The Simons Genome Diversity Project: 300 genomes from 142 diverse populations

A list of authors and affiliations appears at the end of the paper.

Here we report the Simons Genome Diversity Project data set: high quality genomes from 300 individuals from 142 diverse populations. These genomes include at least 5.8 million base pairs that are not present in the human reference genome. Our analysis reveals key features of the landscape of human genome variation, including that the rate of accumulation of mutations has accelerated by about 5% in non-Africans compared to Africans since divergence. We show that the ancestors of some pairs of present-day human populations were substantially separated by 100,000 years ago, well before the archaeologically attested onset of behavioural modernity. We also demonstrate that indigenous Australians, New Guineans and Andamanese do not derive substantial ancestry from an early dispersal of modern humans; instead, their modern human ancestry is consistent with coming from the same source as that of other non-Africans.

To obtain a complete picture of human diversity, it is necessary to sequence the genomes of many individuals from diverse locations. To date, the largest whole-genome sequencing survey, the 1000 Genomes Project, analysed 26 populations of European, East Asian, South Asian, American, and sub-Saharan African ancestry\(^1\). However, this and most other sequencing studies have focused on demographically large populations. Such studies tend to ignore smaller populations that are also important for understanding human diversity. In addition, many of these studies have sequenced genomes to only 4–6-fold coverage. Here, we report the Simons Genome Diversity Project (SGDP): deep genome sequences of 300 individuals from 142 populations chosen to span much of human genetic, linguistic, and cultural variation (Supplementary Data Table 1).

Data set and catalogue of novel variants

We sequenced the samples to an average coverage of 43-fold (range 34–83-fold) at Illumina Ltd; almost all samples (278) were prepared using the same PCR-free library preparation (https://support.illumina.com/content/dam/illumina-marketing/documents/services/ FastTrackServices_Methods_Tech_Note.pdf). We aligned reads to the human reference genome hs37d5/hg19 using BWA-MEM (BWA-0.7.12)\(^2\) (Supplementary Information section 1). We genotyped each sample separately using the Genome Analysis Toolkit (GATK)\(^3\), with a modification to eliminate bias towards genotypes matching the reference (Supplementary Information section 1). We developed a filtering procedure that generates a sample-specific mask. At ‘filter level 1’ which we recommend for most analyses, we retain an average of 2.13 Gb of sequence per sample and identify 34.4 million single nucleotide polymorphisms (SNPs) and 2.1 million insertion/deletion polymorphisms (indels) (Supplementary Information section 2).

We have made the GATK-processed data available in a file small enough to download by FTP, along with software to analyse these data (Supplementary Information section 3). The SGDP data set highlights the incompleteness of current catalogues of human variation, with the fraction of heterozygous positions not discovered by the 1000 Genomes Project being 11% in the KhoeSan and 5% in New Guineans and Australians (Extended Data Fig. 1; Supplementary Data Table 1). We used FermiKit\(^4\) to map short reads against each other, store the assemblies in a compressed form that retains all the information required for polymorphism discovery and analysis, and identified SNPs by comparing against the human reference. We find that FermiKit has comparable specificity and sensitivity to GATK for SNP discovery and genotyping, and is more accurate for indels (Supplementary Information section 4). FermiKit also identified 5.8 Mb of contigs that are present in the SGDP but absent in the human reference genome presumably because they are deleted there; these contigs which we have made publicly available can be used as ‘decoys’ to improve read mapping (Supplementary Information section 5). Finally, we called copy number variants\(^5\) and used lobSTR\(^6,7\) to genotype 1.6 million short tandem repeats (STRs) (Supplementary Information section 6). The high quality of the STR genotypes (\(r^2 = 0.92\) to capillary sequencing calls) is evident from their accurate reconstruction of population relationships, even for difficult-to-genotype mononucleotide repeats (Extended Data Fig. 2).

The structure of human genetic diversity

To obtain an overview of population relationships, we carried out ADMIXTURE\(^8\) (Extended Data Fig. 3) and principal component analysis\(^9\) (Extended Data Fig. 4a). We also built neighbour-joining trees based on pairwise divergence per nucleotide (Fig. 1a) and \(F_{ST}\) (Extended Data Fig. 4b) whose topologies are consistent with previous findings that the deepest splits among human populations are among Africans. We computed heterozygosity—the proportion of diallelic genotypes per base pair—and recapitulate previous findings that the highest genetic diversity is found in sub-Saharan Africa and that there is a much lower ratio of X-to-autosome diversity in non-Africans than in Africans (Fig. 1b)\(^10\). A surprise is that African ‘pygmy’ hunter-gatherers have reduced X-to-autosome diversity ratios relative to all other sub-Saharan Africans. This pattern is just as strong even after we remove the third of chromosome X known to be subject to the strongest natural selection, suggesting that the finding is driven by demographic history rather than by natural selection (Supplementary Information section 7). It has been suggested that the reduced X-to-autosome heterozygosity ratio in non-Africans is due to ongoing male-driven admixture\(^10,11\). Male non-pygmy admixture into pygmies is well-documented\(^12,13\), so this process could explain these findings.

Comparisons of ancient to present-day human genomes have shown that all non-Africans today possess Neanderthal ancestry\(^14\) with...
more in eastern non-Africans, and that Australo-Melanesians, and to a lesser extent other eastern non-Africans, possess Denisovan ancestry. However, these studies only analysed genomes from a handful of populations. We computed statistics informative about Neanderthal and Denisovan ancestry and provide a fine-scale view of these ancestry distributions worldwide (Fig. 1c, d; Supplementary Data Table 1; Supplementary Information section 8). We do not detect any population with a higher proportion of Neanderthal ancestry than is present in East Asians. However, we do find suggestive evidence of an excess of Denisovan ancestry in some South Asians compared to other Eurasians. This signal may not have been detected before because earlier surveys of archaic introgression largely excluded South Asians (Fig. 1d; Supplementary Data Table 1).

**The time course of human population separation**

We studied demographic history by leveraging the fact that variation across the genome in divergent sites per base pair can be used to reconstruct population size changes and separations. We used the pairwise sequential Markovian coalescent (PSMC) and the multiple sequentially Markovian coalescent (MSMC) to study the time course of population separations. We infer that the population ancestral to all present day humans began to develop

![Figure 1 | Genetic variation in the SGDP. a, Neighbour-joining tree of relationships based on pairwise divergence. b, Plot of autosomal heterozygosity against the X-to-autosome heterozygosity ratio, showing the reduction in this ratio in non-Africans and pygmies. c, Estimate of Neanderthal ancestry with a heat map scale of 0–3%. d, Estimate of Denisovan ancestry with a heat map scale of 0–0.5% to bring out subtle differences in mainland Eurasia (Oceania groups with as much as 5% Denisovan ancestry are saturated in bright red).](image-url)
substructure at least 200 thousand years ago (kya), which is most apparent when comparing the ancestors of some present-day African hunter–gatherers (southern African KhoeSan and central African Mbuti pygmies) to other populations (Fig. 2a). However, it is also clear that this substructure developed slowly, as all pairs of present-day populations including African hunter–gatherers share a substantial subset of their ancestors as recently as a hundred thousand years ago.22–25

Quoting the time at which MSMC infers that more than 50% (25–75%) of lineages for a pair of populations are descended from the same ancestor, we estimate that non-Africans separated substantially from the KhoeSan 87 (58–120) kya; from the Mbuti 56 (32–84) kya; and from the Dinka 19 (9–25) kya. We estimate a relatively rapid 21 (21–26) kya separation of northern and southern KhoeSan23,26 potentially reflecting isolation since the last glacial maximum; and 38 (27–44) kya separation between western (Biaka) and eastern (Mbuti) pygmies, confirming very old substructure between these two central African hunter–gatherer groups27. Outside Africa, the most ancient structure dates to around 50 kya (Fig. 2c) during or shortly after the deepest part of the shared non-African bottleneck 40–60 kya, consistent with the archaeological evidence of the dispersal of modern humans into Eurasia during this period. We are not confident about the estimates of the date of

Figure 2 | Cross-coalescence rates and effective population sizes for selected population pairs. a–c, Cross-coalescence rates as a function of time in thousands of years ago (kya) estimated using MSMC, with four haplotypes per pair. In each subfigure legend, we give the point estimate of the date at which 25%, 50% and 75% of lineages in the pair of populations have coalesced into a common ancestral population. We generated these plots using data phased with the 1000 Genomes reference panel (method PS1 described in Supplementary Information section 9), but only show pairs of populations for which the cross-coalescence rates are relatively insensitive to the phasing approach. a, Selected African cross-coalescence rates. b, Central African rainforest hunter–gatherer cross-coalescence rates. c, Ancient non-African cross coalescence rates. d–f, Effective population sizes inferred using PSMC, using one diploid genome per population, for the same populations that we used in a–c.

the Yoruba separated substantially from the KhoeSan 87 (58–120) kya; from the Mbuti 56 (32–84) kya; and from the Dinka 19 (9–25) kya. We estimate a relatively rapid 21 (21–26) kya separation of northern and southern KhoeSan23,26 potentially reflecting isolation since the last glacial maximum; and 38 (27–44) kya separation between western (Biaka) and eastern (Mbuti) pygmies, confirming very old substructure between these two central African hunter–gatherer groups27. Outside Africa, the most ancient structure dates to around 50 kya (Fig. 2c) during or shortly after the deepest part of the shared non-African bottleneck 40–60 kya, consistent with the archaeological evidence of the dispersal of modern humans into Eurasia during this period. We are not confident about the estimates of the date of

substructure at least 200 thousand years ago (kya), which is most apparent when comparing the ancestors of some present-day African hunter–gatherers (southern African KhoeSan and central African Mbuti pygmies) to other populations (Fig. 2a). However, it is also clear that this substructure developed slowly, as all pairs of present-day populations including African hunter–gatherers share a substantial subset of their ancestors as recently as a hundred thousand years ago.22–25

Quoting the time at which MSMC infers that more than 50% (25–75%) of lineages for a pair of populations are descended from the same ancestral population, we estimate that non-Africans separated substantially from KhoeSan 131 (82–173) kya and almost as recently from the Mbuti around 112 (66–171) kya. Within Africa (Fig. 2a, b), we infer that

the Yoruba separated substantially from the KhoeSan 87 (58–120) kya; from the Mbuti 56 (32–84) kya; and from the Dinka 19 (9–25) kya. We estimate a relatively rapid 21 (21–26) kya separation of northern and southern KhoeSan23,26 potentially reflecting isolation since the last glacial maximum; and 38 (27–44) kya separation between western (Biaka) and eastern (Mbuti) pygmies, confirming very old substructure between these two central African hunter–gatherer groups27. Outside Africa, the most ancient structure dates to around 50 kya (Fig. 2c) during or shortly after the deepest part of the shared non-African bottleneck 40–60 kya, consistent with the archaeological evidence of the dispersal of modern humans into Eurasia during this period. We are not confident about the estimates of the date of

substructure at least 200 thousand years ago (kya), which is most apparent when comparing the ancestors of some present-day African hunter–gatherers (southern African KhoeSan and central African Mbuti pygmies) to other populations (Fig. 2a). However, it is also clear that this substructure developed slowly, as all pairs of present-day populations including African hunter–gatherers share a substantial subset of their ancestors as recently as a hundred thousand years ago.22–25

Quoting the time at which MSMC infers that more than 50% (25–75%) of lineages for a pair of populations are descended from the same ancestral population, we estimate that non-Africans separated substantially from KhoeSan 131 (82–173) kya and almost as recently from the Mbuti around 112 (66–171) kya. Within Africa (Fig. 2a, b), we infer that
Figure 3 | Present-day populations have negligible ancestry from an early dispersal of modern humans out of Africa. Best-fitting admixture graph model of relationships among Australians, New Guineans, Andamanese and other diverse populations. Present-day populations are shown in blue, ancient samples in red, and select inferred ancestral nodes in green. Dotted lines indicate admixture events, all of which involve archaic humans. All f-statistic relationships are accurately fit to within 2.1 standard errors. Inset, results of adding putative early dispersal admixture to the graph model for different assumptions about when the early lineage split off. We specify the split time in terms of the genetic drift above the 'Non-African' node, with 0.01 units of drift representing on the order of ten thousand years. The (approximate) model likelihood is maximized with zero early dispersal ancestry, and no more than a few per cent is consistent with the data.

Accelerated mutation accumulation in non-Africans

The SGDP data provide an opportunity to compare the rates at which mutations have accumulated across populations. We restricted our analyses to samples for which our genotypes are likely to be most reliable (this included restricting to samples which were all processed in the same way), and we used the highest level of filtering (‘level 9′) (Supplementary Information section 7). We pooled samples by region to increase power, and for all pairs of regions, computed the expected number of positions where, if we picked a random chromosome from both, region A would mismatch chimpanzee and region B would be identical to chimpanzee (or vice versa). If the rate of accumulation of mutation has been the same since the two populations diverged, these numbers are expected to be equal35. However, when we compute the ratio of mutations on one lineage or the other since separation, we find a subtle (average of 0.5%) but significant excess of mutations in non-Africans relative to sub-Saharan Africans (3.3 < Z < 9.4 standard errors from zero; Extended Data Table 1). Because any difference must reflect events since non-African/African population divergence, which is a less than a tenth of average genetic divergence (Fig. 2a), this implies a greater difference in mutation accumulation rates since population divergence (~5%). We were concerned that these results might be biased by the fact that the human genome reference sequence is more closely related to non-Africans than to Africans, or by higher levels of heterozygosity in Africans, as both of these issues could make detection of divergent sites in Africans more difficult. However, we replicated the findings after remapping to chimpanzee, which is equally distant to all present populations, and after restricting analyses to the X chromosome in males (as males only have a single X chromosome, this procedure avoids bias due to different error rates in detecting heterozygous genotypes in populations with different rates of heterozygosity) (Extended Data Fig. 5). These observations are most likely to be explained by acceleration in the rate of mutation accumulation in non-Africans, since the same signal appears in comparisons to sub-Saharan Africans related in different ways to non-Africans (Extended Data Table 1). It is known that the rate of CTC > CCT mutations differs across human populations. However, this particular mutation class was found to be enriched relative to Africans in Europeans but not in East Asians, and thus cannot explain our signal36. One of several possible explanations for these findings is a decrease in the generation interval in non-Africans compared to Africans since separation37.

No species-wide sweeps in modern humans

Finally, we used the SGDP data set to address the hypothesis that the widespread appearance of modern human behaviour in the

Early modern human dispersals contributed little
to non-Africans

There is intense debate about whether present-day Australians, New Guineans and Andamanese from other populations because we find that these inferences change depending on the computational method we use for phasing, probably due to these populations not being represented in the 1000 Genomes haploid genome reference panel (Supplementary Information section 9). We caution that the date estimates also do not take into account uncertainty about the true value of the human mutation rate, which could plausibly be 30% higher or lower than the point estimate we use28.

Non-Africans more than about 10–20 thousand years before the separation of European and East Asian ancestors, we obtain an upper bound of a few per cent for the possible contribution to Australians and New Guineans (Fig. 3 inset; Supplementary Information section 11). These results are at odds with an inference of substantial early dispersal ancestry in a previous analysis of an Australian genome31; however, that study used a less complete model that, notably, did not include the known Denisovan admixture into Australo-Melanesians17. The findings for Australians are also unlikely to be due to some unusual feature of the individuals we sequenced, as when we compared three different groups of Australian samples for which there is published genome-wide data, we found them all to be consistent with descending from a common homogeneous population since separation from New Guineans (Supplementary Information section 10). These results are not in conflict with skeletal and archaeological evidence of an early modern human presence outside of Africa29,33, as early migrations could have occurred but not contributed substantially to present-day populations. The possibility of populations that once flourished but did not contribute substantially to living groups is especially plausible now that ancient DNA from the ~45 kya Ust'-Ishim28 and the ~40 kya Oase 1 individuals34 has documented their existence.
archaeological record after ~50 kya was driven by one or a few changes in neurological genes that swept through the population shortly before this time\(^8\). We first applied the 3P-CLR method\(^9\) to search for locations in the genome with low allele frequency differentiation between KhoeSan and other modern humans, combined with high differentiation between modern and archaic (Neanderthal and Denisovan) humans, as might be expected from a selective sweep in the ancestors of all modern humans (Supplementary Information section 12) (Extended Data Fig. 6). We found no strong outlier signals, although a caveat is that the scan has limited power and we could not apply it to filtered sections of the genome. We also applied the PSMC method\(^{10}\) to estimate the average time since the most recent common ancestor (TMRCA) of individuals’ two chromosomes in the genomic regions within the largest 3P-CLR peaks (38 peaks corresponding the top 0.1%). In none of the regions did we infer that the great majority of all pairs of modern humans share a common ancestor <100 kya, as would be expected for a sweep just before ~50 kya years ago (Supplementary Data Table 2).

As a second approach to scanning for species-wide selective sweeps, we applied the PSMC to infer TMRCA for SGDP samples across the entire genome. This analysis found no regions where the great majority of pairs of human genomes are inferred to share a common ancestor <100 kya (the largest fraction seen anywhere in the genome is 68%; Extended Data Fig. 7).

Taken together, these results do not rule out the possibility that genetic changes contributed in a meaningful way to changes in human behaviour after 50 kya; for example, changing selection can produce shifts in the frequencies of pre-existing mutations to bring a population to a new and advantageous set-point for a phenotype as occurred in the case of height differences between northern and southern Europeans\(^{40}\). For polygenic selection, however, genetics is not a creative force, and instead responds to selection pressures imposed by novel environmental conditions or lifestyles. Thus, our results provide evidence against a model in which one or a few mutations were responsible for the rapid developments in human behaviour in the last 50,000 years. Instead, changes in lifestyles due to cultural innovation or exposure to new environments are likely to have been driving forces behind the rapid transformations in human behaviour in the last 50,000 years\(^{41,42}\).

Online Content Methods, along with any additional Extended Data display items and transformations in human behaviour in the last 50,000 years\(^{41,42}\). Instead, genetic changes contributed in a meaningful way to changes in human archaeological record after ~50 kya was driven by one or a few changes in neurological genes that swept through the population shortly before this time\(^8\). We first applied the 3P-CLR method\(^9\) to search for locations in the genome with low allele frequency differentiation between KhoeSan and other modern humans, combined with high differentiation between modern and archaic (Neanderthal and Denisovan) humans, as might be expected from a selective sweep in the ancestors of all modern humans (Supplementary Information section 12) (Extended Data Fig. 6). We found no strong outlier signals, although a caveat is that the scan has limited power and we could not apply it to filtered sections of the genome. We also applied the PSMC method\(^{10}\) to estimate the average time since the most recent common ancestor (TMRCA) of individuals’ two chromosomes in the genomic regions within the largest 3P-CLR peaks (38 peaks corresponding the top 0.1%). In none of the regions did we infer that the great majority of all pairs of modern humans share a common ancestor <100 kya, as would be expected for a sweep just before ~50 kya years ago (Supplementary Data Table 2).

As a second approach to scanning for species-wide selective sweeps, we applied the PSMC to infer TMRCA for SGDP samples across the entire genome. This analysis found no regions where the great majority of pairs of human genomes are inferred to share a common ancestor <100 kya (the largest fraction seen anywhere in the genome is 68%; Extended Data Fig. 7).

Taken together, these results do not rule out the possibility that genetic changes contributed in a meaningful way to changes in human behaviour after 50 kya; for example, changing selection can produce shifts in the frequencies of pre-existing mutations to bring a population to a new and advantageous set-point for a phenotype as occurred in the case of height differences between northern and southern Europeans\(^{40}\). For polygenic selection, however, genetics is not a creative force, and instead responds to selection pressures imposed by novel environmental conditions or lifestyles. Thus, our results provide evidence against a model in which one or a few mutations were responsible for the rapid developments in human behaviour in the last 50,000 years. Instead, changes in lifestyles due to cultural innovation or exposure to new environments are likely to have been driving forces behind the rapid transformations in human behaviour in the last 50,000 years\(^{41,42}\).
Extended Data Figure 1 | Heat map of fraction of heterozygous sites missed in the 1000 Genomes project. For each sample, we examine all heterozygous sites passing filter level 1, and compute the fraction included as known polymorphisms in the 1000 Genomes project.
Extended Data Figure 2 | Worldwide variation in human short tandem repeats. a, Mean STR length is reported as the average of the length difference (in base pairs) from the GRCh37 reference for each genotype. Bubble area scales with the number of calls compared at each point. b, c, The first two principal components after performing principal component analysis on tetranucleotide and homopolymer genotypes, respectively. Colours represent the region of origin of each sample. d, Pairwise $F_{ST}$ values between populations computed using only SNPs versus using combined SNP + STR loci. e, Block jackknife standard errors for the SNP versus SNP + STR $F_{ST}$ analysis. The red dashed lines give the best-fit line, described by the formula in red. The black dashed line denotes the diagonal.
Extended Data Figure 3 | ADMIXTURE analysis. We carried out unsupervised ADMIXTURE 1.23 analysis over the 300 SGDP individuals in 20 replicates with randomly chosen initial seeds, varying the number of ancestral populations between $K = 2$ and $K = 12$ and using default fivefold cross-validation (–cv flag). We used genotypes of at least filter level 1, and restricted analysis to sites where at least two individuals carried the variant allele (as singleton variants are non-informative for population clustering). After further filtering of sites with at least 99% completeness and performing linkage-disequilibrium-based pruning in PLINK 1.94 with parameters (–indep-pairwise 1000 100 0.2), a total of 482,515 single nucleotide polymorphisms remained. This figure shows the highest likelihood replicate for each value of $K$. We found that log likelihood monotonically increases with $K$, while the value $K = 5$ minimizes cross-validation error (not shown). The solution at $K = 5$ corresponds to major continental groups (Sub-Saharan Africans, Oceanians, East Asians, Native Americans, and West Eurasians), but we show the full range of $K$ here as they illustrate finer-scale population structure that may be useful to users of the data.
Extended Data Figure 4 | Principal component analysis and neighbour joining tree. a, Principal component analysis. b, Neighbour-joining tree based on $F_{ST}$ values for all populations with at least two samples.
Extended Data Figure 5 | Fewer accumulated mutations in Africans than in non-Africans confirmed by mapping to chimpanzee.

We compute a statistic \( D(\text{Population A, Population B, Chimp}) \), measuring the difference in the rate of matching to chimpanzee in Population A compared to Population B. The evidence of mismatching to chimpanzee is seen when we restrict to the male X chromosome to eliminate possible effects due to differences in heterozygosity across populations, and map to the chimpanzee genome which is phylogenetically symmetrically related to all present-day humans. We find that in 78 randomly chosen Population A = African and Population B = non-African pairs of males, transversion substitutions show no consistent skew from zero, but transition substitutions do.
Extended Data Figure 6 | 3P-CLR scan for positive selection. The red line denotes the 99.9% quantile cut-off. The genes in the top five regions are labelled. **a**, Scan for selection on the San terminal branch. **b**, Scan for selection on the non-San terminal branch. **c**, Scan for selection on the ancestral modern human branch.
Extended Data Figure 7 | Scan for genomic locations where the great majority of present-day humans share a recent common ancestor.

We carried out PSMC analysis on 40 pairs of haploid genomes chosen to sample some of the most deeply divergent present-day human lineages. We recorded the time since the most recent common ancestor (TMRCA) at each position, and rescaled to obtain an estimate of absolute time (Supplementary Information section 12).

a, Distribution across the genome of the fraction of TMRCAs below specified date cut-offs. For the 100 kya cut-off, the maximum fraction observed anywhere in the genome is 68%.

b, Distribution across the genome of the date T at which specified fractions of sample pairs are inferred to have a TMRCA less than T.

c, Percentile points of the cumulative distribution function of B.

| X=Fraction of loci below threshold | Y=80% of TMRCAs less than this date | Y=95% of TMRCAs less than this date | Y=100% of TMRCAs less than this date |
|------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| 0.01%                              | 120 kya                             | 300 kya                             | 430 kya                             |
| 0.1%                               | 320 kya                             | 500 kya                             | 620 kya                             |
| 1%                                 | 580 kya                             | 810 kya                             | 980 kya                             |
Extended Data Table 1 | Fewer accumulated mutations in Africans than in non-Africans

| Population A  | Population B    | All autosomes Dx100 | Z     | All X chromosome Dx100 | Z     | Lowest B quintile Dx100 | Z     | Highest B quintile Dx100 | Z     |
|---------------|-----------------|---------------------|-------|------------------------|-------|------------------------|-------|--------------------------|-------|
| Khoesan       | Oceania         | -0.35               | -8.2  | -0.70                  | -2.7  | -0.68                  | -6.4  | -0.14                    | -1.7  |
| Africa        | America         | -0.33               | -9.4  | -0.73                  | -2.8  | -0.65                  | -7.3  | -0.18                    | -2.6  |
| Khoesan       | WestEurasia     | -0.30               | -7.5  | -0.68                  | -3.1  | -0.63                  | -6.3  | -0.17                    | -2.1  |
| Africa        | Oceania         | -0.29               | -8.5  | -0.66                  | -3.2  | -0.55                  | -6.6  | -0.07                    | -1.0  |
| Africa        | WestEurasia     | -0.25               | -8.5  | -0.66                  | -3.1  | -0.49                  | -6.4  | -0.11                    | -1.8  |
| Khoesan       | SouthAsia       | -0.24               | -6.0  | -0.56                  | -2.7  | -0.61                  | -6.3  | -0.11                    | -1.4  |
| Africa        | EastAsia        | -0.20               | -6.6  | -0.65                  | -2.5  | -0.42                  | -5.2  | -0.10                    | -1.5  |
| Africa        | CentralAsiaSiberia | -0.20             | -6.2  | -0.55                  | -2.2  | -0.48                  | -6.3  | -0.05                    | -0.7  |
| Pygmy         | WestEurasia     | -0.19               | -4.8  | -0.46                  | -1.4  | -0.43                  | -4.6  | -0.04                    | -0.5  |
| Africa        | SouthAsia       | -0.18               | -6.4  | -0.50                  | -2.0  | -0.46                  | -6.3  | -0.03                    | -0.5  |
| CentralAsiaSiberia | Oceania       | -0.13               | -3.9  | -0.15                  | -0.6  | -0.09                  | -1.1  | -0.03                    | -0.4  |
| Pygmy         | SouthAsia       | -0.13               | -3.3  | -0.38                  | -1.1  | -0.38                  | -4.2  | 0.02                     | 0.2   |
| EastAsia      | Oceania         | -0.13               | -4.1  | 0.00                   | 0.0   | -0.17                  | -2.1  | 0.04                     | 0.6   |
| Khoesan       | Pygmy           | -0.10               | -2.6  | -0.14                  | -0.4  | -0.16                  | -1.6  | -0.12                    | -1.5  |
| SouthAsia     | WestEurasia     | -0.08               | -4.3  | -0.20                  | -1.2  | -0.05                  | -1.0  | -0.10                    | -2.7  |
| CentralAsiaSiberia | WestEurasia | -0.06               | -2.2  | -0.16                  | -0.8  | -0.01                  | -0.2  | -0.09                    | -1.6  |
| EastAsia      | WestEurasia     | -0.06               | -2.1  | -0.00                  | -0.0  | -0.08                  | -1.0  | -0.02                    | -0.3  |
| CentralAsiaSiberia | EastAsia   | -0.00               | -0.2  | -0.18                  | -1.1  | 0.07                   | 1.2   | -0.08                    | -1.8  |
| Africa        | Pygmy           | -0.00               | -0.1  | -0.06                  | -0.2  | 0.03                   | 0.4   | -0.06                    | -0.8  |
| EastAsia      | SouthAsia       | 0.02                | 0.7   | 0.22                   | 1.7   | -0.04                  | -0.7  | 0.08                     | 1.7   |
| CentralAsiaSiberia | SouthAsia   | 0.02                | 0.7   | 0.05                   | 0.3   | 0.02                   | 0.4   | -0.00                    | -0.0  |
| America       | Oceania         | 0.03                | 0.9   | 0.11                   | 0.4   | 0.10                   | 1.1   | 0.13                     | 1.7   |
| Oceania       | WestEurasia     | 0.08                | 2.3   | -0.03                  | -0.1  | 0.10                   | 1.1   | -0.04                    | -0.6  |
| Africa        | Khoesan         | 0.10                | 2.9   | 0.17                   | 0.7   | 0.23                   | 2.6   | 0.07                     | 1.0   |
| America       | WestEurasia     | 0.11                | 3.6   | 0.11                   | 0.4   | 0.19                   | 2.2   | 0.08                     | 1.3   |
| CentralAsiaSiberia | Pygmy       | 0.14                | 3.4   | 0.32                   | 0.9   | 0.43                   | 4.5   | -0.04                    | -0.4  |
| Oceania       | SouthAsia       | 0.14                | 4.8   | 0.22                   | 0.9   | 0.13                   | 1.7   | 0.04                     | 0.7   |
| EastAsia      | Pygmy           | 0.15                | 3.6   | 0.49                   | 1.4   | 0.37                   | 3.9   | 0.04                     | 0.5   |
| America       | EastAsia        | 0.18                | 5.9   | 0.09                   | 0.3   | 0.28                   | 3.6   | 0.11                     | 1.8   |
| America       | CentralAsiaSiberia | 0.18              | 6.2   | 0.34                   | 1.7   | 0.23                   | 2.9   | 0.18                     | 3.1   |
| America       | SouthAsia       | 0.18                | 6.4   | 0.34                   | 1.5   | 0.22                   | 3.0   | 0.18                     | 3.1   |
| Oceania       | Pygmy           | 0.24                | 5.4   | 0.46                   | 1.3   | 0.45                   | 4.6   | 0.02                     | 0.2   |
| CentralAsiaSiberia | Khoesan     | 0.25                | 6.0   | 0.57                   | 2.9   | 0.64                   | 6.3   | 0.09                     | 1.1   |
| EastAsia      | Khoesan         | 0.25                | 6.2   | 0.68                   | 3.2   | 0.59                   | 5.9   | 0.14                     | 1.7   |
| America       | Pygmy           | 0.26                | 5.9   | 0.58                   | 1.6   | 0.58                   | 5.7   | 0.09                     | 1.0   |
| America       | Khoesan         | 0.37                | 8.7   | 0.76                   | 3.3   | 0.77                   | 7.3   | 0.22                     | 2.5   |

We compute a statistic \( D \) (Population A, Population B, Chimp), measuring the difference in the rate of matching to chimpanzee in Population A compared to Population B. For all the autosomes, we observe highly significant signals (\( 3.3 < |Z| < 9.4 \)) of excess mismatching to chimpanzee in non-Africans compared to Africans, using a standard error from a block jack-knife. We highlight \( |D| > 0.002 \) in blue, and \( |Z| > 3 \) in yellow. The deviations from zero are greatest in subsets of the genome where the time since two populations split comprises a relatively larger fraction of the total genetic divergence time between the populations; this is the direction expected from a mutation accumulation change since divergence. Compared to all the autosomes as a baseline, a least squares fit indicate that the deviations are 2.2 times higher on chromosome X, 2.0 times higher in the quintile of lowest B-statistic (closest to functionally important regions), and 0.43 times as high in the quintile of lowest B-statistic (furthest from functional regions).