Population Structure, Stratification and Introgression of Human Structural Variation in the HGDP

Mohamed A. Almarri[1*], Anders Bergström[1,2], Javier Prado-Martinez[1], Alistair S. Dunham[1,3], Yuan Chen[1], Chris Tyler-Smith[1], Yali Xue[1*]

1. Wellcome Sanger Institute, Hinxton, CB10 1SA, UK
2. Francis Crick Institute, London, NW1 1AT, UK
3. EMBL-EBI, Hinxton, CB10 1SD, UK

*Correspondence: ma17@sanger.ac.uk (M.A.A.); ylx@sanger.ac.uk (Y.X).

Abstract

Structural variants contribute substantially to genetic diversity and are important evolutionarily and medically, yet are still understudied. Here, we present a comprehensive analysis of deletions, duplications, inversions and non-reference unique insertions in the Human Genome Diversity Project (HGDP-CEPH) panel, a high-coverage dataset of 910 samples from 54 diverse worldwide populations. We identify in total 61,801 structural variants, of which 61% are novel. Some reach high frequency and are private to continental groups or even individual populations, including a deletion in the maltase-glucoamylase gene *MGAM*, involved in starch digestion, in the South American Karitiana and a deletion in the Central African Mbuti in *SIGLEC5,* potentially increasing susceptibility to autoimmune diseases. We discover a dynamic range of copy number expansions and find cases of regionally-restricted runaway duplications, for example, 18 copies near the olfactory receptor *OR7D2* in East Asia and in the clinically-relevant *HCAR2* in Central Asia. We identify highly-stratified putatively introgressed variants from Neanderthals or Denisovans, some of which, like a deletion within *AQR* in Papuans, are almost fixed in individual populations. Finally, by *de novo* assembly of 25 genomes using linked-read sequencing we discover 1631 breakpoint-resolved unique insertions, in aggregate accounting for 1.9 Mb of sequence absent from the GRCh38 reference. These insertions show population structure and some reside in functional regions, illustrating 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 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, 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 & 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)-CEPH panel (Figure 1A), a dataset composed of 910 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 910 whole-genome sequences at an average depth of 36x and mapped reads to the GRCh38 reference (Bergström et al., 2019). As the dataset is generated from lymphoblastoid cell lines, we searched for potential cell-line artefacts by analysing coverage across the genome and excluded samples containing multiple aneuploidies, while masking regions which 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 result in proliferation once duplicated in culture. Nevertheless, these cell line artefacts can readily be recognised, and are excluded from the results below.

We identified 61,801 structural variants relative to the reference. We compared our dataset to published structural variation catalogues (Sudmant, et al. 2015a; Sudmant, et al. 2015b), and find that ~61% of the variants identified in our dataset are not present in the previous studies. Despite having a smaller sample size compared to the 1000 Genomes release (Sudmant, et al. 2015a), we discover a higher total number of variants across all different classes of SVs investigated. These novel 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 S5). The increased sensitivity reflects the higher coverage, longer reads and the large number of diverse populations in our study. Collectively, this illustrates that a substantial amount of global structural variation was previously undocumented, emphasizing the importance of studying underrepresented human populations.

Population Structure

Deletions show clear geographical clustering using principal component (PC) analysis (Figure S1A, S2A). A uniform manifold approximation and projection (UMAP) of the top 10 PCs shows clear separation of continental groups (Figure 1B). Noticeably, populations with known admixture such as the Hazara and Uygur form a separate cluster away from the Central & South Asian and East Asian clusters. Within each continental cluster, we observe examples of finer structure with samples from individual populations appearing closer to themselves relative to other populations (Figure S2C). The drifted Kalash population are clearly differentiated within the main Central & South Asian cluster, while the Mbuti, Biaka and San form their own clusters away from the rest of the African populations.

Duplications, inverted duplications and inversions show some degree of population structure, although less defined in comparison to deletions (Figures 1C and S3). 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.

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 (Figure 1D). 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-Sanchez et al., 2014). Both deletions and biallelic duplications show similar distributions, with deletions showing higher stratification. We do see some striking outliers, for example the Lowland/Sepik Papuans are almost fixed (86%) for a deletion in *HBA2*, which is absent in Papuan Highlanders. High frequencies of α-globin deletions in this region have been suggested to be protective against malaria (Yenchitsomanus et al., 1985, Flint et al., 1986). We also find a deletion within *MYO5B* that is particularly common (88%) in the Lahu, a population shown to have high numbers of private single nucleotide variants in addition to carrying rare Y-chromosome lineages (Bergström et al., 2019).

The large number of samples per population allowed us to investigate population-private variants (Figure S4A). We searched for functional effects of such variants and found a 14kb deletion in the South American Karitiana population at 40% frequency, which removes the 5' upstream region of *MGAM* up to the first exon. This gene encodes Maltase-glucoamylase, an enzyme highly expressed in the small intestine and involved in the digestion of dietary starches (Nichols et al., 2003). Interestingly, a recent ancient DNA study of South Americans has suggested that

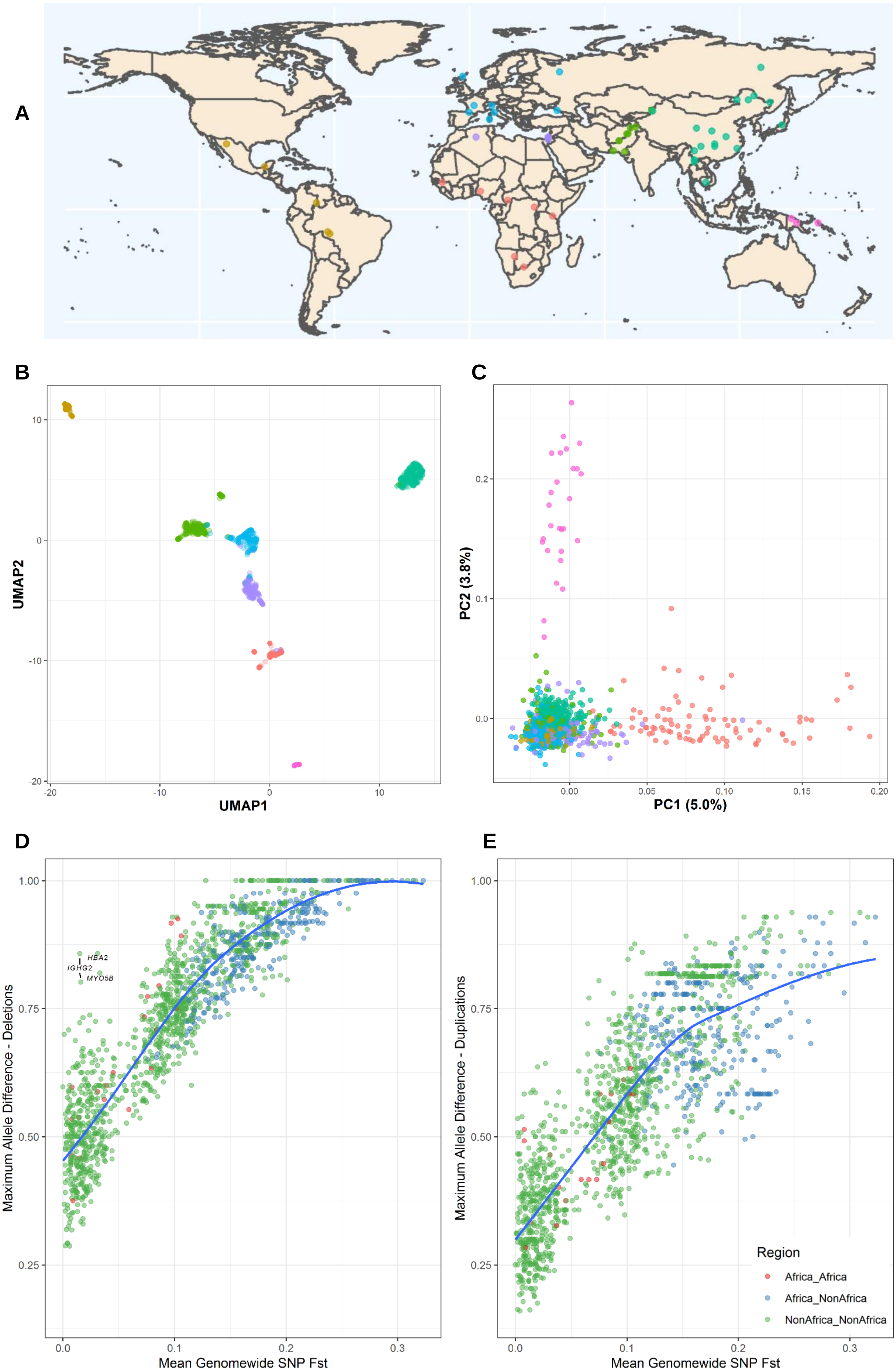
A
B
C
UMAP1
UMAP2
PC1 (5.0%)
PC2 (3.8%)
D
E
HBA2
IGHG2
MYO5B
Maximum Allele Difference - Deletions
Maximum Allele Difference - Duplications
Mean Genomewide SNP Fst
Mean Genomewide SNP Fst
Region
Africa_Africa
Africa_NonAfrica
NonAfrica_NonAfrica

Figure 1: Population structure and stratification of structural variants. **A:** The HGDP dataset, each point and colour represents a population and its regional label, respectively. Colours of regional groups are consistent throughout the study. **B:** UMAP of the top 10 principal components of biallelic deletions genotypes. See Figure S2B-C for more details. **C:** First two principal components of biallelic duplication genotypes. **D:** Maximum allele frequency difference as a function of population differentiation for 1431 pairwise population comparisons. Blue curve represents loess fits. Deletions (left) and Biallelic Duplications (right). Three outlying stratified variants are illustrated. *HBA2* deletion in Papuan lowlands, a deletion within *MYO5B* particularly common in Lahu, and a deletion downstream of *IGHG2* almost fixed in Dai (86% frequency).

selection acted on this gene in ancient Andean individuals, possibly as a result of their transition to agriculture (Lindo et al., 2018). However, 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.

We discovered a deletion that is private and at 54% frequency in the Central African Mbuti hunter-gatherer population that deletes *SIGLEC5* without removing its adjacent paired receptor *SIGLEC14*. 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 which is 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 evolved to fine-tune immune responses; and the loss of an inhibitory receptor is hypothesized to result in immune hyperactivity and autoimmune disease (Lübbers et al., 2018).

Archaic Introgression

We genotyped our calls in the high coverage Neanderthal and Denisovan archaic genomes (Meyer et al., 2012; Prufer et al., 2017; Prufer et al., 2014), and find 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, potentially resulting from adaptive introgression (Table 1).

Position	Variant	EUR	CSA	EA	ME	AMR	OCE	Gene	Neanderthal	Denisova
chr1:64992622-64993000	DEL	0	0	0	0	0	0.32	*JAK1*	REF	DEL
chr2:3684113-3690212	DEL	0.02	0.003	0.05	0.03	0	0.26	*ALLC*	DEL $_{Vin}$	REF
chr8:23124835-23130567	DEL	0	0.02	0.002	0	0	0.36	*TNFRSF10D*	DEL	REF
chr8:23134649-23164796	DUP	0	0	0	0	0	0.48	*TNFRSF10D*	DUP	DUP
chr11:60460681-60461880	DEL	0	0	0.02	0	0.17	0	*MS4A1*	DEL	REF
chr12:101882163-101883377	DEL	0.02	0.08	0.32	0.007	0.01	0.33	*DRAM1*	DEL	REF
chr12:104799951-104803150	DUP	0.003	0.009	0	0.01	0	0.33	*SLC41A2*	DUP	REF
chr15:34920811-34925992	DEL	0	0	0	0	0	0.63	*AQR*	REF	DEL
chr17:3038851-3041981	DEL	0	0	0	0	0	0.16	*RAP1GAP2*	DEL	DEL
chr19:42529806-42531042	DEL	0	0	0	0	0	0.54	*CEACAM1*	DEL	DEL

Table 1: Allele frequencies of regionally stratified variants shared with high coverage archaic genomes but not found in African populations. Neanderthal refers to both published high coverage genomes. Variants lie within or near the genes listed. 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 LD in Oceanians (r^2 = 0.48). EA - East Asia, ME - Middle East, AMR - America, CSA - Central South Asia, OCE - Oceania.

We replicated the putatively Denisovan introgressed duplications at chromosome 16p12 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 the Bouganville 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 3 Oceanian populations (~82%). These duplications form the most extreme regional-specific variants (Figure S4B), and their unusual allele distribution suggests that they may have remained at high frequencies after archaic introgression due to positive selection.

We discovered multiple Oceanian-specific variants shared with archaic genomes. A deletion within *AQR*, an RNA helicase gene, is present at 63% frequency and shared only with the Altai Denisovan. The highest expression of this gene is in EBV-transformed lymphocytes (GTEx Consortium, 2013). RNA helicases play an important role in the detection of viral RNAs and mediating the antiviral immune response, in addition to being necessary host factors for viral replication (Ranji & Boris-Lawrie, 2010). AQR has been reported to be involved in the recognition and silencing of transposable elements (Akay et al., 2017), and is known to regulate HIV-1 DNA integration (Konig et al., 2008). Another Denisovan shared deletion is in *JAK1*, a kinase important in cytokine signalling. We additionally find an intriguing Neanderthal-shared deletion-duplication combination at *TNFRSF10D,* a 5.7 kb deletion upstream, and a 30kb duplication that encompasses the whole gene, that is common in Oceanians but rare globally.

In the Americas we identify a deletion, shared only with Neanderthals, that reaches ~26% frequency in both the Surui and Pima. This variant removes an exon in *MS4A1,* 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.

Both Neanderthals and Denisovans thus appear to have contributed potentially functional structural variants to different modern human populations. As many of the identified variants are involved in immune processes (Table 1), it is tempting to speculate that they are associated with adaptation to pathogens after modern humans expanded into new environments outside of Africa.

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

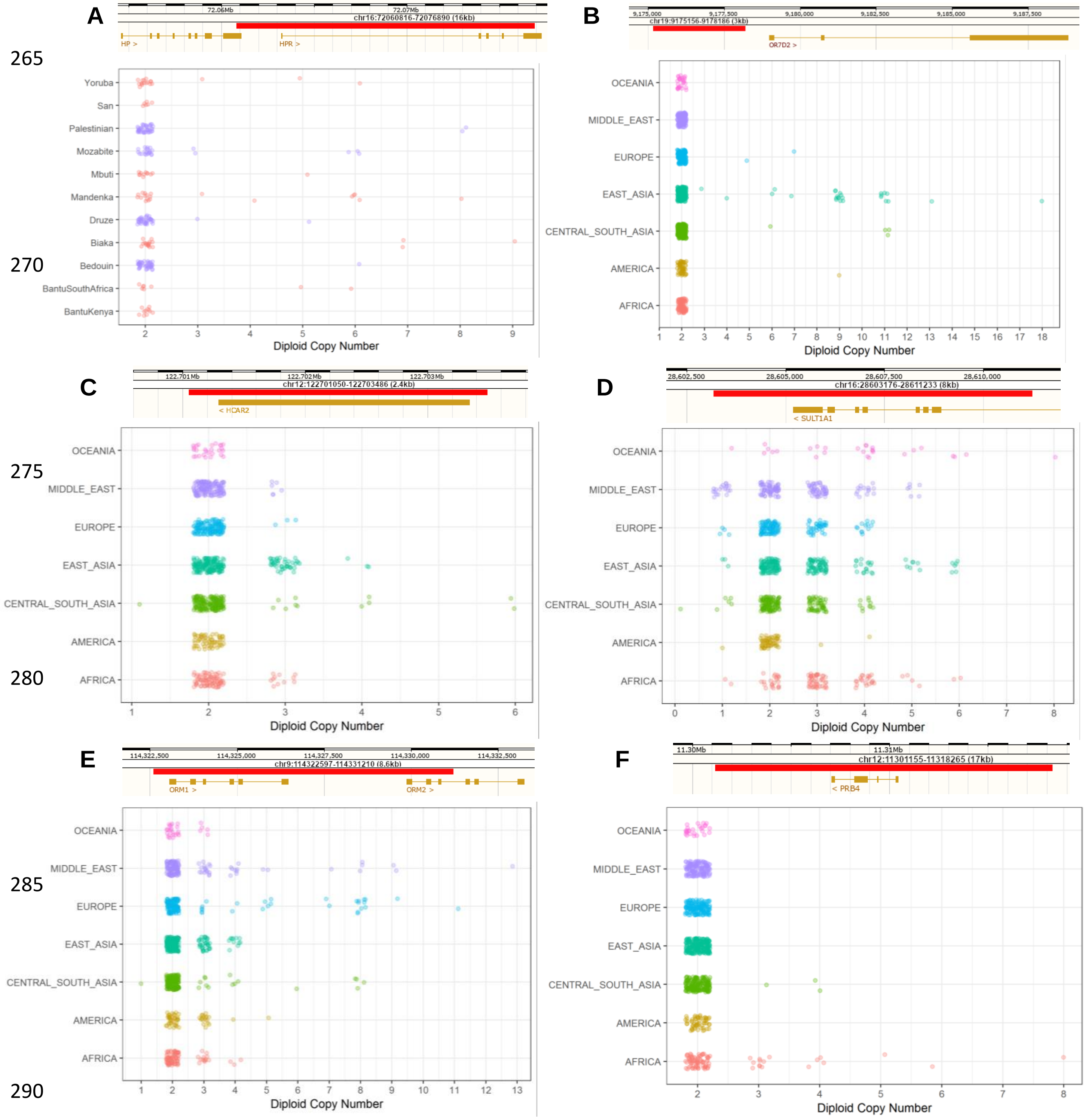

Figure 2: Copy Number Expansions and Runaway Duplications. Red bar illustrates the location of the expansion. Additional examples are shown in Figure S7. **A:** Expansion in *HPR* in Africans and Middle Eastern samples. **B:** Expansions upstream *OR7D2* that is mostly restricted to East Asia. The observed expansions in Central & South Asian samples are all in Hazara samples, an admixed population. **C:** Expansions within *HCAR2* which are particularly common in the Kalash population. **D:** Expansions in *SULT1A1* which are pronounced in Oceanians (median copy number, 4; all other non-African continental groups, 2; Africa ,3). **E:** Expansions in *ORM1/ORM2*. This expansion was reported previously in Europeans (Handsaker et al., 2015), however we find it in all regional groups and particularly in Middle Eastern populations. **F:** Expansions in *PRB4* which are restricted to Africa and in Central & South Asian samples with significant African admixture (Makrani and Sindhi).

globally, but have expanded to high copy numbers in certain populations, possibly in response to regionally-restricted selection events (Figure 2).

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 S7). We replicated the previously identified *HPR* expansions (Figure 2A), and find 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 infections (Smith et al., 1995). We find 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 2B), 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 due to non-allelic homologous recombination (Handsaker et al., 2015).

We discovered expansions in *HCAR2* in Asians which are especially prominent in the Kalash group (Figure 2C), 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 due to its key role in mediating anti-inflammatory effects in multiple tissues and diseases (Offermanns 2017). Another clinically-relevant expansion is in *SULT1A1* (Figure 2D), which encodes a sulfotransferase involved in the 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 ~50x and generated *de novo* assemblies using the Supernova assembler (Weisenfeld et al., 2017) (Table S1). By comparing our assemblies to the GRCh38 reference, we identified 1631 breakpoint-resolved unique, non-repetitive insertions across all chromosomes which in aggregate account for 1.9Mb of sequences missing from the reference (Figure 3A). A San individual contained the largest number of insertions, consistent with their high divergence from other populations. However, we note that the number of identified insertions is correlated with the assembly size and quality (Figure S8), suggesting there are still additional insertions to be discovered.

We find that the majority of insertions are relatively small, with a median length of 513bp (Figure 3B). They are of potential functional consequence as they fall within or near 549 protein coding genes, including 10 appearing to reside in exons (Supplementary Methods). These genes are involved in diverse cellular processes, including immunity (*NCF4*), regulation of glucose (*FGF21*), and a potential tumour 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 the reference genome may harbour rare deletion alleles at these sites. These variants show population structure, with Central Africans and Oceanians showing most differentiation (Figure 3C), reflecting the deep divergences within Africa and the effect of drift, isolation and possibly Denisovan introgression in Oceania.

While the number of *de novo* assembled genomes using linked or long reads is increasing, they are mostly representative of urban populations. Here, we present a resource containing a diverse set of assemblies with no access or analysis restrictions.

Discussion

In this study we present a comprehensive catalogue 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. The relatively large number of high-coverage genomes in each population allowed us to identify and estimate the frequency of population-specific variants, providing insights into potentially geographically-localized selection events, although further functional work is needed to elucidate their effect, if any. Our finding of common clinically-relevant regionally private variants, some of which appears to be introgressed from archaic hominins, argues for further efforts generating genomes without data restrictions from under-

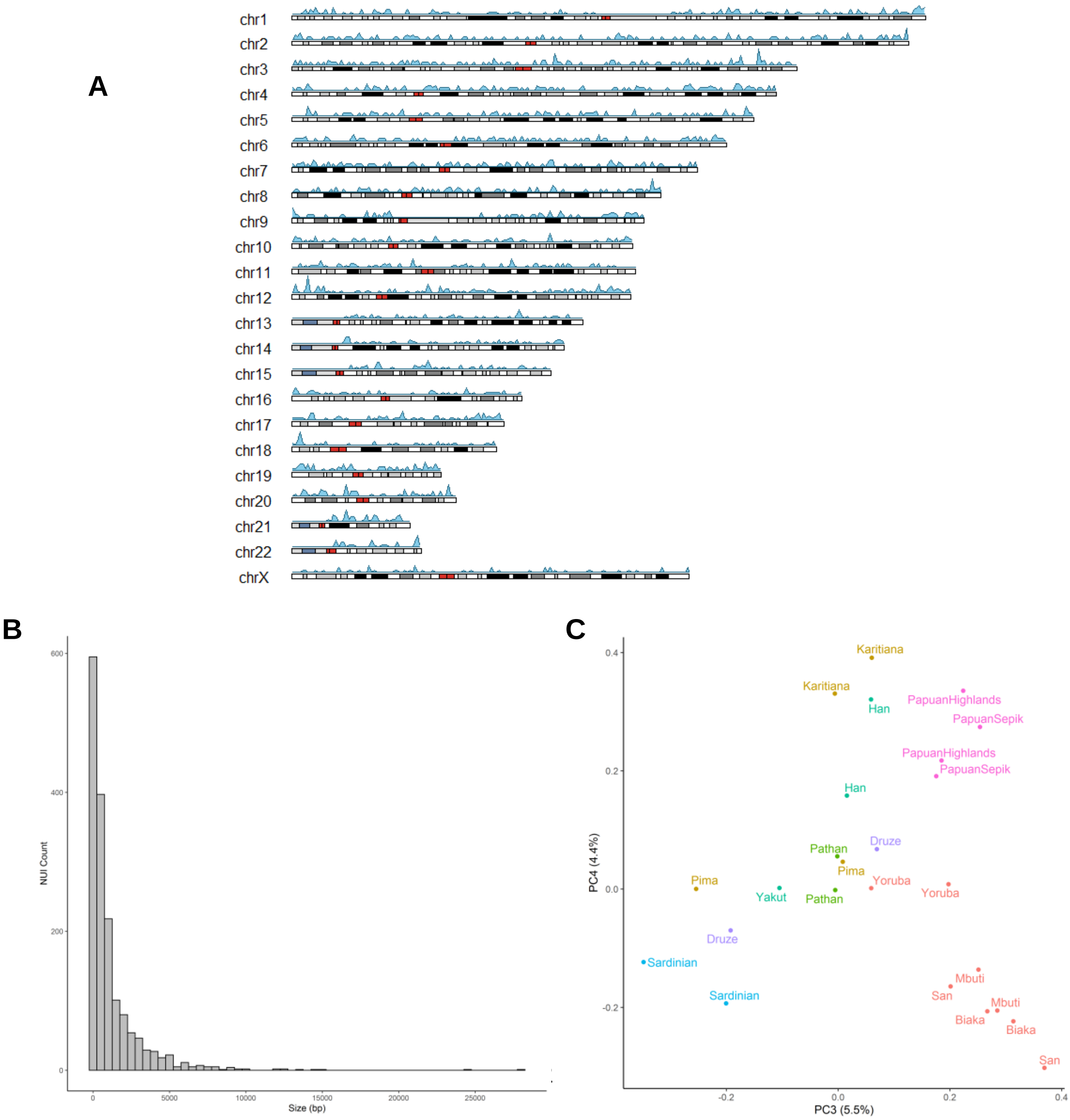

Figure 3: Non-Reference Unique Insertions (NUIs). **A:** Ideogram illustrating the density of identified NUI locations across different chromosomes using a window size of 1 Mb. Colours on chromosomes reflect chromosomal bands with red for centromeres. **B**: Size distribution of NUIs using a bin size of 500bp. **C**: PCA of NUI genotypes showing population structure (PC3-4). Previous PCs potentially reflect variation in size and quality of the assemblies.

represented populations. We note that despite the diversity found in the HGDP panel, considerable geographic gaps remain in Africa, the Americas and Australasia.

The use of short-reads in this study restricts the discovery of complex structural variants, demonstrated by recent reports which 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 accurately representing 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 the 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.

Data availability

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. Structural variant calls, Supernova *de novo* assemblies and NUI fastas will be available on ftp://ngs.sanger.ac.uk/production/hgdp.

Acknowledgments

We thank Richard Durbin, Hélène Blanché, Thomaz Pinotti and members of the Tyler-Smith group for advice and discussions. We thank Robert Handsaker and Arda Söylev for technical advice on the structural variant discovery algorithms. We would like to thank all the individuals who donated samples for this study and the CEPH

Biobank at Fondation Jean Dausset-CEPH for the maintenance and distribution of the HGDP samples. 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.

References

1. Akay, A. et al. "The helicase aquarius/EMB-4 is required to overcome intronic barriers to allow nuclear RNAi pathways to heritably silence transcription." *Developmental cell* 42.3 (2017): 241-255.
2. Akkaya, M. and Barclay, A.N. "How do pathogens drive the evolution of paired receptors?." *European journal of immunology* 43.2 (2013): 303-313.
3. Angata, T. et al. "Discovery of Siglec-14, a novel sialic acid receptor undergoing concerted evolution with Siglec-5 in primates." *The FASEB Journal* 20.12 (2006): 1964-1973.
4. Audano, P.A. et al. "Characterizing the major structural variant alleles of the human genome." *Cell* 176.3 (2019): 663-675.
5. Bergström, A. et al. "Insights into human genetic variation and population history from 929 diverse genomes." *bioRxiv* (2019): 674986.
6. Cann, H.M. et al. "A human genome diversity cell line panel." *Science* 296.5566 (2002): 261-262.
7. Chaisson, M.J.P. et al. "Multi-platform discovery of haplotype-resolved structural variation in human genomes." *Nature communications* 10 (2019).
8. Coop, G. et al. "The role of geography in human adaptation." *PLoS genetics* 5.6 (2009): e1000500.
9. Flint, J. et al. "High frequencies of α-thalassaemia are the result of natural selection by malaria." *Nature* 321.6072 (1986): 744.
10. Franco, J.R. et al. "Epidemiology of human African trypanosomiasis." *Clinical epidemiology* 6 (2014): 257.
11. Garrison, E. et al. "Variation graph toolkit improves read mapping by representing genetic variation in the reference." *Nature biotechnology* (2018).
12. Handsaker, R.E. et al. "Large multiallelic copy number variations in humans." *Nature genetics* 47.3 (2015): 296.
13. Hebbring, S.J. et al. "Sulfotransferase gene copy number variation: pharmacogenetics and function." *Cytogenetic and genome research* 123.1-4 (2008): 205-210.
14. Huddleston, J. and Eichler E.E. "An incomplete understanding of human genetic variation." *Genetics* 202.4 (2016): 1251-1254.
15. Huerta-Sánchez, E. et al. "Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA." *Nature* 512.7513 (2014): 194.
16. König, R. et al. "Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication." *Cell* 135.1 (2008): 49-60.
17. Kuijpers, T.W. et al. "CD20 deficiency in humans results in impaired T cell–independent antibody responses." *The Journal of clinical investigation* 120.1 (2010): 214-222.
18. Lindo, J. et al. "The genetic prehistory of the Andean highlands 7000 years BP though European contact." *Science advances* 4.11 (2018): eaau4921.

19. Lonsdale, J. et al. "The genotype-tissue expression (GTEx) project." *Nature genetics* 45.6 (2013): 580.
20. Lübbers, J. et al. "Modulation of immune tolerance via Siglec-sialic acid interactions." *Frontiers in immunology* 9 (2018).
21. Marshall, M.J.E. et al. "Therapeutic antibodies: what have we learnt from targeting CD20 and where are we going?." *Frontiers in immunology* 8 (2017): 1245.
22. McLaren, W. et al. "The ensembl variant effect predictor." *Genome biology* 17.1 (2016): 122.
23. Meyer, M. et al. "A high-coverage genome sequence from an archaic Denisovan individual." *Science* 338.6104 (2012): 222-226.
24. Nichols, B.L. et al. "The maltase-glucoamylase gene: common ancestry to sucrase-isomaltase with complementary starch digestion activities." *Proceedings of the National Academy of Sciences* 100.3 (2003): 1432-1437.
25. Offermanns, S. "Hydroxy-carboxylic acid receptor actions in metabolism." *Trends in Endocrinology & Metabolism* 28.3 (2017): 227-236.
26. Prüfer, K. et al. "A high-coverage Neandertal genome from Vindija Cave in Croatia." *Science* 358.6363 (2017): 655-658.
27. Prüfer, K. et al. "The complete genome sequence of a Neanderthal from the Altai Mountains." *Nature* 505.7481 (2014): 43.
28. Ranji, A. and Boris-Lawrie, K. "RNA helicases: emerging roles in viral replication and the host innate response." *RNA biology* 7.6 (2010): 775-787.
29. Redon, R. et al. "Global variation in copy number in the human genome." *Nature* 444.7118 (2006): 444.
30. Schneider, V.A. et al. "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly." *Genome research* 27.5 (2017): 849-864.
31. Sherman, R.M. et al. "Assembly of a pan-genome from deep sequencing of 910 humans of African descent." *Nature genetics* 51.1 (2019): 30.
32. Sirugo, G. et al. "The missing diversity in human genetic studies." *Cell* 177.1 (2019): 26-31.
33. Smith, A.B. et al. "Killing of trypanosomes by the human haptoglobin-related protein." *Science* 268.5208 (1995): 284-286.
34. Sudmant, P.H. et al. "An integrated map of structural variation in 2,504 human genomes." *Nature* 526.7571 (2015a): 75.
35. Sudmant, P.H. et al. "Global diversity, population stratification, and selection of human copy-number variation." *Science* 349.6253 (2015b): aab3761.
36. Weisenfeld, N.I. et al. "Direct determination of diploid genome sequences." *Genome research* 27.5 (2017): 757-767.
37. Wong, K.H.Y. et al. "De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations." *Nature communications* 9.1 (2018): 3040.
38. Yenchitsomanus, P.T. et al. "Extremely high frequencies of alpha-globin gene deletion in Madang and on Kar Kar Island, Papua New Guinea." *American journal of human genetics*37.4 (1985): 778.