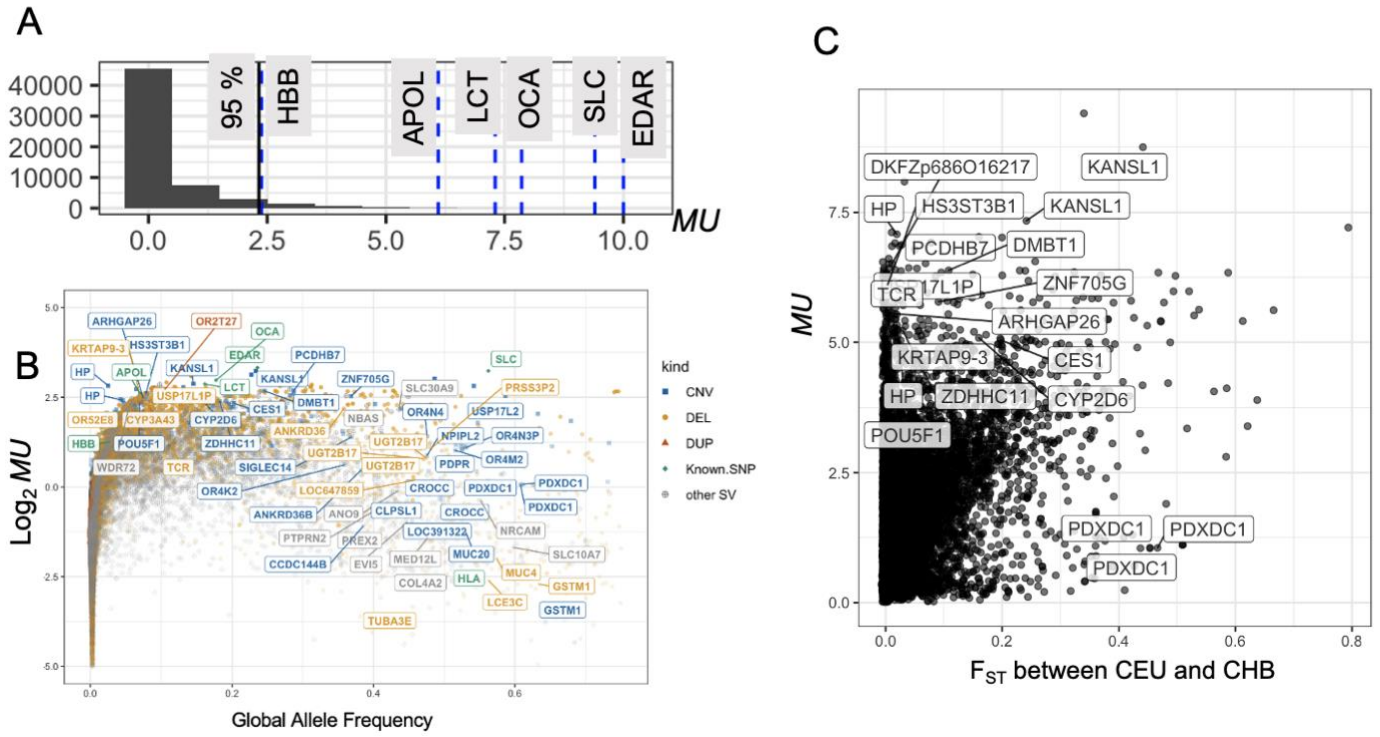


Figure 2. An overview of the study pipeline and results. A. The histogram of *MU* of the 1,008 randomly selected SNPs and the six positively selected SNPs that were found to be under selection in previous studies (see **Methods**). The latter group was indicated by blue vertical dashed lines on the histogram, and the genes affected by these variants were labeled. The 95th percentile of the distribution was marked by a black vertical line. **B.** The relationship between allele frequency and *MU*. The horizontal axis indicates the global alternative allele frequency. The vertical axis indicates the logarithm of the *MU* value. The exonic structural variants with *MU* > 5 or global allele frequency > 0.25 are labeled, as well as the six positively selected SNPs" and one HLA SNP which was shown to have evolved under balancing selection. Some gene names are shown multiple times (e.g., *HP*, *KANSP1*, and *PDXDC1*); this happens because multiple structural variants overlapping these genes were reported in the 1000 Genomes Project Phase 3 data set. Colors represent different types of variants. The abbreviation is from the 1000 Genome Project phase 3 structural variants data set., CNV; copy number variants (multiallelic variants), DEL: deletion, DUP: duplication, Known. SNP: SNPs from previous studies (see Methods), other SVs (insertion, inversion, Alu, Long interspersed nuclear element, SINE-VNTR retrotransposons). **C.** Comparison of F_{ST} (Weir & Cockerham 1984) between CEU and CHB populations and *MU*. Bi-allelic structural variants with F_{ST} (between CEU and CHB) >0.4 and *MU* > 5 were labeled. The shade in blue represents the density of the structural variants (see **fig. S4**).

Simulation and Empirical Confirmation

To validate the accuracy and sensitivity, we conducted forward simulations using SLiM 3.6 (Messer 2013; Haller & Messer 2019) (**Methods**). We modeled stepwise copy number gains or losses in each locus under different mutation rates/selection coefficients in three populations (YRI, CEU, and CHB) for which the demographic parameters were previously established (Gravel et al. 2011). In each simulation, we generated 2970 potentially variable neutral loci, and 10 x 3 potentially variable loci under population-specific selection in each population (with a range of selection coefficients), respectively. We used these data to calculate *MU* and assess the accuracy and sensitivity of our approach (**fig. S2**). Our results suggest that our approach is unable to distinguish between drift and selection if the mutation rate is higher than 10^{-7} mutations per locus per generation. Further, we found that the population for which the selection is acting is an important parameter in determining the power and accuracy of *MU*. Specifically, we found that if the selection is acting on the YRI

population, our power to detect selection increases, possibly because the effect of drift is lower in African populations due to their higher effective population sizes (Tenesa et al. 2007). Although these simulations are useful in the general assessment of our approach, they have two major limitations. First, we were not able to simulate *MU* for all 26 populations because the demographic parameters for most of these populations were not established. Thus, it is likely that *MU* is more powerful when applied to a larger number of populations. Second, the mutation rates of structural variants are highly variable (Lin & Gokcumen 2019). Thus, without having a better sense of the mutation rates of structural variants, it is difficult to assess the power of *MU* for the whole range of structural variants in the human genome. Thus, we argue that an empirical comparison to known variants with well-established population-selection signatures may be currently a better benchmark than simulation-based methods for understanding the power of *MU*.

To assess the accuracy of our approach empirically, we calculated MU for 1,008 random single nucleotide polymorphisms (**tableS1, Methods**). We reasoned that the control SNPs will provide an additional marker set whereby differentiation is determined by near-neutral forces. Using this data set, we confirmed that the distribution of *MU* measured for structural variants is not significantly different from that measured for the uniformly randomly chosen SNPs from the 1000 Genome Phase 3 data set (Sudmant, Rausch, et al. 2015) (p = 0.88, Mann-Whitney test, **fig. S3A**), suggesting that similar to single nucleotide variants, the overall distribution of *MU* for SVs is neutral-like. We replicated this analysis using 5000 neutral SNPs reported in (Pouyet et al. 2018) (**fig. S3B**) and found similar results (p = 0.41, Mann-Whitney test, **fig. S3C**).

Next, we investigated the sensitivity of our method by measuring *MU* from six SNPs that have repeatedly been reported to be under population-specific positive selection and their functional relevance was well established (**Methods**). Thus, they provide an appropriate gold standard to test the sensitivity of our method. We found that all of these six positively selected SNPs were shown to be in the top 5% of the *MU* distribution (**fig. 2A**), improving our confidence in our method. In addition to those 6 SNPs, we also included in our analysis rs1129740, which resides in the *HLA* locus and has been one of the handfuls of variants in humans that are thought to have evolved under long-term balancing selection (Teixeira et al. 2015). This allowed us to observe how *MU* behaves for such variants even though MU was not designed to test balancing selection. We found that this non-synonymous mutation shows low *MU* values when considering its allele frequency (**fig. 2B**). We found it noteworthy that a structural variant, *LCE3BC* gene deletion, that we speculated previously to have evolved under balancing selection (Pajic et al. 2016) shows similarly low *MU* values despite their high allele frequency (**fig. 2B**). Thus, it is plausible that high allele frequency structural variants that may have been evolving under balancing selection may exhibit unusually low *MU* values (**table S2**).

To understand the differences between more traditional methods of measuring population differentiation and our method, we compared *MU* with direct allele-frequency-based F_{ST} between representative continental populations

(**fig. 2C, fig. S4**). As expected, we found a significant correlation between these two measures (Spearman rank correlation coefficient > 0.49 and $p < 10^{-15}$ for all comparisons). However, we also found notable discrepancies. We noted 125 variants that are in the top 1^{ST} percentile for *MU*, but show $F_{ST} < 0.2$ in any pairwise comparison of European (CEU), East Asian (CHB), and African (YRI) populations (**table S3**). Closer inspection of these variants suggests that *MU* captures multi-allelic variations and within-continent variation, which fell below the detection threshold of standard pairwise population comparisons. In addition, we noted structural variants that have large values of F_{ST} but do not stand out in terms of *MU*. When we investigated the allele frequency distribution of structural variants with high F_{ST} (>0.25) but low *MU* (<1), we consistently observed a clinal distribution of allele frequencies across the continents (**fig. S5**), likely due to serial founder effects that define human genetic variation as described previously (Ramachandran et al. 2005). In other words, we argue that variants that have high among-continental differences that the standard F_{ST} detects often reflect the effects of major bottleneck/drift events, such as out-of-Africa migrations, rather than adaptive sweeps. That suggests that a gradual population differentiation may not lead to a high MU value in our measure. Instead, our method is sensitive to deviations from such expected clinal allele frequency changes, including unusually low or high allele frequency in a single population as compared to its neighbors. This is an advantage of our method because local ecological and cultural variation often underlies adaptive evolution in humans (Rees et al. 2020). Thus, our method shows promise in capturing hundreds of novel putatively adaptive variants that have not been captured by traditional SNP-based pairwise population comparisons.

MU Identifies Dozens of Structural Variants Invisible to Traditional Selection Scans

There are several outstanding questions concerning the enrichment of specific properties of adaptive structural variants, including their functional relevance, the mutation mechanisms through which the variants are generated, and their size distribution. However, there are tremendous technical biases inherent in the short-read sequencing-based characterization of these variants, especially concerning extremely high false-negative rates in the discovery of certain types of structural variants, such as tandem and dispersed duplications and inversions (Kronenberg et al. 2015). Thus, instead of searching for general trends in our data set (e.g., adaptively evolving structural variants are larger or smaller than neutrally evolving ones), we focused on resolving the evolutionary forces shaping individual structural variants with functional implications.

In this spirit, we first investigated structural variants that overlap with coding sequences. We identified 39 structural variants with the top 1% *MU* value that contain one or more entire exon (**fig. 3A, table S4**). Regardless, many of these exonic structural variants were associated with metabolic traits and diseases in previous locus-specific analyses and include members of cytochrome p450 (*CYP3A43*, *CYP2D6*), solute carrier (*SLC30A9*, *SLC51A*), olfactory receptor (OR2T27, OR52E8) gene families. For example, *DMBT1* gene copy number was noted for its population differentiation and associated with dietary subsistence strategies (Polley et al. 2015).

Similarly, the copy number variation affecting the *CES1 (Zhu & Markowitz 2013), CYP2D6 (Candiotti et al. 2005), HS3ST3B1 (Kim et al. 2010),* and *SULT1 (Hebbring et al. 2008)* are associated with differences in metabolizing of xenobiotic substances, primarily described within a pharmacogenomics context. Interestingly, we found that 24 (~65%) of these exonic structural variants are multiallelic (**fig. 3B, table1),** more than five times higher than genome-wide expectations (p = 0.0005, Chi-square test). We found that intervals that overlap with multiallelic SVs are enriched for "defense response to Gram-negative bacterium" function (FDR Q value = 1.09×10^{-3}), concordant with previous literature linking adaptive structural variants with immune-related functions.

The exonic structural variants with the highest *MU* have been invisible to previous genome-wide association studies and selection scans. We argue that this is primarily because the current GWAS pipelines interrogate single nucleotide variants. Single nucleotide variants may not tag multiallelic structural variants due to gene conversion, recurrence, and potential genotyping errors, as discussed in the introduction. This phenomenon was diligently dissected by Boettger et al. (Boettger et al. 2016) for the haptoglobin (HP) locus, which harbors two recurrent and multiallelic exonic copy number variants that we found to show unusually high *MU*. They described a novel way to use a combination of single nucleotide variants to impute these structural variants in the locus. Their reanalysis of the genome-wide association studies revealed a previously hidden association between copy number variation affecting HP gene function and blood cholesterol levels. Based on our results, we argue that the effects of multiallelic structural variants on human evolution and phenotypic variation remain underappreciated.

Multiallelic SVs are often genotyped inaccurately (Mills et al. 2011). If such inaccuracies are systemic to given populations, it may lead to errors in identifying spurious genome-wide signals pertaining to population differentiation (Anderson-Trocmé et al. 2020). Thus, some of our observations of unusual allele frequency distribution may be due to such batch effects and genotyping errors inherent in the 1000 Genomes Phase 3 dataset. To address this issue, we used mrCaNaVaR (https://github.com/BilkentCompGen/mrcanavar) to estimate the copy number of individual genes using new high-coverage (~30X) sequencing data from the same samples in the 1000 Genomes Phase 3 dataset (Byrska-Bishop et al. 2021). These new gene copy number estimates are different from the SV calls from the 1000 Genomes in three ways. First, the gene copy number is estimated for individual samples and no population-level SV discovery or genotyping was performed, eliminating batch effects. Second, it provides continuous values of normalized read-depth, rather than discrete categories of different types of SVs as was reported from the 1000 Genomes Phase 3 dataset. This allows us to measure, for each gene, the distance between arbitrary two populations directly by the earth mover's distance, which directly uses the difference between the normalized read-depth value for an individual in one population and that for an individual in the other population. In this manner, we can avoid technical biases in the 1000 Genomes Phase 3 dataset introduced by the categorization of SVs into discrete types (**Methods**). Third, this new dataset

allowed us to estimate copy number variations in genes that were not discovered by the conservative discovery pipeline of 1000 Genomes Phase 3.

This approach allowed us to conduct a parallel assessment of the MU approach in detecting putatively adaptive SVs (**table S5, fig. S6**). We were able to assess 21 copy number variable genes (see **Methods**) that we identified to show unusually high *MU* in our original pipeline and found that 8 (~38%) and 13 (~62%) of these also have unusually high *MU* values in the mrCaNaVaR database at the 99th and 95th percentile, respectively. These include copy number variations of *KANSL1*, *HP*, and *DMBT1* genes with well-described likely adaptive functions. We individually investigated the remaining genes for which the copy number variation showed disparate *MU* percentiles in our original analysis and mrCaNaVaR analysis. We found that all of them are large (18kb-259kb) and relatively rare (<7.5% global allele frequency) variants that fall into regions rich in segmental duplications. These regions are prone to both genotyping errors and recurrence. In addition, using the mrCaNaVaR dataset, we were able to identify several additional candidates exonic SVs, including *AMY1*, *SIGLEC14*, and multiple *CCL* genes, among others, that were not included in the 1000 Genomes Phase 3 SV dataset but were noted because of their relevance to human evolution (Hollox et al. 2021). Thus, our results confirm that the *MU* provides a robust and reliable approach to identify putatively adaptive SVs. However, genotyping errors are a considerable factor in determining the false positive and negative rates in our approach and we argue that it is imperative to conduct follow-up analyses of the candidate adaptive SVs to validate deviations from neutrality at the haplotype level. We provide examples for such analysis below.

Table 1. Twenty-four multiallelic exonic structural variants (CNV, copy number variation in **fig. 2C**) with the top 1% *MU* (higher than 4.23). Variant information is retrieved from the 1000 Genomes Project phase 3 data set (Sudmant, Rausch, et al. 2015). Start and End refer to the starting and ending locations of variants on the chromosome, respectively. We described the gene name if the structural variant contains one or more entire exon(s) of UCSC Genes. Gene function was retrieved from OMIM (https://www.omim.org/).

ID	MU	chr	start	end	size	Allele.freq	gene name	OMIM
esv3640680;esv3640681	8.75	chr17	44230893	44262697	31804	0.22723629	KANSL1	Chromatin Modification
esv3640677;esv3640678	7.34	chr17	44165338	44211686	46348	0.1453675	KANSL1	Chromatin Modification
esv3638992;esv3638993;esv3638994;esv3638995;esv3638996;esv3638997	7.08	chr16	72094527	72110961	16434	0.02515976	HP	Glycoprotein
esv3606964;esv3606965;esv3606966;esv3606967;esv3606968;esv3606969;esv3606970	6.43	chr5	140554408	140558942	4534	0.287539932	PCDHB7	Protocadherin
esv3624777;esv3624778;esv3624779;esv3624780	6.38	chr10	124344431	124353237	8806	0.24580733	DMBT1	Brain Tumor
esv3640025;esv3640026	6.11	chr17	14224374	14483419	259045	0.078474481	HS3ST3B1	Heparan Sulfate
esv3616116;esv3616117;esv3616118;esv3616119	5.78	chr8	7212582	7227421	14839	0.36841027	ZNF705G	Zing Finger
esv3607012;esv3607013	5.56	chr5	142263109	142447062	183953	0.074480842	ARHGAP26	Leukemia
esv3638989;esv3638990;esv3638991	5.33	chr16	72080868	72098986	18118	0.04572683	HP	Glycoprotein
esv3603782;esv3603783;esv3603784;esv3603785	5.22	chr5	814446	825367	10921	0.18250742	ZDHHC11	Zing Finger
esv3647809;esv3647810;esv3647811;esv3647812	5.09	chr22	42523949	42533891	9942	0.160942323	CYP2D6	Metabolism
esv3608531;esv3608532	5.07	chr6	31131451	31272307	140856	0.069289142	POU5F1	Transcription Factor
esv3638688;esv3638689;esv3638690	5.04	chr16	55832207	55864521	32314	0.202276441	CES1	Metabolism
esv3624140;esv3624141	4.98	chr10	90551092	90632203	81111	0.061900981	LIPM	Signal Peptide
esv3638686;esv3638687	4.69	chr16	55798890	55822423	23533	0.20127776	CES1P1	Metabolism
esv3621839;esv3621840	4.68	chr9	132463983	132648102	184119	0.062899361	PRRX2	Homeobox
esv3644233;esv3644234	4.61	chr19	35851718	35863310	11592	0.12879353	FFAR3	Fatty Acid
esv3585247;esv3585248;esv3585249	4.50	chr1	12901370	12921250	19880	0.149361003	LOC649330	Unknown
esv3608684;esv3608685	4.46	chr6	35521984	35568895	46911	0.03674116	FKBP5	Binding
esv3638338;esv3638339;esv3638340;esv3638341;esv3638342	4.40	chr16	28614507	28626916	12409	0.26936938	SULT1A1	Metabolism
esv3641584;esv3641585	4.38	chr18	3200017	3415245	215228	0.064696481	MYOM1	Muscle
esv3599276;esv3599277	4.30	chr3	195954431	196022808	68377	0.060303542	SLC51A	Solute Carrier
esv3599572;esv3599573;esv3599574	4.26	chr4	9418201	9457405	39204	0.05271566	DEFB131	Defensin
esv3607010;esv3607011	4.24	chr5	142174919	142260351	85432	0.061701281	ARHGAP26	Leukemia

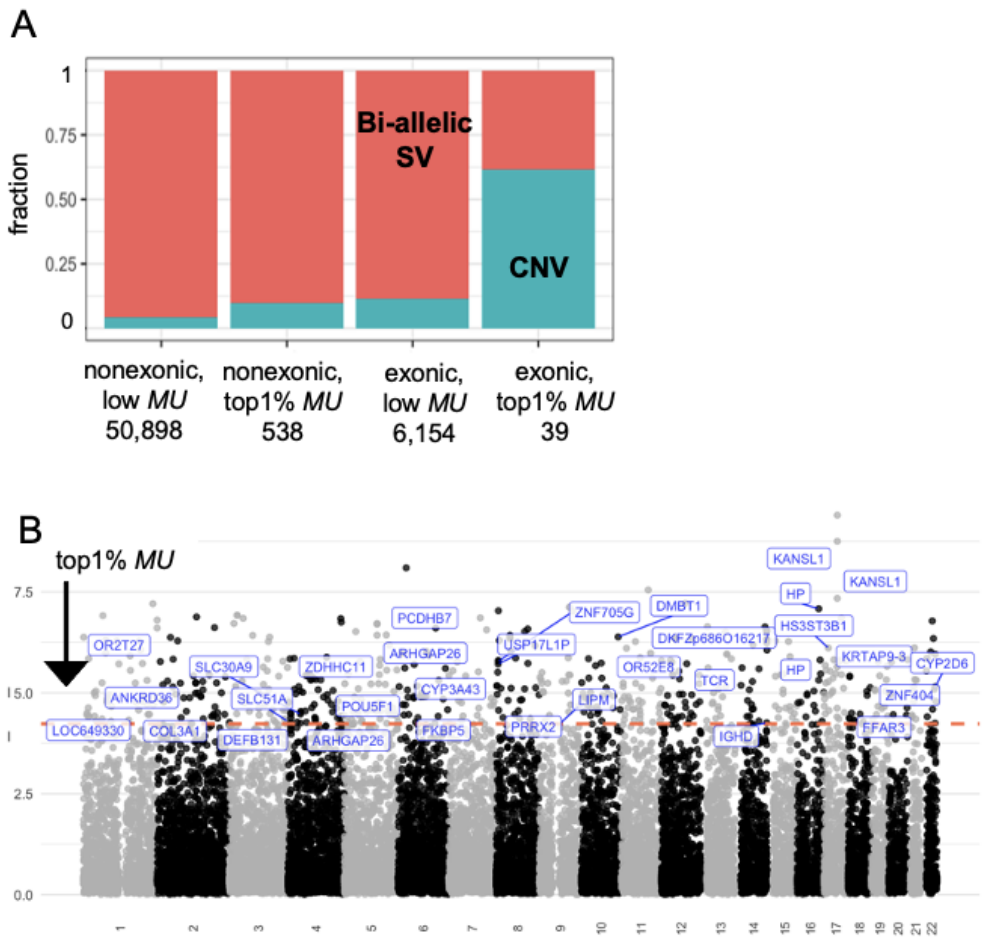


Figure 3 Examples of unusually distributed variants. A. The global allele frequency and *MU* of the highlighted variants (black points). The shade in blue represents the density of all the structural variants. **B** The relative ratio of bi-allelic structural variants and multiallelic copy number variants in each grouping based on exonic overlap (structural variants that contain at least one exon vs. those that do not) and *MU* values (top 1% vs. the rest). Bi-allelic SV: biallelic structural variants. CNV: multiallelic copy number variants with three or more alleles. **F.** Manhattan plots of *MU* of structural variants. The horizontal axis shows the chromosomal location of structural variants, and the vertical axis shows the *MU* value. The exonic variants with the top 1% *MU* (higher than 4.23) were labeled.

Resolving the Haplotypes of Putatively Adaptive Structural Variants

The complex evolutionary dynamics of structural variants often do not fit classical population genetics expectations, such as complete classical sweeps. Thus, we argue that careful investigation of the evolutionary histories of a few examples can provide valuable insights that can later be generalized at the genome-wide scale. Therefore, we wanted to resolve the haplotypes that harbor structural variants in order to investigate functional associations, coalescence times, and signatures of selection concerning these variants in more detail. There are 344 bi-allelic structural variants that are in the 1st percentile in terms of *MU* (> 4.23) and have strong linkage disequilibrium with nearby single nucleotide variants (R^2 > 0.95) (**table S6**).

Among these, we identified 20 haplotypes that are significantly associated with phenotypes (nominal $p < 10^{-9}$; GWAS Atlas (https://atlas.ctglab.nl/; last accessed 3.23.2020) (**table 2**). The selected 20 loci provided us with a

means to further investigate the evolutionary and functional effects of structural variants that show unusual geographical distribution.

Using the linked haplotypic variation for the 20 structural variants, we retrieved allele ages from the Human Genome Dating database (Albers & McVean 2020) (https://human.genome.dating; last accessed, 3.23.2020). Under neutrality, an allele's age is expected to positively correlate with its allele frequency (Patterson 2005). Given that we are explicitly investigating variants that are putatively evolving under population-specific adaptive forces, we expect deviations from this expectation. **Fig. 4A-C** shows the estimated age of the allele and its frequency in European, East Asian, and African populations, respectively. If a variant has emerged recently, but its frequency is common (left upper side) in a given population, it suggests a potential recent selective sweep (i.e., a new allele is rapidly favored and increases its frequency). In contrast, if a variant is old and its frequency is rare, these are candidates for recent negative selection against the allele in that particular population. To more formally interrogate this line of inquiry, we calculated how long it takes for a new allele to reach a given frequency in each population under neutrality using formula (15) in (Kimura & Ohta 1973) assuming previously published demographic parameters (Schaffner et al. 2005) (**fig. S7**). In a manner similar to allele frequency expectations, the age estimate of a variant older than the neutral estimation may suggest a faster increase in allele frequency and a recent selective sweep (**fig. 4A-4C**). In parallel, we calculated Tajima's D scores of 5k upstream and downstream regions of the 20 structural variants of interest and the iHS scores of the tag single nucleotide variants of the target structural variants (**Methods**). We summarized these values in **fig. 4D**.

We found that the flanking haplotypes of putatively adaptive SVs predicted by *MU* do not show consistent trends of haplotypic variation, extended homozygosity, or population differentiation. Rather, our observations fit the emerging consensus in evolutionary genomics that the adaptive structural variants are shaped by complex evolutionary trajectories that change over time and space (Mérot et al. 2020). As an example of the complicated nature of the evolutionary histories of adaptive structural variants, we highlight esv3642017. This variant is recorded as a deletion compared to the reference genome in the 1000 Genomes Phase 3 data set. However, a closer inspection reveals that this variant is a human-specific retro-insertion of the *DHFR* gene (Anagnou et al. 1988; Schrider et al. 2013; Conrad et al. 2010). The haplotype that harbors this insertion is associated with decreased height ($p < 10^{-16}$). Even though deletion seems to be predominantly found in Africa, the derived retrogene inserted is predominantly found in Eurasia. The locus that harbors the insertion shows unusually low Tajima's D in the European population and unusually low genetic diversity in another European-ancestry cohort as reported in Schrider et al. (2013), which altogether suggest a Eurasian-specific sweep of a recent insertion. Based on such locus-specific analyses, we identified incomplete population-specific sweeps and recent population-specific negative selection as the two main drivers for shaping the allele frequency distribution of putatively adaptive structural variants.

Table 2. Structural variants with the top 1 % *MU* value and observed phenotypic effects through tag SNP (R^2 > 0.95) in GWAS Atlas.

All the structural variants in the table are deletions. ID is the same as **fig. 4**. We described the GWAS traits in the "Trait" column if the structural variant shows phenotypic effects through tag SNPs on GWAS Atlas. We described the gene name in the "Exon" column if the structural variant contains one or more entire exons of UCSC Genes. "Afr/EAsia/Eur.freq" are the frequency of the alternative allele in each population, Africa, East Asia, and Europe. The "Selection" column describes estimated natural selection based on the neutrality tests (**fig. 4**). Specifically, "negative" indicates cases where we found recent selection favoring the ancestral allele, while "positive" indicates cases where we found recent selection favoring the derived alleles. *This "deletion (esv3642017)" found in the 1000 Genomes samples as compared to the reference genome is actually a derived insertion that happens to be represented in the reference genome (Anagnou et al. 1988; Schrider et al. 2013; Conrad et al. 2010). Thus, even though the putative action of selection is on the non-deleted haplotypes, given that this haplotype carries the derived allele, we categorized the selection as "positive" in this case. Two variants highlighted in yellow are discussed specifically as examples.

ID	esv number	Trait	*MU*	Exon	Afr.freq	EAsia.freq	Eur.freq	Selection
1	esv3592798	Age started wearing glasses or contact lenses	4.91		0.08620	0.6478	0.4245	positive
2	esv3587652	White blood cells	5.92		0.08090	0.5913	0.3956	positive
3	esv3647661	Height, Waist-hip ratio (adjusted for BMI)	6.78		0.07190	0.5238	0.6809	positive
4	esv3590788	Hot drink temperature	5.14		0.04920	0.5317	0.3231	positive
5	esv3634671	Height, Fat, Vertical cup-disc ratio, Weight, Sexual maturity	4.60		0.02800	0.4573	0.2227	positive
6	esv3626213	BMI, Alcohol intake, Neuroticism, Walking pace	5.23		0.08400	0.4375	0.4314	positive
7	esv3597888	Total bilirubin, Schizophrenia, Worry	6.38		0.05750	0.7857	0.4493	positive
8	esv3629184	Heart pulse rate, Morning person	5.44		0.02800	0.0347	0.3827	positive
9	esv3629182	Heart pulse rate, Morning person	5.44		0.02800	0.0347	0.3827	
10	esv3629434	Hematocrit, Hemoglobin concentration, Red blood cell count	6.35		0.01890	0.0109	0.326	
11	esv3591931	Hair type, Platelet volume	4.88		0.23220	0.0238	0.4016	positive
12	esv3616249	Heel bone mineral density, Platelet, Red cell	6.30		0.74050	0.0208	0.3887	negative
13	esv3639698	Height	4.85		0.03100	0.4851	0.2823	
14	esv3640666	Red blood cell, Bone mineral density, Baldness, brain volume,	4.62	*ARL17A*	0.01510	0.001	0.2416	
15	esv3636701	Heel bone mineral density	4.24		0.08700	0.623	0.1421	
16	esv3625336	Water intake	4.33		0.07870	0.3194	0.4523	
17	esv3609042	Gamma-glutamyl-transferase	5.70		0.09080	0.5089	0.674	
18	esv3636957	FEV1/FVC ratio, Impedance, Height	5.90		0.06880	0.3056	0.6759	
19	*esv3642017	Height	4.29		0.69820	0.2817	0.0984	positive
20	esv3587631	White blood cells	8.50		0.52120	0.002	0	negative

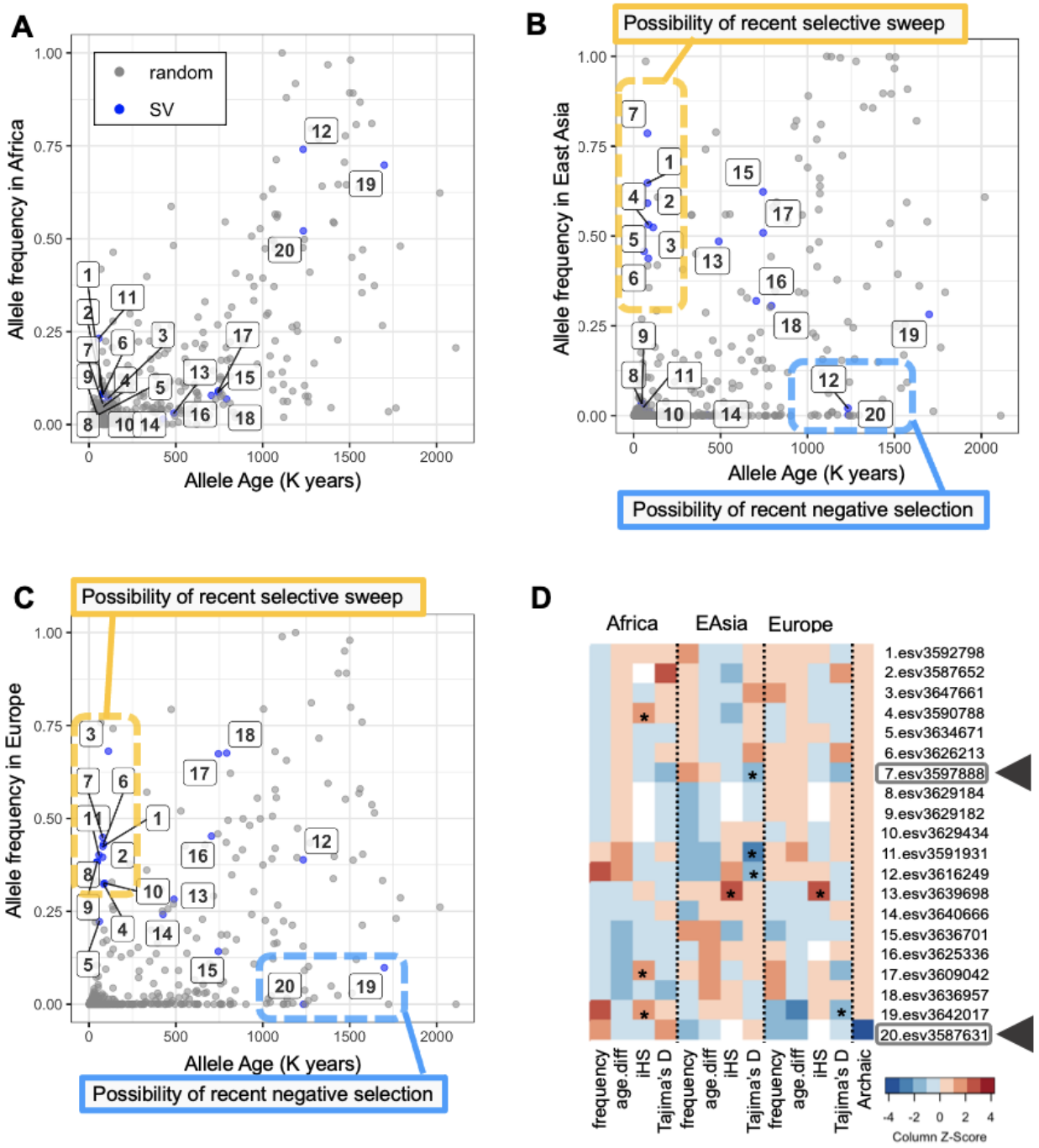


Figure 4. Neutrality tests on the 20 structural variants with phenotypic effects. (A-C) Age and frequency of the structural variants in African **(A)**, East Asian **(B)**, and European **(C)** populations. The variant's ID is the same as **table 2**. The blue dots represent the tag SNPs associated with the 20 structural variants (SV). The gray dots represent 599 random SNPs. The horizontal axis shows the age of allele in the Human Genome Dating database; the vertical axis is the alternative allele frequency in each population. **D.** A heatmap summarizing neutrality tests of the 20 structural variants in each population. It shows allele frequency, the allele age difference between haplotype-based estimation in the Human Genome Dating database (Albers & McVean 2020) and neutral expectation based on the allele frequency (Kimura & Ohta 1973), iHS value, Tajima's D, and if the allele is observed in the Denisovan genome. (We did not observe these variants in the Neanderthal genomes.) Warmer colors indicate higher values. Similarly, for the allele age difference, colder and warmer colors show that the allele is older and newer than neutral expectation, respectively. For the archaic genome, blue shows that the allele is shared with the Denisovan. The asterisk indicates that the value is less than 5 percentile or more than 95 percentile when compared to at least 500 uniformly randomly selected variants or windows across the genome. (See Methods for detail.) Two variants highlighted by the arrows (7 and 20) are discussed specifically as examples.

Incomplete, Population-Specific Sweeps: The example of the PCC Gene

The classical scenario for population-specific adaptive evolution is characterized by (i) high frequency of the variant in the specific population compared with other populations, (ii) deviations in the site frequency spectrum suggesting rapid expansion of the selected allele, resulting in an excess of rare variants in the locus, (iii) lower than expected allele age, and (iv) long haplotype homozygosity suggesting rapid expansion of the selected allele (Rees et al. 2020). We look for signatures of this scenario among the 20 haplotypes that we highlight because they harbor structural variants with unusual allele frequency distributions (*MU* in the 1st percentile, > 4.23) and because they are associated with GWAS traits (**table 2**). We found that 12 (60%) of them fit the scenario of a recent population-specific adaptive sweep (**fig. 4D**).

The haplotype harboring esv3597888 provides an informative example of the population-specific incomplete sweep scenario. The haplotype has a lower than 5% allele frequency in most African populations but reaches near 75% allele frequency in East Asian populations (**fig. 5A**). Further, the Median Joining network of the haplotypic variation in this locus shows a dramatic reduction of haplotypic diversity beyond the expected reduction due to drift in the East Asian population as compared with the African population, which is consistent with a recent selective sweep (**fig. 5B**). The Tajima's D values retrieved from the flanking sequences of the deletion are lower than genome-wide expectations in all three continental populations (**fig. 5C**). Last but not least, the estimated age of the allele is much more recent than what is expected based on its frequency, especially in the East Asian population (**fig. 4B, fig. S7**). Collectively, these results suggest a recent selective sweep in Eurasian populations. However, even for this locus, not all the neutrality tests capture this sweep. For example, in a traditionally defined recent sweep, we expect to find high iHS values. Instead, for this locus, the iHS is relatively low, mirroring the surprisingly high overall haplotypic variation in this locus. Regardless, esv3597888 remains one of the best candidates for a derived structural variant that has recently been swept to higher allele frequency in a population-specific manner.

The phenotypic effects of the haplotype harboring this variant further support the potential adaptive relevance of esv3597888. This 5.4kb deletion overlaps with the intronic region of the propionyl-CoA carboxylase (PCC) gene, which encodes for an enzyme that metabolizes specific amino acids and lipid species (Wongkittichote et al. 2017). The haplotype harboring esv3597888 (tagged by rs556788) is associated with the expression of the PCCB gene in the adrenal gland ($p = 7.8 \times 10^{-10}$ in the GTEx Analysis Release V8 (GTEx Consortium 2013)). Moreover, the haplotype is associated with total bilirubin, a cardiometabolic signaling molecule ($p = 6.6 \times 10^{-16}$), and neuroticism ($p = 1.5 \times 10^{-12}$) (**fig. 5D**). We argue that it is likely that the 5.4kb deletion esv3597888 is the causal variant in these associations, given its size and that it overlaps with a well-documented binding site for the abundant transcription factor CTCF (**fig. S8**). The haplotype has pleiotropic functional effects, and thus the exact reasons why it confers an adaptive advantage in East Asia particularly remains to be seen.

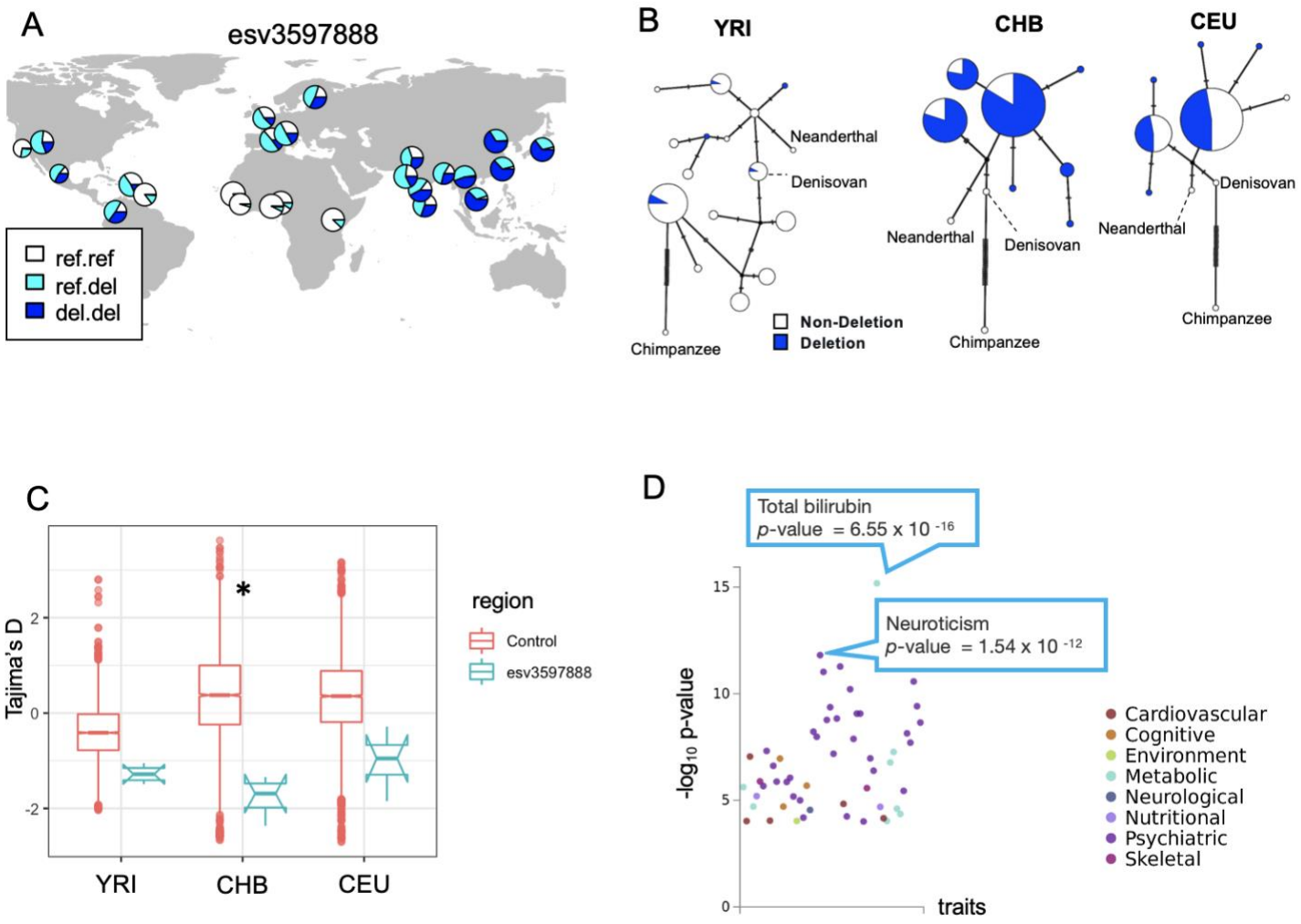


Figure 5. The evolutionary analysis of esv3597888, the deletion overlapping the intronic region of the propionyl-CoA carboxylase (PCC) gene. A. The geographic distribution of the esv3597888. **B.** Haplotype networks constructed from the 5kb upstream and downstream sequences from the esv3597888 location of three modern human populations (Yoruban (YRI), Han Chinese (CHB), and European (CEU),), the Altai Neanderthal sequence (Prüfer et al. 2014) and the Denisovan sequence (Reich et al. 2010) that are mapped to hg19 reference genome, and the chimpanzee reference genome (panTro4). The haplotypes that harbor the deletion are indicated by white and those that do not are indicated by blue. **C.** Tajima's D value in the 5kb upstream and downstream regions of esv3597888 (10kb in total) (Tajima 1993). Asterisk shows that Tajima's D of the esv3597888 flanking region is lower than the bottom five percentile of Tajima's D of 5000 random regions. The asterisk shows that the mean value of esv3597888 tag region is lower than the five percentile of the control region. **D**. The PheWAS result of rs556788, which tags esv3597888 (**table S4**). Each dot indicates a trait. The vertical axis shows the -log10 p-value of the association between the genotype and phenotype. The color indicates the phenotype category in GWAS ATLAS.

Recent Population-Specific Negative Selection: The DAP3 Gene

Among the haplotypes that we highlighted, we noticed that some show unexpectedly low allele frequency in Eurasian populations compared to the expectation based on their estimated age. We hypothesize that these haplotypes, and by proxy the structural variants that they harbor, may have been subjected to recent, population-specific negative selection, favoring the ancestral allele. Variations that have emerged early in human evolution and remain in extant populations are often found at high allele frequencies in all extant human populations under neutrality (Lin et al. 2015). Thus, if *recent* negative selection on the derived structural variant is acting in a

population-specific manner, we expect to observe in that population i) an unusual reduction in allele frequency of the variant that cannot be explained by drift alone and ii) a shift in the allele frequency spectrum towards rare variants in the locus.

A striking example of population-specific negative selection is provided by the haplotypes harboring esv3587631, which shows one of the highest *MU* values in the genome (*MU* = 8.50). This deletion is the major allele (i.e., > 50% allele frequency) in most sub-Saharan African populations but almost absent in non-African populations (**fig. 6A**). Human Genome Dating database estimates the age of a single nucleotide variant tagging esv3587631 to be 1,1-1,3 Million years old (Albers & McVean 2020). Thus, the deletion has emerged prior to human-Neanderthal divergence. Consistent with this result, we found that Denisovan but not Altai Neanderthal carries this deletion (**fig. 6B**). The haplotype network showed that the haplotypes harboring the deletion are similar to those from archaic hominins, consistent with our observation that this deletion is present in archaic human genomes (**fig. 6C**). Collectively, it is clear that the deletion has evolved before Human Neanderthal divergence and increased in allele frequency in African populations to more than 75%, harbored by diverse haplotypes. However, none of the haplotypes that harbor the deletion is found in Eurasian populations. Furthermore, the locus shows significantly negative Tajima's D values in the East Asian population, further supporting non-neutral forces acting on the deletion (**fig. 4D**).

Functionally, this ~4.8kb deletion overlaps with one of the introns of the well-studied and highly conserved *DAP3* gene (**fig S8**). DAP3 is a mitoribosome protein that regulates apoptosis at the cellular level and is linked to multiple developmental, immune-related, and biomedically relevant phenotypes at the organismal level (Kim et al. 2017; Greber & Ban 2016). Specifically, the deletion overlaps with a conserved regulatory region comprising multiple transcription factor binding sites. Consistent with these observations, the haplotypes harboring the deletion (tag variant, rs348195) were strongly associated with the increased expression of the *DAP3* gene in various tissues ($p < 10^{-6}$), with the effect size exceeding 0.4 in some cases. Moreover, the deletion (through the analysis of tag SNP rs348195) is strongly associated with decreased levels of white blood cells (nominal $p = 2.1 \times 10^{-37}$, **Fig. 6D**). Collectively, these results are consistent with a scenario where an ancient deletion variant that has been either neutral or beneficial in African populations has become detrimental to fitness in Eurasian populations, perhaps due to adaptive constraints concerning immune function.

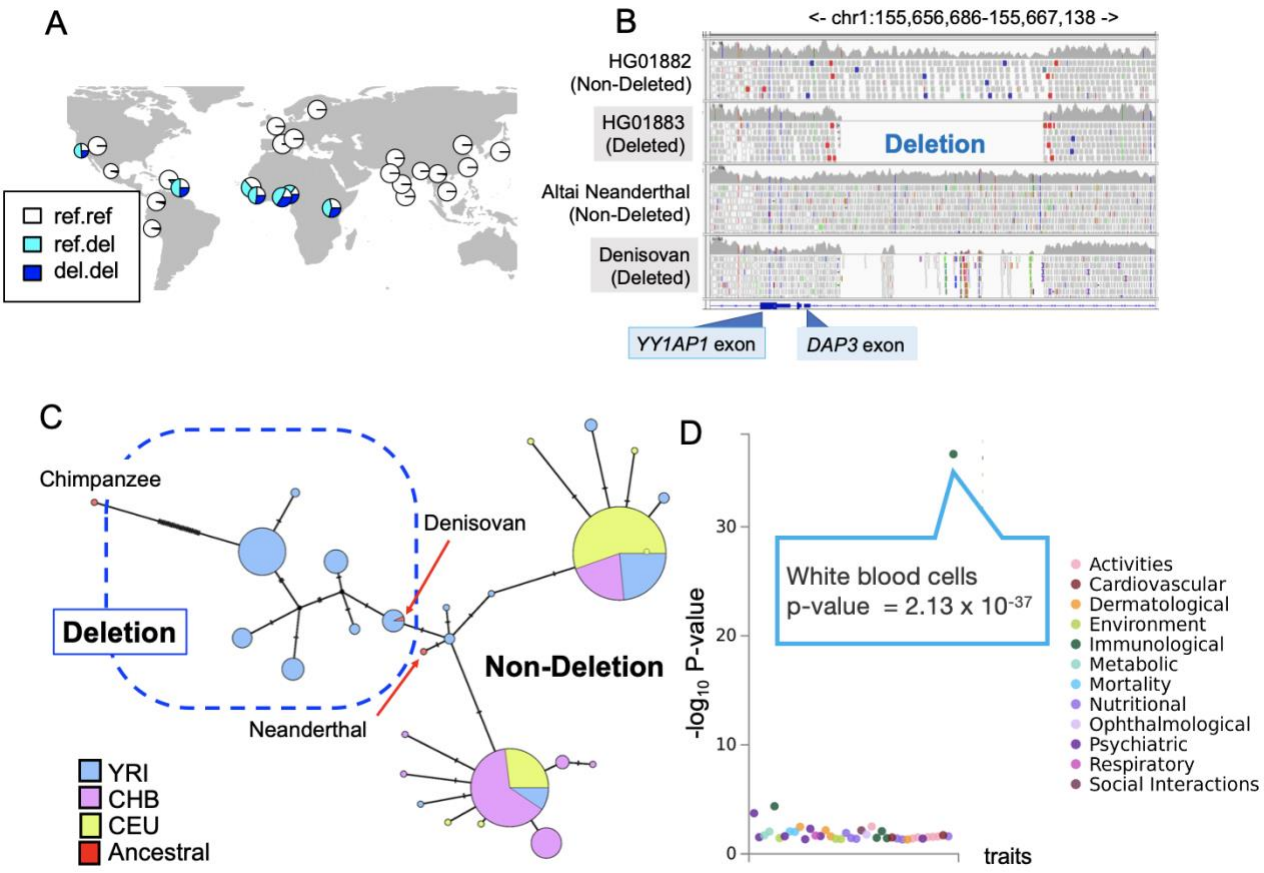


Figure 6. The evolutionary analysis of esv3587631, the intronic deletion polymorphism of the death-associated protein gene3 (DAP3). A. The geographic distribution of the esv3587631. **B**. esv3587631 in modern and ancient hominin genomes. These Integrated Genome Browser snapshots show the genome assembly (Hg19) of a human with the ancestral, homozygous non-deleted genotype and another with a homozygous deleted genotype that shows no reads mapping to the deletion region (top two rows). Similarly, sequences from Neanderthal and Denisovan genomes were mapped to this region. The Denisovan genome shows a clear signature of the deletion with breakpoints indistinguishable from the deletion observed in modern humans. **C.** A haplotype network of three modern human populations, Yoruba (YRI), Han Chinese (CHB), European (CEU), as well as Altai Neanderthal and the Denisovan, and Chimpanzee, constructed from the flanking sequences from the esv3587631 location. **D.** The Phewas result of rs348195, the tag SNP of esv3587631 at GWAS Atlas. Each dot indicates a trait and the y-axis shows the -$\log_{10}$ p-value of the association between the genotype and phenotype.

Conclusion

Although several putatively adaptive structural variants have been reported in previous studies, a genome-wide selection scan of structural variants has remained challenging. In this study, we built a network-based analysis of population differentiation among 26 populations in the 1000 Genome Project data set to identify putatively adaptive structural variants including multi-allelic variants. Our method assumes that drift is the major force that shapes the distributions of genomic variants among human populations as articulated by others (Ramachandran et al. 2005; Coop et al. 2009). In identifying the most common allele frequency distribution combinations across the 26 populations, our method parallels the recent variant-centric integrative analysis method proposed by Biddanda et al (2020). We argue that such direct, empirical scrutiny of the geographical distribution of variants

will provide a valuable and relatively unbiased picture of demographic and non-neutral trends that shape human genetic variation.

Our method is designed to identify structural variants with population differentiation that deviate from neutral expectations without any *a priori* adaptive model. It identified hundreds of putatively adaptive structural variants with unusual genotype frequency distributions in humans. The majority of these structural variants were hidden from traditional selection scans which mainly focus only on single nucleotide variants. Our study identified 24 putatively adaptive exonic multiallelic structural variants, the majority of which were not discussed within an adaptive context in humans. In addition to incomplete sweeps of derived structural variants, we found that recent population-specific negative selection is a considerable force shaping the geographic distribution of functional structural variants in humans. Overall, our study supports the emerging notion that structural variants significantly contribute to non-neutral and biomedically relevant phenotypic variation in humans (Radke & Lee 2015; Mérot et al. 2020) and highlight specific trajectories underlying the evolution of such variants.

From an evolutionary genomics perspective, the prominence of exonic multiallelic copy number variants among the putatively adaptive structural variants is not surprising. Cross-species analyses have repeatedly revealed the outsized role of recurrent gain and losses in gene families in shaping phenotypic characteristics in a variety of species, with recurrent evolution of caffeine in plants (Denoeud et al. 2014), salivary amylase in mammals (Pajic et al. 2019), and venom in snakes (Casewell et al. 2020) providing notable examples. Moreover, studies in humans reported that multiallelic copy number variants have seven times more effect on gene dosage than the combined effect of biallelic deletions and duplications (Handsaker et al. 2015). The same multiallelic structural variants, however, are hidden in the majority of GWAS and selection analyses. Multiallelic variants are not necessarily tagged by nearby single nucleotide variants, and they often reside in the genomic regions with enriched segmental duplications where identifying variants can be problematic. Thus, we expect that better genotyping of multiallelic structural variants with long-read sequencing platforms will dramatically increase our ability to identify multiallelic structural variants and their previously unknown adaptive roles.

A surprising result from our study is the identification of recent negative selection favoring ancestral alleles as a notable force determining the allele frequency distribution of putatively adaptive structural variants. Selective sweeps are often thought to increase the allele frequency of the *derived* and not *ancestral* variant. In this work, we found that at least 10% of the putative adaptive structural variants show recent sweeps favoring the ancestral allele. It is plausible that recent human adaptive evolution involves repeated adaptation to similar environmental conditions across time and geography as reported in (Bergey et al. 2018). Thus, an ancestral adaptive variant that confers a smaller fitness advantage than the derived variant may become adaptively beneficial again if environmental pressures revert back to an earlier state. This scenario is particularly applicable to immune-system-related traits within the context of an evolutionary arms race as articulated previously (Key et al. 2014).

Similarly, adaptive landscapes concerning metabolic traits have drastically changed multiple times for human populations due to technological advances (e.g., agricultural transition) (Hancock et al. 2010) and migrations to new ecologies (e.g., arctic populations) (Marciniak & Perry 2017). Thus, under the assumption that neither the ancestral nor derived alleles are fixed, it is not surprising that ancestral structural variants are favored in certain geographies and instances. Such cases will appear as negative selection against the derived allele. We reported in detail one such case involving the exonic deletion of the growth hormone receptor in another study (Saitou et al. 2021). The current study identifies several other cases, suggesting that recent, geography-specific negative selection is a considerable force shaping allele frequency distribution and population differentiation of functional structural variants.

There are caveats to our study and to the investigation of adaptive structural variants in general. First, it is clear from existing literature that the current data sets suffer from significant false-positive rates, potentially missing up to 80% of the structural variants (Mahmoud et al. 2019). Moreover, current technologies can discover certain types of structural variants (e.g., large biallelic deletions) much more sensitively than other types of variants (e.g., duplications, inversions). It is telling that one of the structural variants most relevant to human evolution, amylase copy number variation, are not cataloged by the 1000 Genomes Phase 3 data set because of alignment issues in the locus. Even when such multiallelic variants are discovered, it is not uncommon that their exact genotypes (e.g., exact copy number) may not be accurately documented. Second, the genotyping platforms commonly used in genetic association studies mostly focus on bi-allelic single nucleotide variants only. In fact, even this study, which is aware of these limitations, highlighted bi-allelic variants, for which the haplotype can be readily resolved, and thus trait associations can be investigated. The true contribution of most structural variants, including multiallelic variants, to phenotypic variation, remains mostly unknown. Third, most structural variant maps, including the dataset we use in our study are not an ideal representation of human variation. For example, a more powerful and adequate sampling would involve hypothesis-driven efforts where specific adaptive pressures are in mind (Scheinfeldt & Tishkoff 2013; Rees et al. 2020). Further, ascertainment bias in GWAS studies (Sirugo et al. 2019), which still comprise primarily European cohorts, limits our power to link evolutionary trends shaping the SV allele frequency distributions to their functional effects. Overall, the current picture of the evolutionary effects of structural variants, including those revealed in this study, remains incomplete and should be treated as a theoretical and methodological framework for future studies with more comprehensive data sets. We believe that as long-read sequencing-based discovery and later genotyping become affordable, the full impact of structural variants on human evolution and diversity will be better revealed.

Methods

1000 Genomes Phase 3 data set

As the input data set, we used 1000 Genome Project phase 3 data sets (Sudmant, Rausch, et al. 2015) for the following three reasons. First, the genotyping is based on whole-genome sequencing and multiple detection methods such as Delly (Rausch et al. 2012), which combines short insert paired-ends, long-range mate-pairs, and split-read alignments, and GenomeSTRiP (Handsaker et al. 2011), which uses read-depth and read pairs for structural variant identification to improve accuracy. Thus, this data set provides a highly accurate structural variant genotype. Second, it contains approximately 100 individuals from each population. Therefore, one can increase the power to detect geographically-differentiated structural variants due to population-specific adaptation by assessing deviations from expected population differentiation. Third, it provides phased genotype information not only of the structural variants but also of the SNPs from the same individuals. This allows us to apply our methods for identifying population differentiation to known SNPs to assess their performance and to carry out the subsequent haplotype-based analysis on a subset of structural variants.

Pre-processing structural variants and selection of known SNPs in the analysis

We selected 57,629 autosomal structural variants with annotations in the 1000 Genomes project phase 3 data set (Sudmant, Rausch, et al. 2015) since variants in sex-chromosomes are differently described from autosomes due to the smaller number of the observed number of chromosomes and can not be analyzed in the same pipeline as autosomal variants. As controls, we also used 1008 uniformly randomly selected single nucleotide variants from the same data set and six single nucleotide variants that have undergone putative natural selection, including rs334 in HBB (Ding et al. 2013), rs73885319 in APOL (Ko et al. 2013), rs4988235 in LCT(Smith et al. 2009), rs12913832 in OCA (Wilde et al. 2014), rs3827760 in *EDAR*, which is common in East Asian populations and associated with hair and dental traits (Mou et al. 2008; Wu et al. 2016; Kimura et al. 2009), rs1426654 in SLC24A5, which is associated with skin color (Basu Mallick et al. 2013; Deng & Xu 2018; Norton et al. 2007). In addition to these SNPs, we included rs1129740 that falls into *HLA-DQA1*, which is one of the few variants in the human genome that showed classical signatures of balancing selection (Teixeira et al. 2015). This *HLA* allele showed unusually low *MU* (0.22) despite the global allele frequency of 0.52 (**fig. 2**).

To verify that 1008 randomly chosen SNPs indeed represent a neutral dataset, we redid our analysis with 5000 SNPs shown to be evolving under near-neutrality by Pouyet et al. (2018). Briefly, we calculated *MU* in the same manner as our previous analysis, where we include all the structural variants and the 5,000 neutral SNPs. We found that this reanalysis did not change our results. Specifically, we found that the MU values calculated for structural variants when the SNP dataset are nearly identical to those that were calculated with our initial dataset (Spearman's correlation = 0.997) (**fig. S3B)**. Further, we replicated our finding that the MU values for all structural variants are not significantly different from those calculated for the neutral SNPs (p = 0.41, Mann-Whitney test, **fig. S3C**), indicating that the majority of structural variants are evolving neutrally.

Calculation of the *MU* and MDS plot

We characterize each locus ℓ as an $N \times N$ similarity matrix, denoted by S_ℓ, where N = 26 is the number of populations in the 1000 Genome Project Phase 3 data set (**fig. 1A**). The entries of matrix S_ℓ represent the similarity between pairs of populations in terms of the frequency of each allele at a locus. Specifically, for each locus ℓ and population i, the genotype count is given by the 1000 Genome Project Phase 3 data set. In general variant call format (VCF), genotype (i.e., the allele status of a pair of chromosomes of one individual) is denoted by $x_1|x_2$. Genotypes 1|0 and 0|1 in general variant call format are effectively the same and mean that one individual has one reference (i.e., 0) allele and one alternative (i.e., 1) allele at the locus. Therefore, we summarized both 1|0 and 0|1 into 0|1 in the following analysis. The maximum (alternative + reference) allele number at a locus was nine, in which case the allele number ranges from 0 to 8 (**fig. S9**). Therefore, in general, we summarized $x_1|x_2$ and $x_2|x_1$ into $x_1|x_2$, where x_1, x_2 = 0,1, . . . , 8 and $x_1 \leq x_2$. We denote the frequency of genotype $x_1|x_2$ at locus ℓ and population i by $p_{\ell,i}(x_1|x_2)$, where $0 \leq x_1 \leq x_2 \leq 8$. Note that $\sum_{x_1=0}^{8} \sum_{x_2=x_1}^{8} p_{\ell,i}(x_1|x_2) = 1$. To increase the sensitivity to identify the population differentiation of multiallelic variants (i.e., variants with more than two alleles), especially, with large copy number variation (such as a multiallelic variant with copy number one to eight, even if the frequency of copy number eight is rare), we use a weighted variant of the Bhattacharyya similarity metric, which modifies the Bhattacharyya similarity metric (Cha & Srihari 2002; Bhattacharyya 1943), as follows (**fig. 1B**).

First, we transform the original distribution, $\{p_{\ell,i}(x_1|x_2);\ 0 \leq x_1 \leq x_2 \leq 8\}$ to $\{\widetilde{p}_{\ell,i}(x_1|x_2);\ 0 \leq x_1 \leq x_2 \leq 8\}$, where

$$\widetilde{p}_{\ell,i}(x_1|x_2) = C(x_1 + x_2 + 0.5)p_{\ell,i}(x_1|x_2) \qquad (1)$$

and

$$C = \frac{1}{\sum_{x_1=0}^{8} \sum_{x_2=x_1}^{8} (x_1 + x_2 + 0.5)\, p_{\ell,i}(x_1|x_2)}. \qquad (2)$$

This transformation magnifies the frequency of genotype and its differences between populations at large x_1 and x_2 values (i.e., large copy number variation). Second, we measure the Bhattacharyya metric between each pair of populations i and j based on the transformed probability distribution, *i.e.*,

$$s_\ell(i,j) = \sqrt{\sum_{x_1=0}^{8} \sum_{x_2=x_1}^{8} \widetilde{p}_{\ell,i}(x_1|x_2)\widetilde{p}_{\ell,j}(x_1|x_2)}. \qquad (3)$$

If the two distributions $\widetilde{p}_{\ell,i}$ and $\widetilde{p}_{\ell,j}$ are identical, then s_ℓ (*i, j*) = 1, which is the largest value of s_ℓ (*i, j*). In this manner, for each locus *ℓ,* we obtain an $N \times N$ similarity matrix S_ℓ whose (*i, j*) entry is given by Equation (1).

To analyze the organization of *M* = 58,644 loci (which is composed of 57,629 structural variants, 1,008 random single nucleotide polymorphisms, and seven single nucleotide polymorphisms under adaptive evolution), we constructed a distance matrix for the *M* loci by comparing similarity matrix S_ℓ across the loci. We define the distance between S_ℓ and $S_{\ell'}$ using the Frobenius norm, denoted by d_F, which is given by

$$d_{\mathrm{F}}(S_\ell, S_{\ell'}) = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}[(S_\ell)_{ij} - (S_{\ell'})_{ij}]^2}. \qquad (4)$$

Note that $(S_\ell)_{ij} = s_\ell(i, j)$. Also note that one can replace the summation by $\sum_{i=1}^{N}\sum_{j=i+1}^{N}$ without loss of generality because the similarity matrix S_ℓ is symmetric and all of its diagonal elements are equal to 1. The $M \times M$ Frobenius distance matrix, denoted by *F,* tabulates the difference between each pair of loci, and its (*ℓ*, *ℓ'*) entry is given by $d_F(S_\ell,S_{\ell'})$. Finally, we ran a multidimensional scaling (MDS) algorithm on *F* to map out relationships between the *M* loci on a two-dimensional space. We used the Python package *manifold*, which is part of *scikit-learn* (Pedregosa et al. 2011), to estimate the MDS. To empirically measure these relationships, we calculated the distance between the origin and each variant in the MDS space and defined it as "*Measure of Unusualness (MU)*", or the degree of unusual allele frequency distribution (**Methods, fig. 1**). This measure informs on the unusualness of global population differentiation of a given structural variant, as compared to the entirety of the data set. We only used one initial condition due to the long time required for the computation. The analyses have been done on the server of the University at Buffalo Center for Computational Research (http://www.buffalo.edu/ccr.html).

Simulations and MU calculations

To model the evolution of copy number variation in humans, we modified recipe 14.11 (modeling microsatellites) and recipe 5.4 (model of human evolution) in SLiM 3.6 (Haller & Messer 2019). We modeled stepwise copy number gains or losses in each locus under different mutation rates/selection coefficients in three populations (YRI, CEU, and CHB) for which the demographic parameters were previously established (Gravel et al. 2011). In each simulation, we generated 2970 potentially variable neutral loci, and 10 x 3 potentially variable loci under population-specific selection in each population (with a range of selection coefficients), respectively. All the variable loci were initially neutral, but at generation 78084, just after the split of CEU and CHB, the 10 loci became adaptive in each population. We used 0.5, 0.05, 0.005 as selection coefficient and 10^{-5}, 10^{-6}, 10^{-7} for mutation rate. We also ran simulations with the mutation rate 10^{-8}, however, most of them did not produce polymorphic sites. So we did not eventually use the mutation rate of 10^{-8}. We simulated each model 100 times per condition

(three mutation rates x three selection coefficients). The allowed copy number range was 1 - 8. The script can be found on our GitHub webpage.
https://github.com/mariesaitou/Network_humanpop_SV

mrCaNaVaR copy number estimates and MU calculations

We used mrCaNaVaR (https://github.com/BilkentCompGen/mrcanavar) to estimate the normalized read-depth for each gene in the human genome using high-coverage (~30X) sequencing data (Byrska-Bishop et al. 2021) available for the same samples we used in our original analysis. mrCaNaVaR remaps short-reads promiscuously and measures the normalized read depth of intervals across the genome. Using this approach we calculated the read-depth of each gene across the human genome and use this data to calculate the MU values for each human gene as follows. First, we discarded the genes on chrM (mitochondrial DNA), chrX, chrY, which left us 36486 genes. We denote by $y_{g,p,i}$ the read-depth associated with gene *g* (*g* = 1, ..., 36486) in the *i* th individual in population *p* (*p* = 1, ..., 26), from the 1000 Genome phase 3 dataset. Second, for each gene, we divided each $y_{g,p,i}$ by the average over all individuals and populations. We applied this normalization to enable a gene-to-gene comparison of the copy number variation in a single population and across multiple populations without being affected by the typical copy number, which substantially depends on the gene. We denote the normalized $y_{g,p,i}$ by $\overline{y}_{g,p,i}$. Third, for the given gene *g*, we calculated the distance between two populations by the earth mover's distance(Levina & Bickel 2001; Pérez-Barbería et al. 2007) as follows. The distribution of the normalized copy number for population *p* is given by assigning probability $1/n_p$ to each value of $\overline{y}_{g,p,i}$ that appears in the data, where n_p is the number of individuals in population *p*. If there are two individuals that have the same value of $\overline{y}_{g,p,i}$, for example, then the probability of this normalized copy number is $2/n_p$. The earth mover's distance between populations *p* and *p*' is the minimal cost to move the distribution of the normalized copy number for population *p* to that for population *p*'. The cost of moving a unit probability mass from a location $\overline{y}_{g,p,i}$ in one distribution (corresponding to population *p*) to another location $\overline{y}_{g,p',j}$ in another distribution (corresponding to population *p*') is given by $|\overline{y}_{g,p,i} - \overline{y}_{g,p',j}|$. By calculating the earth mover's distance between each pair of populations, for each gene *g*, we obtain a 26 x 26 distance matrix, denoted by D_g, which represents how similar/different each pair of populations is in terms of the copy number distribution. Fourth, we calculate the distance between each pair of genes using the Frobenius norm using Eq. (4), but with matrix S_ℓ being replaced by matrix D_g. The remaining steps for calculating *MU* are the same as those for the 1000 Genome Project Phase 3 data set. The script is available on GitHub. https://github.com/mariesaitou/Network_humanpop_SV For comparing the results from mrCaNaVaR calls and our original analysis, we identified 21 genes that are multiallelic SVs in the 1000 Genomes Phase 3 dataset with high MU values (99th percentile) that are also presented in mrCaNaVaR analysis.

Allele frequency, *MU,* and structural variants

Since *MU* is related to alternative allele frequency by definition, we categorized the variants into four groups in terms of the allele frequency using ranges 1-0.25, 0.25-0.5, 0.5-0.75, and 0.75-1, and investigated the distribution of *MU* of the variants in each group. In intermediate allele frequency groups, multiallelic copy number variants (CNV) had higher *MU* than SNPs (Wilcoxon test, $p = 0.0021$ and $p = 0.00095$ for allele frequency ranges 0.25-0.5 and 0.5-0.75, respectively). This result indicates that our methods may detect unusual population differentiation due to the excess of multi-copy alleles than biallelic variants. In addition, we noticed that the small number of variants that have allele frequencies higher than 0.5 show smaller *MU* values than those variants that are less than 0.5. There is no mathematical reason for this observation. Thus, this observation may be due to various other reasons, including potential genotyping errors which may be increased among this group of variants. We want to further acknowledge here other studies that have employed tools to capture variation at the multiallelic SV locus (e.g., V_{ST} (Redon et al. 2006) and use thoughtful hypothesis-driven population sampling (e.g., (Huerta-Sánchez et al. 2014)). Regardless, the majority of genome-wide scans of selection employ biallelic locus and use available continental populations such as the 1000 Genomes dataset and almost none captures variation across dozens of populations.

Functional genomics analysis

We retrieved exonic content from UCSC Genome Browser (http://genome.ucsc.edu/; UCSC Genes, table: knownCanonical) and examined if each structural variant contained one or more entire exon using bedtools (Quinlan & Hall 2010). To find functionally relevant loci, we first calculated linkage disequilibrium between the top 1% high *MU* biallelic structural variants and neighboring regions (5kb upstream and downstream) with vcftools (Danecek et al. 2011). We searched the resulting tag SNPs ($r^2 > 0.95$) in GWAS Atlas Phewas database (https://atlas.ctglab.nl/PheWAS), defining that $p < 10^{-9}$ as a statistically significant association. Of the 576 top 1% structural variants in terms of *MU*, we found that 500 variants were bi-allelic, which were suitable for haplotype analysis. Among the 500 variants, 344 structural variants showed R^2 larger than 0.95 with neighboring variant(s). Of these 344 structural variants, 20 of them have flanking tag SNPs that are significantly associated with phenotypic variation (**fig. 2A, table 2**). Further, we used the GTEx portal (GTEx Consortium 2013) for associating these SNPs to variation in gene expression levels **(fig. S4).** We used Shiny.Go (Ge et al. 2020) or GREAT (McLean et al. 2010) to conduct functional enrichment analysis for genes and intervals overlapping structural variants with the whole genome as the background. These tools search for enrichment in multiple databases and provide multiple-hypotheses-corrected false discovery rates.

Evolutionary Genomics Analysis of the haplotypes harboring putatively adaptive structural variants

We used the 344 structural variants that are among the top 1% in terms of *MU* and have flanking tag SNPs with R^2 value larger than 0.95 for the following evolutionary analysis. We used the age estimates from Human Genome Dating database (Albers & McVean 2020) using tag SNPs as proxies to the adaptive SVs. This database documents the allele age estimates based on the analysis of pairwise haplotype identical tracts in 1000

Genomes (The 1000 Genomes Project Consortium et al. 2015) and Simons Genome Diversity Projects (Brazma 2019) (https://human.genome.dating; last accessed, 3.23.2020). We also calculated how long it takes for a new allele to reach a given frequency in each population under neutrality using equation (15) of (Kimura & Ohta 1973). For this calculation, we used demographic models for each population detailed in (Schaffner et al. 2005) (**fig. 3D**).

For haplotype-level population genetics measures, we targeted 5k upstream and downstream regions of the structural variants and calculated Tajima's D scores of the tag SNP of the target structural variants using VCFTools (Danecek et al. 2011). We retrieved the iHS score from the 1000 Genomes Selection Browser (Pybus et al. 2014). For comparative purposes, we calculated the same scores for ~500 random regions generated with bedtools (Quinlan & Hall 2010) across the genome. Ancient human (Altai Neanderthal and Denisovan) genomic bam files are published on the Max-Planck Institute website (https://www.eva.mpg.de/index.html) (Prüfer et al. 2014; Reich et al. 2010). We used samtool-based (Li et al. 2009) read-depth analysis to genotype deletions in archaic genomes (**fig. 5B**). We generated haplotype networks using VCFtoTree (Xu et al. 2017) and POPArt (Clement et al. 2002; Leigh & Bryant 2015) by Minimum Spanning Network method (Bandelt et al. 1999).

Data visualization and availability

For the data visualization, we used Rstudio (v1.2.1335), R(v3.5.3), and ggplot2 (Wickham 2009). Codes will be available at https://github.com/mariesaitou/Network_humanpop_SV.

Acknowledgments

We thank Drs. John Novembre and Simen Rød Sandve for the careful reading of this manuscript. NM acknowledges support from Japan Science and Technology Agency (JST) Moonshot R&D (under grant no. JPMJMS2021). OG acknowledges support from the National Science Foundation (under grant numbers 2123284). We thank Can Alkan for generously running mrCaNaVaR to estimate copy numbers. We thank Fanny Pouyet for generously providing us with a high-recombination region map.

SUPPLEMENTARY Materials

Table S1 All the variants and *MU* values

Table S2 Notable variants identified in this study (Low *MU* and High frequency; bottom 25% *MU* and 0.4 - 0.6 allele frequency,)

Table S3 MU and F_{ST} comparison for bi-allelic structural variants

Table S4 High *MU* (i.e., top 1%) exonic variants

Table S5 MU values for mrCaNaVaR results (top 10%)

Table S6 All the variants with high *MU* (top 1%) and high LD (R^2 > 0.95) with neighbouring variants

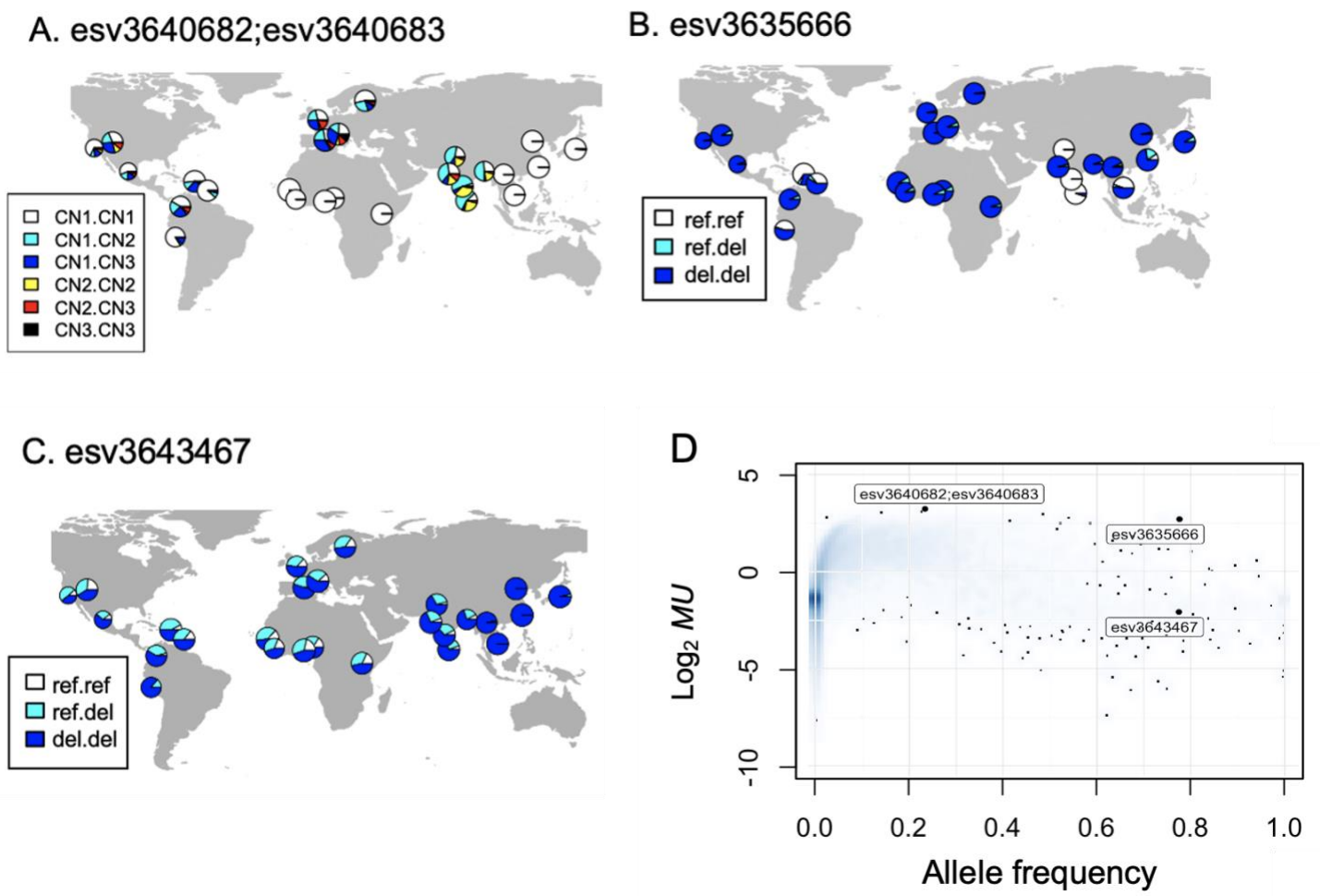


Figure S1 The geographical distribution of several identified structural variants as examples. Color represents genotype. White refers to a homozygous reference genotype. Non-white colors represent other genotypes. **A**. A multiallelic copy number variant; esv3640682;esv3640683, a multiallelic structural variant with unusual population differentiation and a top 1% *MU* value. CN means copy number, for example. CN1 is the copy number one allele and CN1.CN1 is the homozygous copy number one genotype. **B**. esv3635666 with unusual population differentiation and a top 1% *MU* value; this variant overlaps with the coding sequence of DKFZp686O16217 gene. **C**. esv3364367 at MUC4 exon is very common among human populations but showed low *MU*.

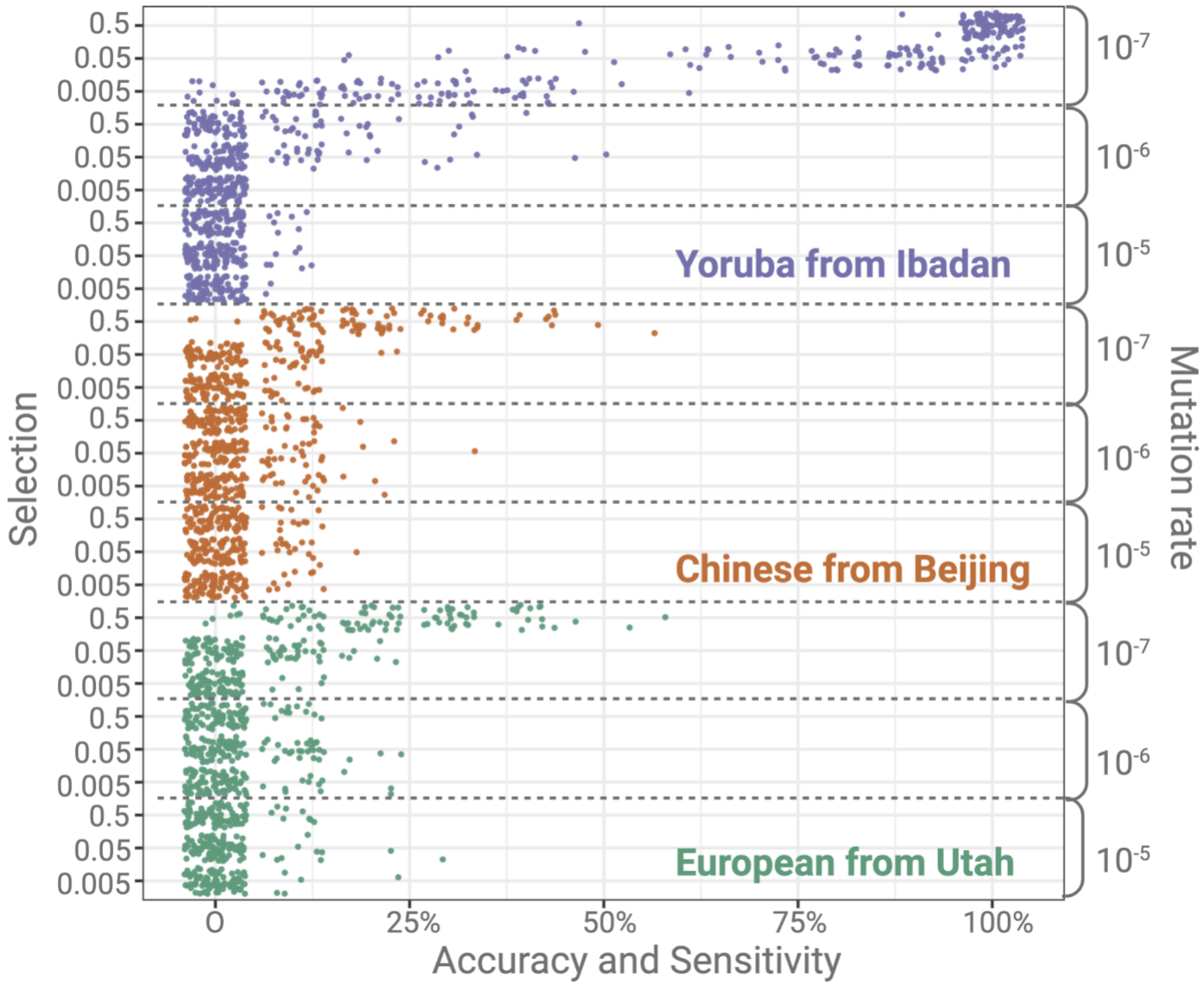


Figure S2. MU values calculated for the simulation results invoking different rates of selection and mutation rates. Each dot indicates one simulation result. For each simulation, we generated 10 alleles (out of 100) that are evolving under population-specific positive selection. The x-axis indicates how many of 10 alleles were correctly detected. Y-axis indicates the selection coefficients (left) and mutation rates (right).

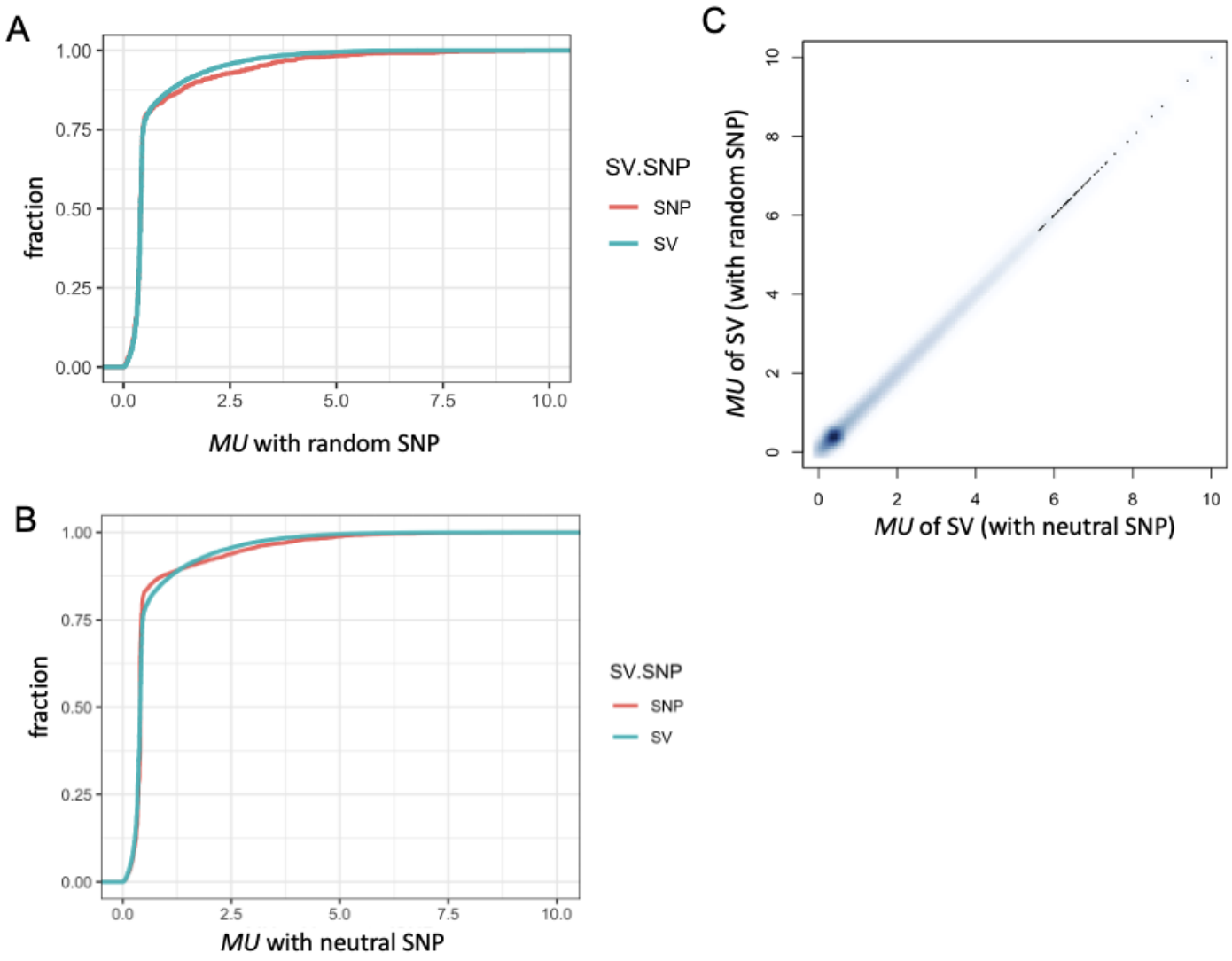


Figure S3. The comparison of MU values calculated for SNPs and structural variants. A. Cumulative fraction of the *MU* of 1008 randomly chosen SNPs and structural variants (SV). **B.** Comparison of the Mu values calculated for structural variants along with 1008 randomly chosen (y-axis) and 5000 “neutral” SNPs (x-axis). **C.** Cumulative fraction of the *MU* of SVs with 5000 “neutral” SNPs and structural variants.

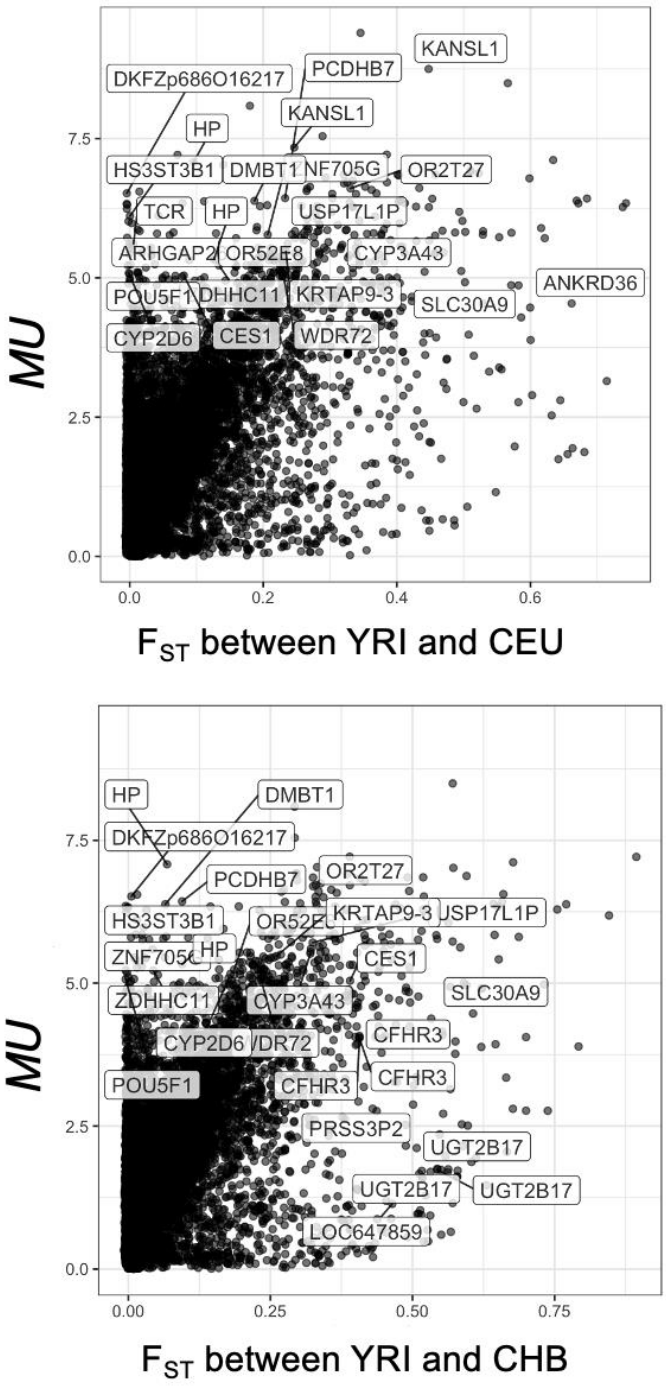


Figure S4 Comparison between F_{ST} and *MU* values of structural variants

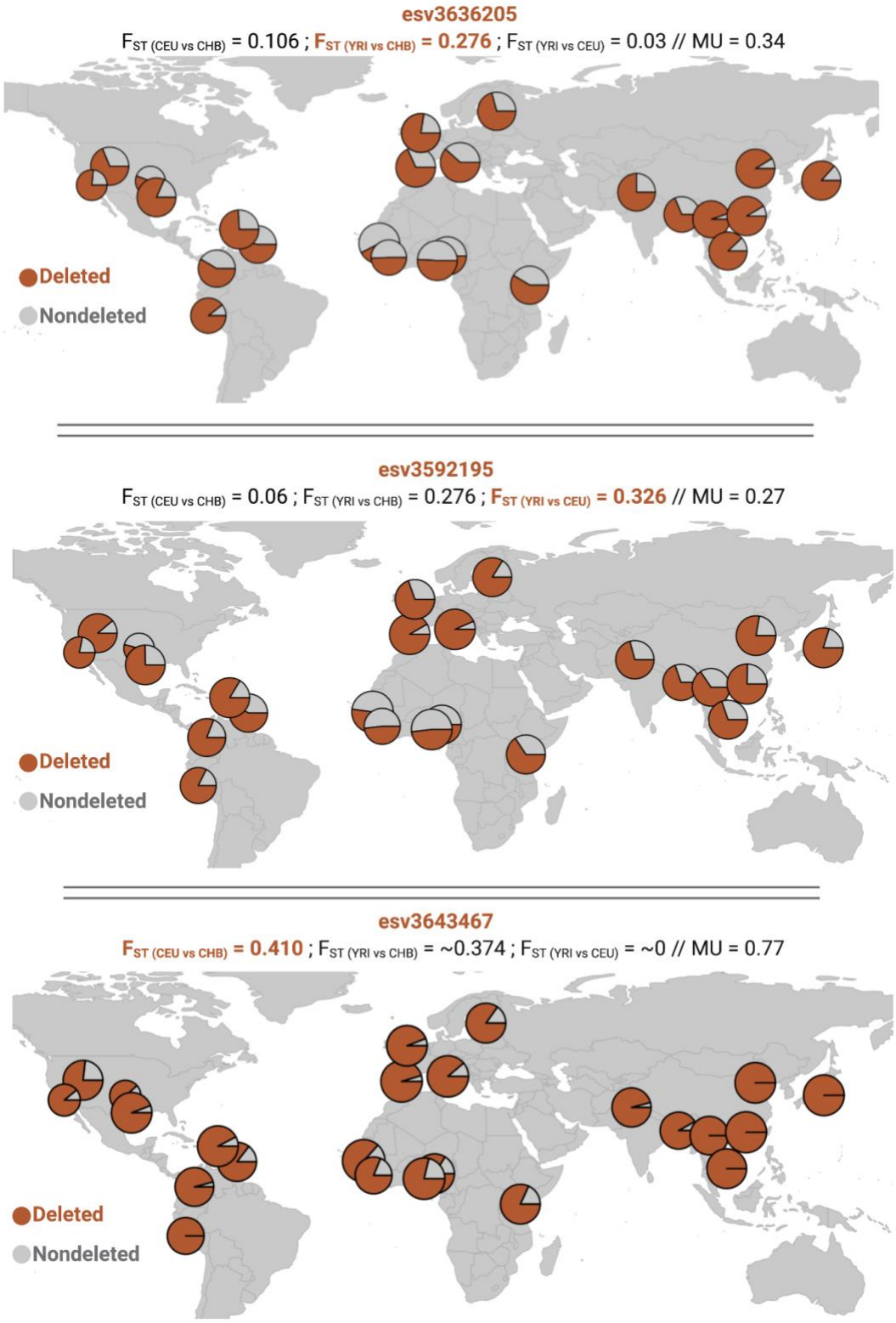


Figure S5. Three examples where variants have high F_{ST} values but low MU values

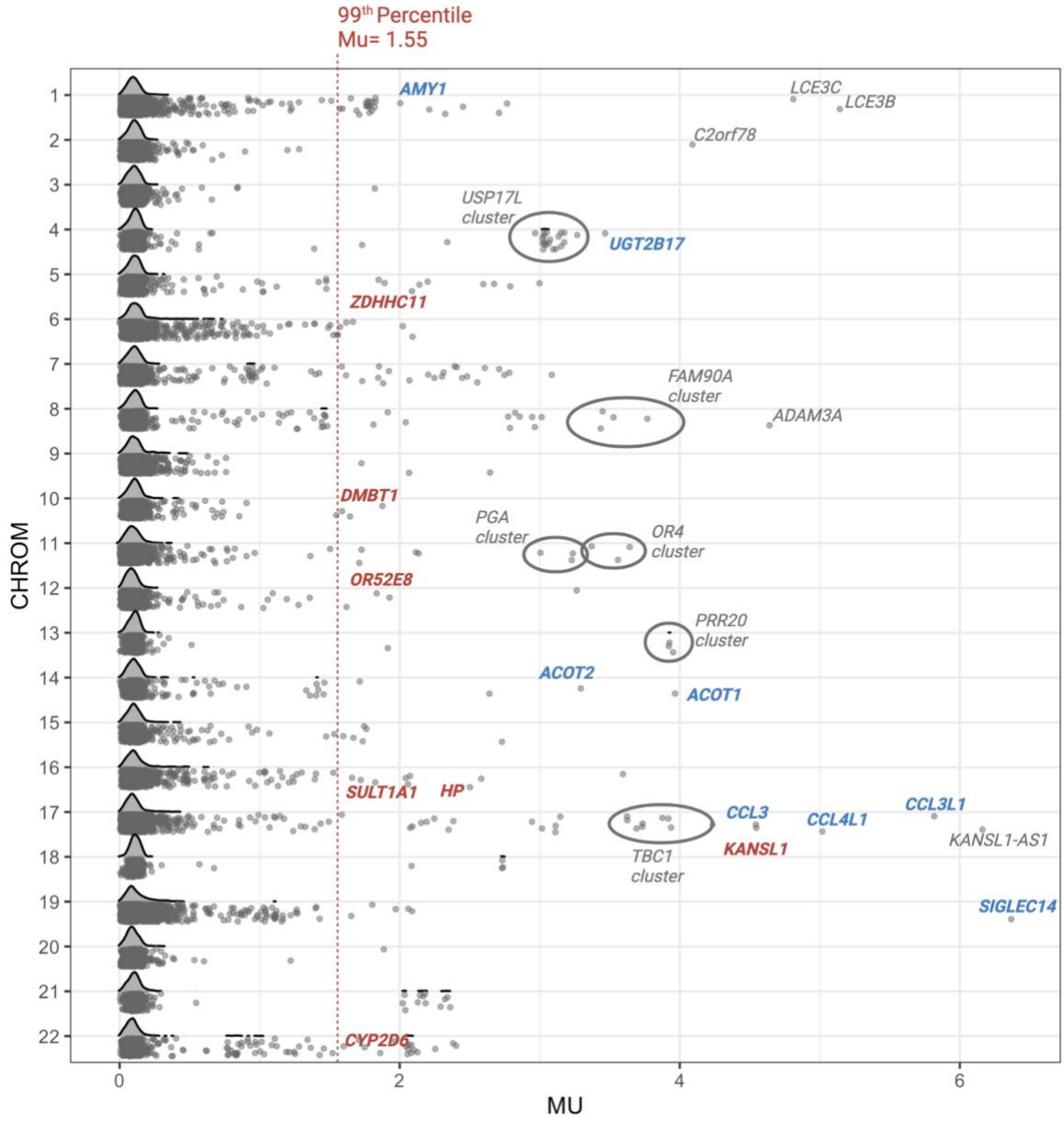


Figure S6. *MU* values calculated from distance matrices based on mrCaNaVaR copy number estimates.

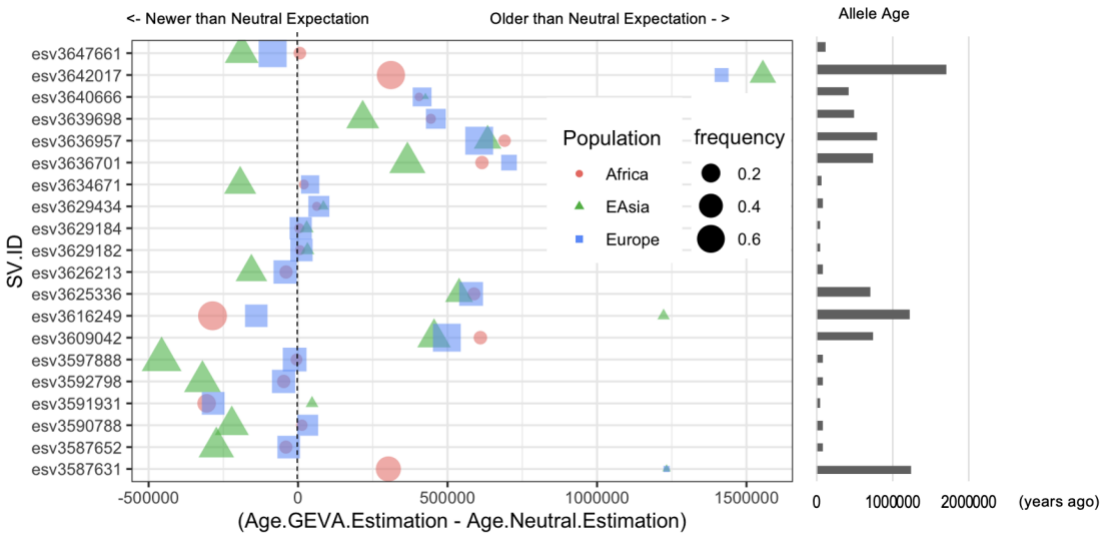
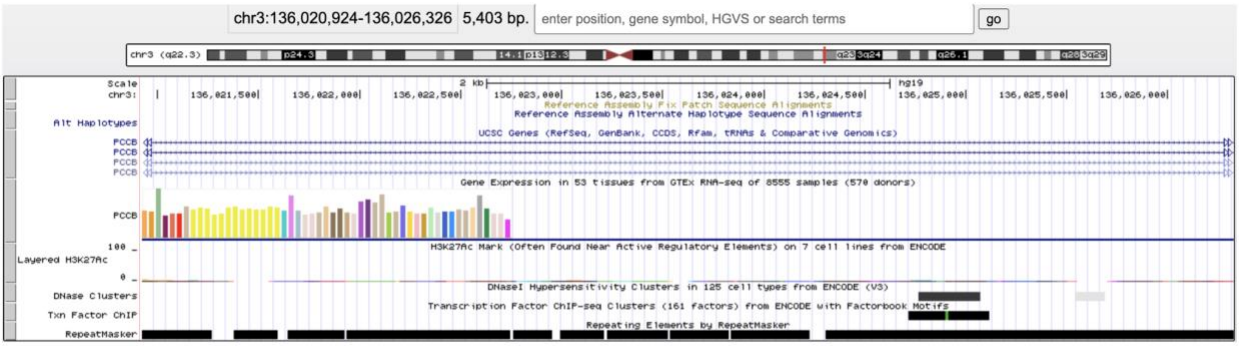


Figure S7 The difference between haplotype-based age estimation and expected age under neutral evolution for the 20 structural variants with phenotypic effects (**table 2, fig. 4D**)

esv3597888



esv3587631

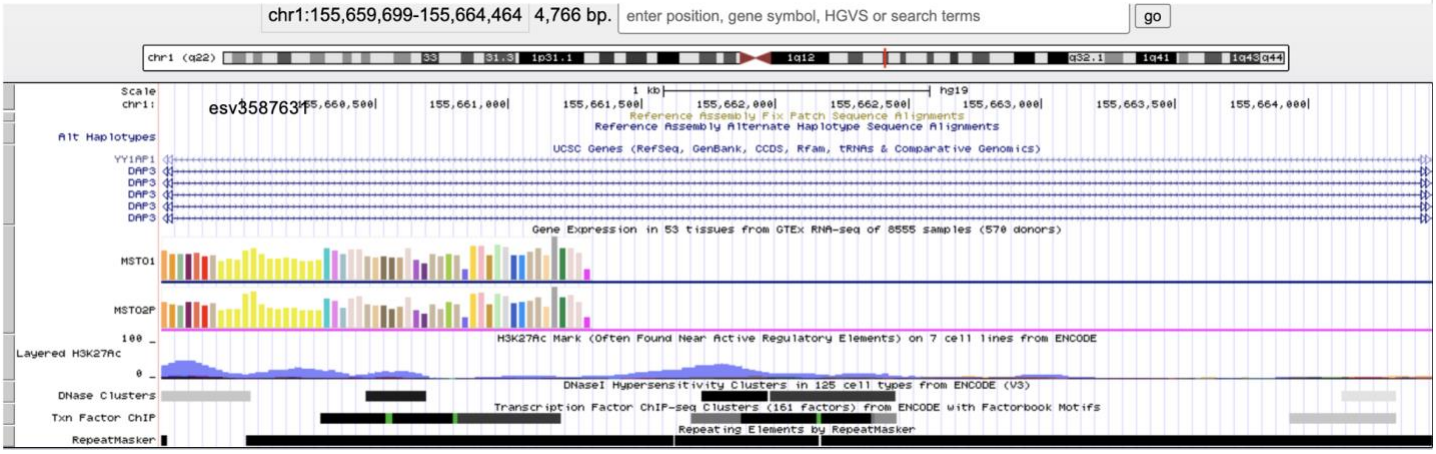


Figure S8 The overlap between functional annotation and structural variants

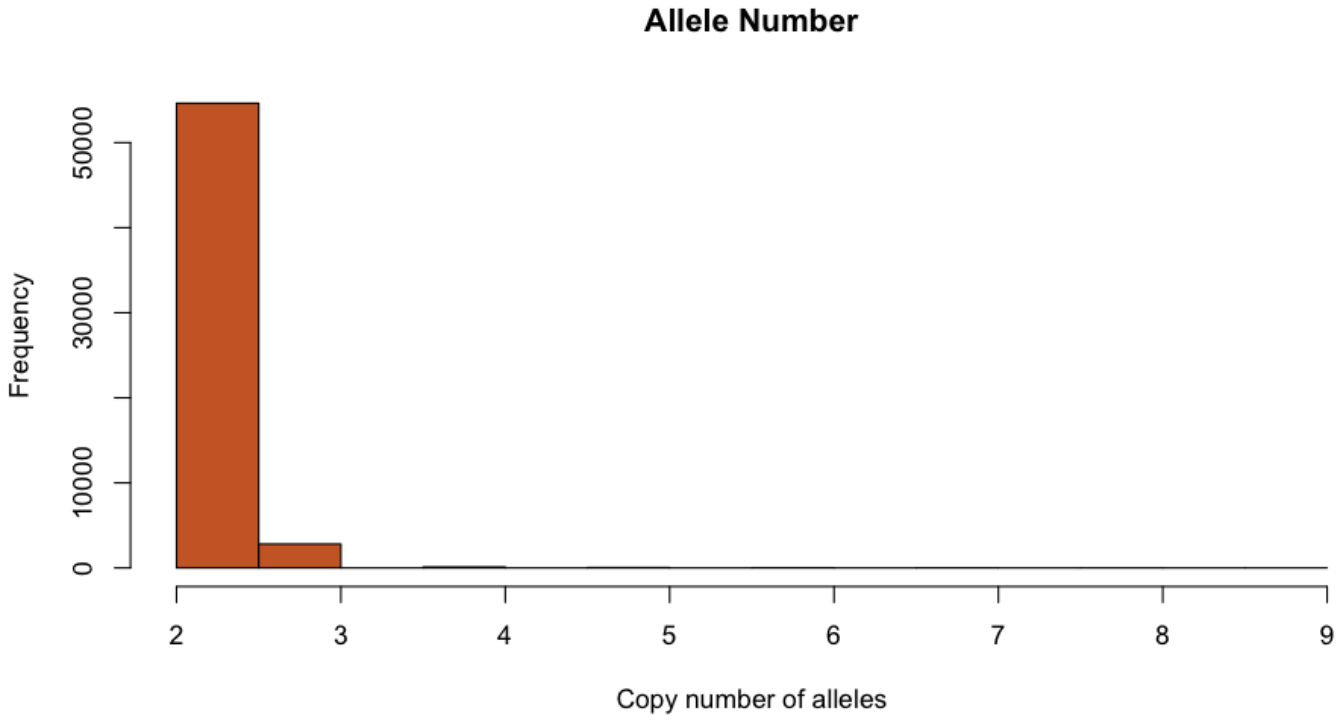


Figure S9. Allele copy number distribution

References

Albers PK, McVean G. 2020. Dating genomic variants and shared ancestry in population-scale sequencing data. PLoS Biol. 18:e3000586.

Almarri MA et al. 2020. Population Structure, Stratification, and Introgression of Human Structural Variation. Cell. 182:189–199.e15.

Anagnou NP, Antonarakis SE, O'Brien SJ, Modi WS, Nienhuis AW. 1988. Chromosomal localization and racial distribution of the polymorphic human dihydrofolate reductase pseudogene (DHFRP1). Am. J. Hum. Genet. 42:345–352.

Anderson-Trocmé L et al. 2020. Legacy Data Confound Genomics Studies. Mol. Biol. Evol. 37:2–10.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution. 16:37–48. doi: 10.1093/oxfordjournals.molbev.a026036.

Basu Mallick C et al. 2013. The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. PLoS Genet. 9:e1003912.

Bergey CM et al. 2018. Polygenic adaptation and convergent evolution on growth and cardiac genetic pathways in African and Asian rainforest hunter-gatherers. Proc. Natl. Acad. Sci. U. S. A. 115:E11256–E11263.

Bergström A et al. 2020. Insights into human genetic variation and population history from 929 diverse genomes. Science. 367. doi: 10.1126/science.aay5012.

Bhattacharyya. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math. Soc. 35:99–109.

Biddanda A, Rice DP, Novembre J. 2020. Geographic patterns of human allele frequency variation: a variant-centric perspective. 2020.07.01.182311. doi: 10.1101/2020.07.01.182311.

Boettger LM et al. 2016. Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat. Genet. 48:359–366.

Brazma A. 2019. F1000Prime recommendation of The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. F1000 - Post-publication peer review of the biomedical literature. doi: 10.3410/f.726761304.793561167.

Byrska-Bishop M et al. 2021. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv. 2021.02.06.430068. doi: 10.1101/2021.02.06.430068.

Candiotti KA et al. 2005. The impact of pharmacogenomics on postoperative nausea and vomiting: do CYP2D6 allele copy number and polymorphisms affect the success or failure of ondansetron prophylaxis? Anesthesiology. 102:543–549.

Carvalho CMB, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. Nat. Rev. Genet. 17:224–238.

Casewell NR, Jackson TNW, Laustsen AH, Sunagar K. 2020. Causes and Consequences of Snake Venom Variation. Trends Pharmacol. Sci. 41:570–581.

Cha S-H, Srihari SN. 2002. On measuring the distance between histograms. Pattern Recognit. 35:1355–1370.

de Cid R et al. 2009. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat. Genet. 41:211–215.

Clement M, Snell Q, Walker P, Posada D, Crandall K. 2002. TCS: estimating gene genealogies. In: Parallel and Distributed Processing Symposium, International.Vol. 2 pp. 0184–0184.

Conrad DF et al. 2010. Origins and functional impact of copy number variation in the human genome. Nature. 464:704–712.

Coop G et al. 2009. The role of geography in human adaptation. PLoS Genet. 5:e1000500.

Crawford NG et al. 2017. Loci associated with skin pigmentation identified in African populations. Science. 358. doi: 10.1126/science.aan8433.

Danecek P et al. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156–2158.

Deng L, Xu S. 2018. Adaptation of human skin color in various populations. Hereditas. 155:1.

Dennis MY, Eichler EE. 2016. Human adaptation and evolution by segmental duplication. Curr. Opin. Genet. Dev. 41:44–52.

Denoeud F et al. 2014. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science. 345:1181–1184.

Ding K et al. 2013. Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: an electronic medical record-based genome-wide association study. G3 . 3:1061–1068.

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB. 2016. Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data. Mol. Biol. Evol. 33:1082–1093.

Ge SX, Jung D, Yao R. 2020. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. 36:2628–2629.

Gravel S et al. 2011. Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. U. S. A. 108:11983–11988.

Greber BJ, Ban N. 2016. Structure and Function of the Mitochondrial Ribosome. Annu. Rev. Biochem. 85:103–132.

GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45:580–585.

Haller BC, Messer PW. 2019. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. Mol. Biol. Evol. 36:632–637.

Hancock AM et al. 2010. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proc. Natl. Acad. Sci. U. S. A. 107 Suppl 2:8924–8930.

Handsaker RE et al. 2015. Large multiallelic copy number variations in humans. Nat. Genet. 47:296–303.

Handsaker RE, Korn JM, Nemesh J, McCarroll SA. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat. Genet. 43:269–276.

Hebbring SJ, Moyer AM, Weinshilboum RM. 2008. Sulfotransferase gene copy number variation: pharmacogenetics and function. Cytogenet. Genome Res. 123:205–210.

Hernandez RD et al. 2011. Classic selective sweeps were rare in recent human evolution. Science. 331:920–

924.

Hollox EJ, Zuccherato LW, Tucci S. 2021. Genome structural variation in human evolution. Trends Genet. doi: 10.1016/j.tig.2021.06.015.

Ho SS, Urban AE, Mills RE. 2020. Structural variation in the sequencing era. Nat. Rev. Genet. 21:171–189.

Hsieh P et al. 2019. Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. Science. 366. doi: 10.1126/science.aax2083.

Huerta-Sánchez E et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 512:194–197.

Hurles ME, Dermitzakis ET, Tyler-Smith C. 2008. The functional impact of structural variation in humans. Trends Genet. 24:238–245.

Key FM, Teixeira JC, de Filippo C, Andrés AM. 2014. Advantageous diversity maintained by balancing selection in humans. Curr. Opin. Genet. Dev. 29:45–51.

Kidd JM et al. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell. 143:837–847.

Kim H-J, Maiti P, Barrientos A. 2017. Mitochondrial ribosomes in cancer. Semin. Cancer Biol. 47:67–81.

Kim HY, Cho S, Yu J, Sung S, Kim H. 2010. Analysis of copy number variation in 8,842 Korean individuals reveals 39 genes associated with hepatic biomarkers AST and ALT. BMB Rep. 43:547–553.

Kimura M, Ohta T. 1973. The age of a neutral mutant persisting in a finite population. Genetics. 75:199–212.

Kimura R et al. 2009. A common variation in EDAR is a genetic determinant of shovel-shaped incisors. Am. J. Hum. Genet. 85:528–535.

Ko W-Y et al. 2013. Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. Am. J. Hum. Genet. 93:54–66.

Kronenberg ZN et al. 2015. Wham: Identifying Structural Variants of Biological Consequence. PLoS Comput. Biol. 11:e1004572.

Leigh JW, Bryant D. 2015. popart: full-feature software for haplotype network construction. Methods Ecol. Evol. 6:1110–1116.

Levina E, Bickel P. 2001. The Earth Mover's distance is the Mallows distance: some insights from statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001.Vol. 2 pp. 251–256 vol.2.

Li H et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

Lin Y-L, Gokcumen O. 2019. Fine-Scale Characterization of Genomic Structural Variation in the Human Genome Reveals Adaptive and Biomedically Relevant Hotspots. Genome Biol. Evol. 11:1136–1151.

Lin Y-L, Pavlidis P, Karakoc E, Ajay J, Gokcumen O. 2015. The evolution and functional impact of human deletion variants shared with archaic hominin genomes. Mol. Biol. Evol. 32:1008–1019.

Lupski JR. 2015. Structural variation mutagenesis of the human genome: Impact on disease and evolution. Environ. Mol. Mutagen. 56:419–436.

Mahmoud M et al. 2019. Structural variant calling: the long and the short of it. Genome Biol. 20:246.

Marciniak S, Perry GH. 2017. Harnessing ancient genomes to study the history of human adaptation. Nat. Rev. Genet. doi: 10.1038/nrg.2017.65.

Martin AR et al. 2017. An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. Cell. 171:1340–1353.e14.

Mathieson S, Mathieson I. 2018. FADS1 and the Timing of Human Adaptation to Agriculture. Mol. Biol. Evol. 35:2957–2970.

McCarroll SA, Hadnott TN, Perry GH. 2005. Common deletion polymorphisms in the human genome. Nature. 38:86–92.

McLean CY et al. 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28:495–501.

Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. Trends Ecol. Evol. 35:561–572.

Messer PW. 2013. SLiM: simulating evolution with selection and linkage. Genetics. 194:1037–1039.

Mills RE et al. 2011. Mapping copy number variation by population-scale genome sequencing. Nature. 470:59–65.

Mou C et al. 2008. Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form. Hum. Mutat. 29:1405–1411.

Mukamel RE et al. 2021. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Science. 373:1499–1505.

Norton HL et al. 2007. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. Mol. Biol. Evol. 24:710–722.

Pajic P et al. 2019. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. Elife. 8. doi: 10.7554/eLife.44628.

Pajic P, Lin Y-L, Xu D, Gokcumen O. 2016. The psoriasis-associated deletion of late cornified envelope genes LCE3B and LCE3C has been maintained under balancing selection since Human Denisovan divergence. BMC Evol. Biol. 16:265.

Pang AW et al. 2010. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 11:R52.

Patterson NJ. 2005. How old is the most recent ancestor of two copies of an allele? Genetics. 169:1093–1104.

Payer LM et al. 2017. Structural variants caused by Alu insertions are associated with risks for many human diseases. Proc. Natl. Acad. Sci. U. S. A. 114:E3984–E3992.

Pedregosa F et al. 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 12:2825–2830.

Pérez-Barbería FJ, Shultz S, Dunbar RIM. 2007. Evidence for coevolution of sociality and relative brain size in three orders of mammals. Evolution. 61:2811–2821.

Perry GH et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nat. Genet. 39:1256–1260.

Polley S et al. 2015. Evolution of the rapidly mutating human salivary agglutinin gene (DMBT1) and population

subsistence strategy. Proc. Natl. Acad. Sci. U. S. A. 112:5105–5110.

Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. 2018. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. Elife. 7. doi: 10.7554/eLife.36317.

Prüfer K et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 505:43–49.

Pybus M et al. 2014. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res. 42:D903–9.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842.

Quinlan AR, Hall IM. 2012. Characterizing complex structural variation in germline and somatic genomes. Trends Genet. 28:43–53.

Radke DW, Lee C. 2015. Adaptive potential of genomic structural variation in human and mammalian evolution. Brief. Funct. Genomics. 14:358–368.

Ramachandran S et al. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. U. S. A. 102:15942–15947.

Rausch T et al. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 28:i333–i339.

Redon R et al. 2006. Global variation in copy number in the human genome. Nature. 444:444–454.

Rees JS, Castellano S, Andrés AM. 2020. The Genomics of Human Local Adaptation. Trends Genet. 0. doi: 10.1016/j.tig.2020.03.006.

Reich D et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 468:1053–1060.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature. 461:489–494.

Rothman N et al. 2010. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat. Genet. 42:978–984.

Sabeti PC et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature. 449:913–918.

Saitou M et al. 2021. Sex-specific phenotypic effects and evolutionary history of an ancient polymorphic deletion of the human growth hormone receptor. Sci Adv. 7:eabi4476.

Saitou M, Gokcumen O. 2020. An Evolutionary Perspective on the Impact of Genomic Copy Number Variation on Human Health. J. Mol. Evol. 88:104–119.

Saitou M, Satta Y, Gokcumen O. 2018. Complex Haplotypes of GSTM1 Gene Deletions Harbor Signatures of a Selective Sweep in East Asian Populations. G3. doi: 10.1534/g3.118.200462.

Saitou M, Satta Y, Gokcumen O, Ishida T. 2018. Complex evolution of the GSTM gene family involves sharing of GSTM1 deletion polymorphism in humans and chimpanzees. BMC Genomics. 19:293.

Schaffner SF et al. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15:1576–1583.

Scheinfeldt LB, Tishkoff SA. 2013. Recent human adaptation: genomic approaches, interpretation and insights. Nat. Rev. Genet. 14:692–702.

Schrider DR et al. 2013. Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet. 9:e1003242.

Schrider DR, Kern AD. 2017. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. Mol. Biol. Evol. 34:1863–1877.

Sekar A et al. 2016. Schizophrenia risk from complex variation of complement component 4. Nature. 530:177–183.

Sirugo G, Williams SM, Tishkoff SA. 2019. The Missing Diversity in Human Genetic Studies. Cell. 177:1080.

Smith GD et al. 2009. Lactase persistence-related genetic variant: population substructure and health outcomes. Eur. J. Hum. Genet. 17:357–367.

Sudmant PH, Rausch T, et al. 2015. An integrated map of structural variation in 2,504 human genomes. Nature. 526:75–81.

Sudmant PH, Mallick S, et al. 2015. Global diversity, population stratification, and selection of human copy-number variation. Science. 349:aab3761.

Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics. 135:599–607.

Teixeira JC et al. 2015. Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees, and bonobos. Mol. Biol. Evol. 32:1186–1196.

Tenesa A et al. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17:520–526.

The 1000 Genomes Project Consortium et al. 2015. A global reference for human genetic variation. Nature. 526:68.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:0446–0458.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 38:1358–1370.

Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. 14:125–138.

Wickham H. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Springer Publishing Company, Incorporated.

Wilde S et al. 2014. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc. Natl. Acad. Sci. U. S. A. 111:4832–4837.

Wongkittichote P, Ah Mew N, Chapman KA. 2017. Propionyl-CoA carboxylase - A review. Mol. Genet. Metab. 122:145–152.

Wu S et al. 2016. Genome-wide scans reveal variants at EDAR predominantly affecting hair straightness in Han Chinese and Uyghur populations. Hum. Genet. 135:1279–1286.

Xu D, Jaber Y, Pavlidis P, Gokcumen O. 2017. VCFtoTree: a user-friendly tool to construct locus-specific

alignments and phylogenies from thousands of anthropologically relevant genome sequences. BMC Bioinformatics. 18:426.

Xue Y et al. 2008. Adaptive evolution of UGT2B17 copy-number variation. Am. J. Hum. Genet. 83:337–346.

Yan SM et al. 2021. Local adaptation and archaic introgression shape global diversity at human structural variant loci. eLife. 10. doi: 10.7554/elife.67615.

Zhu H-J, Markowitz JS. 2013. Carboxylesterase 1 (CES1) genetic polymorphisms and oseltamivir activation. Eur. J. Clin. Pharmacol. 69:733–734.