Local adaptation and archaic introgression shape global diversity at human structural variant loci

Stephanie M. Yan1, Rachel M. Sherman2, Dylan J. Taylor1, Divya R. Nair1, Andrew N. Bortvin1, Michael C. Schatz1,2, Rajiv C. McCoy1*

1Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA
2Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
*Correspondence to: rajiv.mccoy@jhu.edu

Abstract

Large genomic insertions, deletions, and inversions are a potent source of functional and fitness-altering variation, but are challenging to resolve with short-read DNA sequencing alone. While recent long-read sequencing technologies have greatly expanded the catalog of structural variants (SVs), their costs have so far precluded their application at population scales. Given these limitations, the role of SVs in human adaptation remains poorly characterized. Here, we used a graph-based approach to genotype 107,866 long-read-discovered SVs in short-read sequencing data from diverse human populations. We then applied an admixture-aware method to scan these SVs for patterns of population-specific frequency differentiation—a signature of local adaptation. We identified 220 SVs exhibiting extreme frequency differentiation, including several SVs that were among the lead variants at their corresponding loci. The top two signatures traced to separate insertion and deletion polymorphisms at the immunoglobulin heavy chain locus, together tagging a 325 Kbp haplotype that swept to high frequency and was subsequently fragmented by recombination. Alleles defining this haplotype are nearly fixed (60-95%) in certain Southeast Asian populations, but are rare or absent from other global populations composing the 1000 Genomes Project. Further investigation revealed that the haplotype closely matches with sequences observed in two of three high-coverage Neanderthal genomes, providing strong evidence of a Neanderthal-introgressed origin. This extraordinary episode of positive selection, which we infer to have occurred between 1700 and 8400 years ago, corroborates the role of immune-related genes as prominent targets of adaptive archaic introgression. Our study demonstrates how combining recent advances in genome sequencing, genotyping algorithms, and population genetic methods can reveal signatures of key evolutionary events that remained hidden within poorly resolved regions of the genome.

Introduction

Rapid global dispersal and cultural evolution have exposed modern humans to a striking diversity of environments, to which they have developed numerous genetic adaptations (Fan et al., 2016). The vast majority of genomic research on human adaptation has focused on single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels) due to their ease of discovery by high-throughput short-read DNA sequencing (e.g., Frazer et al., 2007; The 1000
Genomes Project Consortium, 2015). This focus overlooks the potential impacts of larger insertions, deletions, and inversions—collectively termed structural variants (SVs)—whose abundance and complexity are now being appreciated with the advent of long-read DNA sequencing (Chaisson et al., 2015; Sedlazeck et al., 2018a; Audano et al., 2019). SVs impact more total sequence per genome than SNPs and may severely disrupt genes and regulatory elements to confer large effects on gene expression (GTEx Consortium et al., 2017). The size, modularity, and functional potential of SVs may offer a rich substrate for natural selection, as supported by several specific examples of adaptive insertions and deletions in diet and pigmentation-related genes (Perry et al., 2007; Kothapalli et al., 2016; Saitou and Gokcumen, 2019; Hsieh et al., 2019), as well as genome-wide evidence from primarily short-read data (Sudmant et al., 2015; Almarri et al., 2020).

More comprehensive analysis of SV evolution has been limited by the long and repetitive nature of many SVs. Short-read approaches to SV discovery depend on abnormalities in depth of coverage or other characteristics of read alignments (Kosugi et al., 2019) but achieve sensitivity of only 10-50% (Chaisson et al., 2019; Jakubosky et al., 2020). In contrast, the recent development of long-read sequencing methods has revolutionized SV discovery, achieving high sensitivity and specificity by spanning SV sequences and their unique flanking regions (Sedlazeck et al., 2018a). This application of long-read sequencing to human samples has dramatically expanded the catalog of known human variation, generating databases of hundreds of thousands of SVs discovered in globally diverse individuals (Audano et al., 2019; Ebert et al., 2020). Yet with rare exceptions (Beyter et al., 2019), these sequencing methods remain impractical for population-scale applications due to their high costs and low throughput.

We addressed this limitation by applying a graph-based method to genotype a catalog of long-read discovered SVs in short-read sequencing data from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015). By intersecting these genotype data with existing RNA-seq data from human cell lines, we discovered 1121 significant associations between SVs and the expression of nearby genes, including several components of the human leukocyte antigen (HLA) gene cluster and its interaction partners. We then applied an admixture-aware method to identify 220 SVs that exhibit strong allele frequency differentiation across human populations. Among the top hits, we discovered an extreme signature of local adaptation tagged by separate insertion and deletion polymorphisms at the immunoglobulin heavy chain locus, which encodes key components of the adaptive immune system. Alternative alleles defining this haplotype achieve high frequencies in certain Southeast Asian populations, but are rare or absent from other global populations. Searching for signatures of archaic introgression within our set of highly differentiated SVs further revealed evidence that the adaptive haplotype entered the modern human population via ancient hybridization with Neanderthals. Together, our work highlights the role of structurally complex and repetitive regions of the genome as hidden sources of functional diversity and evolutionary innovation in the hominin lineage.
Results

Graph genotyping of structural variation

To assess the role of SVs in human adaptive evolution, we sought to combine the accuracy of long-read sequencing with the scale and population diversity of short-read sequencing data. To this end, we used the graph genotyping software Paragraph (Chen et al., 2019) to genotype a set of 107,866 SVs, discovered from long-read sequencing data we reanalyzed from 15 individuals from five continents (Audano et al., 2019), in 2,504 high-coverage (30x) short-read sequenced samples from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) (Fig. 1A; see Methods). Results from recent benchmarking using ground truth data from the Genome in a Bottle Consortium (Zook et al., 2019) support the accuracy of Paragraph compared to other SV genotyping tools (Hickey et al., 2020). Paragraph achieves such accuracy by generating graph representations of SV loci, which include diverging paths for known alternative alleles such as SVs. Short reads are aligned to the graph along the path of best fit, facilitating genotyping even in structurally complex and repetitive regions. Informed by a large catalog of candidate SV alleles discovered by long-read sequencing, graph genotyping thus permits the study of variants that would be difficult or impossible to discover with short-read data alone (Sibbesen et al., 2018; Chen et al., 2019; Hickey et al., 2020; Sirén et al., 2020).

As quality control, we filtered the resulting data based on genotyping rates and adherence to Hardy Weinberg equilibrium, in accordance with (Chen et al., 2019; see Methods). Specifically, we removed SVs that were not successfully genotyped in \( \geq 50\% \) of samples, as well as SVs that violated one-sided Hardy Weinberg equilibrium expectations (excess of heterozygotes) in more than half of the 1000 Genomes populations. The latter scenario, whereby nearly all individuals are genotyped as heterozygous, is a common artifact caused by slightly divergent repetitive sequences that are falsely interpreted as alternative alleles at a single locus (Graffelman et al., 2017). These filtering steps removed 15,580 SVs, leaving 92,286 variants for downstream analysis.

Among this remaining set, global alternative allele frequencies were strongly correlated with the number of long-read-sequenced samples in which they were originally discovered (Audano et al., 2019), broadly supporting the accuracy of our graph genotyping results (Fig. S1). A total of 25,201 (27.3\%) alternative SV alleles were absent from all 1000 Genomes samples, likely reflecting a combination of false negatives and ultra-rare variation within the panel of long-read sequenced genomes. Such an abundance of rare variation is a known feature of human populations, which have experienced rapid demographic expansion over recent history (Keinan and Clark, 2012). In contrast, a total of 1139 (1.2\%) alternative SV alleles were observed to be fixed in the 1000 Genomes sample and likely reflect a combination of assembly errors and rare variation in the reference genome. Relaxing the criterion for fixation to 90\% alternative allele frequency (thus allowing for a modest rate of false negatives) resulted in a total of 4947 (5.4\%) fixed or nearly-fixed alternative SV alleles.

The frequency spectrum of segregating variation was meanwhile shaped by an ascertainment bias caused by variant discovery within a small sample and genotyping within a separate and larger sample. This bias, which is similar to that previously described for genotyping microarrays
(Lachance and Tishkoff, 2013), is characterized by an apparent depletion of rare variation, given that such variants are not shared with the discovery sample. While important to note, our study should be largely unaffected by this bias, due to our focus on individual variants rather than the shape of the allele frequency spectrum. Moreover, positively selected variants are by definition locally common and thus enriched within a small but globally diverse sample (Audano et al., 2019).

A. Graph construction

![Graph construction diagram]

B. Graph alignment and genotyping

![Graph alignment and genotyping diagram]

C. SVs distribution by population

![SVs distribution by population graph]
Figure 1. Variant graph genotyping of SVs with Paragraph. 

A. Genotyping of SVs was performed using a graph-based approach that represents the reference and alternative alleles of known SVs as edges. The SVs used for graph construction were originally identified from long-read sequencing of 15 individuals (Audano et al., 2019). 

B. At candidate SV loci, samples sequenced with short reads are aligned to the graph along the path of best fit, and individuals are genotyped as heterozygous (middle), homozygous for the reference allele (right), or homozygous for the alternative allele (not depicted). We applied this method to the 1000 Genomes dataset to generate population-scale SV genotypes. 

C. Allele frequency spectra of SVs genotyped with Paragraph. The left-most bin represents SVs where the alternative allele is absent from the 1000 Genomes sample (AF = 0). Samples are stratified by their 1000 Genomes superpopulation.

Among SVs that were polymorphic within the 1000 Genomes sample, we observed a negative correlation between SV length and minor allele frequency (Kendall’s τ = -0.140, p-value < 1 x 10^{-10}), which is an expected consequence of purifying selection. We similarly observed that deletions segregated at lower average minor allele frequencies than insertions (β = -0.555, p-value < 1 x 10^{-10}).

Quantifying linkage disequilibrium with known SNPs and short indels

We next investigated the extent of linkage disequilibrium (LD) between the catalog of long-read discovered SVs and known SNPs and short indels from the 1000 Genomes project. For each of the 26 populations, we computed the maximum observed LD between each SV and the nearest 100 variants within a 1 Mb window. Depending on the population subject to measurement, 36-41% of segregating SVs possessed an \( r^2 > 0.8 \) with any known SNP or short indel (Fig. S3). Levels of LD were lowest for African populations, in accordance with known patterns of haplotype structure (Conrad et al., 2006). We emphasize that these observations are strongly affected by the challenge of small variant discovery and genotyping in repetitive regions of the genome that are enriched for SVs. Specifically, 50,917 of all SVs (47.2%) intersect at least partially with one or more genomic intervals deemed inaccessible based on the 1000 Genomes pilot mask, and 5,214 SVs (4.8%) occur within regions of the genome identified as problematic by the ENCODE Consortium (Amemiya et al., 2019). Nevertheless, LD with known SNPs and indels is a meaningful metric for our study, because it quantifies the extent to which SVs represent independent and unexplored variation relative to previous evolutionary studies. Low observed levels of LD between a substantial fraction of SVs and known variants presents an opportunity to discover novel functional associations and signatures of adaptation at loci poorly tagged by easily genotyped markers.

Expression quantitative trait locus (eQTL) mapping

Seeking to first test the functional impacts of common structural variation, we intersected the SV genotype data with RNA-seq data from an overlapping set of 441 samples from the Geuvadis Consortium, which was generated from lymphoblastoid cell lines (LCLs) derived from individuals
from four European and one African population (The Geuvadis Consortium et al., 2013). We tested for associations between levels of gene expression and genotypes for SVs within 1 Mb from the transcription start site (TSS). After filtering the data on genotyping call rate, minor allele frequency, and gene expression level (see Methods), we identified a total of 1121 SV-gene pairs with significant gene expression associations at a 10% false discovery rate (FDR; Fig. 2A), broadly consistent with expectations from previous eQTL studies when scaled to the number of tested SVs (GTEx Consortium et al., 2017). SVs with significant impacts on expression tended to occur near the genes that they regulate, with 62% of significant SV eQTLs occurring within 100 Kb of the corresponding TSS (Fig. 2B).

Top gene expression associations include several genes in the HLA complex (HLA-DQB2, HLA-DQA2, HLA-DRB6, and HLA-P), as well as ERAP2, which encodes an endoplasmic reticulum aminopeptidase that processes HLA ligands to lengths suitable for their binding (Fig. 2C). While gene expression data from LCLs provides a limited snapshot of the functional implications of SVs, it is notable that these immune loci—which are known targets of balancing and diversifying selection (Hughes and Nei, 1988; Parham and Ohta, 1996; Andrés et al., 2010)—also exhibit strong gene expression diversity mediated by genetic variation.

Figure 2. eQTL mapping of SVs. We used RNA-seq data from the Geuvadis Consortium (The Geuvadis Consortium et al., 2013), obtained from LCLs derived from individuals from four European and one African population of the 1000 Genomes dataset, to test for associations between SV genotypes and gene expression. SV-eQTL pairs that were significant at a 10% FDR are depicted in purple. A. Q-Q plot of permutation p-values for all SV-gene pairs tested. B. Distribution of the distance of significant SV eQTLs from the transcription start site (TSS) of their associated genes. C. Volcano plot of eQTLs and the estimated effect of the alternative allele on expression ($\beta$).

Admixture-aware scan for signatures of local adaptation

Approaches to locus-specific scans for local adaptation can be broadly classified into two categories: frequency differentiation-based and LD-based methods (Vitti et al., 2013). While powerful, LD-based methods generally require haplotype phasing and/or dense and accurate
genotyping of the region subject to analysis—a requirement that is infeasible for many of the complex and repetitive regions of the genome investigated in our study. In contrast, frequency differentiation-based approaches can be applied to individual loci and are based on the logic that positive selection tends to cause a particular allele to increase in frequency only in the population(s) where it became established and was advantageous. Because most demographic events will affect the genome in its entirety, outlier loci exhibiting extreme levels of frequency differentiation compared to the genome-wide average serve as candidate targets of local adaptation. Common examples of frequency differentiation-based metrics include Wright’s fixation index \((F_{ST})\) (Wright, 1949), as well as tree-based extensions of this concept such as the locus-specific branch length (LSBL) (Shriver et al., 2004) and population branch statistic (PBS) (Yi et al., 2010). While useful for polarizing frequency changes on particular lineages, these tree-based tests still require the specification of (typically three) populations for comparison. The number of possible comparisons thus grows in a combinatorial manner with the number of populations in the study. Specifically, for the 26 populations of the 1000 Genomes Project, there are 2600 (26 choose 3) possible comparisons. A second limitation of such tests is the definition of population, which may or may not reflect genetic patterns of population structure that occur at multiple scales. Moreover, many human populations exhibit substantial admixture, which is ignored by, and may confound, some frequency differentiation-based tests.

To overcome these limitations, we used Ohana (Cheng et al., 2017; Ilardo et al., 2018; Cheng et al., 2019), a maximum likelihood-based method that models individuals as possessing ancestry from combinations of \(k\) ancestry components, inspired by related methods (Alexander et al., 2009; Falush et al., 2003). The method then tests whether individual variants adhere to this genome-wide null model, or are better explained by an alternative model in which frequencies are allowed to vary in one or more populations by consequence of local adaptation. Following the precedent of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015), we modeled individual genomes as combinations of 8 ancestry components, replicating known patterns of population structure at continental scales as well as prominent signatures of admixture within a subset of populations (e.g., African ancestry in Southwest US [ASW] and admixed American populations [AMR]; Fig. 3). The results of this admixture analysis were qualitatively unaffected by the choice of \(k\) (Fig. S2).

Across all ancestry components, we identified 220 unique SVs exhibiting significant deviation from genome-wide patterns of allele frequency differentiation (99.9th percentile of matched distribution for SNPs and short indels; see Methods; Fig. 4; Table S1). These included 139 SVs with coordinates overlapping with annotated genes, of which 13 intersected with annotated exons. We also identified 7 SVs at these frequency-differentiated loci that were significant eQTLs based on our previous analysis of the Geuvadis LCL data. Only 119 (54.1%) highly differentiated SVs possessed strong LD with known SNPs or short indels \((r^2 > 0.8;\) Fig. S3), indicating that many of these loci constitute novel candidate targets of selection that may have been missed by previous scans.
Figure 3. Global patterns of admixture and population relationships. To conduct an admixture-aware scan for local adaptation, we used Ohana (Cheng et al., 2017; Iliardo et al., 2018; Cheng et al., 2019) to infer genome-wide patterns of ancestry in the 1000 Genomes samples. This method models each individual as a combination of $k$ ancestry components and then searches for evidence of local adaptation on these component lineages. A. Admixture proportions ($k = 8$) for all samples in the 1000 Genomes dataset, grouped by population. Vertical bars represent individual genomes. B. Tree representation of the relationships among the 8 ancestry components based on genome-wide covariances in allele frequencies.
Figure 4. Quantifying allele frequency differentiation among ancestry components. Using Ohana, we searched for evidence of local adaptation by testing whether the allele frequencies of individual variants were better explained by the genome-wide tree (Fig. 3B), or by an alternative tree (e.g., Fig. 5B) where allele frequencies were allowed to vary in one ancestry component. The likelihood ratio statistic (LRS) reflects the relative support for the latter selection hypothesis. For each ancestry component, SVs with LRS > 32 are colored by their maximum linkage disequilibrium ($r^2$) with any known SNP in the corresponding 1000 Genomes superpopulation.
Known and novel targets of local adaptation

Notable examples of highly differentiated loci included rs333, the Δ32 allele of the chemokine receptor CCR5, which is known to confer resistance to HIV infection and progression (Dean et al., 1996). Among our results, this deletion polymorphism is the 14th most frequency-differentiated SV with respect to ancestry component 3, which is highly represented in Europe. The CCR5-Δ32 allele segregates at moderate frequencies in European populations (MAF = 10.9%) and achieves its highest frequency of 15.6% in the Finnish population, but segregates at low frequencies elsewhere. The case for historical positive selection at this locus has been contentious, with initial studies citing a geographic cline in allele frequencies and strong LD with adjacent microsatellite markers (Stephens et al., 1998), potentially driven by epidemics such as the bubonic plague or smallpox (Galvani and Slatkin, 2003). However, subsequent studies argued that patterns of long-range LD (Sabeti et al., 2005) and temporal allele frequency changes based on ancient DNA (aDNA) samples (Bollback et al., 2008) could not exclude models of neutral evolution.

Our study also identified numerous novel hits, such as a 309 bp intronic insertion in TMEM131 (ancestry component 3), which segregates at a frequency of 63% in Finnish populations, but at a mean frequency of 24% in non-European populations. This gene encodes proteins involved in collagen cargo trafficking from the endoplasmic reticulum to the Golgi (Zhang et al., 2020). Collagen is the most abundant protein in the human body and the major component of human skin. Recurrent positive selection shaping skin pigmentation and other phenotypes is one of the best described examples of local adaptation across human populations (Jablonski, 2004; Crawford et al., 2017).

We also identified a 2.8 Kb insertion in an intron of CLEC16A, inferred to be under selection in ancestry component 4, which corresponds to the Peruvian (PEL), Mexican (MXL), and Colombian (CLM) populations of the 1000 Genomes Project. The insertion, which is poorly tagged by linked SNPs and short indels (maximum $r^2 = 0.26$; Fig. S4), segregates at frequencies of 52.4%, 38.3%, and 15.4%, respectively, in these three populations, but is rare in others (AF < 0.04). CLEC16A is thought to impact susceptibility to autoimmune disorders, and SNPs in this gene have been associated with diseases such as type 1 diabetes, multiple sclerosis, and rheumatoid arthritis (Pandey et al., 2019). Notably, this same SV was also identified in a recent study of structural variation, which similarly reported that it segregates at high frequency in Peruvians (Ebert et al., 2020).

In rare cases, multiple SVs in LD with one another captured the same underlying signature of frequency differentiation. One such example from South Asian populations (ancestry component 5) involved a linked intronic insertion and deletion in the cellular growth and morphogenesis related gene CSNK1G1. In this case, the reference genome carries the global minor allele, such that the signature of local adaptation presents as a lower frequency of the alternative allele in South Asian populations (60%-71% for 22980_HG00514_ins; 60%-72% for 25014_HG02106_del) compared to other global populations (91% and 92%, respectively).
Extreme signatures of adaptation at the immunoglobulin locus in Southeast Asian populations

The SV with the strongest evidence of local adaptation across all populations was an insertion polymorphism in an intron of *IGHG4*, which codes for a constant domain of the immunoglobulin heavy chain. These heavy chains pair with light chains, the latter of which include a domain composed of variable (V), diversity (D), and joining (J) segments. Complementing their substantial germline variation, V(D)J loci experience somatic recombination and hypermutation to generate vast antibody repertoires—the defining feature of the adaptive immune system (Watson et al., 2017). This insertion polymorphism identified by our scan exhibits strong allele frequency differentiation in ancestry component 2, which is highly represented in the Chinese Dai in Xishuangbanna, China (CDX) and Kinh in Ho Chi Minh City, Vietnam (KHV) populations, where it achieves frequencies of 65% and 88%, respectively, while remaining at much lower allele frequencies in other global populations (Fig. 5A, 5B).

The SV was originally reported as a 34 bp insertion based on long-read sequencing of the genome of a Vietnamese individual (HG02059) (Audano et al., 2019). Based on realignment to a modified version of the reference genome that includes the alternative allele, we revised the sequence of this insertion to 33 bp, but found that it is well supported by patterns of coverage, split reads and soft clipped alignments at the SV breakpoints (Fig. S5). The sequence of the insertion itself is repetitive within the human reference genome, with two identical copies of the sequence occurring ~44 and ~117 Kb downstream, respectively, also within the *IGH* gene cluster.

Interestingly, the second strongest signature of adaptation across all populations traced to a nearby 135 bp deletion, which overlaps with two transcription factor binding sites for *IGHE*, another component of the constant region of immunoglobulin heavy chains. This deletion, which lies only 33 Kb upstream of the *IGHG4* insertion, again achieves high frequencies in the CDX and KHV populations (70% and 54%, respectively) but segregates at low frequencies in most other global populations. Despite their genomic proximity and similar patterns of frequency differentiation, these two SVs exhibit only modest levels of LD ($r^2 = 0.17$), likely reflecting recombination occurring during or after the episode of selection.

Strength and timing of positive selection at the IGH locus

Despite the short size of the *IGHG4* insertion, its location in a structurally complex and repetitive region of the genome prevented its detection by traditional short-read sequencing approaches, and it was consequently not reported in the 1000 Genomes study of structural variation (The 1000 Genomes Project Consortium et al., 2015). In addition, the challenge of dense genotyping and phasing within this region hinders haplotype-based approaches for inferring the timing of selection. Nevertheless, global patterns of allele frequencies can be interpreted in the context of known population demographic histories, providing a rough estimate of such timing. For
example, pairwise divergence times among East Asian populations inferred by (Wang et al., 2018) constrain the plausible timing of selection at the \textit{IGH} locus between approximately 60 generations (1740 years) and 400 generations (11,600 years) ago (assuming a 29-year generation time [Tremblay and Vézina, 2000]): after the divergence of the Chinese Dai and Vietnamese populations from the Japanese (JPT) population, but before the divergence of CDX and KHV. These divergence time estimates are roughly consistent with those inferred using an alternative approach based on the joint allele frequency spectrum by Jouganous et al. (2017).

Figure 5. Local adaptation at the \textit{IGH} locus. A. Population-specific frequencies of the insertion allele in each of the 1000 Genomes populations, in the style of the Geography of Genetic Variants browser (Marcus and Novembre, 2017). B. Tree representation of the best-fit selection hypothesis for the \textit{IGHG4} insertion polymorphism, as computed by Ohana. C. Five-population demographic model used for simulation and parameter inference via approximate Bayesian computation (ABC). Population sizes and
split times are further described in the Methods. D. Posterior distribution of the selection coefficient ($s$). E. Posterior distribution of the timing of the onset of selection ($T_{\text{adaptive}}$). F. Posterior distribution of the initial allele frequency at the beginning of the simulation ($p_0$). G. Negative relationship between $s$ and $T_{\text{adaptive}}$ for simulations retained by ABC.

Seeking more formal estimates of the parameters of this episode of selection, we applied approximate Bayesian computation (ABC) to forward genetic simulations of a simplified five-population demographic model based on parameters obtained from the literature (The 1000 Genomes Project Consortium, 2015; Jouganous et al., 2017; Wang et al., 2018) (Fig. 5C). For simplicity, our model did not include migration, but we note that this omission should make our estimates of the selection coefficient conservative, as stronger selection is required to generate allele frequency differences between populations exchanging migrants. In line with our previous intuition, the results of this analysis indicated a recent onset of extremely strong selection, with the selection coefficient parameter ($s$) inferred to be 0.06 (95% credible interval [CI] [0.02, 0.12]), the selection onset time parameter ($T_{\text{adaptive}}$) inferred to be 4400 years ago (95% CI [1700, 8400]), and the initial frequency of the adaptive allele in the ancestral Eurasian population ($p_0$) to be 0.18 (95% CI [0.04, 0.35]) (Fig. 5D-F). We observed a strong correlation between estimates of the selection coefficient and timing of selection. Specifically, older, weaker selection could produce the same frequency differences as stronger and more recent selection (Fig. 5G), though even the lower bound on our estimate of $s$ places it among the strongest episodes of positive selection documented in humans.

Evidence of Neanderthal-introgressed origin of the high frequency IGH haplotype

Approximately 2% of the genome of all non-African individuals traces to admixture with Neanderthals between 47 and 65 Kya, while Oceanian populations (and Asian populations to a lesser extent) also possess sequences introgressed from Denisovans (Green et al., 2010; Sankararaman et al., 2012, 2016; Vernot et al., 2016). Introgressed alleles, including some SVs (Almarri et al., 2020; Hsieh et al., 2019), are thought to have conferred both beneficial and deleterious effects on modern human populations, especially with respect to immune-related phenotypes (Rotival and Quintana-Murci, 2020). Motivated by these findings, we tested our set of highly differentiated SVs for evidence of archaic hominin introgression. Using results from the method Sprime (Browning et al., 2018), which leverages patterns of divergence and haplotype structure to classify archaic introgressed sequences, we identified SVs that segregate in LD with putative introgressed SNPs (see Methods). Of the 220 highly differentiated SVs, 26 (12%) exhibited strong LD ($r^2 > 0.5$) with putative introgressed haplotypes while simultaneously exhibiting low allele frequencies (AF < 0.01) within African populations of the 1000 Genomes Project (Fig. S6; Table S2). Notably, this set of candidate introgressed SVs included the $IGHG4$ insertion and nearby deletion, which despite their low LD with one another each tag multiple putative introgressed SNPs within the CDX population. Indeed, the original Sprime publication reported that putative introgressed variants of $IGHA1$, $IGHG1$, and $IGHG3$ achieve high
frequencies in Eurasian populations, but the characteristics of this signature were not further examined, in part due to stringent filtering of aDNA genotypes in this genomic region (Browning et al., 2018).

**Figure 6. Evidence for Neanderthal introgression of the adaptive IGH haplotype.** A. Local analysis of likelihood ratio statistics (LRS) in the region near the 33 bp insertion (red point) reveals a 325 Kb haplotype encompassing 94 SNPs with strong allele frequency differentiation within ancestry component 2. Points where the alternative allele matches an allele observed in the Chagryskaya Neanderthal genome but at a frequency of 1% or less in African populations are highlighted in purple. B. Individual haplotypes defined by the highly differentiated SNPs (LRS > 450). Four archaic hominin genomes are plotted at the top, while 30 randomly sampled haplotypes from each of 6 populations from the 1000 Genomes project are plotted below. Archaic hominins samples are colored according to whether they possess one or more aligned reads supporting the alternative allele at a given site.

Most candidate introgressed SVs fall within complex regions of the genome, such that genotyping directly from fragmented and degraded aDNA data remains infeasible. However, the IGHG4 intronic insertion is short enough to be spanned by individual sequencing reads, thus
allowing us to apply a k-mer based approach to validate our genotyping results and further investigate introgression at this locus. Specifically, we identified a 48 bp sequence that is unique to individuals with the insertion, but not observed in the reference genome (see Methods), thereby facilitating rapid searches for the insertion allele in any sequencing dataset. Application of this approach to the 1000 Genomes data broadly validated our graph genotyping results and supported the strong allele frequency differences that we had previously described (Fig. S7). We then extended this scan to published short-read sequencing data from four high-coverage (~30x) archaic hominin genomes (Meyer et al., 2012; Prüfer et al., 2014, 2017; Mafessoni et al., 2020). While the k-mer was absent from the Denisovan genome, we observed one or more perfect matches in the sequences of each of three Neanderthals, which we found notable given its absence among African populations (Fig. S7).

We expanded our analysis to the genomic region around the IGHG4 insertion, investigating the archaic hominin allelic states at each of the 109 highly differentiated variants (2 SVs, 14 short indels, and 93 SNPs) defining the 325 Kb LRS peak at the IGH locus. Focusing on the 93 SNPs where archaic alleles are easily compared, we observed that the Denisovan genome exhibited a modest degree of matching (39 shared alleles [42%]), while the Altai, Vindija, and Chagyrskaya Neanderthal genomes exhibited near perfect matching over the entirety of the differentiated region (77 [83%], 88 [95%], and 89 [96%] shared alleles, respectively) (Fig. 6). Additionally, of these 93 highly differentiated SNPs, 64 were called as introgressed in CDX by Sprime (Fig. S8).

Combined with the near absence of these alleles from other global populations and the length of the differentiated haplotype, our observations strongly support the conclusion that this sequence originated via ancient introgression from a Neanderthal population related to the Chagyrskaya and Vindija Neanderthals.

Examination of the haplotype structure at the IGH locus revealed four discrete blocks of LD within the highly differentiated region (Fig. S9), consistent with substantial recombination after the original haplotype achieved high frequency. The deletion SV (22231_HG02059_del) falls within the largest LD block, while the insertion SV (22237_HG02059_ins) falls within a smaller LD block that exhibits the greatest allele frequency differences. The latter block includes the tag SNP rs150526114, where the global minor allele matches the Neanderthal genomes and segregates at 91% frequency in CDX, 73% frequency in KHV, and 59% frequency in CHS, but is rare or absent in most other populations from the 1000 Genomes Project. Data from the Human Genome Diversity Panel (HGDP) (Bergström et al., 2020) and Simons Genome Diversity Panel (SGDP) (Mallick et al., 2016) shed additional light on the geographic distribution of the putative Neanderthal-introgressed allele and confirmed its extreme pattern of frequency differentiation specific to Southeast Asian populations (Fig. S10). Notably, the allele is absent from the HGDP populations from the Americas, which are thought to have split from East Asian populations approximately 26 Kya (Moreno-Mayar et al., 2018), further supporting the recency and geographically restricted nature of this positive selection event.
Discussion

Long-read sequencing is starting to provide more comprehensive views of the landscape of human genetic variation, drawing novel links to phenotypes and diseases. However, long-read sequencing methods remain impractical for most population-scale studies due to their low throughput and high cost. We sought to overcome these limitations by applying variant graph genotyping of a large catalog of long-read-discovered SVs to short-read sequencing data from globally diverse individuals of the 1000 Genomes Project. By mapping eQTLs and scanning for evidence of local adaptation and adaptive introgression, we highlighted the role of SVs as largely unexplored contributors to variation in human genome function and fitness.

Despite the scale and diversity of SVs and samples used in our study, we anticipate that additional SV targets of selection remain undiscovered, either because they were not present in the set of long-read sequenced individuals, or because they remain inaccessible to genotyping using graph-based approaches (e.g., tandem repeats) (Chen et al., 2019). Studying the evolutionary impacts of such SVs will require the application of long read-sequencing at population scales (Ebert et al., 2020). Moreover, SV loci exhibiting expression associations or signatures of selection may not themselves be the causal targets, but rather tag nearby causal variation by consequence of LD. Methods such as fine mapping (Schaid et al., 2018) and multiplex reporter assays (Tewhey et al., 2016; van Arensbergen et al., 2019) will be invaluable for disentangling LD to reveal causal relationships and contrast the relative impacts of various forms of genetic variation.

The two strongest signatures of local adaptation in our study traced to the IGH locus. While the precise phenotypic impacts of these variants remain unknown (Fig. S11), their potential effects on adaptive immunity is intriguing given the established role of immune-related genes as common targets of local adaptation in human populations (Barreiro and Quintana-Murci, 2020). The human IGH locus is highly polymorphic (Mikocziova et al., 2020), with examples of SNPs and copy number variants exhibiting frequency differences between populations (Watson et al., 2013). In developing lymphocytes, this locus undergoes somatic V(D)J recombination and hypermutation to produce antibodies that drive the immune response—processes that may be influenced by nearby germline variation (Watson et al., 2017). The combination of these forms of variation makes the region difficult to probe with traditional sequencing methods, in turn highlighting the power of long-read sequencing and graph genotyping.

Our observation of the IGHG4 insertion within the Neanderthal genomes allowed us to connect and build upon anecdotal evidence of selection and introgression at the IGH locus. Specifically, the 1000 Genomes Project previously reported SNPs in several immunoglobulin genes as allele frequency outliers with respect to the CDX population (The 1000 Genomes Project Consortium, 2015). Browning et al. (2018) later noted that a Neanderthal-introgressed haplotype at the IGH locus achieves high frequencies in Eurasian populations, though the magnitude and population-specific nature of this selection event were not further investigated. Our findings add to the growing list of examples of adaptive introgression from archaic hominins (Huerta-Sánchez et al., 2014; Gittelman et al., 2016; Racimo et al., 2017; Hsieh et al., 2019), several of which are thought to have targeted immune-related phenotypes (Abi-Rached et al., 2011; Mendez et al.,
Based on forward genetic simulations, we estimated that selection on the introgressed $IGH$ haplotype initiated between 1700 and 8400 years ago, before the divergence of the CDX and KHV populations and in line with our intuition based on patterns of allele frequencies. This recent onset of selection is intriguing given that introgression from Neanderthals into the ancestors of Eurasian modern human populations dates to 47-65 Kya (Sankararaman et al., 2012). Our findings thus suggest that persisting archaic introgressed haplotypes provided a reservoir of functional variation to the ancestors of CDX and KHV that proved adaptive during a period of environmental change, for example in response to local pathogens (Rasmussen et al., 2015). Recent reports that Neanderthal-introgressed sequences mediate individual outcomes of SARS-CoV-2 infections to this day lend plausibility to this hypothesis (Zeberg and Pääbo, 2020a, 2020b) (Zhou et al., 2020), as does polygenic evidence of adaptation in response to ancient viral epidemics, including in East Asia (Souilmi et al., 2021).

Our simulations additionally demonstrated that a selection coefficient between 0.02 and 0.12 best explains the observed frequency differences, comparable to other episodes of strong selection in humans. These include examples such as lactase persistence mutations near $LCT$ (0.01-0.15) (Bersaglieri et al., 2004) and malaria resistance mutations affecting $DARC$ (0.08) (Hamid et al., 2021) and $HBB$ (0.1) (Elguero et al., 2015). While we caution against overinterpretation of these parameter estimates given the uncertainty in the underlying demographic model, our results are broadly consistent with the observation of extreme allele frequency differences among closely related populations. Future studies incorporating more complex evolutionary models and fully resolved $IGH$ haplotypes (Rodriguez et al., 2020) will be essential for further refining the evolutionary history of the immunoglobulin locus. Nevertheless, the coexistence of V(D)J recombination, somatic hypermutation, and local adaptation at this locus presents a remarkable example of diversifying selection at multiple scales of biological organization, generating allelic diversity both within individuals and across populations.

Together, our study demonstrates how new sequencing technologies and bioinformatic algorithms are facilitating understanding of complex and repetitive regions of the genome—a new frontier for human population genetics. Combined with studies of diverse populations, these technologies are providing a more complete picture of human genomes and the evolutionary forces by which they are shaped.

Methods

Graph genotyping of structural variation

We used published long-read sequencing data from 15 individuals to generate a set of 107,866 SVs for graph genotyping (Audano et al., 2019). Raw reads were downloaded using the accessions provided in the original publication and aligned with NGM-LR (Sedlazeck et al., 2012a, 2012b; Sams et al., 2016; Dannemann et al., 2016; Enard and Petrov, 2018; Gouy and Excoffier, 2020).
2018b) using default PacBio parameters to the main chromosomes of GRCh38. Variants were called with Sniffles (Sedlacek et al., 2018b), requiring a minimum SV length of 30 bp and a minimum of 10 supporting reads. Resulting VCFs were refined with Iris (Alonge et al., 2020) to polish the reported SV sequences. Variants were then merged with SURVIVOR v1.0.7 (Jeffares et al., 2017), using a merge distance of 50 bp and requiring strand and type to match. For each merged variant, a representative variant was then obtained from the original pre-merged call set to improve accuracy. Such representative variants were selected by first prioritizing homozygous over heterozygous calls, and then by prioritizing variants with greater proportions of reads supporting the non-reference allele. To prepare the variants for input into Paragraph, translocations, mitochondrial DNA variants, inversions and duplications over 5 Kb, and variants without a ‘PASS’ filter, were removed from the VCF. This resulted in a set of 107,866 SVs.

High-coverage (30x) short-read sequencing data for the core 2,504 individuals in the 1000 Genomes Project, sequenced by the New York Genome Center, was obtained from ENA (PRJEB31736). We genotyped SVs in these samples with Paragraph v2.2 (Chen et al., 2019). In accordance with Paragraph’s recommendations, we set the maximum permitted read count for variants to 20 times the mean sample depth in order to limit runtime for repetitive regions. Genotypes from all samples were combined using bcftools v1.9 (Danecek et al., 2020).

To obtain a high-quality set of genotyped SVs, we filtered the resulting data based on dataset-wide genotyping rates and within-population Hardy Weinberg equilibrium. We determined an SV’s overall genotyping rate with cvc2 (Pedersen and Quinlan, 2017) and removed variants that were not genotyped in ≥50% of samples. We additionally calculated one-sided Hardy-Weinberg equilibrium p-values (excess of heterozygotes) for variants within each of the 26 1000 Genomes populations, using the HardyWeinberg package from R (Graffelman, 2015). We filtered out SVs that violated equilibrium expectations (Fisher’s exact test, p < 1 x 10^-4) in ≥13 populations. Unfolded, within-population allele frequencies were calculated with PLINK v1.90b6.4 (Purcell et al., 2007).

**Calculating linkage disequilibrium with SNPs and short indels**

To calculate linkage disequilibrium (LD) between SVs and SNPs or short indels in the 1000 Genomes samples, we used small variant genotypes produced by the 1000 Genomes Consortium. These genotypes were generated by aligning the 1000 Genomes Project Phase 3 data to GRCh38 and then calling variants against the GRCh38 reference, and are restricted to biallelic SNVs and indels (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/). Y chromosome genotypes, not included in the former data release, were obtained from Phase 3 variant calls lifted over to GRCh38 (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/). We combined the 1000 Genomes SNP and indel genotypes with our SV genotypes and calculated $r^2$ between all variants within each population, using PLINK v1.90b6.4.
Genome accessibility masks from the 1000 Genomes Project were obtained from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/working/20160622_genome_mask_GRCh38/, and ENCODE blacklisted regions were obtained from https://www.encodeproject.org/files/ENCFF356LFX/.

**eQTL mapping**

To conduct eQTL analysis, we used gene expression data generated by the Geuvadis Consortium (The Geuvadis Consortium et al., 2013), which includes 447 intersecting samples from the following 1000 Genomes project populations: CEU, FIN, GBR, TSI, and YRI. Using the recount3 package from R/Bioconductor (Collado-Torres et al., 2017), we extracted gene expression counts for all corresponding samples. Counts were normalized across samples using the trimmed mean of M values (TMM) method from edgeR (Robinson et al., 2010). TPM values were also computed from raw counts. In accordance with the methods employed for eQTL mapping in the GTEx Project (GTEx Consortium, 2020), we retained all genes with TPM values greater than or equal to 0.1, as well as raw read counts greater than or equal to 6 in at least 20% of samples. We then applied rank normalization to the TMM values for each remaining gene. We then performed cis-eQTL mapping with a modified version of fastQTL (Ongen et al., 2016)(https://github.com/hall-lab/fastqtl), which accounts for SV size when determining the appropriate cis window. We conducted nominal and permutation passes with genotype principal components and sex included as covariates. Beta-distribution-approximated permutation p-values from fastQTL were used as input to estimate q-values and control the false discovery rate (FDR) with the qvalue package (Storey and Tibshirani, 2003; Storey et al., 2019).

**Admixture-aware scan for signatures of local adaptation**

We used the software package Ohana (Cheng et al., 2017; Ilardo et al., 2018; Cheng et al., 2019) to scan SVs for signatures of positive selection, i.e. extreme frequency differentiation between populations. Ohana uses allele frequency to model individuals as an admixed combination of \(k\) ancestral populations, constructing a genome-wide covariance matrix to describe the relationship between these ancestry components. Variants are then assessed to determine whether their allele frequencies are better explained by the genome-wide “neutral” matrix, or by an alternative matrix that allows allele frequencies to vary in one ancestry component.

In accordance with Ohana’s recommended workflow, we conducted admixture inference on the 1000 Genomes dataset with a set of \(~100,000\) variants, downsampling from chromosome 21 of the 1000 Genomes SNV/indel callset used for LD calculations above. Downsampling was performed with PLINK’s variant pruning function. Inference of admixture proportions in the dataset was allowed to continue until increased iterations produced qualitatively similar results.
(50 iterations). The covariance matrix generated from this downsampled dataset was used as a neutral input for downstream scans for selection on SVs.

In order to generate “selection hypothesis” matrices to search for selection in a specific ancestry component, we modified the neutral covariance matrix by allowing one component at a time to have a greater covariance. A scalar value of 10, representing the furthest possible deviation that a variant could have in a population under selection, was added to elements of the neutral covariance matrix depending on the population of interest. For each variant of interest, Ohana then computes the likelihood of the observed ancestry-component-specific allele frequencies under the selection and neutral models, then compares them by computing a likelihood ratio statistic (LRS), which quantifies relative support for the selection hypothesis. We filtered these results to remove extreme outliers in null model log-likelihoods (global log likelihood (LLE) < -1000), which were unremarkable in their patterns of allele frequency and instead indicated a failure of the neutral model to converge for a small subset of rare variants. To calculate p-values, we then compared the LRS to a chi-square distribution with one degree of freedom (Cheng et al., 2019) and adjusted for multiple hypothesis testing using a Bonferroni correction. To further refine the list of candidate selected loci, we also compared the ancestry component-specific LRS computed for each SV to the observed distribution of LRS computed from SNPs and short indels matched on global minor allele frequency (in 1% frequency bins). Specifically, we identified SVs with LRS exceeding the 99.9 percentile of the empirical LRS distribution for frequency-matched SNPs and short indels as calculated using identical methods. These background SNPs and indels were limited to chromosome 1 for computational efficiency, but results were qualitatively unaffected by the choice of chromosome.

**Inference of selection parameters with approximate Bayesian computation**

We used a sequential algorithm for approximate Bayesian computation (Lenormand et al., 2013; Pritchard et al., 1999), implemented with the R package EasyABC (Jabot et al., 2013), to infer the strength and timing of selection at the *IGHG4* locus, as well as the initial frequency of the adaptive allele. This approach consisted of drawing model parameters from prior distributions as input to the forward evolutionary simulation software package SLiM (Haller and Messer, 2019), computing summary statistics from each simulation and comparing to those observed from our data. Simulations with summary statistics most closely matching the observed data are then used to construct posterior distributions of the model parameters. The sequential algorithm automatically determines the tolerance level and uses a predetermined stopping criterion ($\rho_{accmin}$ = 0.05), thereby reducing the necessary number of simulations and improving estimates of the posterior distribution (Lenormand et al., 2013).

We constructed a simplified five-population demographic model based on parameters obtained from (Jouganous et al., 2017; Wang et al., 2018) as well as population size estimates from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) (Fig. 6A). Specifically, our model consisted of an initial divergence event between the CEU and East Asian populations at 46 Kya, subsequent divergence between the population ancestral to CHB/JPT and the
population ancestral to KHV/CHX at 9.8 Kya, and final splits between CHB/JPT and KHV/CHX at 9.0 Kya and 1.7 Kya, respectively. Each new population was drawn from its originating population at the same size of the originating population. The size of the ancestral population was set at 2831 (Jouganous et al., 2017) and allowed to expand exponentially at a rate of $1.25 \times 10^{-3}$ per generation, resulting in a final size of 17883 in each subpopulation, broadly consistent with pairwise sequential Markovian coalescent (PSMC) results presented by the 1000 Genomes Project for these populations (The 1000 Genomes Project Consortium, 2015). We allowed the initial allele frequency of the selected variant ($p_0$) to vary between 0 and 1, the timing of the onset of selection ($T_{\text{adaptive}}$) to vary between 1 and 1686 generations ago (i.e. the entire simulated timespan), and the selection coefficient ($s$) to vary between -0.01 and 0.2—all drawn from uniform distributions.

Assessment of archaic introgression

To identify candidate archaic introgressed SVs among the set of highly differentiated variants, we tested each SV for LD with putative introgressed SNPs as identified based on published results from the method SPrime, which is based on signatures of LD and divergence from an African outgroup population (Browning et al., 2018). Specifically, we computed pairwise LD between the SV and all SNPs in a 100 Kb window, matching the ancestry component to a corresponding population from the 1000 Genomes Project. We additionally computed the allele frequencies of SVs in African populations, excluding the ASW (Americans of African Ancestry in SW USA) and ACB (African Caribbeans in Barbados) populations which exhibit substantial non-African admixture. We reported all highly differentiated SVs with $r^2 > 0.5$ with any putative introgressed SNP and AF < 0.01 in non-admixed African populations (Table S2).

To efficiently search for the $\text{IGHG4}$ insertion in additional datasets, we designed a 48 bp sequence (TGGAGAGAGTGGGGGACAGCGTCAGGGACAGGTGGGGACAGCCTGGGG) that spans the insertion breakpoint, extending across the entire 33 bp of the insertion itself as well as 11 bp and 4 bp, into the respective upstream and downstream flanking regions. BLAST searches for this sequence in the hg19 and GRCh38 human reference genomes returned no exact matches, but the k-mer was sufficiently similar to the reference sequence that reads containing it still mapped to the same locus. Sequence alignments from four high-coverage archaic hominin samples (Altai Neanderthal, Vindija Neanderthal, Chagyrskaya Neanderthal, and Denisovan) were obtained from http://cdna.eva.mpg.de/neandertal and http://cdna.eva.mpg.de/denisova/. Forward strand sequences of unique reads were extracted from reads aligned to the hg19 reference genome, and exact matches to the 48 bp query sequence were identified using grep.

Evidence of introgression at the IGH locus was further examined by counting observed alleles at highly differentiated SNPs from sequenced alignments for each high-coverage archaic sample (see above). Sites with two or more reads supporting the alternative allele were used to define matching and color Figure 6B. Figure 6A further conditions on alternative allele frequency ≤1% within African populations of the 1000 Genomes Project.
**Phenotype-wide association analysis**

We examined potential phenotype associations with the putative Neanderthal-introgressed haplotype at the IGH locus by extracting summary statistics from the pan-ancestry analysis of the UK Biobank (https://pan.ukbb.broadinstitute.org/). Specifically, we obtained association p-values for two SNPs, which each tag one of the two major LD blocks (rs115091999 and rs150526114). We restricted analysis to individuals of East Asian ancestry. No variants were significant after Bonferroni correction (Fig. S11).

**Data and software availability**

All code necessary for reproducing our analysis is available on GitHub (https://github.com/mccoy-lab/sv_selection). SV genotypes, eQTL results, and selection scan results are available on Zenodo (doi: 10.5281/zenodo.4469976).

**Acknowledgments**

We thank Tim O’Connor, Sai Chen, and members of the McCoy lab for feedback and helpful discussions. We also thank the staff at the Maryland Advanced Research Computing Center for computing support. This work is supported by the National Institutes of Health (NIH) grant R35GM133747 to R.C.M and the US National Science Foundation grant DBI-1350041 to M.C.S.

**References**


Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**:559–575. doi:10.1086/519795


Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. 2012. The Date of Interbreeding between Neandertals and Modern Humans. *PLOS Genet* **8**:e1002947. doi:10.1371/journal.pgen.1002947


