bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

Stephanie M. Yan[1], Rachel M. Sherman[2], Dylan J. Taylor[1], Divya R. Nair[1], Andrew N. Bortvin[1], Michael C. Schatz[1,2], Rajiv C. McCoy[1]*

[1]Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA
[2]Department 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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).

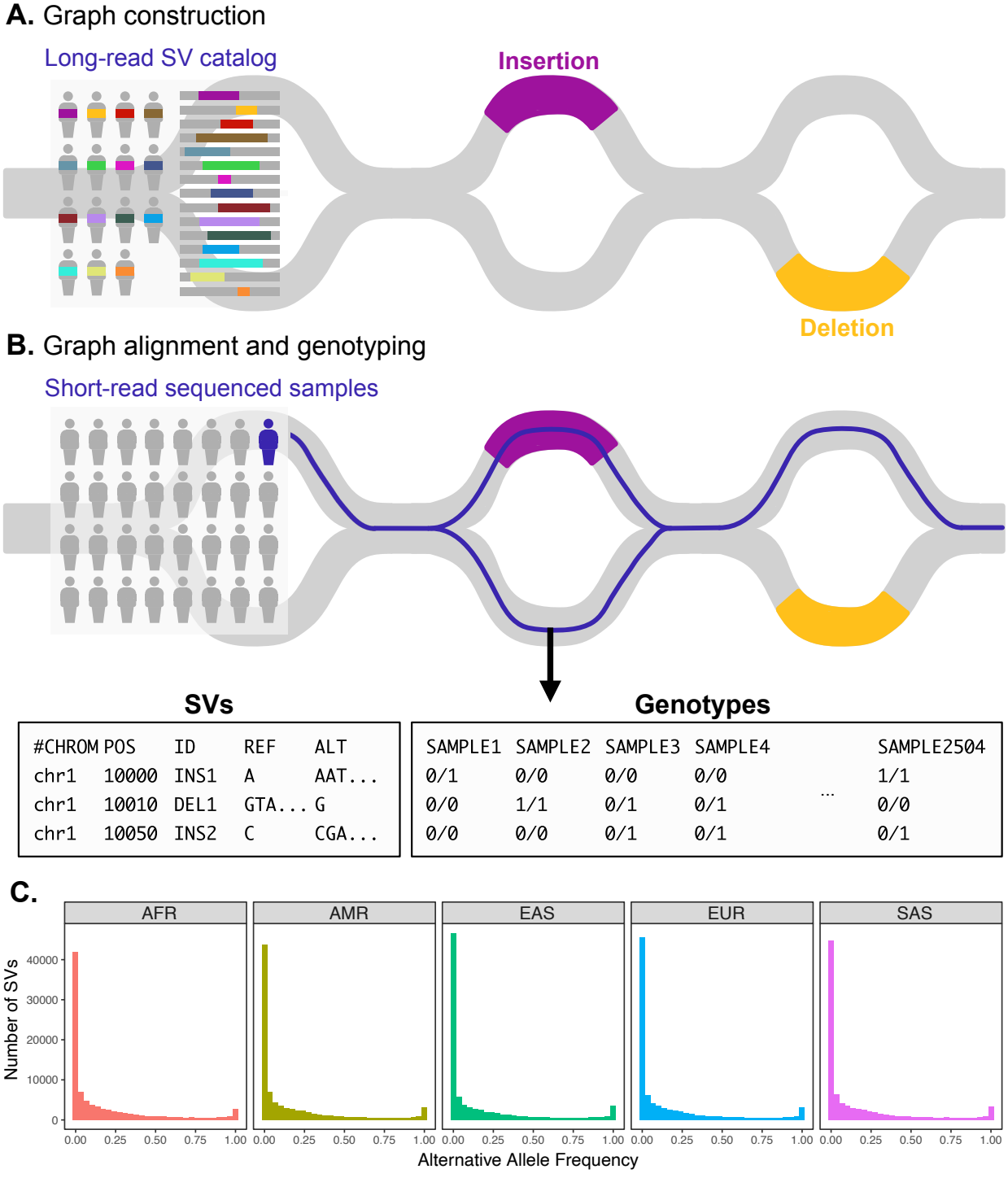
As quality control, we filtered the resulting data based on genotyping rates and adherence to Hardy Weinberg equilibrium, in accordance with (Chen et al., 2019; see **Methods**). Specifically, we removed SVs that were not successfully genotyped in ≥50% of samples, as well as SVs that violated one-sided 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(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).



bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 disequlibrium 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

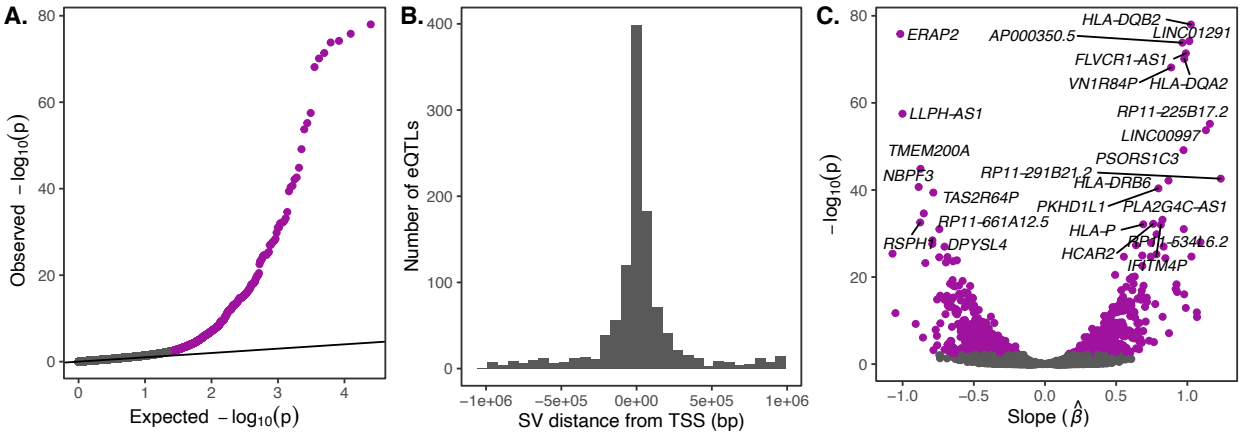


Figure 2. eQTL mapping of SVs. We used RNA-seq data from the Geuvadis Consortium (The Geuvadis Consortium et al., 2013), obtained from LCLs derived from individuals from four European and one African population of the 1000 Genomes dataset, to test for associations between SV genotypes and gene expression. 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 (β).

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

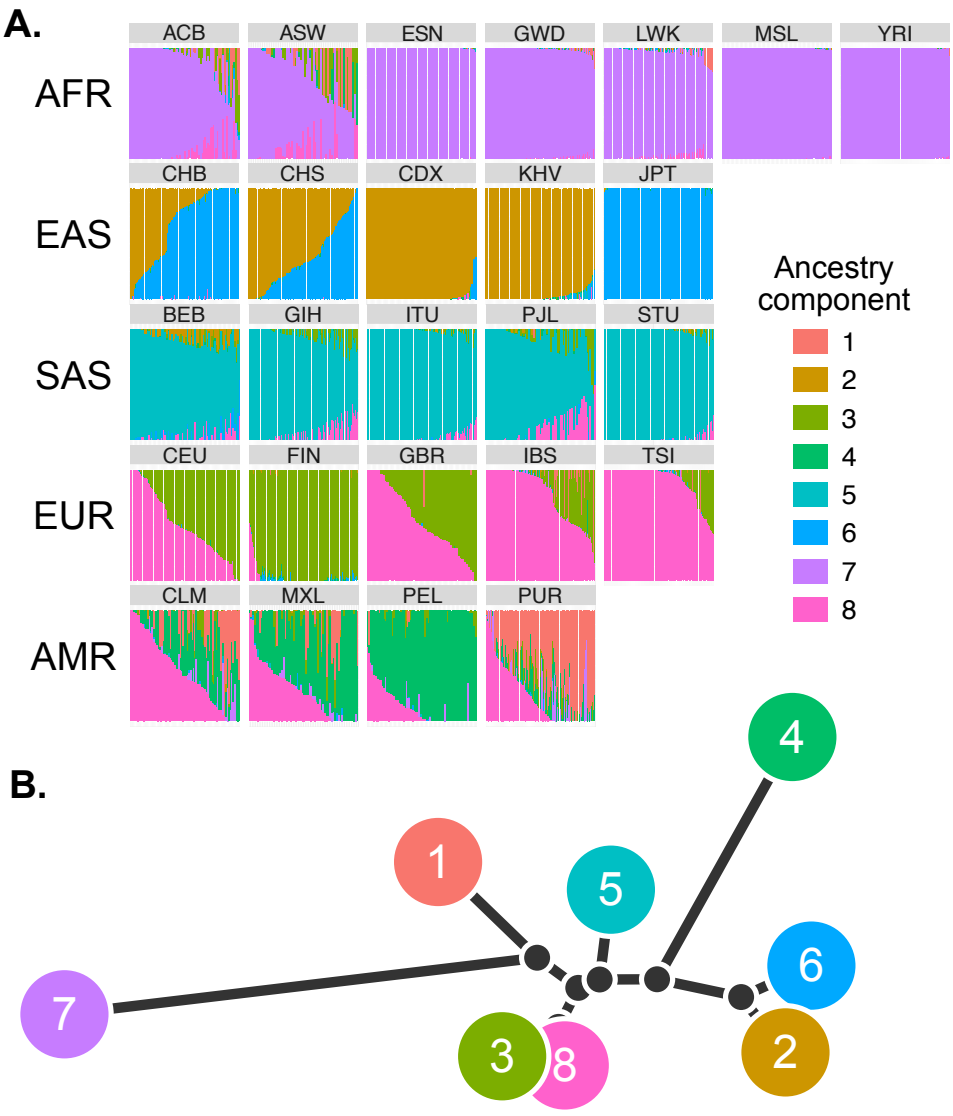


Figure 3. Global patterns of admixture and population relationships. To conduct an admixture-aware scan for local adaptation, we used Ohana (Cheng et al., 2017; Ilardo et al., 2018; Cheng et al., 2019) to infer genome-wide patterns of ancestry in the 1000 Genomes samples. This method models each individual 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

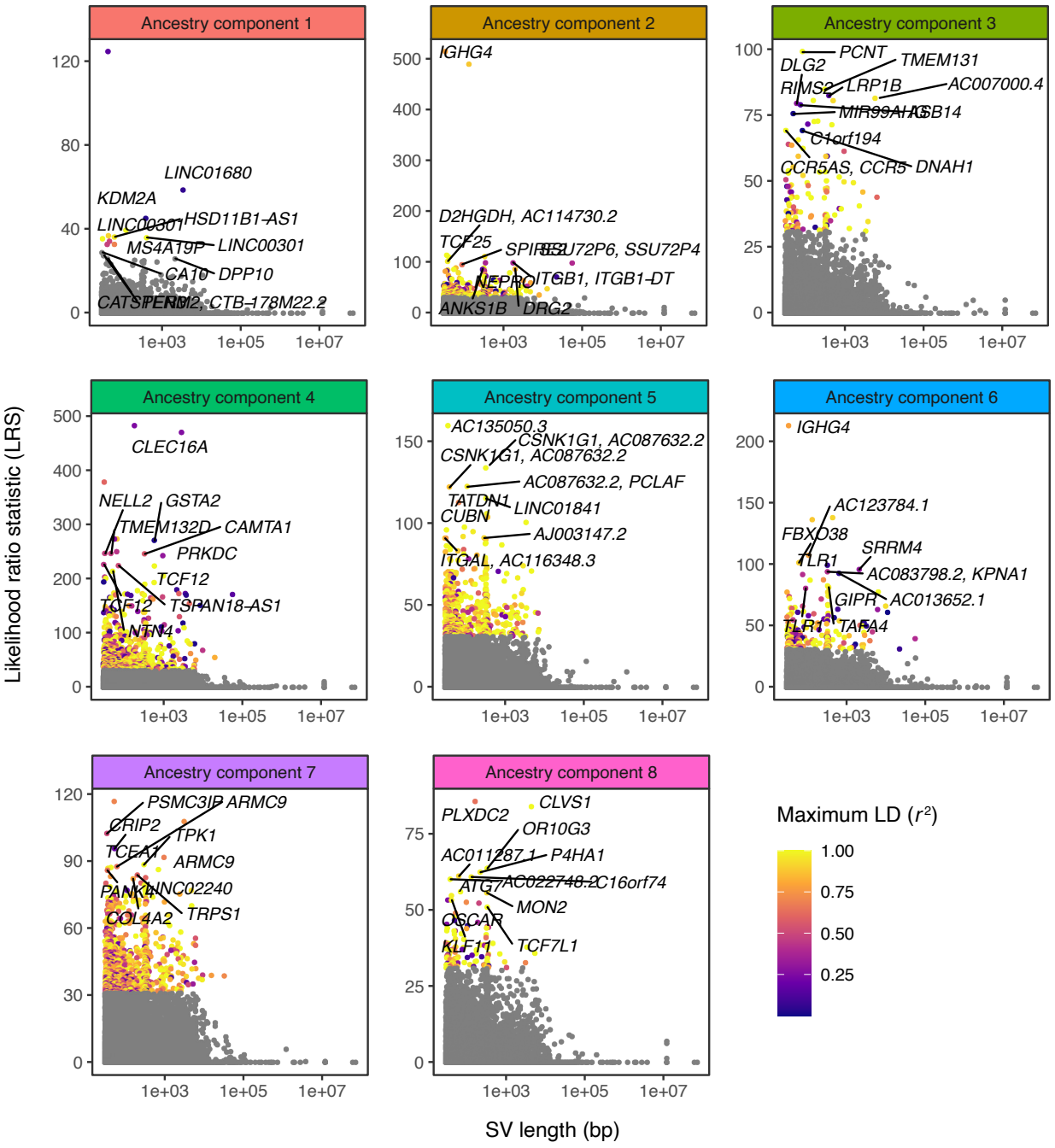


Figure 4. Quantifying allele frequency differentiation among ancestry components. Using Ohana, we searched for evidence of local adaptation by testing whether the allele frequencies of individual variants were better explained by the genome-wide tree (**Fig. 3B**), or by an alternative tree (e.g., **Fig. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 I 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).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 88% and 65%, 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

example, pairwise divergence times among East Asian populations inferred by (Wang et al., 2018) constrain the plausible timing of selection at the *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).

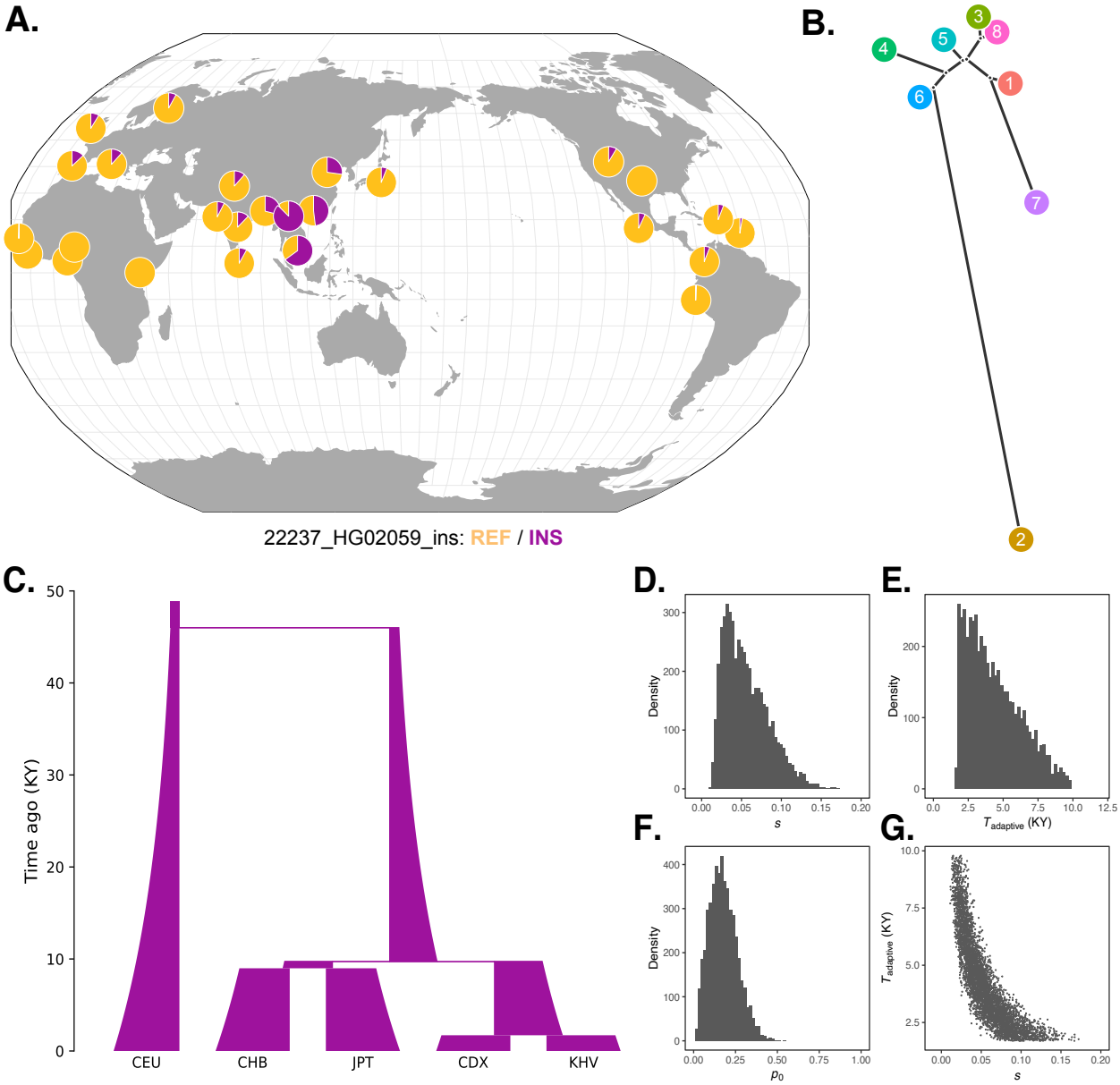


Figure 5. Local adaptation at the *IGH* locus. A. Population-specific frequencies of the insertion allele in each of the 1000 Genomes populations, in the style of the Geography of Genetic Variants browser (Marcus and Novembre, 2017). **B.** Tree representation of the best-fit selection hypothesis for the *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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 ($T_{adaptive}$). **F.** Posterior distribution of the initial allele frequency at the beginning of the simulation (p_0). **G.** Negative relationship between *s* and $T_{adaptive}$ 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 ($T_{adaptive}$) 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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).

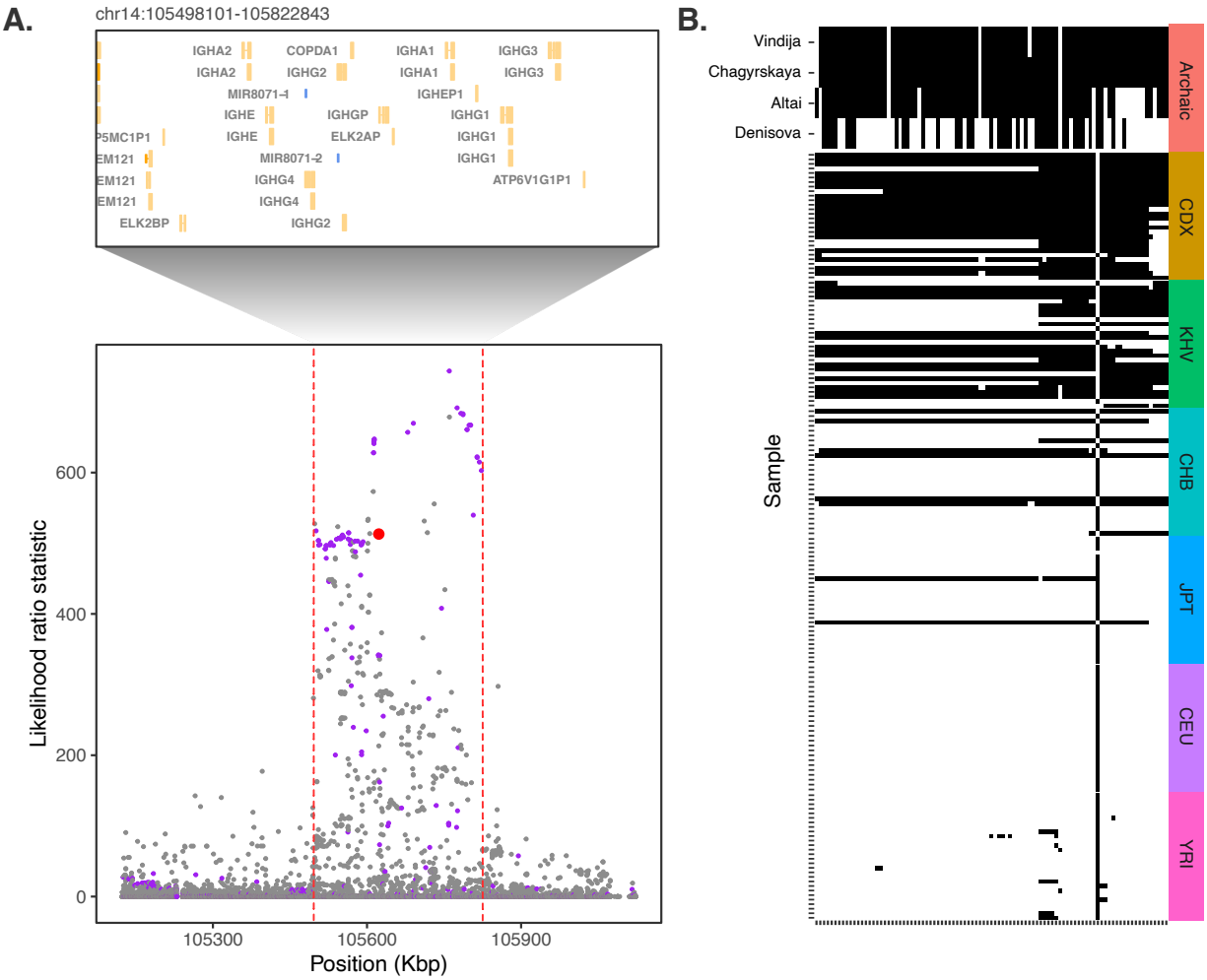


Figure 6. Evidence for Neanderthal introgression of the adaptive *IGH* haplotype. A. Local analysis of likelihood ratio statistics (LRS) in the region near the 33 bp insertion (red point) reveals a 325 Kb haplotype encompassing 94 SNPs with strong allele frequency differentiation within ancestry component 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.,

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2012a, 2012b; Sams et al., 2016; Dannemann et al., 2016; Enard and Petrov, 2018; Gouy and Excoffier, 2020).

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.,

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2018b) using default PacBio parameters to the main chromosomes of GRCh38. Variants were called with Sniffles (Sedlazeck 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 cyvcf2 (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, $p < 1 \times 10^{-4}$) in ≥13 populations. Unfolded, within-population allele frequencies were calculated with PLINK v1.90b6.4 (Purcell et al., 2007).

Quantifying linkage disequilibrium with known 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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, downsampled 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(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 (p_{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

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 x 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 ($T_{adaptive}$) 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 *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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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, John Kim, 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

Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA, Kimani J, Carrington M, Middleton D, Rajalingam R, Beksac M, Marsh SGE, Maiers M, Guethlein LA, Tavoularis S, Little A-M, Green RE, Norman PJ, Parham P. 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* **334**:89–94. doi:10.1126/science.1209202

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**:1655–1664. doi:10.1101/gr.094052.109

Almarri MA, Bergström A, Prado-Martinez J, Yang F, Fu B, Dunham AS, Chen Y, Hurles ME, Tyler-Smith C, Xue Y. 2020. Population Structure, Stratification, and Introgression of Human Structural Variation. *Cell* **182**:189-199.e15. doi:10.1016/j.cell.2020.05.024

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, Levy Y, Harel TH, Shalev-Schlosser G, Amsellem Z, Razifard H, Caicedo AL, Tieman DM, Klee H, Kirsche M, Aganezov S, Ranallo-Benavidez TR, Lemmon ZH, Kim J, Robitaille G, Kramer M, Goodwin S, McCombie WR, Hutton S, Van Eck J, Gillis J, Eshed Y, Sedlazeck FJ, van der Knaap E, Schatz MC, Lippman ZB. 2020. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* **182**:145-161.e23. doi:10.1016/j.cell.2020.05.021

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**:9354. doi:10.1038/s41598-019-45839-z

Andrés AM, Dennis MY, Kretzschmar WW, Cannons JL, Lee-Lin S-Q, Hurle B, NISC Comparative Sequencing Program, Schwartzberg PL, Williamson SH, Bustamante CD, Nielsen R, Clark AG, Green ED. 2010. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet* **6**:e1001157. doi:10.1371/journal.pgen.1001157

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, Warren WC, Magrini V, McGrath SD, Li YI, Wilson RK, Eichler EE. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**:663-675.e19. doi:10.1016/j.cell.2018.12.019

Barreiro LB, Quintana-Murci L. 2020. Evolutionary and population (epi)genetics of immunity to infection. *Hum Genet* **139**:723–732. doi:10.1007/s00439-020-02167-x

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, Blanché H, Deleuze J-F, Cann H, Mallick S, Reich D, Sandhu MS, Skoglund P, Scally A, Xue Y, Durbin R, Tyler-Smith C. 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**. doi:10.1126/science.aay5012

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *Am J Hum Genet* **74**:1111–1120. doi:10.1086/421051

Beyter D, Ingimundardottir H, Eggertsson HP, Bjornsson E, Kristmundsdottir S, Mehringer S, Jonsson H, Hardarson MT, Magnusdottir DN, Kristjansson RP, Gudjonsson SA, Sverrisson ST, Holley G, Eyjolfsson G, Olafsson I, Sigurdardottir O, Masson G, Thorsteinsdottir U, Gudbjartsson DF, Sulem P, Magnusson OT, Halldorsson BV, Stefansson K. 2019. Long read sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease. *bioRxiv* 848366. doi:10.1101/848366

Bollback JP, York TL, Nielsen R. 2008. Estimation of 2Nes From Temporal Allele Frequency Data. *Genetics* **179**:497–502. doi:10.1534/genetics.107.085019

Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. 2018. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173**:53-61.e9. doi:10.1016/j.cell.2018.02.031

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, Eichler EE. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**:608–611. doi:10.1038/nature13907

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, Fan X, Wen J, Handsaker RE, Fairley S, Kronenberg ZN, Kong X, Hormozdiari F, Lee D, Wenger AM, Hastie AR, Antaki D, Anantharaman T, Audano PA, Brand H, Cantsilieris S, Cao H, Cerveira E, Chen C, Chen X, Chin C-S, Chong Z, Chuang NT, Lambert CC, Church DM, Clarke L, Farrell A, Flores J, Galeev T, Gorkin DU, Gujral M, Guryev V, Heaton WH, Korlach J, Kumar S, Kwon JY, Lam ET, Lee JE, Lee J, Lee W-P, Lee SP, Li S, Marks P, Viaud-Martinez K, Meiers S, Munson KM, Navarro FCP, Nelson BJ, Nodzak C, Noor A, Kyriazopoulou-Panagiotopoulou S, Pang AWC, Qiu Y, Rosanio G, Ryan M, Stütz A, Spierings DCJ, Ward A, Welch AE, Xiao M, Xu W, Zhang C, Zhu Q, Zheng-Bradley X, Lowy E, Yakneen S, McCarroll S, Jun G, Ding L, Koh CL, Ren B, Flicek P, Chen K, Gerstein MB, Kwok P-Y, Lansdorp PM, Marth GT, Sebat J, Shi X, Bashir A, Ye K, Devine SE, Talkowski ME, Mills RE, Marschall T, Korbel JO, Eichler EE, Lee C. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**:1784. doi:10.1038/s41467-018-08148-z

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, Kirsche M, Bentley DR, Schatz MC, Sedlazeck FJ, Eberle MA. 2019. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol* **20**:291. doi:10.1186/s13059-019-1909-7

Cheng JY, Mailund T, Nielsen R. 2017. Fast admixture analysis and population tree estimation for SNP and NGS data. *Bioinformatics* **33**:2148–2155. doi:10.1093/bioinformatics/btx098

Cheng JY, Racimo F, Nielsen R. 2019. Ohana: detecting selection in multiple populations by modelling ancestral admixture components (preprint). Bioinformatics. doi:10.1101/546408

Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. 2017. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol* **35**:319–321. doi:10.1038/nbt.3838

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38**:1251–1260. doi:10.1038/ng1911

Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL, Jewett E, Ranciaro A, Thompson S, Lo Y, Pfeifer SP, Jensen JD, Campbell MC, Beggs W, Hormozdiari F, Mpoloka SW, Mokone GG, Nyambo T, Meskel DW, Belay G, Haut J, Program NCS, Rothschild H, Zon L, Zhou Y, Kovacs MA, Xu M, Zhang T, Bishop K, Sinclair J, Rivas C, Elliot E, Choi J, Li SA, Hicks B, Burgess S, Abnet C, Watkins-Chow DE, Oceana E, Song YS, Eskin E, Brown KM, Marks MS, Loftus SK, Pavan WJ, Yeager M, Chanock S, Tishkoff SA. 2017. Loci associated with skin pigmentation identified in African populations. *Science* **358**. doi:10.1126/science.aan8433

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2020. Twelve years of SAMtools and BCFtools. *ArXiv201210295 Q-Bio*.

Dannemann M, Andrés AM, Kelso J. 2016. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *Am J Hum Genet* **98**:22–33. doi:10.1016/j.ajhg.2015.11.015

Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, Goedert JJ, Buchbinder SP, Vittinghoff E, Gomperts E, Donfield S, Vlahov D, Kaslow R, Saah A, Rinaldo C, Detels R, O'Brien SJ. 1996. Genetic Restriction of HIV-1 Infection and Progression to AIDS by a Deletion Allele of the CKR5 Structural Gene. *Science* **273**:1856–1862. doi:10.1126/science.273.5283.1856

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, Yilmaz F, Zhao X, Hsieh P, Lee J, Kumar S, Lin J, Rausch T, Chen Y, Ren J, Santamarina M, Hoeps W, Ashraf H, Chuang NT, Yang X, Munson KM, Lewis AP, Fairley S, Tallon LJ, Clarke WE, Basile AO, Byrska-Bishop M, Corvelo A, Chaisson MJP, Chen J, Li C, Brand H, Wenger AM, Ghareghani M, Harvey W, Raeder B, Hasenfeld P, Regier A, Abel H, Hall I, Flicek P, Stegle O, Gerstein MB, Tubio JMC, Mu Z, Li YI, Shi X, Hastie AR, Ye K, Chong Z, Sanders AD, Zody MC, Talkowski ME, Mills RE, Devine SE, Lee C, Korbel JO, Marschall T, Eichler EE. 2020. De novo assembly of 64 haplotype-resolved human genomes of diverse ancestry and integrated analysis of structural variation. *bioRxiv* 2020.12.16.423102. doi:10.1101/2020.12.16.423102

Elguero E, Délicat-Loembet LM, Rougeron V, Arnathau C, Roche B, Becquart P, Gonzalez J-P, Nkoghe D, Sica L, Leroy EM, Durand P, Ayala FJ, Ollomo B, Renaud F, Prugnolle F. 2015. Malaria continues to select for sickle cell trait in Central Africa. *Proc Natl Acad Sci* **112**:7051–7054. doi:10.1073/pnas.1505665112

Enard D, Petrov DA. 2018. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell* **175**:360-371.e13. doi:10.1016/j.cell.2018.08.034

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**:1567–1587.

Fan S, Hansen MEB, Lo Y, Tishkoff SA. 2016. Going global by adapting local: A review of recent human adaptation. *Science* **354**:54–59. doi:10.1126/science.aaf5098

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao Hongbin, Zhao Hui, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Yan, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Yayun, Sun W, Wang Haifeng, Wang Yi, Wang Ying, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT-F, Galver LM, Fan J-B, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J-F, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P-Y, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L-C, Mak W, Qiang Song Y, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Vernon Smith A, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Steve Qin Z, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Johnson T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao Hui, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Wright Clayton E, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang Hongguang, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Ota Wang V, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, The International HapMap Consortium, Genotyping centres: Perlegen Sciences, Baylor College of Medicine and ParAllele BioScience, Beijing Genomics Institute, Broad Institute of Harvard and Massachusetts Institute of Technology, Chinese National Human Genome Center at Beijing, Chinese National Human Genome Center at Shanghai, Chinese University of Hong Kong, Hong Kong University of Science and Technology, Illumina, McGill University and Génome Québec Innovation Centre, University of California at San Francisco and Washington University, University of Hong Kong, University of Tokyo and RIKEN, Wellcome Trust Sanger Institute, Analysis groups: Broad Institute, Cold Spring Harbor Laboratory, Johns Hopkins University School of Medicine, University of Michigan, University of Oxford, University of Oxford WTC for HG, RIKEN, US National Institutes of Health, US National Institutes of Health National Center for Biotechnology Information, Community engagement/public consultation and sample collection groups: Beijing Normal University and Beijing Genomics Institute, Health Sciences University of Hokkaido EEI and Shinshu University, Howard University and University of Ibadan, University of Utah, Ethical legal

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

and social issues: CA of SS, Genetic Interest Group, Kyoto University, Nagasaki University, University of Ibadan School of Medicine, University of Montréal, University of Oklahoma, Vanderbilt University, Wellcome Trust, SNP discovery: Baylor College of Medicine, Washington University, Scientific management: Chinese Academy of Sciences, Genome Canada, Génome Québec, Japanese Ministry of Education C Sports, Science and Technology, Ministry of Science and Technology of the People's Republic of China, The Human Genetic Resource Administration of China, The SNP Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**:851–861. doi:10.1038/nature06258

Galvani AP, Slatkin M. 2003. Evaluating plague and smallpox as historical selective pressures for the CCR5-Δ32 HIV-resistance allele. *Proc Natl Acad Sci* **100**:15276–15279. doi:10.1073/pnas.2435085100

Gittelman RM, Schraiber JG, Vernot B, Mikacenic C, Wurfel MM, Akey JM. 2016. Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. *Curr Biol* **26**:3375–3382. doi:10.1016/j.cub.2016.10.041

Gouy A, Excoffier L. 2020. Polygenic Patterns of Adaptive Introgression in Modern Humans Are Mainly Shaped by Response to Pathogens. *Mol Biol Evol* **37**:1420–1433. doi:10.1093/molbev/msz306

Graffelman J. 2015. Exploring Diallelic Genetic Markers: The **HardyWeinberg** Package. *J Stat Softw* **64**. doi:10.18637/jss.v064.i03

Graffelman J, Jain D, Weir B. 2017. A genome-wide study of Hardy–Weinberg equilibrium with next generation sequence data. *Hum Genet* **136**:727–741. doi:10.1007/s00439-017-1786-7

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, Hansen NF, Durand EY, Malaspinas A-S, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Ž, Gušic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, Rasilla M de la, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S. 2010. A Draft Sequence of the Neandertal Genome. *Science* **328**:710–722. doi:10.1126/science.1188021

GTEx Consortium. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**:1318–1330. doi:10.1126/science.aaz1776

GTEx Consortium, Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Montgomery SB, Battle A, Conrad DF, Hall IM. 2017. The impact of structural variation on human gene expression. *Nat Genet* **49**:692–699. doi:10.1038/ng.3834

Haller BC, Messer PW. 2019. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Mol Biol Evol* **36**:632–637. doi:10.1093/molbev/msy228

Hamid I, Korunes KL, Beleza S, Goldberg A. 2021. Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde. *eLife*. doi:10.7554/eLife.63177

Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B. 2020. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol* **21**:35. doi:10.1186/s13059-020-1941-7

Hsieh P, Vollger MR, Dang V, Porubsky D, Baker C, Cantsilieris S, Hoekzema K, Lewis AP, Munson KM, Sorensen M, Kronenberg ZN, Murali S, Nelson BJ, Chiatante G, Maggiolini FAM, Blanché H, Underwood JG, Antonacci F, Deleuze J-F, Eichler EE. 2019. Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* **366**:eaax2083. doi:10.1126/science.aax2083

Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, Ni P, Wang B, Ou X, Huasang, Luosang J, Cuo ZXP, Li K, Gao G, Yin Y,

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Wang W, Zhang X, Xu X, Yang H, Li Y, Wang Jian, Wang Jun, Nielsen R. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**:194–197. doi:10.1038/nature13408

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**:167–170. doi:10.1038/335167a0

Ilardo MA, Moltke I, Korneliussen TS, Cheng J, Stern AJ, Racimo F, de Barros Damgaard P, Sikora M, Seguin-Orlando A, Rasmussen S, van den Munckhof ICL, ter Horst R, Joosten LAB, Netea MG, Salingkat S, Nielsen R, Willerslev E. 2018. Physiological and Genetic Adaptations to Diving in Sea Nomads. *Cell* **173**:569-580.e15. doi:10.1016/j.cell.2018.03.054

Jablonski NG. 2004. The Evolution of Human Skin and Skin Color. *Annu Rev Anthropol* **33**:585–623. doi:10.1146/annurev.anthro.33.070203.143955

Jabot F, Faure T, Dumoulin N. 2013. EasyABC: performing efficient approximate Bayesian computation sampling schemes using R. *Methods Ecol Evol* **4**:684–687. doi:https://doi.org/10.1111/2041-210X.12050

Jakubosky D, Smith EN, D'Antonio M, Jan Bonder M, Young Greenwald WW, D'Antonio-Chronowska A, Matsui H, Stegle O, Montgomery SB, DeBoever C, D'Antonio M, Frazer KA. 2020. Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. *Nat Commun* **11**:2928. doi:10.1038/s41467-020-16481-5

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* **8**:14061. doi:10.1038/ncomms14061

Jouganous J, Long W, Ragsdale AP, Gravel S. 2017. Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation. *Genetics* **206**:1549–1567. doi:10.1534/genetics.117.200493

Keinan A, Clark AG. 2012. Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. *Science* **336**:740–743. doi:10.1126/science.1217283

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* **20**:117. doi:10.1186/s13059-019-1720-5

Kothapalli KSD, Ye, Gadgil MS, Carlson SE, O'Brien KO, Zhang JY, Park HG, Ojukwu K, Zou J, Hyon SS, Joshi KS, Gu Z, Keinan A, Brenna JT. 2016. Positive Selection on a Regulatory Insertion–Deletion Polymorphism in FADS2 Influences Apparent Endogenous Synthesis of Arachidonic Acid. *Mol Biol Evol* **33**:1726–1739. doi:10.1093/molbev/msw049

Lachance J, Tishkoff SA. 2013. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it: Prospects & Overviews. *BioEssays* **35**:780–786. doi:10.1002/bies.201300014

Lenormand M, Jabot F, Deffuant G. 2013. Adaptive approximate Bayesian computation for complex models. *Comput Stat* **28**:2777–2796. doi:10.1007/s00180-013-0428-3

Mafessoni F, Grote S, Filippo C de, Slon V, Kolobova KA, Viola B, Markin SV, Chintalapati M, Peyrégne S, Skov L, Skoglund P, Krivoshapkin AI, Derevianko AP, Meyer M, Kelso J, Peter B, Prüfer K, Pääbo S. 2020. A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc Natl Acad Sci* **117**:15132–15136. doi:10.1073/pnas.2004944117

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, Skoglund P, Lazaridis I, Sankararaman S, Fu Q, Rohland N, Renaud G, Erlich Y, Willems T, Gallo C, Spence JP, Song YS, Poletti G, Balloux F, van Driem G, de Knijff P, Romero IG, Jha AR, Behar DM, Bravi CM, Capelli C, Hervig T, Moreno-Estrada A, Posukh OL, Balanovska E, Balanovsky O, Karachanak-Yankova S, Sahakyan H,

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Toncheva D, Yepiskoposyan L, Tyler-Smith C, Xue Y, Abdullah MS, Ruiz-Linares A, Beall CM, Di Rienzo A, Jeong C, Starikovskaya EB, Metspalu E, Parik J, Villems R, Henn BM, Hodoglugil U, Mahley R, Sajantila A, Stamatoyannopoulos G, Wee JTS, Khusainova R, Khusnutdinova E, Litvinov S, Ayodo G, Comas D, Hammer MF, Kivisild T, Klitz W, Winkler CA, Labuda D, Bamshad M, Jorde LB, Tishkoff SA, Watkins WS, Metspalu M, Dryomov S, Sukernik R, Singh L, Thangaraj K, Pääbo S, Kelso J, Patterson N, Reich D. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**:201–206. doi:10.1038/nature18964

Marcus JH, Novembre J. 2017. Visualizing the geography of genetic variants. *Bioinformatics* **33**:594–595. doi:10.1093/bioinformatics/btw643

Mendez FL, Watkins JC, Hammer MF. 2012a. Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations. *Mol Biol Evol* **29**:1513–1520. doi:10.1093/molbev/msr301

Mendez FL, Watkins JC, Hammer MF. 2012b. A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am J Hum Genet* **91**:265–274. doi:10.1016/j.ajhg.2012.06.015

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, Filippo C de, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**:222–226. doi:10.1126/science.1224344

Mikocziova I, Gidoni M, Lindeman I, Peres A, Snir O, Yaari G, Sollid LM. 2020. Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. *Nucleic Acids Res* **48**:5499–5510. doi:10.1093/nar/gkaa310

Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspinas A-S, Sikora M, Reuther JD, Irish JD, Malhi RS, Orlando L, Song YS, Nielsen R, Meltzer DJ, Willerslev E. 2018. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* **553**:203–207. doi:10.1038/nature25173

Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. 2016. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**:1479–1485. doi:10.1093/bioinformatics/btv722

Pandey R, Bakay M, Hain HS, Strenkowski B, Yermakova A, Kushner JA, Orange JS, Hakonarson H. 2019. The Autoimmune Disorder Susceptibility Gene CLEC16A Restrains NK Cell Function in YTS NK Cell Line and Clec16a Knockout Mice. *Front Immunol* **10**:68. doi:10.3389/fimmu.2019.00068

Parham P, Ohta T. 1996. Population Biology of Antigen Presentation by MHC Class I Molecules. *Science* **272**:67–74. doi:10.1126/science.272.5258.67

Pedersen BS, Quinlan AR. 2017. cyvcf2: fast, flexible variant analysis with Python. *Bioinformatics* **33**:1867–1869. doi:10.1093/bioinformatics/btx057

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**:1256–1260. doi:10.1038/ng2123

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* **16**:1791–1798. doi:10.1093/oxfordjournals.molbev.a026091

Prüfer K, Filippo C de, Grote S, Mafessoni F, Korlević P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, Reher D, Hopfe C, Nagel S, Maricic T, Fu Q, Theunert C, Rogers R, Skoglund P, Chintalapati M, Dannemann M, Nelson BJ, Key FM, Rudan P, Kućan Ž, Gušić I, Golovanova LV, Doronichev VB, Patterson N, Reich D, Eichler EE, Slatkin M,

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Schierup MH, Andrés AM, Kelso J, Meyer M, Pääbo S. 2017. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**:655–658. doi:10.1126/science.aao1887

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PLF, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**:43–49. doi:10.1038/nature12886

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

Racimo F, Gokhman D, Fumagalli M, Ko A, Hansen T, Moltke I, Albrechtsen A, Carmel L, Huerta-Sánchez E, Nielsen R. 2017. Archaic Adaptive Introgression in TBX15/WARS2. *Mol Biol Evol* **34**:509–524. doi:10.1093/molbev/msw283

Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren K-G, Pedersen AG, Schubert M, Van Dam A, Kapel CMO, Nielsen HB, Brunak S, Avetisyan P, Epimakhov A, Khalyapin MV, Gnuni A, Kriiska A, Lasak I, Metspalu M, Moiseyev V, Gromov A, Pokutta D, Saag L, Varul L, Yepiskoposyan L, Sicheritz-Pontén T, Foley RA, Lahr MM, Nielsen R, Kristiansen K, Willerslev E. 2015. Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago. *Cell* **163**:571–582. doi:10.1016/j.cell.2015.10.009

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139–140. doi:10.1093/bioinformatics/btp616

Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, Deikus G, Auckland K, Eichler EE, Marasco WA, Sebra R, Sharp AJ, Smith ML, Bashir A, Watson CT. 2020. A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. *Front Immunol* **11**:16.

Rotival M, Quintana-Murci L. 2020. Functional consequences of archaic introgression and their impact on fitness 4.

Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B, Hutcheson HB, Cullen M, Mikkelsen TS, Roy J, Patterson N, Cooper R, Reich D, Altshuler D, O'Brien S, Lander ES. 2005. The Case for Selection at CCR5-Δ32. *PLOS Biol* **3**:e378. doi:10.1371/journal.pbio.0030378

Saitou M, Gokcumen O. 2019. Resolving the Insertion Sites of Polymorphic Duplications Reveals a HERC2 Haplotype under Selection. *Genome Biol Evol* **11**:1679–1690. doi:10.1093/gbe/evz107

Sams AJ, Dumaine A, Nédélec Y, Yotova V, Alfieri C, Tanner JE, Messer PW, Barreiro LB. 2016. Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol* **17**:246. doi:10.1186/s13059-016-1098-6

Sankararaman S, Mallick S, Patterson N, Reich D. 2016. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Curr Biol* **26**:1241–1247. doi:10.1016/j.cub.2016.03.037

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

Schaid DJ, Chen W, Larson NB. 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**:491–504. doi:10.1038/s41576-

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

018-0016-z

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018a. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet* **19**:329–346. doi:10.1038/s41576-018-0003-4

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018b. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**:461–468. doi:10.1038/s41592-018-0001-7

Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, Jones KW. 2004. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum Genomics* **1**:274–286. doi:10.1186/1479-7364-1-4-274

Sibbesen JA, Maretty L, Krogh A. 2018. Accurate genotyping across variant classes and lengths using variant graphs. *Nat Genet* **50**:1054–1059. doi:10.1038/s41588-018-0145-5

Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen J, Hickey G, Chang P-C, Carroll A, Haussler D, Garrison E, Paten B. 2020. Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *bioRxiv* 2020.12.04.412486. doi:10.1101/2020.12.04.412486

Souilmi Y, Lauterbur ME, Tobler R, Huber CD, Johar AS, Enard D. 2021. An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in East Asia. *bioRxiv* 2020.11.16.385401. doi:10.1101/2020.11.16.385401

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley GA, Allikmets R, Schriml L, Gerrard B, Malasky M, Ramos MD, Morlot S, Tzetis M, Oddoux C, di Giovine FS, Nasioulas G, Chandler D, Aseev M, Hanson M, Kalaydjieva L, Glavac D, Gasparini P, Kanavakis E, Claustres M, Kambouris M, Ostrer H, Duff G, Baranov V, Sibul H, Metspalu A, Goldman D, Martin N, Duffy D, Schmidtke J, Estivill X, O'Brien SJ, Dean M. 1998. Dating the Origin of the CCR5-Δ32 AIDS-Resistance Allele by the Coalescence of Haplotypes. *Am J Hum Genet* **62**:1507–1515. doi:10.1086/301867

Storey JD, Bass AJ, Dabney A, Robinson D. 2019. qvalue: Q-value estimation for false discovery rate control.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci* **100**:9440–9445. doi:10.1073/pnas.1530509100

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, Jorde LB, Posukh OL, Sahakyan H, Watkins WS, Yepiskoposyan L, Abdullah MS, Bravi CM, Capelli C, Hervig T, Wee JTS, Tyler-Smith C, Driem G van, Romero IG, Jha AR, Karachanak-Yankova S, Toncheva D, Comas D, Henn B, Kivisild T, Ruiz-Linares A, Sajantila A, Metspalu E, Parik J, Villems R, Starikovskaya EB, Ayodo G, Beall CM, Rienzo AD, Hammer MF, Khusainova R, Khusnutdinova E, Klitz W, Winkler C, Labuda D, Metspalu M, Tishkoff SA, Dryomov S, Sukernik R, Patterson N, Reich D, Eichler EE. 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**. doi:10.1126/science.aab3761

Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, Sabeti PC. 2016. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**:1519–1529. doi:10.1016/j.cell.2016.04.027

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**:68–74. doi:10.1038/nature15393

The 1000 Genomes Project Consortium, Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stütz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK,

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Jasmine Mu X, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer E-W, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**:75–81. doi:10.1038/nature15394

The Geuvadis Consortium, Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Lehrach H, Schreiber S, Sudbrak R, Carracedo Á, Antonarakis SE, Häsler R, Syvänen A-C, van Ommen G-J, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**:506–511. doi:10.1038/nature12531

Tremblay M, Vézina H. 2000. New Estimates of Intergenerational Time Intervals for the Calculation of Age and Origins of Mutations. *Am J Hum Genet* **66**:651–658. doi:10.1086/302770

van Arensbergen J, Pagie L, FitzPatrick VD, de Haas M, Baltissen MP, Comoglio F, van der Weide RH, Teunissen H, Võsa U, Franke L, de Wit E, Vermeulen M, Bussemaker HJ, van Steensel B. 2019. High-throughput identification of human SNPs affecting regulatory element activity. *Nat Genet* **51**:1160–1169. doi:10.1038/s41588-019-0455-2

Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, Dannemann M, Grote S, McCoy RC, Norton H, Scheinfeldt LB, Merriwether DA, Koki G, Friedlaender JS, Wakefield J, Pääbo S, Akey JM. 2016. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**:235–239. doi:10.1126/science.aad9416

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting Natural Selection in Genomic Data. *Annu Rev Genet* **47**:97–120. doi:10.1146/annurev-genet-111212-133526

Wang Y, Lu D, Chung Y-J, Xu S. 2018. Genetic structure, divergence and admixture of Han Chinese, Japanese and Korean populations. *Hereditas* **155**:19. doi:10.1186/s41065-018-0057-5

Watson CT, Glanville J, Marasco WA. 2017. The Individual and Population Genetics of Antibody Immunity. *Trends Immunol* **38**:459–470. doi:10.1016/j.it.2017.04.003

Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, Willsey AJ, Joy JB, Scott JK, Graves TA, Wilson RK, Holt RA, Eichler EE, Breden F. 2013. Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation. *Am J Hum Genet* **92**:530–546. doi:10.1016/j.ajhg.2013.03.004

Wright S. 1949. The Genetical Structure of Populations. *Ann Eugen* **15**:323–354. doi:https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, Zheng Hancheng, Liu T, He W, Li K, Luo R, Nie X, Wu H, Zhao M, Cao H, Zou J, Shan Y, Li Shuzheng, Yang Q, Asan, Ni P, Tian G, Xu J, Liu X, Jiang T, Wu R, Zhou G, Tang M, Qin J, Wang T, Feng S, Li G, Huasang, Luosang J, Wang W, Chen F, Wang Y, Zheng X, Li Z, Bianba Z, Yang G, Wang X, Tang S, Gao G, Chen Y,

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.26.428314; this version posted January 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Luo Z, Gusang L, Cao Z, Zhang Q, Ouyang W, Ren X, Liang H, Zheng Huisong, Huang Y, Li J, Bolund L, Kristiansen K, Li Y, Zhang Y, Zhang X, Li R, Li Songgang, Yang H, Nielsen R, Wang Jun, Wang Jian. 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science* **329**:75–78. doi:10.1126/science.1190371

Zeberg H, Pääbo S. 2020a. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* **587**:610–612. doi:10.1038/s41586-020-2818-3

Zeberg H, Pääbo S. 2020b. A genetic variant protective against severe COVID-19 is inherited from Neandertals. *bioRxiv* 2020.10.05.327197. doi:10.1101/2020.10.05.327197

Zhang Z, Bai M, Barbosa GO, Chen A, Wei Y, Luo S, Wang X, Wang B, Tsukui T, Li H, Sheppard D, Kornberg TB, Ma DK. 2020. Broadly conserved roles of TMEM131 family proteins in intracellular collagen assembly and secretory cargo trafficking. *Sci Adv* **6**:eaay7667. doi:10.1126/sciadv.aay7667

Zhou S, Butler-Laporte G, Nakanishi T, Morrison D, Afilalo J, Afilalo M, Laurent L, Pietzner M, Kerrison N, Zhao K, Brunet-Ratnasingham E, Henry D, Kimchi N, Afrasiabi Z, Rezk N, Bouab M, Petitjean L, Guzman C, Xue X, Tselios C, Vulesevic B, Adeleye O, Abdullah T, Almamlouk N, Chen Y, Chassé M, Durand M, Pollak M, Paterson C, Zeberg H, Normark J, Frithiof R, Lipcsey M, Hultström M, Greenwood CMT, Langenberg C, Thysell E, Mooser V, Forgetta V, Kaufmann DE, Richards JB. 2020. A Neanderthal OAS1 isoform Protects Against COVID-19 Susceptibility and Severity: Results from Mendelian Randomization and Case-Control Studies (preprint). Genetic and Genomic Medicine. doi:10.1101/2020.10.13.20212092

Zook JM, Hansen NF, Olson ND, Chapman LM, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, Sahraeian SME, Huang V, Rouette A, Alexander N, Mason CE, Hajirasouliha I, Ricketts C, Lee J, Tearle R, Fiddes IT, Barrio AM, Wala J, Carroll A, Ghaffari N, Rodriguez OL, Bashir A, Jackman S, Farrell JJ, Wenger AM, Alkan C, Soylev A, Schatz MC, Garg S, Church G, Marschall T, Chen K, Fan X, English AC, Rosenfeld JA, Zhou W, Mills RE, Sage JM, Davis JR, Kaiser MD, Oliver JS, Catalano AP, Chaisson MJ, Spies N, Sedlazeck FJ, Salit M, Consortium the G in a B. 2019. A robust benchmark for germline structural variant detection. *bioRxiv* 664623. doi:10.1101/664623