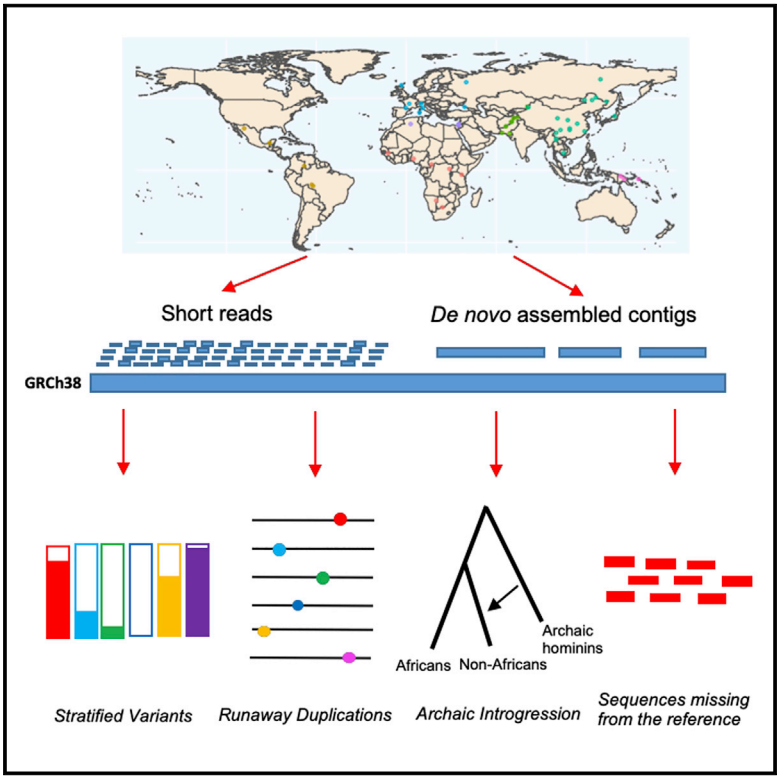
Cell

Resource

Population Structure, Stratification, and Introgression of Human Structural Variation

Graphical Abstract



Highlights

- Catalog of structural variants from a diverse set of human populations
- Identification of high-frequency population-specific variants
- Highly stratified variants putatively introgressed from Neanderthals and Denisovans
- *De novo* genome assemblies uncover and place sequences missing from the reference

Authors

Mohamed A. Almarri, Anders Bergström, Javier Prado-Martinez, ..., Matthew E. Hurles, Chris Tyler-Smith, Yali Xue

Correspondence

ma17@sanger.ac.uk (M.A.A.), ylx@sanger.ac.uk (Y.X.)

In Brief

Almarri et al. generate a structural variation atlas for a geographically diverse set of human genomes, including recovery of sequences missing from the human reference sequence.

Almarri et al., 2020, Cell *182*, 189–199
July 9, 2020
https://doi.org/10.1016/j.cell.2020.05.024

CellPress

Cell

CellPress

Resource

Population Structure, Stratification, and Introgression of Human Structural Variation

Mohamed A. Almarri,[1,4,*] Anders Bergström,[1,2] Javier Prado-Martinez,[1] Fengtang Yang,[1] Beiyuan Fu,[1] Alistair S. Dunham,[1,3] Yuan Chen,[1] Matthew E. Hurles,[1] Chris Tyler-Smith,[1] and Yali Xue[1,*]

[1]Wellcome Sanger Institute, Hinxton CB10 1SA, UK
[2]The Francis Crick Institute, London NW1 1AT, UK
[3]EMBL-EBI, Hinxton CB10 1SD, UK
[4]Lead Contact
*Correspondence: ma17@sanger.ac.uk (M.A.A.), ylx@sanger.ac.uk (Y.X.)
https://doi.org/10.1016/j.cell.2020.05.024

SUMMARY

Structural variants contribute substantially to genetic diversity and are important evolutionarily and medically, but they are still understudied. Here we present a comprehensive analysis of structural variation in the Human Genome Diversity panel, a high-coverage dataset of 911 samples from 54 diverse worldwide populations. We identify, in total, 126,018 variants, 78% of which were not identified in previous global sequencing projects. Some reach high frequency and are private to continental groups or even individual populations, including regionally restricted runaway duplications and putatively introgressed variants from archaic hominins. By *de novo* assembly of 25 genomes using linked-read sequencing, we discover 1,643 breakpoint-resolved unique insertions, in aggregate accounting for 1.9 Mb of sequence absent from the GRCh38 reference. Our results illustrate the limitation of a single human reference and the need for high-quality genomes from diverse populations to fully discover and understand human genetic variation.

INTRODUCTION

Despite the progress in sampling many populations, human genomics research is still not fully reflective of the diversity found globally (Sirugo et al., 2019). Understudied populations limit our knowledge of genetic variation and population history, and their inclusion is needed to ensure that they benefit from future developments in genomic medicine. Whole-genome sequencing projects have provided unprecedented insights into the evolutionary history of our species; however, they have mostly concentrated on substitutions at individual sites, although structural variants (affecting 50 bp or more), which include deletions, duplications, inversions, and insertions, contribute a greater diversity at the nucleotide level than any other class of variation and are important in genome evolution and disease susceptibility (Huddleston and Eichler, 2016).

Previous studies surveying global population structural variation have examined metropolitan populations at low coverage (Sudmant et al., 2015a) or a few samples from a larger number of populations (Sudmant et al., 2015b), allowing broad continental comparisons but limiting detailed analysis within each continental group and population. In this study, we present the structural variation analysis of the Human Genome Diversity Project (HGDP)-Centre d'Etude du Polymorphism CEPH panel (Figure 1A), a dataset composed of 911 samples from 54 populations of linguistic, anthropological, and evolutionary interest (Cann et al., 2002). We generate a comprehensive resource of structural variants from these diverse and understudied populations, explore the structure of different classes of structural variation, characterize regional and population-specific variants and expansions, discover putatively introgressed variants, and identify sequences missing from the GRCh38 reference.

RESULTS

Variant Discovery and Comparison with Published Datasets

We generated 911 whole-genome sequences at an average depth of 36× and mapped reads to the GRCh38 reference (Bergström et al., 2020). As the dataset was generated from lymphoblastoid cell lines, we searched for potential cell line artifacts by analyzing coverage across the genome and excluded samples containing multiple aneuploidies while masking regions that show more limited aberrations. We find many more gains of chromosomes than losses, and, in agreement with a previous cell line based study (Redon et al., 2006), we observe that most trisomies seem to affect chromosomes 9 and 12, suggesting that they contain sequences that enhance proliferation when duplicated in culture. Nevertheless, these cell line artifacts can be readily recognized and are excluded from the results below.

We identified 126,018 structural variants relative to the reference (Figure S1). These included 25,588 (~20% of the total) that are smaller than 100 bp. We compared our dataset with published structural variation catalogs (Sudmant et al., 2015a,

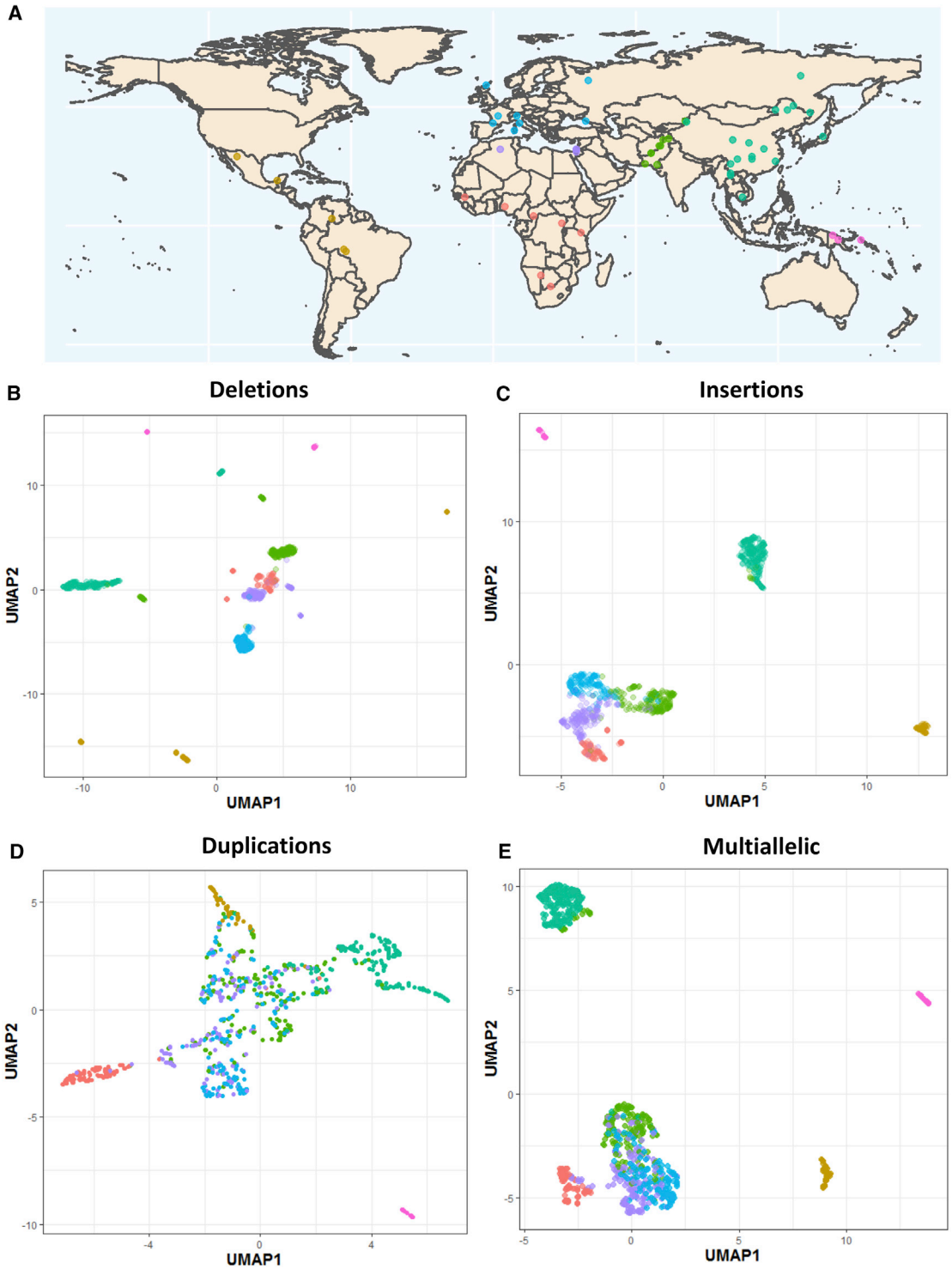


Figure 1. The HGDP Dataset and Population Structure

(A) The HGDP dataset. Each point and color represents a population and its regional label, respectively. Colors of regional groups are consistent throughout the study. See Table S1 for more details.

(B) UMAP of biallelic deletion genotypes.

(C) UMAP of insertions.

(legend continued on next page)

2015b) and find that ~78% of the variants identified in our dataset are not present in the previous studies. Despite having a smaller sample size compared with the 1000 Genomes phase 3 release (Sudmant et al., 2015a), we discover a higher total number of variants across all different classes of variants investigated. These previously unidentified calls are not limited to rare variants, as a considerable number of common and even high-frequency variants are found in regional groups and individual populations (Figure S4). Notably, our resource identifies the abundant but understudied class of small variants (50–100 bp) that were not particularly characterized by the Simons Genome Diversity Project (Sudmant et al., 2015b). At this size range, ~91% of variants in our dataset are not present in either published catalog. Collectively, this illustrates that a substantial amount of global structural variation was previously undocumented, emphasizing the importance of studying underrepresented human populations.

Population Structure

A uniform manifold approximation and projection (UMAP) of deletion genotypes shows clear separation of continental groups, and in many cases, even individual populations are distinguished (Figure 1B). Deeply divergent African populations, such as the Mbuti, Biaka, and San, form their own clusters away from the rest of the African populations. Admixed groups, such as the Hazara and Uygur, cluster separately from the Central and South Asian and East Asian groups, whereas populations that experienced high rates of genetic drift, such as the Kalash, in addition to American and Oceanian populations, are clearly differentiated. For less clearly defined populations projecting into continental clusters, we observe examples of finer structure with samples from individual populations appearing closer to themselves relative to other groups.

Insertions, duplications, multiallelic variants, and inversions also show some degree of population structure, although less defined in comparison with deletions (Figures 1C–1E and S2). Strikingly, the Oceanian populations always remain well differentiated. Consequently, we find that all classes of genetic variation show population structure, with the observed differences likely reflecting the varying mutational patterns generating each class of structural variant in addition to the overall number of discovered variants in each class.

Population Stratification and Selection

Selective pressures can result in highly stratified variants between populations. We assessed the relationship between average population differentiation and the maximal variant allele frequency difference for each population pair (Figures 2A–2C). Outliers in this relationship (i.e., variants that show a higher allele frequency difference than expected) have been proposed to be under selection (Coop et al., 2009; Huerta-Sánchez et al., 2014). Deletions and insertions show similar distributions, whereas biallelic duplications display lower stratification. We do see some notable outliers; for example, the Lowland/Sepik Papuans are almost fixed (86%) for a deletion in *HBA2*, which is absent in Papuan highlanders ($p < 0.001$, population stratification test using 1,000 permutations; STAR Methods). High frequencies of α-globin deletions have been suggested to be protective against malaria, which is not found in the highlands of Papua New Guinea but is present in the lowlands (Yenchitsomanus et al., 1985; Flint et al., 1986). On the other hand, Papuan highlanders have a small insertion (123 bp) near an exon of *VGLL4* at 93% frequency, which is markedly less common in Papuan lowlanders (7%, $p = 0.001$). We also find a deletion within *MYO5B* that is particularly common (88%) in the Lahu from China (Lahu-Hezhen, $p = 0.001$), a population shown to have high numbers of private single-nucleotide variants in addition to carrying rare Y chromosome lineages (Bergström et al., 2020).

The large number of samples per population allowed us to investigate population-private variants (Figure S3). We searched for functional effects of such variants and found a 14-kb deletion in the South American Karitiana population at 40% frequency. This variant removes the 5′ upstream region of *MGAM* up to the first exon, potentially inactivating the gene that encodes maltase-glucoamylase, an enzyme highly expressed in the small intestine and involved in digestion of dietary starches (Nichols et al., 2003). Interestingly, a recent ancient DNA study of South Americans has suggested that selection acted on this gene in ancient Andean individuals, possibly as a result of their transition to agriculture (Lindo et al., 2018). This gene has also been proposed to be under selection in dogs due to adaptation to a starch-rich diet during domestication (Axelsson et al., 2013). The Karitiana are known to have suffered a recent population crash (Bergström et al., 2020), and this has left a genetic consequence: they have one of the highest levels of runs of homozygosity of any human population studied to date (Ceballos et al., 2018). We detect suggestive but not strong evidence for a departure from neutrality at this locus (STAR Methods). The high frequency and presence of individuals homozygous for this deletion suggests that purifying selection on the ability to digest starch has been relaxed in the history of the Karitiana, possibly due to increased drift caused by the population decrease.

We discovered a deletion that is private and at 54% frequency in the Central African Mbuti hunter-gatherer population and deletes *SIGLEC5* without removing its adjacent paired receptor, *SIGLEC14* (Figure 2F). Siglecs, a family of cell-surface receptors that are expressed on immune cells, detect sialylated surface proteins expressed on host cells. Most Siglecs act as inhibitors of leukocyte activation, but *SIGLEC14* is an activating member thought to have evolved by gene conversion from *SIGLEC5* (Angata et al., 2006). This evolution has been proposed to result in a selective advantage of combating pathogens that mimic host cells by expressing sialic acids, providing an additional activation pathway (Akkaya and Barclay 2013). The deletion we identify in the Mbuti, however, seems to remove the function of the inhibitory receptor while keeping the activating receptor intact. This finding is surprising, as paired receptors are thought to have

(D) UMAP of biallelic duplications.
(E) UMAP of multiallelic variants.
See Figure S2 for more details.

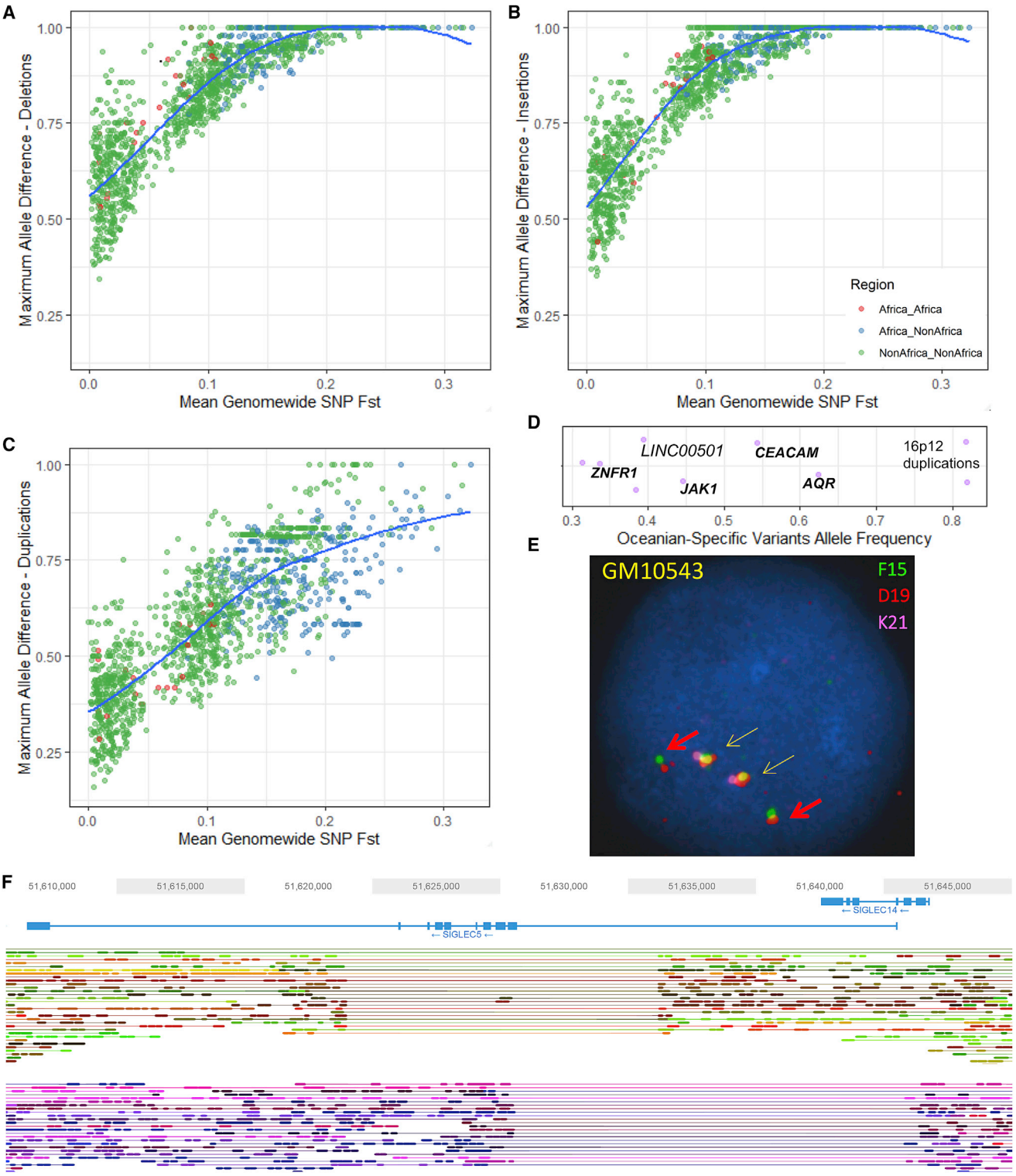


Figure 2. Population Stratification of Structural Variants

(A) Maximum allele frequency difference of deletions as a function of population differentiation for 1,431 pairwise population comparisons. The blue curve represents locally estimated scatterplot smoothing (LOESS) fits.

(B and C) The same as (A) but for insertion (B) and biallelic duplications (C).

(legend continued on next page)

Table 1. Allele Frequencies of Regionally Stratified Variants Shared with High-Coverage Archaic Genomes but Not Found in African Populations

Position	Size (bp)	Variant	EUR	CSA	EA	ME	AMR	OCE	Gene	PBS Rank (%)	Neanderthal	Denisova
chr1: 64992619–64992994	375	DEL	0	0	0	0	0	0.44	*JAK1*[a]	*98.4*	REF	DEL
chr2:3684113–3690212	6,099	DEL	0.02	0.003	0.05	0.03	0	0.26	*ALLC*[a]	*90.3*	DEL $_{Vindija}$	REF
chr3:177287011–177292441	5,430	DEL	0	0	0	0	0	0.39	*LINC00501*	*97.7*	REF	DEL
chr8:23124835–23130567	5,732	DEL	0	0.02	0	0	0	0.36	*TNFRSF10D*	*96.8*	DEL	REF
chr8:23134649–23164796	30,147	DUP	0	0	0	0	0	0.48	*TNFRSF10D*[a]	*99*	DUP	DUP
chr11:60460681–60461880	1,199	DEL	0	0	0.02	0	0.17	0	*MS4A1*[a]	–	DEL	REF
chr12:101882163–101883377	1,214	DEL	0.02	0.08	0.32	0.01	0.01	0.33	*DRAM1*[a]	–	DEL	REF
chr12:104799951–104803150	3,199	DUP	0.003	0.009	0	0.01	0	0.33	*SLC41A2*[a]	*96.8*	DUP	REF
chr15:34920811–34925992	5,181	DEL	0	0	0	0	0	0.63	*AQR*[a]	*99.8*	REF	DEL
chr16p12.2	complex	DUP	0	0	0	0	0	0.82	multiple	*99.99*	REF	DUP
chr16:75059992–75060055	63	DEL	0	0	0	0	0	0.34	*ZNRF1*[a]	*96.4*	DEL	DEL
chr17:3038851–3041981	3,130	DEL	0	0	0	0	0	0.16	*RAP1GAP2*	*86.1*	DEL	DEL
chr19:42529806–42531042	1,236	DEL	0	0	0	0	0	0.54	*CEACAM1*[a]	*99.4*	DEL	DEL

Neanderthal refers to both published high-coverage genomes. The deletion within *ALLC* is only shared with the Vindija Neanderthal. The *TNFRSF10D* duplication common in Oceania is also present at low frequency (5%) in Africa. Africans do not have both deletion and duplication variants, which are in linkage disequilibrium in Oceanians (r^2 = 0.48). The duplications at chr16p12.2 at high frequency in Oceania (82%) are part of a complex structural variant (Figures S6 and S7). Population branch statistic (PBS) rank is presented for stratified variants common only in Oceania. REF, Reference; DEL, Deletion; DUP, Duplication; EUR, Europe; EA, East Asia; ME, Middle East; AMR, America; CSA, Central South Asia; OCE, Oceania.
[a]A variant that lies within or intersects a gene; otherwise, the nearest gene is presented.

evolved to fine-tune immune responses, and loss of an inhibitory receptor is hypothesized to result in immune hyperactivity and autoimmune disease (Lübbers et al., 2018). This variant shows an extreme population branch statistic (PBS; 99.87% rank), potentially indicative of positive selection

Archaic Introgression

We genotyped our calls in the high-coverage Neanderthal and Denisovan archaic genomes (Meyer et al., 2012; Prüfer et al., 2014, 2017) and found hundreds of variants that are exclusive to Africans and archaic genomes, suggesting that they were part of the ancestral variation that was lost in the out-of-Africa bottleneck. We then searched for common, highly stratified variants that are shared with archaic genomes but are not present in Africa. We identify variants across a wide range of sizes, the smallest 63 bp and largest 30 kb, within or near genes, potentially having functional consequences (Table 1).

We replicated the putatively Denisovan introgressed duplication at chromosome 16p12.2 exclusive to Oceanians (Sudmant et al., 2015b). We explored the frequency of this variant in our expanded dataset within each Oceanian population, and despite all of the Bougainville Islanders having significant East Asian admixture, which is not found in the Papuan highlanders, we do not find a dilution of this variant in the former population; it is present at a remarkable and similar frequency in all three Oceanian populations (~82%). These duplications form the most extreme region-specific variants (Figures 2D and S3), and their unusual allele distribution suggests that they may have remained at high frequencies after archaic introgression due to positive selection (Table 1). We characterized this variant in more detail using fluorescent *in situ* hybridization (Figures 2E, S6, and S7) and found that it consists of a region of the reference sequence that has duplicated and inserted into a gene-rich region ~7 Mb away in chr16p11.2, confirming a recent study (Hsieh et al., 2019). This locus is known to exhibit multiple complex recurrent structural rearrangements and is associated with ~1% of autism cases (Weiss et al., 2008). The selective pressure acting on this duplication and its target remain unknown and require further study; however, its similar frequency across the Oceanian populations examined contrasts with the differing frequency of the malaria-associated *HBA2* deletion across Oceania, suggesting that malaria infection is unlikely to be driving the signal we see at the 16p12.2 duplication.

We discover multiple additional high-frequency Oceania-private variants that are shared with the Denisovan genome (Figure 2D), illustrating the separate introgression event in Oceanians and their subsequent isolation (Browning et al., 2018; Jacobs

(D) High-frequency Oceania-specific variants (>30% frequency). See Figure S3 for more details. Each point represents a variant, with the x axis illustrating its frequency. Random noise is added to aid visualization. Almost all variants are shared with the Denisovan genome and are within (bold) or near the illustrated genes.
(E) Fluorescence *in situ* hybridization illustrating the 16p12 Oceania-specific duplication shared with Denisova in a homozygous state (cell line GM10543). Yellow arrows show reference, and red arrows illustrate duplication. See Figure S6 and S7 for more details.
(F) Distinct deletions at the SIGLEC5/SIGLEC14 locus in an Mbuti sample (HGDP00450), resolved using linked reads. Lines connecting reads illustrate that they are linked; i.e., they are from the same input DNA molecule. One haplotype (top) carries the Mbuti-specific variant that deletes most exons in *SIGLEC5* and is present at high frequency (54%), whereas the second haplotype (bottom) carries a globally common deletion that deletes SIGLEC14, creating a fused gene (see STAR Methods for more details).

et al., 2019). Many show unusual PBS (Table 1). A deletion within *AQR*, an RNA helicase gene, is present at 63% frequency and shared only with the Altai Denisovans. The highest expression of this gene is in Epstein-Barr virus-transformed lymphocytes (Lonsdale et al., 2013). RNA helicases are important in detection of viral RNAs and mediating the antiviral immune response in addition to being necessary host factors for viral replication (Ranji and Boris-Lawrie, 2010). AQR has been reported to be involved in recognition and silencing of transposable elements (Akay et al., 2017) and is known to regulate HIV-1 DNA integration (König et al., 2008). Two other notable Denisovan-shared deletions of high frequency are in *JAK1*, encoding a kinase important in cytokine signaling (44%), and *CEACAM1* (also known as CD66a), a glycoprotein part of the immunoglobulin superfamily (54%).

In the Americas, we identify a deletion, shared only with Neanderthals, that reaches ~26% frequency in the Surui and Pima. This variant removes an exon in *MS4A1* (Figure S6), a gene encoding the B cell differentiation antigen CD20, which plays a key role in T cell-independent antibody responses and is the target of multiple recently developed monoclonal antibodies for B cell-associated leukemias, lymphomas, and autoimmune diseases (Kuijpers et al., 2010; Marshall et al., 2017). This deletion raises the possibility that therapies developed in one ethnic background might not be effective in others and that access to individual genome sequences could guide therapy choice.

Multiallelic Variants and Runaway Duplications

We found a dynamic range of expansion in copy numbers, with variants previously found to be biallelic containing additional copies in our more diverse dataset. Among these multiallelic copy number variants, we find intriguing examples of ''runaway duplications'' (Handsaker et al., 2015), variants that are mostly at low copy numbers globally but have expanded to high copy numbers in certain populations, possibly in response to regionally restricted selection events (Figure 3).

We discover multiple expansions that are mostly restricted to African populations. The hunter-gatherer Biaka are notable for a private expansion downstream of *TNFRSF1B* that reaches up to 9 copies (Figure S5). We replicated the previously identified *HPR* expansions (Figure 3A) and found that they are present in almost all African populations in our study (Handsaker et al., 2015; Sudmant et al., 2015b). *HPR* encodes a haptoglobin-related protein associated with defense against trypanosome infection (Smith et al., 1995). We observe populations with the highest copy numbers to be Central and West African, consistent with the geographic distribution of the infection (Franco et al., 2014). In contrast to previous studies, we also find the expansion at lower frequencies in all Middle Eastern populations, which, we hypothesize, is due to recent gene flow from African populations.

We identified a remarkable expansion upstream of the olfactory receptor *OR7D2* that is almost restricted to East Asia (Figure 3B), where it reaches up to 18 copies. Haplotype phasing demonstrates that many individuals contain the expansion on just one chromosome, illustrating that these alleles have mutated repeatedly on the same haplotype background. However, we identify a Han Chinese sample that has a particularly high copy number. This individual has nine copies on each chromosome, suggesting that the same expanded runaway haplotype is present twice in a single individual. This could potentially lead to an even further increase in copy number through non-allelic homologous recombination (Handsaker et al., 2015).

We discovered expansions in *HCAR2* (encoding HCA_2) in Asians that are especially prominent in the Kalash group (Figure 3C), with almost a third of the population displaying an increase in copy number. HCA_2 is a receptor highly expressed on adipocytes and immune cells and has been proposed as a potential therapeutic target because of its key role in mediating anti-inflammatory effects in multiple tissues and diseases (Offermanns 2017). Another clinically relevant expansion is in *SULT1A1* (Figure 3D), which encodes a sulfotransferase involved in metabolism of drugs and hormones (Hebbring et al., 2008). Although the copy number is polymorphic in all continental groups, the expansion is more pronounced in Oceanians.

De Novo Assemblies and Sequences Missing from the Reference

We sequenced 25 samples from 13 populations using linked-read sequencing at an average depth of ~50× and generated phased *de novo* assemblies using the Supernova assembler (Weisenfeld et al., 2017; Table S2). By comparing our assemblies with the GRCh38 reference, we identified 1,643 breakpoint-resolved unique (non-repetitive) insertions across all chromosomes that, in aggregate, account for 1.9 Mb of sequences missing from the reference (Figure 4A). A San individual contained the largest number of insertions, consistent with their high divergence from other populations. However, the number of identified insertions is correlated with the assembly size and quality (STAR Methods), suggesting that there are still additional insertions to be discovered. These variants show population structure, with Central Africans and Oceanians showing the most differentiation (Figure 4B), reflecting the deep divergences within Africa and the effect of drift, isolation, and possibly Denisovan introgression in Oceania.

We find that the majority of insertions are relatively small, with a median length of 513 bp (Figure 4C). They are of potential functional consequence as 10 appear to reside within or near exons. These genes are involved in diverse cellular processes, including immunity (*NCF4*), regulation of glucose (*FGF21*), and a potential tumor suppressor (*MCC*). Although many insertions are rare—41% are found in only one or two individuals—we observe that 290 are present in over half of the samples, suggesting that the reference genome may harbor rare deletion alleles at these sites.

To further understand the origin of the sequences, we compared them with chimpanzee, gorilla, and orangutan reference genomes (Gordon et al., 2016; Kronenberg et al., 2018). We find that the majority of insertions are present in the great ape genomes, with 62% in chimpanzees, 59% in gorillas, and 35% in orangutan. These values are consistent with the evolutionary divergence between humans and the great apes. Overall, 68% are present in at least one great ape genome and 33% in all 3 genomes. Notably, for variants found in more than half of the samples, 85% are also found in the chimpanzee reference, whereas this decreases to 18% for variants only present in two individuals or less. The high percentage of common variants shared with the chimpanzee genome suggests that donors to the reference genome harbored human-specific deletions that occurred after the split from chimpanzees.

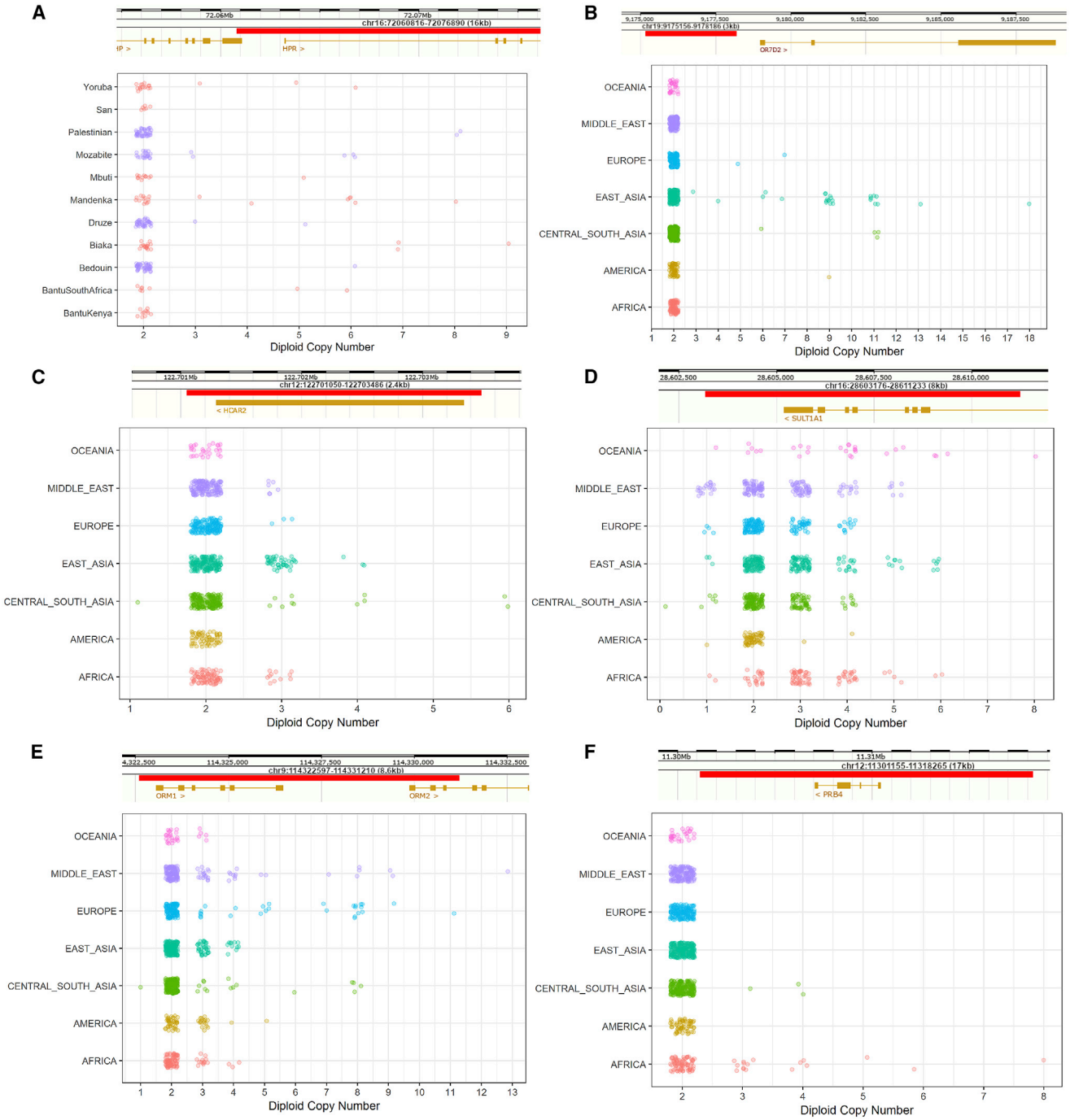


Figure 3. Copy Number Expansions and Runaway Duplications

The red bars illustrate the location of the expansion. Additional examples are shown in Figure S5.

(A) Expansion in *HPR* in African and Middle Eastern samples.

(B) Expansions upstream of *OR7D2* that are mostly restricted to East Asia. The observed expansions in Central and South Asian samples are all in Hazara samples, an admixed population carrying East Asian ancestry.

(C) Expansions within *HCAR2* that are particularly common in the Kalash population.

(D) Expansions in *SULT1A1* that are pronounced in Oceanians (median copy number, 4; all other non-African continental groups, 2; Africa, 3).

(E) Expansions in *ORM1*/*ORM2*. This expansion has been reported previously in Europeans (Handsaker et al., 2015); however, we found it in all regional groups and particularly in Middle Eastern populations.

(F) Expansions in *PRB4* that are restricted to Africa and Central and South Asian samples with significant African admixture (Makrani and Sindhi).

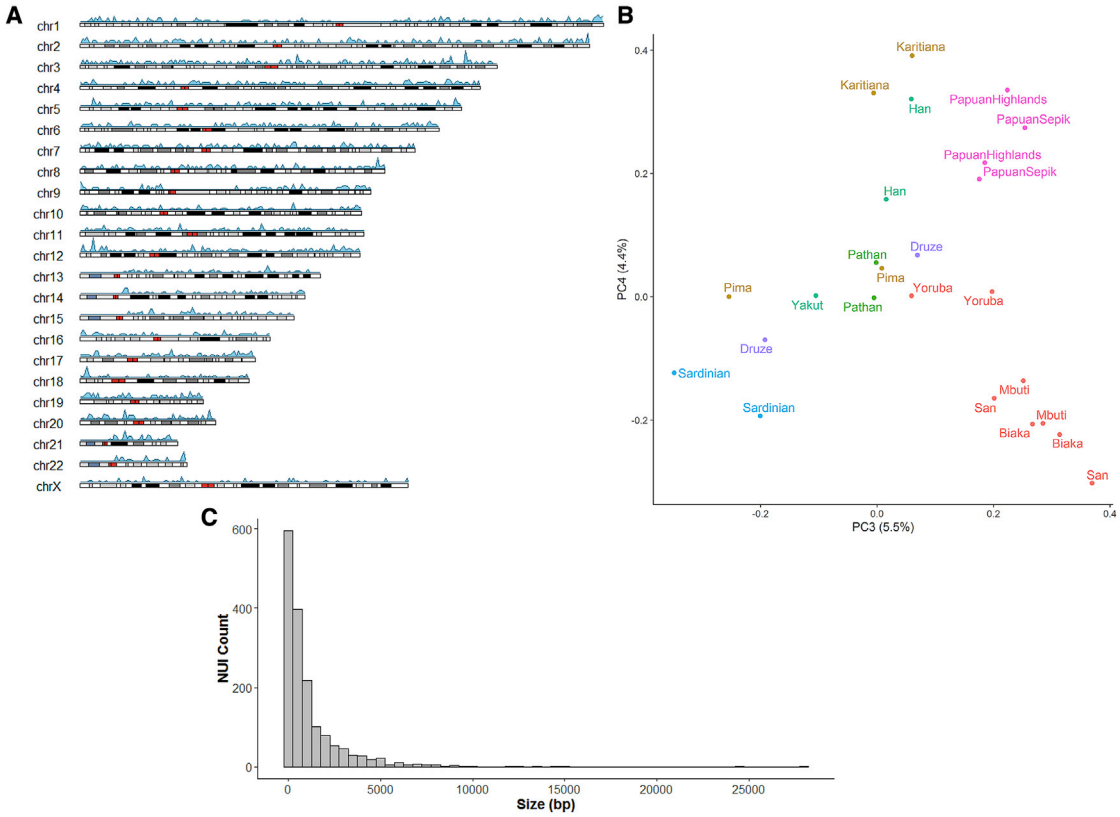


Figure 4. Non-reference Unique Insertions

(A) Ideogram illustrating the density of identified non-reference unique insertion (NUI) locations across different chromosomes using a window size of 1 Mb. Colors on chromosomes reflect chromosomal bands, with red for centromeres.

(B) Principal-component analysis (PCA) of NUI genotypes showing population structure (principal component 3 [PC3] and PC4). Previous PCs potentially reflect variation in size and the quality of the assemblies.

(C) Size distribution of NUIs using a bin size of 500 bp.

DISCUSSION

In this study, we present a comprehensive catalog of structural variants from a diverse set of human populations. Our analysis illustrates that a substantial amount of variation, some of which reaches high frequency in certain populations, has not been documented in previous sequencing projects. Our finding of common clinically relevant, regionally private variants argues for further efforts generating genome sequences without data restrictions from under-represented populations. We note that, despite the diversity found in the HGDP panel, considerable geographic gaps remain in Africa, the Americas, and Australasia.

The relatively large number of high-coverage genomes in each population allowed us to identify and estimate the frequency of population- and region-specific variants, providing insights into potentially geographically localized selection events, although further functional work is needed to elucidate their effect. Our results demonstrate that Neanderthals and Denisovans appear to have contributed potentially functional structural variants to different modern human populations. As many of the identified variants are involved in immune processes, it is tempting to speculate that they are associated with adaptation to pathogens after modern humans expanded into new environments outside of Africa.

By generating one of the most diverse sets of phased *de novo* assemblies to date, we identify and additionally place non-repetitive sequences missing from the human reference. The importance of including such variants in medical studies has been illustrated by a recent analysis in an Icelandic population that found over 100 unique insertions to be in linkage disequilibrium with a genome-wide association study marker, and one was associated with myocardial infarction (Kehr et al., 2017).

The use of short reads in this study restricts the discovery of complex structural variants, demonstrated by recent reports that uncovered a substantially higher number of variants per individual using long-read or multi-platform technologies (Audano et al., 2019; Chaisson et al., 2019). Additionally, comparison with a mostly linear human reference formed from a composite of a few individuals, and mainly from just one person, limits accurate representation of the diversity and analysis of human structural variation (Schneider et al., 2017). The identification of considerable amounts of sequences missing from the reference, in this study and others (Wong et al., 2018; Sherman et al., 2019), argues for generation of reference quality genomes from a diverse

set of the human population and creation of a graph-based pan-genome that can integrate structural variation (Garrison et al., 2018). Such computational methods and further developments in long-range technologies will allow the full spectrum of human structural variation to be investigated.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Sample Quality Control
  - Variant Calling and Quality Control
  - GenomeSTRiP
  - Manta + Graphtyper2
  - Merging of GenomeSTRiP and Manta+Graphtyper callsets
  - Comparison with Published Datasets
  - Archaic Introgression
  - Longranger and Supernova Assembly
  - Non-Reference Unique Insertions
  - Aligning Non-Reference Unique Insertions to great ape genomes
  - Fluorescent *in situ* hybridization (FISH)
  - Population Structure
  - Regional and Population-Specific Variation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Population Stratification

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2020.05.024.

ACKNOWLEDGMENTS

We thank Richard Durbin, Hélène Blanché, Thomaz Pinotti, Klaudia Walter, and members of the Tyler-Smith group for advice and discussions. We also thank Robert Handsaker for technical advice regarding the structural variant discovery algorithm. We would particularly like to thank all individuals who donated or collected samples for this study and the CEPH Biobank at Fondation Jean Dausset-CEPH for maintenance and distribution of the HGDP DNAs. M.A.A., A.B., J.P.-M., A.S.D., Y.C., C.T.-S., and Y.X. were supported by Wellcome grant 098051. M.A.A. was supported by the Government of Dubai – Dubai Police GHQ. A.B. was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001595), the UK Medical Research Council (FC001595), and the Wellcome Trust (FC001595).

AUTHOR CONTRIBUTIONS

Y.X. and C.T.-S. conceived and supervised the study. M.A.A. designed the study. M.A.A., A.B., J.P.-M., A.S.D., and Y.C. performed analyses. F.Y. and B.F. designed, performed, and interpreted FISH results. M.E.H. assisted with interpretation of results. M.A.A. wrote the manuscript with contribution from all authors.

DECLARATION OF INTERESTS

M.E.H. is a co-founder of, director of, share-holder in, and consultant to Congenica.

Received: January 10, 2020
Revised: March 4, 2020
Accepted: May 12, 2020
Published: June 11, 2020

REFERENCES

Akay, A., Di Domenico, T., Suen, K.M., Nabih, A., Parada, G.E., Larance, M., Medhi, R., Berkyurek, A.C., Zhang, X., Wedeles, C.J., et al. (2017). The helicase aquarius/EMB-4 is required to overcome intronic barriers to allow nuclear RNAi pathways to heritably silence transcription. Dev. Cell *42*, 241–255.e6.

Akkaya, M., and Barclay, A.N. (2013). How do pathogens drive the evolution of paired receptors? Eur. J. Immunol. *43*, 303–313.

Ali, S.R., Fong, J.J., Carlin, A.F., Busch, T.D., Linden, R., Angata, T., et al. (2014). Siglec-5 and Siglec-14 are polymorphic paired receptors that modulate neutrophil and amnion signaling responses to group B Streptococcus. Journal of Experimental Medicine *211*, 1231–1242.

Angata, T., Hayakawa, T., Yamanaka, M., Varki, A., and Nakamura, M. (2006). Discovery of Siglec-14, a novel sialic acid receptor undergoing concerted evolution with Siglec-5 in primates. FASEB J. *20*, 1964–1973.

Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E., Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K., et al. (2019). Characterizing the major structural variant alleles of the human genome. Cell *176*, 663–675.e19.

Axelsson, E., Ratnakumar, A., Arendt, M.L., Maqbool, K., Webster, M.T., Perloski, M., Liberg, O., Arnemo, J.M., Hedhammar, A., and Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature *495*, 360–364.

Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*, eaay5012.

Browning, S.R., Browning, B.L., Zhou, Y., Tucci, S., and Akey, J.M. (2018). Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. Cell *173*, 53–61.e9.

Cameron, D.L., Di Stefano, L., and Papenfuss, A.T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat. Commun. *10*, 3240.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.

Ceballos, F.C., Joshi, P.K., Clark, D.W., Ramsay, M., and Wilson, J.F. (2018). Runs of homozygosity: windows into population history and trait architecture. Nat. Rev. Genet. *19*, 220–234.

Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. *10*, 1784.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222.

Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M., Cavalli-Sforza, L.L., Feldman, M.W., and Pritchard, J.K. (2009). The role of geography in human adaptation. PLoS Genet. *5*, e1000500.

Durham-Pierre, D., Gardner, J.M., Nakatsu, Y., King, R.A., Francke, U., Ching, A., Aquaron, R., del Marmol, V., and Brilliant, M.H. (1994). African origin of an intragenic deletion of the human P gene in tyrosinase positive oculocutaneous albinism. Nat. Genet. *7*, 176–179.

Eggertsson, H.P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M.T., Gudbjartsson, D.F., Stefansson, K., Halldorsson, B.V., and Melsted, P. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. Nat. Commun. *10*, 5402.

Flint, J., Hill, A.V.S., Bowden, D.K., Oppenheimer, S.J., Sill, P.R., Serjeantson, S.W., Bana-Koiri, J., Bhatia, K., Alpers, M.P., Boyce, A.J., et al. (1986). High frequencies of α-thalassaemia are the result of natural selection by malaria. Nature *321*, 744–750.

Franco, J.R., Simarro, P.P., Diarra, A., and Jannin, J.G. (2014). Epidemiology of human African trypanosomiasis. Clin. Epidemiol. *6*, 257–275.

Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. *36*, 875–879.

Gel, B., and Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics *33*, 3088–3090.

Gordon, D., Huddleston, J., Chaisson, M.J., Hill, C.M., Kronenberg, Z.N., Munson, K.M., Malig, M., Raja, A., Fiddes, I., Hillier, L.W., et al. (2016). Long-read sequence assembly of the gorilla genome. Science *352*, aae0344.

Gribble, S.M., Wiseman, F.K., Clayton, S., Prigmore, E., Langley, E., Yang, F., Maguire, S., Fu, B., Rajan, D., Sheppard, O., et al. (2013). Massively parallel sequencing reveals the complex structure of an irradiated human chromosome on a mouse background in the Tc1 model of Down syndrome. PLoS ONE *8*, e60482.

Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., Kashin, S., Boettger, L.M., and McCarroll, S.A. (2015). Large multiallelic copy number variations in humans. Nat. Genet. *47*, 296–303.

Hebbring, S.J., Moyer, A.M., and Weinshilboum, R.M. (2008). Sulfotransferase gene copy number variation: pharmacogenetics and function. Cytogenet. Genome Res. *123*, 205–210.

Howe, B., Umrigar, A., and Tsien, F. (2014). Chromosome preparation from cultured cells. J. Vis. Exp. *83*, e50203.

Hsieh, P., Vollger, M.R., Dang, V., Porubsky, D., Baker, C., Cantsilieris, S., Hoekzema, K., Lewis, A.P., Munson, K.M., Sorensen, M., et al. (2019). Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. Science *366*, eaax2083.

Huddleston, J., and Eichler, E.E. (2016). An incomplete understanding of human genetic variation. Genetics *202*, 1251–1254.

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature *512*, 194–197.

Jacobs, G.S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C.C., Lawson, D.J., Mondal, M., Pagani, L., Ricaut, F.X., Stoneking, M., et al. (2019). Multiple deeply divergent Denisovan ancestries in Papuans. Cell *177*, 1010–1021.e32.

Kehr, B., Helgadottir, A., Melsted, P., Jonsson, H., Helgason, H., Jonasdottir, A., Sigurdsson, A., Gylfason, A., Halldorsson, G.H., Kristmundsdottir, S., et al. (2017). Diversity in non-repetitive human sequences not found in the reference genome. Nature Genetics *49*, 588.

König, R., Zhou, Y., Elleder, D., Diamond, T.L., Bonamy, G.M., Irelan, J.T., Chiang, C.Y., Tu, B.P., De Jesus, P.D., Lilley, C.E., et al. (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. Cell *135*, 49–60.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420–426.

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. *20*, 117.

Kronenberg, Z.N., Fiddes, I.T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O.S., Underwood, J.G., Nelson, B.J., Chaisson, M.J.P., Dougherty, M.L., et al. (2018). High-resolution comparative analysis of great ape genomes. Science *360*, eaar6343.

Kuijpers, T.W., Bende, R.J., Baars, P.A., Grummels, A., Derks, I.A., Dolman, K.M., Beaumont, T., Tedder, T.F., van Noesel, C.J., Eldering, E., and van Lier, R.A. (2010). CD20 deficiency in humans results in impaired T cell-independent antibody responses. J. Clin. Invest. *120*, 214–222.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Lindo, J., Haas, R., Hofman, C., Apata, M., Moraga, M., Verdugo, R.A., Watson, J.T., Viviano Llave, C., Witonsky, D., Beall, C., et al. (2018). The genetic prehistory of the Andean highlands 7000 years BP though European contact. Sci. Adv. *4*, u4921.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585.

Louzada, S., Komatsu, J., and Yang, F. (2017). Fluorescence in situ hybridization onto DNA fibres generated using molecular combing. Fluorescence In Situ Hybridization (FISH) (Springer), pp. 275–293.

Lübbers, J., Rodríguez, E., and van Kooyk, Y. (2018). Modulation of immune tolerance via Siglec-sialic acid interactions. Front. Immunol. *9*, 2807.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature *538*, 201–206.

Manga, P., Kromberg, J., Turner, A., Jenkins, T., and Ramsay, M. (2001). In Southern Africa, brown oculocutaneous albinism (BOCA) maps to the OCA2 locus on chromosome 15q: P-gene mutations identified. Am. J. Hum. Genet. *68*, 782–787.

Marks, P., Garcia, S., Barrio, A.M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., et al. (2019). Resolving the full spectrum of human genome variation using Linked-Reads. Genome Res. *29*, 635–645.

Marshall, M.J.E., Stopforth, R.J., and Cragg, M.S. (2017). Therapeutic antibodies: what have we learnt from targeting CD20 and where are we going? Front. Immunol. *8*, 1245.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv, arXiv:1802.03426. https://arxiv.org/abs/1802.03426.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. Genome Biol. *17*, 122.

Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.

Morgulis, A., Gertz, E.M., Schäffer, A.A., and Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J. Comput. Biol. *13*, 1028–1040.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics *28*, 1919–1920.

Nichols, B.L., Avery, S., Sen, P., Swallow, D.M., Hahn, D., and Sterchi, E. (2003). The maltase-glucoamylase gene: common ancestry to sucrase-isomaltase with complementary starch digestion activities. Proc. Natl. Acad. Sci. USA *100*, 1432–1437.

Offermanns, S. (2017). Hydroxy-carboxylic acid receptor actions in metabolism. Trends Endocrinol. Metab. *28*, 227–236.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.

Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., et al. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. Science *358*, 655–658.

Ranji, A., and Boris-Lawrie, K. (2010). RNA helicases: emerging roles in viral replication and the host innate response. RNA Biol. *7*, 775–787.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. *27*, 849–864.

Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M.P., Chavan, S., Vergara, C., Ortega, V.E., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat. Genet. *51*, 30–35.

Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The missing diversity in human genetic studies. Cell *177*, 26–31.

Smith, A.B., Esko, J.D., and Hajduk, S.L. (1995). Killing of trypanosomes by the human haptoglobin-related protein. Science *268*, 284–286.

Soylev, A., Le, T.M., Amini, H., Alkan, C., and Hormozdiari, F. (2019). Discovery of tandem and interspersed segmental duplications using high-throughput sequencing. Bioinformatics *35*, 3923–3930.

Stevens, G., Ramsay, M., and Jenkins, T. (1997). Oculocutaneous albinism (OCA2) in sub-Saharan Africa: distribution of the common 2.7-kb P gene deletion mutation. Hum. Genet. *99*, 523–527.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.Y., et al.; 1000 Genomes Project Consortium (2015a). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., et al. (2015b). Global diversity, population stratification, and selection of human copy-number variation. Science *349*, aab3761.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., and Jaffe, D.B. (2017). Direct determination of diploid genome sequences. Genome Res. *27*, 757–767.

Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T., et al.; Autism Consortium (2008). Association between microdeletion and microduplication at 16p11.2 and autism. N. Engl. J. Med. *358*, 667–675.

Wong, K.H.Y., Levy-Sakin, M., and Kwok, P.Y. (2018). De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. Nat. Commun. *9*, 3040.

Yamanaka, M., Kato, Y., Angata, T., and Narimatsu, H. (2009). Deletion polymorphism of SIGLEC14 and its functional implications. Glycobiology *19*, 841–846.

Yenchitsomanus, P.T., Summers, K.M., Bhatia, K.K., Cattani, J., and Board, P.G. (1985). Extremely high frequencies of alpha-globin gene deletion in Madang and on Kar Kar Island, Papua New Guinea. Am. J. Hum. Genet. *37*, 778–784.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
HGDP-CEPH Human Genome Diversity Cell Line Panel	CEPH (Centre d'Etude du Polymorphism)	http://www.cephb.fr/en/hgdp_panel.php
Melanesian lymphoblastoid cell lines	Coriell Institute for Medical Research	GM10543 and GM10540
Critical Commercial Assays		
Phase-Prep BAC DNA kit	Sigma-Aldrich	NA0100
the GenomePlex Complete Whole Genome Amplification kit	Sigma-Aldrich	WGA2
GenomePlex Complete Whole Genome Reamplification kit	Sigma-Aldrich	WGA3
Deposited Data		
1000 Genomes Project Structural Variation Callset	Sudmant et al., 2015a	https://www.internationalgenome.org/phase-3-structural-variant-dataset
SGDP Structural Variation Callset	Sudmant et al., 2015b	Supplementary table from study
HGDP SNV Callset	Bergström et al., 2020	ftp://ngs.sanger.ac.uk/production/hgdp
HGDP Structural Variation Callset	This study	ftp://ngs.sanger.ac.uk/scratch/project/team19/HGDP_SV/
Software and Algorithms		
GenomeSTRiP v2.00	Handsaker et al., 2015	http://software.broadinstitute.org/software/genomestrip/
Manta v1.6	Chen et al., 2016	https://github.com/Illumina/manta
GraphTyper v2.0	Eggertsson et al., 2019	https://github.com/DecodeGenetics/graphtyper
svimmer	N/A	https://github.com/DecodeGenetics/svimmer
LiftOver	UCSC Genome Browser	https://genome.ucsc.edu/cgi-bin/hgLiftOver
plink2	Chang et al., 2015	https://www.cog-genomics.org/plink/2.0/
R-3.6.0	https://www.r-project.org/	https://www.r-project.org/
EIGENSTRAT	Price et al., 2006	https://github.com/DReichLab/EIG/tree/master/EIGENSTRAT
Vcflib	N/A	https://github.com/vcflib/vcflib
bwa v0.7.12	Li & Durbin 2009	https://github.com/lh3/bwa
Picard v2.6.0	N/A	http://broadinstitute.github.io/picard/
IGV	Thorvaldsdóttir et al., 2013	http://software.broadinstitute.org/software/igv/
Long Ranger v2.12	Marks et al., 2019	https://support.10xgenomics.com/genome-exome/software/downloads/latest
Supernova v2.1.1	Weisenfeld et al., 2017	https://support.10xgenomics.com/de-novo-assembly/software/downloads/latest
NUI pipeline	Wong et al., 2018	https://github.com/wongkarenhy/NUI_pipeline
Variant Effect Predictor	McLaren et al., 2016	http://asia.ensembl.org//useast.ensembl.org/info/docs/tools/vep/index.html?redirectsrc=//asia.ensembl.org%2Finfo%2Fdocs%2Ftools%2Fvep%2Findex.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bcftools	N/A	http://samtools.github.io/bcftools/
Loupe Genome Browser	N/A	https://support.10xgenomics.com/genome-exome/software/downloads/latest

RESOURCE AVAILABILITY

Lead Contact

Further information and requests should be directed to Mohamed A. Almarri (ma17@sanger.ac.uk).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The datasets generated during this study are available at ftp://ngs.sanger.ac.uk/scratch/project/team19/HGDP_SV/

Raw read alignments are available from the European Nucleotide Archive (ENA) under study accession number PRJEB6463. The 10x Genomics linked-reads data are available at ENA under study accession PRJEB14173.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Extracted DNA of samples sequenced in this project were provided by the HGDP-CEPH (Cann et al., 2002). Number of samples per population and regional labels are presented in Table S1. For more detailed information on the population labels, sequencing and mapping process of the samples analyzed in this dataset, refer to Bergström et al. (2020). In brief, 10 samples (PCR) were sequenced in a previous study for comparison with the Denisovan genome (Meyer et al., 2012), all using PCR-based libraries (subsequently called "Meyer" samples). An additional 142 samples were sequenced as part of the Simon Genome Diversity Project ("SGDP"), mostly using PCR-free methods (Mallick et al., 2016). The remaining 808 samples were sequenced at the Wellcome Sanger Institute using either library preparation method, and in some cases both on the same sample, resulting in 823 genome sequences ("Sanger" samples). Twelve SGDP and two Meyer samples were also independently sequenced at Sanger. Each of the Sanger, SGDP and Meyer samples used sequencing technologies with different read lengths (2 × 151 bp, 2 × 100 bp, and 94+100 bp or 95+101), mean depth (35x, 42.4x, 28x) and insert sizes (447 bp, 310 bp, 264 bp) respectively. All sample reads were processed through the automated pipeline of the Wellcome Sanger Institute sequencing facility and mapped to the GRCh38 reference.

In addition, for the fluorescent *in situ* hybridization analysis (see Method Details), Melanesian lymphoblastoid cell lines were purchased from Coriell Institute for Medical Research (samples GM10543 and GM10540)

METHOD DETAILS

Sample Quality Control

As the whole dataset is derived from lymphoblastoid cell lines, we searched for potential cell-line artifacts by analyzing local coverage of each sample. Coverage was calculated at ~300,000 single positions across the genome and a rolling mean was plotted, normalized by the genome-wide median. Each chromosome in all samples was manually inspected for variation in depth.

In the SNP analysis (Bergström et al., 2020), a total of 929 samples remained after quality control, including some samples exhibiting copy number gains, as these were observed to have minor effects on genotyping accuracy. Here, we subsequently excluded an additional 10 samples that show putative cell-line artifacts across multiple chromosomes. For samples showing more limited putative artifacts, we masked such regions and set any calls within them to missing. A total of 74 samples contained masked regions. This included the sex chromosomes, where we identify many instances of partial loss of Y chromosomes in addition to observing a single XXY male, which could be a natural occurrence rather than an artifact. This resulted in a total dataset of 919 samples composed of 644 Sanger PCR-free, 147 Sanger PCR, 111 SGDP PCR-free, 9 SGDP PCR and 8 Meyer.

Variant Calling and Quality Control

Two recent studies have comprehensively evaluated different short-read structural variant callers and provided recommendations and best practices (Cameron et al., 2019; Kosugi et al., 2019). We choose to use Manta (Chen et al., 2016), an assembly-based caller, as it performed well in these studies. Additionally, to complement Manta, we also used GenomeSTRiP (GS, Handsaker et al., 2015), which uses read-depth and read-pair information to identify copy number variants, and the copy number state of each variant. Manta only identifies tandem-duplications, and not interspersed ones, while GS which is able to identify both types. GS can also be used to

genotype the copy number variants in the archaic genomes (see Archaic Introgression section). GS identifies deletions, duplications (tandem or interspersed) and multiallelic variants > 1kb. Manta identifies deletions, insertions, inversions, tandem duplications and interchromosomal translocations > 50bp. Manta is not able to detect expansions or contractions of a reference tandem repeats, interspersed duplications, large insertions (maximum fully-assembled insertion is approximately twice read-pair fragment size). GenomeSTRiP v2.00 and Manta v1.6 were run using default parameters.

GenomeSTRiP

We first ran the algorithm jointly on all libraries, including libraries not passing quality control for short variant calling. We subsequently found the Meyer libraries to have lower quality of calls and re-ran the algorithm excluding them.

Duplicate samples prepared using both PCR and PCR-free libraries were created for quality control purposes. We ran GenomeSTRiP twice, once including the PCR prepared duplicates, and the second with PCR-free duplicates, together with the rest of the dataset. Comparing both callsets revealed that PCR-based libraries contained a higher number of shared heterozygous calls that were missing from the PCR-free libraries. These calls were excluded using the excessive heterozygosity tag calculated by bcftools v1.9 (ExcHet < 0.0001) separately for each library preparation and sequencing location set (i.e., SGDP PCR, SGDP PCR-free, Sanger PCR and Sanger PCR-free). For the SGDP PCR samples we used ExcHet < 0.05 due to this callset only having 9 samples. After this quality control, a VCF with 50,474 CNVs from 911 samples was generated. We find no detectable batch effects, with the top PCs displaying continental variation and subsequently population variation. We show PC1-4 for some examples in Figure S2.

We examined the callset and identified instances where the algorithm splits single variants into multiple shorter entries which are not always overlapping. This a known behavior of the GenomeSTRiP CNV pipeline which seems to occur if a low quality variant is found within a larger CNV or when there are variants with different copy numbers across different individuals within a sub-segment of a larger variant. To more accurately estimate the total number of identified CNVs in our dataset accounting for these issues, we merged high quality (CNQ > 12) calls that have the same diploid copy number and are within 50 kb of each other, for each sample separately. At this step we found one sample (HGDP01254) that, although not showing any observable chromosomal abnormalities, contained slightly elevated numbers of variants compared to the rest of the samples. These calls had relatively low genotype quality. We chose to be conservative and subsequently excluded this sample from the GenomSTRiP callset, leaving 910 individuals. All variants were then merged using bedmap v2.4.35 (Neph et al., 2012) based on 100% overlap. This identified 39,634 autosomal variants, 1,102 variants on the X chromosome and 289 variants on the Y chromosome. 22,914 variants were composed of biallelic deletions, 16,012 were duplications, and 2,099 were variants with both deletion and duplication alleles (Figure S1).

Manta + Graphtyper2

We ran Manta v1.6 (Chen et al., 2016) on the 911 libraries discussed above to generate individual VCFs for each sample. We then extracted variants that 'PASS' all the quality thresholds of the algorithm. In Manta v1.6, inversions are reported as breakends (BND), we subsequently used a script provided with the Manta download (convertInversion.py) to convert them into single inverted sequence junctions, as represented in previous versions. We masked the potential cell-line artifact regions identified in the samples as in the previous step. We subsequently merged all samples using svimmer (https://github.com/DecodeGenetics/svimmer) under default conditions, as performed in Eggertsson et al. (2019). The merged dataset comprised 160,958 variants.

As the Manta call set is not joint-called, differences in read lengths, insert sizes, coverage and library preparation in the HGDP dataset may create batch effects. Additionally, a variant found in one sample may be present but missed in another sample due to the differing variables mentioned above. To address this, we discarded the original genotypes identified by Manta for each sample and jointly regenotyped the merged dataset across all samples concurrently using Graphtyperv2.0 (Eggertsson et al., 2019). This algorithm creates an acyclic mathematical graph structure to represent the reference genome and identified structural variants, to which reads are then re-aligned and genotyped. The algorithm provides three different genotyping models: 'coverage', 'breakpoint' and also an 'aggregate' model that uses information from the two previous models. We extracted the 'aggregate' model as suggested for all variants (Eggertsson et al., 2019), except inversions which we used the breakpoint model (no aggregate model was identified for inversions). We excluded variants with size over 10 Mb and set all variants with GQ < 20 to missing. We also set variant genotype calls that have a 'FAIL1' tag to missing. For duplications, we also so 'FAIL2′ and 'FAIL3′ to missing. We excluded variants with (ExcHet < 0.00001) across the entire dataset and also separately for each library and location samples set (SangerPCR < 0.001, SangerPCRfree < 0.00001, SGDP PCR < 0.05, SGDP PCRfree < 0.00001). Finally, we removed any monomorphic variants. To test for batch effects, we ran a principal component analysis separately for each class of variant identified (DEL, DUP, INS, INV). We find no observed batch effects across all classes, with the top PCs displaying continental variation and subsequently population variation. The final analyzed Manta callset included 68,089 deletions, 25,084 insertions, 7,290 duplications, 1,895 inversions and 1,667 translocations.

Merging of GenomeSTRiP and Manta+Graphtyper callsets

To identify non-overlapping variants in both callsets we used bedmap v2.4.35 (Neph et al., 2012) with a threshold of 50% reciprocal overlap. This identified 126,018 unique variants. To evaluate the accuracy of genotype calling, we extracted regional-specific variants present in both GenomeSTRiP and Manta+Graphtyper callsets (2,140 variants). We observe high correlation of variant allele frequencies between both callsets (r = 0.97; Figure S1), with the slight differences partly due to varying missingness.

Comparison with Published Datasets

We compared our dataset with two global-scale structural variation callsets:

1) The 1000 Genomes Phase 3 Structural Variation Dataset (1000G, Sudmant et al., 2015a), downloaded from: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/GRCh38_positions/
2) The copy number analysis of the Simons Genome Diversity Project (Sudmant et al., 2015b).

As the SGDP callset is mapped using GRCh37, we used the UCSC LiftOver function to GRCh38 using default parameters (https://genome.ucsc.edu/cgi-bin/hgLiftOver). We observe that 294 variants failed LiftOver and were not further considered. The README file for the downloaded 1000G dataset notes that these variants were lifted over to GRCh38, leading to the exclusion of 121 variants. We lifted over variants from our dataset to GRCh37, excluding translocations, and found 4,495 that fail. As these variants will increase our novelty estimate, we chose to exclude them for comparison with the published datasets. We did not include translocations in the comparison.

We used a threshold of 30% reciprocal overlap between variants identified in our dataset and either published callset to classify them as the same variant. Using higher thresholds results in a higher novelty rate for our dataset, and since the earlier studies used older discovery algorithms and are based on shorter reads, this results in larger confidence interval around the position of the SV. Therefore we chose a more relaxed lower threshold to be more conservative. We note that the SGDP SV study also used 30% to assess the novelty of their dataset (Sudmant et al., 2015). This was implemented using bedmap v2.4.35 (Neph et al., 2012). For the comparison, we chose to be conservative by assessing whether a locus is structurally variable, rather than comparing the class of variant between the callsets. The reasoning for this is the possible misclassification of variant class (e.g., insertion versus duplication, in addition some inversions identified in the 1KG have since been shown to be inverted duplications and deletions (Soylev et al., 2019)). This analysis shows 78% of variants in our callset not to be present in either the 1000G or SGDP callsets. Some of these variants reach high frequency in regional groups or individual populations (Figure S3). We also compared the number of variants in the SGDP and 1000G that is not present in our dataset, and find this to be 53% and 64% respectively. Though note in this comparison we looked at all variants in the 1000G and whether a locus is structurally variable in our dataset, this included classes of variants that were not investigated in this study but were in the 1000G, such as relatively large polymorphic mobile element insertions. Of the 1000G variants not present in our dataset, 93% are < 1% frequency in the 1000G.

To further evaluate the quality of our callset, we extracted variants common in African populations in the 1KG that overlap African variants in our dataset, based on 50% reciprocal overlap (> 5% minor allele frequency in 1KG). This resulted in 1,671 variants. Deletions were chosen because it had the highest sensitivity in the 1KG dataset, in addition because the 1KG dataset is mostly composed of deletions (82%, excluding mobile element insertions). Although we expect some variation due to the different African populations in the two datasets, we should see a correlation at common variation at a continental level. Indeed, we do find high correlation of allele frequencies between the two callsets ($r = 0.95$; Figure S1).

Archaic Introgression

We genotyped the identified CNVs from this study (GenomeSTRiP calls) in the three published high coverage archaic genomes: Altai Denisova (Meyer et al., 2012), Altai Neanderthal (Prüfer et al., 2014) and Vindija Neanderthal (Prüfer et al., 2017). In these previous studies, sequencing reads for each sample were aligned to GRCh37. For our analyses, we downloaded these mapped reads and remapped them to GRCh38 using bwa aln v0.7.12 (Li & Durbin 2009), with parameters tuned for ancient DNA ("-l 16500 -n 0.01 -o 2"), and marked duplicates using the MarkDuplicates tool from Picard v2.6.0 (http://broadinstitute.github.io/picard/). Each ancient genome was joint-called separately using a site VCF with 30 Sanger-PCR samples using GenomeSTRiP. This was done as we were concerned that the different library preparations of the HGDP dataset may affect calling in the archaic genomes, so we limited joint calling to single library (Sanger-PCR) and a single archaic genome. We then investigated variants that were highly stratified (Vst > 0.2) and shared with any archaic genome but missing from African populations, and restricted analysis to high quality archaic variant calls (CNQ $\geq$ 13). All identified putative introgressed variants were then checked and confirmed manually in IGV (Thorvaldsdóttir et al., 2013). In the Manta dataset we identified a relatively small deletion within *JAK1* (375 bp) which is specific to Oceania at 44% frequency. We checked the archaic genomes in IGV and find the Denisovan genome to be homozygous for the deletion, and the Altai Neanderthal to be homozygous reference. The Vindija Neanderthal shows a less clear genotype: we do see a reduction in depth relative to flanking regions; however, due to the small size of the region it is difficult to ascertain if it is heterozygous for the deletion or if the reduction in depth is due to stochastic noise. To be conservative we do not consider the Vindija Neanderthal genotype in Table 1 in the main text. We also identify small deletion (63 bp) within *ZNRF1* specific to Oceania at 34% frequency. Manually checking the variant using IGV in the archaic genomes illustrated that all three are homozygous for the deletion.

For the chr16p12 duplications exclusive to Oceanians, we used the estimated individual ancestry from SNV analysis (Bergström et al., 2020) and find that all 11 Bougainville samples have appreciable East Asian ancestry (average 19%, minimum 16%, maximum 21%), one out of the eight Sepik/Lowlanders had 20%, while all eight Highlanders show no East Asian component.

Longranger and Supernova Assembly

In the SNV analysis (Bergström et al., 2020), 26 HGDP samples from 13 populations (two per population) were sequenced using 10x Genomics linked-reads. For the present study, we performed an additional lane of sequencing for these 26 samples from the same library preparation to increase coverage for structural variants analysis and additionally for the *de novo* assembly using the Supernova assembler v2.1.1. We used the Long Ranger v2.12 pipeline (Marks et al., 2019) which generated phased VCFs of structural variants. This was performed twice, once for single-lane and another for two-lane (higher coverage) libraries. In this study, we use the linked reads to validate variants we identified in the standard Illumina WGS and present it as a resource for the scientific community. We also used linked reads from two lanes as input to Supernova v2.1.1 and selected the pseudohap2 output which extracts both pseudohaplotypes (Weisenfeld et al., 2017). Assembly statistics are presented in Table S2. We observe variable contiguity and assembly sizes for the assemblies, likely reflecting the initial average molecular size for each sample. One sample had a markedly smaller assembly size compared to the rest (HGDP00954), and was excluded for assembly based analysis, leaving 25 samples.

Non-Reference Unique Insertions

To identify non-reference unique (non-repetitive) insertions (NUIs), we used the NUI pipeline which compared each of the Supernova assemblies to the GRCh38.p12 reference (Wong et al., 2018, https://github.com/wongkarenhy/NUI_pipeline). We followed the definition of NUI as proposed by Wong et al., 2018: "full-length insertions that harbor at least 50 bp of non-repetitive sequences not found in the hg38 reference set, including alternative haplotypes and patches." Sequencing reads were extracted from BAM files generated from samples sequenced using one lane by the Long Ranger v2.12 pipeline. Briefly, the pipeline takes poorly mapped, unmapped and discordant reads from the Longranger output and maps them to the Supernova assemblies. It subsequently identifies read clusters and extends the contig ends to use as anchors, which are then aligned against GRCh38. Breakpoints are subsequently identified and filtered to identify NUIs. Sequences are masked using RepeatMasker (http://www.repeatmasker.org) and dustmasker (Morgulis et al., 2006) and only sequences with 50 or more unique (i.e., non-masked) bases are kept. These NUIs are then blasted to GRCh38p.12, including all the patches, to confirm they are not present in the reference. The number of NUIs per sample is shown in Table S2. To assess the potential functional effect of each identified insertion we used the Variant Effect Predictor (McLaren et al., 2016). We set the "Upstream/Downstream distance (bp)" = 0 and extracted "canonical" transcripts from the predicted results. To identify if coding sequences are affected, we filtered for "coding_sequence_variant." Some insertions affected more than one transcript. For PCA we excluded variants that are present in four or less individuals or that are present in more than 23 individuals and used the prcomp function in R-3.6.0 with default parameters. We show PC3-4 as the top two PCs likely represent variation in assembly size and quality, with correlation observed between PC2 values and contig N50 ($r = 0.63$). Additionally, the number of identified insertions is correlated with the contig N50 ($r = 0.91$). NUIs density across chromosomes was plotted using karyoploteR v1.10.4 (Gel and Serra, 2017).

Aligning Non-Reference Unique Insertions to great ape genomes

To identify whether the NUIs identified in this project are present in other closely-related great apes, we downloaded chimpanzee (panTro6), gorilla (gorGor5) and Orangutan (ponAbe3) reference genomes from UCSC (Gordon et al., 2016; Kronenberg et al., 2018). We then used blastn to align the sequences, including 50 bp flanking the NUI sequences, using the options –task megablast and –dust no. We considered the sequences to be present in the great ape genomes if they aligned with at least 95% identity and 95% query coverage.

Fluorescent *in situ* hybridization (FISH)

Melanesian lymphoblastoid cell lines were purchased from Coriell Institute for Medical Research (GM10543 and GM10540) while fosmid and bacterial artificial chromosome (BAC) clones used in this study were provided by the clone archive team of the Wellcome Sanger Institute (Table S3). Fosmid/BAC DNA was prepared using the Phase-Prep BAC DNA kit (Sigma-Aldrich) following the manufacturer's protocol. For fiber-FISH, stretched chromatin and DNA fibers were prepared by alkaline lysis of lymphoblastoid cells deposited on Thermo Scientific Polysine adhesion slides (Fisher Scientific) as described previously (Korbel et al., 2007). Purified fosmid/BAC DNA were first amplified using the GenomePlex® Complete Whole Genome Amplification kit (WGA2) (Sigma-Aldrich) and then labeled with either biotin-16-dUTP, Dinitrophenol (DNP)-11-dUTP or Digoxigenin (DIG)-11-dUTP (Jena Bioscience) using the GenomePlex® Complete Whole Genome Reamplification kit (WGA3) (Sigma-Aldrich) as described in Louzada et al. (2017). The DNP-labeled probes were detected with rabbit anti-DNP and Alexa 488 conjugated goat anti-rabbit IgG (Invitrogen). The DIG-labeled probes were detected with monoclonal mouse anti-DIG IgG (Sigma-Aldrich) and Texas red conjugated donkey anti-mouse IgG (Invitrogen). The biotin-labeled probes were labeled with biotin-16-dUTP and detected with one layer of Cy3-streptavidin (Sigma-Aldrich). After detection, slides were mounted with SlowFade Gold® (Invitrogen) mounting solution containing 4′, 6-diamidino-2-phenylindole (Invitrogen). Metaphase chromosomes were prepared from lymphoblastoid cell lines following standard procedure (Howe et al., 2014). Metaphase- and interphase-FISH essentially followed Gribble et al. (2013). Probes directly labeled ChromaTide Texas Red®-12-dUTP (Invitrogen), Green-dUTP (Abbott), Cy3-dUTP and Cy5-dUTP (Enzo) were used in this study. Images were captured on a Zeiss AxioImager D1 fluorescent microscope and processed with the SmartCapture software (Digital Scientific UK).

Population Structure

We ran PCA using plink2 v2.00a2LM (Chang et al., 2015). We set variants with GQ < 20 to missing, included variants with minor allele frequency > 1%, missingness < 1% and pruned for linkage disequilibrium using the option–indep-pairwise 50 5 0.2. For the GenomeSTRiP dataset, we extracted biallelic deletions and biallelic duplications and ran the PCA separately. We excluded a single variant from the pruned duplication set due to it likely being affected by genotyping error (Hardy-Weinberg equilibrium (HWE) test = 1.73e-24). For multiallelic variants, we used a newer version of plink2 (v2.00a3LM) which can run PCA for multiallelic variants, and using the same parameters above.

Due to the relatively large number of PCs with observed patterns of structure, reflecting the diversity of our dataset, we ran a Uniform Manifold Approximation and Projection (UMAP) on the top 10 PCs that show population structure in the GenomeSTRiP deletion PCA (McInnes et al., 2018). This was implemented in R-3.6.0 using the package uwot (v0.1.3; https://cran.r-project.org/web/packages/uwot/index.html) setting initialization for the coordinates as 'spca', min_dist = 0.001, and n_neighbors = 16.

For biallelic duplications, we see structure limited to the first four PCs. However, the first two PCs separate Africans and Oceanians from the rest of the samples, in contrast to deletions. To further investigate this, we looked at the variant loadings in the PCA and find two variants with particularly high loadings, which when excluded, returns a similar, albeit much less defined, pattern to deletions. Those two variants were found to be the Oceanic-specific duplication on chr16p12 putatively introgressed from Denisovans and the highly differentiated *TNFRSF10D* variant. The relatively small number of large biallelic duplications identified renders the PCA sensitive to the few highly stratified variants found in Oceanians. A UMAP was run on the top 4 PCs as described above. The multiallelic variant UMAP was run on the top 10 PCs.

We also ran a PCA using the same parameters above on all classes identified in the Manta+Graphtyper callset. We additionally excluded variants with HWE < 0.0001, and similarly to the GenomeSTRiP callset, we see population structure across all classes. However, we find more defined structure in the deletion Manta+Graphtyper callset in comparison to the GenomeSTRiP callset, which is likely due to the larger number variants identified by Manta. We ran a UMAP on the top 20 PCs deletion genotypes as implemented above and, as expected, see a more defined pattern of structure (Figure S2). We also ran a UMAP for insertions (top 10 PCs) using the same parameters. In Figure 1 of main text, the deletion and insertion UMAP was constructed from the Manta dataset, while the biallelic duplication and multiallelic variant UMAP was run using the GenomeSTRiP callset.

Regional and Population-Specific Variation

We explored the total number and frequency of variants that are specific to continental and geographic regions (Figure S3). As this analysis is sensitive to individuals with recent admixture, we used previous estimated individual ancestry from SNV analysis and excluded samples that show such admixture (for more details refer to Bergström et al., 2020). To further conservatively avoid over-counting single variants that have been called as multiple adjacent entries, potentially as a result of a complex structural event, we merged variants with similar allele frequencies and the same copy number lying within 25 kb of each other.

For multiallelic copy number variants, we restricted the analysis to high quality variants that have CNQ $\geq$ 13. This score is phred-scaled, with CNQ $\sim$13 representing $\sim$95% confidence of diploid copy number. In the expansion plots presented (Figure 3 and Figure S5), the highlighted regions (red bar) illustrate the expanded regions. However, in some cases the discovery algorithm finds the expanded region to vary in size and can be slightly larger in different samples. To be conservative we display the smallest overlapping region consistent across samples.

Similarly, we explored the total number and frequency of variants that are only found in a specific population (Figure S3). Here, we did not exclude samples based on known admixture as in the regional analysis. We note that this analysis is sensitive to the sampling location and sample size of each population, i.e., if a region is more comprehensively sampled we would expect a lower number of population-private variants in contrast to more sparsely sampled regions. In addition, even if we find a variant that seems population specific, it may be present at a lower frequency in another population but was not captured due to sample size. Nevertheless, we still identify examples of high-frequency population-specific variants that are not found in geographically nearby populations.

OCA2 deletion in BantuSouthAfrica

We find a 2.7 kb deletion in *OCA2* (also known as *P* gene) to be of surprising frequency (44%) and an outlier exclusive in the Bantu South African population (Figure S3). This deletion has been reported previously in African populations, and is known to cause Brown Oculocutaneous Albinism following a recessive mode of inheritance (Durham-Pierre et al., 1994; Manga et al., 2001). We find homozygotes for this deletion in our dataset, suggesting that samples with albinism were donated to the HGDP collection. We contacted CEPH (Centre d'Etude du Polymorphism) about this observation and were informed that Trefor Jenkins (now deceased) was the researcher who provided samples from this population. As he has a history of working with African populations with albinism (Stevens et al., 1997), we conclude that this variant in the HGDP dataset is likely to result from the particular sample ascertainment rather than being representative of its frequency in the Bantu South African population.

SIGLEC5 deletion in Mbuti

In the main text, we report a deletion that is specific and high frequency (54%) in the Mbuti population that deletes the inhibiting receptor *SIGLEC5* without removing its paired activating receptor *SIGLEC14* (Ali et al., 2014). We also find a previously reported deletion which removes the function of the activating Siglec-14, to be common in all populations (global frequency 38%), with particularly high frequency in East Asians (63%). This common deletion removes the activating receptor, by fusing *SIGLEC5* and *SIGLEC14*, creating a gene that has the *SIGLEC5* coding sequence and expressed under the promoter of *SIGLEC14* (Yamanaka et al., 2009).

We discover a single Mbuti sample (HGDP00450) that has both deletions on separate haplotypes. By looking at depth in this region the two deletions appear complex, but we were able to resolve them using 10x linked reads (Figure 1).

QUANTIFICATION AND STATISTICAL ANALYSIS

Population Stratification

We calculated the maximal allele frequency difference for each population pair (total 1431 pairwise comparisons) and assessed this in relation to the average SNV differentiation between each population (SNV Fst). SNV Fst was calculated using EIGENSTRAT on all SNPs within the accessibility mask defined in Bergström et al. (2020) (Price et al., 2006). The distribution of values for SNVs should provide a conservative null distribution for other classes of variants, as the expected fate of a neutral variant in a population does not depend on the nature of the variant (SNV or SV), and most SNPs are neutral, or nearly so, and should drift to the same degree. Using SNPs as the null distribution would only lead to false positive outliers if structural variants experienced stronger drift than SNPs, and there is no reason to believe this (if anything, they might on average be under stronger purifying selection and thus experience less drift). Recurrent mutations at unstable structural variants might lead to a departure from this expectation, but is conservative in this context because it should only lead to decreased differentiation relative to what's observed at SNPs, not increased differentiation.

We calculated structural variant allele frequency and missingness in each population separately setting variants with GQ < 20 to missing, and excluded variants with missingness > 25% in each population. We then calculated the maximal variant allele frequency difference for each population pair, separately for deletions (which include biallelic deletions and deletions in multiallelic sites), biallelic duplications and insertions. As these values are sensitive to the sample size of each population, we assessed this relationship in Figure S4. To assess whether the observed outlier variants are significantly stratified between two closely-related populations, we used the pVst option in the vcflib package (https://github.com/vcflib/vcflib) which tests for stratification and outputs an empirical p value using 1000 permutations. For the *HBA2* deletion we find almost fixed in the PapuanSepik population, we find the Papuan-Highlanders, who do not have the deletion, have high missingness at this variant. The variant GQ is 19 for almost all samples in this population, just missing the threshold we set. However, closer inspection of the variant quality shows that the copy number genotype quality (CNQ) was high for these individuals (all CNQ > 70, except one CNQ = 18). Thus this variant was subsequently included in this population for analysis.

We calculated population branch statistics (PBS) for populations showing highly-stratified and high-frequency private variants. The PBS distribution for SNVs was found to be very similar, and slightly more conservative, than structural variants. For e.g., the PBS statistic calculated for (Oceanians; Europeans, East Asian): the 99% PBS threshold for SNPs: 0.65, DEL: 0.63, DUP: 0.59. Using (Karitiana; Surui,Han) shows a 99% rank of 0.65 for SNPs and 0.65 for deletions, while for (Mbuti; Biaka, Han): SNPs 0.6 and deletions 0.59. We subsequently used the SNVs PBS as a conservative distribution to assess whether a stratified variant shows a departure from neutrality. For all variants, we filtered for a minor allele frequency threshold of 1% and removed variants showing > 10% missingness. We used a rank of 99% as a threshold for evidence of departure from neutrality. We calculated PBS using (Oceanians; Sardinians, Han), (Karitiana; Surui,Han) and (Mbuti; Biaka, Han). The Karitiana deletion in *MGAM* discussed in the main text shows a PBS rank of 98.7%.

Supplemental Figures

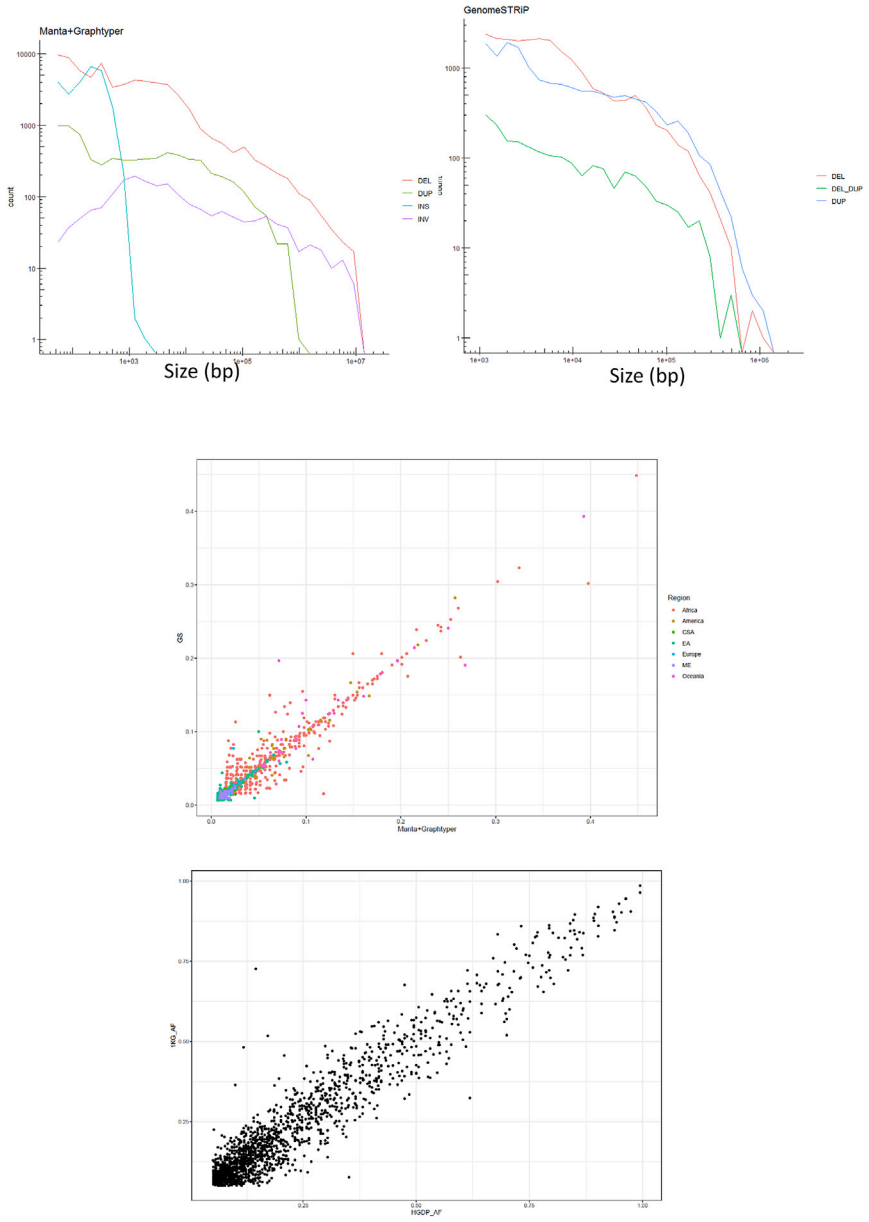


Figure S1. Dataset Quality Control, Related to STAR Methods

Top: Size distribution of identified variants that passed all filters and were included in the final callset. Note the differences in scales between the two plots. Left: Manta+Graphtyper. Right: GenomeSTRiP – green line shows variants that have both deletion and duplication alleles. Centre: Correlation of allele frequency of variants identified by both Manta+Graphtyper and GenomeSTRiP within the HGDP dataset (Regional-specific variants, colored by region). Bottom: Allele frequency correlations between deletions identified in the 1000G and the HGDP Manta+Graphtyper callset (using African variants > 5% frequency in 1KG).

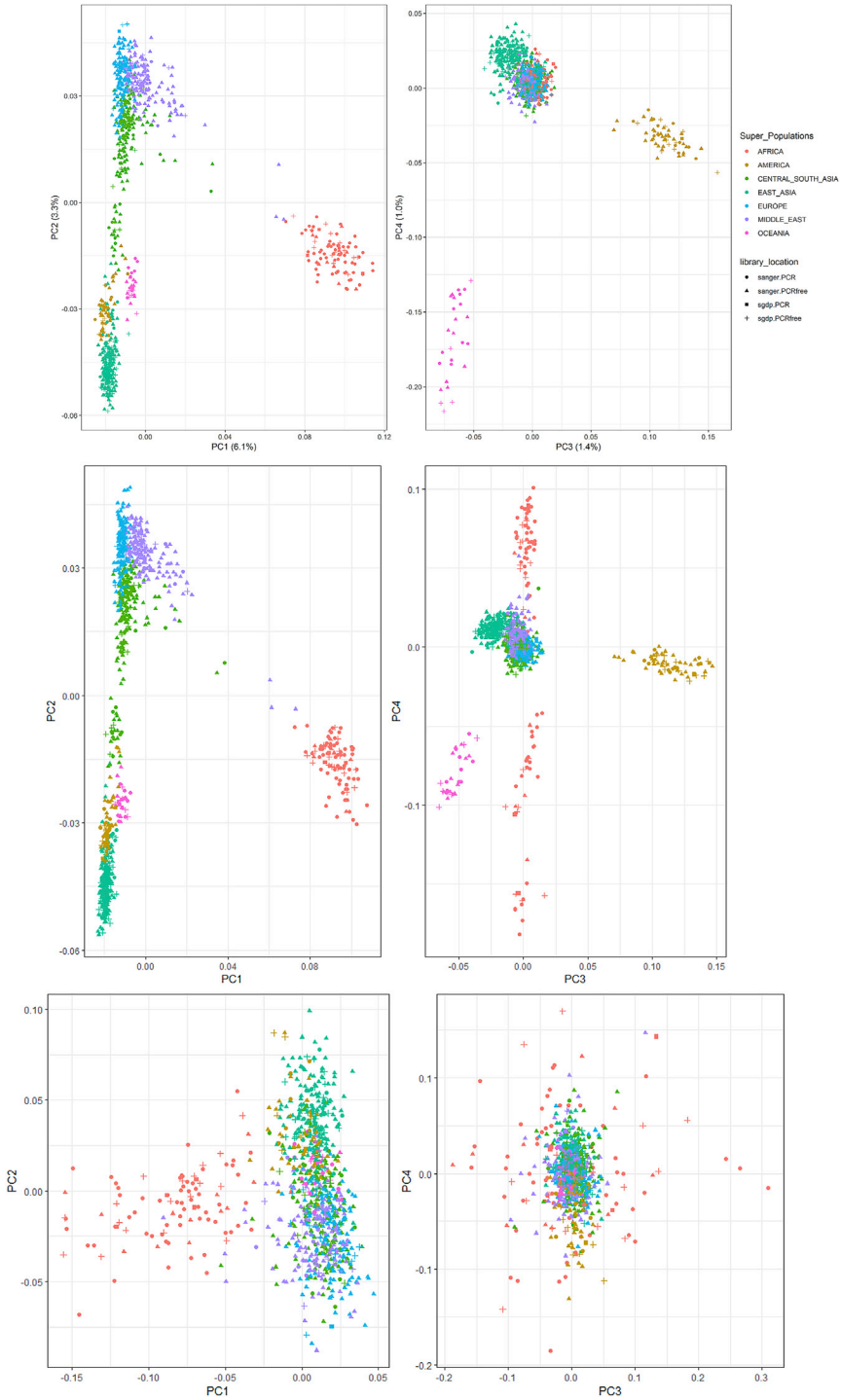


Figure S2. Population Structure, Related to Figure 1 and STAR Methods

No batch effects identified between samples prepared using different library preparations and sequenced in different centers. Top: PCA (1-4) of GenomeSTRiP biallelic deletion genotypes by sample library preparation and sequencing location. Centre: PCA1-4 of Manta+Graphtyper deletion genotypes by sample library preparation and sequencing location. Bottom: PCA1-4 of Manta+Graphtyper inversion genotypes by sample library preparation and sequencing location.

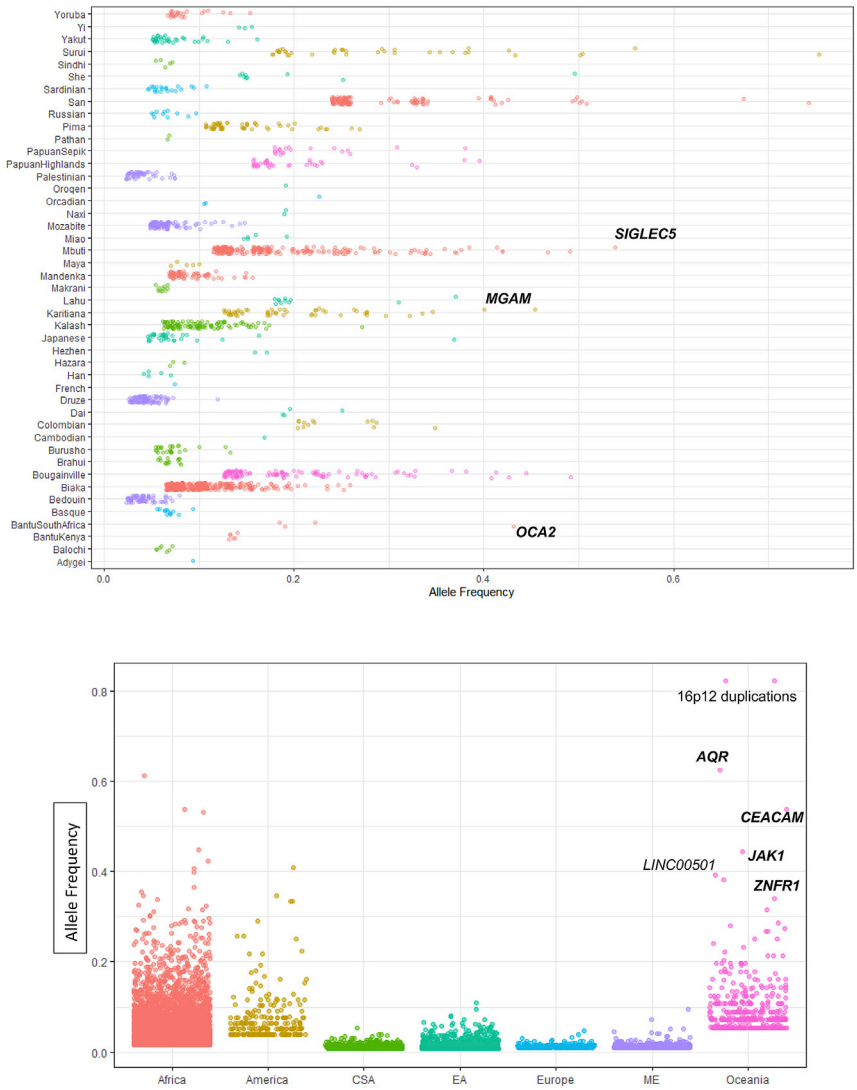


Figure S3. Population- and Region-Specific Variation, Related to Figures 2D–2F

Top: Population-specific variation - Each point represents a variant private to a population (n > 2) with the x axis reflecting its frequency. Colors represent regional labels and random noise is added to aid visualization. High-frequency variants discussed in the text are highlighted. Bottom: Regional-Specific Variation – Each point represents a variant private to a regional group (n > 2) with the y axis illustrating its frequency. Random noise is added to aid visualization. The distribution reflects the ancestral diversity in Africa, the connectivity of Eurasia, the isolation & drift of the Americas and Oceania, and the separate Denisovan introgression event in Oceania. Oceania is notable for having private high-frequency variants that are all shared with the Denisovan genome and are within (bold) or near the illustrated genes, four of which are newly identified in this study (*AQR*, *CEACAM*, *JAK1*, *ZNFR1*). The Americas contain high frequency variants which are not shared with any archaic genomes, suggesting they arose and increased to high-frequency after they split from other populations. EA: East Asia, CSA: Central & South Asia, ME: Middle East.

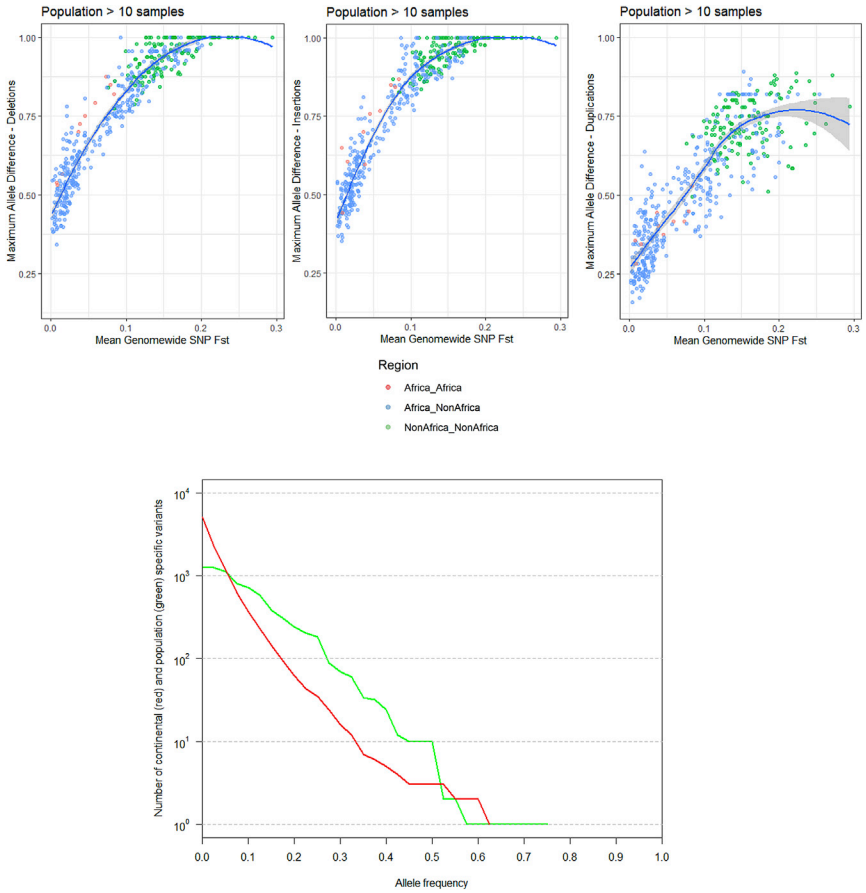
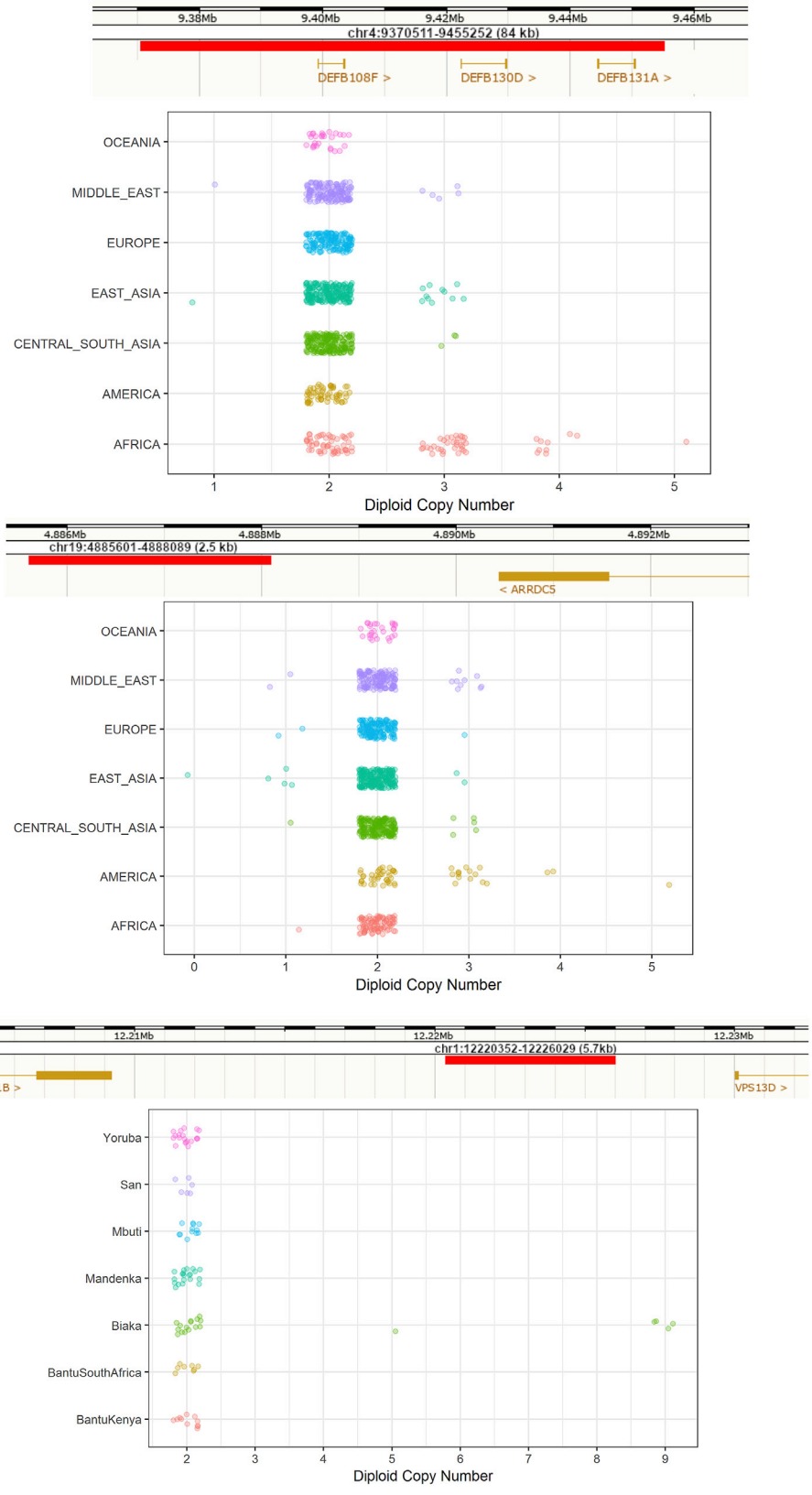


Figure S4. Population Stratification and Unreported Variants, Related to Figures 2A–2C and STAR Methods

Top: Population Stratification: Maximum allele frequency difference as a function of population differentiation. Blue line is loess fits after excluding populations with 10 samples or less. Deletions (Left), Insertions (Centre), Duplications (Right). Bottom: Variants not present in 1000G or SGDP. Continental (red) or Population (green) specific variants (n > 2) in the HGDP not found in 1000G or SGDP SV callsets binned by allele frequency. The same variant can be present in both distributions.



(legend on next page)

Figure S5. Additional Copy Number Expansions, Related to Figure 3

Red bar illustrates region expanded. Top: Expansions in beta-Defensin genes. Centre: Expansions downstream of *ARRDC5* prominent in Americans. Bottom: Expansion downstream *TNFRSF1B* private to Biaka.

Figure S6. Putatively Introgressed Variants, Related to Figure 2E and Table 1

Top: fiber-FISH of chr16 Oceanian-specific expansion shared with Denisovan genome at ~82% frequency in all three Oceanian populations. Cartoon illustration of location of original (16p12.2) and inserted site 7Mb away (16p11.2) into clone RP11-368N21 (green). Bottom: *MS4A1* deletion: IGV screenshot of a deletion in an exon of *MS4A1,* which encodes the B cell differentiation antigen CD20. The deletion is shared by both Neanderthals (Altai top, Vindija middle track) and American populations (reaches ~26% in Surui and Pima). The deletion is not present in the Denisovan genome (lower track). Bottom track shows Loupe screenshot of the region in HGDP01043 showing the two haplotypes resolved using 10x linked-reads, with one carrying the deletion.

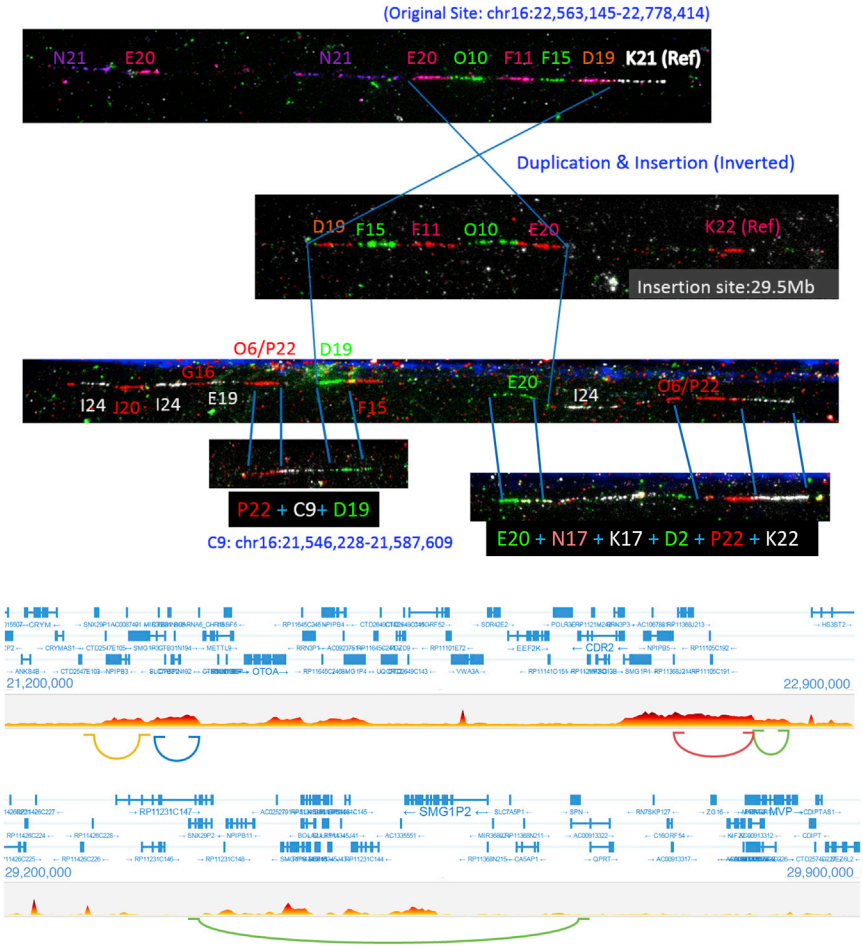


Figure S7. Chr16 Oceanian-Specific Expansion, Related to Figure 2E and Table 1

Top: Fiber-FISH illustrating the original site (top), the (inverted) insertion sites (center) and the region surrounding the insertion site (bottom). Region flanking the insertion site (C9) is a sequence 1Mb away from the original site, consistent with GenomeSTRiP calling a second duplication at this site in perfect LD with the initial duplication. Manta also identifies a Papuan-specific inversion at this locus. This suggests a complex event involving a duplication-inverted-insertion, an inversion and a deletion. Bottom: 10X-linked reads barcode overlap in region. Longranger also identifies a complex event at this locus. Top plot shows the original site barcode overlap and the regions of structural rearrangements, including the region of C9 (on the left). Bottom shows the insertion site. Note that this region is gene rich, and the candidate gene(s) under selection is not known and requires further study.