Global diversity, population stratification, and selection of human copy number variation

Peter H. Sudmant[1], Swapan Mallick[2,3], Bradley J. Nelson[1], Fereydoun Hormozdiari[1], Niklas Krumm[1], John Huddleston[1,39], Bradley P. Coe[1], Carl Baker[1], Susanne Nordenfelt[2,3], Michael Bamshad[4], Lynn B. Jorde[5], Olga L. Posukh[6,7], Hovhannes Sahakyan[8,9], W. Scott Watkins[10], Levon Yepiskoposyan[9], M. Syafiq Abdullah[11], Claudio M. Bravi[12], Cristian Capelli[13], Tor Hervig[14], Joseph TS Wee[15], Chris Tyler-Smith[16], George van Driem[17], Irene Gallego Romero[18], Aashish R. Jha[18], Sena Karachanak-Yankova[19], Draga Toncheva[19], David Comas[20], Brenna Henn[21], Toomas Kivisild[22], Andres Ruiz-Linares[23], Antti Sajantila[24], Ene Metspalu[8,25], Jüri Parik[8], Richard Villems[8], Elena B. Starikovskaya[26], George Ayodo[27], Cynthia M. Beall[28], Anna Di Rienzo[18], Michael Hammer[29], Rita Khusainova[30,31], Elza Khusnutdinova[30,31], William Klitz[32], Cheryl Winkler[33], Damian Labuda[34], Mait Metspalu[8], Sarah A. Tishkoff[35], Stanislav Dryomov[26,36], Rem Sukernik[26,37], Nick Patterson[2,3], David Reich[2,3,38], and Evan E. Eichler[1,39]

1. Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
2. Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
3. Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
4. Department of Pediatrics, University of Washington, Seattle, WA 98119, USA
5. Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84112, USA
6. Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, Novosibirsk, 630090, Russia
7. Novosibirsk State University, Novosibirsk, 630090, Russia
8. Estonian Biocentre, Evolutionary Biology group, Tartu, 51010, Estonia
9. Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences of Armenia, Yerevan, 0014, Armenia
10. Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA
11. RIPAS Hospital, Bandar Seri Begawan, Brunei Darussalam
12. Laboratorio de Genética Molecular Poblacional, Instituto Multidisciplinario de Biología Celular (IMBICE), CCT-CONICET & CICPBA, La Plata, B1906APO, Argentina
13. Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK
14. Department of Clinical Science, University of Bergen, Bergen, 5021, Norway
15. National Cancer Centre Singapore, Singapore
16. The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambs. CB10 1SA, UK
17. Institute of Linguistics, University of Bern, Bern, CH-3012, Switzerland
18. Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA
19. Department of Medical Genetics, National Human Genome Center, Medical University Sofia, Sofia, 1431, Bulgaria
20. Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, 08003, Spain
21. Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794, USA
22. Division of Biological Anthropology, University of Cambridge, Fitzwilliam Street, Cambridge, CB2 1QH, UK

23. Department of Genetics, Evolution and Environment, University College London, WC1E 6BT, UK
24. University of Helsinki, Department of Forensic Medicine, Helsinki, 00014, Finland
25. University of Tartu, Department of Evolutionary Biology, Tartu 5101, Estonia
26. Laboratory of Human Molecular Genetics, Institute of Molecular and Cellular Biology, Siberian Branch of Russian Academy of Sciences, Novosibirsk, 630090, Russia
27. Center for Global Health and Child Development, Kisumu, 40100, Kenya
28. Department of Anthropology, Case Western Reserve University, Cleveland, OH 44106-7125, USA
29. ARL Division of Biotechnology, University of Arizona, Tucson, AZ 85721, USA
30. Institute of Biochemistry and Genetics, Ufa Research Centre, Russian Academy of Sciences, Ufa, 450054, Russia
31. Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa, 450074, Russia
32. Integrative Biology, University of California, Berkeley, CA 94720-3140, USA
33. Basic Research Laboratory, Center for Cancer Research, NCI, Leidos Biomedical Research, Inc., Frederick National Laboratory, Frederick, MD 21702, USA
34. CHU Sainte-Justine, Pediatrics Departement, Université de Montréal, QC, H3T 1C5, Canada
35. Department of Biology and Genetics. University of Pennsylvania, Philadelphia, PA 19104, USA
36. Department of Paleolithic Archaeology, Institute of Archaeology and Ethnography, Siberian Branch of Russian Academy of Sciences, Novosibirsk, 630090, Russia
37. Altai State University, Barnaul, 656000, Russia
38. Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA
39. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
Correspondence to: Evan E. Eichler, Department of Genome Sciences, University of Washington School of Medicine, Foege S413A, Box 355065, 3720 15th Ave NE, Seattle, WA 98195-5065 E-mail: eee@gs.washington.edu

Structured Abstract

Introduction: Most studies of human genetic variation have focused on single nucleotide variants (SNVs). Copy number variants (CNVs), however, affect more base pairs of DNA among humans and yet our understanding of CNV diversity among human populations is limited.

Rationale: We aimed to understand the pattern, selection and diversity of copy number variation by analyzing deeply sequenced genomes representing the diversity of all humans. We compared the selective constraints of deletions versus duplications to understand population stratification in the context of the ancestral human genome and to assess differences in CNV load between African and non- African populations.

Results: We sequenced 236 individual genomes from 125 distinct human populations and identified 14,467 autosomal CNVs and 545 X-linked CNVs with a sequence read-depth approach. Deletions exhibit stronger selective pressure and are better phylogenetic markers of population relationships than duplication polymorphisms. We identify 1,036 population-stratified copy number variable regions, 295 of which intersect coding regions and 199 of which exhibit signatures of differentiation. Strikingly, duplicated loci were 1.8-fold more likely to be stratified than deletions but were poorly correlated with flanking genetic diversity. Among these, we highlight a duplication polymorphism restricted to modern Oceanic populations yet also present in the genome of the archaic Denisova hominin. This 225 kbp duplication includes two microRNA genes and is almost fixed among human Papuan–Bougainville genomes.

The data allow us to reconstruct the ancestral human genome and create a more accurate evolutionary framework for the gain and loss of sequence during human evolution. We identified 571 loci that segregate in the human population and another 2,026 loci of fixed copy 2 in all human genomes but absent from the reference genome. The total deletion and duplication load between African and non-African population groups showed no difference after we account for ancestral sequences missing from the human reference. However, we did observe that the relative number of base pairs affected by CNVs compared to SNPs is higher among non-Africans than Africans.

Conclusion: Deletions, duplications and CNVs have shaped, to different extents, the genetic diversity of human populations by the combined forces of mutation, selection and demography.

Figure Legend: Counterclockwise from the top: The geographic coordinates of populations sampled are indicated on a world map (colored dots). The pie charts show the continental population allele frequency of a single ~225 kbp duplication polymorphism found exclusively among Oceanic populations and an archaic Denisova. The ancestral structure of this duplication locus (1) and the Denisova duplication structure (2) are shown in relation to their position on chromosome 16. We estimate that the duplication emerged ~440 thousand years ago (kya) in the Denisova and then introgressed into ancestral Papuan populations ~40 kya.

Abstract

In order to explore the diversity and selective signatures of duplication and deletion human copy number variants (CNVs), we sequenced 236 individuals from 125 distinct human populations. We observed that duplications exhibit fundamentally different population genetic and selective signatures than deletions and are more likely to be stratified between human populations. Through reconstruction of the ancestral human genome, we identify megabases of DNA lost in different human lineages and pinpoint large duplications that introgressed from the extinct Denisova lineage now found at high frequency exclusively in Oceanic populations. We find that the proportion of CNV base pairs to single nucleotide variant base pairs is greater among non-Africans than it is among African populations but we conclude that this difference is likely due to unique aspects of non-African population history as opposed to differences in CNV load.

Introduction

In the past decade, genome sequencing has provided insights into demography and migration patterns of human populations (*1-4*), ancient DNA (*5-7*), *de novo* mutation rates (*8-10*), and the relative deleteriousness and frequency of coding mutations (*11*, *12*). Global human diversity, however, has only been partially sampled and the genetic architecture of many populations remains uncharacterized. To date, the majority of human diversity studies have focused on single nucleotide variants (SNVs) although copy number variants (CNVs) have contributed significantly to hominid evolution (*13*, *14*), adaptation and disease (*15-18*). Much of the research into CNV diversity has been performed with SNP microarray and array comparative genomic hybridization (aCGH) platforms (*19-22*), which provide limited resolution. In addition, comparisons of population CNV diversity with heterogeneous discovery platforms may lead to spurious population-specific trends in CNV diversity (*22*, *23*). Although there are many other forms of structural variation (e.g., inversions or mobile element insertions) in this study, we focused on understanding the population genetics and normal pattern of copy number variation by deep sequencing a diverse panel of human genomes.

Results

CNV discovery: We sequenced to high coverage a panel of 236 human genomes representing 125 diverse human populations from across the globe (**Fig. 1, Table S2**). Sequencing was performed to a mean genome coverage of 41-fold from libraries prepared using a standard PCR-free protocol on the HiSeq 2000 Illumina sequencing platform (*24*). The panel includes representation from a broad swathe of human diversity, including individuals from across Siberia, the Indian subcontinent, and Oceania. We also analyzed the high-coverage archaic Neanderthal (*25*) and Denisova (*26*) as well as three ancient human genomes to refine the evolutionary origin and timing of CNV differences (*24*). We applied a read-depth-based digital comparative genomic hybridization (dCGH) approach (*13*, *24*) to discover 14,467 autosomal CNVs and 545 X-linked CNVs among individuals relative to the reference genome (**Table 1, Table S1**), which we estimate provides breakpoint resolution to ~210 bp (*24*). CNV calls were validated with SNP microarrays and a custom aCGH microarray that targeted all CNVs identified in 20 randomly selected individuals (*24*).

The median CNV size was 7,396 bp with 82.2% of events (n=12,338) less than 25 kbp (*24*). CNVs mapping to segmental duplications were larger on average (median of 14.4 kbp), than CNVs mapping to the unique portions of the genome (median of 6.2 kbp). Almost one-half of CNV base pairs mapped within previously annotated segmental duplications (a 10-fold enrichment) (**Table 1**). In total, 217.1 Mbp (7.01%) of the human genome is variable due to CNVs in contrast to 33.8 Mbp (1.1%) due to single-nucleotide variation (**Table 1**). Deletions (loss of sequence) were less common (representing 85.6 Mbp or 2.77% of the genome) compared to duplications (gain of sequence, 136.1 Mbp or 4.4% of the genome). Furthermore, comparing

our dataset with other studies of CNVs (*21*, *27*) 67-73% of calls we report are unique to our study while we capture 68-77% of previously identified CNVs (*24*).

CNV diversity and selection: African populations are broadly distinguished from non-African populations by a principal component analysis (PCA) for either deletions (**Figs. 2a, S20,** (*24*)) or duplications (**Fig. 2b**). In this analysis, we limited the variants to bi-allelic deletions or bi-allelic duplications (diploid genotypes of 2, 3 or 4) to eliminate difficulty of inferring phase from multicopy CNVs. For deletions, PC1 (6.8% of the variance) and PC2 (3.94%) distinguish Africans, West Eurasians, East Asians and Oceanic populations. PC3 and PC4, describing 2.8% and 2.0% of the total variance, cluster Papuans and populations of the Americas, respectively. Many other populations were predictably distributed along clines between these clusters (e.g., Northern Africans, Siberian, South Asian, Amerindian and indigenous peoples of Philippines and North Borneo). PCAs generated from SNVs showed similar patterns as those from deletions. Africans also show much greater heterozygosity (**Fig. 2c**, **Table 2)**, for instance, ~25% more heterozygous bi-allelic deletions and more than a twofold difference when compared to Amerindians ($\theta_{African}$=535 vs $\theta_{Americas}$=209). The archaic Neanderthal and Denisova genomes form an out-group to all humans (*24*).

Duplication heterozygosity and PCA in general show similar trends (**Fig. 2d**); albeit with far less definition. Interestingly, Oceanic populations, especially those from Papua New Guinea, Australia, and Bougainville showed the greatest separation on PC1 by duplication. Bi-allelic duplications appear somewhat less informative markers of human ancestry in contrast to SNVs, which provide the greatest resolution (e.g., SNV PCs 1-4 describe 5.8, 3.4, 2.6 and 1.7% of the variance, respectively). This difference is also seen when comparing SNV and CNV heterozygosity (**Fig. 2e,f**). While heterozygous bi-allelic deletions were strongly correlated (R=0.88) with SNV heterozygosity, the correlation between SNVs and duplications was much weaker (R=0.27). We compared this correlation for duplications located adjacent to segmental duplications (within or proximal 150 kbp) in contrast to those occurring in unique regions of the genome and therefore less likely to be subject to recurrent mutation. Heterozygous duplications occurring in unique regions were better correlated with heterozygous SNVs (r=0.29) than those adjacent or within segmental duplications (r=0.17), though the difference was not significant (two-sided Williams' test P<0.1).

Studies of larger (>100 kbp) deletion and duplication events indicate that deletions are more deleterious than duplications (*28*). We reasoned that this may be reflected in the allele frequency spectrum (AFS) of normal genetic variation and compared the AFS of genic versus intergenic deletions and duplications for smaller events (**Fig. 3a,b**). Genic deletions were significantly rarer than intergenic deletions (Wilcoxon rank sum test, P=1.84e-9), but genic duplications showed no such skew (Wilcoxon rank sum test, P=0.181). Size also had a significant impact on the AFS of CNVs. Deletions increased in rarity as a function of size (F-test, P=5.02e-11, **Fig. 3c**), but only a

nominally significant trend was observed for duplications (P=0.031, **Fig. 3d**). These data suggest that selection has shaped the extant diversity of deletions and duplications differently during human evolution.

Population stratification: As population stratification can be indicative of loci under adaptive selection, we calculated V_{st} statistics for each CNV among all pairs of continental population groups, a metric analogous to F_{st} (the fixation index) (*29*). V_{st} and F_{st} statistics compare the variance in allele frequencies between populations with V_{st} allowing comparison of multi-allelic or multicopy CNVs. We identified 1,036 stratified copy number variable loci (CNVRs with maximum population V_{st}>0.2, ~10% of the total), 295 of which intersected the exons of genes and 199 that exhibited extreme stratification (V_{st}>0.5, **Table S3**). After correcting for copy number, duplicated loci were 1.8-fold more likely to be stratified than deletions. This finding is more remarkable in light of the fact that duplications were less discriminatory by PCA suggesting that a subset of multi-allelic duplicated CNVs show large allele frequency differences between different populations (see discussion below). The V_{st} of stratified duplicated CNVs was weakly correlated with the F_{st} of flanking SNVs (R^2=0.03, P=3.27e-12) in contrast to deletions (R^2=0.2, p<2e-16). Stratified duplication loci, thus, are far less likely to be tagged by adjacent SNPs through linkage disequilibrium.

Many of the population-differentiated loci were multi-allelic and mapped to segmental duplications including the repeat domain of *ANKRD36*, and the DUF1220 domain of *NBPF* (*24*) (**Table 3**). Several of these population differences involve genes of medical consequence, such as the multi-allelic duplication of *CLPS*, a pancreatic colipase involved in dietary metabolism of long chain triglyceride fatty acids (**Fig. 4a**). Increased expression in mouse models of this gene is negatively correlated with blood glucose levels (*30*). A duplication of the haptoglobin and haptoglobin-related (*HP and HPR*) genes expanded exclusively in Africa. The duplication has recently been associated with a possible protective effect against trypanosomiasis in Africa, though only copy 3 and 4 alleles were reported (*31*). We find this locus has further expanded to five and six copies in Esan, Gambian, Igbo, Mandenka and Yoruban individuals (**Fig. 4a)**. We also compared the location of our CNVs with disease loci identified by GWAS (*32*) and sites of potential positive selection (*33*). Although only a small fraction of our CNVs (1-6%) overlapped such functional annotation, we note that 21% of putative adaptive loci intersected with a CNV when compared to 6% of disease GWAS loci (**Table S4**). Because many of the intervals are large, further refinement and investigation are needed to determine the significance of such overlaps.

Denisovan CNVs are retained and expanded in Oceanic populations: We further searched for highly stratified population-specific CNVs sharing alleles with the archaic Neanderthal and Denisovan individuals assessed in our study. While no Neanderthal-shared population-specific CNVs were identified, five Oceanic-specific CNVs were identified that shared the Denisova allele at high frequency (*24*). Papuan genomes have previously been reported to harbor 3-6%

Denisovan admixture (*6*, *26*). CNVs of putative Denisovan ancestry were at remarkably high frequency in Papuan individuals (all >0.2 allele frequency), with one ~9 kbp deletion lying 2 kbp upstream of the long noncoding RNA *LINC00501*, another 5 kbp duplication lying 8 kbp upstream of the *METTL9* methyltransferase gene, and a 73.5 kbp duplication intersecting the *MIR548D2* and *MIR548AA2* microRNAs (**Fig. 4b**).

We determined that the latter two are part of a larger composite segmental duplication that appears to have almost fixed among human Papuan–Bougainville genomes (AF=0.84) but has not been observed in any other extant human population (**Fig. 4b,c**). We noted three additional duplications proximal to this locus exhibiting strikingly correlated copy number, despite being separated by >1 Mbp in the reference genome (**Fig. 4c,** (*24*)). We suggest that these constitute a single, larger (~225 kbp) complex duplication composed of different segmental duplications. Using discordantly mapping paired end reads, we resolved the organization of two duplication architectures not represented in the human reference (**Fig. 4d**). The first of which (architecture A/C) is present in all individuals assessed in this study (5,625 discordant paired-end reads supporting) but not in the human reference genome. The second (B/D) corresponds to the Denisova–Papuan-specific duplication and is only present in these individuals and the Denisova genome. 70 paralogous sequence variants (markers distinct to paralogous locus (*34*, *35*)) distinguish the Papuan duplication of which 65/70 (92.9%) were shared with the archaic Denisova genome. On the basis of single-nucleotide divergence we estimate that the duplication emerged ~440 kya and rose to high frequency in Papuan (>0.80 AF) but not Australian genomes probably over the last 40,000 years after introgression from Denisova (**Fig. 4e**). This duplication polymorphism represents the largest introgressed archaic hominin locus in modern humans.

The ancestral human genome: The breadth of the dataset allowed us to reconstruct the structure and content of the ancestral human genome prior to human migration and subsequent gene loss. To identify ancestral sequences potentially lost by deletion, we identified a set of sequences present in chimpanzee and orangutan reference genomes but absent from the human reference genome (20,373 nonredundant loci corresponding to 40.7 Mbp of sequence). Of these, 9,666 (27.6 Mbp) were unique (i.e., not composed of common repeats). Due to the inability to accurately genotype copy number for unique segments less than 500 bp by read-depth analysis, we limited our ancestral reconstruction to non-repetitive sequences greater than this length threshold. While the majority represented deletions specifically lost in the human lineage since divergence from great apes (6,341 loci) or else reference genome artifacts (2,026 loci fixed-copy 2 in all individuals assessed, 6.2 Mbp), a small subset of these (n=571 or 1.55 Mbp) segregate as bi-allelic polymorphisms in human populations (**Fig. 5a**). As expected, Africans were more likely to show evidence of these ancestral sequences compared to non-African populations, as the latter have experienced more population bottlenecks and thus retained less of the ancestral human diversity. A comparison to archaic genomes allowed us to identify sequences (50 loci or 104 kbp) that were present in Denisova or Neanderthal but lost in all contemporary humans as

well as ancestral sequences present in all humans but not found in Denisova or Neanderthal (17 loci or 33.3 kbp).

No difference in the CNV load between Africans and non-Africans: The high coverage and uniformity allowed us to contrast putatively deleterious, exon-removing CNVs among human populations, of interest in disease studies (*36-38*). In our callset we identified 2,437 CNVRs intersecting exons. The distribution of allele counts of these tended towards lower frequency events with, again, deletions more rare than duplications (Wilcoxon rank sum test, P=1.25e-5). Collectively, individuals harbor a mean of 19.2 exon-intersecting deletions per genome (22.8 per diploid genome), with African individuals exhibiting, on average, a mean of 22.4 deletions compared to 18.6 in non-Africans (26.1and 22.1 per diploid genome, respectively), consistent with the increased diversity of African populations and consistent with data observed for loss-of-function SNVs ((*12, 39*), ~122 LoF SNVs in Africans vs. ~104 in non-Africans).

While non-African individuals exhibited more homozygous deletion variants compared to Africans, among exon-intersecting deletions no such pattern was observed. Exon-intersecting duplications were much more balanced with African populations showing only a slight excess when compared to non-Africans (98.4 vs. 95.2 events per genome). Studies of SNVs have not found consistent evidence of difference in load between African compared to non-African populations (*40-42*). We compared the difference in load between African and non-African populations for deletions and duplications, respectively. Here, we defined the difference in load as the difference in the sum of derived allele frequencies between African and non-African populations, $\mathrm{L(Afr)} - \mathrm{L(nAfr)} = \sum_{\forall \mathrm{i}} \mathrm{P_{Afr}(i)} - \sum_{\forall \mathrm{i}} \mathrm{P_{nAfr}(i)}$ where $\mathrm{P_{Afr}(i)}$ is the derived allele frequency of a variant i. Prima facie Africans exhibited an apparent higher deletion load than non-African populations (**Fig. 5b,** P=0.0003, block bootstrap test), though only a nominal difference in the load of exonic deletions (P=0.0482). Duplications showed no such effect.

We reasoned that this striking difference might potentially be driven by high-frequency derived alleles, absent from the human reference genome, which was enriched for clone libraries of non-African ancestry (*5*). Approaches that rely on identifying CNVs based on read placements to the reference genome would necessarily miss these CNVs, decreasing the number of variants identified in individuals more closely resembling the reference, i.e., non-Africans. To test this hypothesis we incorporated the bi-allelic 571 non-repetitive human CNV loci described above. Copy numbers were estimated for these sequences in each of the individuals assessed by remapping raw reads against an ancestral human reference genome. As expected, the deletion allele of this sequence was at a high frequency (mean derived allele frequency, DAF=0.58). After including these sequences we observed no difference in the CNV load between Africans and non-Africans (95% confidence interval -18.4 to 8.8 load difference as defined above, **Fig. 5b**) underscoring the importance of an unbiased human reference for such population genetic assessments.

Although we found no CNV or SNV load differences between populations, we examined whether the relative proportion of base pairs differing among individuals derived from CNVs versus SNVs showed any population-specific trends. We calculated the number of base pairs varying between all pairs of individuals assessed in our study contributed either from SNVs or from deletions calculating the DEL-bp/SNV-bp ratio. As expected, the number of base pairs differing between individuals by deletions or by SNVs independently was always higher among African individuals when compared to other populations. Surprisingly, the ratio of deletion-bp to SNV-bp was substantially higher within non-African populations (mean 1.27 compared to 1.14, **Fig. 5c,d**). This relative increase in deleted base pairs was most pronounced among non-African populations, which have experienced more recent genetic bottlenecks (e.g., Siberian and Amerindian). Given the absence of a significant difference in the deletion load comparing African and non-African populations, there is no reason to believe that this finding is due to differences in the effectiveness of selection against deletions since the populations separated. However, selection places a downward pressure on the allele frequencies of both deletions and SNVs, with the pressure being stronger for deletions because the selection coefficients are stronger on average. As has been previously shown for SNVs, different allele frequency spectra for deletions in contrast to SNVs has the potential to interact with the differences in demographic history across populations—even without differences in the effectiveness of selection after population separation—to contribute to observed differences in the apportionment of genetic variation among human populations (*41*).

Discussion

While the mutational properties and selective signatures of SNVs have been explored extensively, similar analyses of CNVs have lagged behind. As a class, duplications show generally poor correlations with SNV density, have poor linkage disequilibrium to SNVs (*43*, *44*), and are less informative as phylogenetic markers but are more likely to be stratified than deletions among human populations. This observation may be explained by the fact that directly orientated duplications show a gradient of elevated mutation rates due to non-allelic homologous recombination and, as such, can change their copy number state more dynamically over short periods of time. This property also makes this class of variation, similar to highly mutable loci such as minisatellites (*45*), particularly susceptible to homoplasy—i.e., identity by state as opposed identity by descent. Deletions, in contrast, recapitulate most properties of SNVs because they are more likely to exhibit identity by descent as a result of single ancestral mutation event.

We have provided here sequencing data for the study of human diversity and utilize this resource to explore patterns of human CNV diversity at a fine scale of resolution (>1 kbp). As expected, human genomes differ more with respect to CNVs than SNVs and almost one-half of these CNV differences map to regions of segmental duplication. Both deletion and duplication analyses consistently distinguish African, Oceanic, and Amerindian human populations. Africans show

the greatest deletion and duplication diversity and have the lowest rate of fixed deletions with respect to ancestral human insertion sequences. Oceanic and Amerindian, in contrast, show greater CNV differentiation likely as a result of longer periods of genetic isolation and founder effects (*46*). Among the Oceanic, the Papuan–Bougainville group stands out in sharing more derived CNV alleles in common with Denisova, including a massive interspersed duplication that rose to high frequency over a short period of time.

We find that duplications and deletions exhibit fundamentally different population-genetic properties. Duplications are subjected to weaker selective constraint and are four times more likely to affect genes than deletions (**Table 1**) indicating that they provide a larger target for adaptive selection. After controlling for reference genome biases, we find no difference in CNV load between human populations when measured on a per-genome basis which is what matters to disease risk assuming that CNVs act additively. However, we find that the proportion of human variation that can be ascribed to CNVs rather than to SNVs is greater among non-Africans than among Africans. The biological significance of this difference should be interpreted cautiously and will require association studies to determine its relevance to disease and other phenotypic differences.

Tables

Table 1: CNVs and SNVs broken down by their intersection with genomic region. The number of Mbp of exonic and segmentally duplicated CNVs reflects the amount of exonic and segmental duplication sequence affected, respectively, not the total sum of the intersecting CNVs.

Class	Autosomal (Mbp)	X chromosome (Mbp)	Exonic (Mbp)	Segmentally Duplicated (Mbp)
Deletions	7,233 (78.99)	278 (6.61)	636 (0.32)	331 (8.47)
Duplications	7,234 (129.62)	267 (6.46)	2,093 (1.56)	4,462 (96.93)
Subtotal	14,467 (204.54)	545 (12.61)	2,729 (1.84)	4,793 (99.84)
SNVs	32,630,650 (32.63)	1,175,170 (1.18)	314,872 (0.31)	1,559,158 (1.56)
All	32,645,117 (237.17)	1,175,715 (13.79)	317,601 (2.15)	1,563,951 (101.4)

Table 2: Summary statistics of bi-allelic CNV deletions versus SNVs by continental population group.

continental population group	n	segregating SNVs	segregating CNVs	CNVs / individual (median)	heterozygous CNVs / individual (median)	continental population group specific CNVs (allele	θ_{CNV} / genome

						count ≥2)	
West Eurasian (WEA)	58	13610715	1728	279.0	209.0	688 (89)	324.42
Oceanic (OCN)	21	9467426	1022	263.0	173.0	353 (84)	237.51
East Asian (EA)	45	17452049	1463	271.0	191.0	525 (59)	288.48
Siberian (SIB)	23	9644914	1102	285.0	205.0	214 (30)	250.74
South Asian (SA)	27	11308883	1405	279.0	208.0	418 (43)	308.32
Americas (AMR)	21	8127639	899	266.0	169.0	208 (25)	208.93
African (AFR)	41	21698517	2663	319.0	261.0	1772 (702)	534.97

Table 3: CNVs differentiated between human populations. CNVs intersecting genes that show dramatic difference in copy number (as measured by Vst) between human populations (see Figure 1 for definition of populations).

Locus	Genes	V_{st}	Copy range	Description
chr2:97849921-97899292	*ANKRD36*	0.49 (OCN-WEA)	30-41	Repeat domain expanded to 45 copies in Papuans.
chr1:144146792-144224420	*NBPF*	0.32 (AFR-EA)	185-271	Expansion of the *DUF1220* repeat domain in Africans and Amerindians. Copy number associated with cognitive function and autism severity (*47*).
chr6:35749042-35767153	*CLPS*	0.29 (AMR-SA)	2-6	Pancreatic colipase involved in dietary metabolism of long chain triglyceride fatty acids. Increased expression is negatively correlated with blood glucose in mice (*30*).
chr16:72088031-72119241	*HP, HPR*	0.25 (AFR-WEA)	1-6	Haptoglobin and haptoglobin-related genes are expanded exclusively in Africa and associated with a possible protective effect against trypanosomiasis (*31*).
chr12:64011854-64015265	*DPY19L2*	0.32 (OCN-SA)	5-7	*DPY* genes are required for sperm head elongation and acrosome formation during spermatogenesis and *DPY19L2* homozygous deletions have been identified as a major cause of globozoospermia (*48*).
chr1:74648583-74664195	*LRRIQ3*	0.23 (AMR-WEA)	2-3	*LRRIQ3* is duplicated exclusively in Siberian and Amerindian populations.
chr17:43692284-43708692	*CRHR1*	0.25 (EA-WEA)	4-7	Deletions of corticotropin-releasing hormone receptor 1 result in reduced anxiety and neurotransmission impairments in mice (*49*).
chr5:150201231-150223428	*IRGM* promoter	0.25 (AFR-WEA)	0-2	The *IRGM* promoter CNV is a Crohn's disease risk factor (*50*).
chr3:195771149-195776591	*TFRC promoter*	0.57 (AFR-EA)	0-2	Transferrin receptor is a cellular receptor for New World haemorrhagic fever arenaviruses (*51*).

Figure Legends

Figure 1 – Analysis of CNVs in several world populations: The geographical locations of the 125 human populations, including two archaic genomes, assessed in this study. Populations are colored by their continental population groups and archaic individuals are indicated in black.

Figure 2 – Population structure and CNV diversity: Principal component analysis (PCA) of individuals assessed in this study plotted for bi-allelic deletions (a) and duplications (b) with colors and shapes representing continental and specific populations, respectively. Individuals are projected along the PC1 and PC2 axes. The deletion (c) and duplication (d) heterozygosity plotted and grouped by continental population. The relationship between SNV heterozygosity and deletion (e) or duplication (f) heterozygosity is compared.

Figure 3 – Selection on CNVs: Folded allele frequency spectra of exon-intersecting deletions (a) and duplications (b). While deletions intersecting exons are significantly rarer than intergenic deletions, exon-intersecting duplications show no difference compared to intergenic duplications. The mean frequency of CNVs beyond a minimum size threshold is plotted for deletions (c) and duplications (d). A strong negative correlation between size and allele frequency is observed for deletions but less so for duplications.

Figure 4 – Population-stratified CNVs and archaic introgression: a) Four specific examples of population-stratified CNVs intersecting genes are shown, including *LRRIQ3*, the pancreatic collipase *CLPS*, the sperm head an acrosome formation gene *DPY19L2*, and the haptoglobin and haptoglobin-related genes *HP* and *HPR*. Dot-plots indicating the copy of the locus in each individual and pie charts with colors depicting the continental population distribution per copy number (see text for details and Figures 1 and 2 and dot plots for color scheme). b) Predicted copy number on the basis of read-depth for a 73.5 kbp duplication on chromosome 16. It is observed in the archaic Denisovan genome and at 0.84 allele frequency in Papuan and Bougainville populations, yet absent from all other assessed populations. The duplication intersects two microRNAs. The orange arrow corresponds to the position and orientation of this duplication as further highlighted in c and d. c) A heatmap representation of a ~1 Mbp region of chromosome 16p12 (chr16:21518638-22805719). Each row of the heatmap represents the estimated copy number in 1 kbp windows of a single individual across this locus. Genes, annotated segmental duplications, and arrows highlighting the size and orientation in the reference of the Denisova/Papuan-specific duplication locus (locus D) and three other duplicated loci (A, B, C) of interest are shown below. d) The structure of duplications A, B, C and D (as shown in 4c over the same locus) in the reference genome and the discordant paired-end read placements used to characterize two duplication structures. Structure A/C is found in all individuals, though not present in the reference genome, while structure B/D is only found in Papuan and Bougainville individuals indicating a large complex, duplication (~225 kbp) composed of different segmental duplications. Both the A/C and B/D duplication architectures exhibit inverted orientations compared to the reference. The number of reads in all Oceanic and non-Oceanic individuals supporting each structure are indicated. e) Maximum likelihood tree of the 16p12 duplication locus (duplication D in 4b, 4c, 4d) constructed from the locus in Orangutan, Denisova, the human reference and the inferred sequence of the Papuan duplication (*24*). All bootstrap values are 100%.

Figure 5 – The ancestral human genome and CNV burden: a) A heatmap of the allele frequency of 571 (1.55 Mbp) non-repetitive sequences absent from the human reference genome yet segregating in at least one human population ordered in humans by a maximum likelihood tree (*49*). Four groups of interest are highlighted: G1 – ancestral sequences that have almost been completely lost from the human lineage, G2 – ancestral sequences that are largely fixed but rarely deleted (also absent in human reference), G3 – ancestral sequences that have become copy number variable since the divergence of humans and Neanderthals/Denisovans ~700 kya, and G4 – sequences potentially lost in Neanderthals and Denisovans since their divergence from humans. b) The resulting distributions of 10,000 block-bootstrapped estimates of the difference in load between African (AFR) and non-African (nAFR) populations considering only the reference genome (GRCh37) and supplemented by sequence absent from the human reference genome (GRCh37 + NHP) included (see text for details). c) Violin plots of the distribution of the ratio of deletion base pairs to SNV base pairs differing between every pair of African individuals (AFR-AFR), all pairs of non-African individuals (nAFR-nAFR) and every non-African, African pair (nAFR-AFR). d) Heatmap representation of the mean ratio of deletion to SNV base pairs differing between individuals from pairs of populations.

Acknowledgements
We are grateful to the volunteers who donated the DNA samples used in this study. This work was supported, in part, by a U.S. National Institutes of Health (NIH) grant 2R01HG002385 and a grant (11631) from The Paul G. Allen Family Foundation to E.E.E. The sequencing for this study was supported by a grant from the Simons Foundation to D.R. (SFARI 280376) and by a HOMINID grant from The National Science Foundation to D.R. (BCS-1032255). T.K. is supported by ERC Starting Investigator grant FP7 - 261213. R.S. and S.D. received support from The Ministry of Education and Science, Russian Federation (14.Z50.31.0010). H.S., E.M. R.V. and M.M. are supported by Institutional Research Funding from the Estonian Research Council IUT24-1 and by the European Regional Development Fund (European Union) through the Centre of Excellence in Genomics to Estonian Biocentre and University of Tartu. S.A.T. is supported by the grants 5DP1ES022577 05, 1R01DK104339-01 and 1R01GM113657-01. C.T.S. is supported by The Wellcome Trust grant 098051. C.M.B. is supported by the National Science Foundation (award numbers 0924726 and 1153911). E.E.E. and D.R. are investigators of the Howard Hughes Medical Institute. Data are deposited into ENA and variant calls are deposited in dbVar (PRJEB9586, PRJNA285786). E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc. and is a consultant for Kunming University of Science and Technology (KUST) as part of the 1000 China Talent Program.

References

1. S. C. Schuster *et al.*, Complete Khoisan and Bantu genomes from southern Africa. *Nature*. **463**, 943–947 (2010).

2. K. M. Steinberg *et al.*, Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature Genetics*. **44**, 872–880 (2012).

3. S. Gravel *et al.*, Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. *PLoS Genet*. **9**, e1004023 (2013).

4. M. Raghavan *et al.*, The genetic prehistory of the New World Arctic. *Science*. **345**, 1255832 (2014).

5. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *Science*. **328**, 710–722 (2010).

6. D. Reich *et al.*, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. **468**, 1053–1060 (2010).

7. M. Rasmussen *et al.*, Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*. **463**, 757–762 (2010).

8. D. F. Conrad *et al.*, Variation in genome-wide mutation rates within and between human families. *Nature Genetics*. **43**, 712–714 (2011).

9. C. D. Campbell *et al.*, Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics*. **44**, 1277–1281 (2012).

10. A. Kong *et al.*, Rate of de novo mutations and the importance of father's age to disease risk. *Nature*. **488**, 471–475 (2012).

11. W. Fu *et al.*, Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. **493**, 216–220 (2013).

12. J. A. Tennessen *et al.*, Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. **337**, 64–69 (2012).

13. P. H. Sudmant *et al.*, Evolution and diversity of copy number variation in the great ape lineage. *Genome Research*. **23**, 1373–1382 (2013).

14. T. Marques-Bonet *et al.*, A burst of segmental duplications in the genome of the African great ape ancestor. *Nature*. **457**, 877–881 (2009).

15. R. A. Kumar *et al.*, Recurrent 16p11.2 microdeletions in autism. *Human Molecular Genetics*. **17**, 628–638 (2008).

16. A. J. Sharp *et al.*, A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genetics*. **40**, 322–328 (2008).

17. J. Sebat *et al.*, Large-scale copy number polymorphism in the human genome. *Science*. **305**, 525–528 (2004).

18. L. A. Weiss *et al.*, Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).

19. S. A. McCarroll *et al.*, Common deletion polymorphisms in the human genome. *Nature Genetics*. **38**, 86–92 (2006).

20. S. A. McCarroll *et al.*, Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*. **40**, 1166–1174 (2008).

21. D. F. Conrad *et al.*, Origins and functional impact of copy number variation in the human genome. *Nature*. **464**, 704–712 (2010).

22. M. Jakobsson *et al.*, Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*. **451**, 998–1003 (2008).

23. A. Itsara *et al.*, Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84**, 148–161 (2009).

24. *Information on materials and methods is available at the Science Web site*.

25. K. Prüfer *et al.*, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*. **505**, 43–49 (2014).

26. M. Meyer *et al.*, A high-coverage genome sequence from an archaic Denisovan individual. *Science*. **338**, 222–226 (2012).

27. R. E. Mills *et al.*, Mapping copy number variation by population-scale genome sequencing. *Nature*. **470**, 59–65 (2011).

28. G. M. Cooper *et al.*, A copy number variation morbidity map of developmental delay. *Nature Genetics*. **43**, 838–846 (2011).

29. R. Redon *et al.*, Global variation in copy number in the human genome. *Nature*. **444**, 444–454 (2006).

30. J. Zhang, K. Kaasik, M. R. Blackburn, C. C. Lee, Constant darkness is a circadian metabolic signal in mammals. *Nature*. **439**, 340–343 (2006).

31. R. J. Hardwick *et al.*, Haptoglobin (HP) and Haptoglobin-related protein (HPR) copy number variation, natural selection, and trypanosomiasis. *Human Genetics*. **133**, 69–83 (2014).

32. L. A. Hindorff *et al.*, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362–9367 (2009).

33. S. R. Grossman *et al.*, Identifying recent adaptations in large-scale genomic data. *Cell*. **152**, 703–713 (2013).

34. J. E. Horvath, S. Schwartz, E. E. Eichler, The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. *Genome Research*. **10**, 839–852 (2000).

35. J. Cheung *et al.*, Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol*. **4**, R25 (2003).

36. N. Krumm *et al.*, Copy number variation detection and genotyping from exome sequence data. *Genome Research*. **22**, 1525–1532 (2012).

37. M. Fromer *et al.*, Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).

38. Deciphering Developmental Disorders Study, Large-scale discovery of novel genetic causes of developmental disorders. *Nature*. **519**, 223–228 (2015).

39. D. G. MacArthur *et al.*, A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. **335**, 823–828 (2012).

40. W. Fu, R. M. Gittelman, M. J. Bamshad, J. M. Akey, Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *Am. J. Hum. Genet.* **95**, 421–436 (2014).

41. R. Do *et al.*, No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genetics*. **47**, 126–131 (2015).

42. Y. B. Simons, M. C. Turchin, J. K. Pritchard, G. Sella, The deleterious mutation load is insensitive to recent population history. *Nature Genetics*. **46**, 220–224 (2014).

43. C. D. Campbell *et al.*, Population-genetic properties of differentiated human copy-number polymorphisms. *Am. J. Hum. Genet.* **88**, 317–332 (2011).

44. D. P. Locke *et al.*, Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).

45. A. J. Jeffreys, V. Wilson, S. L. Thein, Hypervariable "minisatellite" regions in human DNA. *Nature*. **314**, 67–73 (1985).

46. A. T. Duggan *et al.*, Maternal history of Oceania from complete mtDNA genomes: contrasting ancient diversity with recent homogenization due to the Austronesian expansion. *Am. J. Hum. Genet.* **94**, 721–733 (2014).

47. J. M. Davis, V. B. Searles Quick, J. M. Sikela, Replicated linear association between DUF1220 copy number and severity of social impairment in autism. *Human Genetics*. **134**, 569–575 (2015).

48. I. Koscinski *et al.*, DPY19L2 deletion as a major cause of globozoospermia. *Am. J. Hum. Genet.* **88**, 344–350 (2011).

49. D. Refojo *et al.*, Glutamatergic and dopaminergic neurons mediate anxiogenic and anxiolytic effects of CRHR1. *Science*. **333**, 1903–1907 (2011).

50. S. A. McCarroll *et al.*, Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature Genetics*. **40**, 1107–1112 (2008).

51. S. R. Radoshitzky *et al.*, Transferrin receptor 1 is a cellular receptor for New World haemorrhagic fever arenaviruses. *Nature*. **446**, 92–96 (2007).

52. N. A. Rosenberg *et al.*, Genetic structure of human populations. *Science*. **298**, 2381–2385 (2002).

53. B. P. Coe *et al.*, Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nature Genetics*. **46**, 1063–1071 (2014).

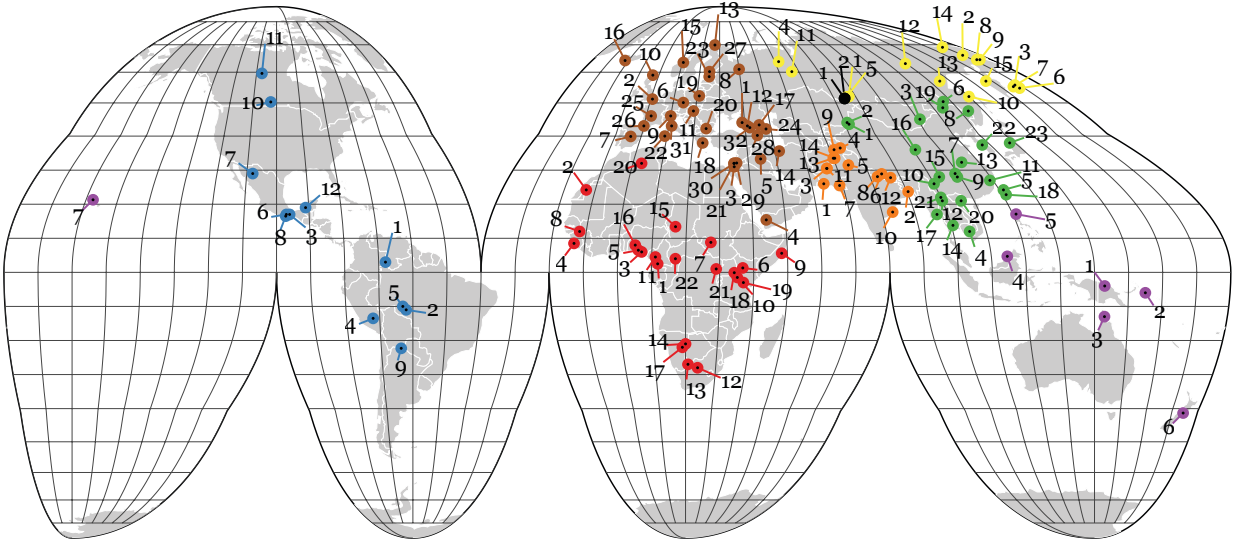
54. P. H. Sudmant *et al.*, Diversity of human copy number variation and multicopy genes. *Science*. **330**, 641–646 (2010).

55. I. Lazaridis *et al.*, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. **513**, 409–413 (2014).

56. Q. Fu *et al.*, Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*. **514**, 445–449 (2014).

57. J. M. Zook *et al.*, Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*. **32**, 246–251 (2014).

Figure 1 - Sudmant 2015



1 - Neanderthal
2 - Denisova
1 - Piapoco
2 - Surui
3 - Mixe
4 - Quechua
5 - Karitiana
6 - Mixtec
7 - Pima
8 - Zapotec
9 - Chane
10 - Cree
11 - Chipewyan
12 - Maya
1 - Xibo
2 - Uygur
3 - Mongola
4 - Cambodian
5 - Atayal
6 - Oroqen
7 - Tujia
8 - Hezhen
9 - Miao
10 - Naxi
11 - She
12 - Dai
13 - Han
14 - Thai
15 - Yi
16 - Tu
17 - Burmese
18 - Ami
19 - Daur
20 - KinhKVH
21 - Lahu
22 - Korean
23 - Japanese
1 - Papuan
2 - Bougainville
3 - AustraliaECCAC
4 - Dusun
5 - Igorot
6 - Maori
7 - Hawaiian
1 - Adygei
2 - EnglandGBR
3 - Bedouin
4 - YemeniteJew
5 - IraqiJew
6 - Czechoslovakian
7 - SpainIBS
8 - Russian
9 - Bergamo
10 - Orcadian
11 - Hungarian
12 - Abkhasian
13 - Saami
14 - Iranian
15 - Norwegian
16 - Icelander
17 - Chechen
18 - Greek
19 - Polish
20 - Bulgarian
21 - Palestinian
22 - Sardinian
23 - Estonian
24 - Lezgin
25 - French
26 - Basque
27 - FinlandFIN
28 - Armenian
29 - Jordanian
30 - Druze
31 - Tuscan
32 - Georgian
1 - Kongo
2 - Saharawi
3 - Igbo
4 - MendeMSL
5 - EsanESN
6 - LuhyaLWK
7 - Dinka
8 - Mandenka
9 - Somali
10 - MasaiMKK
11 - Lemande
12 - BantuSETswana
13 - KhomaniSan
14 - San
15 - GambiaGWD
16 - Yoruba
17 - BantuSEHerero
18 - Luo
19 - BantuKenya
20 - Mozabite
21 - Mbuti
22 - Biaka
1 - Makrani
2 - BengaliBEB
3 - Balochi
4 - Burusho
5 - PunjabiPJL
6 - Tibetan
7 - Sindhi
8 - Kusunda
9 - Kalash
10 - Madiga
11 - Brahui
12 - Sherpa
13 - Hazara
14 - Pathan
1 - Tubalar
2 - EskimoNaukan
3 - AleutBeringIsland
4 - MansiSosvaRiver
5 - Altaian
6 - TlingitPreobrazhenskoye
7 - TlingitNikolskoye
8 - EskimoChaplino
9 - EskimoSireniki
10 - Ulchi
11 - MansiKondaRiver
12 - Yakut
13 - Even
14 - Chukchi
15 - Itelman

Figure 2 - Sudmant 2015

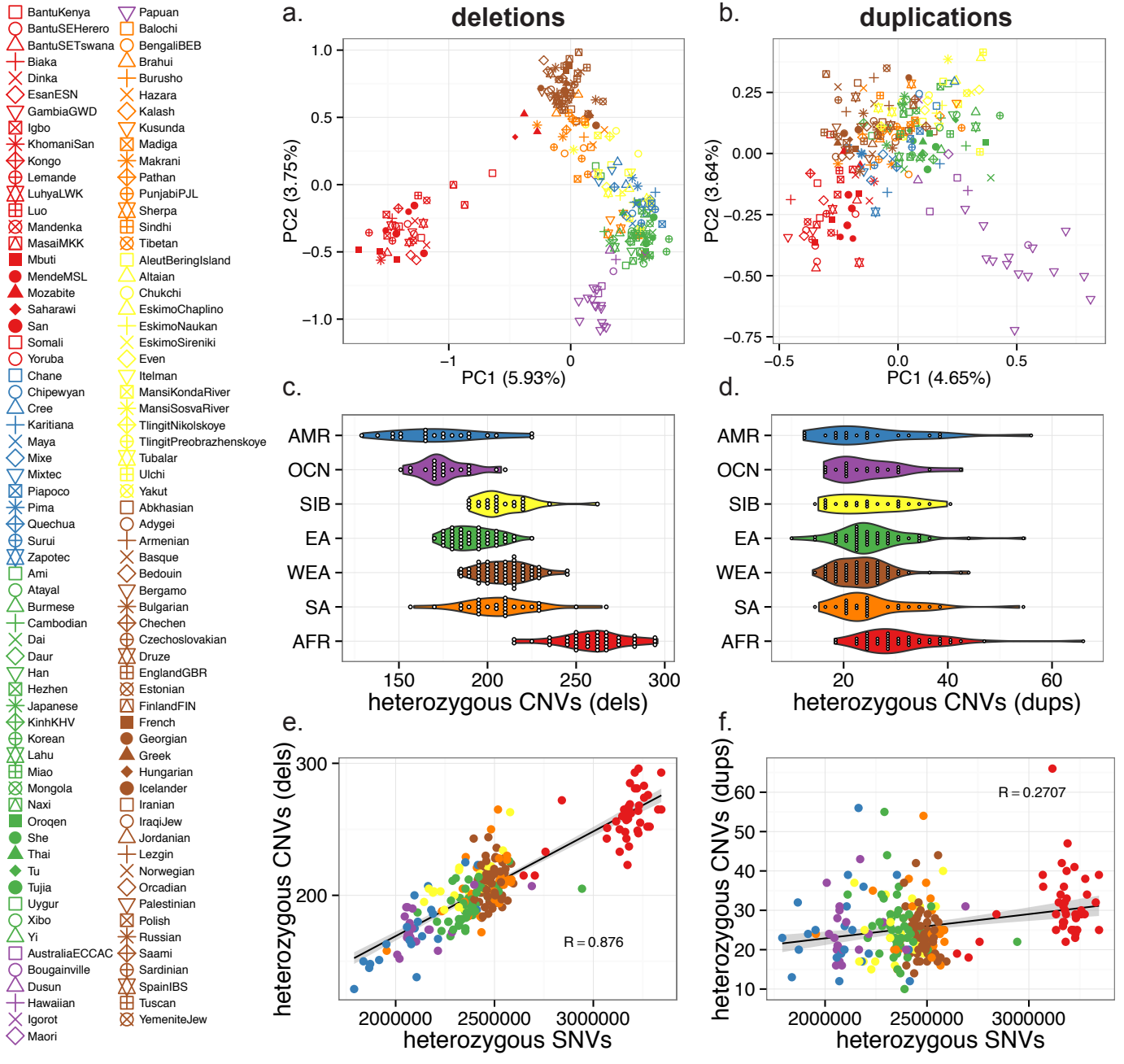


Figure 3 - Sudmant 2015

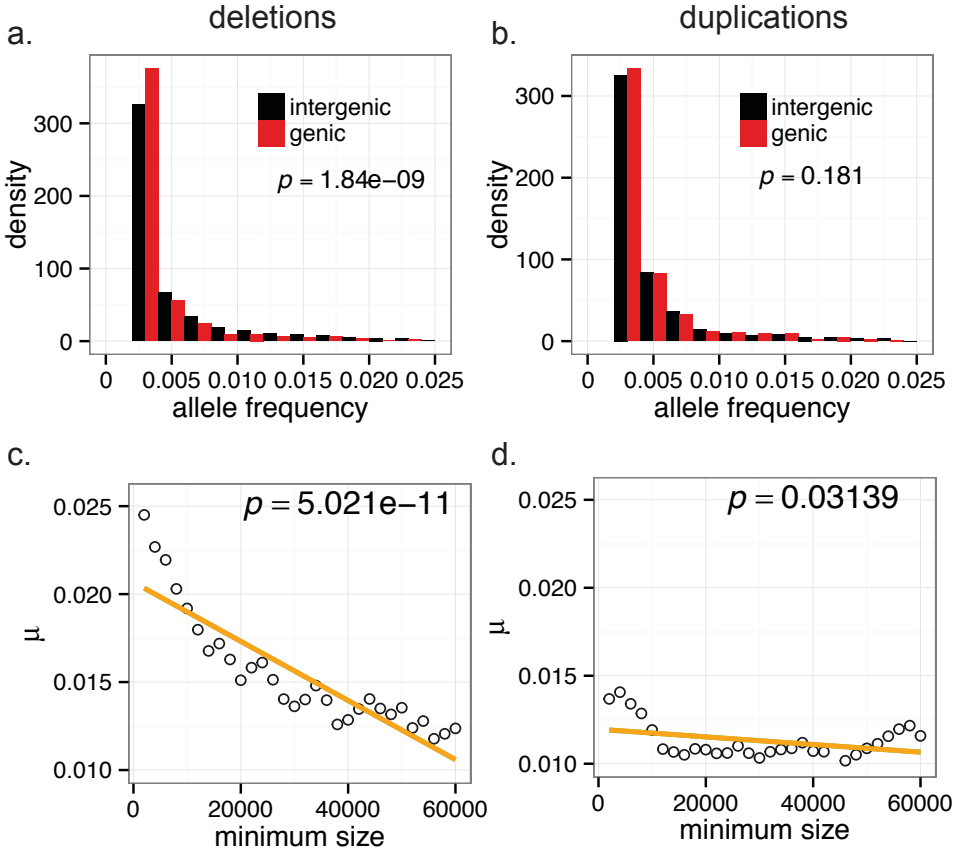


Figure 4 - Sudmant 2015

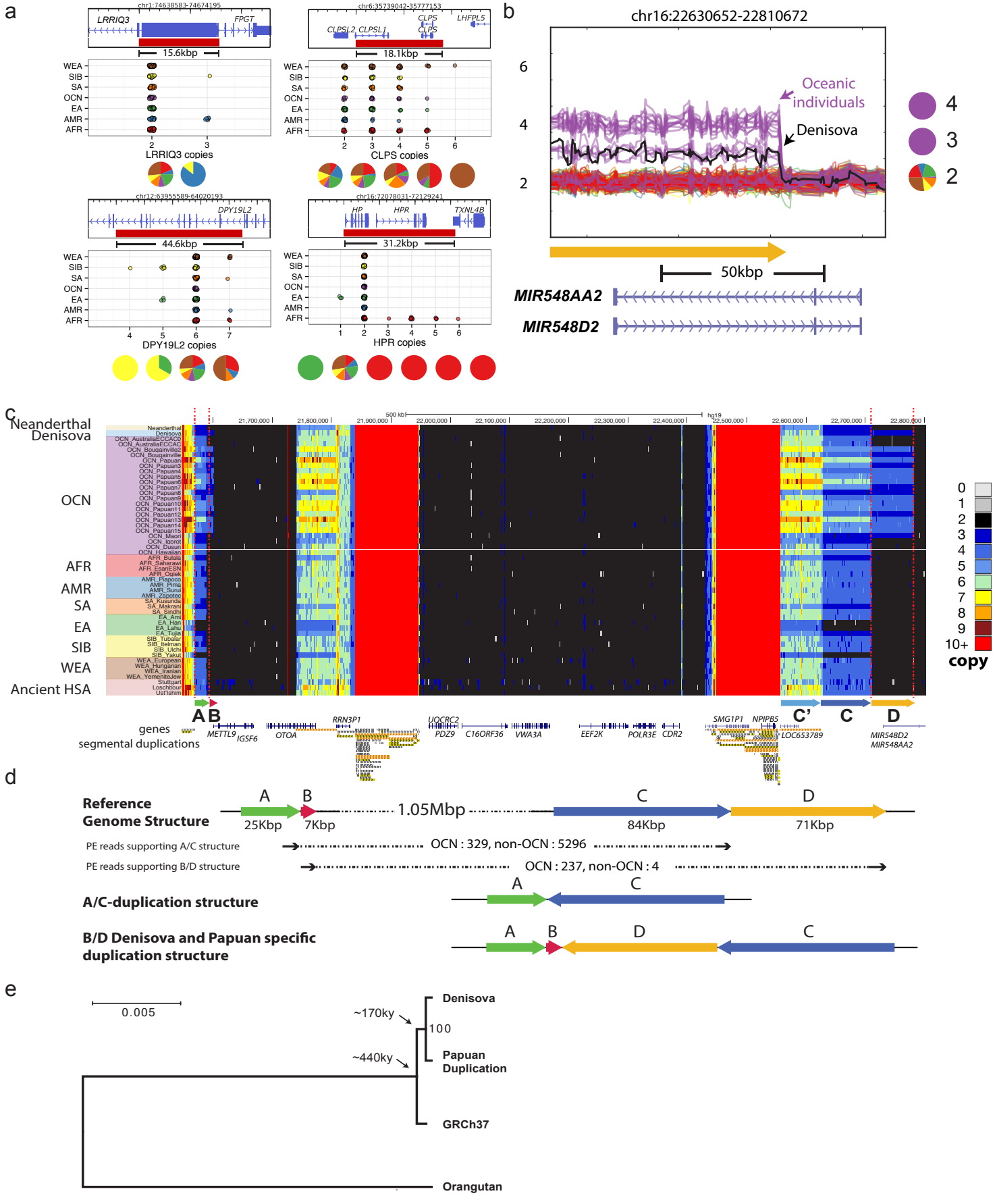
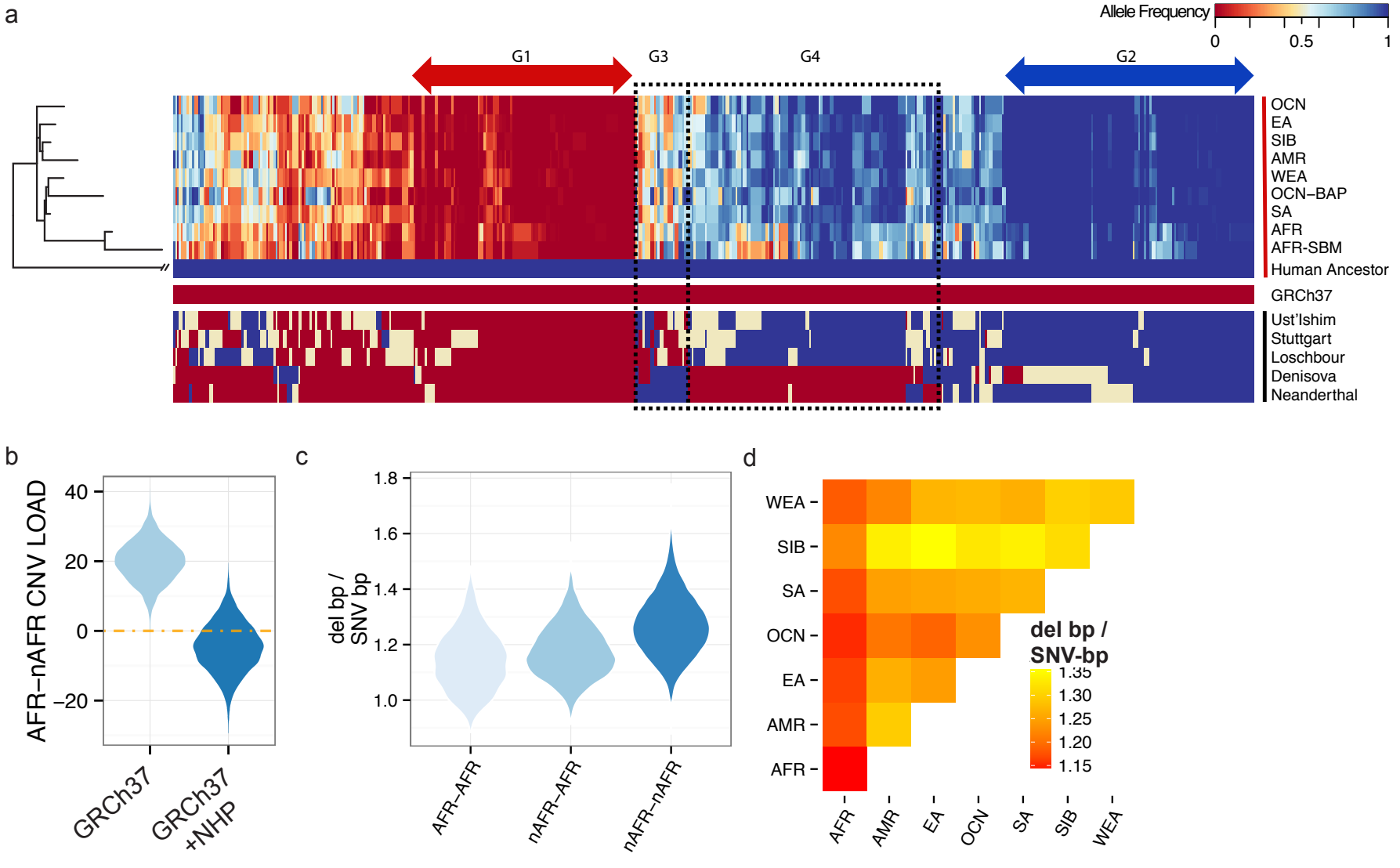
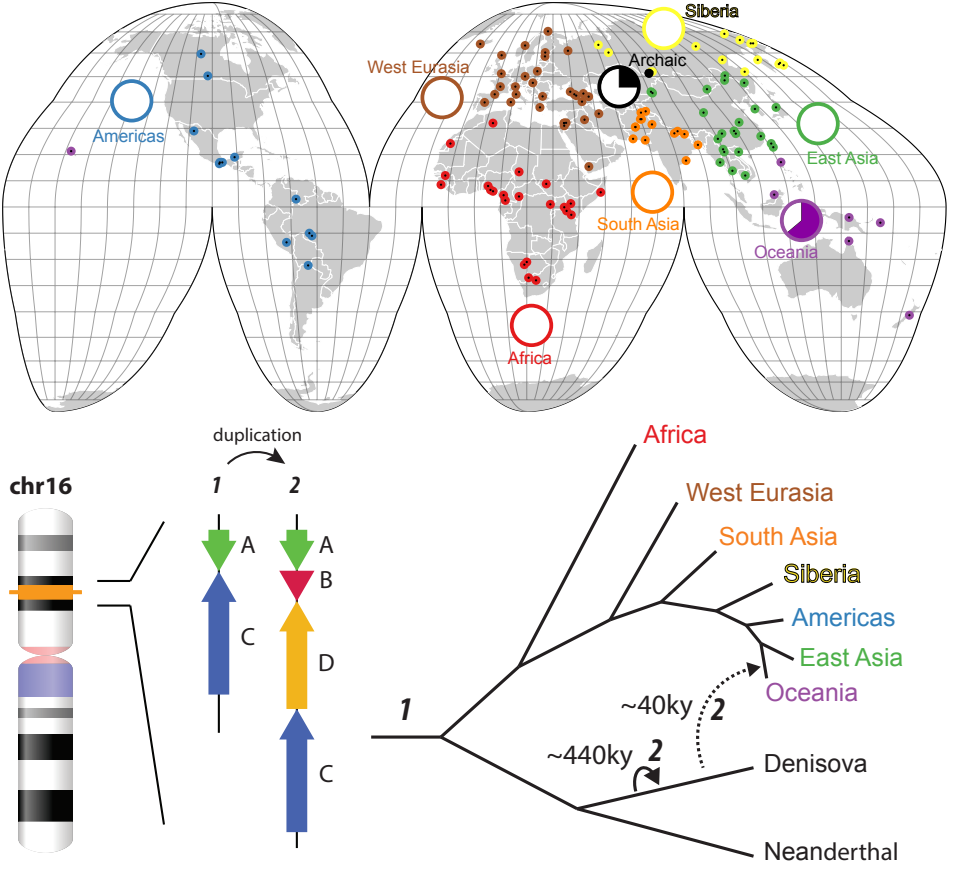


Figure 5 - Sudmant 2015




Siberia
West Eurasia
Archaic
Americas
East Asia
South Asia
Oceania
Africa
chr16
duplication
1
2
A
A
B
C
D
C
Africa
West Eurasia
South Asia
Siberia
Americas
East Asia
Oceania
1
~40ky
2
~440ky
2
Denisova
Neanderthal