www.sciencemag.org/cgi/content/full/science.aab3761/DC1

Supplementary Materials for

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

Peter H. Sudmant, Swapan Mallick, Bradley J. Nelson, Fereydoun Hormozdiari, Niklas Krumm, John Huddleston, Bradley P. Coe, Carl Baker, Susanne Nordenfelt, Michael Bamshad, Lynn B. Jorde, Olga L. Posukh, Hovhannes Sahakyan, W. Scott Watkins, Levon Yepiskoposyan, M. Syafiq Abdullah, Claudio M. Bravi, Cristian Capelli, Tor Hervig, Joseph T. S. Wee, Chris Tyler-Smith, George van Driem, Irene Gallego Romero, Aashish R. Jha, Sena Karachanak-Yankova, Draga Toncheva, David Comas, Brenna Henn, Toomas Kivisild, Andres Ruiz-Linares, Antti Sajantila, Ene Metspalu, Jüri Parik, Richard Villems, Elena B. Starikovskaya, George Ayodo, Cynthia M. Beall, Anna Di Rienzo, Michael Hammer, Rita Khusainova, Elza Khusnutdinova, William Klitz, Cheryl Winkler, Damian Labuda, Mait Metspalu, Sarah A. Tishkoff, Stanislav Dryomov, Rem Sukernik, Nick Patterson, David Reich, Evan E. Eichler*

*Corresponding author. E-mail: eee@gs.washington.edu

Published 6 August 2015 on *Science* Express
DOI: 10.1126/science.aab3761

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/science.aab3761/DC1)

Supplementary Materials for

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

INDEX

Note: Tables S1-S5 are included as separate Excel sheets

Section S1: Data set generation

This paper reports deep genome sequences from a total of 236 de-identified human samples from 123 populations (**Table S2**; see separate Excel sheet). We aimed to sequence two samples from most populations, although for 31 populations we report data for one sample, for nine populations we report data for three samples, and for

Papuan we report data from 14 samples. A total of 63 DNA samples were extracted directly from blood or saliva, and the remaining 173 DNA samples were extracted from cell lines. Supplementary Data **Table S2** provides information on each of the samples included in the study, as well as information on population affiliation, sampling location, gender, and DNA source. This study was reviewed by the Harvard Medical Institution Review Board (Protocol Title M11381, Protocol Number MOD14-4442-01) and was determined not to constitute Human Subjects Research as the DNA samples are all de-identified and no phenotype information is available for any of the samples.

A total of 168 samples came from cell line repositories that distribute DNA from de-identified individuals for the purpose of research into human genetic variation. This includes 119 samples from the CEPH - Human Genome Diversity Cell Line Panel, 41 samples from the Coriell Cell Repositories, 4 from Tel Aviv University, and 4 from the European Collection of Cell Cultures (ECACC). In the case of the ECACC samples, we requested a formal re-review to determine whether the samples we analyzed whose contributors we could not contact were collected in a way that made it appropriate to carry out whole-genome sequencing (WGS) and public data distribution, and this was determined to be the case.

A total of 68 samples were provided by individual investigators who are co-authors of this study. For each set of samples, a representative of each set of contributing investigators filed a signed letter affirming that, to the best of their knowledge, the following conditions were met for the samples that they were involved in contributing. Specifically, these conditions are the criteria that were developed for the purpose of reviewing samples as appropriate for inclusion in the CEPH - Human Genome Diversity Cell Line Panel *(52)*, namely:

(i) That the samples were collected with informed consent.

(ii) That the samples were collected with consent that was not merely for a specific biomedical research project but that included as one of its stated purposes a broader scientific use for the study of human population genetics, human evolution, human history, or "human genetics" in general.

(iii) That the samples were collected in a way that is consistent with the then-applicable laws and regulations of the jurisdictions in which they had been collected.

All DNA samples were sent to Illumina Ltd. for deep genome sequencing on Illumina HiSeq 2000 sequencers and were prepared using the same PCR-free protocol ensuring uniformity of data processing over the entire dataset.

Section S2: Quality control analysis of genomes

Identification of cell line artifacts

Passaged cell lines are often subject to artifacts, such as large-scale CNVs and aneuploidies. We, thus, assessed each of the genomes analyzed in this study independently for the presence of artifactual variants, in particular chromosomal aneuploidies, microdeletions, and microduplications that are suspect due to their size and unlikely placement based on a map of large CNVs developed from ~20,000 humans *(53)*. We identified four individuals containing likely cell line artifacts (**Table S6**) and excluded these from further analysis.

Table S6: Putative cell line artifacts.

Individual ID	Aberration
OCN_Bougainville_HGDP00660_F	Chromosome 12 trisomy
EA_She_HGDP01335_F	Chromosome 12 trisomy
EA_Dai_HGDP01315_F	12q telomeric deletion
WEA_Palestinian_HGDP00725_M	XO genotype – likely a cell line artifact as individual is annotated as male

Sex assignments

Amongst the 267 individuals initially assessed, 68 genomes lacked sex information. To identify the sex of these individuals and confirm previously assigned sex designations, we genotyped two loci on the X and Y chromosomes, respectively: *chrX:4320168-4627799* and *chrY:16808875-16824630*. The genotypes of these segments are reflective of the ploidy of the individual chromosomes. All genotypes confirmed predefined sex designations with one exception: WEA_Palestinian_HGDP00725_M exhibited an XO genotype, as noted in the table above, likely the result of a cell line artifact.

GC content associated sequencing coverage biases

To assess the quality of individual genome sequence data, we analyzed regions of known copy number, including 4,836 diploid invariant regions encompassing 1.1 Gbp of sequence and 4.1 Mbp of regions of fixed increased copy number state *(54)*. We first computed a simple correlation between read-depth and regions of known copy showing a strong simple relationship between depth and copy (r>0.9). We next assessed the total fraction of diploid invariant sequence correctly estimated to be copy number 2 amongst individuals. We found 10 individuals in which <98% of loci were correctly assigned the correct copy number state. To better assess these individuals, we binned the fraction of correctly determined diploid regions by their %GC content (**Figure S1**). This analysis highlights GC-associated sequencing biases in a subset of individuals. Plotting the total fraction of correctly predicted diploid invariant regions (**Figure S1**) identifies 31 outlier individuals (falling below a 90% threshold). We excluded these individuals from further analyses using the remaining 236. Additionally, the archaic Denisova and Neanderthal genomes were assessed along with three ancient human genomes totaling 241 individuals *(25,26,55,56)*.

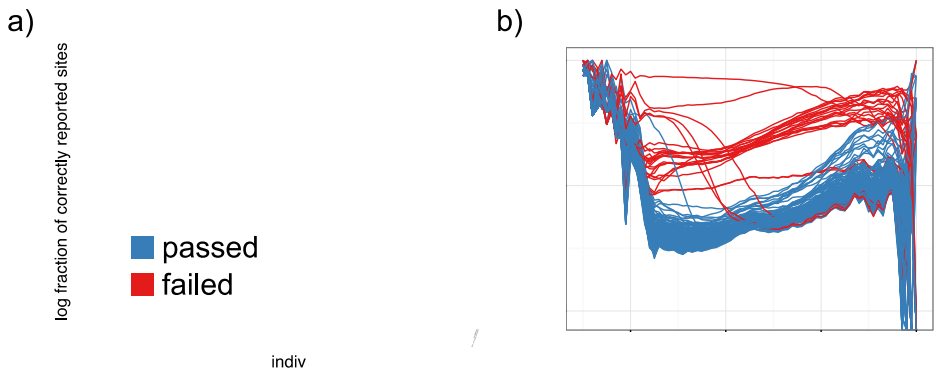

Figure S1: The total fraction of correctly determined windows (a) colored by pass/fail status. (b) The GC-associated coverage biases are colored by pass/fail status of individual.

Section S3: Variant calling

CNV calling methods

To call CNVs genome-wide, we utilized digital comparative genomic hybridization (dCGH), a method we previously described *(54).* Briefly, copy numbers were estimated genome-wide in 500 base-pair (bp) windows of unmasked sequence spaced at overlapping intervals of 100 bp of unmasked sequence. Masking was performed using RepeatMasker and Tandem Repeat Finder thus excluding simple repeats (e.g., *Alu*s, LINEs, microsatellites) but not segmental duplications (SDs). A complete description of genome masking, mapping and copy number calling can be found in references *(13,54)*. Individuals were then compared independently to

each of 17 diverse reference genomes (**Table S7**) representing individuals with the lowest overall variance in their sequence coverage. The $\log_2$ ratio of the signal between test and reference individuals was then computed and analyzed for deviations from 0, which may represent CNVs. We used a scale-space filtering-based CNV detection algorithm *(13)* to identify putative CNVs from dCGH signal. This method assesses the first and second derivatives of the dCGH signal at various scales approximating CNV breakpoints and defining an initial segmentation of the signal. The segmentation is then iteratively refined by clustering adjacent segments that do not differ by more than a predetermined threshold. Performing the dCGH segmentation algorithm against multiple reference genomes increases both the sensitivity and specificity of the callset as each genome is assessed for variation multiple times and rediscovering a CNV strengthens its confidence and provides additional breakpoint resolution.

Table S7: Reference individuals used for dCGH.

AFR_BantuSETswana_HGDP01030_M	WEA_Bergamo_HGDP01153_M	EA_Uygur_HGDP01297_M
AFR_MasaiMKK_NA21490_M	EA_Ami_NA13616_M	OCN_Papuan_HGDP00545_M
AMR_Karitiana_HGDP01012_M	EA_Japanese_HGDP00749_M	SA_BengaliBEB_HG03006_M
AMR_Mixtec_Mixa0099_M	EA_Korean_NA00726_F	SA_Hazara_HGDP00125_M
AMR_Zapotec_zapo0098_M	EA_Lahu_HGDP01320_M	SIB_EskimoNaukan_Nesk22_F
WEA_EnglandGBR_HG00126_M	EA_Tu_HGDP01350_M	

SNV calling methods

We additionally generated an SNV callset for all individuals. Reads were first mapped to the 1000 Genomes Project version of HG19 using the BWA-MEM aligner with default parameters. SNV calls were then generated using Freebayes with calling performed across all individuals simultaneously and requiring the alternative allele at any position to be observed at least five times in any one individual before attempting to evaluate the site. SNVs were then filtered as suggested from analyses of NA12878 (*(57)*, http://bit.ly/1ipCkyJ) filtering for depth (DP > 20) and quality (QUAL > 20). In total, we identified 35,088,808 SNVs among the human genomes assessed with a mean of 2,496,512 heterozygous and 1,454,025 homozygous sites identified per individual. A breakdown of the number of SNVs identified in each individual population surveyed is presented in **Table S8.**

Table S8: The number of SNVs identified among different populations assessed. Note that two European individuals excluded from the analysis of CNVs were included in the combined SNV calling. Abbreviations are as follows: AFR, African; AMR, Americas; EA, East Asian; OCN, Oceanic; SA South Asian; SIB, Siberian; WEA, West Eurasian.

population	n	mean hets / individual	mean homs / individual	segregating sites

AFR	41	3157096	1445966	21698517
AMR	21	2090876	1586176	8127639
EA	45	2367204	1470599	17452049
OCN	21	2135780	1634563	9467426
SA	27	2452334	1409978	11308883
SIB	23	2358096	1459305	9644914
WEA	62	2483263	1355457	13610715
ALL	238	2496512	1454025	35088808

Section S4: Validation of CNV callset

2.1M SNP microarray validation

To evaluate the quality of our callset, we generated CNV calls from Illumina 2.1M SNP microarrays run on 116 of the same individuals assessed for CNVs by WGS. Calls were made on each individual genome using the CNV-Partition software, Version 3.2 from Illumina (http://res.illumina.com/documents/products/technotes/technote_cnv_algorithms.pdf), which has been shown to accurately call large deletions and duplications from SNP arrays. The algorithm was run using default parameters and provides estimated copy numbers for each of the calls in addition to a confidence score. We considered only autosomal calls ≥1 kbp with confidence scores ≥50 and encompassing ≥5 probes resulting in a total of 3,305 calls, a median of 22 CNVs per individual.

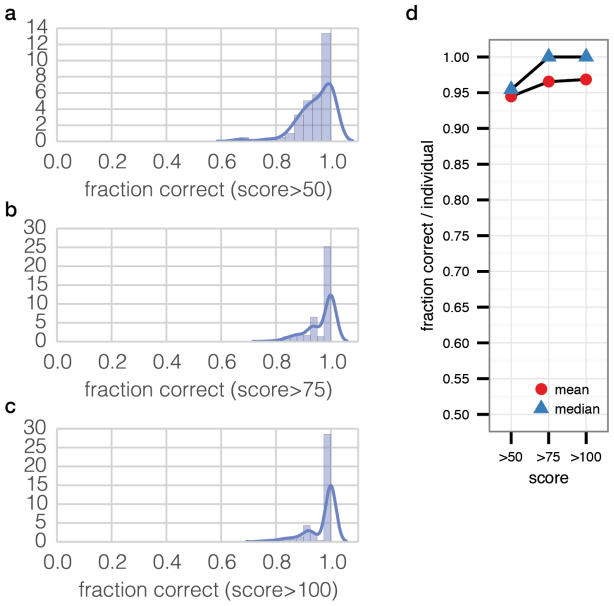

Figure S2: a, b, c - The distribution of the calls detected by CNVpartition correctly identified from WGS for CNVpartition calls with a score >50 (a), >75 (b) and >100 (c). d - The mean and median fraction of correct calls per individual plotted as a function of the CNVpartition score.

Application of the CNV partition identified an exceedingly high number of calls for a subset of individuals (8 individuals with >40 calls). The corresponding intersection of these calls with our WGS-derived callset was much lower than the other individuals, indicating that these particular microarrays likely suffer from an excess of background noise. Excluding this subset, the total concordance between calls made from CNVpartition and dCGH was 94% (2,493/2,652) and increased to 96.5% and 97.1% considering CNVpartition calls with scores >75 and >100, respectively (1,760/1,824 and 1,397/1,439) (**Figure S2**). On a per-individual basis the median fraction of calls identified by CNVpartition additionally called by dCGH was 95.5% for calls with a score >50 and 100% for calls with a score >75 (**Figure S2d**). Calls made by the CNVpartition algorithm had a mean and median size of 84.6 kbp and 23.9 kbp, respectively, with a minimum call size of 1.1 kbp (**Figure S3**). Most of the discordancy occurred for CNVs that were <35 kbp, below which SNP microarrays have reduced sensitivity. These results suggest a per-individual false negative rate of <5%.

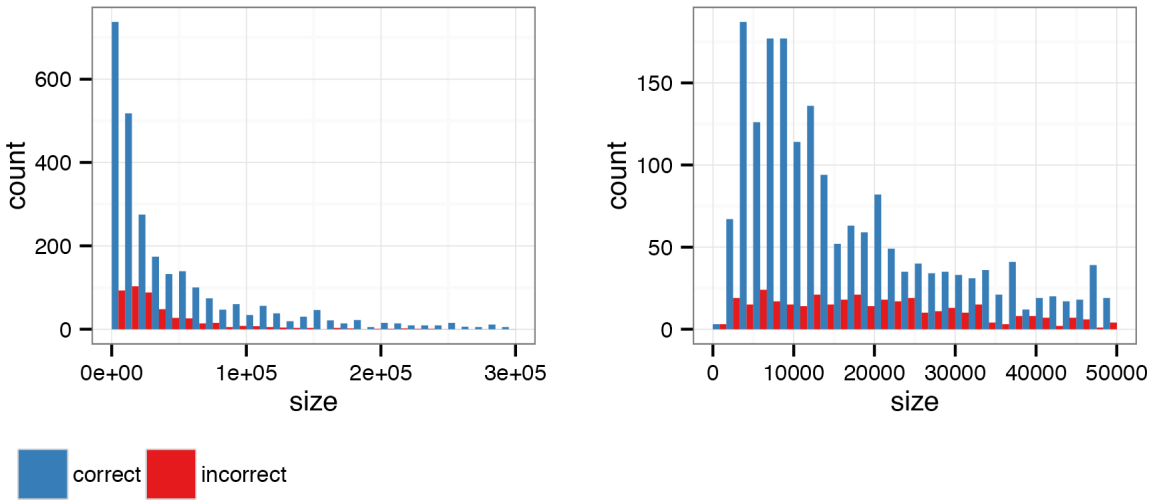

Figure S3: Distribution of calls made by CNVpartition and additionally identified from WGS data.

Targeted array comparative genomic hybridization (aCGH) validations

To further validate our callset we selected 20 individuals at random from our dataset to comprehensively assess; the set consisted of 7 African individuals, 3 individuals from the Americas, 3 Eurasians, 1 Oceanic individual, 3 South Asians, and 3 West Eurasians. We designed a custom 4x180K Agilent probe microarray targeted to every call made in each of these 20 individuals attempting to design a minimum of 20 probes for each putative CNV. The remainder of the microarray was filled with backbone probes resulting in a median target probe spacing of 331 bp and a median backbone probe spacing of 24.5 kbp (**Figure S4**).

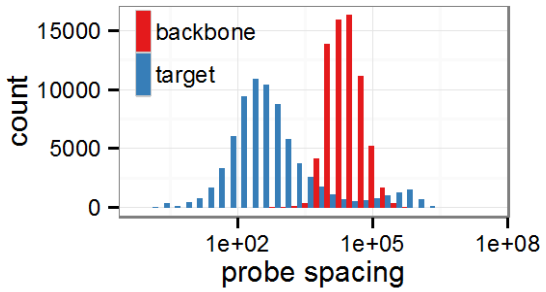

Figure S4: Microarray probe density—distribution of the interprobe spacing for targeted loci and the genomic backbone (x-axis plotted in logspace).

aCGH experiments were performed using NA12878 as a reference against each of the 20 test individuals. DNA from NA12878 and each test individual were labeled with Cy3 or Cy5 dye, (NimbleGen labeling kit), and equal proportions of test and reference DNA were then mixed and hybridized to Agilent 4x180K arrays for 24 hours at 65°C. Finally, slides were washed and scanned. To determine the $\log_2$-ratio thresholds at which a CNV is considered validated by the CGH signal, we constructed receiver operator characteristic (ROC) curves using calls identified from the 2.1M SNP microarrays and confirmed by read-depth as our truth-set (**Figure S5**).

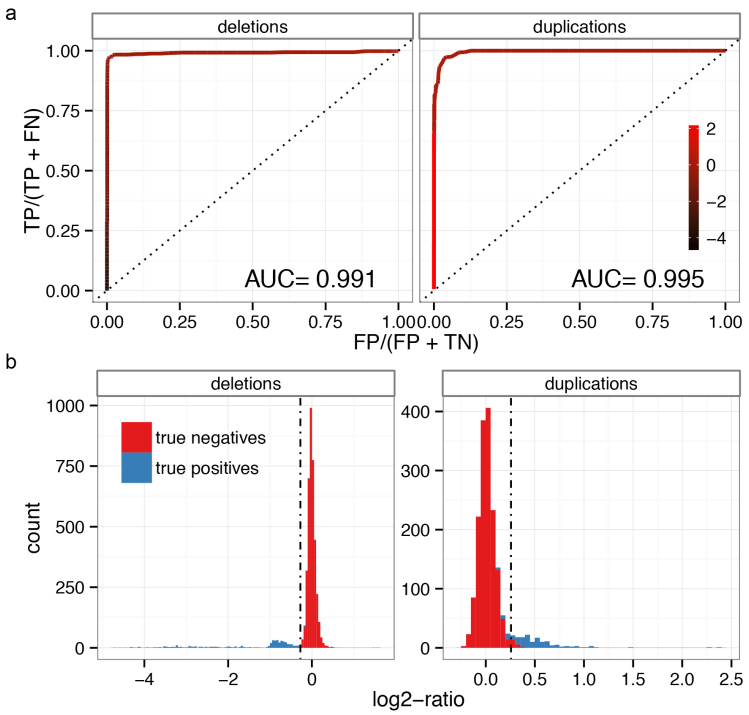

Figure S5: a) Receiver operator characteristic (ROC) curves for varying $\log_2$-ratio cutoff thresholds for CGH signals from deletions and duplications, respectively. Cutoff values are shown in red gradient. b) Distributions

of $\log_2$ ratios for true-negatives and true-positives of deletions and duplications, respectively, with the optimal precision/recall indicated by the dotted line.

ROC curves exhibited areas (area under curve, AUC) >0.99 for both duplications and deletions suggesting the aCGH signal provided a near-perfect classifier of true-positives and true-negatives. Cutoffs were finally chosen to optimize both the precision and the recall, -0.275 for deletions and 0.257 for duplications.

Finally, all calls in each of the 20 test individuals for which we were able to design at least five unique probes were assessed and classified as true-positives or false-positives (**Table S9**). Overall, the accuracy of deletions in the callset was 98.5% and the accuracy of duplications was 92.2%. The median per-sample accuracies were similar, 98.3% and 92.7%, respectively. Overall, the total accuracy of the callset was 97.5%.

Table S9: Callset accuracy.

	total calls	correct	fraction correct	per-sample median
deletions	3564	3511	0.985	0.983
duplications	606	559	0.922	0.927
TOTAL	4170	4070	0.976	-

Section S5: Properties of CNV callset and assessment of novelty

Genomic features and size distribution of CNVs

Amongst the 236 individuals assessed, we identified a total of 14,467 autosomal CNVs and 545 X-linked CNVs (**Table S10**). These CNVs intersected a total of 217.1 Mbp and 7.01% of the human genome with deletions making up 85.6 Mbp (2.77% of the genome) and duplications accounting for 136.1 Mbp (4.4% of the genome). In total, these CNVs intersected 1.84 Mbp of exonic sequence and 99.84 Mbp of segmentally duplicated sequence.

Table S10: CNVs broken down by their intersection with a particular genomic region. The number and Mbp of exonic and segmentally duplicated CNVs reflect the amount of exon-containing and SD-containing affected, respectively, not the total sum of the intersecting CNVs. For example, 636 CNVs intersect exons and this set represents 320 kbp of sequence.

	autosomal CNVs (Mbp)	X chromosome CNVs (Mbp)	exonic CNVs (Mbp)	segmentally duplicated CNVs (Mbp)
deletions	7233 (78.99)	278 (6.61)	636 (0.32)	331 (8.47)
duplications	7234 (129.62)	267 (6.46)	2093 (1.56)	4462 (96.93)
all	14467 (204.54)	545 (12.61)	2729 (1.84)	4793 (99.84)

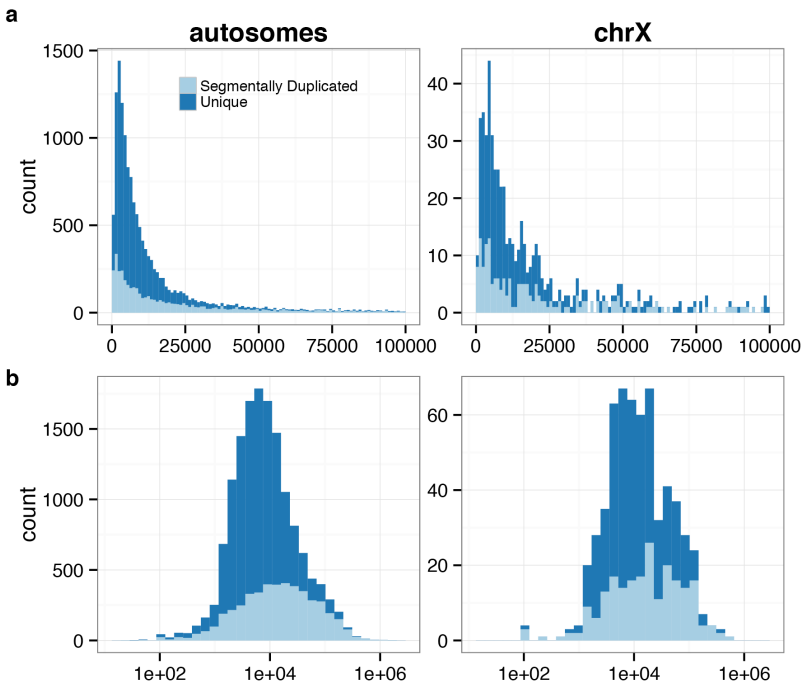

Figure S6: Size distribution of CNVs plotted for autosomes and the X chromosome in (a) linear space and (b) in log space.

The size distribution of CNVs (**Figures S6, S7**) is skewed toward smaller events with a median CNV size of 7,396 bp; 82.2% of events (12,338) are less than 25 kbp in size. While fewer CNVs were present on the X chromosome, the distribution of CNV sizes was similar. The median size of CNVs overlapping SDs was considerably larger (14,358 bp) compared to CNVs in unique space, which had a median size of 6,212 bp (P<2.2e-16). Previous analyses of copy number variation have been performed across HGDP individuals—many of which intersect our sample set *(22)*; however, these analyses were based on SNP arrays and, as such, only reported 396 large CNVs.

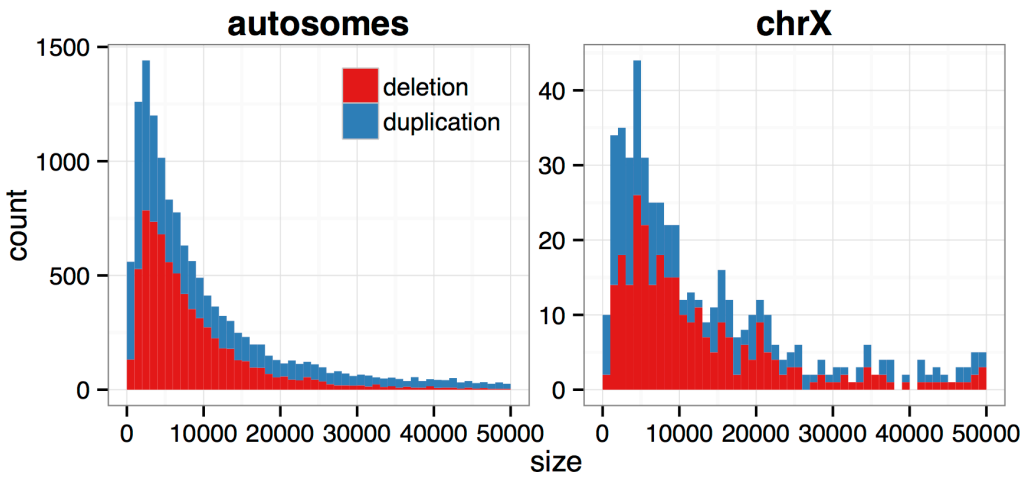

Figure S7: Size distribution of deletion and duplication CNVs plotted for autosomes and the X chromosome.

Comparison to previously published datasets

To estimate the additional yield of CNVs based on our more global diversity panel, we compared our callset to a set of validated CNVs generated by the 1000 Genomes Project *(27)* and to a set of calls generated by Conrad et al *(21)* limiting our comparison to calls encompassing at least 500 bp of unmasked sequence. Each of these studies focused on HapMap populations (YRI, CEU, JPN/CHB individuals) compared to the diverse worldwide focus of our study. We used a threshold of 30% reciprocal overlap between calls to determine if calls intersecting between the two sets were identical. We capture 77% (4894/6336) of calls made by Conrad et al, and 68% (4017/5895) of calls made by Mills et al. Additionally, 67-73% of calls we report are unique to our study (**Figure S8**). The disproportionately high number of CNVs identified in our study compared to both the Conrad et al. and Mills et al. callsets suggests our diverse set of individuals has identified a large number of rare and population-stratified CNV calls.

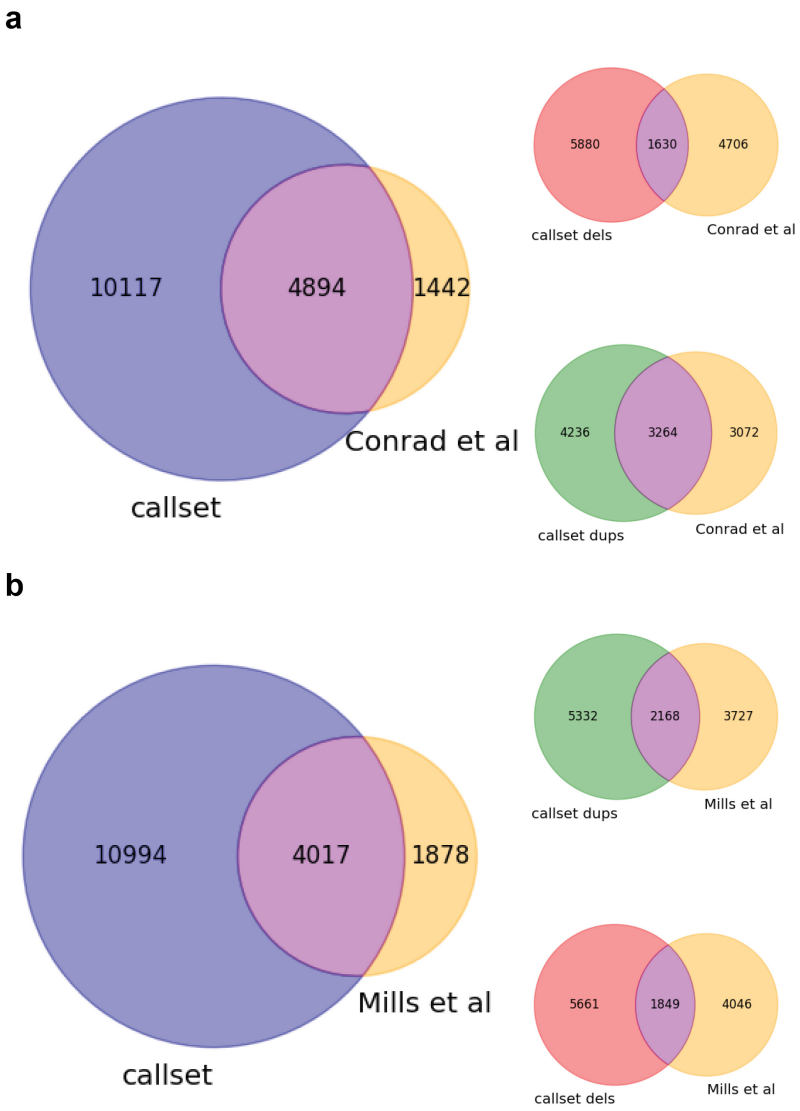

Figure S8: Intersection of CNVs from our callset with CNVs from Conrad et al. and Mills et al. Inset we subset our callset to deletions or duplications and compare to the total callsets of Conrad et al. and Mills et al.

We assessed the precision of breakpoints (**Figure S9**) of the calls that confidently overlapped between our callset and that of Mills et al. and Conrad et al. (70% reciprocal overlap). The median distance between breakpoints of overlapping calls was 443 bp and 537 bp for the Mills and Conrad datasets, respectively, with the peak of the distribution falling at 210 bp.

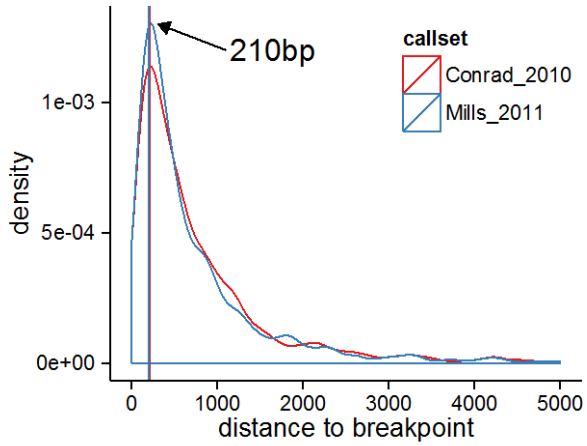

Figure S9: Distribution of the distance between breakpoints of overlapping calls in our dataset and those made by Conrad et al. 2010 or Mills et al. 2011. The median distance between inferred breakpoints was 443 bp and 537 bp for the Mills et al. and Conrad et al. callsets, respectively.

Section S6: Population genetic properties of CNVs

Allele frequency spectrum of bi-allelic events

Using simple bi-allelic 0,1,2 copy number states, we estimated the folded site frequency spectrum of all CNVs (**Figure S10**). The allele frequency spectrum (AFS) is skewed towards rare events, a pattern that we observed to a greater extent with larger calls suggesting selection against large deletions (**Figure S10**). We plotted allele frequency as a function of size (**Figure S11a,b**) demonstrating as deletions increase in size they rapidly become increasingly rare. A kernel density estimate of this trend is shown in **Figure S11c** demonstrating that the allele frequency of events drops precipitously as they increase in size.

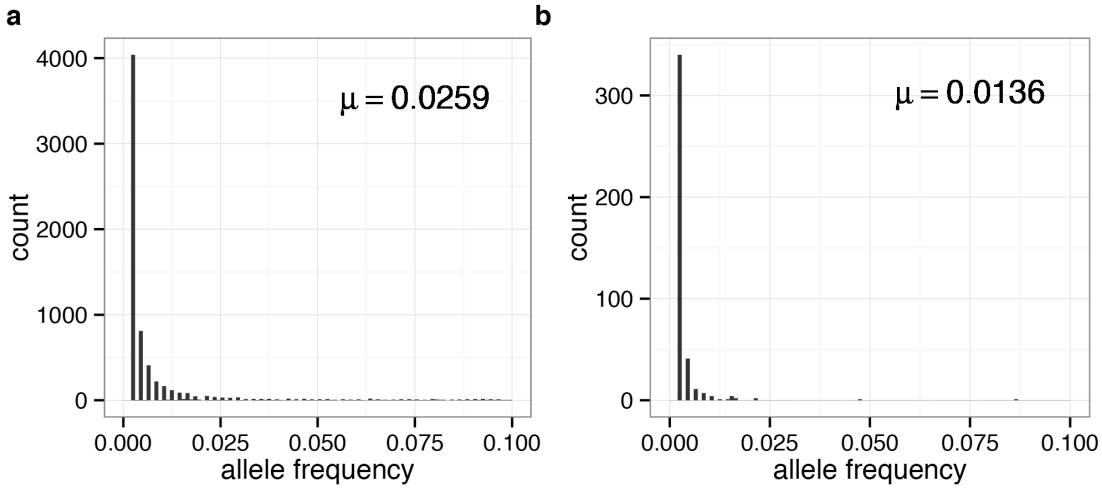

Figure S10: Folded site frequency spectra of all bi-allelic 0,1,2 copy number state calls (a) and CNVs >30 kbp (b) demonstrating a propensity of larger calls to be more rare than CNVs intersecting genes. Mean frequency (μ) is shown for each.

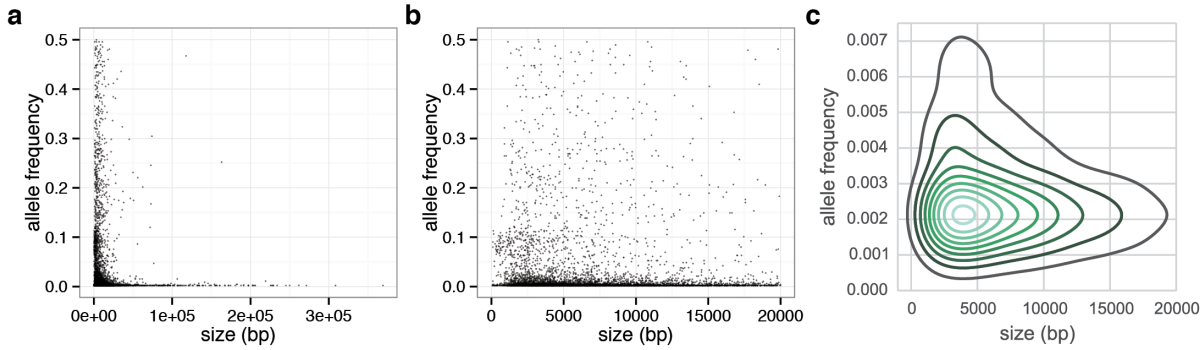

Figure S11: Allele frequency is plotted as a function of size (a) with a zoom of 0-20 kbp shown (b) and the two-dimensional kernel density estimate of the distribution of allele frequency and size for 0-20 kbp is shown (c).

To quantify the relationship between allele frequency and size, we modeled the folded site frequency spectrum of bi-allelic deletions as an exponential distribution, parameterized by the rate variable λ. For increasing size cutoffs, we identified the maximum-likelihood fit of an exponential distribution to the site frequency spectrum (i.e., $1/\bar{f}$ where $\bar{f}$ is the mean frequency) (**Figure S12a**). The λ rate estimators correlate positively as a function of the size cutoff (**Figure S12b**, p=1.153e-14) indicating that indeed CNVs of increasing sizes become increasingly rare.

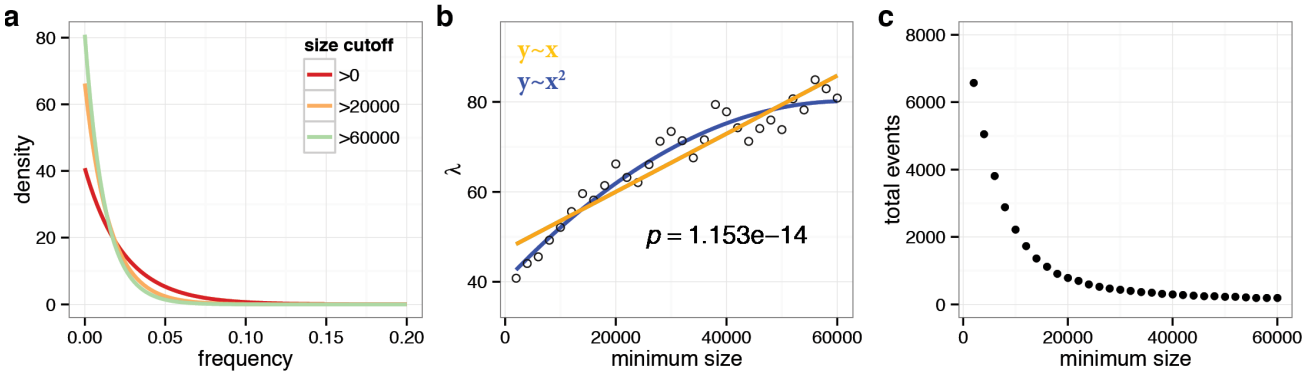

Figure S12: The relationship between the size of a CNV and allele frequency was explored by modeling the folded site frequency spectrum as an exponential distribution. The maximum-likelihood exponential distribution was estimated for the site frequency spectrum estimated at varying minimum CNV size cutoffs. Exponential distributions for 0 bp, 20 kbp and 60 kbp cutoffs are shown in *a*. The resulting λ estimators were then plotted as a function of the minimum size cutoff (b) exhibiting a significant positive correlation (p<1.153e-14). The number of events at each size cutoff is plotted in (c).

We reasoned that the AFS of deletions is skewed towards increasing rarity of larger events as these events are more likely to intersect functional elements and, thus, more likely to be selected against. If this is indeed the case, bi-allelic deletions intersecting genes should be rarer than those that are located between genes. To test this, we grouped bi-allelic deletions into those that intersected exons (genic) and those which do not (intergenic) (**Figure S13**). The mean allele frequency of intergenic bi-allelic CNVs was 0.0267 compared to 0.0172 for genic CNVs, demonstrating genic CNVs to be significantly more rare (p=1.842e-9, Mann-Whitney test).

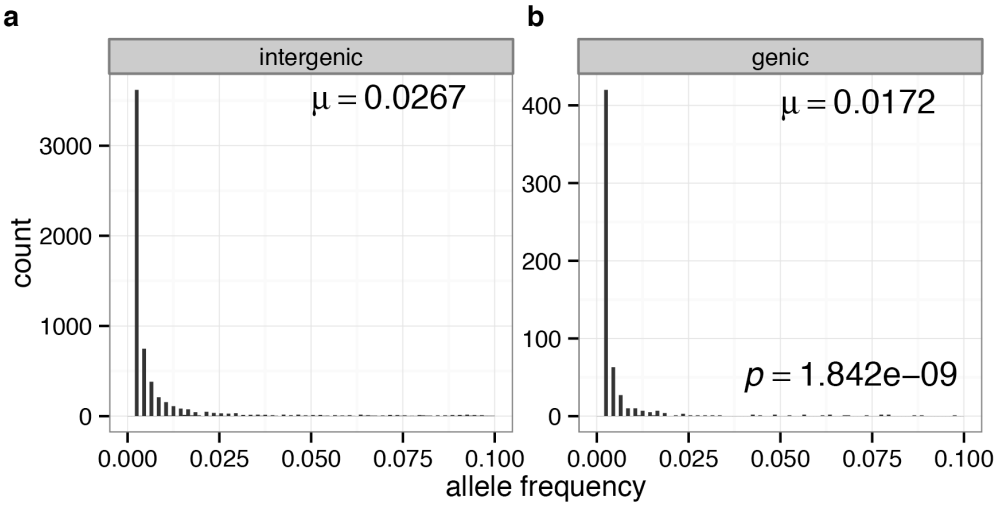

Figure S13: Folded AFS of all bi-allelic 0,1,2 copy number state calls for intergenic events (a) and events intersecting the exons of genes (b). Gene-intersecting events are significantly more rare (mean frequency of 0.0172 compared to 0.0267, p=1.842e-9).

We next analyzed the AFS of bi-allelic duplications, i.e., regions with 2,3,4 copy number genotypes. Strikingly, the relationship between the frequency spectrum and the size of duplications was very different from that observed for deletions (**Figure S14**). Only a very weak significant relationship between size and allele frequency was observed (p=0.04966), consistent with the reduced impact of selection on duplications compared to deletions. To further test this, we again stratified the AFS by events intersecting genes and those not intersecting genes (**Figure S15**). Of note, no significant difference was observed between the AFS of gene-intersecting duplications and intergenic duplications (P=0.181) providing strong evidence to suggest that deletions are deleterious and selected against while duplications are generally more neutral.

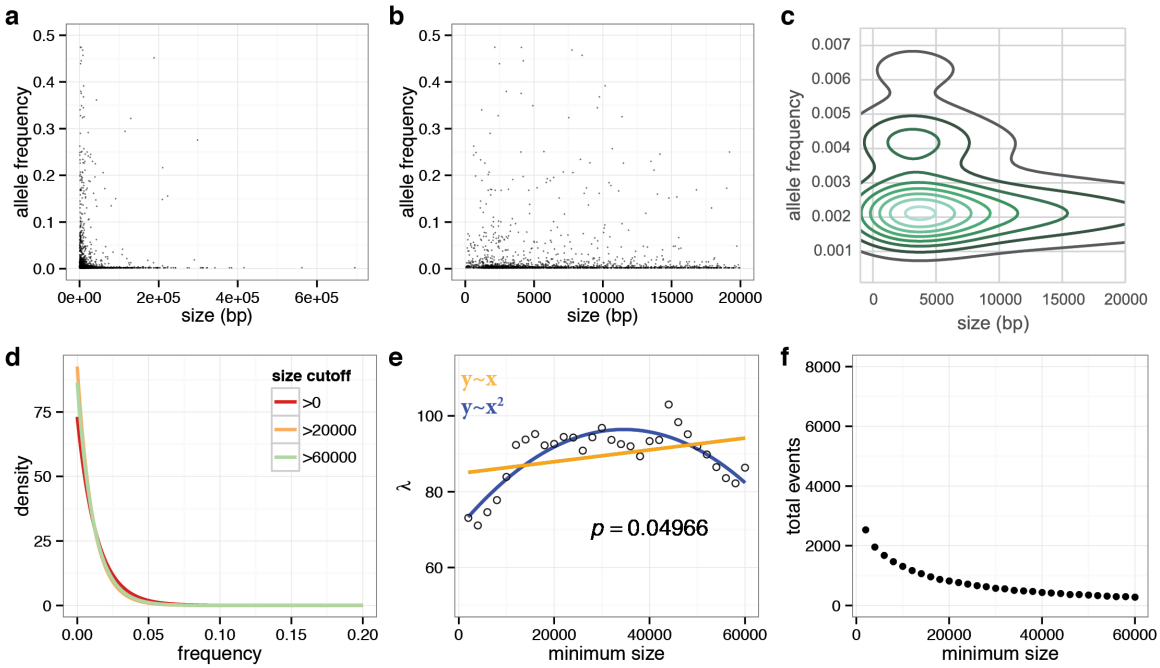

Figure S14: Allele frequency is plotted as a function of size (a) with a zoom of 0-20 kbp shown (b) and the two-dimensional kernel density estimate of the distribution of allele frequency and size for 0-20 kbp shown (c). Exponential distributions for 0 bp, 20 kbp and 60 kbp cutoffs are shown d) and λ estimators are plotted as a function of the minimum size cutoff (e); however, no significant correlations are observed (p=0.7842). The number of events at each size cutoff is plotted (f).

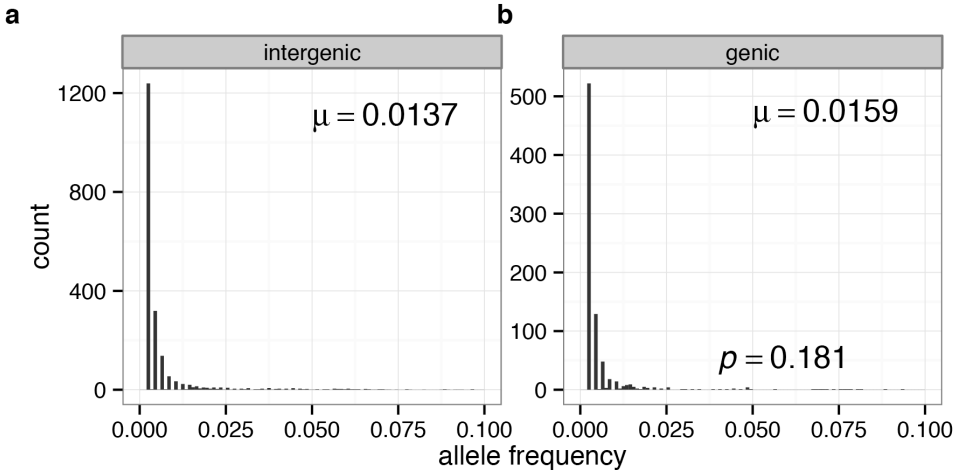

Figure S15: Folded AFS of all bi-allelic 2,3,4 copy number state calls for intergenic events (a) and events intersecting the exons of genes (b). Gene-intersecting duplication events show no significant difference in their frequency compared to intergenic events (mean frequency of 0.0142 compared to 0.0127, p=0.07925).

Human CNV diversity

Heterozygosity: To explore how human demography may have shaped the underlying diversity of CNVs in the genomes of different populations, we assessed the number of

heterozygous deletions (copy number 1) and the total number of bi-allelic CNVs (copy 0 or 1) in each individual of each population we assessed. As expected from an African origin of the human species, African individuals exhibited ~25% more heterozygous sites per individual on average than non-Africans (**Figure S16**). Individuals from the Americas exhibited the lowest diversity with the South American Karitiana, Surui, and Piapoco along with the Native American Pima peoples exhibiting the least diversity. Oceanic populations also exhibited very low levels of diversity.

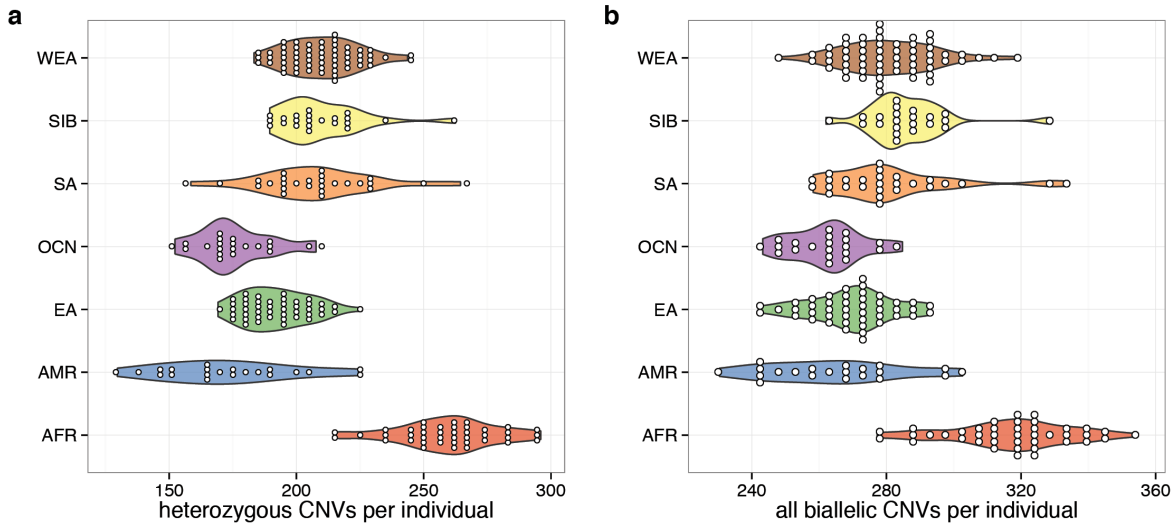

Figure S16: Violin plots of the distribution of heterozygous bi-allelic CNVs (a) and all bi-allelic CNVs per individual (b) overlaid with a binned dot-plot of counts for populations surveyed in this study (WEA – West Eurasian, SIB – Siberian, SA – South Asian, OCN – Oceanic, EA – East Asian, AMR – Americas, AFR – Africans). African populations exhibit increased diversity compared to non-Africans as a result of a human population bottleneck out of the African continent. Populations from the Americas in particular show decreased diversity in addition to Oceanic populations.

We compared the number of heterozygous bi-allelic deletions in each individual to the number of bi-allelic SNVs (**Figure S17a**) identifying the two metrics to be highly significantly correlated (R=0.876, p=4.658e-76). Heterozygous bi-allelic duplications were also significantly correlated with SNV heterozygosity though the correlation was significantly reduced (**Figure S17b**) (R=0.271, p=2.485e-5). We reasoned that this might be the result of a higher rate of mutability and thus increased rates of homoplasy in duplications compared to deletions. To test this, we binned duplications by those that intersect or are proximal to SDs (within 150 kbp) versus those that do not (**Figure S18**). SDs and proximal loci are known to be highly dynamic and susceptible to homoplasy (21). We found that duplication CNVs in unique space were more highly correlated to SNV diversity than duplication CNVs lying within and proximal to highly dynamic SDs (R=0.285 compared to R=0.173). They were also more correlated than those lying proximal to SDs though not intersecting SDs themselves.

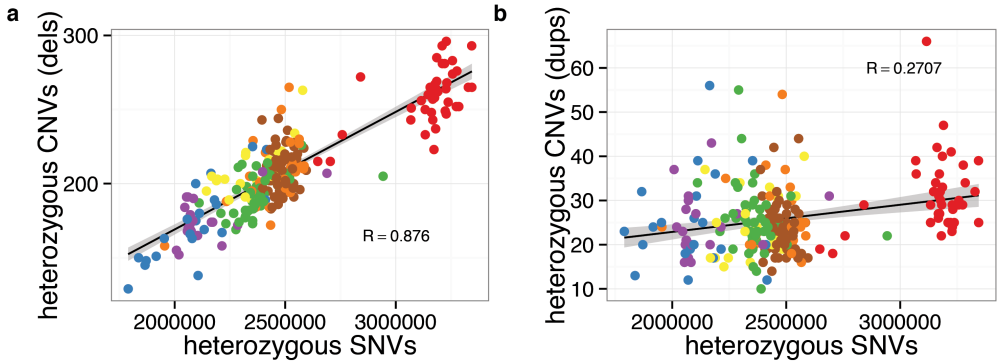

Figure S17: The number of heterozygous CNVs plotted as a function of heterozygous SNVs identified in the same genomes (a) and the number of heterozygous duplicated CNVs as a function of heterozygous SNVs (b). Colors represent continental populations and are identical to those shown in Figure S16.

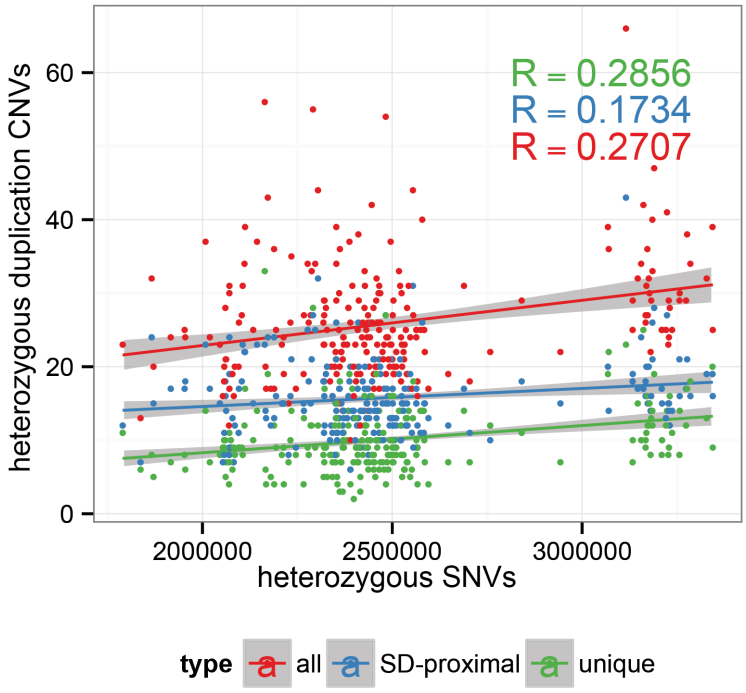

Figure S18: The number of heterozygous duplicated CNVs plotted as a function of heterozygous SNVs identified in the same genomes for all duplication CNVs (red), those proximal and intersecting SDs (blue), and those in unique space (green). Duplication CNVs in unique space, which are less likely to undergo homoplasy, are more strongly correlated with SNV diversity than those proximal to SDs.

We quantified diversity statistics for each of the populations assessed (**Table S11**) calculating an estimate of Waterson's theta (θ_w) based on the number of segregating deletion CNV sites observed in the chromosomes of the individuals from each of the populations. Commensurate with their increased diversity, the estimate of θ_w for African populations was

twofold that of the non-African populations. We also observed 1,772 African-specific CNVs (only in African individuals) compared to an average of 401 for each of the non-African populations. Of the 1,772 African-specific CNVs, 702 were observed in at least two individuals compared to an average of 55 in non-Africans.

Table S11: Summary statistics of bi-allelic deletion CNVs by population. *OCN-PAB represents the Oceanic subset of Papuan, Australian and Bougainville individuals.

population	n individuals	CNVs / individual (median)	heterozygous CNVs / individual (median)	segregating sites	population-specific CNVs (allele count ≥2)	θ_w / genome
SIB	23	285	205	1102	214 (30)	250.74
WEA	58	279	209	1728	688 (89)	324.42
OCN	21	263	173	1022	353 (84)	237.51
OCN-PAB	17	262	170	838	304 (81)	204.95
SA	27	279	208	1405	418 (43)	308.32
AMR	21	266	169	899	208 (25)	208.93
EA	45	271	191	1463	525 (59)	288.48
AFR	41	319	261	2663	1772 (702)	534.97

Principal Component Analysis: Using this same set of bi-allelic sites, we performed a principal component analysis (PCA) (**Figures S19, S20**). Principal component 1 (PC1) explained 6.18% of the total variance of the dataset and separated the continental populations from each other, primarily distinguishing the African individuals from the others. PC2 explained 3.94% of the variance of the dataset and primarily separated the West Eurasian individuals from African and East Eurasian and Oceanic populations. Oceanic populations form the furthest cluster away from West Eurasian populations along PC2. Siberian, South Asian and peoples from the Americas lie intermediate along PC2 between Western Eurasian and East Asian populations. PC3 distinguishes the Oceanic populations from all others while PC4 separates the peoples from the Americas. We noted a number of cases of single individuals clustering distal to their annotated population group. For instance, three African individuals extend into the West Eurasian cluster. Upon closer examination, two of these individuals are members of the Mozabite population from the Northern Sahara in Algeria and one is a Saharawi individual from the Western Sahara. The history of these populations and their close proximity to West Eurasians in the PCA strongly suggests West Eurasian admixture is likely influencing their placement on the PCA.

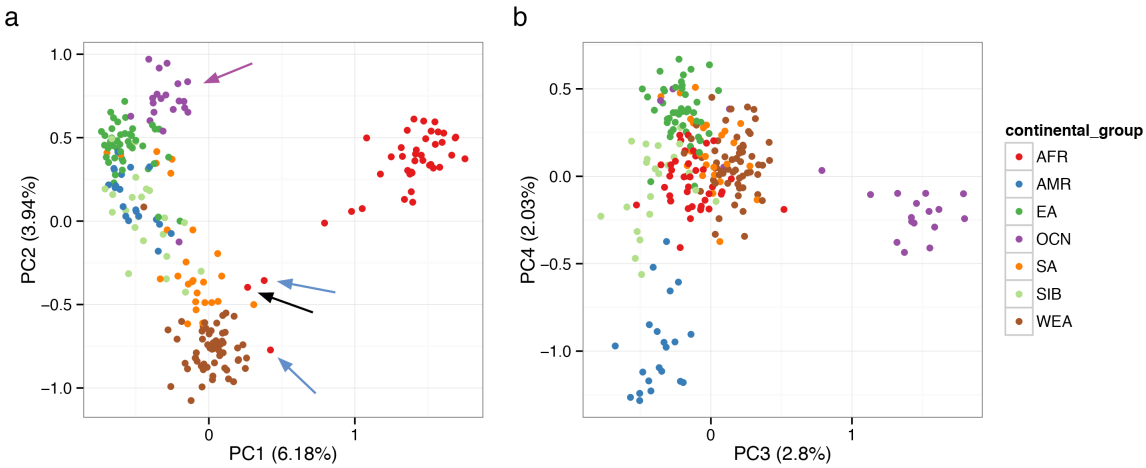

Figure S19: PCA of bi-allelic 0,1,2 deletion CNVs. Arrows note Saharawi (black), Mozabite (blue), and Oceanic Papuan/Australian/Bougainville (purple) individuals.

We performed a similar PCA on bi-allelic duplications identified in our callset (**Figure S21**). The first principal component reconstituted only loosely the general population structure observed for deletions amongst individuals. PC1 and PC2 explained 6.07% and 4.56% of the variance, respectively. Of note, the Oceanic populations (specifically the Bougainville, Australian and Papuan) are separated by PC2. PC3 and PC4 add virtually no additional information and do not appear to distinguish populations as was observed for deletions. Our results suggest that duplications, unlike binary deletions, provide very little ancestry information. We suggest that this difference in the PCA is due to reversion or homoplasy—i.e., duplications can more rapidly reestablish the diploid state by unequal crossing over between tandem duplicates—therefore creating identity by state as opposed to identity by descent. We hypothesize that population signal can only be observed when specific groups have experienced long periods of genetic isolation as is the case for the San and Oceanic Papuans.

Figure S20: Detailed PCA of deletion CNVs where each population group is labeled.

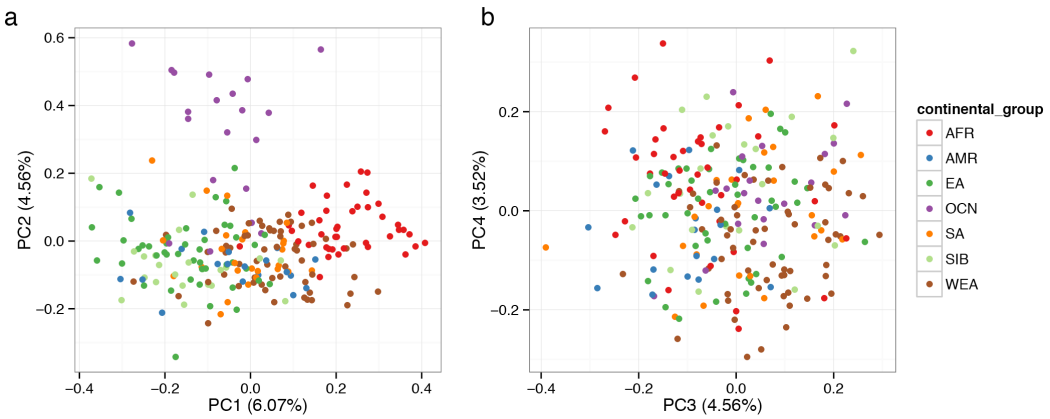

Figure S21: PCA of bi-allelic 2,3,4 duplication copy number state CNVs. PC1 and PC2 reconstruct the population structure of individuals, however, less clearly than deletions. PC3 and PC4 do not appear to distinguish population differences as was observed for deletions.

As an alternate approach to PCA, we also visualized the genetic relationship of deletions among individuals by employing the t-distributed stochastic neighbor embedding (t-SNE) method (http://lvdmaaten.github.io/tsne/), a machine learning-based dimensionality reduction technique (**Figure S22**). The t-SNE method has the advantage over PCA and other associated methods in that it captures both local and long-range structure in data. In other words, t-SNE allows visualization in two dimensions of structure elucidated in both PC1 and PC2 of PCA, in addition to further PCs. t-SNE, in general, results in greater discretization of individuals into their representative continental population groups. It clearly identifies, for example, the close clustering of Australian, Papuan and Bougainville Oceanic individuals in addition to the close relationship between individuals from Siberia and those from the Americas. South Asian individuals, in contrast, form two groups distributed between West Eurasian and East Asian populations. The South East Asian individuals fall into two clusters, with those from the Indian subcontinent largely clustering with West Eurasian individuals and individuals from further east, such as the Sherpa, clustering with East Asians. The application of t-SNE to the duplication calls did not yield informative clustering.

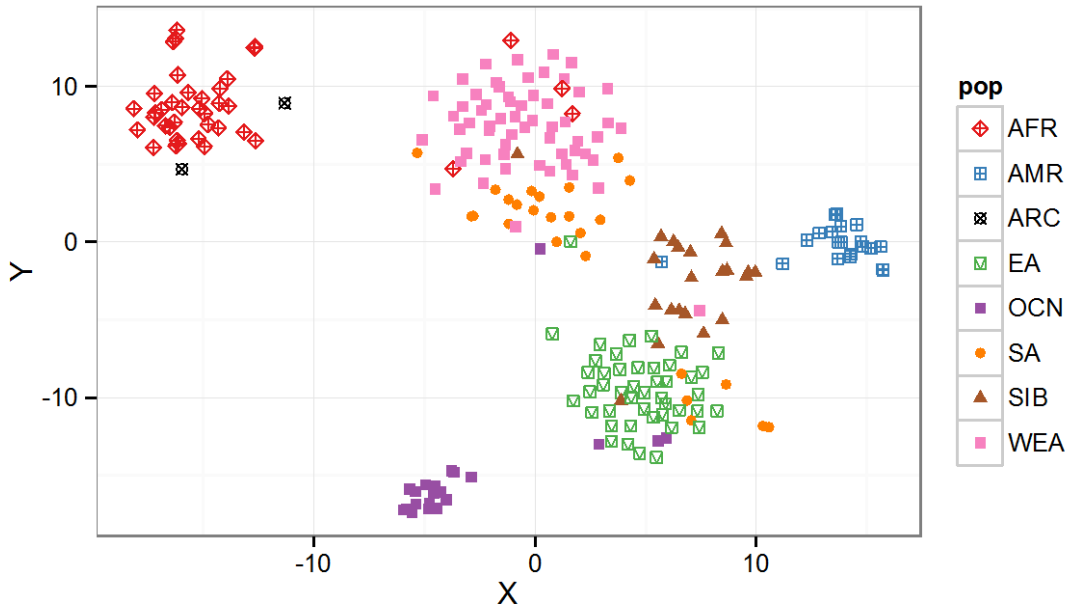

Figure S22: t-distributed stochastic neighbor embedding of deletion CNV genotypes.

As a final test to validate the population structure identified from bi-allelic CNVs and the potential effect of a smaller number of variants, we performed PCA on SNVs identified in the same individuals. We subsampled 166,044 SNVs coding the variants as 0,1,2 and performed PCA as described above. We additionally subsampled 5,000 SNVs and performed PCA to determine how the number of input variants affected the resulting PCA plot **(Figure S23**). These PCAs exhibited identical relationships among individuals as compared to those generated from deletion CNVs; however, the PCAs generated from 166,044 SNVs showed tighter clustering and greater discrimination than those generated from fewer SNVs.

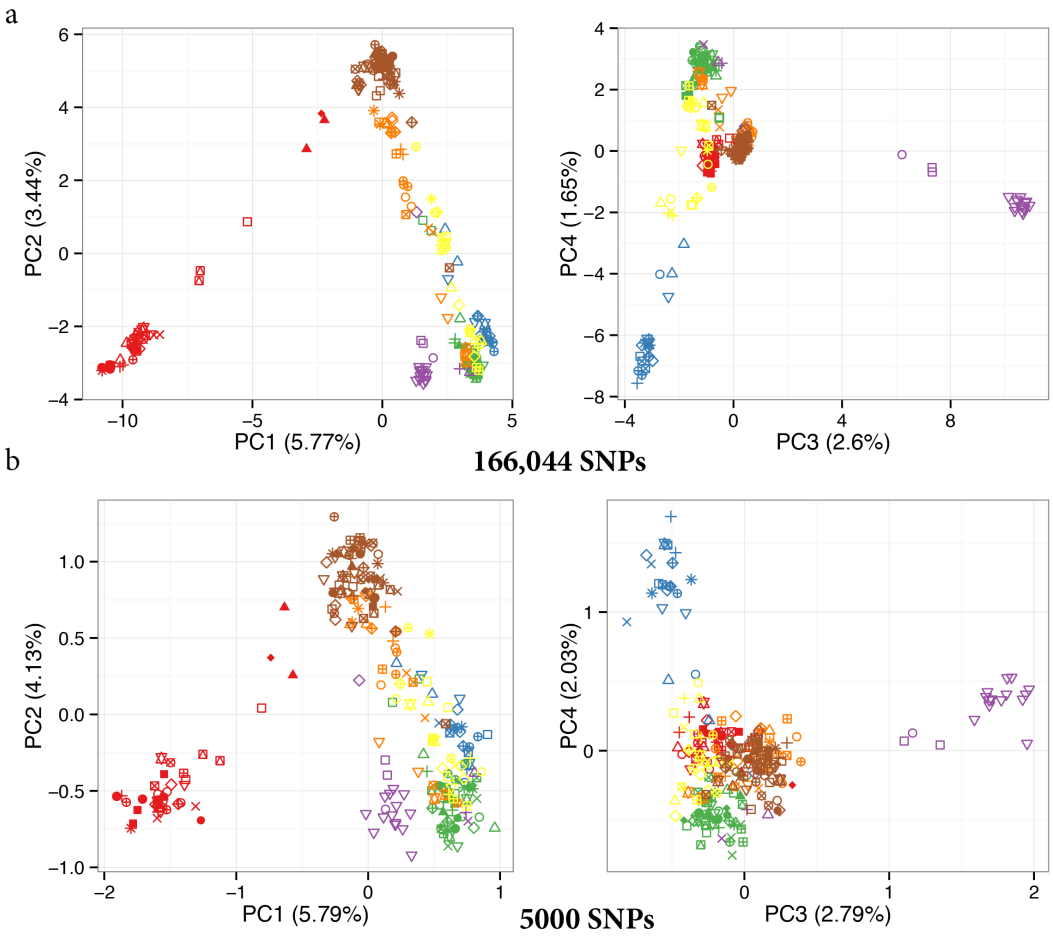

Figure S23: PCA plots generated from 166,044 SNVs (a) and 5,000 SNVs (b). Relationships among individuals are identical to those determined from CNVs. NOTE – legend for each population group is as shown in S6.11.

Phylogenetic Analysis: Using bi-allelic deletions we constructed unrooted neighbor-joining trees of all the individuals assessed in this study (**Figures S24 and S25)**. Branch length represents the number of CNV alleles. A consensus tree was also generated by subsampling 90% variants 100X (**Figure S26**). The phylogeny distinguishes each of the major populations in addition to defining some unexpected relationships between individuals and populations. Based on this genome-wide deletion pattern, the archaic Neanderthal and Denisova individuals emerge as a clear out-group to all humans and cluster as sister groups with much longer branch lengths. As expected, Africans represent the first and deepest branches with the San, Bantu, and Mbuti clustering together. Longer branches are clearly observed for each of the African individuals when compared to the non-African individuals, reflecting their increased deletion diversity.

The Papuan individuals form a single clade with the Australian and Bougainville. Of all out-of-African populations, this group stands out as having a distinct, common longer branch reflecting, perhaps, an extended period of genetic isolation. Interestingly, the Bougainville-Papuan-Australian group is distinct from three other Oceanic individuals—the Hawaiian, Igorot and Dusun—which cluster together with East Asians. The Igorot and Dusun are native peoples of the Philippines and North Borneo, respectively. Siberian, South American and Native American populations form a distinct clade as would be expected given the recent origins of Native American populations from Northern Asia. Additionally, two Native American individuals, Cree and Chipewayan, actually cluster more closely with the Siberians than other Amerindians—an observation consistent with the PCA. Similarly, two East Asian individuals representative of the Oroquen and Hezhen cluster with the Siberian group. We find that the Uygur samples, in general, do not form a single clade but instead are distributed among South East Asians, Europeans and Amerindians consistent with the PCA. Three African individuals—two Mozabites and a Saharawi individual—group more closely with individuals from the Middle East, such as the Palestinians, Yemenites, Jordanians, Druze and Bedouins. Throughout the tree there are individuals that do not cluster perfectly with their continental assignment. Although this may be due, in part, to long branch-length attraction, in most cases this atypical clustering is consistent with other analyses and unique aspects of their population history.

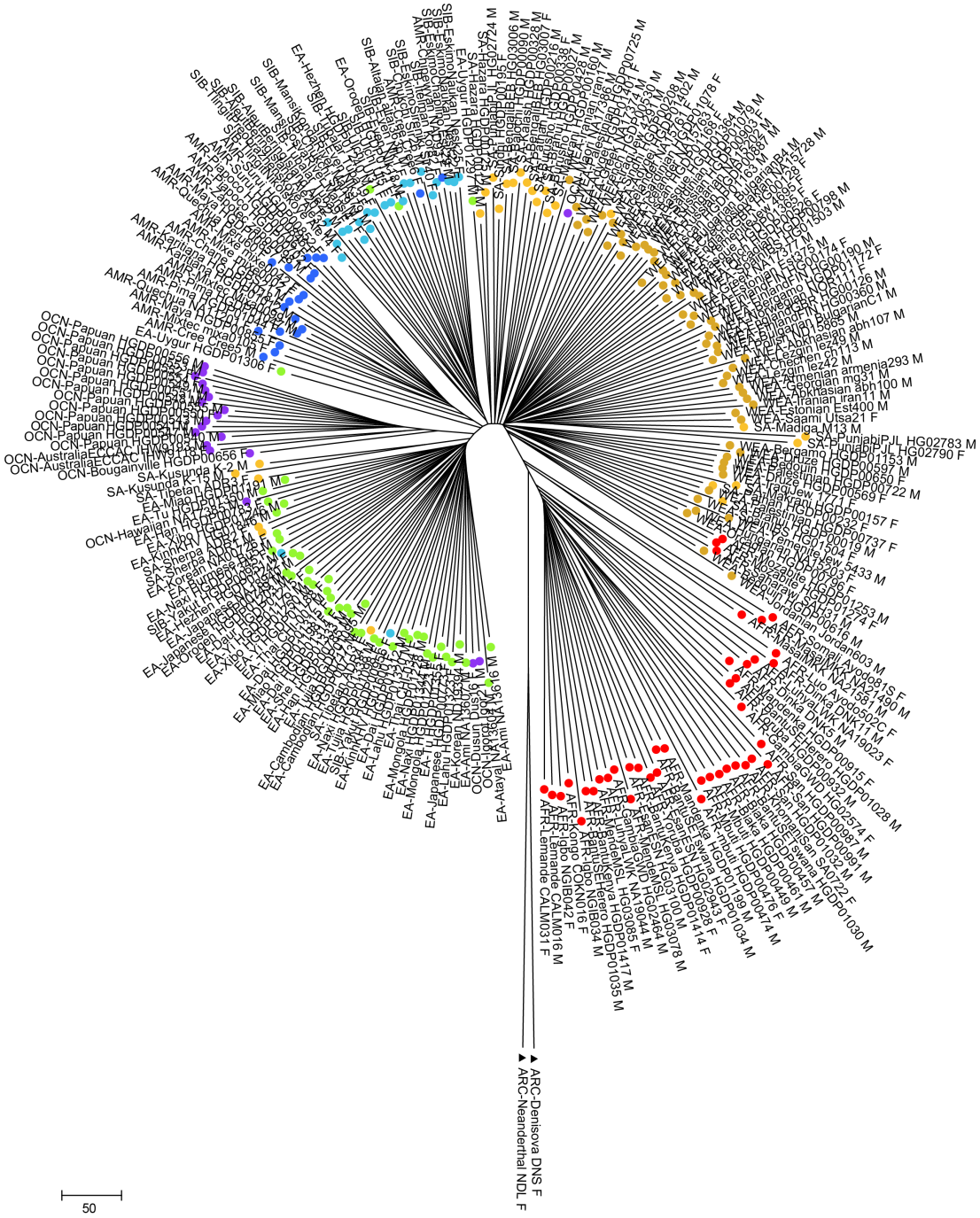

Figure S24: Neighbor-joining tree generated from bi-allelic deletion CNVs. Neanderthals and Denisovans emerge as sister out-groups to all humans with much longer branch lengths.

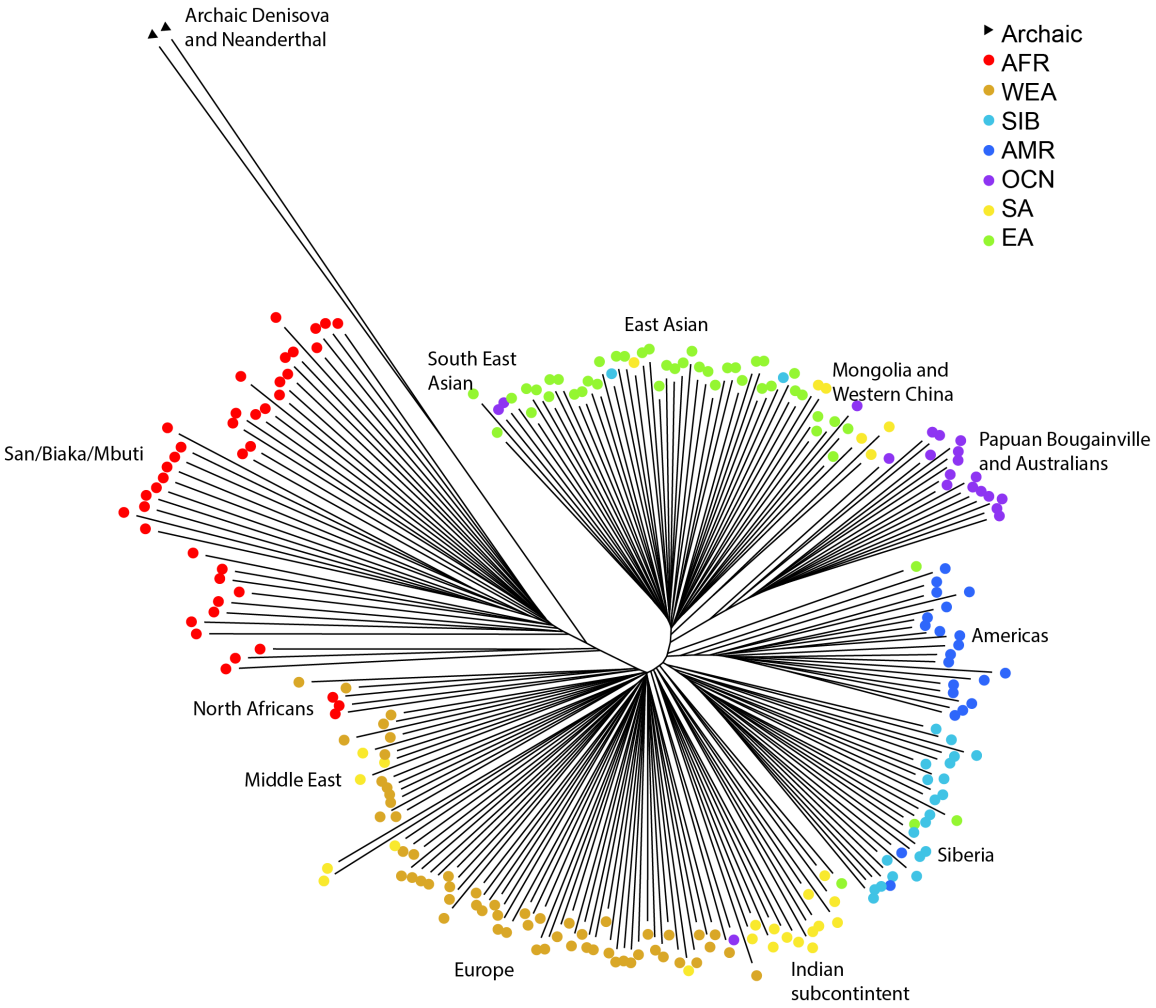

Figure S25: Neighbor-joining tree without labels generated from bi-allelic deletion CNVs with clades associated with geographic regions highlighted.

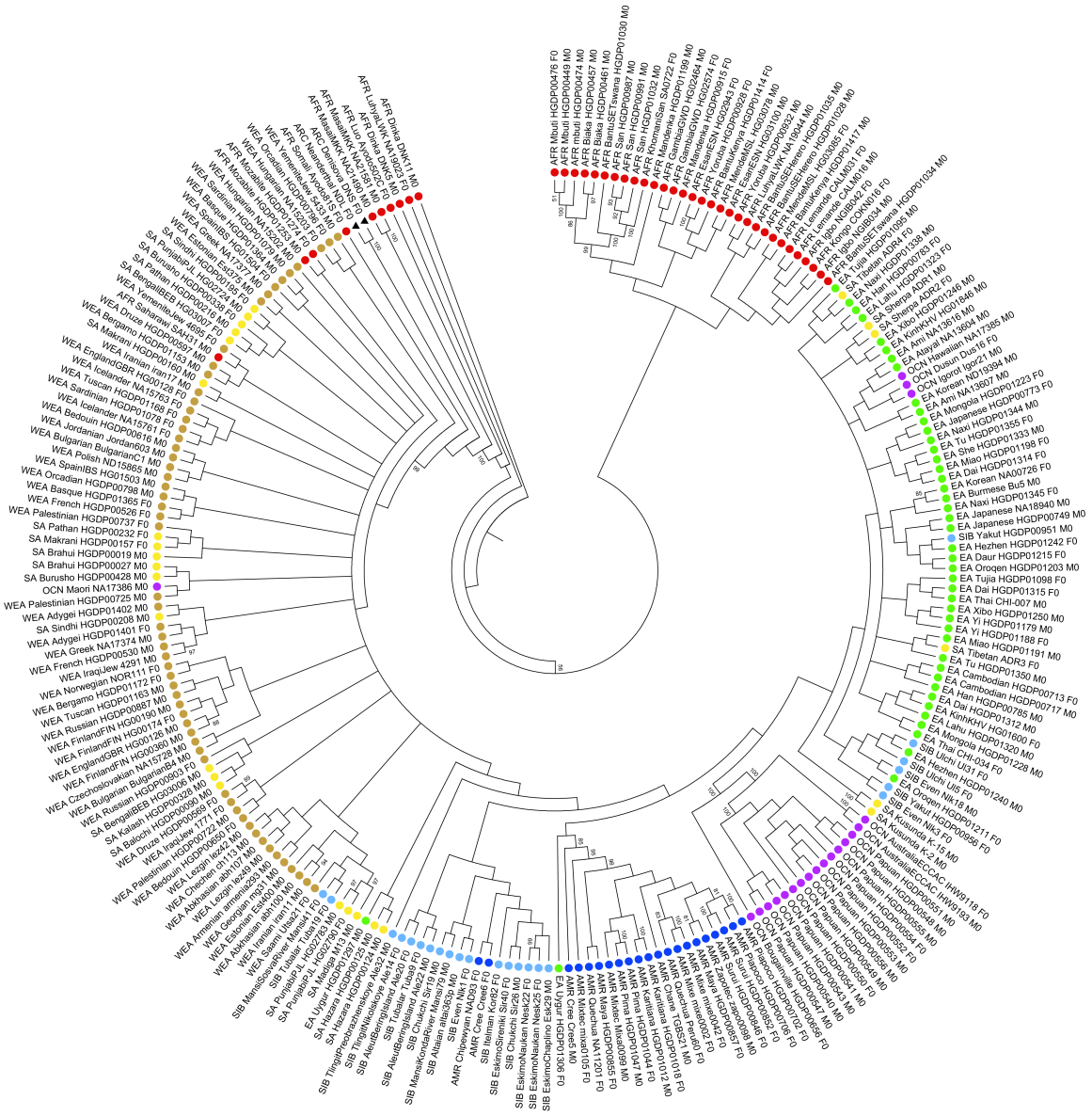

Figure S26: A consensus tree with bootstrap values reported for branches with >80% support.

Section S7: CNVs intersecting genes and the density of deleterious mutations

In our callset we identified 2,437 CNVRs intersecting exons. The distribution of allele counts of these exonic CNVs is skewed towards singletons and rare events, with the distribution of deletions tending to significantly lower frequency events compared to duplications (p=1.245e-5, **Figure S27**). These trends confirm previous suggestions *(13,21)* that deletions are more deleterious than duplications and thus may be under stronger selection.

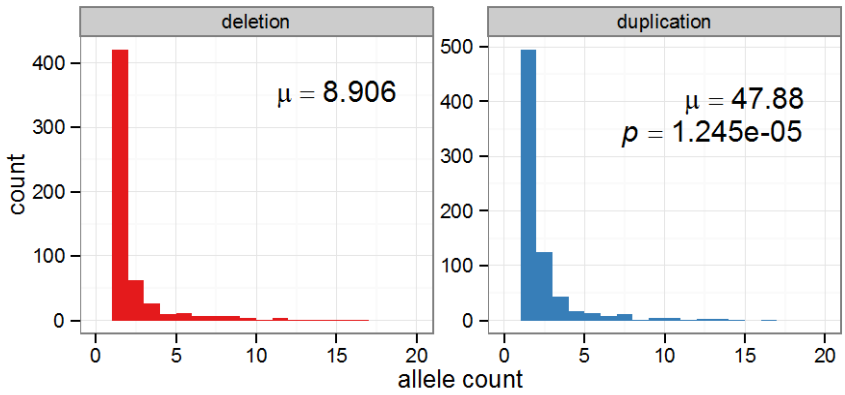

Figure S27: Allele count of CNVs intersecting exons for deletions and duplications (0,1,2 and 2,3,4 genotypes, respectively). Exonic CNVs are rare with most occurring as singletons. Exon-intersecting deletions are fivefold more rare than exon-intersecting duplications (mean of 8.9 alleles compared to 47.9, respectively, p=1.245e-5, Wilcoxon rank sum test).

We also compared the distribution of intergenic to putatively more deleterious genic CNVs among the different populations assessed in this study. Collectively, individuals harbor a mean of 19.2 exon-intersecting deletions per genome; however, African individuals exhibit more exonic deletions on average than non-African individuals with a mean 22.4 deletions compared to 18.6 in non-Africans, a 1.2-fold enrichment, and a mean total number of 26.1 exonic deletion alleles compared to 22.1 in non-Africans, commensurate with increased African diversity (**Figure S28b**). Exonic deletions also tended to be longer than non-exonic deletions (mean of 18.8 kbp versus 10.9 kbp, P=0.0001, Wilcoxon rank sum test), likely simply because longer elements have a greater chance of intersecting a functional element.

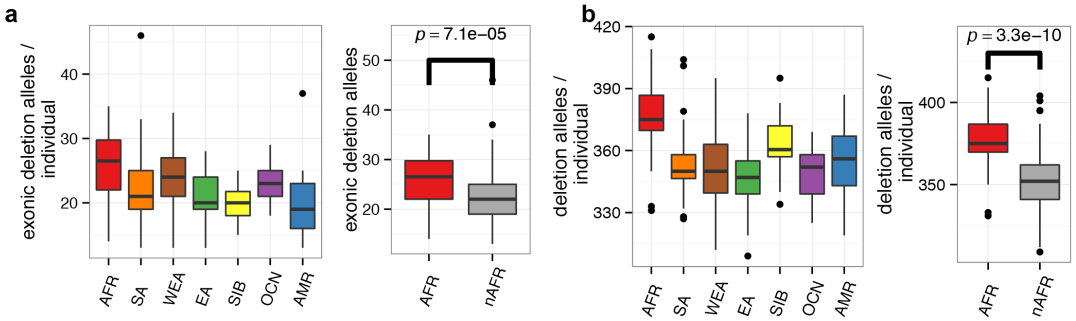

Figure S28: Boxplots of the number of exonic and intergenic deletion alleles per individual in different human populations are plotted in a and b, respectively, demonstrating more putatively deleterious deletions in African individuals.

We next investigated how the distribution of CNVs duplicating exons compares to the distribution of CNVs deleting exons among different populations. Collectively, individuals exhibited an average of 95.7 duplications intersecting genes with African individuals exhibiting

only a slight elevation of 98.4 exonic duplications on average and non-African populations exhibiting 95.2 exonic duplications on average. Africans harbored on average 174.8 exonic duplication alleles compared to 171.2 in non-Africans (**Figure S29a**). Exonic duplications, similar to deletions, also tended to be longer than non-exonic duplications (mean of 36.8 kbp versus 17.6 kbp, P<2.2e-16, Wilcoxon rank sum test), again suggesting longer elements simply have a greater chance of intersecting a functional element.

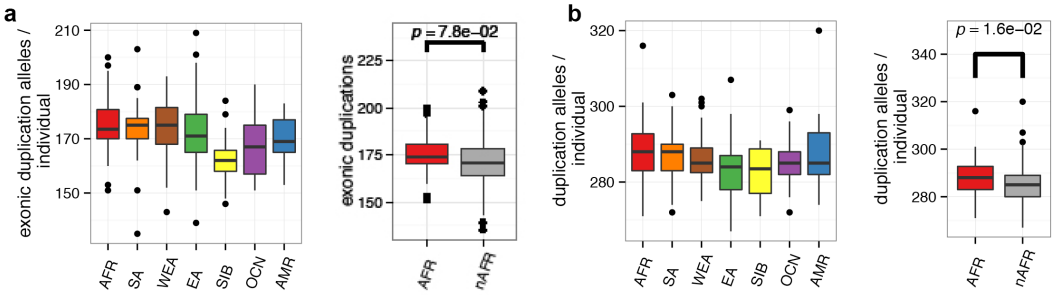

Figure S29: Boxplots of the number of exonic and intergenic duplication alleles per individual in different human populations are plotted in a and b, respectively, demonstrating no difference between African and non-African individuals.

Section S8: The ancestral human genome

The diversity panel assessed in this study provides a unique opportunity to examine novel genomic sequence segregating in human populations that is not included in the human reference genome. To identify putative sequences absent from the human reference genome but segregating in human populations, we first aligned chimpanzee and orangutan reference sequences (panTro3 and ponAbe2) to the human reference to identify regions absent from the human reference as described previously *(13)*. In total we identified 20,373 nonredundant regions >500 bp encompassing 40.7 Mbp of sequence absent from the human reference but present in nonhuman primate genomes. Sequences were masked with RepeatMasker and Tandem Repeats Finder and regions with <500 bp of unmasked sequence were discarded resulting in 9,666 loci encompassing 27.96 Mbp. Raw shotgun reads from each genome in our diversity panel were then mapped to these sequences in addition to a set of conserved copy number 2 chimpanzee, gorilla and orangutan control sequences. Using these control sequences, copy numbers were estimated for each locus as described above.

Of the 9,666 loci analyzed, 6,341 (15.8 Mbp) were deleted (copy 0) in all individuals assessed (**Table S12**) indicating that they were lost in the human lineage after separation from the African great apes before the diaspora of modern day humans. Of the remaining 3,325 loci, 728 (4.22 Mbp) were duplicated (>3 copies), all of which were copy number variable, while

571 (1.56 Mbp) were found to be segregating with bi-allelic deletions among the individuals assessed. The remaining 2,026 loci (6.2 Mbp) were copy number 2 in all individuals assessed. We include the genotypes and locations of these 3,325 loci as **Table S1**.

Table S12: Sequence absent from the human reference genome assayed across diversity panel.

type	loci	base pairs (kbp)
fixed deletions	6341	15843.86
fixed copy 2	2026	6243.06
duplicated	728	4216.2
segregating bi-allelic deletions	571	1555.88
total	**9666**	**27859.01**

We assessed the 571 segregating loci among individuals (**Figure S30**) and population groups (**Table S13, Figure S31**). African populations were much more likely to exhibit the presence of a site while all other populations showed no evidence for the variant (allele frequency 0 in all other populations). We additionally identified 11 sites (18.6 kbp) specifically present in Neanderthals but not in any humans and 33 sites (73.5 kbp) specifically present in Denisovans yet absent from all other humans. We identified an additional three sites present in both Denisovans and Neanderthals though not in any of the human populations assessed and 17 sites absent from both Denisovans and Neanderthals but present in humans. Both the archaic Ust'Ishim and Loschbour individuals exhibited two sites not found in any other extant human populations.

Table S13: Sequences absent from the human reference genome and population specifically, copy number variable, present, or absent. Africans exhibit the most population-specific copy number variable sites and the most population-specific present sites, i.e., the most sites that are absent from all other populations (allele frequency, AF=0).

		Population Specific	
population(n)	**CNV**	**presence**	**absence**
WEA(57)	6	11	0
SIB(23)	0	7	0
EA(45)	0	6	0
SA(27)	2	5	0
OCN(4)	0	0	0
OCN-BAP(17)	6	2	0
AMR(21)	2	2	0
AFR(33)	14	8	0
Biaka/San/Mbuti(8)	4	0	0
UstIshim(1)	2	0	4
Stuttgart(1)	0	6	5
Loschbour(1)	2	2	5
Denisova(1)	33	2	22
Neanderthal(1)	11	1	10

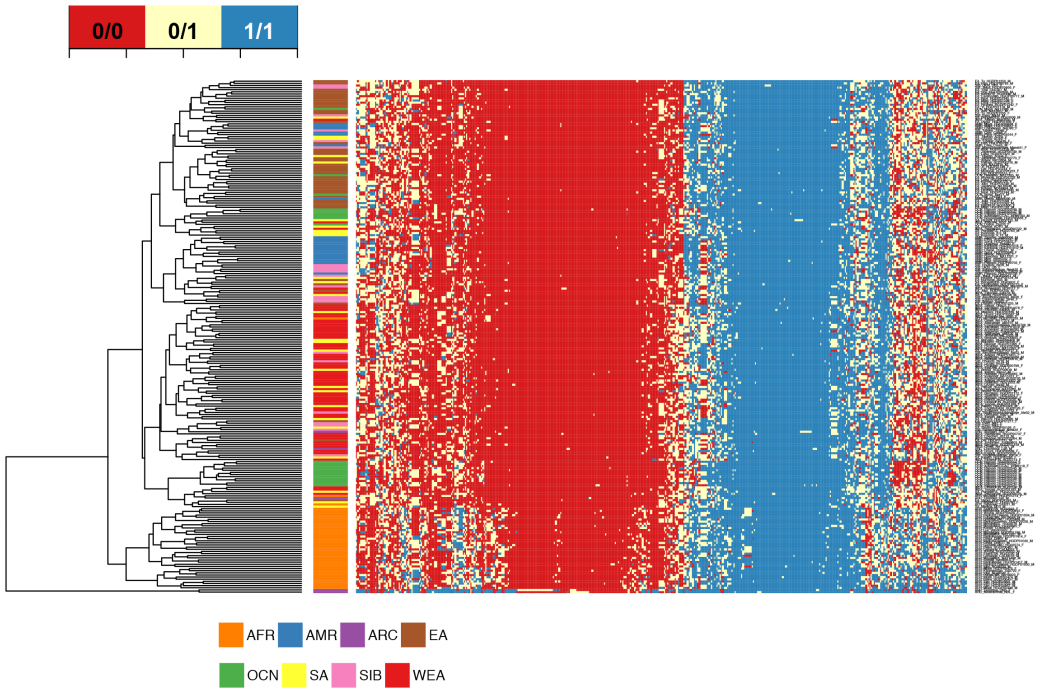

Figure S30: Genotypes of all segregating sequence absent from the reference genome amongst all individuals assessed in this study and hierarchically clustered. Colors along the rows to the right of the cladogram represent the continental population designation of the individual adjacent along the row.

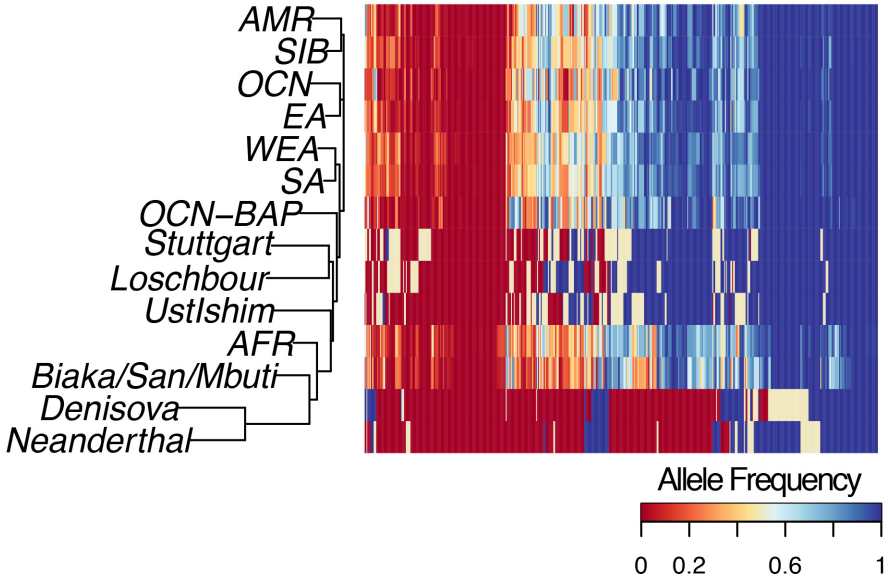

Figure S31: A heatmap of the allele frequencies of segregating sequence absent from the human reference genome clustered by neighbor joining according to the cladogram on the left.

Section S9: CNV load comparisons between population groups

We compared the difference in the deletion and duplication load between African and non-African populations as defined as: $L(Afr) - L(nAfr) = \sum_{\forall i} P_{Afr}(i) - \sum_{\forall i} P_{nAfr}(i)$ where $P_{Afr}(i)$ is the derived allele frequency of a variant i among African individuals (and analogously for *nAfr,* non-African individuals). The copy number of all sites was estimated from data generated from 23 chimpanzee genomes *(13)*; however, in all cases where the chimpanzee allele was not 2, the underlying CNV exhibited different breakpoints than the human CNV and thus the ancestral copy number was assumed to be 2. To obtain a confidence interval and p-values of the difference in load between African and non-African individuals, we performed 10,000 block-bootstraps, dividing the genome into 5 Mbp non-overlapping bins and sampling with replacement.

The difference in CNV load between African and non-African populations was computed originally using only calls made against the human reference genome GRCh37. It was repeated including sequences absent from the reference (see above section, **Figure S32**). Notably, for deletions, Africans exhibited a significant excess deletion load only when compared to the reference. When ancestral sequences absent from GRCh37 were included, we observed no difference in CNV load for deletions. In the case of duplications, no difference in load was observed between Africans and non-Africans in either analysis.

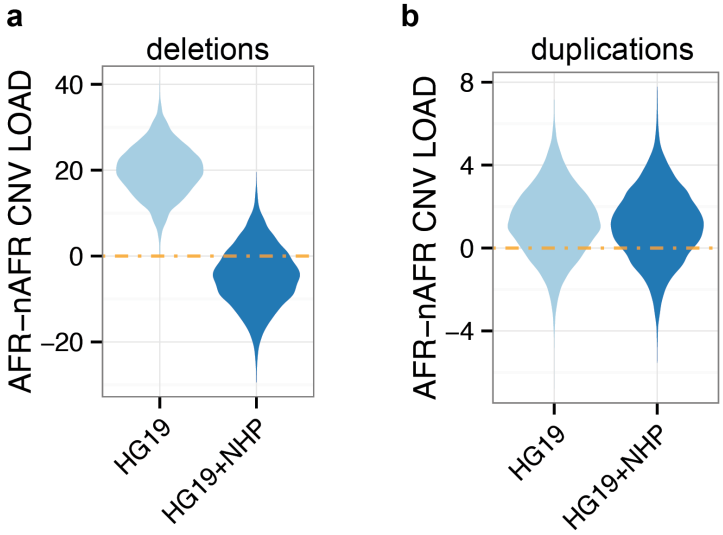

Figure S32: The distribution of differences in CNV load between Africans and non-Africans for deletions (a) and duplications (b) (10,000 block-bootstrap replicates). The plots compare the difference when dependent only on the reference (GRCh37/HG19) in contrast to a pan-genome where ancestral sequences have been included in the analysis (HG19+NHP), and where the significant difference disappears.

We further tested if there was any size dependence in the CNV load difference between Africans and non-Africans computing the distribution of load differences for all CNVs above specific size thresholds. Even as the minimum CNV size threshold was increased, no significant difference was observed in the load difference between African and non-African individuals (**Figure S33**). Similar analyses were performed between all continental populations No significant differences were observed.

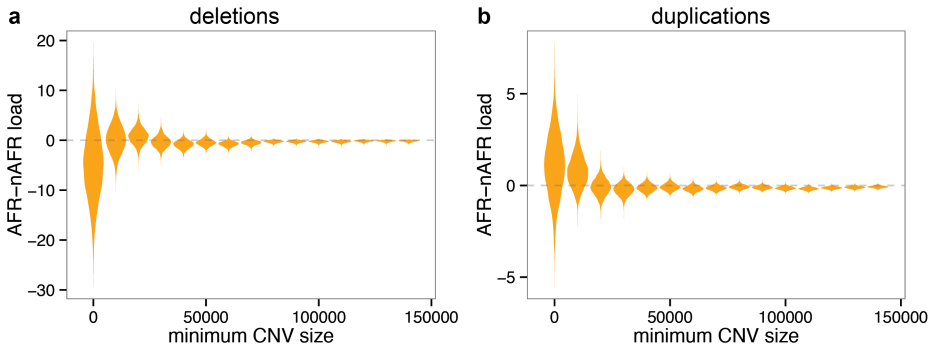

Figure S33: The distribution of differences in CNV load between Africans and non-Africans computed from 10,000 block-bootstrap replicates for all CNVs greater than specific size thresholds for deletions (a) and duplications (b).

Although no difference in the CNV load was identified between populations, we also examined if the relative contribution of base pairs differing between individuals as a result of CNVs varied when compared to SNVs. We then calculated the pairwise number of base pairs differing between all pairs of individuals contributed by CNVs and SNVs, respectively. As expected, the number of pairwise base-pair differences between individuals was highest within African populations and between African populations and other populations (**Figure S34**). Interestingly, the ratio of base pairs contributed by CNVs to those contributed by SNVs was consistently higher among non-African individuals (**Figure S35**).

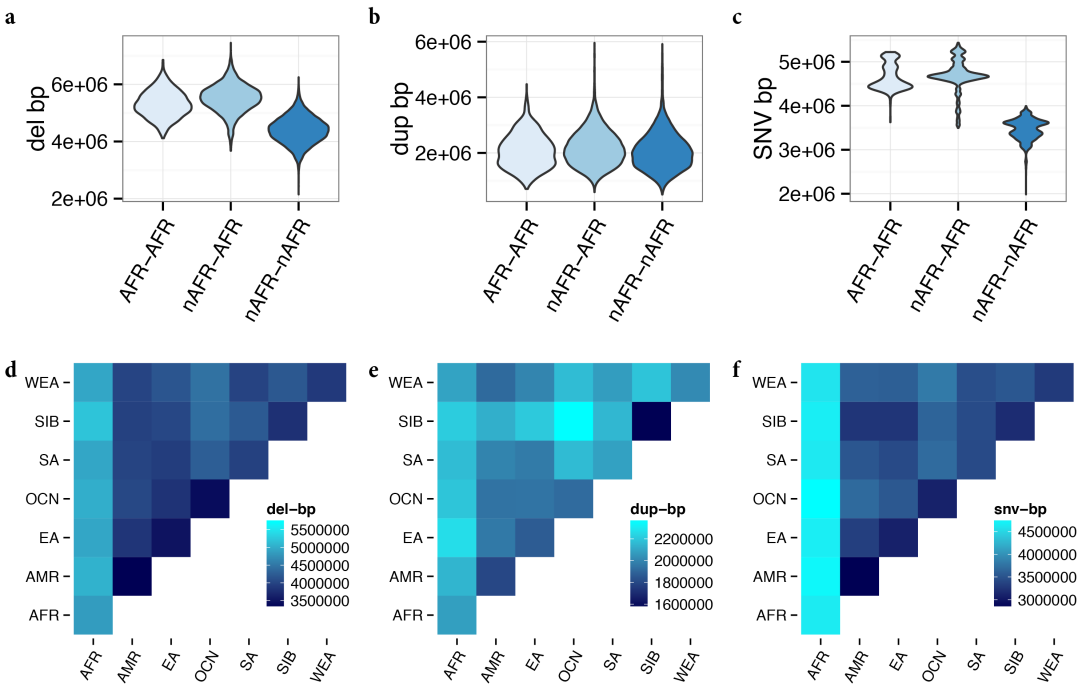

Figure S34: The number of base pairs differing between individuals from African and non-African populations are shown as violin plots for deletions (a), duplications (b) and SNVs (c). The mean pairwise base-pair differences plotted between continental populations as a heatmap for deletions (d), duplications (e) and SNVs (f).

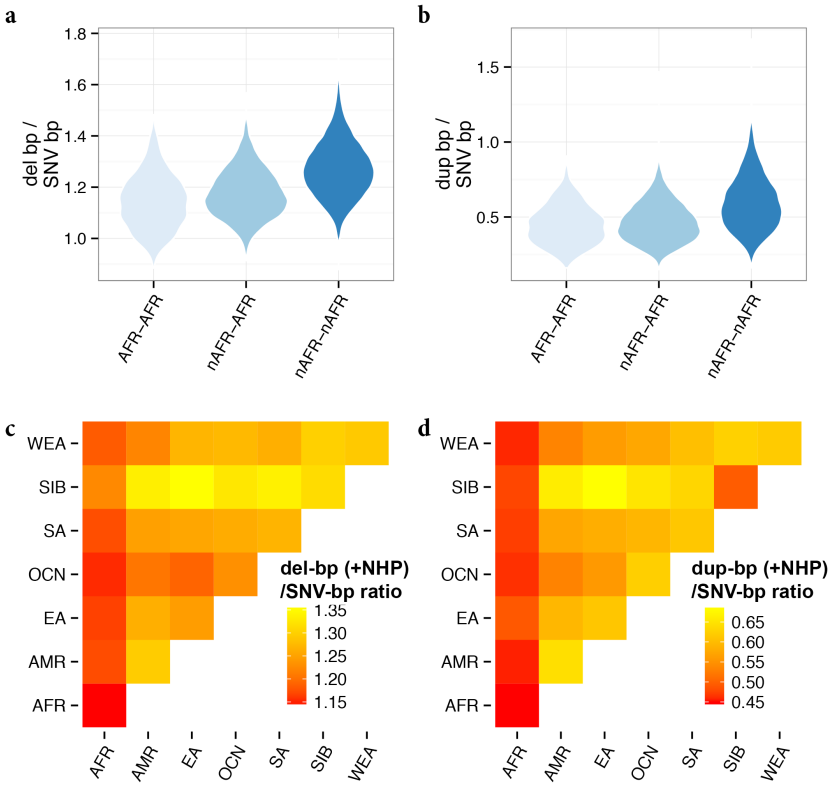

Figure S35: The ratio of the number of base pairs differing between populations contributed by CNVs compared to SNVs plotted as violin plots for deletions (a) and duplications (b) and plotted in heatmap form for intercontinental population means for deletions (c) and duplications (d).

Section S10: Population-stratified loci

We identified population-stratified CNVs using the Vst statistic (29), a measure of the variance of a CNV between populations, for each locus computing pairwise Vst values between all populations and retaining the highest Vst and associated population-pair (**Table S3,** see separate Excel sheet, **example Figures S36, S37**). The stratification of adjacent SNVs flanking each CNV was quantified using the Weir and Cockerham estimator of Fst (fixation index) implemented in VCFtools. As overlapping CNVs and CNVs mapping to duplicated loci would be over counted, we also constructed a collapsed table of Vst for CNVRs merging such sites and reporting the maximum Vst and population pair (**Table S3,** see separate Excel sheet). Analyzing the distribution of Vst values for bi-allelic deletions, bi-allelic duplications and multi-allelic duplication CNVRs (**Figure S38**), we found duplications to be more stratified than deletions; however, this signal was driven by multi-allelic duplications. When multi-copy duplication CNVs mapping to SDs were excluded, bi-allelic deletions were more stratified than bi-allelic duplications.

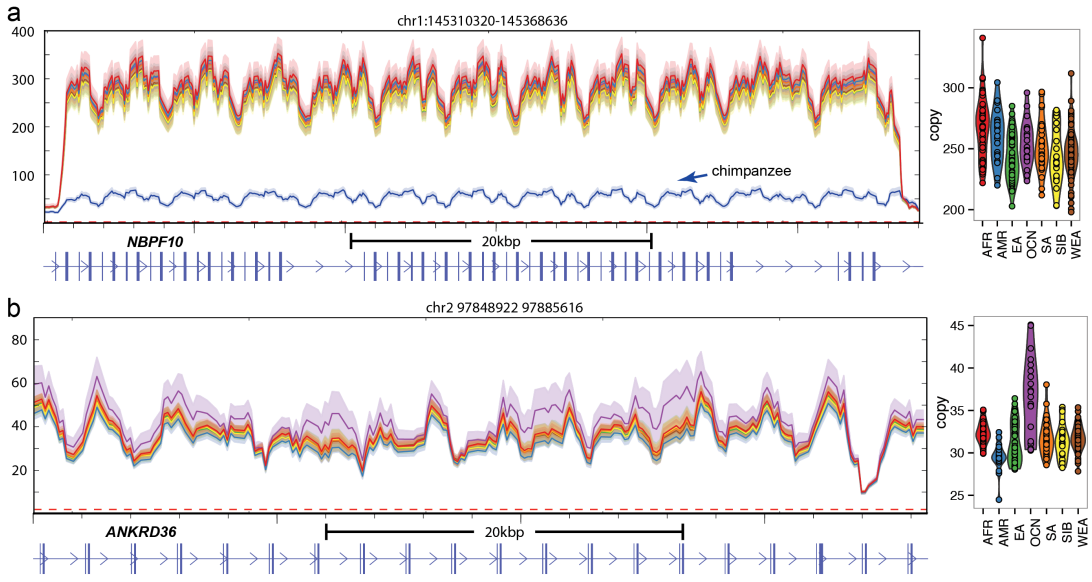

Figure S36: Mean copy number per population (with standard deviation shaded) estimated over the DUF1220 repeat domain of *NBPF10* (a) and over the repeat domain of *ANKRD36* (b) with adjacent violin plots of copy number estimates over the entire locus.

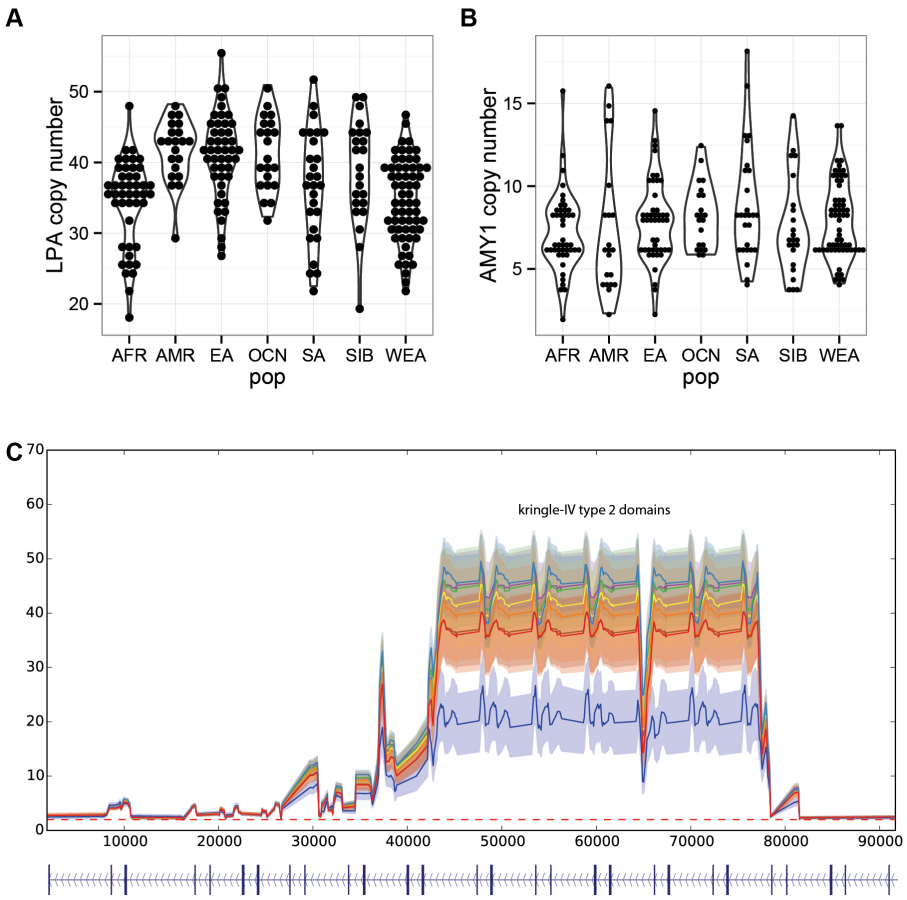

Figure S37: Copy number distributions per population of the number of Kringle IV type-2 repeats in LPA (a and c) and the copy number of the AMY1 locus (b).

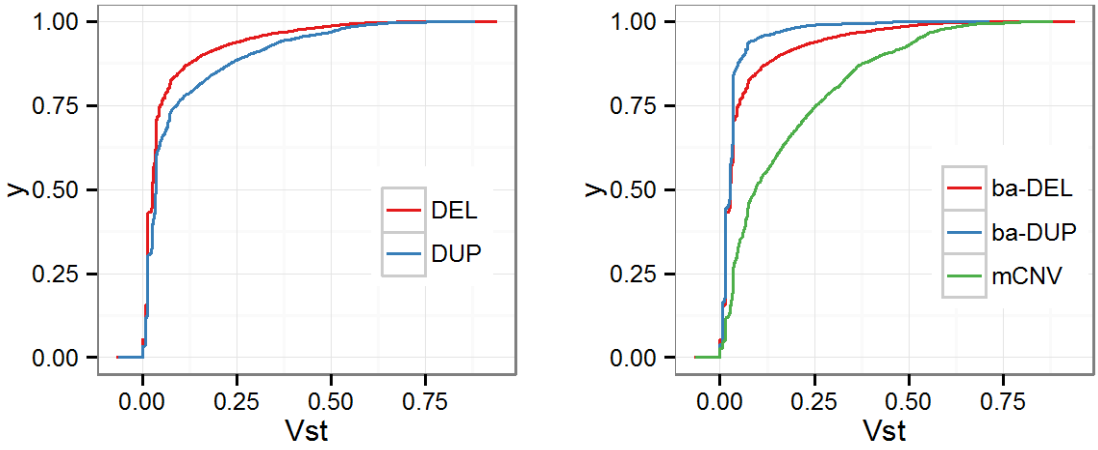

Figure S38: Cumulative distributions of Vst for all deletions and duplications (left) and for bi-allelic deletions, bi-allelic duplications vs. multi-allelic CNVs (mCNVs) (right).

As we analyzed the archaic Denisova and Neanderthal genomes in concert with the diverse human individuals assessed, we sought to determine if we could identify any highly stratified CNVs sharing alleles with these archaic genomes suggesting putative introgression. While we found no highly stratified loci sharing the Neanderthal allele, we identified five Oceanic-specific CNVs sharing the Denisova allele (**Table S14**). One variant in particular, chr16:22710040-22783557, was of interest due to its large size and high frequency in Papuan and Bougainville individuals (27 alleles / 32 chromosomes, 0.84 AF).

Table S14: Population-specific CNVs in Oceanic populations with the same allele as Denisovans with allele frequency indicated for all populations.

contig	start	end	size	type	Vst	SIB	OCN	EA	AFR	SA	AMR	WEA	Denisova	Neanderthal
chr16	21596721	21601719	4998	Dup	0.64	0	0.64	0	0	0	0	0	0.5	0
chr16	22710040	22783557	73517	Dup	0.32	0	0.31	0	0	0	0	0	0.5	0
chr13	104595073	104600742	5669	Del	0.24	0	0.24	0	0	0	0	0	1	0
chr3	110158791	110182347	23556	Del	0.28	0	0.21	0	0	0	0	0	1	0
chr3	177002636	177011374	8738	Del	0.52	0	0.45	0	0	0	0	0	1	0

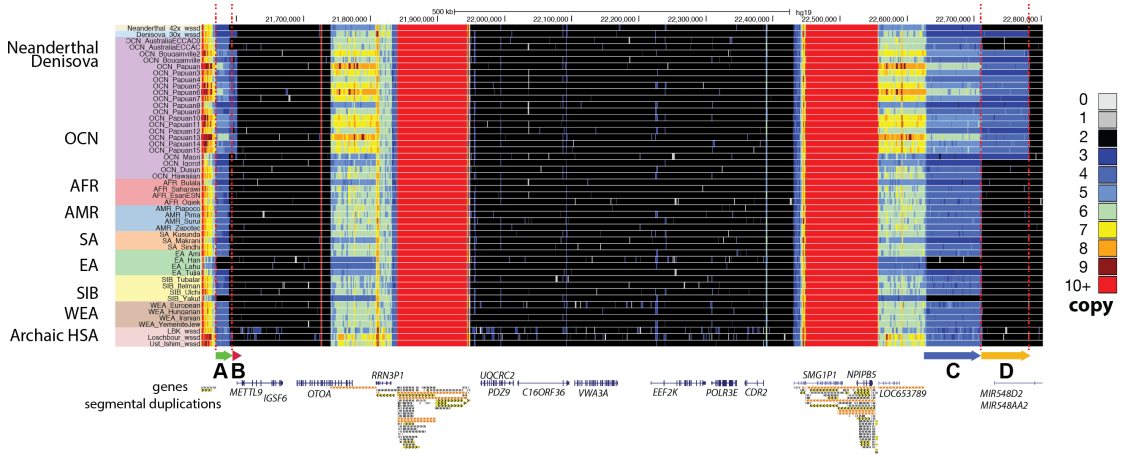

Figure S39: Heatmap representation of copy number over 16p12 (chr16:21518638-22805719) highlighting population-specific duplications in Oceanic populations with the same allele as Denisovans. Four regions of interest, A, B, C and D, are highlighted. Papuan and Bougainville Oceanic samples carry duplications in common with Denisova that are absent from other humans including Oceanic samples more closely allied with East Asians.

Further analysis of this locus identified four duplicated loci (A,B,C and D) exhibiting correlated copy number (**Figures S39, S40**), suggesting alternate duplication structural polymorphism. By manually inspecting paired-end reads, we hypothesized two alternate structural haplotypes not present in the reference genome (**Figure S41**) and searched for discordant paired-end reads supporting either the *AC* or *BD* structures in 209 individuals (**Table S15**). We identified 5,625 reads supporting the A/C duplication structure in all individuals and all populations

assessed, demonstrating this structure to likely be present in all humans, despite its absence from the reference genome. In contrast, while supporting reads for the BD structure were found in all Papuan and Bougainville genomes containing the Denisova-Papuan-specific duplication (237 reads total), no other individuals or populations showed strong support for this structure. Further analysis demonstrated this architecture likely extends further distally for ~70 kbp (C′ locus, **Figure S42**). Thus, we estimate that 225 kbp of the duplication (B-D-C-C′) structure is unique to Papuan-Bougainville and the archaic Denisova genome.

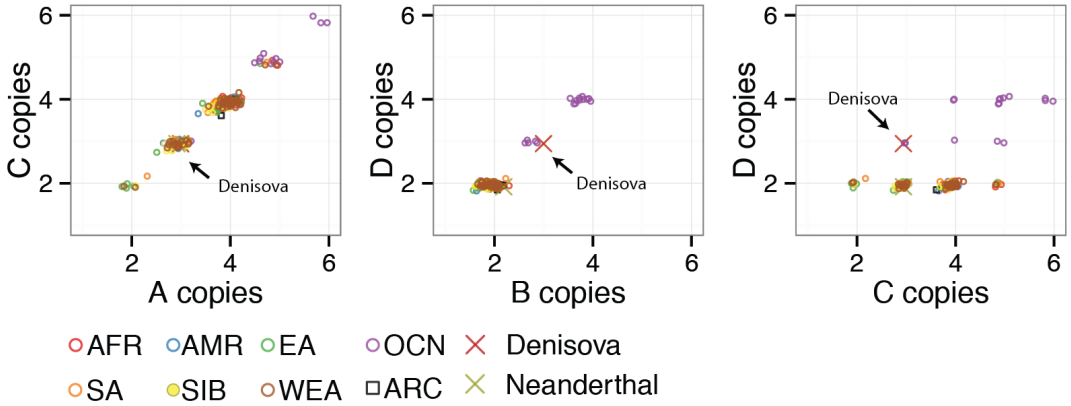

Figure S40: Pairwise plots of the copy number between A,C, B,D and C,D. Data show that individual duplications duplicate in concert transitively indicating they are part of a larger cassette that has expanded specifically in Denisova and Oceanic samples.

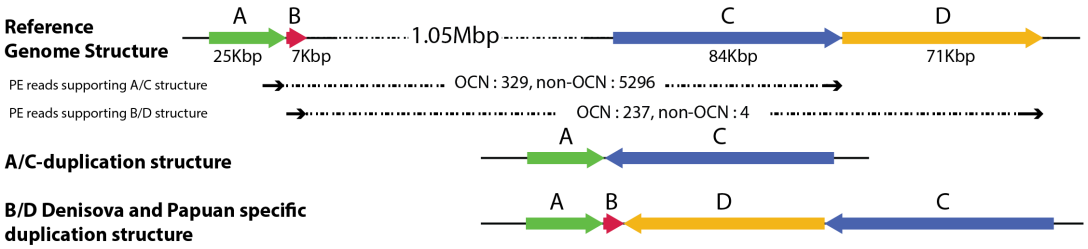

Figure S41: The reference genome structure highlighting chromosome 16p12 duplications A,B,C and D and the postulated alternate ABDC structure and Denisova/Papuan-specific structure. Paired-end reads supporting either A/C or B/D are diagrammed with the number of read supports in Oceanic and non-Oceanic individuals indicated.

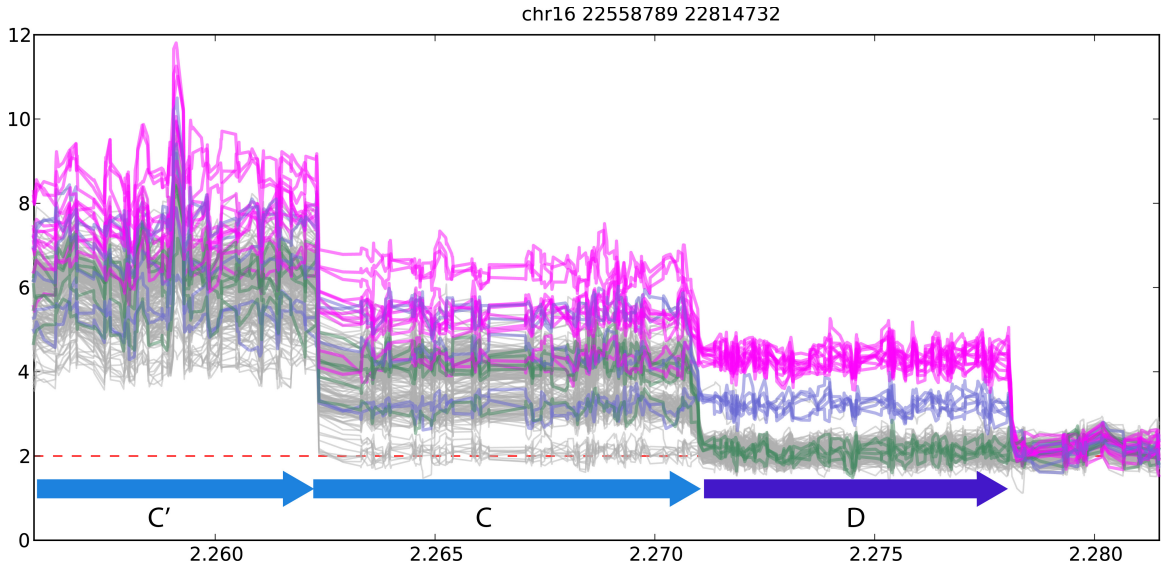

Figure S42: Focus on the CD duplication architecture and the adjacent locus (C′). Individuals with 2, 3 and 4 copies of D are represented in green, blue and pink, respectively. The highest copy numbers in C′ and C are found in individuals with four copies of D suggesting the duplications extends across all three segments.

Table S15: A table of the number of discordant paired-end reads in 209 individuals from varying populations supporting either the AC duplication structure or the BD duplication structure. While the AC structure is present in all humans, though not represented in the reference genome, the BD structure is only present in Denisovans and Papuan individuals.

Structure	Population (n)	Supporting Reads
AC	AFR (33)	1068
	AMR (19)	506
	EA (40)	1014
	OCN (17)	329
	SA (20)	573
	SIB (20)	575
	WEA (52)	1560
BD	AFR (33)	1
	EA (40)	3
	OCN (17)	237
	SIB (20)	1

The B/D duplication was present either heterozygously or homozygously in all of the Papuan and Bougainville individuals assessed. We used this property to identify paralog-specific variants (PSVs) unique to the duplication haplotype by calling SNVs using Freebayes with an input CNV map specifying the copy number of the duplication in each individual. We identified 70 likely PSVs by identifying SNVs which in all individuals harboring the duplication (copy 3 or copy 4) exhibited the same number of non-reference genotypes as copies (e.g., phased for copy number status). For example, for a particular PSV site all individuals with three copies exhibited 0/0/1 (or 1/1/1) genotypes and all individuals with four copies exhibited 0/0/1/1 (or 1/1/1/1) genotypes (**see Table S16**).

We also called SNVs in the archaic Denisova individual (heterozygous for the duplication, copy 3) to identify Denisova variants present either on the duplication or at the diploid ancestral locus. We identified a total of 114 SNVs/PSVs of which 65 were shared with the 70 Papuan duplication PSVs (92.8% of the Papuan duplication PSVs thus were shared with Denisova SNVs/PSVs) (**Figure S43, Table S16**).

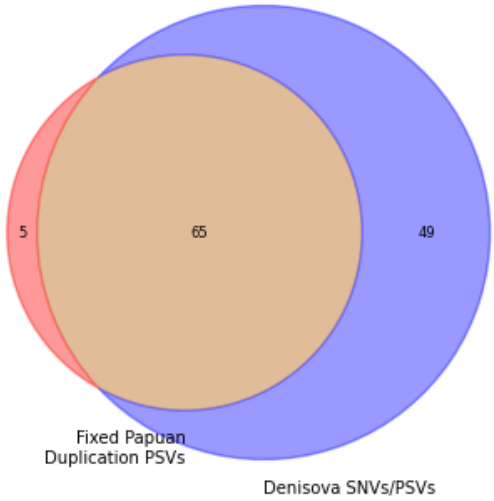

Figure S43: A Venn diagram of the number of fixed PSVs identified in the Papuan duplication and non-reference SNVs/PSVs identified in the three Denisova copies. 65/70 of the fixed Papuan duplication PSVs were also found on at least one the Denisova haplotypes.

Table S16: Genotypes of the reference, fixed Papuan duplication PSVs, and the three Denisova alleles (the Denisova is heterozygous for the duplication).

position	reference	Fixed_Papuan_Dup	Denisova
22712161	G	T	TTT
22712902	T	C	CCC
22715413	T	C	CCC
22716182	C	T	TTT
22716533	G	A	AAA
22718399	T	C	CCC
22718613	C	T	TTT
22718643	G	A	AAA
22718938	G	A	AAA
22719254	G	A	GGG
22720091	C	T	TTT
22720120	C	T	TTT
22720189	G	A	AAA
22723946	G	G	CCC
22724022	A	C	CCC
22724075	A	T	TTT
22725630	T	G	GGG
22725853	A	G	GGG
22726214	G	A	AAA
22726893	A	G	GGG
22727401	T	C	CCC
22727845	A	G	GGG
22727922	C	T	TTT
22728639	C	C	CCT
22729846	A	A	GGG
22730166	C	T	TTT
22730771	G	G	GGA
22730776	G	G	AAA
22731324	T	T	CCC
22731503	T	T	TTC
22731885	C	T	TTT

22732413	C	C	TTT
22732471	A	T	TTT
22732587	T	G	GGG
22732645	G	A	AAA
22732976	C	G	GGG
22732984	C	C	CAA
22733047	T	T	CCC
22733278	A	A	GGG
22733450	A	A	GGG
22733467	C	C	AAA
22733480	T	T	TTC
22734213	C	G	GGG
22734381	G	C	CCC
22734444	C	C	TTT
22735217	G	G	CCC
22735561	G	G	AAA
22736085	T	T	AAA
22736181	C	A	AAA
22736311	G	A	AAA
22737967	A	A	AGG
22738536	A	G	AAA
22738818	T	C	TTT
22739058	A	A	GGG
22740160	G	C	CCC
22740872	G	G	AAA
22741438	T	C	CCC
22742915	C	C	TTT
22743803	C	T	TTT
22743958	T	T	AAA
22744494	G	G	GGA
22744538	G	G	CCC
22744714	T	C	TTT
22745316	G	A	AAA
22746289	T	T	TTC
22746902	C	C	CTT
22747017	T	T	CCC
22747280	G	A	AAA
22748398	T	C	TTT
22748603	A	C	CCC
22750715	C	C	TTT
22750828	G	T	TTT
22750885	C	C	CCT
22751331	G	C	CCC
22751443	G	A	AAA
22752966	C	T	CCT
22753373	A	C	AAC
22753570	C	C	CAA
22754416	A	G	GGG
22754679	A	A	AGG
22754716	A	A	AGG
22755459	C	T	TTT
22757369	C	C	CCT
22757748	C	T	CCT
22757931	C	C	TTT
22758215	A	G	GGG
22758342	C	G	GGG
22758938	G	G	CCC
22760379	A	A	AAG
22760667	A	A	CCC
22761671	A	T	AAT
22762065	G	A	AAA
22763113	G	A	GGA
22763137	C	T	TTT
22764018	A	T	AAT
22764118	A	G	GGG
22765984	T	C	CCC
22768213	C	T	TTT
22768821	A	A	GGG
22770150	C	A	AAA
22770618	G	A	AAA
22771000	T	C	CCC
22771592	A	T	TTT
22771649	T	G	GGG
22771733	T	C	CCC
22771753	C	C	TTT
22772329	A	A	AAG
22773354	C	T	TTT
22773375	G	G	CCC
22773497	C	T	TTT
22773583	T	T	TGG
22774186	T	T	CCC
22774436	A	A	CCC
22774526	G	A	AAA
22778043	C	C	TTT
22778086	G	G	AAA
22779337	C	C	TTT
22780592	A	A	GGG
22781794	A	G	GGG

As the phase of these variants in Denisova cannot be readily deduced, they were randomly distributed amongst three haplotypes and phylogenetically compared to the inferred duplication sequence, the reference genome or the orangutan reference (the locus intersects an independent duplication in chimpanzees and gorillas). Sampling random Denisova haplotypes, the mean number of shared non-reference variants also found as fixed in the Papuan duplication was 49.1 (**Table S17**). The remaining 128 non-fixed duplication variants found in each of the Papuan/Bougainville individuals were assigned as heterozygous or homozygous; however, we could not determine whether they occurred on the duplicate or at the ancestral location. We thus constructed random duplication haplotypes for each Papuan individual using the fixed duplication variants and randomly sampled non-fixed variants and constructed a maximum likelihood tree. This tree exhibited 100% bootstrap support clustering the Papuan duplications with the Denisova branch providing strong support that this is a case of introgression (**Figure S44**). Selecting a single Denisova and Papuan haplotype, we used the orangutan and human reference genomes to calibrate timing estimates assuming a human-orangutan divergence time of 13 million years ago (**Figure S45**). A full table of variants called across the duplication is attached separately in **Table S5** (attached as a separate Excel file**)**.

Table S17: Summary of the number of nucleotides differing between pairs of randomly sampled Denisova haplotypes and the reference or the fixed Papuan duplication.

Sequence 1	Sequence 2	Delta (mean)
GRCh37	Papuan duplication	70
GRCh37	Denisova	100.91
Papuan duplication	Denisova	49.10

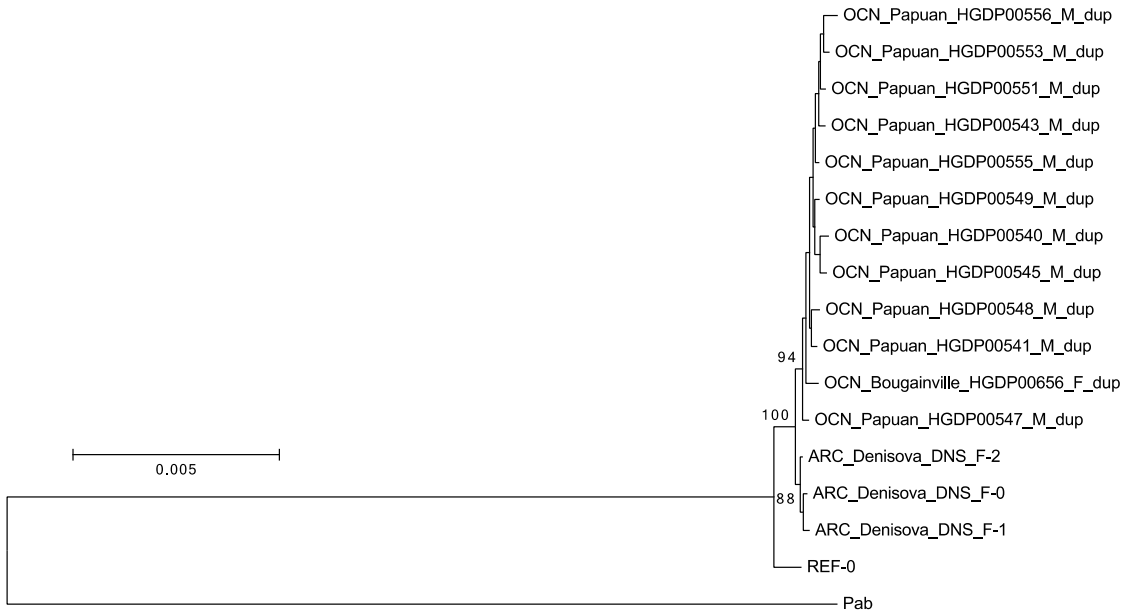

Figure S44: A maximum likelihood tree of randomly generated Papuan duplication haplotypes, the reference genome and three randomly assigned Denisova haplotypes over the duplicated locus out-grouped on the orangutan reference genome (Pab). Bootstrap supports >85% are shown. The Papuan duplication haplotypes and Denisova haplotypes cluster together with 100% bootstrap support.

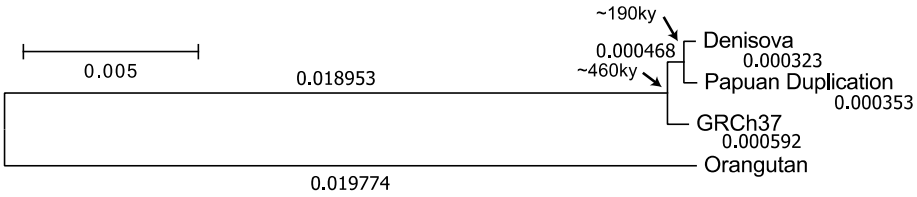

Figure S45: A maximum likelihood tree of a single random Denisova haplotype, a random duplication haplotype, the reference genome and orangutan with branch lengths. Split times were estimated using a human-orangutan divergence time of 13 million years ago.

Section S11: Genome-wide distribution of CNVs and SNVs and comparison to GWAS SNPs and positively selected loci

In order to assess the genome wide distribution of CNVs ascertained in our study we plotted ideograms with the locations of deletions, duplications and mCNVs (**Figure S46**) and assessed the density CNV heterozygosity with SNP heterozygosity (**Figures S47, S48**). We additionally assessed the intersection of GWAS SNPs and sites of positive selection with our CNVs (**Table S18**).

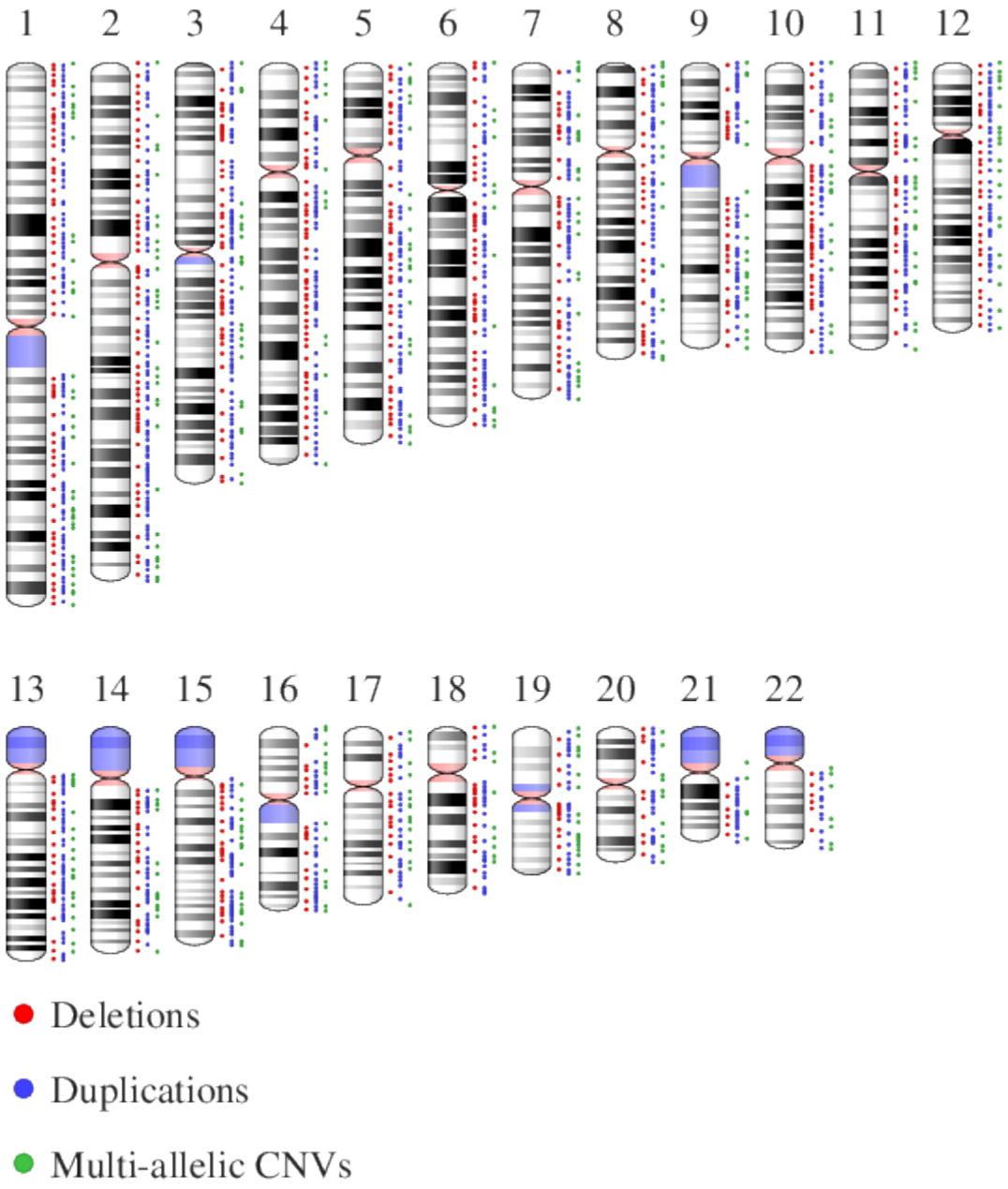

Figure S46. Genomic distribution of deletions, duplications, and multi-allelic CNVs (mCNVs) in GRCh37. Events of the same type occurring within 1 Mbp of each other are represented by a single point.

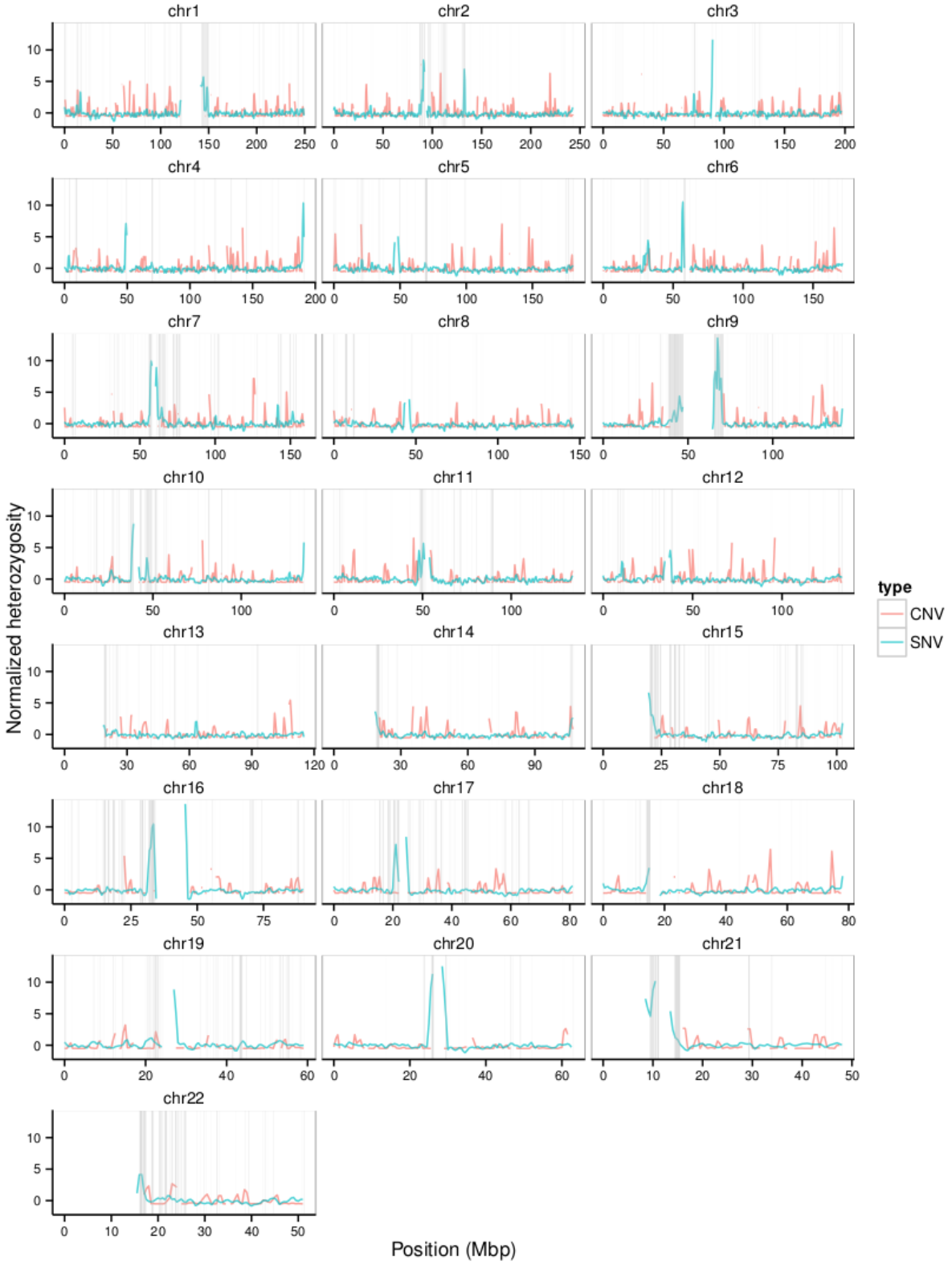

Figure S47. Normalized heterozygosity (genome-wide z-score) for CNVs (bi-allelic deletions) and SNVs for 1 Mbp, windows slid by 500 kbp across chromosomes. Genomic regions containing SDs >5 kbp are highlighted in gray.

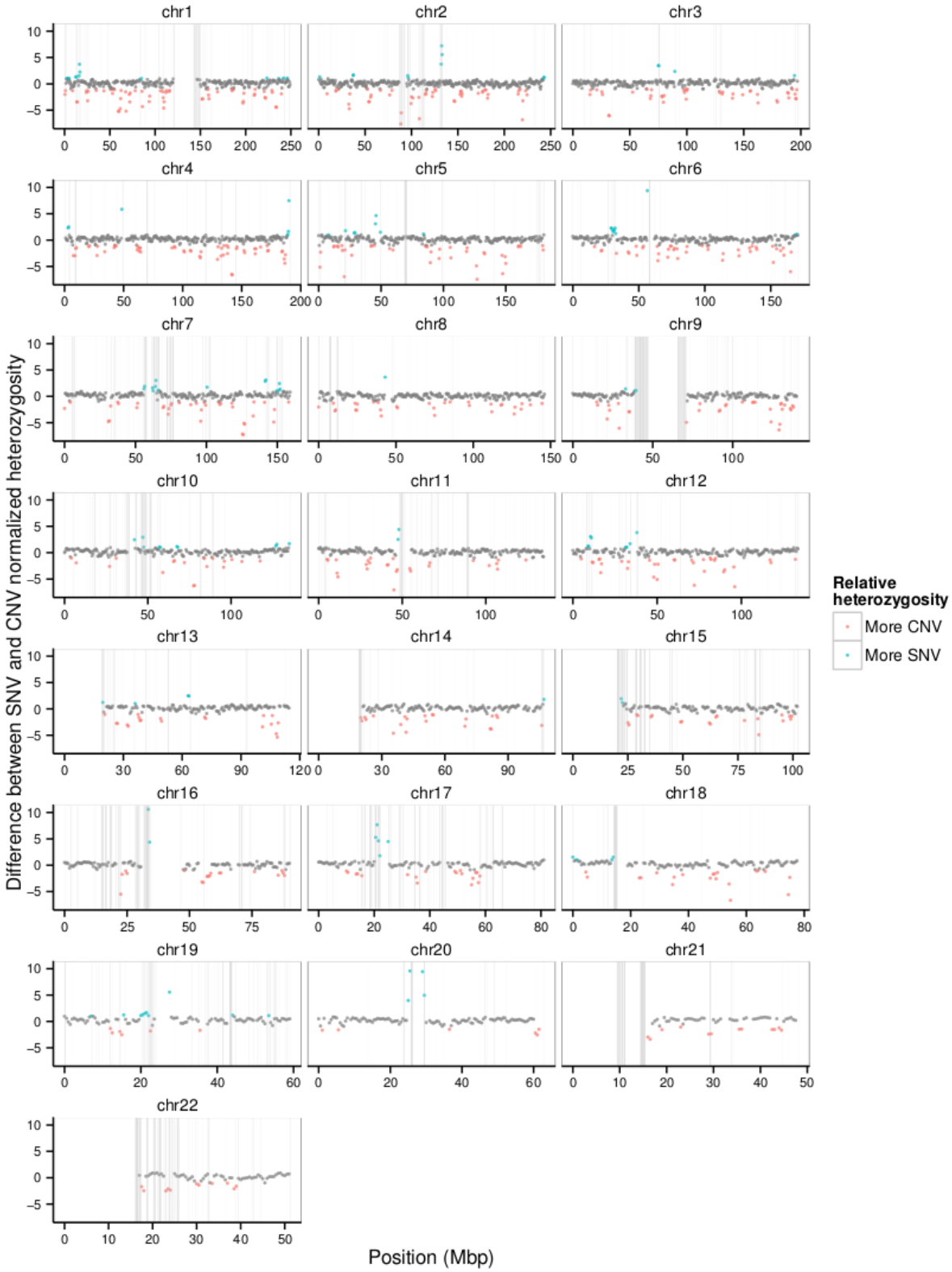

Figure S48. Difference between SNV and CNV (deletion) normalized heterozygosity (z-score) for bi-allelic events measured in 1 Mbp windows slid by 500 kbp across chromosomes. Regions of excess CNV heterozygosity

relative to SNVs (z-score difference >1) are shown in red and those with excess SNV heterozygosity are shown in blue. Genomic regions containing SDs >5 kbp are highlighted in gray.

Table S18: CNVs intersecting positively selected autosomal loci *(33)* and GWAS SNPs *(32)*.

	CNVs intersecting - positively selected loci (loci intersecting CNVs)	CNVs intersecting – GWAS SNP regions (GWAS regions intersecting CNVs)
Deletions (7,511)	73	400
Bi-allelic Duplications (2,990)	58	273
Multi-allelic duplications (4,511)	84	192
Total (15,012)	215 (77/364)	865 (691/10,918)