doi:10.1038/nature18964

Table of Contents 1

SI 1 – Processing of sequencing data 2-5

SI 2 – Filter construction and quality control 6-10

SI 3 – Data access, formats and tools 11-13

SI 4 – Comparison of genotypes obtained by different methods 14-20

SI 5 – Characterization of sequences missing from the reference genome GRCh38 21-25

SI 6 – Worldwide variation in human short tandem repeats 26-33

SI 7 – Variability in heterozygosity and mutation load across populations 34-43

SI 8 – The worldwide landscape of Neanderthal and Denisova introgression 44-49

SI 9 – Demographic inference 50-62

SI 10 – Sequenced Australians form a clade with previously published Australians 63-64

SI 11 – Australo-Melanesians have little ancestry from early dispersals out of Africa 65-72

SI 12 – No evidence for a shared human selective sweep in the last ~100,000 years 73-81

Supplementary Information section 1

Processing of sequencing data

Swapan Mallick*, Heng Li, Susanne Nordenfelt, Pontus Skoglund, Arti Tandon, Mengyao Zhao and David Reich

*To whom correspondence should be addressed: (shop@genetics.med.harvard.edu)

Sequencing

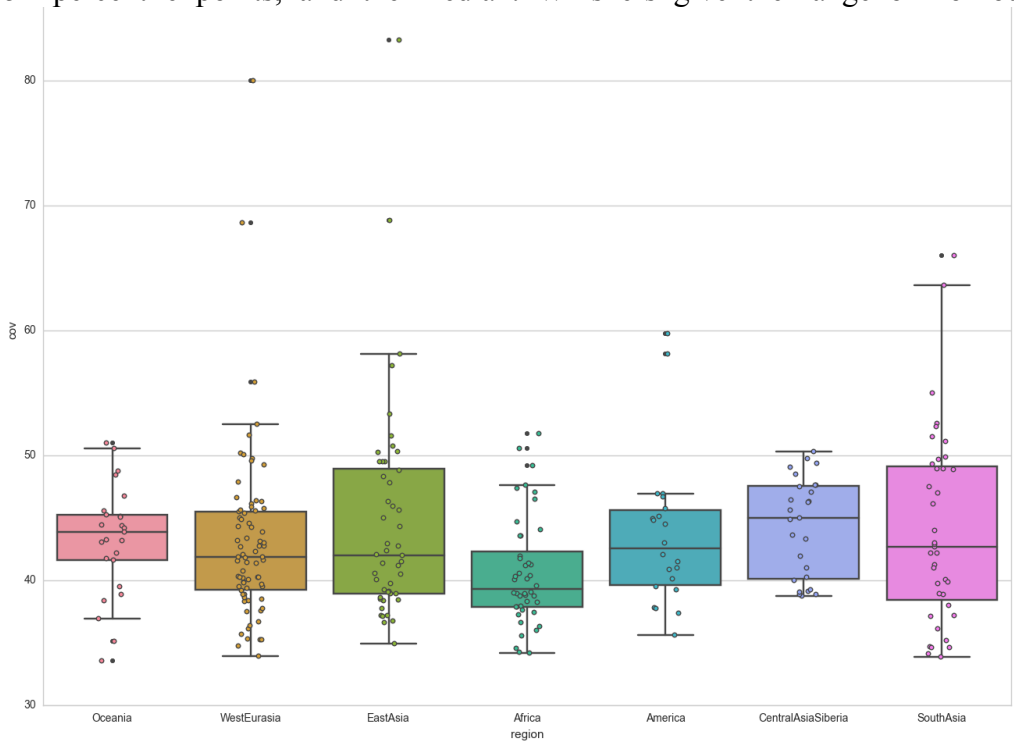
For all samples, we submitted a minimum of 2.5 micrograms of DNA to Illumina Ltd. for their standard high coverage sequencing service.

For Panel C (278 samples), all samples were all processed using the same PCR-free library preparation and sequencing protocol (all library preparation and sequencing took place between the dates of Feb 27 and Oct 16, 2013), minimizing the danger that systematic differences in processing could cause artifactual differences among samples. The samples were sequenced using 100 base pair paired-end sequencing on HiSeq2000 sequencers with an insert length distribution of 314 ± 20 bases.

The Panel B samples (22 samples, of which 14 were previously reported[1]) were submitted to Illumina at an earlier time than Panel C. The libraries were prepared using a PCR-based library preparation protocol.

After sequencing, the 300 samples reported in this study had a range of 33.59-83.23 fold coverage (median 42.0-fold) (Fig. S1.1; Supplementary Data Table 1).

Figure S1.1. Sequence coverage of the samples analyzed. Boxes give the 25th to the 75th percentile points, and the median. Whiskers give the range of non-outliers.



Preprocessing and alignment

We discovered that the raw data files supplied by Illumina retained some adapter sequences, which have the potential to contaminate genotyping results. We estimated that 0.27% of reads in the raw dataset retained adapters (range across samples of 0.014% to 1.87%). To address this problem, we implemented an adapter trimming step. We extracted reads from the raw bams by shuffling with *htslib*[1], which groups read pairs together while avoiding a computing-intensive sorting step. We converted the resulting file into an interleaved fastq format. We trimmed reads using *trimadap*[2], and then aligned using *bwa mem*[3] *(v0.7.10-r1005-dirty)* to the 'decoy' version of the human reference sequence (*hs37d5*). This reference is based on *hg19*, but contains an additional 35.4 Mb of decoy sequences, which improves alignment and subsequent variant calling in regions that are misassembled in the human reference genome or affected by common copy number variations. We added read groups during the alignment step in order to facilitate downstream analysis. We marked optical duplicates using *samblaster*[4], though it has a small chance (~1%) of mislabeling alignments as PCR duplicates for our PCR-free data, thus slightly reducing coverage. We converted the output reads in sam format to bam format, and then sorted.

The pipeline for this procedure is:

```
./htscmd bamshuf -Oun128 in.bam tmp-pre \
| ./htscmd bam2fq -as aln-se.fq.gz - \
| ./trimadap \
| ./bwa mem -pt8 hs37d5.fa - \
| ./samblaster \
| samtools view -uS - \
| samtools sort -@4 -m512M - out-pre
```

For researchers who wish to repeat our processing, but whose input dataset is in standard fastq format, the first two steps can be replaced with: “seqtk mergepe read1.fq.gz read2.fq.gz”.

Genotyping

Most analyses in this paper are based on single-sample genotypes determined using a reference-bias free modification of GATK[5]. We did not perform multi-sample genotyping as we were concerned that this could induce biases in population genetic analyses. Specifically, we were concerned that the GATK *UnifiedGenotyper* has a built-in prior for Bayesian SNP calling that assumes that the site is more likely to be homozygous for the reference allele than homozygous for the variant allele. For a diploid sample, the default priors for a homozygous reference, heterozygote and homozygous non-reference genotypes are (0.9985, 0.001, 0.0005), respectively. When there is ambiguity in a heterozygote, GATK prefers the reference homozygote. This is a reference bias, and while this bias is not typically problematic for medical studies, it can complicate interpretation of population genetics signals. With the Genome Sequencing and Analysis Group at the Broad Institute, we developed an alternative model that was integrated into the *UnifiedGenotyper*, allowing reference-bias free priors to be specified. We are using a prior (0.4995, 0.001, 0.4995). Details are at: https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_genotyper_UnifiedGenotyper.php#--input_prior.

Once we aligned bams, we performed reference-bias free genotyping on a per-chromosome basis using the following command (where CHR_ID is the chromosome). We did not treat chromosome Y or mtDNA in a special way

(java -Xmx2g -jar /home/sm213/src/src_extended/gatk/GenomeAnalysisTK-2.5-2-gf57256b/GenomeAnalysisTK.jar -T UnifiedGenotyper -I srt.aln.bam -L CHR_ID -R /groups/reich/reference-genomes/h\
s37d5/unzipped/hs37d5.fa -dcov 600 -glm SNP -out_mode EMIT_ALL_SITES -stand_call_conf 5.0 -stand_emit_conf 5.0 -inputPrior 0.0010 -inputPrior 0.4995 -D /groups/reich/sw/gatk/bundle/2.8/b37/dbsnp_138.b37.\
vcf -o CHR_ID.vcf -A GCContent -A BaseCounts >& CHR_ID.vcf.oe ; bgzip CHR_ID.vcf; tabix -p vcf CHR_ID.vcf.)

Polymorphism discovery and comparison to published SNP call sets

We determined genotypes for each sample as described above, and restricted to positions passing filter level 1 (Supplementary Information section 2) where there was polymorphism among humans or comparing human to chimpanzee (panTro2). This results in a dataset of 62.60M sites. Restricting to sites that pass universal filters and are polymorphic in the samples (rather than being panTro2 specific differences) produced 32.50M sites on autosomes and 0.93M sites on chromosome X. Comparing autosomal with public datasets at autosomal sites (Table S1.1), we find:

Table S1.1. Comparison of SGPD to other datasets at autosomal sites.

Comparison set	#sites specific to SGDP	#sites overlapping comparison set	#sites unique to comparison set
GoNL[9]	23,486,021	9,015,983	8,303,703
dbSNP (v137)[8]	25,244,229	7,257,775	6,544,167
1kg[7] (phase3)	10,582,114	21,919,890	46,554,050

Note: This analysis restricts sites that pass the universal filter (Supplementary Information section 2).

SGDP adds more than 10.5M autosomal SNPs to the 1000 Genomes Project dataset at positions passing the universal filter. The number of unique variants to this dataset (compared with 1000 Genomes Project) is evident in Supplementary Data Table 1 and Extended Data Fig. 1, where we present the faction of heterozygous genotype calls in each individual in the SGDP that are not known SNPs in the 1000 Genomes Project dataset. As expected, relatively few new variants are found for the large census size populations that have been analyzed extensively for medical genetics projects. However, in more isolated populations a substantial fraction of heterozygous positions are not in the database (Onge 4%; Papuans 5%; KhoeSan 11%).

The Transition-to-Transversion ratio (Ti/Tv) of the SNPs in the dataset is 2.02 for all positions, consistent with high quality and low error rate.

Principal Component Analysis (PCA)

We started with the filter level 1 calls in the previous step, and obtained genotypes for 1,152,838 autosomal SNPs, chosen based on the Panel 1 and Panel 2 SNP sets described in ref. 6. Extended Data Fig. 4 shows a plot of the first and second Principal Components, which maximizes differences between sub-Saharan and non-African populations on PC1, and differences between West Eurasian and East Eurasian populations on PC2. This qualitative picture is typical of previous studies of worldwide human population structure.

References

1 http://www.htslib.org/

2 https://github.com/lh3/trimadap

3 Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics (2010) 26 (5): 589-595.*

4 Faust G. and Hall I. (2014) *SAMBLASTER*: fast duplicate marking and structural variant read extraction , *Bioinformatics 30 (17): 2503-2505*.

5 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Research 20:1297-303*.

6 Fu Q., Hajdinjak M., Moldovan O., Constantin S., Mallick S., Skoglund P., Patterson N., Lazaridis I., Nickel B., Viola B., Pruefer K., Meyer M., Kelso J., Reich D. and Pääbo S. (2015) An early modern human from Romania with a recent Neanderthal ancestor *Nature* (accepted, to be published August 13 2015).

7 McVean et al (2012) An integrated map of genetic variation from 1,092 human genomes, *Nature 491, 56–65.*

8 Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., Sirotkin K. (2001) dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res. Jan 1;29(1):308-11.*

9 The Genome of the Netherlands Consortium (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population, *Nat Genet. Jun 29.*

Supplementary Information 2

Filter construction and quality control

Nick Patterson*, Mengyao Zhao, Heng Li, Niru Chennagiri, Arti Tandon, David Reich and Swapan Mallick

*To whom correspondence should be addressed: (nickp@broadinstitute.org)

Overview

Our basic strategy for filter design was to use divergence from chimpanzee as a figure-of-merit. We interpreted higher divergence from chimpanzee as evidence of a higher error rate. We designed cutoffs that minimized this divergence, and secondarily checked that divergence to the human reference genome was reasonable based on our understanding of patterns of human genetic variation. A consequence of this filtering strategy is that we may be biasing toward regions of lower true mutation rate. This means that estimates of mutation rate per base pair based on the literature may tend to be too high for the regions passing our filters. For analyses where we wish to convert genetic divergence estimates to absolute time, we therefore use a mutation rate that we recalibrate to be exactly appropriate to the parts of the genome we are actually analyzing as described in Supplementary Information section 9.

We assigned filter levels at each nucleotide in the genome for each sample as a single character (0-9), both in fasta format and as an annotated field ("FL") in per-sample vcf files (we provide annotation in both formats to permit both fasta and vcf style processing). The characters "?" or "N" indicate that the site should not be used.

We believe that for most applications, including for SNP discovery, filter level 1 is likely to be most useful, achieving a good balance between sensitivity and low error rate. For applications in which the goal is to drive down errors rates as far as possible, filter level 9 is recommended (but it loses a substantially higher fraction of the genome), as shown in Figure S2.1. If positions are already known to be polymorphic, filter level 0 is reasonable.

The filtering strategy used here is highly specific to the SGDP dataset. On a different dataset with different mean coverage for instance, we would not expect it to work as well without some modifications.

Inputs to filtering algorithm

Several inputs are required for the filtering engine, which are either *sample-specific* or *all-samples* (valid for all samples):

(i) rawvcf: a *sample-specific* file, produced from GATK genotyping (Supplementary Information section 1). It includes per-base metrics such as DP (depth), MQ (root mean square of mapping quality) and MQ0. MQ0 counts the number of reads with MAPQ = 0 for the sample, with high counts tending to occur in regions where it is difficult to make confident calls. The rawvcf file is produced using "-out_mode EMIT_ALL_SITES" which produces metrics at non-variant and variant sites.

(ii) hetfa: a *sample-specific* file that encodes the genotypes into a fasta type file using IUB encoding. This is constructed using the tool *vcf2hetfa*.

(iii) cnv: a *sample-specific* file that encodes copy number variant data. This is constructed using the tool *bam2cnv*.

(iv) an *all-samples* filter which comprises of 3 parts: i) a structural filter, which contains problematic regions mostly caused by CNVs identified by 1kg data, ii) a compositional filter, including low-complexity regions, and iii) a mapability filter where the 75-mer centered at the base is unique in the human reference genome. Details of the construction of this filter are described in Supplementary Information section 3.

The inputs used for these filters are either available for download from our ftp site as described in Supplementary Information section 6, or are outputs of the Unified Genotyper from GATK[1].

Filter construction on a per-sample basis

Step 1 – Filtering out of nucleotides based on non-GATK filters:
Bases that fail the *all samples* filter, or the *sample-specific* cnv filter, or that do not have a valid reference allele (human) are marked as N, as we do not think that they should be used in most analyses. Also, bases without a valid diploid call are marked N. Subsequent filters are based on three fields of the vcf file (MQ, MQ0, DP).

Step 2 – Depth-based-filtering based on GATK outputs:
(i) For each possible read depth interval [l,h], restricting to nucleotides with MQ=60, MQ0=0 and depth $l \le d \le h$, we compute:
(a) Coverage, that is, the proportion of bases as a total of the genome
(b) # matches and mismatches to chimpanzee (PanTro2).
We insist that the boundary values l and h each cover ≥1000 positions.
(ii) For each targeted coverage, find the [l, h] interval that has the minimal divergence to chimpanzee. This builds a filter that is optimized to a targeted coverage level.
(iii) Prune the resulting list of "optimal" intervals. We first insist that for each coverage X and coverage Y (X < Y) the interval for X is a subset of the interval for Y. This avoids paradoxical behavior, especially for extreme choices of coverage. We now set y_0 to be the maximum coverage (expressed as a fraction of all nucleotides in the genome). y_0 will be <1 as nucleotides are discarded by a number of filters (for example, the *all samples* filter). We set $y_1 = 0.5$. At filter levels (k=1…9), set the desired coverage to be $C(k) = (y_1 k + y_0(9-k))/9$. We set the interval (on the DP field) for level k to be [l(k), h(k)] as the "optimal" interval for coverage C(k). This allows us to determine 10 filter intervals that attain increasingly low divergence to chimpanzee at the expense of reduced coverage.

Step 3 - Annotation:
This requires a second pass through the vcf, in which we annotate each nucleotide according to the highest filter level that it achieves according to the depth intervals [l(k), h(k)] learned in Step 2. We further mark nucleotides as filter level 0 if they are not already marked N, and also have MQ0 = 0 and MQ > 50. We record the mask values in a fasta file and compute a .fai index.

For most applications, we recommend filter level 1. If only the best quality sites are required, we recommend filter level 9. For pulling down genotypes onto a known set of polymorphic SNPs, it is reasonable to use filter level 0, and this is the default for *cpulldown*. Statistics of divergence to chimpanzee and the human genome reference sequence by filter level are generated by the software, but note that the human reference genome is not used to set the filter level values (except insofar as requiring it to be covered). Furthermore, although chimpanzee is used to set the filter parameters, we do NOT require chimpanzee data when setting the filter value.

The whole filtering procedure is implemented in the program *cmakefilter*, which is available for download (see Supplementary Information section 6).

Example of command line for a sample (20G of memory is required):

```
cmakefilter -p S_Eskimo_Sireniki-1.par
```

cmakefilter is driven by a parameter (.par) file; example is S_Eskimo_Sireniki-1.par:

```
XXX:            LP6005443-DNA_B03
SDIR:           /n/data1/hms/genetics/reich/1000Genomes/cteam_remap/A-samples/XXX
sdir:           SDIR
vcfdir:         SDIR/rawvcf
vcfsuffix:      vcf
hetfa:          SDIR/rawvcf/rawvcf.hetfa.fa
cnv:  /n/data1/hms/genetics/reich/1000Genomes/cteam_remap/A-
samples/XXX/depth_filt/hs37d5_maskCNV_soft.fa
gender:         M
fixeddbase:     /home/np29/cteam/release/fixedfilters
sampname:       S_Eskimo_Sireniki-1
debug:          NO
```

The gender of the sample must be supplied; this may be obtained by examining the ratio of reads aligning to the X and Y chromosomes, or a gender prediction tool[2].

Post-filtering Quality Control

The filtering engine produces files sample.mask.fa and sample.mask.fa.fai, which specify the filter level at each nucleotide of the genome.

After we generate the .mask.fa files by running *cmakefilter*, we run a quality control (QC) step to make sure the .mask.fa files have expected features. We do this by running the program *filtstats* (available for download, see Supplementary Information section 6), with command:

```
filtstats -p S_Eskimo_Sireniki-1.par
```

Parameter files are the same as that for *cmakefilter*. For a sample.mask.fa, this outputs the divergence to both chimpanzee and human reference genomes for each filter level.

One example output of filtstats (the output of the above command) is as follows. The “base count” here gives the total number of alleles in each mask category. Because humans are diploid except for males on chromosomes X and Y, this is two times the length of the analyzable reference genome. The “divergence” number here gives the divergence as a to the reference genome specific to the sample being analyzed.

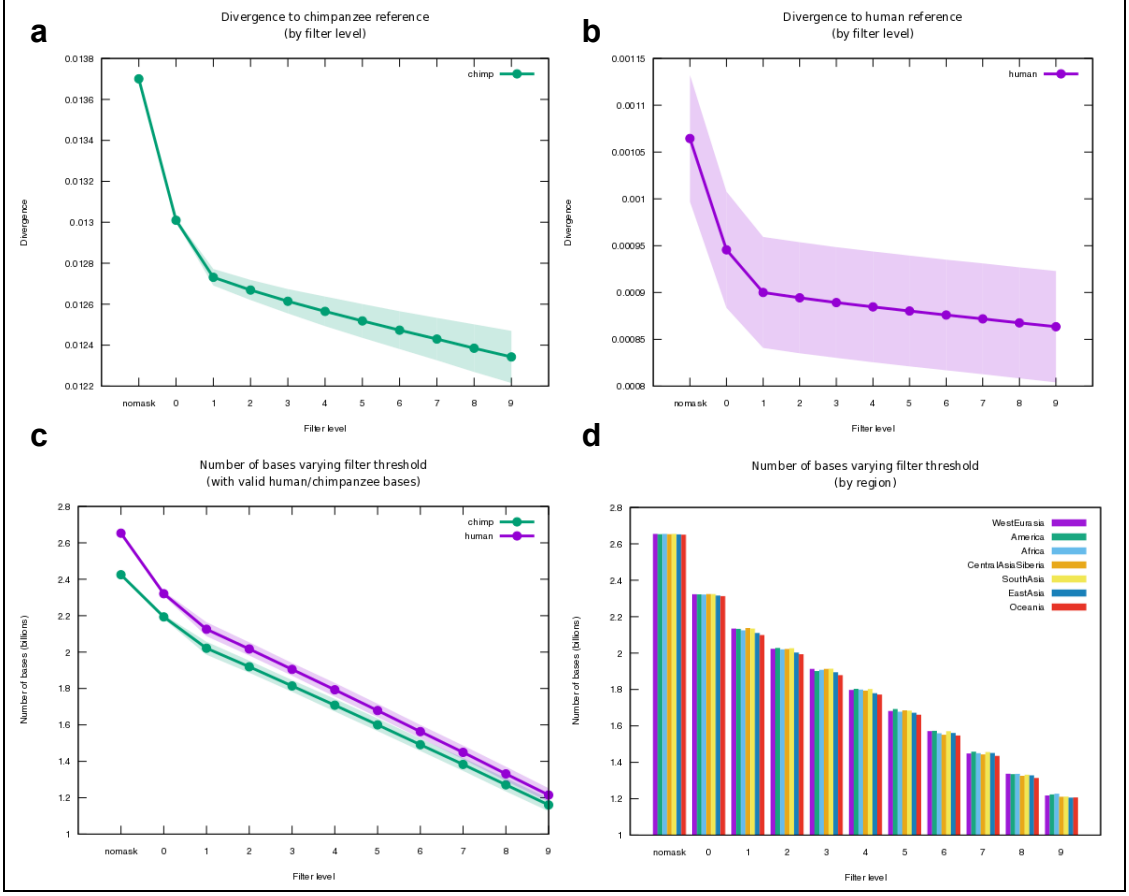
chimp_divergence

#1_mask_level	2_base_count	3_divergence
no mask:	500576324	0.019997
0	632059598	0.015103
1	132733944	0.014027
2	196627594	0.013657
3	124172884	0.013414
4	300574334	0.013211
5	174208320	0.012998
6	186528610	0.012905
7	201977194	0.012784
8	0	0.000000
9	2405268900	0.012272
Total:	4854727702	0.013700

Href_divergence

no mask:	705187112	0.001783
0	698686562	0.001143
1	143095222	0.001027
2	210454102	0.000996
3	132282554	0.000977
4	318804508	0.000957
5	184104838	0.000931
6	196666104	0.000917
7	212586160	0.000907
8	0	0.000000
9	2509201540	0.000841
Total:	5311068702	0.001036

Figure S2.1: Filtering results over all samples; shaded area indicates one standard deviation. (a) By design, increasing filter levels minimize divergence to chimpanzee. (b) Divergence to the human reference genome also decreases monotonically with filter level, even though the human reference genome is not used in filter design. (c) The number of bases retained at each filter level. (d) Stratifying by region shows only a slight variation which might be indicative of alignment bias to the reference, which is comprised mostly sequences of European, African and East Asian descent.



Annotating vcf files

We added the filter level in the "FORMAT" field of the vcf. For each sample, this takes the FL value from the .mask.fa file created above and annotates the vcf. This is done using the tool: *annotate.pl* (available by ftp, see Supplementary Information section 6).

input: raw vcf files (one vcf file for each chromosome), gzipped.

output: sample.annotated.vcf.bgz

Example of command line for each sample:

```
annotate.pl S_Eskimo_Sireniki-1.par | bgzip -c > S_Eskimo_Sireniki-1.annotated.vcf.bgz
```

The parameter files is the same as that for *cmakefilter*.

References

1 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Research 20:1297-303*.

2 Skoglund P., Storå J., Götherström A., Jakobsson M. (2013) Accurate sex identification of ancient human remains using DNA shotgun sequencing, *Journal of Archaelogical Science 40:4477-4482*

Supplementary Information section 3

Data access, formats and tools

Mengyao Zhao, Swapan Mallick, Arti Tandon, David Reich and Nick Patterson*

*To whom correspondence should be addressed: (nickp@broadinstitute.org)

Overview

The 300 samples that constitute the SGDP dataset are divided into two categories: (i) 279 (of which 263 are in Panel C) that are fully publicly available, and (ii) 21 (of which 15 are in Panel C) that require a signed letter for access (Table S3.1).

Table S3.1: Samples by geographic region

Region	C Panel fully public	B Panel fully public	Signed letter	Total
Africa	39	5	11	55
America	20	2		22
CentralAsiaSiberia	27			27
EastAsia	45	2		47
Oceania	22	3		25
SouthAsia	39		10	49
WestEurasia	71	4		75
Total	**263**	**16**	**21**	**300**

Repositories for the raw data

Raw data, results bams, annotated vcfs (which include filtering annotations), and hetfa (encoded genotypes) are available in one of two repositories:

(i) Fully public samples: EBI (http://www.ebi.ac.uk/)
Accession number: PRJEB9586, secondary accession number: ERP010710.

(ii) Signed letter samples: dbGAP (http://www.ncbi.nlm.nih.gov/gap)
dbGAP will provide raw data using their mechanism for controlled data sharing.
[Accession number to be released upon publication]

In addition to the per-sample files, we are distributing two versions of all multi-sample files (such as multi-sample vcfs and cteam-lite; see below), corresponding to the fully public data (n=279 samples), and the complete dataset (n=300 samples).

Distribution of packed files that we expect will be of greatest value to many users

Most analyses of interest to users will be possible using reduced versions of the dataset that are available on the Reich laboratory website. This consists, most importantly, of "cteam-lite", a dataset that consists of packed hetfa files and masks allowing both variant and non-variant sites to be analyzed at user-specified filter levels. This may be used for example to extract raw data for divergence calculations, which is not possible with typical vcfs restricted to variant sites. In the mask files, we assign filter levels at each nucleotide in the genome for each sample as a single

character (0-9, N, ?) in fasta format. The nucleotides with filter level 9 have the highest quality. The characters "?" or "N" indicate that the base should not be used.

In addition to the SGDP, cteam-lite also provides two reference genomes and five ancient genomes with enough coverage to allow diploid calls. The references include the human reference build 19 (https://genome.ucsc.edu/cgi-bin/hgGateway?db=hg19) and a chimpanzee genome in human coordinates (from EPO alignments of panTro2, http://hgdownload.cse.ucsc.edu/goldenPath/panTro2/bigZips/). The five ancient genomes are the Altai Neanderthal sequenced to 52-fold coverage[1], the Siberian Denisovan genome sequenced to 31-fold coverage[2], the Upper Paleolithic Siberian Ust-Ishim genome at 42-fold coverage[3], the Mesolithic European Loschbour genome at 22-fold coverage[4], and the early European farmer Stuttgart genome at 19-fold coverage[4]. For the ancient genomes, we used previously published filters, and therefore include mask files for the five ancient genomes at only two levels (0 and 1). Thus, users who wish to analyze these samples using cteam-lite must not specify *minfilterlevel* as more than 1. We hope that these 5 fasta files will provide a convenient way to access these genome sequences in conjunction with the SGDP.

The entire *cteam-lite* dataset takes up 129 Gb and thus may be downloaded by ftp [address upon publication], thus eliminating the practical difficulty in accessing raw whole genome genotype data for hundreds of samples.

Fast tools for processing hundreds of whole genome sequences

Cteam-lite is supplied with software (https://github.com/DReichLab/cTools): "cTools". This is comprised of three major components:

(i) *cascertain* which allows discovery of sites from whole genome data according to user specified ascertainment rules. Rules may be quite complex and are built from a mini-language, allowing for queries such as: "Identify sites where S_Yoruba_1 is heterozygous and either Altai or Denisova has a derived allele (chosen at random) EXCEPT where both Altai and Denisova are both heterozygous";

(ii) *cpulldown* which allows genotype calls to be extracted at user-specified filtering levels, given a set of known positions, such as those in a genotype array;

(iii) *cpoly*, which pulls down all the SNP sites that are polymorphic in a specified sample list from one or multiple bams.

The *cTools* software suite is designed to access data without unzipping individual *cteam-lite* files, though users may wish to do so for their own needs in which case the dataset expands to ~10 Tb.

Additional tools and data available

Our lab website (http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html) also supplies the full suite of software (the mapping and filtering and QC pipeline) used to generate the dataset, which we hope will allow researchers who generate their own data to combine their data with the data reported here.

The individual files available from the ftp site are:

Filtering tools:
CTEAM_TOPDIR/tools/filters/
- For generating inputs:
- cnv: bam2cnv
- hetfa: vcf2hetfa
- mappability75: datafile: a *universal* filter indicating bases where the 75-mer centered at the base is unique in human reference
- hs37d5 (Href): human reference
- Chimp: in human coordinates

Programs for running the filter engine:
- cmakefilter

Programs for postprocessing the filters:
- filtstats
- annotate.pl

cteam-lite dataset
- CTEAM_TOPDIR/cteam_lite/
- cteam-lite: access

Multi-sample VCF file
- CTEAM_TOPDIR/data/multi_sample_vcf/
- from the processing of Supplementary Information sections 1 and 2, where calls are made using single-sample genotyping using GATK
- from the process of Supplementary Information section 3, where calls are made based on *de novo* assembly, and multi-sample processing

References

1 Prufer, K. *et al*. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49, doi:10.1038/nature12886 (2014).
2 Meyer, M. *et al*. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226, doi:10.1126/science.1224344 (2012).
3 Fu, Q. *et al*. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449, doi:10.1038/nature13810 (2014).
4 Lazaridis, I. *et al*. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-413, doi:10.1038/nature13673 (2014).

Supplementary Information section 4

Comparison of genotypes obtained by different methods

Heng Li*

*To whom correspondence should be addressed: H.L. (hengli@broadinstitute.org)

Universal mask construction

To allow a fair comparison of several methods of SNP calling, we defined a *universal mask*. The *universal* mask is a sample independent mask that identifies complex regions in the human reference genome where variant calling can be challenging.

The *universal mask* has three components: (a) mapability mask; (b) low complexity regions; and (c) regions enriched with aberrant SNP calls from the 1000 Genomes Project. The final mask is the union of the three components. This *universal mask* is used as one of the inputs for the filters in Supplementary Information section 2.

(a) Mapability mask

At each position in the human reference genome, we extracted all possible 75-mers overlapping the position and mapped them back to the reference genome with BWA. We kept the position unmasked if 38 or more overlapping 75-mers cannot be mapped elsewhere with at most one mismatch or gap. The rest of positions are masked.

(b) Low-complexity mask

The low-complexity mask has three sub-components: a) low-complexity regions identified by the mDUST program [1] (command options "-w 7 -v 28"); b) homopolymers 7 bp or longer; c) DNA satellites and low-complexity regions as identified by RepeatMasker (extracted from the file "rmsk.txt.gz" from the UCSC Genome Browser: http://genome.ucsc.edu). These regions are merged together with 10 bp flanking added to each end. This gives the final low-complexity mask.

(c) Regions enriched with aberrant SNP calls

We acquired the samtools pre-filtered SNP calls on the 1000 Genomes Project phase III data. For each SNP, we computed from the genotype likelihoods the inbreeding coefficient, and the P-value under the Hardy-Weinberg equilibrium assumption. We focused on SNPs with negative inbreeding coefficient (i.e. excessive heterozygotes). We further selected SNPs satisfying one of two criteria: (1) $P<10^{-10}$; or (2) the P-value is below 10^{-5} and the total read depth is above 22,000 on the autosomes or 19,000 on the X chromosome. We clustered the selected SNPs that are close to each other, added 150bp of flanking bases to each cluster, and merged the resulting intervals to generate the final mask. This mask identifies regions susceptible to mis-assemblies in the human reference genome or common copy number variations (CNVs).

Properties of the *universal mask*

The unmasked region covers 87% of A/C/G/T bases in GRCh37, 93% of GenCode protein coding regions and 96% of pathogenic variants in the ClinVar database. The

[1] ftp://occams.dfci.harvard.edu/pub/bio/tgi/software/seqclean/

universal mask is a mild mask. It retains the majority of reference genomes, especially functional regions and variants.

Another commonly used mask for SNP calling is the regions used by the Genome-In-A-Bottle project[1]. The unmasked regions here covers 78% of A/C/G/T bases in the GRCh37 autosomes, 74% of GenCode coding regions and 79% of ClinVar pathogenic variants. This mask imposes a significant penalty on functional regions and is specific to sample NA12878. Our *universal mask* is more permissive (albeit at the cost of a higher error rate) and not influenced by variants in a single genome.

Comparison of Variant Call Sets

Additional variant call sets

The GATK call set (Supplementary Information section 1) excludes Indels and multi-allelic SNPs. To account for these additional short variants, we called variants with FermiKit-0.8[2] for the 263 fully public samples from Panel C of the SGDP. FermiKit assembles short reads into unitigs, maps them to a reference genome, and then calls variants without using a complex statistical model. This procedure is very different from a typical mapping-based SNP calling pipeline. We have also called SGDP samples from an earlier BWA-based mapping of a subset of samples using the Platypus[3] software. This call set is only discussed here only for the comparison purposes (we do not release the calls, as the underlying read mappings are out of date). The Platypus call set contains 259 out of the 263 public samples.

Comparison of call sets

Figure S4.1 shows the Euler diagram of the three call sets, restricted to the region not covered by the *universal mask*. The call sets largely agree, suggesting they are all of high quality. The GATK call set is smaller probably due to the more aggressive filtering (we retain SNPs at which at least one sample passes filter level 1). Of the FermiKit-, Platypus- and GATK-specific SNPs, 73%, 76% and 68%, respectively, are singletons or doubletons. The percentages are higher than the overall rate of singletons and doubletons: about 57% in all three call sets. In the following sections, we use the FermiKit call set which includes both indels and multi-allelic SNPs.

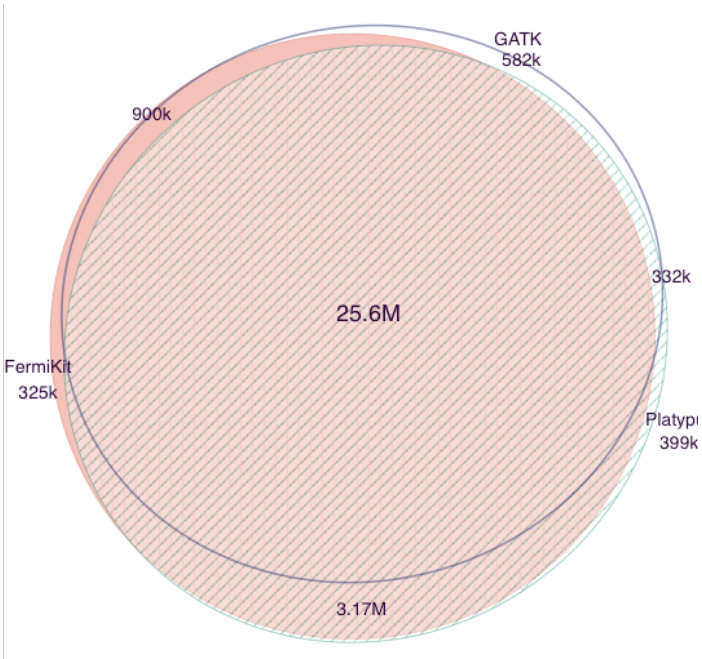


Figure S4.1. Comparing SNPs discovered by different methods. We compare GATK, FermiKit, and Platypus calls for samples and sites that overlap.

Characteristics of Variants

Overall statistics

Across all the autosomes, FermiKit called 37.4M SNPs, 3.0M insertions and 3.8M deletions relative to the human reference genome. In the high quality unmasked regions, it called 30.0M SNPs, 0.7M insertions and 1.4M deletions. The number of indels is affected more by masking because the *universal mask* includes low-complexity regions that are enriched in short tandem repeats.

Notably, the *universal mask* filters 20% of SNPs, although it has only masked 13% of the human genome, including long recent segmental duplications and centromeric regions that are inaccessible to 100 bp short reads. We speculate that a substantial fraction of the 7.4M (=37.4-30.0) SNPs in the masked regions are false positives due to CNVs or misassemblies in the human genome that have been filtered out by the *universal mask*. For this reason, we focus on the unmasked variants in the following analyses. It is important to note that our *universal mask* excludes most short-tandem repeats (STRs). Supplementary Information section 5 characterizes the STRs.

Allele frequencies

Suppose ϕ_{nk} is the rate of observing k non-reference alleles out of n haplotypes. Let $\psi_{nk}=k\phi_{nk}/\Sigma_j j\phi_{nj}$. Then, ψ_{nk} is the fraction of non-reference alleles on a single haplotype which occur k times out of n. If the sample haplotypes and the reference haplotype all come from a single uniform population, the Wright-Fisher expectation of ψ_{nk} is $1/n$. In the continuous form, if $\phi(x)$ is the allele frequency spectrum, let:

$$\psi(x) = \frac{x\phi(x)}{\int_0^1 y\phi(y)dy}$$

Here $\psi(x)dx$ is the probability of seeing, on a single haplotype, a non-reference allele of frequency in the range of $[x,x+dx)$. The Wright-Fisher expectation is $\psi(x)=1$.

In practice, population demography is not Wright-Fisher as the SGDP samples do not come from a homogeneous population and populations have not been constant in size over time. We therefore do not expect $\psi(x)=1$. Figure S4.2 shows the empirical $\psi(x)$ of each of seven regionally grouped populations for both SNPs and short indels. A point above 1 indicates excess in comparison to the ideal Wright-Fisher expectation. The empirical $\psi(x)$ of SNPs and indels are broadly similar. We observe an excess of rare variants, which could either be due to population expansions or the fact that our samples from multiple populations have substantial population structure. In addition, for non-African samples, variants at frequency of around 5% are depleted. We hypothesize that this is caused by the common bottleneck shared in non-Africans.

A related question is for a single sample, what is the fraction of heterozygous positions that have substantial non-reference allele frequency among other samples. Figure S4.3 shows that given a new East Asian and European sample, ~97% of heterozygous SNPs are seen at >1% frequency in the rest of SGDP. The percentage drops to 96% for Papuans, 85-90% for sub-Saharans of non-hunter-gatherer background, and as low as 80% for San samples.

Figure S4.2: Ratio of empirically observed rate of sites in a given frequency band vs. Wright-Fisher expectation. (a) SNPs. (b) Indels. Horizontal line at $\psi(x)=1$ shows the Wright-Fisher expectation.

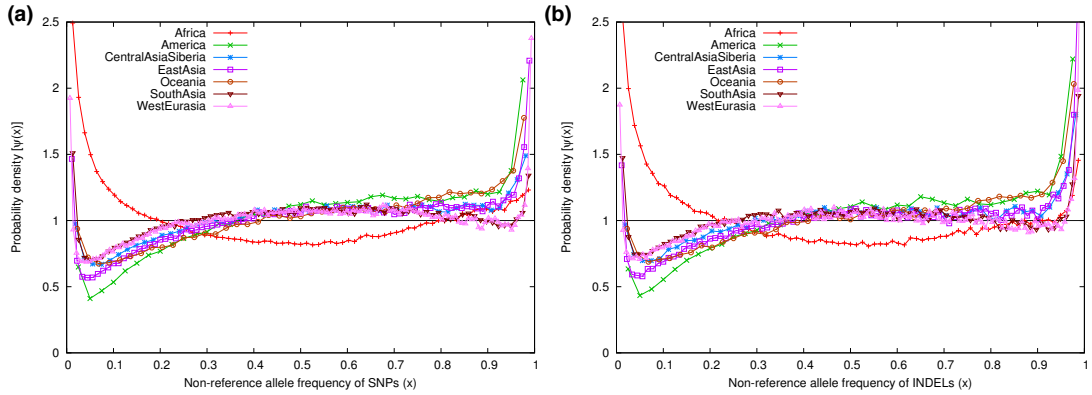
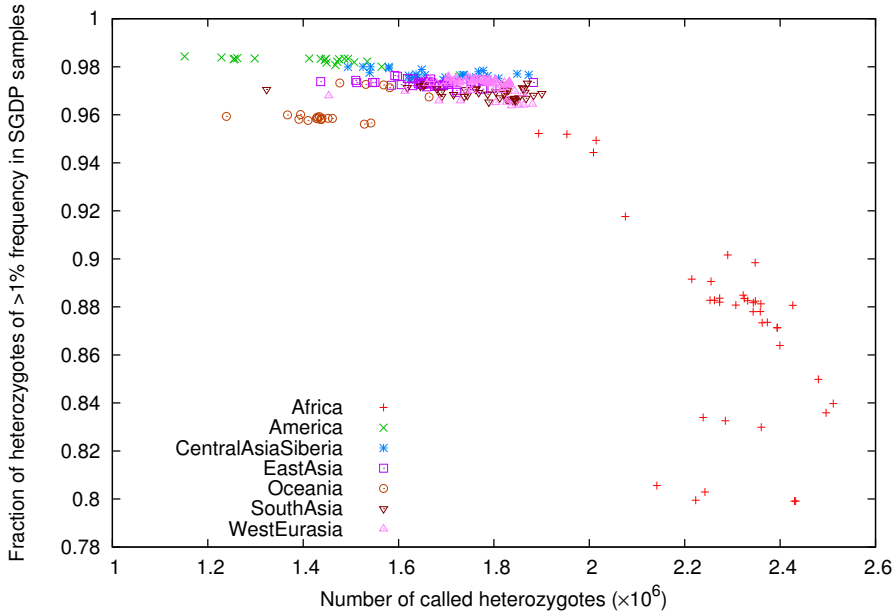


Figure S4.3: Fraction of heterozygous positions observed in a new sample that are at >1% frequency in the other SGDP samples.



Compositional biases

Figure S4.4a shows the fraction of CpG-related transitions as a function of minor allele frequency. We observe an excess of instances in which the ancestral state in characteristic CpG mutations is the major allele at the low-frequency end. This is unsurprising given that the mutation rate is known to be higher at CpG sites, which leads to more young mutations that have lower frequency in the population. Due to this effect, more CpG-related SNPs have C/G as the major allele at low frequency (blue dots in Figure S4.4b).

For transversions and non-CpG transitions, we observe a signal in the same direction albeit a weaker one: C/G tends to be the major allele at low frequency (Figure S4.4b). To understand what the ratio means, it is illustrative to examine a population at Wright-Fisher equilibrium with C/G-to-A/T mutation rate θ and A/T-to-C/G

mutation θ_w across both C/G and A/T sites. In this model, the number of C/G-major alleles at minor allele count $k<n/2$ is proportional to:

$$\frac{n-k}{n}\cdot\frac{\theta_s}{k}+\frac{k}{n}\cdot\frac{\theta_w}{n-k}$$

Similarly we can derive the number of A/T-major alleles. The ratio in Figure S4.4b is:

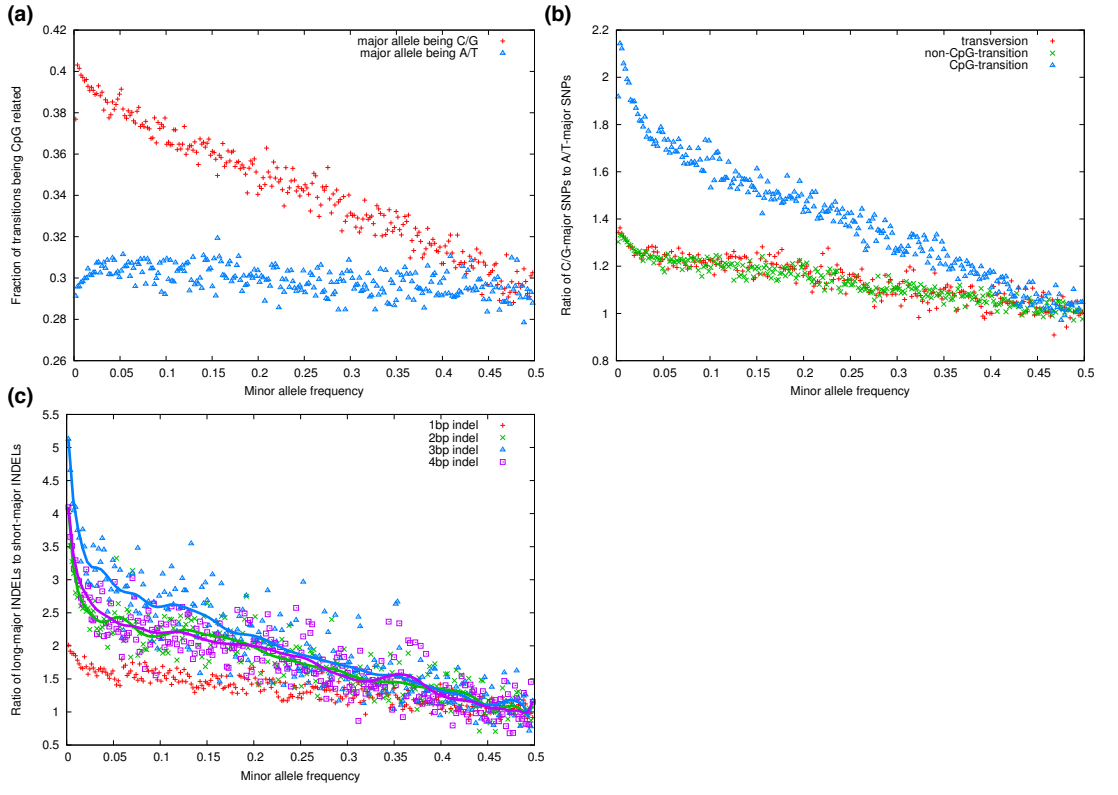
$$\frac{(n-k)^2\theta_s+k^2\theta_w}{k^2\theta_s+(n-k)^2\theta_w}$$

Or if we let $f=k/n$ being the minor allele frequency, the equation above becomes:

$$\frac{(1-f)^2\theta_s+f^2\theta_w}{f^2\theta_s+(1-f)^2\theta_w}$$

This ratio equals 1 at f=0.5 and approaches θ_s/θ_w at low f. Therefore, Figure 3.4b implies that even at non-CpG sites, there is a mutational bias favoring C/G-to-A/T mutations, in agreement with previous studies[4-6].

Figure S4.4: Over-representation of different types of mutation as a function of minor allele frequency.



It is important to point out that GC-biased gene conversation—which has the effect of causing a shift in the allele frequency spectrum toward GC-major alleles at sites that are GC/AT polymorphisms[7]—could also be playing a role in the patterns observed in

Fig. S4.4b. However, GC-biased gene conversation cannot be explaining entire signal, and in particular, cannot explain the stronger signal at CpG transitions.

We finally observe a bias in which the major allele tends to be longer than the minor at Indel sites (Figure S4.4c). This signal is even stronger that the compositional biases observed at SNPs, and could similarly be explained by a scenario in which mutation favors deletions over insertions.

To further probe these compositional biases, we examined whether the C/G-to-A/T bias is the same across populations. Given two populations, we draw a haplotype from each population and then collect C/G vs. A/T single-nucleotide differences between them. Let A be the number of times the first population has the C/G allele and B the number of times the first population has the A/T allele. Define $D=(A-B)/(A+B)$. The neutral expectation is $D=0$. We can compare all pairs in two populations to derive D between two populations[8].

We computed D for all pairs of the seven super-populations. Between any pair of non-African populations, D is below 0.001. The maximum Z-score is 3.2, which is not significant after accounting for the number of hypotheses tested. Between African and non-African populations, the D value is around 0.002. The Z-scores become significant (between 5.0 and 8.5), but at such small D values, the apparently significant Z-scores may be caused by reference biases or the higher heterozygosity of African samples, which generally make variant calling harder. Thus, we view these results as interesting, but do not view them as compelling evidence for a difference in C/G-to-A/T bias across populations.

We applied a similar approach to 1-4bp biallelic Indels and asked whether a population prefers the longer allele in an INDEL variant. We observed differences in this analysis: D values reach 0.062 to 0.100 between African and non-African populations (Z-scores over 50), suggesting that African samples tend to have longer alleles. We tried a more aggressive *universal mask* to filter out more potential low quality variant calls. The D values stay the same, though the Z-scores become smaller as there are fewer variants. We have also performed a similar analysis on the indel calls from the 1000 Genomes Project. The D values between African other populations are 0.026 (Z>24), 0.067 (>51), 0.059 (>51) and 0.053 (>46), when the other population is American, East Asian, South Asian and European, respectively. The magnitude of D is smaller. We speculate that the asymmetry of allele lengths is caused by artifact, though we have not been able to identify the source.

Functional analysis

We acquired ClinVar database version 20150806 and found that 509 out of 26,411 pathogenic non-reference alleles are present in the fully publically available samples. Each sample has 21.7 ± 9.2 (mean ± 2SD) potential pathogenic alleles, and 4.8 ± 4.7 pathogenic homozygotes on average. ClinVar may include common variants identified from GWAS. Having a pathogenic allele or even a homozygote does not necessarily imply disease status.

We annotated SNPs and Indels with Ensembl Variant Effect Predictor (VEP) version 80. The effect of a variant is associated with the transcript that harbors it. We always select the most significant effect if there are multiple associated transcripts. Thus if an

indel falls in the intron of one transcript but causes a frame-shift in another transcript, the effect of this indel is marked as frame-shift.

In average, each sample has 55.2 ± 13.2 gains of stop codons and 81.7 ± 17.8 frame-shift events. Our numbers are about twice as large as those obtained by MacArthur et al.[9] We can see four potential explanations for this difference. Firstly, loss-of-function (LoF) variants are more susceptible to artifacts. Before aggressive filtering and experimental validation, MacArthur et al. [9] identified over 300 LoF variants per sample. The error rate of our LoF variants may be higher than the overall rate, too. Secondly, gene annotations, the effect predicting software and the choice of transcripts may have a sizable influence on LoF[10]. MacArthur et al. were using a different pipeline and set of gene annotation. Thirdly, LoF variants are rare: 70% are singletons among our analyzed 263 samples. With 36 bp reads at <4-fold coverage (the authors were not using exome sequencing data from the 1000 Genomes Project), the previous study should have low power on singletons which cannot be rescued by imputation, either. Fourth, the authors aggressively filtered LoF variants, which may affect sensitivity (MacArthur, personal communication). In conclusion, our number of LoF variants per sample may not be inconsistent with previous studies. Our results are achieved without aggressive filtering and without experimental validation.

References

1 Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* **32**, 246-251, doi:10.1038/nbt.2835 (2014).

2 Li, H. FermiKit: assembly-based variant calling for Illumina resequencing data. *arXiv* **arXiv:1504.06574**.

3 Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46**, 912-918, doi:10.1038/ng.3036 (2014).

4 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475, doi:10.1038/nature11396 (2012).

5 Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A* **101**, 13994-14001, doi:10.1073/pnas.0404142101 (2004).

6 Sueoka, N. Directional mutation pressure, selective constraints, and genetic equilibria. *Journal of molecular evolution* **34**, 95-114 (1992).

7 Glemin, S. *et al.* Quantification of GC-biased gene conversion in the human genome. *Genome Res*, doi:10.1101/gr.185488.114 (2015).

8 Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet* **47**, 126-131, doi:10.1038/ng.3186 (2015).

9 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828, doi:10.1126/science.1215040 (2012).

10 McCarthy, D. J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome medicine* **6**, 26, doi:10.1186/gm543 (2014).

Supplementary Information section 5

Characterization of sequences missing from the reference genome GRCh38

Heng Li *

*To whom correspondence should be addressed: H.L. (hengli@broadinstitute.org)

Generating the dataset

We performed a *de novo* assembly of 254 SGDP samples using a pre-released version of FermiKit (https://github.com/lh3/fermikit).

For each sample and for each fosmid and BAC clone from GenBank (v203), we extracted >500bp segments on which each 101-mer is absent from the human reference genome GRCh38, including all ALT contigs. We mapped these segments to GRCh38 plus *HLA* (from IMGT/HLA 3.18.0), and dropped sequences that had fewer than 20 mismatches/gaps or over 99% identity to GRCh38+HLA. We dropped a subset of 15 samples that we found provided far more novel sequences than did other samples in the study. Manual BLAST checking indicated to us that most novel sequences identified from these samples are due to microbial contamination.

After obtaining individual novel sequences, we merged them (4.5Gbp sequence in total) and dropped sequences that were substrings of another novel sequence (1.0Gbp left). We further discarded sequences highly enriched in the GGAAT motif, a characteristic centromeric repeat, as well as sequences that were fully low-complexity according to DUST[1] (85Mbp left). We aligned the remaining novel sequences against each other in four rounds with different mapping settings and thresholds, in order to minimize redundant sequences (68Mbp left). We mapped the low-redundancy contigs to the "nt" database and removed those whose best matches were to unicellular organisms, which could reflect microbial contamination (65Mbp left). We applied two additional rounds of all-vs-all mapping between contigs to identify additional redundant segments. This gave 12,296 contigs over 500bp, totaling 13Mbp in length.

We ran RepeatMasker-4.0.5 (on RepBase-19.07) and classified the contigs into three categories: low repeat content, enriched with interspersed repeats, and enriched with centromeric repeats. We set a length threshold of 1000bp for the first two categories and 2000bp for the last. This left us with 2,385 contigs totaling 5.8Mbp. These SGDP-derived sequences, which we propose can be used as new decoy sequences to improve mapping efficiency, have been submitted to NCBI and made available at the GRC FTP site for public download.

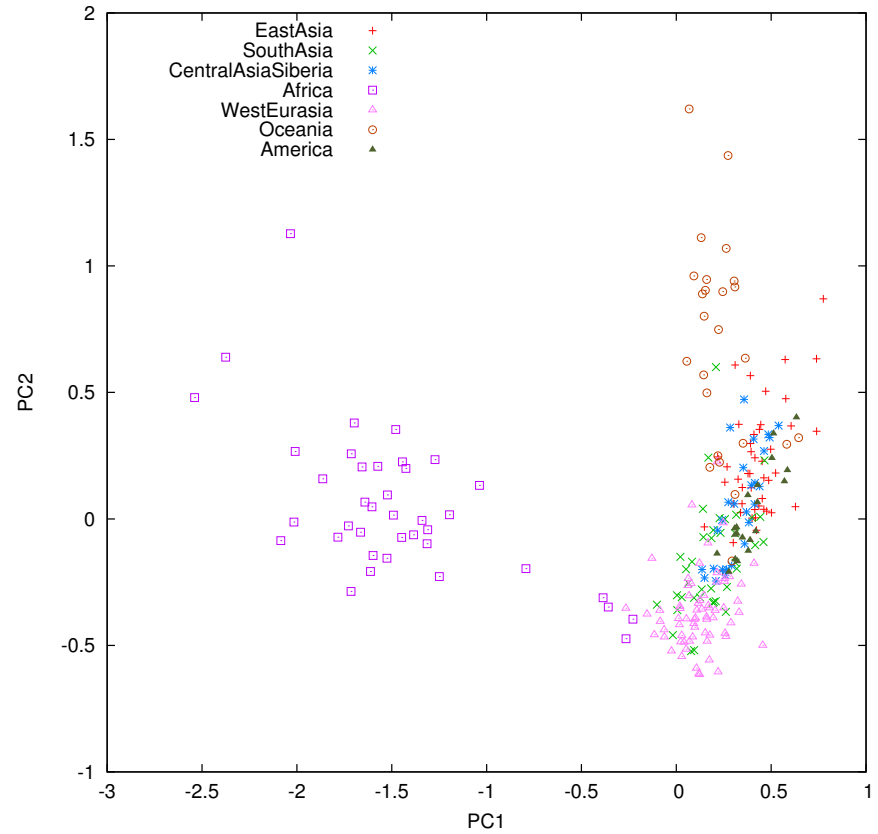
Population structure of previously unknown human sequences

We assembled 261 public SGDP samples with a more accurate version of FermiKit-0.8 (Supplementary Information section 3). We mapped the new contigs to GRCh38 plus the decoy and calculated the coverage of each contig for each sample. We define a sequence as *present* in a sample if over 90% of the sequence is covered, *absent* if less than 10% of the sequence is covered, and otherwise call its genotype *ambiguous*. With this classification, we find that 264 previously unreported sequences are present in all 261 samples. A total of 4 sequences, all from GenBank clones, are completely absent; these could be rare sequences in humans or contamination in GenBank. We

also found 1,515 sequences that are present in some samples but absent in others. They represent variation across populations. The rest of the previously unreported sequences are either present in some samples but ambiguous in others, or absent from some samples but ambiguous in others.

We performed principal component analysis (PCA) for the 1,515 previously unreported sequences that are variable among samples. To prepare a cleaner dataset, we dropped sequences that are ambiguous in more than 130 samples or that have <1% minor allele frequency among samples. We also identified two clusters of contigs that are strongly linked. All the contigs in one cluster map to a 120kbp CHM1 PacBio contig LCYE01013315.1. The SGDP samples either have the whole PacBio contig or do not have it, explaining why contigs in this cluster are strongly linked. We hypothesize that the strong linkage in the other cluster is due to a similar reason, but the haploid samples CHM1 and CHM13 do not carry this haplotype. For population genetic analyses, we excluded contigs in these two clusters, leaving 950 sequences for PCA (Figure S5.1). The plot is broadly similar to a PCA of SNPs: the first principal component (PC1) separates Africans from others, and PC2 separates West Eurasians, East Asians and Oceanians. The most extreme populations on PC2 are Oceanians, who are not typically the most extreme population in SNP PCAs. We hypothesize that this is due to the fact that they are least closely related to the individuals whose DNA was used to construct the human genome reference sequence of the non-Africans in SGDP, and thus their contribution in terms of never-before-reported sequences is larger than for West or East Eurasians[2]. Overall, these results show how variation in unreported sequence, like other types of variation, is sensitive to population structure.

Figure S5.1. PCA of 261 public SGDP samples across 950 sequences.



Variation at the Immunoglobulin heavy chain (IgH) locus

The human reference genome GRCh38 replaces the IgH locus in GRCh37 with the haplotype from the CHM1 haploid genome[3]. However, the GRCh37 haplotype is not added back to GRCh38 as an ALT contig. It is thus unsurprising that our analysis finds the GRCh37 sequence at the IgH locus. Its coordinates are chr14:106531321-25074982 in GRCh37 and chr14:106075079-106112834 in GRCh38.

For these two ~38kb sequences, we genotyped the 261 public samples by mapping their assemblies to the two haplotypes. We define a haplotype as called if at least 20,000bp of the contig is covered. With this threshold, 42 samples have both haplotypes, 125 have the GRCh37 haplotype only, 49 have the GRCh38 haplotype only and 45 samples have neither haplotype. If we decrease the coverage threshold by 10-fold to 2,000bp, the four numbers become 69, 128, 47 and 17, respectively. Thus, varying the coverage threshold does not have a qualitative effect on our assessment. Consistent with the observation of ref. [3], the GRCh37 haplotype has higher frequency.

The observation of the lack of both haplotypes in some samples suggests that there might be other rarer IgH haplotypes different from those in GRCh37 and GRCh38. However, assembling Illumina short reads in the IgH region is technically challenging due to multiple gene copies and segmental duplications. It is tempting to think that KI270846, which is listed as a third IgH haplotype in GenBank, could be present in some of the individuals who have neither the GRCh37 nor the GRCh38 haplotype. However, KI270846 is identical to the GRCh38 haplotype in the 38kb region; indeed, the relevant subsequence of KI270846 is made of CHM1 BAC AC247036 according to its tiling path. Thus, KI270846 is not an independent observation.

Variation around the *HLA-A* gene

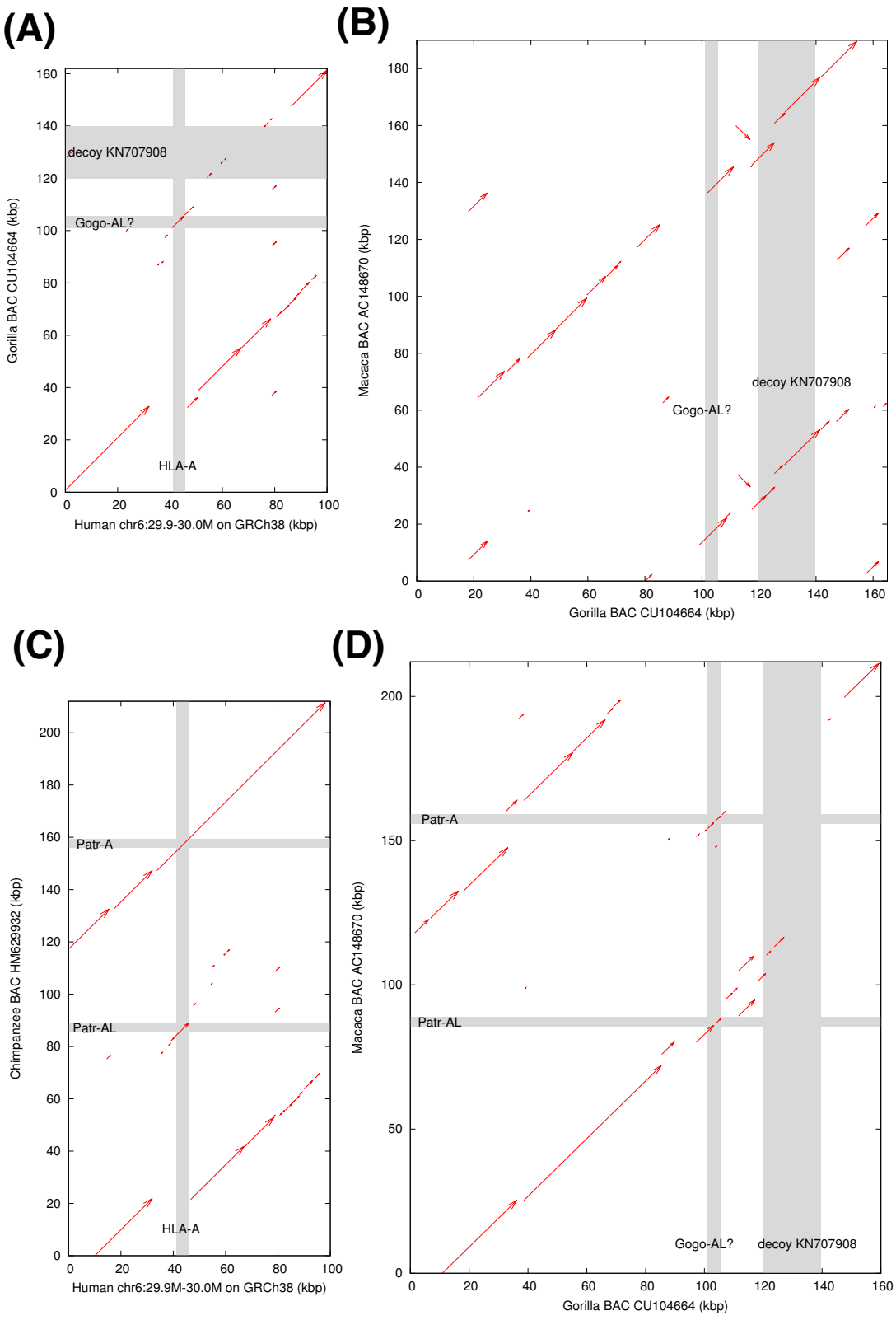
One of the sequences we detect, KN707908, is homologous to gorilla BAC CU104664 in the *MHC* region. According to a dot plot (Figure S5.2), this BAC lacks the *HLA-A* counterpart, but has an *HLA-A*-like gene *Gogo-AL* that is not present in humans[4]. It has two homologs in macaque BAC AC148670, but is absent from a chimpanzee BAC HM629932 in MHC (Figure S5.2). KN707908 is thus likely to reflect an ancient copy number variation (CNV) that arose over 20 million years ago and has persisted to the present. Among 261 public samples, 82 have the KN707908 haplotype, and it is easy to genotype: if a sample does not have this haplotype, the coverage of the decoy is <0.6%, and if a sample has it, the coverage is >97.5%.

KN707908 may also be linked with the *HLA-Y* pseudogene[5], believed to be a non-functional ortholog of *Patr-AL*/*Gogo-AL*. Out of 82 samples with KN707908, 32 have contigs that are homologous to *HLA-A* but are different from all known *HLA-A* alleles in the IMGT/HLA database. In contrast, of the 179 samples not carrying KN707908, only one sample has such contigs. Given that *Gogo-AL* is close to KN707908 on the gorilla BAC, linkage between *HLA-Y* and KN707908 seems plausible.

Conclusions

We identified over two thousand sequences that are absent from or highly divergent to the latest human reference genome GRCh38. The majority are variable among SGDP samples and are informative about population structure. We have shown that these sequences capture complex events in the HLA and IgH regions, suggesting that there are likely to be still undetected large-scale structural variations in these highly polymorphic regions. Long-read or clone sequencing might shed light in these cases.

Figure S5.2. Dot plot at HLA-A reveals a 20 million year old structural variant.



References

1 Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology* **13**, 1028-1040, doi:10.1089/cmb.2006.13.1028 (2006).

2 Green, R. E. *et al*. A draft sequence of the Neandertal genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).

3 Watson, C. T. *et al*. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet* **92**, 530-546, doi:10.1016/j.ajhg.2013.03.004 (2013).

4 Gleimer, M. *et al*. Although divergent in residues of the peptide binding site, conserved chimpanzee Patr-AL and polymorphic human HLA-A*02 have overlapping peptide-binding repertoires. *Journal of immunology* **186**, 1575-1588, doi:10.4049/jimmunol.1002990 (2011).

5 Williams, F., Curran, M. D. & Middleton, D. Characterisation of a novel HLA-A pseudogene, HLA-BEL, with significant sequence identity with a gorilla MHC class I gene. *Tissue antigens* **54**, 360-369 (1999).

Supplementary Information section 6

Worldwide variation in human short tandem repeats

Melissa Gymrek, Thomas Willems, David Reich and Yaniv Erlich*

*To whom correspondence should be addressed: Y.E. (yaniv@cs.columbia.edu) or M.G. (mgymrek@mit.edu)

Summary

We generated the most comprehensive catalog of short tandem repeat (STR) genotypes to date, based on 300 deeply sequenced human genomes. Genotypes show strong concordance with capillary electrophoresis and accurately recover population structure. We used this call set to characterize allele frequency spectra, analyze sequence determinants of STR variation, and to identify common loss of function alleles. STR genotypes are available in raw and interactive format at strcat.teamerlich.org.

Genotyping STRs

We analyzed STRs using lobSTR[1], a custom algorithm for genotyping short tandem repeats. We modified lobSTR's allelotyping tool to be able to call STRs directly from alignments generated by tools besides the lobSTR aligner. This greatly reduces the run time and allows rapid STR genotyping from large sequencing panels that have already been aligned using alternative indel-sensitive methods. We used raw reads aligned to GRCh37 using BWA-MEM (http://bio-bwa.sourceforge.net/) (version 0.7.10) with default parameters (Supplementary Information section 1). These alignments were used as input to lobSTR's allelotyper (Github revision 3.0.3.24-892e). We used optional parameters "--filter-mapq0 --filter-clipped --max-repeats-in-ends 3 --min-read-end-match 10" and a noise model trained on PCR-free sequencing data. We jointly genotyped samples at sites in lobSTR's reference panel: 1.6 million loci with motif lengths ranging from 1-6bp. The reference is part of the GRCh37 lobSTR resource bundle available at http://lobstr.teamerlich.org/download.html. Table S6.1 provides a summary of the reference panel.

Table S6.1 Composition of GRCh37 lobSTR reference panel. We list motifs that occur >5,000 times in the reference, in order from most to least common.

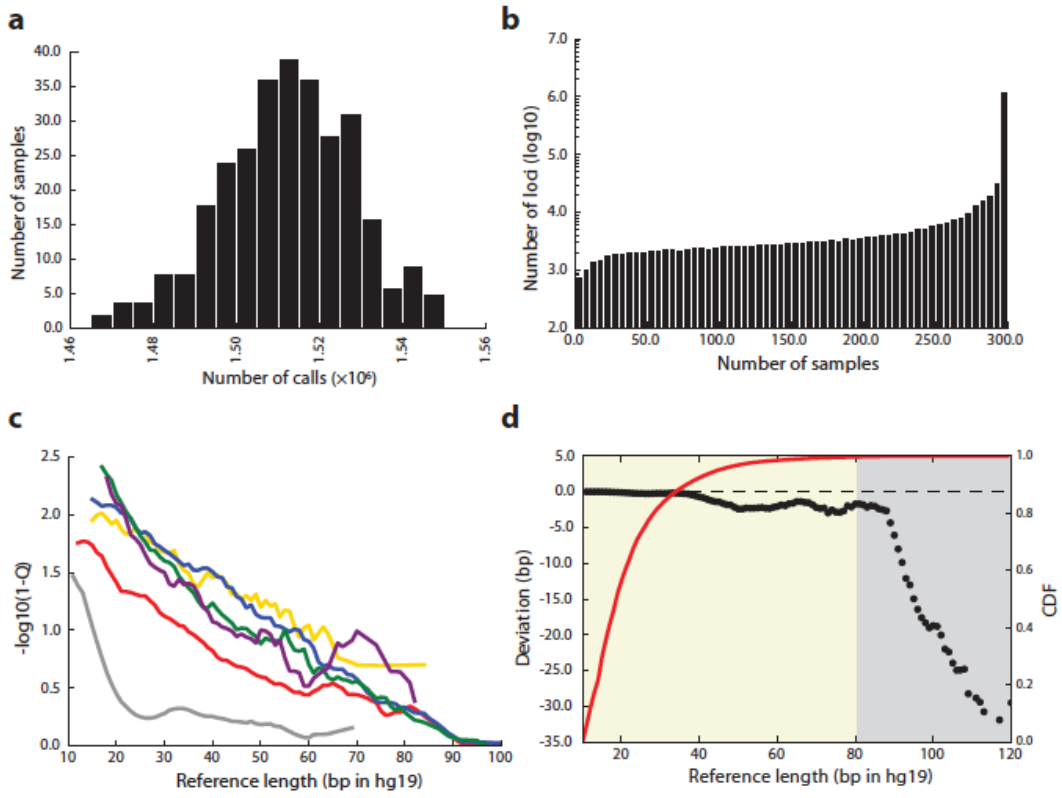
Motif length	No. of loci	% in reference	Common motifs	% genotyped (after filtering)
1	795,043	48.5	A, C	99.9 (70.3)
2	310,761	19.0	AC, AT, AG	96.2 (88.5)
3	84,869	5.2	AAT, AAC, AGG, AAG, ATC	97.6 (95.6)
4	262,179	16.0	AAAT, AAAC, AAAG, AAGG, AATG, AGAT, AGGG, ATCC, ACAT	94.3 (91.8)
5	106,481	6.5	AAAAC, AAAAT, AAAAG	97.4 (93.1)
6	79,246	4.8	AAAAAC, AAAAAT, AAAAAG	97.4 (93.3)
All	**1,638,516**	**100.0**		**97.9 (81.1)**

Quality controls

lobSTR generated genotypes for an average of 1.5 million loci per sample (Figure S6.1a) with an average of 15.3 informative reads (reads that completely span the

repeat region) for each autosomal call. All samples had call rates within 3 standard deviations of the mean. For population genetic analysis, we removed individuals from the Bergamo and Hazara populations, as some of these individuals were outliers. Each locus had genotype calls for an average of 280 samples (95%) (Figure S6.1b). We were not able to genotype 2% of loci in our reference. Most of these loci have allele lengths greater than 100bp that could not be spanned by Illumina reads. Genotype quality scores, which report the likelihood of the genotype call divided by the sum of likelihoods of all considered genotypes, tended to decrease for longer STRs and increase with motif length, with homopolymers showing significantly lower quality scores than other classes (Figure S6.1c). For the majority of loci, we found no directional bias in allele length compared to the reference allele. However, as the reference track increases, calls become biased toward shorter alleles, again reflecting the limitation of calling STR genotypes using 100bp reads (Figure S6.1d).

Figure S6.1: STR call set quality metrics. a. Distribution of the number of STR calls per sample. **b.** Distribution of the number of samples with calls for each STR. **c.** Mean genotype quality score decreases with the length of the STR. Each line represents a different repeat motif length (gray = homopolymers, red = dinucleotides, yellow = trinucleotides, blue = tetranucleotides, green = pentanucleotides, purple = hexanucleotides). **d.** Mean length deviation from the reference allele as a function of reference length (black). As the reference track increases in length, calls tend to be biased toward alleles shorter than the reference allele (black). The red line gives the Cumulative Distribution Function (CDF) of calls vs. reference length. Gray shading: loci that were filtered from analysis. Beige: loci retained for downstream analysis.



We subjected the resulting genotypes to stringent filtering to ensure high quality calls. We based our filters on coverage, call rate (percent of samples with a genotype call

for a given locus), and the metrics Q and DISTENDS reported in the VCF file generated by lobSTR. Q reports the genotype quality score as described above. DISTENDS reports the mean distance between the STR boundary and the end of the read. Specifically, we calculate the difference in distance between the STR and the left and right read ends, and take the average difference across all reads for a given call. We find that high quality calls tend to have DISTENDS close to 0, meaning there is no bias towards a specific end of the read on which the STR occurs. On the other hand large positive or negative DISTENDS scores often indicate that a locus has problematic alignments.

The specific filters listed below are described in the "Best practices for using BWA-MEM alignments with lobSTR" section of the lobSTR website. We filtered loci with the following properties:

- Average coverage <5×
- Average $-\log_{10}(1-Q)<0.8$
- Call rate <0.8
- Reference allele length >80bp

After filtering loci we additionally filtered individual calls with:

- Coverage <5×
- $-\log_{10}(1-Q)<0.8$
- Absolute value of DISTENDS score >20

After filtering, 1.3 million loci remained for analysis.

Validation

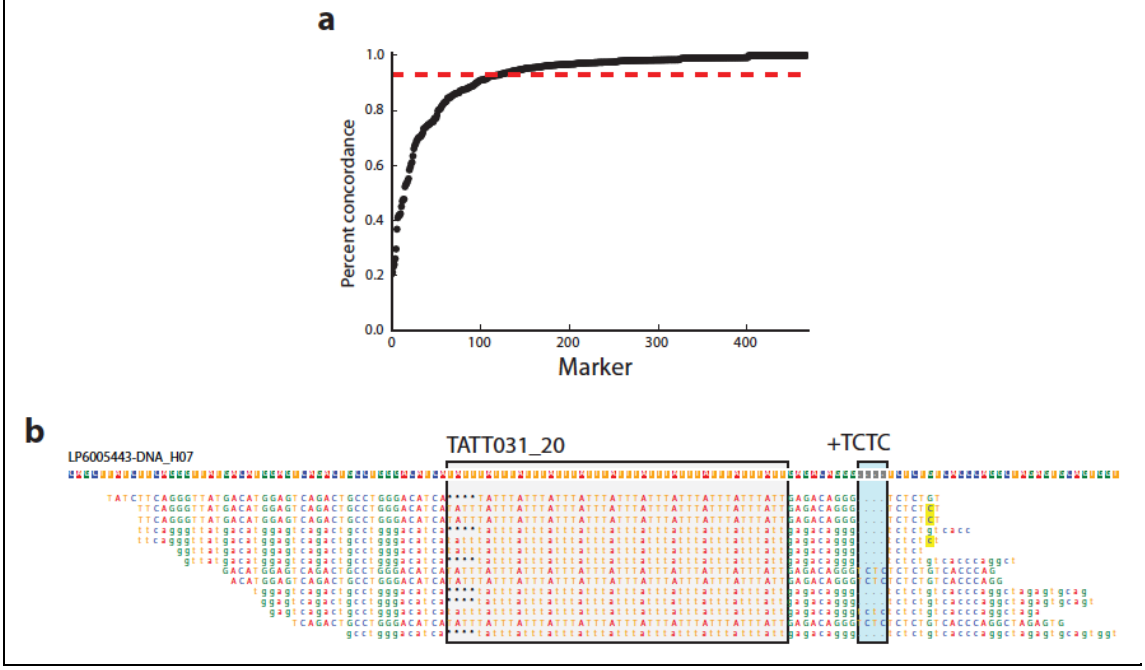
We compared lobSTR results to genotypes generated using capillary electrophoresis, the gold standard for STR genotyping. We evaluated concordance with two panels: Y chromosome STRs (mostly tetranucleotide loci), and the Marshfield set of mostly di- and tetranucleotide autosomal loci.

We obtained Y-STR genotypes for 39 loci for which there was data from capillary genotyping from the CEPH-HGDP website (ftp://ftp.cephb.fr/hgdp_supp9/genotype-supp9.txt). We calibrated capillary calls to the lobSTR format using the reference alleles annotated in Supplementary Table 5 of Gymrek *et al*.[2] As reported there, markers DYS481 and DYS594 are off by one unit in the CEPH data, and we corrected the lobSTR calls to reflect this. We discarded marker DYS640 due to ambiguous nomenclature. For 74 samples that overlapped between the SGDP dataset and the dataset on the HGDP website, we observed a genotype concordance of 99%.

We downloaded genotypes and additional metadata for the Marshfield markers for 627 loci from the Rosenberg lab website by Pemberton *et al*[3] (https://web.stanford.edu/group/rosenberglab/diversity.html) of which we were able to convert 468 capillary genotyped loci to loci in the lobSTR GRCh37 reference. Capillary genotypes were reported as the size of the PCR product and we converted these to lobSTR format as described on the lobSTR webpage (http://lobstr.teamerlich.org/validation-sets.html). We rounded all genotypes to the nearest repeat unit. A total of 127 samples overlapped between the SGDP dataset and this capillary dataset. The overall genotype concordance rate was 93%. We compared

STR dosage, defined as the sum of lengths of the two alleles, across methods and found strong correlation (r^2=0.92) between the two datasets (Extended Data Figure 2a). In discrepant calls, lobSTR tended to underestimate the true allele length compared to the capillary data, again reflecting a bias toward detecting shorter alleles due to the read length limitation. Notably, the majority of errors originated from a small set of loci (Figure S6.2a), with many errors potentially due to discrepancies in STR annotations between the datasets. For instance, marker TATT031_20 has a 4bp indel nearby the annotated STR sequence that is strongly linked to particular STR alleles. lobSTR only considers variation within the annotated sequence when making calls, whereas the capillary calls consider all length variation contained in the product amplified by PCR during genotyping, resulting in discordant genotypes. Thus, both methods are correct by their own definitions, despite the apparent discrepancy. An example discordant call affected by this issue is shown in Figure S6.2b.

Figure S6.2: Concordance between lobSTR and capillary genotypes. a. Concordance by marker, ordered from the marker with lowest to highest concordance. The red dashed line gives the overall concordance. **b.** Example marker with poor concordance between lobSTR and capillary data due to an annotation error. In this sample, marker TATT031_20 has a genotype of "-4,0" reported by lobSTR. However, the capillary data reports "-4,4", due to an extra 4bp "TCTC" indel (blue box) in the flanking regions that is linked with the STR allele "0". Because this indel is not included in the annotated STR sequence (gray box) lobSTR does not consider it when making a genotype call. We visualized the alignment using PyBamView[4].



We next sought to assess the accuracy of homopolymers in our data. These markers are not part of the capillary data discussed above and were excluded in previous studies of STR variation[5]. To this end, we tested whether the lobSTR calls from these loci could recapitulate known differences among population groups based on principal component analysis (PCA). As a positive control, we first analyzed autosomal tetranucleotides with heterozygosity greater than 10% that were called in at least 90% of samples. These loci represent a relatively high quality STR call-set. The 28,403 tetranucleotides passing the above filters were able to accurately recover

known population differences in these samples (Extended Data Figure 2b), with the first principal component separating non-African from African samples and the second primarily separating European and Asian samples. Remarkably, repeating the same analysis with 53,002 homopolymer loci, we were able to recover the majority of the structure seen by tetranucleotides (Extended Data Figure 2c), a testament to the quality of the calls in our catalog for these difficult-to-genotype loci.

STRs improve resolution of population structure inference

Encouraged by the ability of STR calls to distinguish population structure, we sought to determine whether STRs increase the resolution of population inference beyond that which can be obtained by genome-wide SNPs. We used the smartpca tool from the EIGENSOFT[6] package to compute F_{ST} and block jackknife standard errors between all pairs of populations. We first computed F_{ST} and standard errors using SNPs (Supplementary Information section 2). We obtained genotypes for 1,152,838 autosomal sites from a panel of SNPs known to be informative for population structure, built from a union of SNP Panels 1 and 2 of ref. [7]. We then repeated this analysis using a dataset that combined SNP and STR genotype data. To encode STRs in bi-allelic format, we followed the convention suggested by Patterson et al.[8], and encoded each STR allele in the frequency range of 5-95% as a separate bi-allelic marker. This gave 357,863 STR "markers" from 160,530 unique STR loci for a total of 1.51 million markers for the combined SNP+STR analysis. Whereas the two datasets gave highly concordant F_{ST} values (slope of best fit line = 0.96, Pearson r^2=0.999) (Extended Data Figure 2d), the combined dataset has decreased standard errors compared to SNP variation alone (slope = 0.86), documenting the added value provided by STRs for discerning population structure (Extended Data Figure 2e).

Patterns of STR variation

We used our catalog to examine overall trends in polymorphism at STRs. Of the 1.3 million genotyped loci, 32.2% show more than two common alleles (defined as having an allele frequency ≥ 0.01), and some loci have more than 20 common alleles. The remaining loci are either fixed across all individuals (47.6%) or have only two common alleles (20.5%). These patterns changed significantly when stratifying by motif length, with longer motif lengths showing less variability. For instance, only 23% of homopolymers are fixed compared to 70% of tetranucleotides. (Figure S6.3).

As has been previously shown, we found that STR variability depends strongly on properties of the STR itself and on local sequence features. We examined the ability of these features to explain differences in variability for all STR loci with at least two common alleles. We used heterozygosity as a metric of variation, which is defined as $1 - \sum_{i=1}^{n} p_i^2$, where p_i is the frequency of allele i and n is the total number of alleles. As mentioned above, heterozygosity tends to decrease with motif length. Additionally, we found that heterozygosity is positively correlated with STR sequence purity ($r = 0.21$, $p<10^{-200}$) and reference track length ($r = 0.17$, $p<10^{-200}$) (Figure S6.4). Both these observations agree with previously reported results[5,9]. We also observed a positive correlation with local recombination rate ($r = 0.028$, $p<10^{-209}$) (deCODE recombination maps[10] available on the UCSC genome browser). A joint linear model including all of these features explained 53% of variation in heterozygosity across loci. When restricting to STRs with no sequence imperfections (interruptions in the STR), these features explained 70% of variation.

Figure S6.3: Allele frequency spectra of STRs. a. Distribution of the number of common alleles per locus. **b**. Stratification by motif length (gray=homopolymers, red = dinucleotides, yellow = trinucleotides, blue = tetranucleotides, green = pentanucleotides, purple = hexanucleotides).

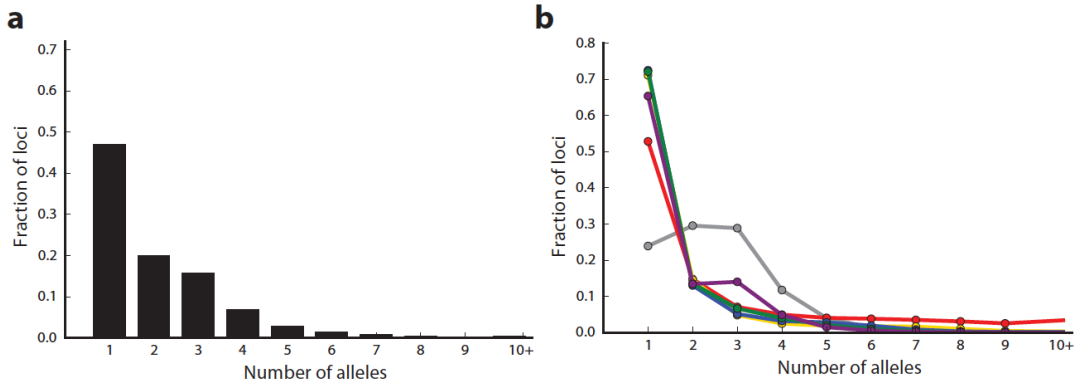
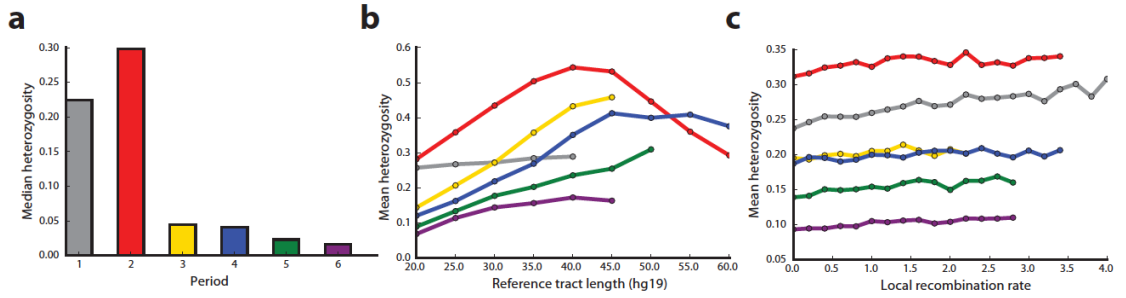


Figure S6.4: Sequence determinants of STR variation. a. Median heterozygosity by motif length. STRs with longer motif lengths tend to be less polymorphic. **b., c.** Mean heterozygosity as a function of reference track length and local recombination rate (gray = homopolymers, red = dinucleotides, yellow = trinucleotides, blue = tetranucleotides, green = pentanucleotides, purple = hexanucleotides).



Potential loss-of-function variants at STRs

We used our catalog to identify STRs in coding regions with common loss-of-function (LoF) variants, which we identified as frameshifting variants in coding exons as defined by Refseq. We restricted to alleles found in at least 10 individuals. Seventeen loci with potential common frameshifts passed these criteria, five of which have a frameshift as the major allele (Table S6.2). Four of the five common LoF alleles with periods 2-6 reported by Willems *et al*. using an independent dataset are included in our list (*TMEM254*, *GP6*, *FAM166B*, and *DCHS2*), and more than half were reported in dbSNP, suggesting that these putative LoF do not represent genotyping errors.

In 13 of the 17 cases, the potential LoF variant occurs in the last exon of the gene or toward the end of a single-exon gene, reducing its potential impact on protein function. The variants in *TMEM254* and *LFNG* occur in an internal exon. In both cases there are alternative transcript annotations that do not contain the affected exons. The putative LoF variants for *PTEN* and *RYK* occur in the first exons of these genes. On visual inspection, the CCG repeat for *RYK* occurs in a difficult-to-align GC-rich area and likely represents an alignment artifact. The variant in *PTEN* is fixed

at a 1bp deletion from the reference sequence adjacent to the CGG repeat. This deletion is annotated as a 1bp intron in Refseq (Figure S6.5). Notably this region is not annotated as coding by Ensembl, Gencode, or UCSC and the frameshift allele is fixed across all samples, suggesting an error in gene annotation. In conclusion, most common STR frameshift variants are unlikely to affect protein function.

Figure S6.5 The major allele at an STR in *PTEN* is an apparent frameshift from the reference sequence. The red box denotes the CGG repeat. The 1bp deletion at the adjacent "T" nucleotide is fixed across samples, has poor conservation compared to surrounding bases, and is not annotated as a coding region in other gene annotations, suggesting it may in fact be a misannotation and not a true frameshift variant.

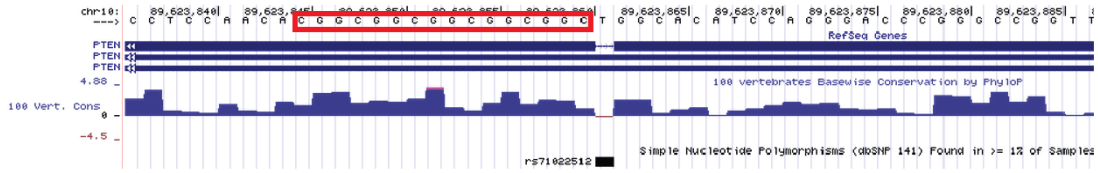


Table S6.2 Common loss-of-function alleles at STRs. We give the combined allele frequencies of all frameshift alleles for each locus. dbSNP data is from versions 141 and 142. * entries are LoF alleles previously reported by Willems, *et al*. [+] entries are low confidence alleles likely due to alignment artifacts or stutter errors.

STR Locus	Gene	Motif	LoF allele(s)	dbSNP
chr13:51530580	RNASEH2B	A	1bp (0.030)	rs200320729 (-/A)
chr14:23528485	ACIN1[+]	AGAGGG	-2bp (0.030)	
chr10:81841429	TMEM254*	AAAG	-4bp (0.034)	rs143538725 (-/AAAG)
chr3:133969414	RYK[+]	CCG	1bp (0.036)	
chr15:83677271	C15orf40	A	1bp (0.078)	
chr19:55526092	GP6*	ACAG	4bp (0.093)	rs138680589 (-/CAGA)
chr5:147861098	HTR4[+]	AAAAAG	1bp, -1bp (0.095)	
chr12:55820959	OR6C76	A	-1bp (0.218)	
chr20:3026346	GNRH2	CCCCG	5bp (0.320)	
chr16:58577316	CNOT1	A	-1bp (0.367)	
chr9:35561913	FAM166B*	ACCC	1bp, -8bp (0.402)	rs143266743 (-/CCCACCCT)
chr6:31380147	MICA	AGC	-1bp, -4bp, 2bp, 11bp (0.810)	rs547446871 (-/G) and rs41293539 (-/CT/CTGCTGCT /CTGCTGCTGCT)
chr7:2552851	LFNG	ATCC	4bp, -4bp (0.422)	
chr4:155244402	DCHS2*	AAAC	-4bp (0.846)	rs140019361 (-/TTTG)
chr10:125780753	CHST15	C	-1bp (0.895)	rs5788645 (-/C)
chr10:89623845	PTEN	CCG	-1bp (1.000)	rs71022512 (-/A)
chr5:72743281	FOXD1	CCG	2bp (1.000)	rs587745355 (-/GC)

Conclusion

We have presented the highest quality catalog of STR variation to date, which can serve as a reference panel of STR polymorphisms across diverse populations. Additionally, our dataset provides unprecedented opportunities to study STR variation that were not possible using previous studies either due to the small number of markers or to the low quality of individual genotypes[5]. Importantly, it contains the first panel of previously inaccessible homopolymer genotypes and allows in-depth study of these extremely polymorphic loci for the first time. We envision that this dataset will be an invaluable resource for future studies of STR polymorphism.

References

1 Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome research* **22**, 1154-1162, doi:10.1101/gr.135780.111 (2012).

2 Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339**, 321-324, doi:10.1126/science.1229566 (2013).

3 Pemberton, T. J., Sandefur, C. I., Jakobsson, M. & Rosenberg, N. A. Sequence determinants of human microsatellite variability. *BMC genomics* **10**, 612, doi:10.1186/1471-2164-10-612 (2009).

4 Gymrek, M. PyBamView: a browser-based application for viewing short read alignments. *Bioinformatics* **30**, 3405-3407, doi:10.1093/bioinformatics/btu565 (2014).

5 Willems, T. *et al.* The landscape of human STR variation. *Genome Res*, doi:10.1101/gr.177774.114 (2014).

6 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-909, doi:10.1038/ng1847 (2006).

7 Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, doi:10.1038/nature14558 (2015).

8 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS genetics* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006).

9 O'Dushlaine, C. T. & Shields, D. C. Marked variation in predicted and observed variability of tandem repeat loci across the human genome. *BMC genomics* **9**, 175, doi:10.1186/1471-2164-9-175 (2008).

10 Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature genetics* **31**, 241-247, doi:10.1038/ng917 (2002).

Supplementary Information section 7
Variability in heterozygosity and mutations load across populations

David Reich*, Heng Li and Nick Patterson

*To whom correspondence should be addressed: (reich@genetics.med.harvard.edu)

Statistics

Define the derived allele frequency at position i in a sample (a single individual or a group of individuals) from population A as p_A^i, and in a sample from population B as p_B^i. The expected number of alleles that are derived in population A but not B is $p_A^i(1-p_B^i)$, and derived in population B but not A is $(1-p_A^i)p_B^i$. This allows us to define the expected number of alleles $L_{A,not\ B}$ that are derived in population A but not B and similarly $L_{B\ not\ A}$ integrating over all N nucleotides i in the genome:

$$L_{A,\ not\ B} = \sum_{i=1}^{N} p_A^i(1-p_B^i) \quad \text{(S7.1)}$$

$$L_{B,\ not\ A} = \sum_{i=1}^{N} (1-p_A^i)p_B^i \quad \text{(S7.2)}$$

We define the normalized differences of accumulated mutations on the sample A lineage to that on the sample B lineage since they diverged in a compartment of the genome as:

$$D(A,B,Chimp) = \frac{L_{B,\ not\ A} - L_{A,\ not\ B}}{L_{B,\ not\ A} + L_{A,\ not\ B}} \quad \text{(S7.3)}$$

This is similar to the quantity analyzed in [1], although there the statistic was expressed as a ratio $R(A,B,Chimp) = L_{A,\ not\ B}/L_{B,\ not\ A}$ and the question of interest was if this quantity was consistent with 1 (here, we instead test whether $D(A,B,Chimp)$ is consistent with 0).

We computed this on both chromosome X (excluding the pseudo-autosomal regions) and on the autosomes for all pairs of samples in the dataset.

We also computed divergence per base pair over all sites that passed filters for both samples being compared. For the numerator, we analyzed all sites that were called as polymorphic in the dataset. For the denominator, the number of sites was very large making computation impractical (given the large number of sample pairs), so we sampled every 100th nucleotide, and then multiplied by 100. The total number of counts in the denominator is an order of magnitude larger than the numerator even after the 100-fold downsampling, so we do not expect the fact that we did not count all sites in the denominator to add substantial noise.

$$Div(A,B) = \frac{L_{A,\ not\ B} + L_{B,\ not\ A}}{Number\ of\ sites\ passing\ filters\ in\ both\ samples\ A\ and\ B} \quad \text{(S7.4)}$$

Data curation

For each pairwise comparison of samples, we initially restricted to nucleotides where both individuals had a genotype passing filter level 9, and where we had an ancestral allele assignment from comparison to chimpanzee (*PanTro2*). We use chimpanzee to determine an ancestral allele instead of a consensus based on multiple primates[2], since we were concerned that the algorithm used to determine the consensus could produce a bias in population genetic

analyses due to its reliance on the human reference sequence (which is of predominantly European ancestry[3]). Use of the chimpanzee genome is expected to produce an incorrect assignment at about a percent of SNPs, too small to cause a substantial bias for the analysis reported here or for population genetic analyses such as *D*-statistic tests of admixture[3]. For the convenience of users of this dataset, in the combined VCF file (Supplementary Information section 3), we not only provide an ancestral allele assignment based on chimpanzee, but also based on multiple primates.

To obtain as clean a dataset as possible, we restricted all analyses to populations that were represented in our dataset by at least two samples, and for which we could identify subsets of at least two samples that had similar statistical profiles.

We defined a statistic $Q_{VW}(Stat)$ that measured how similar two samples *V, W* were with respect all other samples in the dataset. For each statistic *Stat* of interest, we looped over all samples in the dataset (excluding *V* and *W*), computing the sum of the squared difference between V and that sample, and W and that sample.

$$Q_{VW}(Stat) = \sum_{Z=1}^{all\ samples\ except\ for\ V\ and\ W} (Stat(V,Z) - Stat(W,Z))^2 \qquad \text{(S7.5)}$$

This gives a measurement of how similar a given statistic "*Stat*" (either *D(A,B,Chimp)* or *Div(A,B))* is for samples *V* and *W* from the same population. If the data were error-free, the statistics would be expected to be consistent for the two samples assuming the populations were homogeneous. Thus, requiring that they be similar restricts to pairs of samples that likely both have low error rates (or alternatively, error rates of the same magnitude).

We rank-ordered Q_{VW} over all possible pairwise comparisons, for all four statistics of interest, (*D(V,W,Chimp)* and *Div(V,W)* on both chromosome X and the autosomes). We used human judgment to identify a cut-point beyond which there was evidence for samples with systematic difference in their genotype calls relative to other sample pairs in the dataset. When a pair of samples had a Q_{VW} in the tail and were the only two from that population in the dataset, we filtered out both. When there were more than two samples from a population, we looked for individuals overrepresented in the tail, and filtered those out.

Our filtering steps took us from 300 genomes to 235 (covering 108 distinct populations). The reasons for filtering were as follows. These filters were applied in order, so that once a sample was filtered for one reason we did not filter it out for another:

22 samples that were from panel B and thus processed differently from the majority of others: B_Australian-3, B_Australian-4, B_Crete-1, B_Crete-2, B_Dai-4, B_Dinka-3, B_French-3, B_Han-3, B_Ju_hoan_North-4, B_Karitiana-3, B_Mandenka-3, B_Mbuti-4, B_Mixe-1, B_Papuan-15, B_Sardinian-3, B_Yoruba-3, BR_Kashmiri_Pandit-1, BR_Kharia-1, BR_Kurumba-1, BR_Mala-1, BR_Onge-1, BR_Onge-2

17 samples based on only having one sample per population: S_Albanian-1, S_Altaian-1, S_Atayal-1, S_Chane-1, S_Chechen-1, S_Chukchi-1, S_Czech-2, S_Eskimo_Chaplin-1, S_Hawaiian-1, S_Itelman-1, S_Khonda_Dora-1, S_Kongo-2, S_Maori-1, S_Norwegian-1, S_Polish-1, S_Samaritan-1, S_Somali-1, S_Daur-2

12 samples based on a very different autosomal divergence vector to other samples relative to others from the same population: S_BantuTswana-1, S_Gambian-1, S_Mixtec-1,

S_Mozabite-1, S_Thai-1, S_Tlingit-1, S_BantuTswana-2, S_Gambian-2, S_Mixtec-2, S_Mozabite-2, S_Thai-2, S_Tlingit-2

4 samples based on a very different autosomal D(A,B,Chimp) vector relative to others from the same population: S_Russian-1, S_Finnish-3, S_Naxi-2, S_Russian-2

2 samples based on a very different X chromosome divergence vector to others from the same population: S_Masai-1, S_Masai-2

2 samples based on a very different X chromosome D(A,B,Chimp) vector relative to others from the same population: S_Jordanian-1, S_Papuan-3

5 samples based on missing X chromosome data in an initial processing, for themselves or a second sample: S_Finnish-1, S_Finnish-2, S_Mansi-1, S_Mansi-2, S_Palestinian-2

Our filtering meant that all samples within the same population had a correlated vector of *D(A,B,Chimp)* relative to samples from other populations. Thus, we pooled samples from each, and averaged all statistics from individual genome comparisons for a given pair of populations, to obtain a slightly more precise statistic to represent that pair of populations.

Worldwide variation in heterozygosity on chromosome X and the autosomes

We restricted all subsequent analyses to the 235 samples passing the filters of the previous section, and to nucleotides where both individuals being compared had a genotype passing the strongest filter (filter level 9), and where there was an ancestral allele.

For each population, we computed expected heterozygosity (number of differences per base pair between pairs of chromosomes from the same population) on chromosome X and the autosomes (Fig. S7.1). For this analysis we estimated heterozygosity using the *Div(A,B)* statistic comparing across samples from the same population. This computation is not based on the number of differences between two chromosomes of the same individual, which adds robustness to our analyses, as accurately calling heterozygous genotypes within diploid individuals is a difficult problem with substantial rates of false-negatives and false-positives. Encouragingly, we find that in our chromosome X data curation (previous section), we do not tend to find outliers at a higher rate in comparisons of males and females than in comparisons of two individuals of the same sex. Thus, the genotyping appears to be sufficiently accurate that the profound difference in females and males on chromosome X—with one sex being diploid and the other being haploid—is not causing a measurable bias.

Lower X-to-autosome ratio in pygmies than in other sub-Saharan Africans

Figure S7.2 plots the ratio of heterozygosity on chromosome X to that on the autosomes. We replicate the previous finding of a higher X-to-autosome heterozygosity ratio in sub-Saharan Africans than in non-Africans[4], while generalizing this result by showing its applicability to a much wider diversity of sub-Saharan Africans than has previously been analyzed (including Khoesan) and a much wider diversity of non-Africans (including New Guineans, Australians, Native Americans, Near Easterners, and indigenous Siberians).

The one exception to the uniformly higher X-to-autosome heterozygosity ratio in sub-Saharan Africans than in non-Africans is in Pygmies (eastern Mbuti and western Biaka). This is illustrated in Fig. 1b and Fig. S7.3, which shows a scatterplot of heterozygosity on the autosomes, against the X-to-autosome heterozygosity ratio. There are two primary clusters: sub-Saharan Africans, and all other populations. Within these clusters, there is no visually

evident differentiation among groups in X-to-autosome ratio with the exception of the two Pygmy groups, who have high autosomal heterozygosity, but relatively low X-to-autsome ratios (for Mbuti, closer to non-Africans than to Africans).

Figure S7.1. *Heatmap of heterozygosity per base pair on the autosomes*

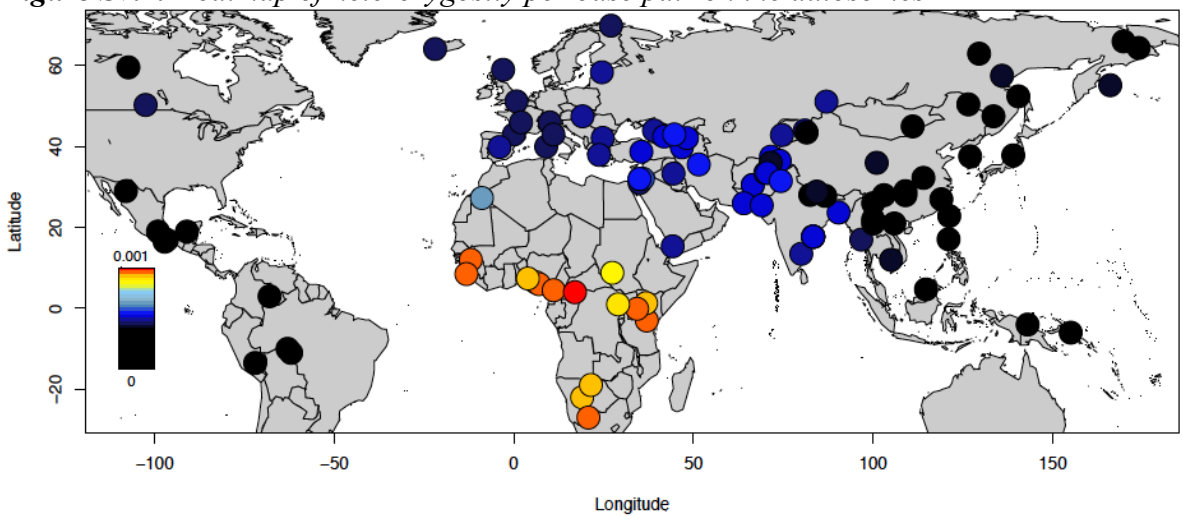
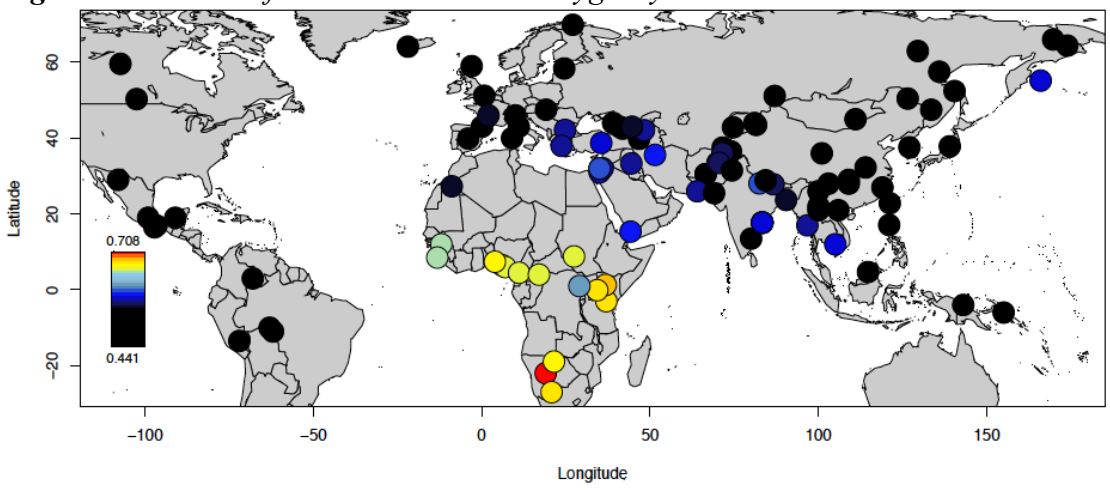


Figure S7.2. *Ratio of X-to-autosome heterozygosity*



To evaluate the robustness of this observation, we repeated the analysis removing 36% of chromosome X that falls within regions that have been identified in two independent publications as consistently affected by strong linked selection in great apes[5] or in the human-chimpanzee, gorilla ancestral population[6]. Specifically, we masked out from our X chromosome analysis the union of all regions identified as under strong selective constraint in those two studies, with the exception that we did not include in the mask the list of regions identified in ref. [5] as discovered as being under constraint in humans (we wished our mask to be built in a way that was blinded to patterns of genetic variation in humans). After applying this mask, we find that the X-to-autosome ratio is 1.21-fold higher on average across human populations, reflecting the profound selective constraint in the masked regions. However, the empirical patterns of differences across human populations persist (Figure S7.3). In particular, as shown in Table S7.1, when we remove the selectively constrained regions, we find that the reduction in X-to-autosome heterozygosity ratio in non-Africans compared to Africans, and in Pygmies compared to non-Pygmies, if anything grow larger.

Figure S7.3. Heterozgyosity on the autosomes vs. X-to-autosome heterozygosity ratio. *There are two main clusters with respect to the X-to-autosome heterozygosity ratio: sub-Saharan Africans and all others. However, the pygmies shown in red triangles have a pattern that is surprising for sub-Saharan Africans of high heterozygosity and relatively low X-to-autosome ratio. (A) Analysis with pseudoautosomal regions of chromosome X removed. (B) Analysis with regions of functional constraint across diverse ape populations removed.*

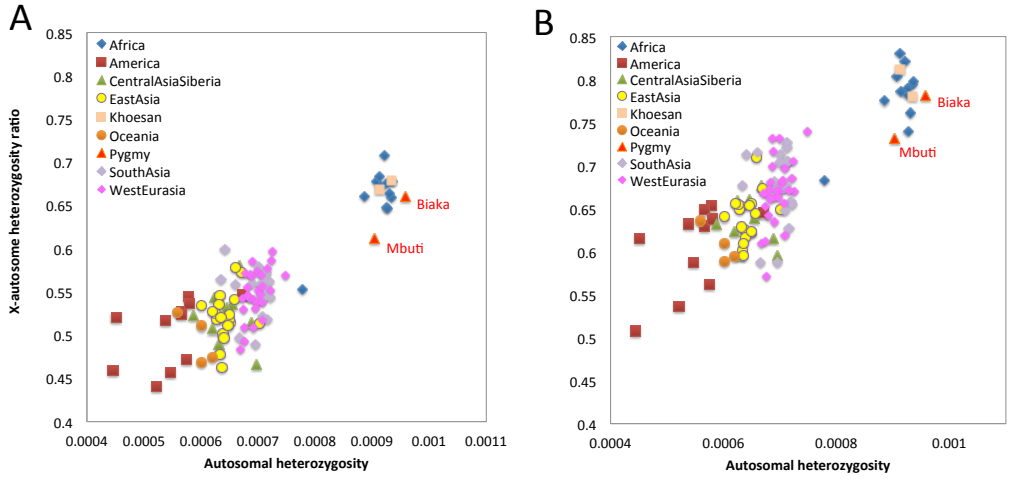


Table S7.1. Disparities in the X-to-autosome heterozygosity ratios are not reduced by restricting to regions that are not selectively constrained.

Population comparison	All chromosome X excluding PARs	Excluding 36% of chromosome X known to be selectively constrained
All non-African / All African	0.836	0.821
All pygmy / All non-pygmy Africans	0.967	0.966
Mbuti / All non-pygmy Africans	0.936	0.929

These results based on excluding the third of chromosome X most affected by linked selection provide new support for previous claims that the reduction in the X-to-autosome heterozygosity ratio in non-Africans relative to Africans is driven at least in part by demographic history [4,7-9]. The claim was initially made based on sub-dividing chromosome X and the autosomes based on bins of B-value (a proxy for selective constraint at linked sites[10]), and then computing the empirical ratios of X-to-autosome heterozygosity in Africans and non-Africans in each bin and observing that the ratios did not change[8,9]. A potential critique is that levels of selective constraint are difficult to compare on chromosome X and the autosomes, so it is not clear whether the analyses in [8,9] adequately control for similar levels of selective constraint. Our analysis, by contrast, cannot be confounded by differences in the scales of selective constraint on the autosomes and chromosome X. We simply mask strongly constrained loci on chromosome X, and find that the overall difference between Africans and non-Africans remains the same or possibly becomes greater, suggesting that the observed patterns are unlikely to be explained by selection.

The novel observation that comes from our analysis is that there also appears to be a substantial reduction in the X-to-autosome ratio in pygmies relative to other sub-Saharan Africans (Figure S7.4 and Table S7.1). What history could explain these observations?

It is known that Pygmy populations are admixed with ancestry of non-Pygmy origin[11,12], and as shown in Figure 2 in western Biaka pygmies the admixture is from people related to

present-day Bantu speakers and dates to the last few thousand years. From anthropological and genetic studies of mitochondrial DNA and the Y chromosome, it is known that this mixture is highly sex-biased, with matings between non-Pygmy fathers and Pygmy mothers producing offspring who are raised among the Pygmies[12-15]. The direction of our observations is concordant with these previous observations. Male-biased non-Pygmy admixture into the Pygmies would introduce greater genetic diversity in the parts of the genome that harbor equal amounts of ancestry from males and females (the autosomes) than into the parts of the genome that harbor more ancestry from females (chromosome X). The direction of this effect would be expected to drive down the X-to-autosome heterozygosity ratio in such populations, compared to populations of similar heterozygosity, as we observe in Pygmies.

In two previous simulation studies reported by Keinan and colleagues, it was shown that a scenario of repeated waves of male mixture into an already mixed population could be responsible for the extreme reduction in the X-to-autosome ratio in all non-African compared to most African populations [7,16]. Such a scenario is actually much more speculative for the history of all non-African populations than it is for Pygmies. In the deep shared history of non-Africans, there is no anthropological evidence for such a sex-based gene flow history. The model was only proposed to explain a genetic observation. In the case of the Pygmies, there is strong anthropological support for sex-biased gene flow[15].

Fewer accumulated divergent sites in Africans than in non-Africans

To understand whether some human populations have accumulated mutations at a higher rate than others, we computed $D(Pop_A,Pop_B,Chimp)$, which compares the accumulated number of mutations in two lineages Pop_A and Pop_B, restricting to the same set of 235 samples from 108 populations analyzed in the previous section.

To increase the power of this analysis, we pooled samples into 8 worldwide groupings. We divide Africans into three categories: “Pygmy” (Mbuti and Biaka, n=5), “Khoesan” (Khomani San and Ju_hoan_North, n=5), and “Africa” (n=22), which includes all other African samples except populations north of the Sahara because of their West Eurasian related ancestry (we exclude Saharawi and Mozabite). We group the non-African populations into the categories “America” (n=23), “CentralAsiaSiberia” (n=21), “EastAsia” (n=42), “WestEurasia” (n=61), and “Oceania” (n=19). We performed analyses separately on the autosomes and chromosome X (excluding pseudoautosomal regions). We used a Weighted Block Jackknife [17] to compute a standard error and a Z-score for being different from zero.

This analysis reveals significant evidence of differences in the rate of accumulation of mutation across populations (Extended Data Table 1), and specifically fewer accumulated mutations in sub-Saharan Africans that in non-Africans. Depending on which population pool comparison we analyze, the significance ranges from $3.3<|Z|<9.4$, with $0.0013<D(Africa,Non\text{-}Africa,Chimp)<0.0037$. The average value of $D(Africa,Non\text{-}Africa,Chimp)$ is 0.0025, which corresponds to a $0.5\%=2\times0.0025$ higher rate of accumulation of mutations in non-Africans than in Africans since divergence. This might seem small, but considering that all the differences in accumulation of mutations in non-Africans and in Africans must have occurred since population divergence, and that population divergence is less than a tenth of average genetic divergence (Fig. 2), these results reflect a quite substantial difference in the accumulation of mutations since population divergence of $>5\%=10\times0.005$.

Two lines of evidence suggest that these patterns are due to a real difference in the rate of accumulation between Africans and non-Africans since they separated.

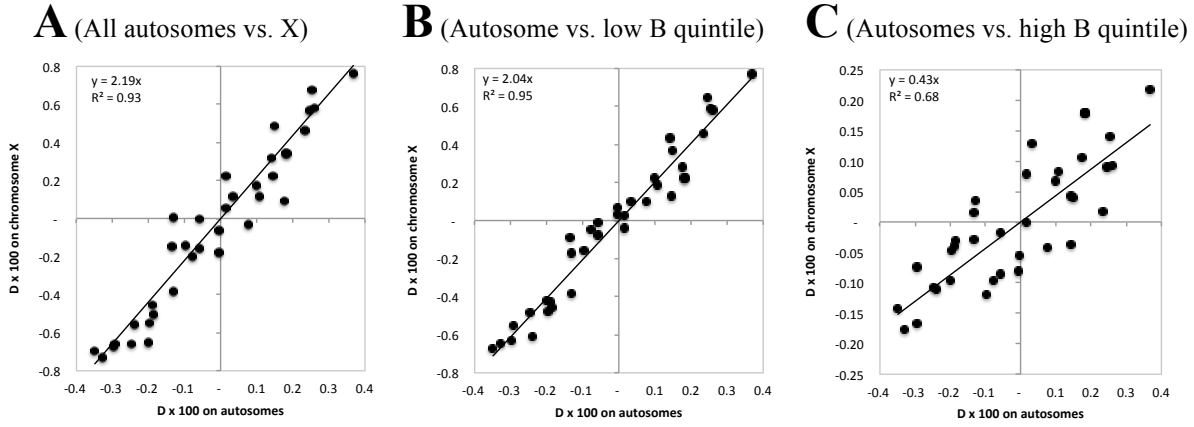
First, we tested a population genetic prediction. If the difference in the rate of accumulation of mutations is driven by events since Africans and non-African populations separated, we would expect the observed signal to be greater in subsets of the genome where the time since populations split (and thus the time since mutation accumulation rates have been different) is a larger proportion of the total history. We validated this in three subsets of the genome.

(A) X-chromosome: 2.19× enhancement. The average time since the common ancestor is predicted to be larger on chromosome X than on the autosomes due to a lower effective population size. As expected, chromosome X skews in $D(Pop_A,Pop_B,Chimp)$ are substantially larger than autosomal ones (Fig. S7.4A) (Extended Data Table 1).

(B) Lowest B quintile: 2.04× enhancement. The average time since the common ancestors is known to be reduced closest to functional elements[10]. As expected, skews in $D(Pop_A,Pop_B,Chimp)$ in the fifth of the genome closest to functional elements are substantially larger than autosomal ones (Fig. S7.4B) (Extended Data Table 1).

(C) Highest B quintile: 0.43× shrinkage. The average time since the common ancestors is known to be increased furthest from functional elements[10]. As expected, skews in $D(Pop_A,Pop_B,Chimp)$ in the fifth of the genome furthest from functional elements are substantially smaller than autosomal ones (Fig. S7.4C) (Extended Data Table 1).

Figure S7.4. $D(Pop_A,Pop_B,Chimp)$ is most extreme in subsets of the genome where the population split time comprises a larger fraction of the total time since the most recent common ancestor. *As a result, any differences in mutation rate accumulation since the population split have a larger proportional effect.*



Second, we tested whether there is evidence that our findings could be arising as artifacts of the bioinformatics analysis. One area of concern is that the human reference genome is primarily of West Eurasian ancestry[3]. If short reads map more easily to a reference sequence to which they are more closely related, a subtle difference in the rate of detection of sites that differ from the reference sequence could arise. A second area of concern is that when heterozygous positions are misread (the most common mode of genotyping error), they tend to be miscalled as the allele matching the reference genome because of reference mapping bias. This error mode is expected to occur more often in Africans than in non-Africans because of the higher rates of heterozygous positions in Africans (African are the most diverse present-day humans). We were concerned that this could be an artifactual explanation for the evidence of reduced accumulation of mutations in Africans.

To eliminate the concern about the predominantly West Eurasian ancestry of the reference sequence being more closely related to some human populations than to others, we remapped sequencing reads to the chimpanzee genome *PanTro2*.

To eliminate the concern that differences in heterozygosity across populations could be causing bias, we performed this remapping analysis on the X chromosome in males. Males have only one X chromosome copy, so there are no heterozygous positions (outside of the pseudo-autosomal regions). Thus differences in heterozygosity across individuals from different populations are not expected to bias these analyses.

To implement this approach, for a large number of pairs of human male samples (S_1, S_2) we performed the following analysis. For each 51-mer on the chimpanzee X chromosome that is unique within the entire chimpanzee genome, we retrieved all human reads containing the 51-mer and counted each type of read base next to the 51-mer. We ignored the 51-mer if:

(1) The most common allele in either sample is supported by ≤4 reads
(2) The second most common allele count of either sample is ≥3
(3) The most common allele is the same for two samples.

Requirements #1 and #2 make sure that we are mostly looking at well-supported haploid regions of human chromosome X.

We determined an allele to represent each individual at each nucleotide by majority rule. This allowed us to compute $D(S_1, S_2, Chimp)$.

We randomly selected 78 African-non-African male sample pairs and computed *D(African,non-African,Chimp)* stratified by transition/transversion. Extended Data Fig. 5 shows that at transversion sites, *D(African,non-African,Chimp)*is around zero, the expected value. However, at transition sites, *D(African,non-African,Chimp)*is usually positive. Thus, the observation of non-African samples having more accumulated mutations is not an artifact due to human reference bias or different heterozygosities across populations.

Assuming that the effects we are observing are real, is there evidence that they are due to an acceleration of the rate of mutation accumulation in non-Africans, or a deceleration within Africa? An acceleration in non-Africans is most parsimonious, as it could explain the observations by a single historical/biological process. In contrast, a deceleration in Africans is not parsimonious. Given the phylogeny of Africans in which the KhoeSan and Pygmy branch most deeply, there is no single population in which a deceleration could explain the fact that the strongest signals always involve non-Africans (Extended Data Table 1).

What type of process could cause an acceleration of mutation accumulation in non-Africans? We discuss four possibilities.

(1) After the dispersal of modern humans out of Africa, there could have been changes in life history traits such as the generation interval, which affected mutation rates[18].

(2) Living at higher latitudes or in colder climates could have resulted in accelerated mutation rates in non-Africans.

(3) GC-biased gene conversion against newly arising A or T alleles could have worked more effectively in Africans than in non-Africans after population separation. An enhanced impact of gene conversion could arise because Africans are more genetically diverse that non-

Africans, which leads to more heterozygous positions per genome that can be acted upon by gene conversion. In addition, the larger average effective population size in Africans than in non-Africans since population separation would have made the trajectories of alleles under the pressure of gene conversion more deterministic, and thus could have made gene conversion more effective.

(4) We considered—but ruled out as unlikely—the possibility that mutation rates in Neanderthals have been higher than those in modern humans since separation. In this case, Neanderthal admixture into the ancestors of non-Africans but not into the ancestors of sub-Saharan Africans ~50,000 years ago could contribute to the observed greater accumulation of mutations in non-Africans. To explain our observations, the rate of accumulation of mutations in Neanderthals must on average have been ~20% higher than in modern humans since separations since only a few percent of mutations accumulated on non-African lineages compared to African lineages owe their origin to Neanderthal admixture, and thus a greatly increased rate of accumulation of mutations on these segments is needed to explain the genome-wide excess of ~0.5% of mutations in non-Africans[19]. There is no evidence for a higher mutation rate in Neanderthals, however, as Prüfer et al. showed that the number of mutations accumulated in a deeply sequenced Neanderthal genome from the Altai mountains compared to present-day human genomes is not more than one would expect from the hypothesis of the mutation rate having been constant in both taxa since separation; indeed, if anything it is less than what one would expect given the date of the sample[19].

References

1 Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature genetics* **47**, 126-131, doi:10.1038/ng.3186 (2015).

2 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

3 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).

4 Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat Genet* **41**, 66-70, doi:10.1038/ng.303 (2009).

5 Nam, K. *et al.* Extreme selective sweeps independently targeted the X chromosomes of the great apes. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 6413-6418, doi:10.1073/pnas.1419306112 (2015).

6 Julien Y Dutheil, K. M., Kiwoong Nam, Thomas Mailund, Mikkel Schierup. Strong selective sweeps on the X chromosome in the human-chimpanzee ancestor explain its low divergence. *bioRxiv* http://dx.doi.org/10.1101/011601, http://dx.doi.org/10.1101/011601 (2015).

7 Keinan, A. & Reich, D. Can a sex-biased human demography account for the reduced effective population size of chromosome X in non-Africans? *Molecular biology and evolution* **27**, 2312-2321, doi:10.1093/molbev/msq117 (2010).

8 Gottipati, S., Arbiza, L., Siepel, A., Clark, A. G. & Keinan, A. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nature genetics* **43**, 741-743, doi:10.1038/ng.877 (2011).

9 Arbiza, L., Gottipati, S., Siepel, A. & Keinan, A. Contrasting X-linked and autosomal diversity across 14 human populations. *American journal of human genetics* **94**, 827-844, doi:10.1016/j.ajhg.2014.04.011 (2014).

10 McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics* **5**, e1000471, doi:10.1371/journal.pgen.1000471 (2009).

11 Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035-1044, doi:10.1126/science.1172257 (2009).

12 Quintana-Murci, L. *et al.* Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 1596-1601, doi:10.1073/pnas.0711467105 (2008).

13 Verdu, P. *et al.* Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Molecular biology and evolution* **30**, 918-937, doi:10.1093/molbev/mss328 (2013).

14 Verdu, P. *et al.* Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current biology : CB* **19**, 312-318, doi:10.1016/j.cub.2008.12.049 (2009).

15 Joiris, D. V. THE FRAMEWORK OF CENTRAL AFRICAN HUNTER-GATHERERS AND NEIGHBOURING SOCIETIES. *African Study Monographs,* **Suppl.28: x**, 57-79 (2003).

16 Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nature genetics* **41**, 66-70, doi:10.1038/ng.303 (2009).

17 Patterson, N. J. *et al.* Ancient Admixture in Human History. *Genetics*, doi:10.1534/genetics.112.145037 (2012).

18 Segurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15**, 47-70, doi:10.1146/annurev-genom-031714-125740 (2014).

19 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49, doi:10.1038/nature12886 (2014).

Supplementary Information section 8

The worldwide landscape of Neanderthal and Denisova introgression

Swapan Mallick*, Nick Patterson and David Reich

*To whom correspondence should be addressed: (shop@genetics.med.harvard.edu)

Approach

Previous work has established that present-day humans outside of Africa harbor appreciable amounts of ancestry from archaic humans, both Neanderthals[1-4] and Denisovans[2,4,5]. However, these inferences have been based on genome-sequences from only a handful of populations. To date, there has been no fine-grained assessment of variation in the proportion of Neanderthal and Denisova ancestry.

To carry out a survey of how Neanderthal and Denisovan ancestry varies across populations—and to distinguish these two sources of ancestry—we took the approach of studying genome-wide rates of sites that are diagnostic of deriving one of these sources of archaic ancestry or the other. We caution that these are not intended as absolute estimates of Neanderthal and Denisovan ancestry such as have been reported previously, and instead as relative estimates.

To identify sites that are diagnostic of Neanderthal or Denisovan ancestry, we use the strategy first described in Supplementary Information section 13 of Prüfer et al. 2014[3]. At each of Z_T positions on the autosomes (chr. 1-22) for which we have coverage from chimpanzee allowing us to infer the ancestral allele, for which there is a valid call passing the Map35_50% filter of ref. [3] in both Neanderthal and Denisova, and for which there is a *Test* population sample passing filter level 1, we compute:

p_i^N = frequency of the derived allele in the Altai Neanderthal (0, 0.5 or 1) at position i

p_i^D = frequency of the derived allele in Denisova (0, 0.5 or 1) at position i

p_i^T = frequency of the derived allele in the *Test* population at position i

p_i^Y = frequency of the derived allele in a panel of 107 Yoruba individuals that have been sequenced to medium coverage (average of 7.1×) by the 1000 Genomes Project[6]. We restrict to sites where each Yoruba called have coverage from at least three reads with a map quality of MAPQ≥37 and base quality of ≥30, and then call a single allele to represent each of these individuals based on majority rule.

δ_i^Y = indicator variable: δ_i^Y=1 if p_i^Y=0, and δ_i^Y =0 otherwise.

We are interested in the rate of observing in the *Test* population x two classes of sites:

nd_{10} sites diagnostic of Neanderthal ancestry

These are sites where a randomly drawn chromosome from Neanderthal is derived, a randomly drawn chromosome from Denisova is ancestral, and all sub-Saharan African chromosomes are ancestral (this filters out alleles that were

polymorphic in modern humans before archaic introgression). We have previously shown that such sites are likely to derive from Neanderthals and not Denisovans.

nd_{01} sites diagnostic of Denisova ancestry

These are sites where a randomly drawn chromosome from Denisova is derived, a randomly drawn chromosome from Neanderthals is ancestral, and all sub-Saharan African chromosomes are ancestral. We have previously shown that these are highly likely to derive from Denisovans and not Neanderthals.

In the entire dataset, the expected number of nd_{10} and nd_{01} sites across the genome is quite large:

$$Expected_Count_{nd10} = \sum_{i=1}^{n} p_i^N (1 - p_i^D)\, \delta_i^Y = 373{,}637$$

$$Expected_Count_{nd01} = \sum_{i=1}^{n} (1 - p_i^N) p_i^D\, \delta_i^Y = 445{,}320$$

However, in any one *Test* modern human population only a small fraction of these sites are expected to be derived, reflecting the fact that most of the alleles that any modern human individual carries are not of archaic origin. For example, in Papuans, a population with a substantial proportion of both Neanderthal and Denisovan ancestry, the mean and standard deviation of alleles that are expected to be derived, measured on a per-haploid genome basis, are 9232±270 for nd_{10} and 7273±330 for nd_{01}.

In practice, we study rates per base pair of derived nd_{10} and nd_{01} sites in any *Test* population, as we need to normalize by the number of nucleotides that pass filter level 1, which varies from population to population.

$$Rate_{nd10}^T = \frac{1}{Z_T} \sum_{i=1}^{n} p_i^T\, p_i^N (1 - p_i^D)\, \delta_i^Y$$
$$Rate_{nd01}^T = \frac{1}{Z_T} \sum_{i=1}^{n} p_i^T\, (1 - p_i^N) p_i^D\, \delta_i^Y$$

As described previously[3,7], this approach provides a high resolution estimate of a quantity that is proportion to the fraction of a population's ancestry they inherit from Neanderthals (proportional to $Rate_{nd10}^T$) or Denisovans (proportional to $Rate_{nd01}^T$). To convert to an absolute estimate, we need to use information on the proportion of ancestry in a specified reference population, for example taking it as a given that the proportion of Neanderthal ancestry in French is N_{French}=2% (following the example of ref. [7]), and that the proportion of Denisovan ancestry in Papuan is D_{Papuan}=5%. To convert *Rate* to an absolute estimate, we can then subtract out the background rate of false-positives, which we infer in practice by analyzing a sub-Saharan African populations (we use Dinka following ref. [7]):

$$N_T = N_{French} \left(\frac{Rate_{nd10}^T - Rate_{nd10}^{Dinka}}{Rate_{nd10}^{French} - Rate_{nd10}^{Dinka}} \right)$$

$$D_T = D_{French} \left(\frac{Rate_{nd01}^T - Rate_{nd01}^{Dinka}}{Rate_{nd01}^{Papuan} - Rate_{nd01}^{Dinka}} \right)$$

If the Yoruba population used in the ascertainment of the sites was symmetrically related to all modern human populations, this strategy would provide an unbiased estimate of a quantity proportional to Neanderthal and Denisova ancestry. However,

this assumption is not fully accurate. One caveat is that Yoruba are known to have on the order of a couple percent West Eurasian ancestry[3], and thus requiring that Yoruba always have the ancestral allele will tend to filter out SNPs that are indicative of Neanderthal ancestry in West Eurasians and result in underestimates of Neanderthal ancestry in West Eurasians relative to populations that have not contributed ancestry to Yoruba[8]. A second issue is that some human populations have received gene flow from Yoruba related populations; for example, populations in the Near East and North Africa that have received historical period sub-Saharan African gene flow[9]. A third issue is that for sub-Saharan Africans that are most distant from Yoruba, such as Khoesan and Pygmy, the filtering to remove sites that were present in the modern human ancestral population since separation from Neanderthals is expected to be least effective, and we thus expect a higher background rate of false-positive nd_{10} and nd_{01} sites relative to other sub-Saharan Africans (and non-Africans). We believe that the apparent excesses of nd_{10} and nd_{01} in these hunter-gatherer African populations relative to other sub-Saharan Africans, documented in Table S8.1, is due to this artifact. Despite these caveats, these statistics give relatively accurate estimates of quantities informative about archaic ancestry, and we show results from them here.

Results

Rates of nd_{10} and nd_{01} sites per world region are summarized in Table S8.1, showing the mean rate and standard deviation across individuals within each group.

Table S8.1: Rates of nd_{10} and nd_{01} sites in present-day humans (by world region)

	$Rate^{T}_{nd10}$ ($\times 10^{-6}$)		$Rate^{T}_{nd01}$ ($\times 10^{-6}$)	
	Mean	Std. Dev.	Mean	Std. Dev.
Africa	0.659	0.681	0.37	0.333
Africa excluding HG*	0.573	0.745	0.203	0.075
Africa HG*	0.927	0.291	0.886	0.288
America	3.929	0.222	0.456	0.04
Central Asia & Siberia	4.216	0.261	0.474	0.053
East Asia	4.291	0.231	0.52	0.052
Oceania	3.937	1.379	2.47	1.437
South Asia	3.834	0.264	0.501	0.089
West Eurasia	3.507	0.307	0.307	0.029

* HG = African Hunter Gatherers

We converted the rate of derived nd_{10} (likely Neanderthal-derived) sites to a by-sample estimate in Supplementary Data Table 1 (Figure S8.1) by assuming that the French value is N_{French}=2%. The qualitative patterns are consistent with previously documented evidence of far more Neanderthal ancestry in non-Africans than in Africans, and more Neanderthal ancestry in eastern non-Africans than in West Eurasians. Fig. 2c shows the by-sample plots. There is clear evidence of Neanderthal ancestry in northeastern Africa, which can be ascribed to the documented evidence of West Eurasian ancestry in these populations[10]. We do not plot African hunter gatherers (Khoesan and Pygmy), since as discussed above and shown in Table S8.1, the elevations in these populations (and similar ones for Denisovan ancestry) are likely due to the fact that we use Yoruba to screen out sites that are likely derived in modern humans, and this is less effective in the sub-Saharan African populations that are most distantly related to Yoruba.

Figure S8.1: Proportion of Neanderthal ancestry estimated per sample

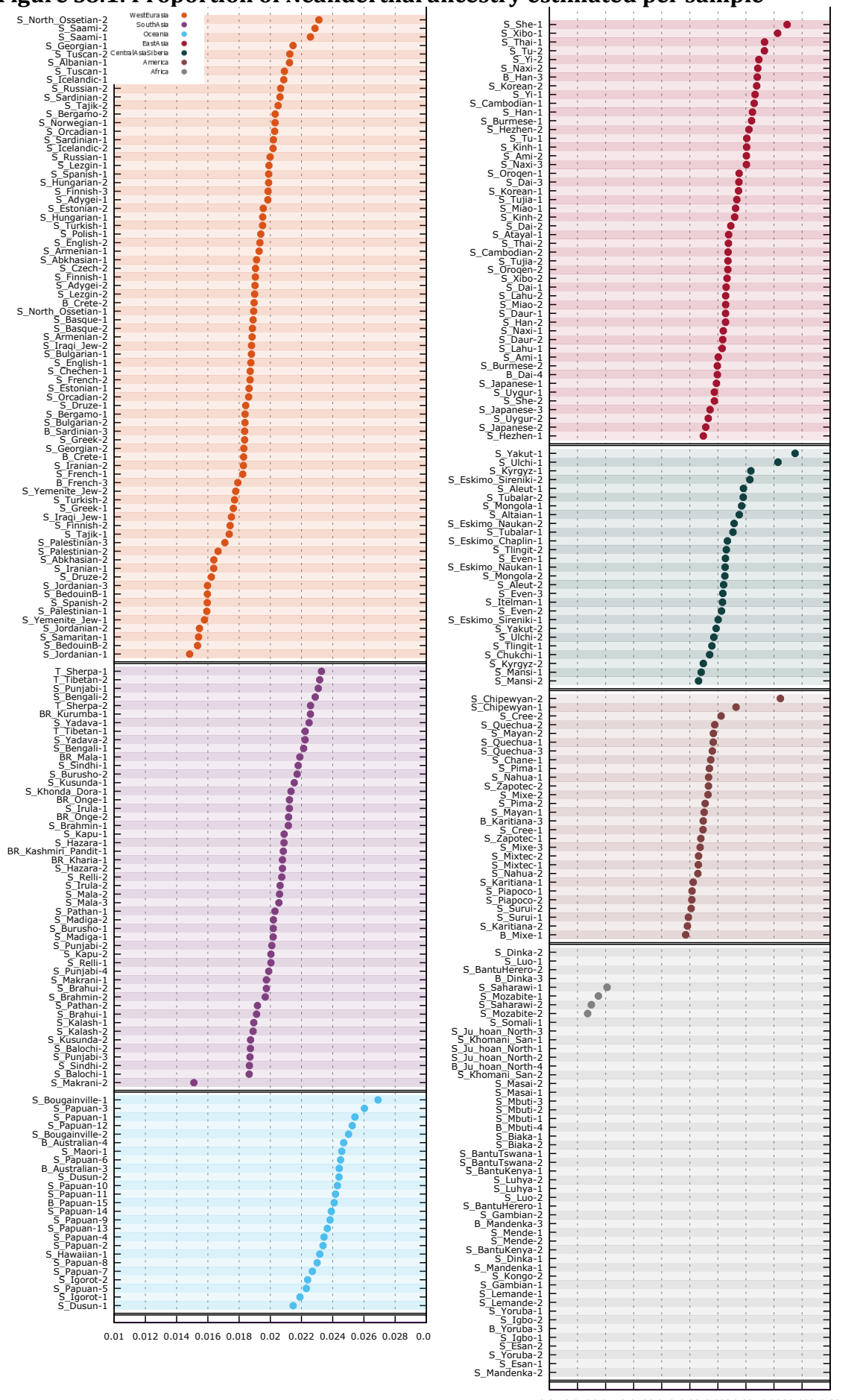
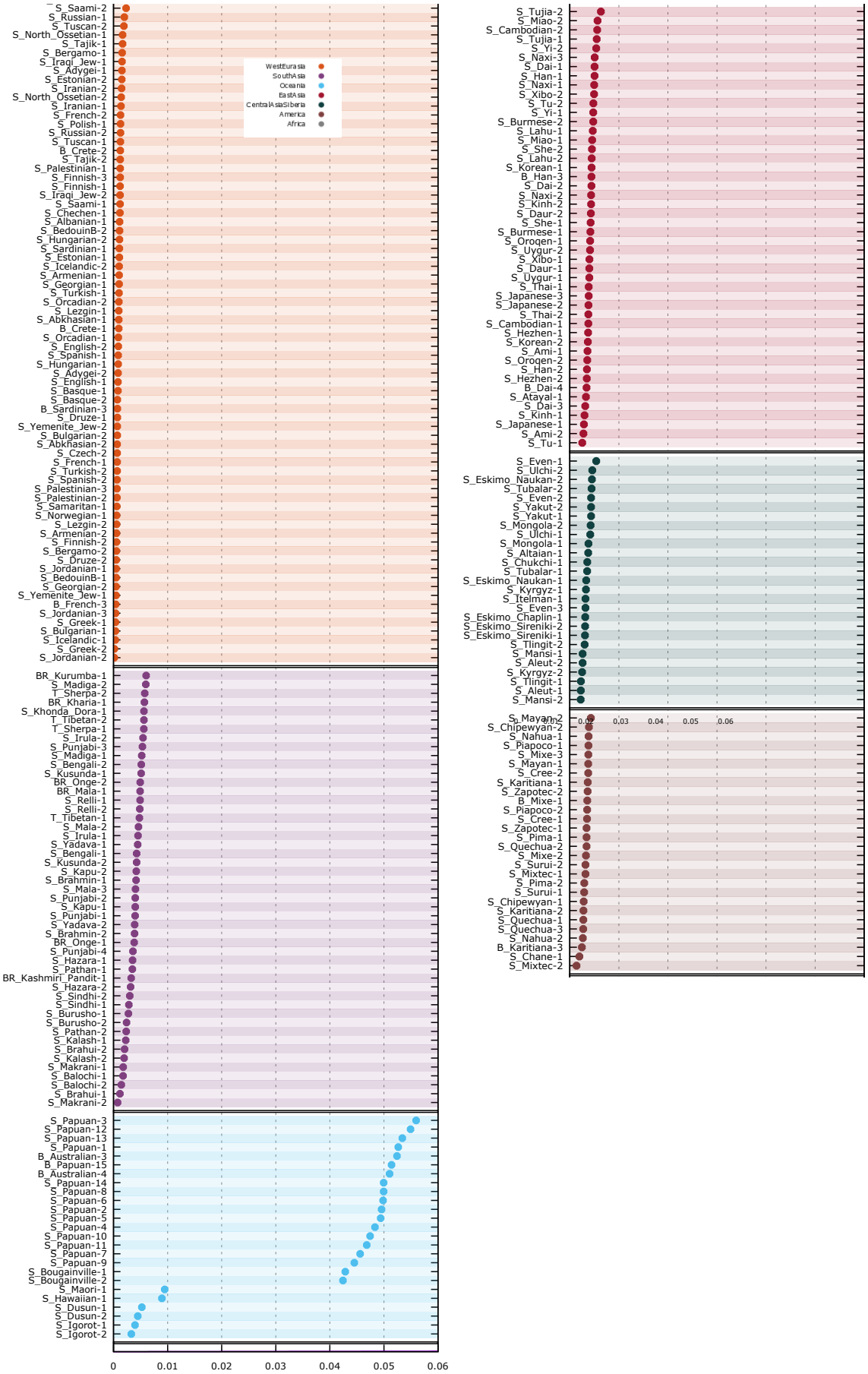


Figure S8.2: Proportion of Denisovan ancestry estimated per sample.



A heat map of the rate of derived nd_{01} (likely Denisova-derived) sites by population is shown in Fig. 2d. We recapitulate previous evidence of more Denisova ancestry in Australia and New Guinea than in mainland Eurasians[4,5], as well as in Oceanian populations like Maori and Hawaiians known to have New Guinean admixture. We also recapitulate the finding of more Denisova ancestry in eastern than in western Eurasians[3,11]. A novel finding is the observation of a peak of Denisovan ancestry in a subset of South Asian populations, which is most evident in a heatmap (Fig. 1d). The per-sample estimate of Denisova ancestry is shown in Figure S8.2.

References

1 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).

2 Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, doi:10.1126/science.1224344 (2012).

3 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49, doi:10.1038/nature12886 (2014).

4 Reich, D. *et al.* Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American journal of human genetics* **89**, 516-528, doi:10.1016/j.ajhg.2011.09.005 (2011).

5 Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060, doi:10.1038/nature09710 (2010).

6 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

7 Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, doi:10.1038/nature14558 (2015).

8 Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354-357, doi:10.1038/nature12961 (2014).

9 Moorjani, P. *et al.* The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS genetics* **7**, e1001373, doi:10.1371/journal.pgen.1001373 (2011).

10 Wang, S., Lachance, J., Tishkoff, S. A., Hey, J. & Xing, J. Apparent variation in Neanderthal admixture among African populations is consistent with gene flow from Non-African populations. *Genome Biol Evol* **5**, 2075-2081, doi:10.1093/gbe/evt160 (2013).

11 Skoglund, P. & Jakobsson, M. Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 18301-18306, doi:10.1073/pnas.1108181108 (2011).

Supplementary Information section 9

Demographic inference

Iain Mathieson*, Jeffrey P. Spence, Yun S. Song

*To whom correspondence should be addressed (iain_mathieson@hms.harvard.edu)

9.1 Overview

We used PSMC[1] and MSMC[2] to infer population sizes and split times for selected SGDP populations. Split time estimation requires phased haplotypes, so we phased the SGDP samples using both SHAPEIT (v2.r837)[3] and IMPUTE2 (version 2.3.2)[4]. Though the relative ordering of more ancient population splits is fairly robust to the phasing and inference method used, we found that there was considerable uncertainty in the absolute split times inferred, limiting our ability to make strong statements about history.

9.2 Phasing

We first made a list of all SNPs that passed filter level 1 in any sample, and then genotyped each of these SNPs in all SGDP samples, removing any sites with more than two alleles in the data. We phased the samples at all of these sites, and later restricted to sites that passed sample-specific filters for downstream analysis. We used three different phasing strategies, which we refer to as PS1-3:

(PS1) We used SHAPEIT together with the 1000 Genomes phase 3 haplotypes[a] as a reference panel, phasing each sample separately using SHAPEIT (using the --input-ref and --no-mcmc options). We left heterozygous sites not in the 1000 Genomes data as unphased.

(PS2) We used SHAPEIT without a reference panel, phasing all samples together. This phases all sites in the sample.

(PS3) We used IMPUTE2 to phase the 1000 Genomes reference panel, but also phasing all sites in the sample (using the -no_remove and -fill_holes options). We split up chromosomes into 5 Mb chunks and then joined them randomly, introducing switch errors at a rate of 1 per 10 Mb (2-3 orders of magnitude less than the rate of switch errors from statistical errors).

Ideally we would use SHAPEIT (which is faster and more accurate than IMPUTE2) to phase using a reference panel and within-sample, as in PS3, but this option is not available. In principle this is possible by using the output of PS1 as a scaffold and then running SHAPEIT with the -call and --input-scaffold options, but this failed for us with numerical underflow errors, even for small regions. Therefore we are left with a tradeoff between accuracy, completeness of phasing, and potential systematic differences between populations (for example, when we use a reference panel,

[a] Downloaded from "https://mathgen.stats.ox.ac.uk/impute/1000GP Phase haplotypes 6 October 3 2014.html" on 12th October 2014

samples from SGDP populations that are closely related to populations in the 1000 Genomes Project are likely to be better phased than populations that are not).
To compare rates of phasing error, we used previously generated experimentally phased data for three samples which are included in the present study[5] (Table S9.1).

	PS1	PS2	PS2*	PS3
B_Australian-4	0.012	0.014	0.009	0.019
B_Australian-3	0.029	0.046	0.019	0.056
B_Mixe-1	0.026	0.031	0.025	0.032

Table S9.1. Switch error rates per kb estimated using experimentally phased samples. PS1-3 are as described above. PS2* is the same as PS2, but with sites not found in 1000 Genomes excluded.

We find that PS3 (IMPUTE2 phasing) is less accurate than SHAPEIT phasing, even when using SHAPEIT without a reference panel. We also find that switch errors are concentrated at sites that are not included in 1000 Genomes, even without a reference panel (compare PS2 and PS2*). These results should be interpreted with caution, since these samples, for which experimentally phased data is available, are not representative of the SGDP. In particular, the Australian samples are not closely related to any 1000 Genomes populations. However, this analysis is consistent with our expectations that SHAPEIT performs better than IMPUTE2, and that using a reference panel improves phasing for samples that have a closely related population in the reference panel.

In summary, we could not find an optimal phasing strategy. PS1 likely produces the best phasing, but leaves many unphased sites for populations that are not in 1000 Genomes (for example Australians and San). PS2 and PS3 both perform worse than PS1, but this is largely because population-private sites are poorly phased. PS2 performs better than PS3 for the three tested samples, but we cannot rule out the possibility that PS3 would perform better for populations that are close to populations in the 1000 Genomes Project study. Finally, it is possible that the different strategies produce qualitatively different patterns of phase errors, with differential effects on demographic inference. For this reason, we compared results from all three strategies in downstream analysis.

9.3 Uncertainty in mutation rate and generation interval

There is substantial uncertainty about the human mutation rate, which we parameterize by the mean autosomal per-base per-generation mutation rate μ, the mean generation interval g or, equivalently, the mean per-base per-year mutation rate v, where $\mu=vg$. We do not model their dependence on time, sex and genomic context[6]. Here we briefly describe this uncertainty and its consequences for our analysis.

Generation interval g: The average human generation interval g is typically assumed to be in the range of 27-31 years, based on data from contemporary hunter-gatherer societies, although there is considerable variation across populations[7,8].

Mutation rate μ, v: Human mutation rates have been estimated in a number of ways. Relatively high mutation rates based on human-chimpanzee divergence and split times based on calibration to the fossil record are now generally considered to be

unreliable, likely due to shifts in mutation rate hundreds of thousands or millions of years ago[9]. The per-generation mutation rate μ can be estimated directly by counting mutations among parent-child trios[10-12], which gives estimates ranging from 0.96-1.20×10^{-8}, corresponding to ν=3.1-4.4×10^{-10} if we assume generation times between 27 and 31 years. An alternative approach is to use the recombination clock to calibrate the mutation rate[13,14] which gives higher estimates of μ=1.61×10^{-8}, corresponding to ν=5.2-6.0×10^{-10}. Finally, ν has been estimated to be 3.9-4.7×10^{-10} by comparing a 45,000 year-old modern human genome from Siberia to present-day genomes[15].

Genomic context and filtering: Per-base mutation rates vary substantially along the genome. *De novo* rates vary because of factors like sequence context[12] and replication timing[16], while long-term rates also vary due to differential selection and recombination rates across the genome. This means that studies that apply different filters, and access different parts of the genome, are likely to report different mutation rates. One way to make studies that look at different regions comparable is to scale by the reported heterozygosity in some reference population, for example northern Europeans.

At filter level 1, we estimate the effective heterozygosity in our filtered regions for the French population to be 0.68×10^{-3} by computing the observed heterozygosity in our MSMC input files. Two studies that reported French heterozygosity in their called genomes both reported higher values of 0.77×10^{-3} and 0.75×10^{-3}, suggesting that, to apply these mutation rates to our data, we might wish to use a mutation rate approximately 11% lower than they report. We summarize these results in Table S9.2.

Method	Ref	μ	ν	π(French)	Rescaled ν	Notes
Direct	Kong *et al.*[11]	1.20e-08	3.9-4.4e-10		3.5-4.0e-10	Called 2.5Gb. Assume π=7.5e-4
Ancient	Fu *et al.*[15]		3.9-4.7e-10	7.7e-4	3.4-4.1e-10	Table S14.2
LD	Lipson *et al.*[13]	1.61e-08	5.2-6.0e-10	7.5e-4	4.7-5.4e-10	

Table S9.2. Mutation rate estimates rescaled using heterozygosity to be comparable to the filtering using in this study. Three studies that estimate μ or ν, rescaled to give a value of ν appropriate to our dataset where the heterozygosity of French is 0.68×10^{-3}. For these computations we assume the generation interval averaged 27 to 31 years.

However, rescaling by realized heterozygosity may not always be correct. Regions of high heterozygosity that are filtered out by strict filters may be regions with high error rates, rather than regions with high mutation rates. More subtly, removing regions of high heterozygosity may remove regions with ancient but not recent coalescence that, for MSMC analysis, would predict a large effect on estimates for ancient but not recent times. Thus the appropriate scaling factor would be a function of time, rather than a constant. For this reason, we do not rescale our results in the main text.

Parental age and sex-specific effects: Most *de novo* mutations arise paternally, due to the larger number of cell divisions to produce male gametes[11]. Since cells in the male germline, unlike the female germline, continue to divide over a man's lifetime, the average number of mutations increases with paternal age. Thus, the sex-averaged mutation rate depends in a complicated way on the exact values of the male and female mutation rates and the average male and female generation interval. There is still considerable uncertainty about the values of these parameters and about the

appropriate mutational model[6], so we do not attempt to take these details into account here. However, we caution that it is possible that these vary across populations, which might complicate interpretation of MSMC analyses.

Summary: Based on these estimates, we report times scaled by $\nu=4.3\times10^{-10}$. This is the midpoint of the PSMC-based branch-shortening estimate calibrated to a directly dated ancient genome[15], which we consider to be the most relevant for this analysis (and also is very similar to the estimates from *de novo* mutation rate studies). When we plot MSMC results, we show the times scaled to this range on the x-axis. Times can easily be converted to alternative mutation rates by scaling by a constant (for example, it would be reasonable to rescale to $\nu=3.4\text{-}6.0\times10^{-10}$, the most extreme values of ν from Table S9.1).

9.4 MSMC analysis

We processed our phased genomes into MSMC format[b]. For each analysis here, we used two phased genomes in four-haplotype MSMC format. We included only sites that passed filter level 1 or higher. Following the MSMC paper, we report the cross-coalescence rate: the ratio of the between- and mean within-population inferred coalescence rates. We ignored unphased sites using the "--skipAmbiguous" option. If the n^{th} variant site in one sample is unphased, this excludes all sequence between the $n\text{-}1^{th}$ and n^{th} site.

MSMC produces population size estimates that are size-scaled by twice the mean autosomal per-base per-generation mutation rate μ, and time-scaled by the mean per-base per-year mutation rate v, where $\mu=vg$, and g is the mean generation interval. We assume all of these parameters are constant over time. For cross-coalescence rate estimation, only v is required.

Reporting MSMC results: MSMC makes no assumptions about the structure of population splits. However to summarize results in a way that can be compared with other estimates that assume instantaneous splits, we convert the MSMC results into a range of split times by taking the most recent time at which the cross coalescence is above 0.25 and 0.75. If we give a point estimate without a range, or a midpoint, we quote the time at which the cross-coalescence rate rises above 0.5. Sometimes we quote the time of "initial split", by which we mean the oldest time for which the cross-coalescence rate is <1. When we report MSMC estimates of effective population size, we scale by $2\mu=2.5\times10^{-8}$, assuming $\nu=4.3\times10^{-10}$ and g=29 years.

Effect of phasing errors on simulations: to understand how phasing would affect inferences from MSMC, we simulated a simple two-population split model using scrm[17], and added phasing errors to the simulated haplotypes. We then ran MSMC to infer cross coalescence rates. We found (Figure S9.2) that phasing errors tend to make splits look older by breaking up long recent haplotypes. The effect is more severe for more recent splits, suggesting that inference of the time of more ancient events will be more accurate than for recent events, which are more prone to being overestimated.

[b] We modified multihetsep.py from https://github.com/stschiff/msmc-tools to take into account our site- and sample-specific filters.

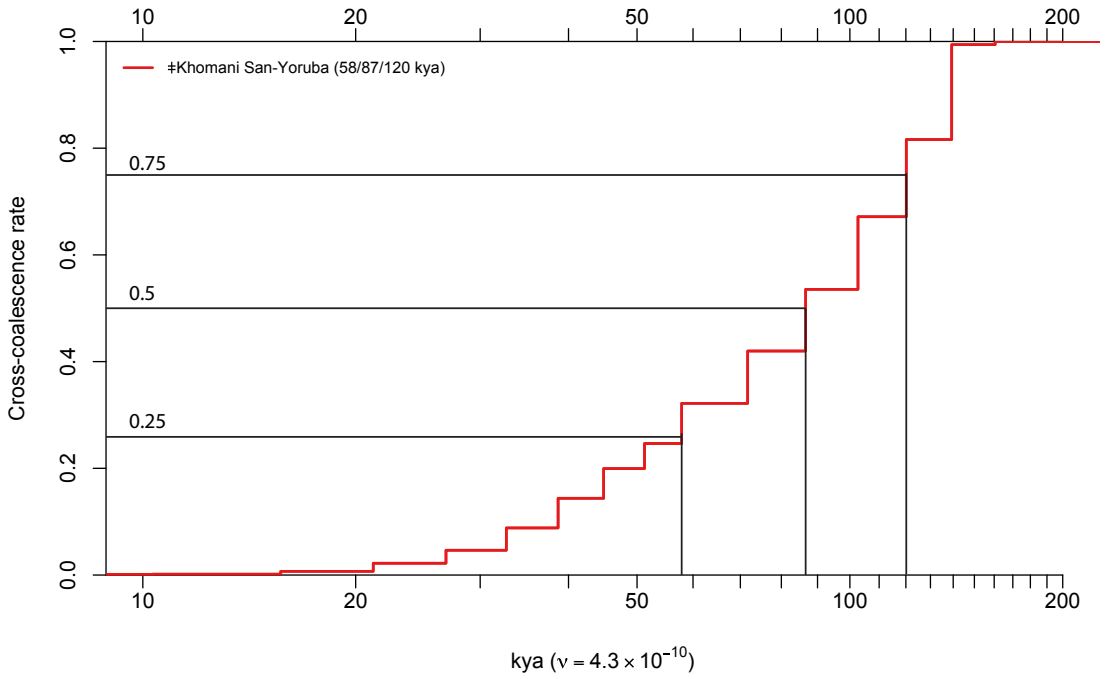


Figure S9.1. Reporting MSMC results. This figure shows an example of how we report MSMC results, in this case, the cross-coalescence rate for Yoruba and ǂKhomani San. In the figure legend we report the times at which the cross-coalescence rate crosses 0.25, 0.5, and 0.75, in this case as (57/87/120 kya).

Effect of phasing strategy in practice: We compared the three phasing strategies on different pairs of populations. We found that in cases where the split is old, and the populations are closely related to 1000 Genomes populations (for example, Figure S9.3; French-Yoruba split) the MSMC cross-coalescence rates are insensitive to phasing. This is true even for more recent splits (Figure S9.4; French-Han). Even when one population is not closely related to a 1000 Genomes population, the MSMC results are still fairly robust to phasing for old splits (Figure S9.5 French-Mbuti) but can be dramatically different for recent splits (Figures S9.6 and S9.7; French-New Guinean and Australian-New Guinean). In all cases, recent effective population size estimates are lower for PS1, suggesting that removing sites not in 1000 Genomes leads to an underestimate of the population size in recent times. However, it is also possible that these sites are enriched with genotype errors.

In summary, MSMC results are robust to phasing for old splits, or more recent splits where the populations are represented in the 1000 Genomes Project and therefore well-phased, but not for recent splits, particularly when the populations are not represented in 1000 Genomes. For the plots in the main text (Fig. 2), we report cross-coalescence rates inferred using PS1 and effective population sizes inferred using PSMC (PSMC does not require phasing, and provides qualitatively similar population size reconstructions as MSMC).

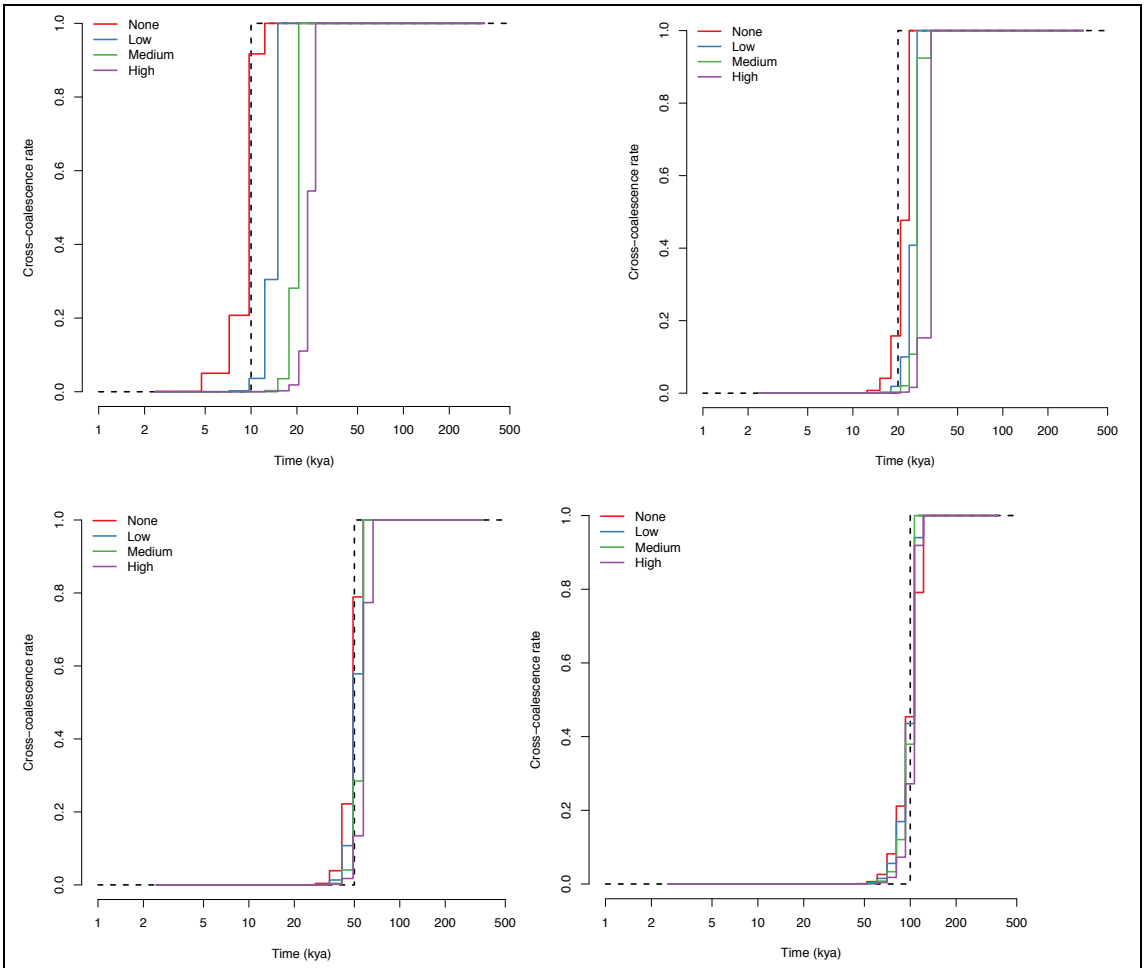


Figure S9.2. Effect of switch errors on MSMC inference of split times. Inferred cross-coalescence rates for simulated data with added phasing errors. We simulated ten 100Mb chromosomes for each of two diploid individuals from two populations that split abruptly at 10, 20, 50 and 100 kya (Ne=14,000 and $\mu=1.2\times10^{-8}$ per-base per-generation). Low, medium and high error rates correspond to setting both the single- and multi-site switch rates to 0.5×10^{-5}, 1.0×10^{-5} and 1.5×10^{-5} per base.

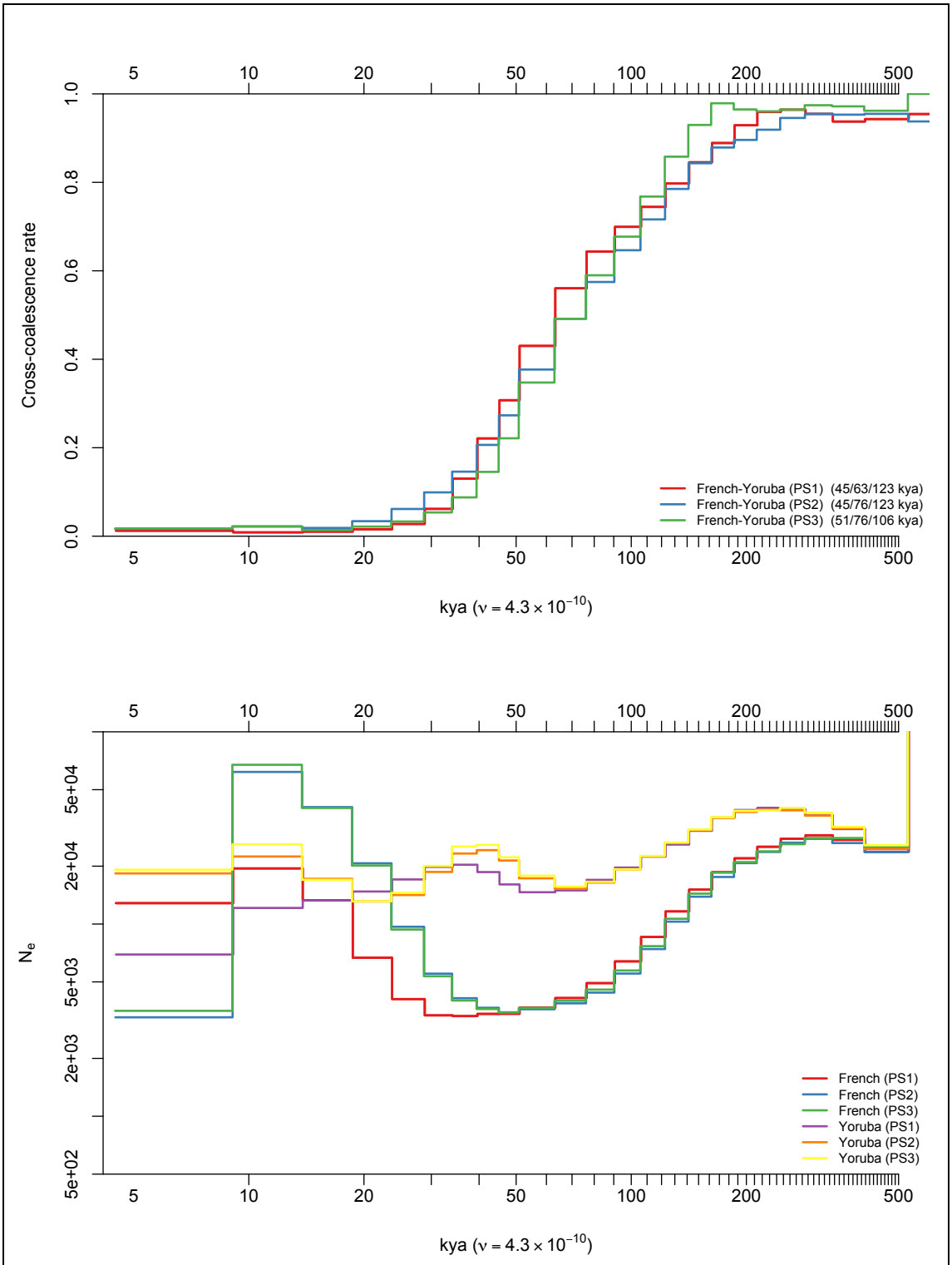


Figure S9.3: French-Yoruba cross coalescence rates (upper panel) and effective population size (lower panel) inferred using MSMC with different phasing strategies.

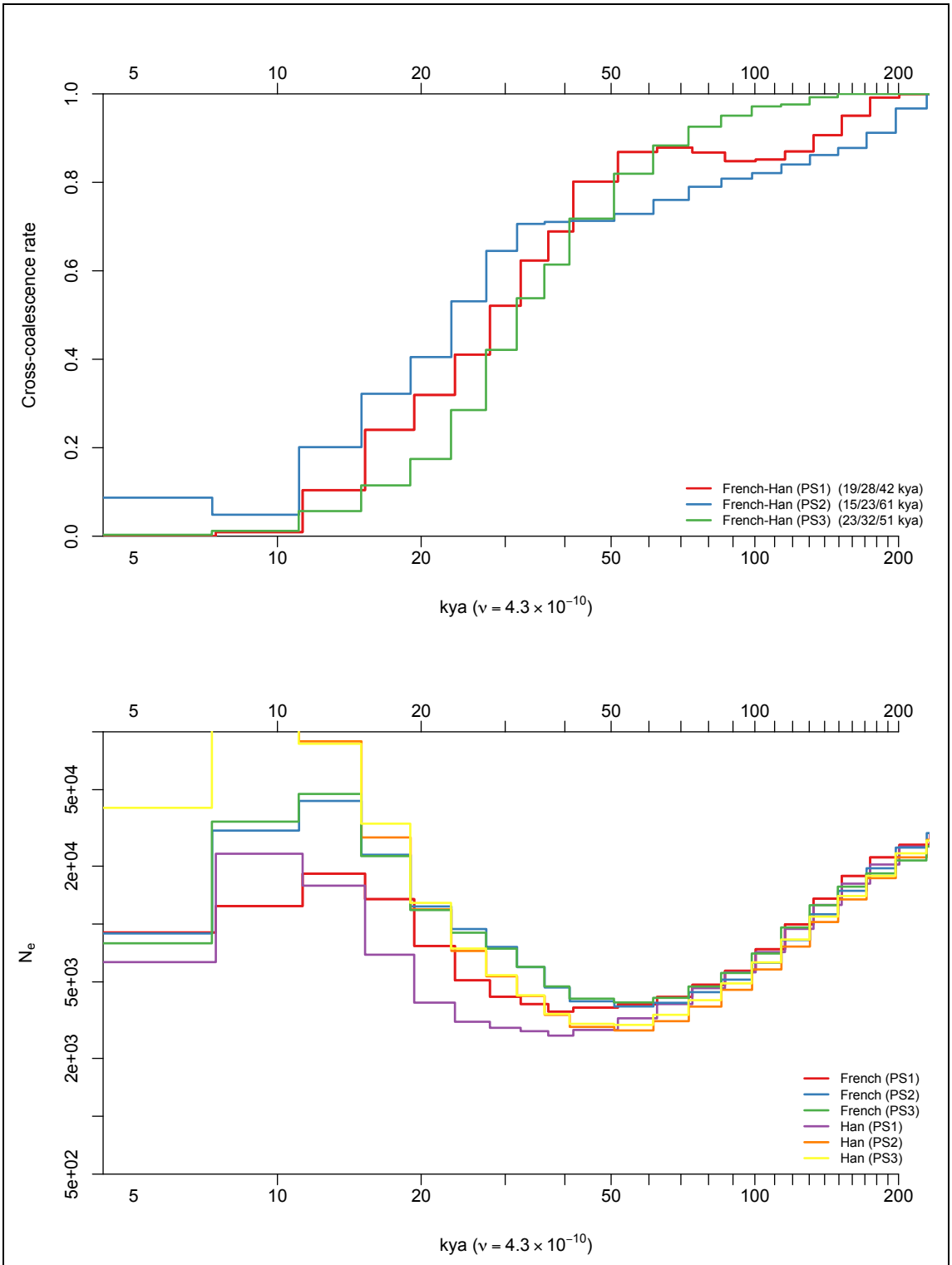


Figure S9.4 French-Han cross coalescence rates (upper panel) and effective population size (lower panel) inferred using MSMC with different phasing strategies.

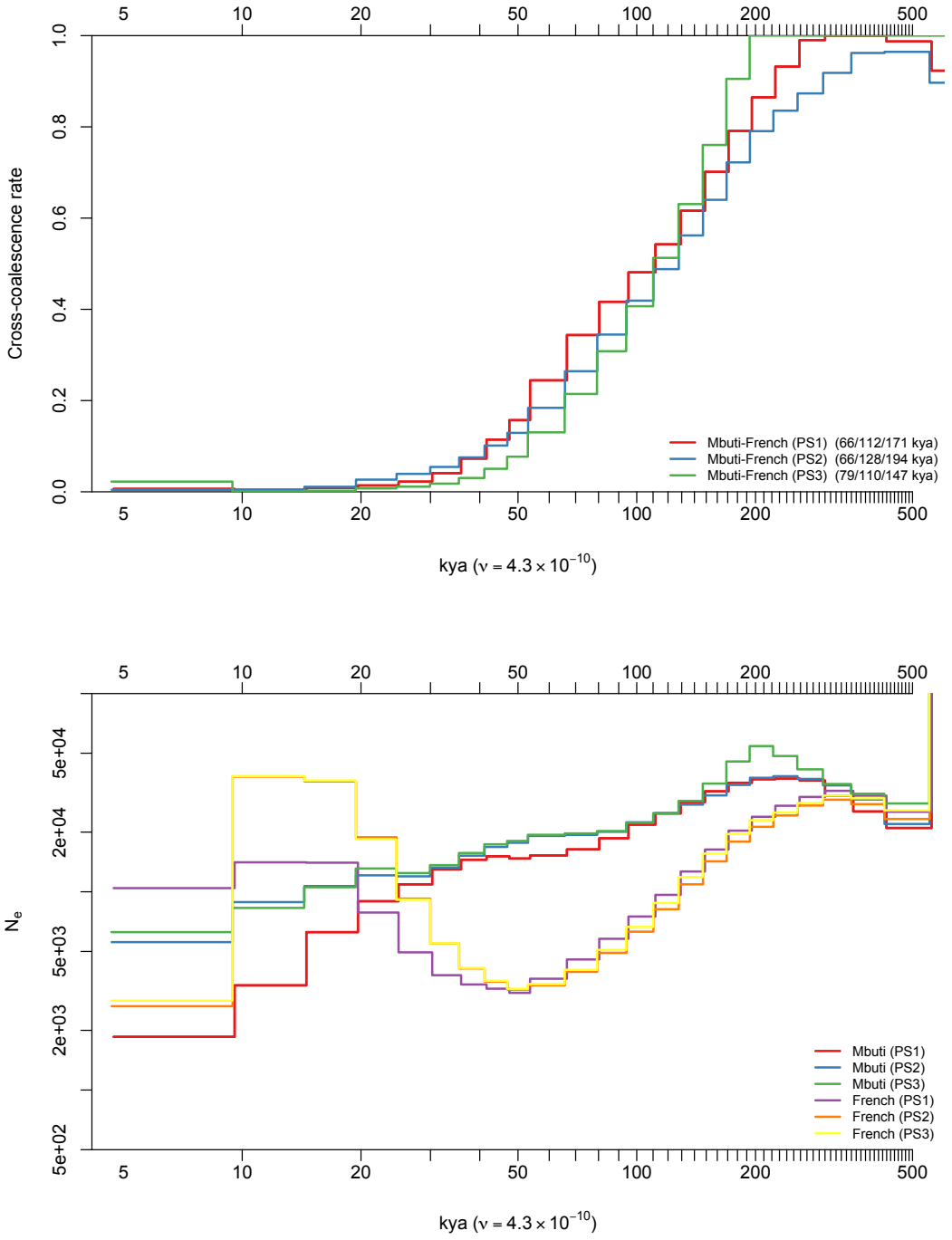


Figure S9.5 French-Mbuti cross coalescence rates (upper panel) and effective population size (lower panel) inferred using MSMC with different phasing strategies

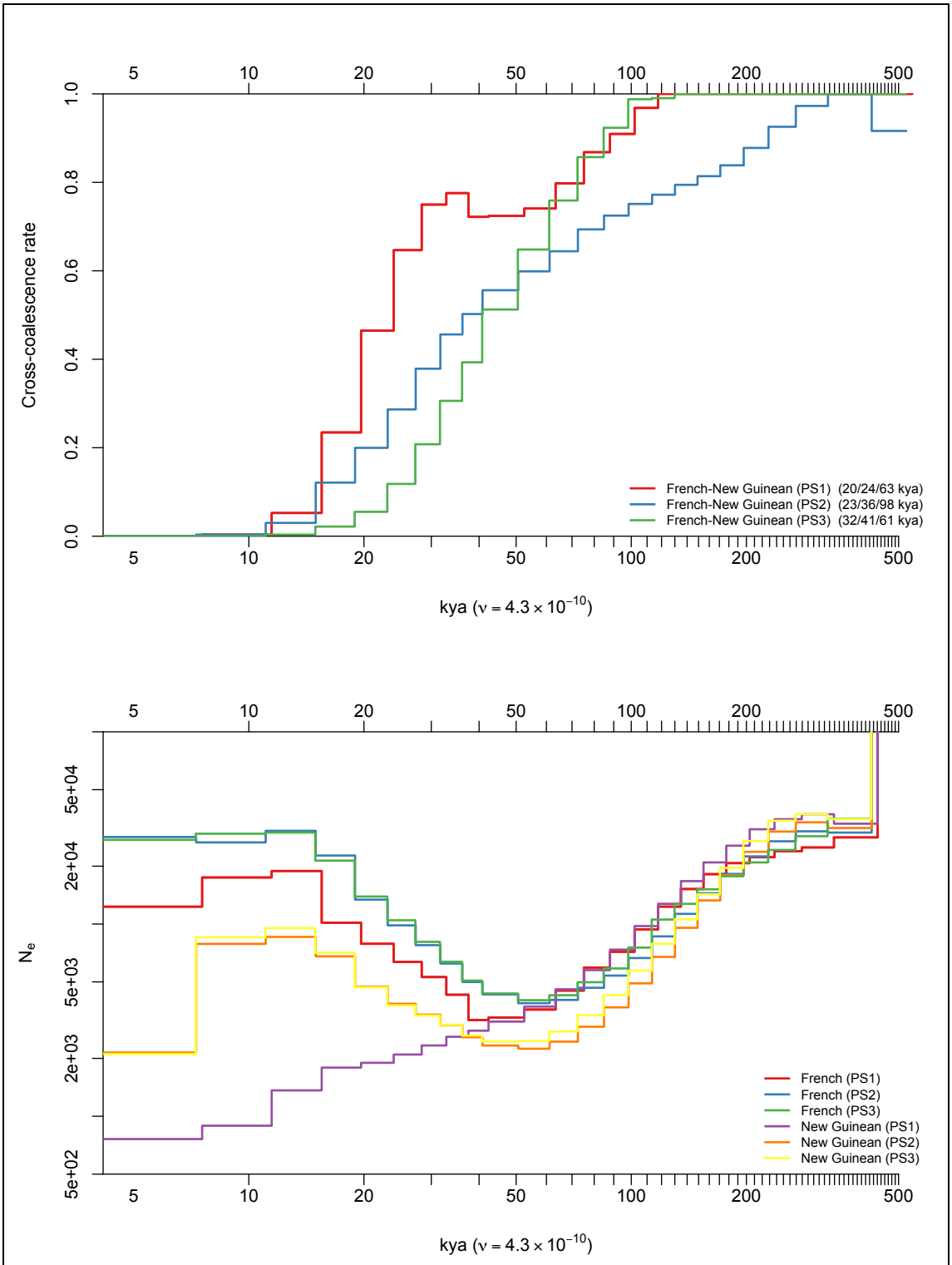


Figure S9.6 French-New Guinean cross coalescence rates (upper panel) and effective population size (lower panel) inferred using MSMC with different phasing strategies

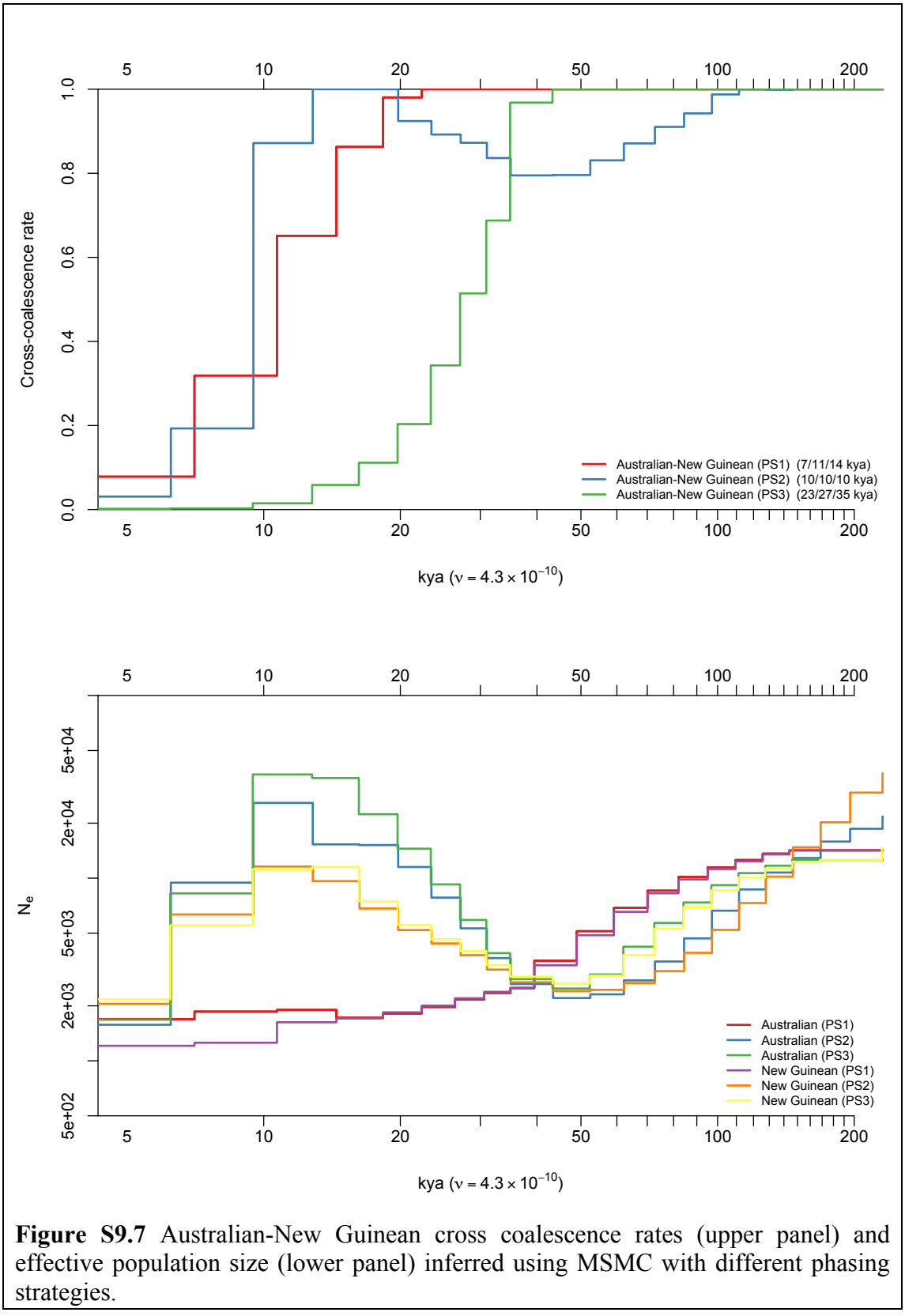


Figure S9.7 Australian-New Guinean cross coalescence rates (upper panel) and effective population size (lower panel) inferred using MSMC with different phasing strategies.

Effect of filter level: We investigated whether the filter level affected the analysis. This might have two effects, first by restricting to higher quality sites, and second by changing the overall heterozygosity and thus changing the appropriate mutation rate (although, as discussed in Section 9.3, this might not be linear). In practice we found that the effect of changing filter levels was minimal (Figure S9.3) and certainly much less than the uncertainty due to phasing error. Thus, we performed all further analysis at filter level 1.

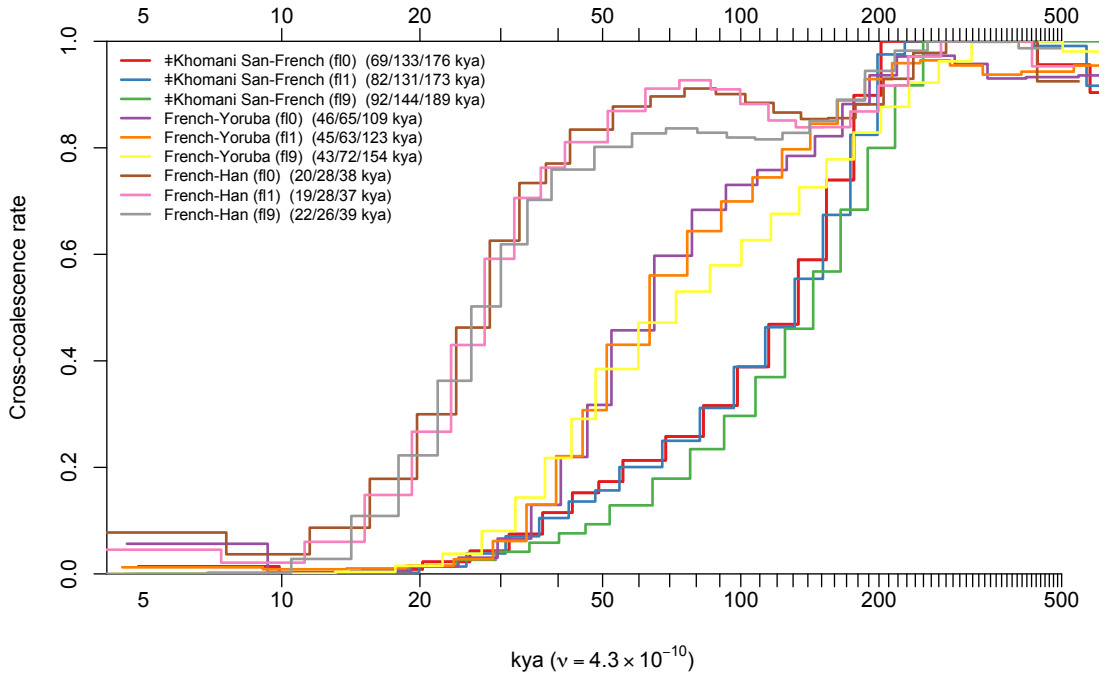


Figure S9.8 Effect of filter levels. We ran MSMC for three pairs of populations at filter levels 0,1 and 9 (using PS1 phasing). Heterozygosity in our data at these levels (0/1/9) is 0.89/0.88/0.84 for ǂKhomani San, 0.96/0.94/0.89 for Yoruba, 0.70/0.68/0.64 for French and 0.68/0.66/0.62 for Han. Filtering to level 9 reduced heterozygosity by around 10% compared to level 0 and increases some split time estimates by up to 10%; however not all estimates are equally affected. For this reason, we do not scale by heterozygosity when reporting our results in the main text. All results in the main text are reported at filter level 1.

References

1 Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496, doi:10.1038/nature10231 (2011).

2 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature genetics* **46**, 919-925, doi:10.1038/ng.3015 (2014).

3 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2012).

4 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

5 Saxena, R. *et al.* Large-Scale Gene-Centric Meta-Analysis across 39 studies Identifies Type 2 Diabetes Loci. *American journal of human genetics*, doi:10.1016/j.ajhg.2011.12.022 (2012).

6 Segurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15**, 47-70, doi:10.1146/annurev-genom-031714-125740 (2014).

7 Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415-423, doi:10.1002/ajpa.20188 (2005).

8 Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44**, 1161-1165, doi:10.1038/ng.2398 (2012).

9 Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature reviews. Genetics* **13**, 745-753, doi:10.1038/nrg3295 (2012).

10 Campbell, C. D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44**, 1277-1281, doi:10.1038/ng.2418 (2012).

11 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475, doi:10.1038/nature11396 (2012).

12 Michaelson, J. J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431-1442, doi:10.1016/j.cell.2012.11.019 (2012).

13 Lipson, M. *et al.* Calibrating the Human Mutation Rate via Ancestral Recombination Density in Diploid Genomes. *BiorXiv preprint* (2015).

14 Palamara, P. F. *et al.* Leveraging distant relatedness to quantify human mutation and gene conversion rates. *BiorXiv preprint* (2015).

15 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449, doi:10.1038/nature13810 (2014).

16 Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nature genetics* **41**, 393-395, doi:10.1038/ng.363 (2009).

17 Staab, P. R., Zhu, S., Metzler, D. & Lunter, G. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* (2015).

Supplementary Information section 10

Sequenced Australians form a clade with previously studied Australians

Pontus Skoglund* and David Reich

*To whom correspondence should be addressed: P.S. (skoglund@genetics.med.harvard.edu)

We considered the possibility that the Australian samples, from the European Collection of Cell Cultures cell line diversity panel[1], might not be a typical population of Australians.

The possibility that the Australian genomes we sequenced have different ancestry from other Australians that have been previously analyzed is important, as in this paper we show that the Australian samples we sequenced are consistent with descending from the same modern human dispersal out of Africa as other non-Africans today (Supplementary Information section 11). A previous claim that Australians harbor ancestry from a distinct migration into Asia[1] was based on a different Australian individual than the ones sequenced as part of the SGDP, and thus it is important to determine if that individual had similar ancestry to the one we sequenced.

To evaluate this question, we took advantage of the fact that there is published genome-wide data from three different indigenous Australians groups from different geographic locations across Australia whose phylogenetic relationship to New Guineans we can compare:

(1) Australian_ECCAC (3 individuals, 2 of which are in the SGDP)
Sampling location unknown. We use genotyping data from the Affytmetrix Human Origins SNP array[2].

(2) Australian_WGA (5 individuals)
Sampled in Arnhem Land in northeastern Australia. We use genotyping data from the Affymetrix Human Origins SNP array[2].

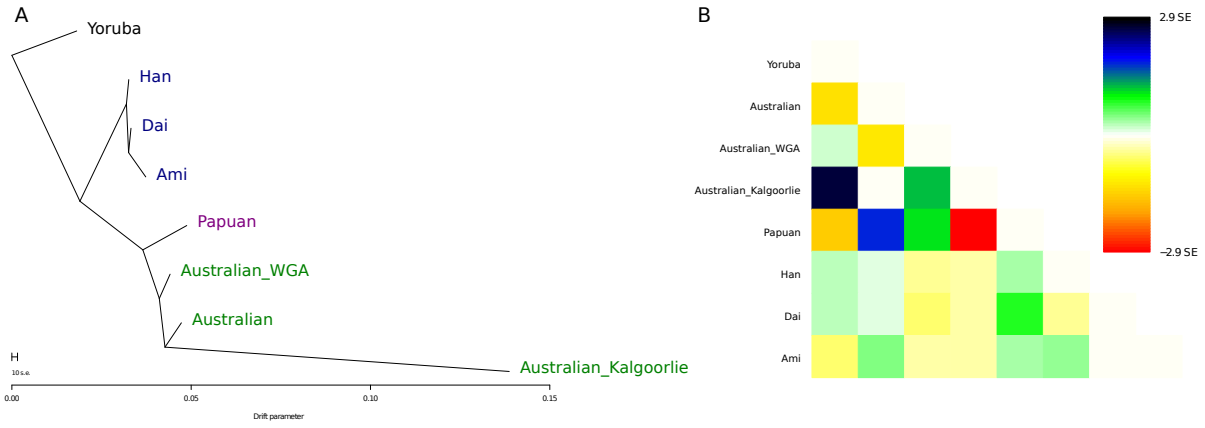
(3) Australian_Kalgoorie (1 individual)
This is the low-coverage genome reported by Rasmussen et al[1], derived from hair sampled in the early 20th century close to Kalgoorie in southwestern Australia. We restricted to reads with a phred-scaled mapping quality of ≥30 and to sites on the reads with base quality of ≥30. At each site in the Affymetrix Human Origins SNP array covered at least once by a base passing these filters, we selected the allele with the highest count, or in case there was a tie we randomly selected a single read to represent the site. We obtained coverage on the individual for 605,986 SNPs (97.8% of all targets).

To test whether the three Australian samples were symmetrically related to New Guineans, we computed all possible *D*-statistics of the form $D(Han, New\ Guinean; Australian_i, Australian_j)$, using the HGDP Papua New Guineans from which the SGDP New Guinean genomes are drawn. This statistic tests whether the two Australian groups are symmetrically related to Han Chinese and New Guineans. If there are differences in relatedness to New Guineans among the three Australian groups (e.g. due to recent gene flow), we would expect the allele frequency differences between the Australian groups to be correlated to be allele frequency differences between Han and New Guineans. However, we find no evidence of such differences, since the *Z*-scores (the number of standard errors that the *D*-statistic is from zero according a Block Jackknife) for all pairs of Australian groups tested are $|Z| < 3$.

[1] https://www.phe-culturecollections.org.uk/pages/M074%20EDP-1%202ug%20Data%20Sheet%20.pdf

We also used TreeMix 1.12[3] to automatically fit a maximum likelihood phylogenetic tree to the three different Australian groups, the New Guineans, eastern non-Africans, and sub-Saharan Africans, dividing the data into 720 contiguous blocks to obtain standard errors. Figure S10.1 shows that a TreeMix tree without admixture events provides a fit to the data within the resolution of the analysis, with the Australian genomes consistent with being a clade with respect to New Guineans. Within Australians, the samples included in SGDP (Australian_ECCAC) are more closely related to the Australian_Kalgoorlie population than either is to the Australian_WGA. Thus, the TreeMix analysis suggests that the Australian_ECCAC genome is not from a population with an unusually close relationship to New Guineans compared to Australian groups with known sampling locations.

Figure S10.1: The sequenced Australians are not atypical in relatedness to New Guineans. We used TreeMix to automatically fit a phylogenetic tree to three different sampled Australian populations, New Guineans, other eastern non-Africans, and sub-Saharan African outgroups. There is no significant evidence for admixture (all |Z|-scores <3). Note that the long external branch leading to Australian_Kalgoorlie is due to only haploid genotypes being used for this individual.



We conclude with two clarifications about what we have and have not established.

(1) What we have established is that the Australians we sequenced form a clade with Australian groups with known geographic locations for which genome-wide data are available (Kalgoorlie and Arnhem Land). The samples we sequenced are typical of other Australian genomes that have been studied in the sense that all are consistent with descending from common ancestors since separation from the ancestors of Papuans.

(2) What we have not established is that the patterns we observe in the genomes we sequenced are typical of all Australians. Thus, it is possible that there are Australian populations that we have not sampled that might have different ancestry, such as speakers of non-Pama-Nyungan languages or indigenous Tasmanians.

References

1 Rasmussen, M. *et al*. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**, 94-98, doi:10.1126/science.1211177 (2011).

2 Lazaridis, I. *et al*. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-413, doi:10.1038/nature13673 (2014).

3 Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS genetics* **8**, e1002967, doi:10.1371/journal.pgen.1002967 (2012).

Supplementary Information section 11

Australo-Melanesians have little ancestry from an early dispersal of modern humans

Mark Lipson*, Iain Mathieson, Pontus Skoglund, Nick Patterson and David Reich

*To whom correspondence should be addressed: (mlipson@genetics.med.harvard.edu)

Summary
Using the SGDP data, we study Australians, New Guineans, and Andamanese and do not find evidence that they have ancestry from a deeply diverging modern human source population. Furthermore, we show that if such an ancestral population did exist, its contribution to the ancestry of these groups is bounded at a few percent.

Motivation
It is widely agreed that anatomically modern humans arose in Africa ~200 thousand years ago (kya). However, the timing and mode of the migrations out of Africa is controversial [1].

Among models proposing multiple dispersals into Eurasia, one of particular interest is the "southern route" hypothesis, which suggests that an early migration through southern Asia brought modern humans as far as Australia, and moreover that some or all of the ancestry of present-day indigenous Australians, New Guineans, and "Negritos" (of the Andaman Islands, the Philippines, and Malaysia) can be traced to this migration [2].

The first line of evidence that has been proposed to support this hypothesis is from physical anthropology, in particular morphological similarities between Australians, New Guineans, and Andamanese and some African groups [2-4]. However, morphological characters can be discordant with population histories (due either to shared retained features or convergent evolution). For example, two ancient DNA studies showed that skeletal remains argued on morphological grounds to reflect distinct migrations into the Americas were in fact derived from the same ancestral population as most other Americans [5-6].

Archaeological evidence has also been interpreted as providing evidence of an early southern route dispersal. Modern humans inhabited Australia by ~47 kya [7], and a new analysis suggests occupation before 50 kya [8], earlier than all known modern human remains in Europe and northern Asia [1]. There is also evidence (thus far only lithic [1]) that modern humans may have spread into Arabia >100 kya [9] and into India >74 kya [10].

From a genetic perspective, the most notable study supporting the southern route hypothesis was that of Rasmussen et al. [11], who analyzed a historical period Aboriginal Australian genome sequence by modeling allele frequency correlation and linkage disequilibrium patterns and estimated an earlier split of Australians from Europeans and East Asians than of the latter two from each other. Two subsequent papers fit genetic data to spatial migration models [3] and a different set of divergence statistics [12] and inferred an early split of New Guineans from mainland Eurasians. However, the methods used in these studies are all potentially confounded by more recent demographic events, including population size changes, archaic introgression [13-16], and other admixture. In particular, if the several percent of Denisovan ancestry present in Australians and New Guineans is not accounted for (as in [11] and [3]), then these populations may appear inaccurately deeply diverged.

In this note, we describe the details of our tests for evidence of an early dispersal of modern humans out of Africa using the SGDP data. We ask whether the patterns in the genetic data are best explained by a model in which Australians, New Guineans, and Andamanese have ancestry from an early dispersal predating the later diversification of northern Eurasian lineages, or alternatively by a model in which these populations and northern Eurasians descend from the same dispersing population. We also place an upper bound on the proportion of ancestry that could derive from an early dispersal.

Difficulty of interpreting MSMC results

We first explored whether we could use MSMC [17] to test for evidence of an early dispersal of modern humans out of Africa that differentially affected present-day French, Han, Australians, and New Guineans (Fig. 2). However, inferences involving Australians and New Guineans were very different depending on phasing strategy (Supplementary Information section 9). In addition, we did not know how the 3-6% Denisovan ancestry in Australians and New Guineans [14] would affect MSMC inferences. Because of these considerations, we were concerned that we could not reliably use MSMC to investigate the early dispersal hypothesis.

Admixture graph construction

We fit an admixture graph (a phylogenetic tree based on shared genetic drift, augmented with point admixture events) relating Australians, New Guineans, and Andamanese to other present-day and ancient populations. Unlike previous tests of early modern human dispersal hypotheses [3, 11-12], we modeled archaic admixture. In all, we co-modeled 10 groups: Chimpanzee, Denisova [15], Altai Neanderthal [16], Dinka, Kostenki 14 [18], Ami, Dai, Onge, New Guinean, and Australian. We represented West Eurasians by Kostenki 14 rather than by a present-day population because of known complications of later admixture [18-19].

We used diploid genotype calls from SGDP individuals and merged with data from the other samples on the set of SNPs (~2.9 million) ascertained as polymorphic among the four SGDP Mbuti. While this means that the SNPs are not polymorphic at the root of the tree (as would be desirable), this property does hold within the modern human sub-tree (disregarding archaic introgression), for which Mbtui serves as an outgroup [20]. The graph is built from shared drift relationships (*f*-statistics) among the populations, as computed on this set of SNPs [21].

Admixture graph model fits well without an early modern human dispersal

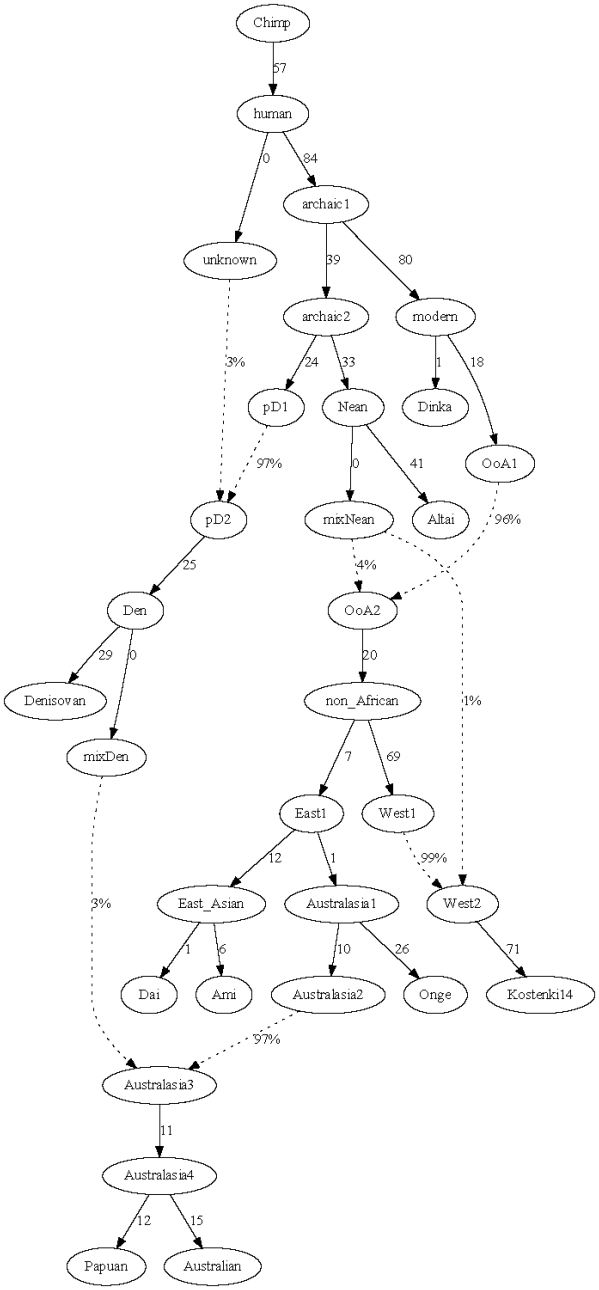
We used ADMIXTUREGRAPH [21] (downloadable as part of the ADMIXTOOLS package: http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html) to optimize the parameters of the admixture graph. After the user specifies a graph topology (including admixture events), the program finds the best-fitting branch lengths and admixture proportions and returns a list of outlier statistics (*f*-statistics relating populations in the graph that deviate significantly from their model expectations) as well as an approximate log-likelihood for the graph as a whole [21]. While the program does not search over the space of possible topologies, its behavior when the specified branching order is incorrect is usually to create an artificial trifurcation (where one split point should be on a different branch), in which case we manually adjust the topology.

We find that four archaic admixture events are necessary to provide a good fit to the data:
(1) Shared Neanderthal introgression into all non-Africans;
(2) Denisovan gene flow into the ancestors of Australians and New Guineans;

(3) Extra Neanderthal ancestry in Kostenki 14 (which dates to 36-39 kya), consistent either with more extensive admixture or with purifying selection against Neanderthal ancestry (which has had more time to operate in present-day non-Africans [22] than Kostenki 14);
(4) Unknown basal archaic admixture in Denisova [16].

Given these admixtures, we optimized the positions of the 10 populations and found that all other relationships were consistent with a tree-like history, as the best-fitting graph (Fig. 3; Fig. S11.1) correctly predicted all *f*-statistics to within 2.1 standard errors.

Figure S11.1: Admixture graph relating Australians, New Guineans, Andamanese, and diverse other populations. Dotted lines denote admixture events, with proportions shown. Edge labels are branch lengths in drift units: a constant (1000) times f_2 distances (average squared differences in allele frequencies [21]). Terminal nodes are sampled populations, and internal nodes are hypothesized ancestral populations. Fig. 3 is a redrawn version of this.



Notably, the topology of the tree does not specify an early divergence of the lineage leading to Australians, New Guineans, and Andamanese. Apart from the Denisovan introgression, these populations are unambiguously placed in an eastern Eurasian clade together with East Asians ($p<10^{-15}$), to which the ancient West Eurasian Kostenki 14 is an outgroup. In an early dispersal scenario, by contrast, at least a portion of the modern human ancestry in these populations would be an outgroup to a clade consisting of the other East and West Eurasians. We speculate that one reason why some studies have not inferred Australians and New Guineans to be sister groups to East Asians is that their Denisovan ancestry causes them to appear more deeply diverged. However, other studies have recovered an eastern clade (East Asians plus New Guineans) even without allowing archaic gene flow [23-24].

We also tested the robustness of this model to two perturbations outside of the eastern Eurasian clade (see next section for further robustness checks). First, as mentioned above, we believe that Kostenki 14 is more appropriate as a representative West Eurasian than later populations because of several layers of complicated admixtures experienced by later Europeans [19]. However, we did re-fit our model using the ~7 kya Neolithic Stuttgart farmer – which lacks some of the most recent gene flow into Europe but is closely related to present-day Sardinians [19] – in place of Kostenki 14, and the graph topology was unchanged. All *f*-statistics were correct to within 2.9 standard errors, with the poorer quality of fit due to admixture in Stuttgart (all *f*-statistics without Stuttgart remained correct to within 2.1 standard errors).

Second, we experimented with manually changing the proportion of Neanderthal gene flow into non-Africans. While the inferred proportion of 3.2% Denisovan introgression into Australians and New Guineans in our best-fitting model is reasonable, the inferred shared Neanderthal introgression is inflated at 4.1%. We believe that this may be due to the ascertainment of SNPs as polymorphic in present-day Africans; as noted previously, this scheme is desirable for studying modern human ancestry in non-Africans, but it could cause some branch lengths leading to archaic humans to be underestimated. Even if this is the case, however, the algebraic shared drift relationships in other parts of the graph should be unaffected by the larger Neanderthal mixture proportion. To test if this reasoning is correct, we manually set the shared Neanderthal gene flow into non-Africans at 2.5%, a reasonable value for eastern Eurasians. The topology of the model is unchanged, with all *f*-statistics remaining correct to within 2.4 standard errors. Finally, we verified that, as in the next section, adding a putative early dispersal ancestry component to Australians and New Guineans (with or without Onge) in the model with lower Neanderthal gene flow did not improve the fit (see below).

Including admixture from an early modern human dispersal does not improve model fit

Although the graph shown in Fig. S11.1 fits the data to within the limits of our resolution, it is possible that adding an additional admixture event could further improve the fit. Specifically, we considered the possibility that Australians, New Guineans, and Andamanese are inferred to be a clade with East Asians because of later gene flow; that is, the former groups have ancestry from an early dispersal that has been diluted due to admixture with populations related to East Asians. To explore this scenario, we built more complex graphs in which Australians, New Guineans, and Andamanese are descended from an admixture involving an East Asian-related component as well as a deeper, early dispersal component that we model as absent in East Asians. While this might not be accurate in the event of bidirectional gene flow, we hypothesize that the early dispersal component would be present

in a smaller proportion in East Asians. Thus, our analysis can be viewed as studying the difference in early dispersal ancestry between the two groups.

First, we added early dispersal admixture into the common ancestor of Australians, New Guineans, and Andamanese and allowed its position and proportion to vary (only constraining the source to split somewhere on the branch above the "non-African" node). In this case, the best fit was obtained with the split at the "non-African" node itself, which is algebraically equivalent to the graph in Fig. S11.1. In other words, no parameters for a model specifying early dispersal ancestry provide a better fit than the simpler model. This was true for the full graph, the full graph without Onge, and the full graph without Australian and New Guinean. We also replicated the result for the full graph using Mbuti as the African outgroup population in place of Dinka. The topology was such that it was more straightforward to model the early dispersal admixture prior to the Denisovan gene flow, but the order of these events (including potential Denisovan admixture into the early dispersal lineage) does not affect the shared drift relationships.

Comparison with Basal Eurasian ancestry in Neolithic Europeans

As a positive control, we carried out an analogous modeling procedure with the Stuttgart Neolithic farmer, which has been shown to be descended in part from a "Basal Eurasian" population that diverged prior to the main eastern/western Eurasian split point [19]. First, we added Stuttgart to our main graph model, with Kostenki 14 present as well. When assumed to be unadmixed, it fit best as a sister group to Kostenki 14, but several significant *f*-statistic outliers were present, with Z-scores up to 4.1. All of the most significant of these statistics indicated that Stuttgart shares excess alleles with outgroups to the main non-African clade (for example, f_4(Dinka, Onge; Kostenki 14, Stuttgart) << 0, |Z| > 4). We then allowed Stuttgart to be admixed, with one component splitting closest to Kostenki 14 and the other (the Basal Eurasian component) splitting above the "non-African" node (as in the previous section). This new model fit better, with no outliers above |Z| = 2.9 and the overall log-likelihood 10.5 higher ($p<10^{-4}$, likelihood ratio test). This is in contrast to our finding that no such admixture for Australians, New Guineans, and/or Andamanese improved the model fit. While the ability to detect admixtures depends in part on the reference populations available, we have three topologically distinct sets of references in both cases: Kostenki 14, eastern Eurasians, and outgroups (Dinka, archaic humans, and chimp) for Stuttgart, and equivalently East Asians, Kostenki 14, and outgroups for Australians, New Guineans, and Andamanese.

Upper bound of a few percent early dispersal ancestry

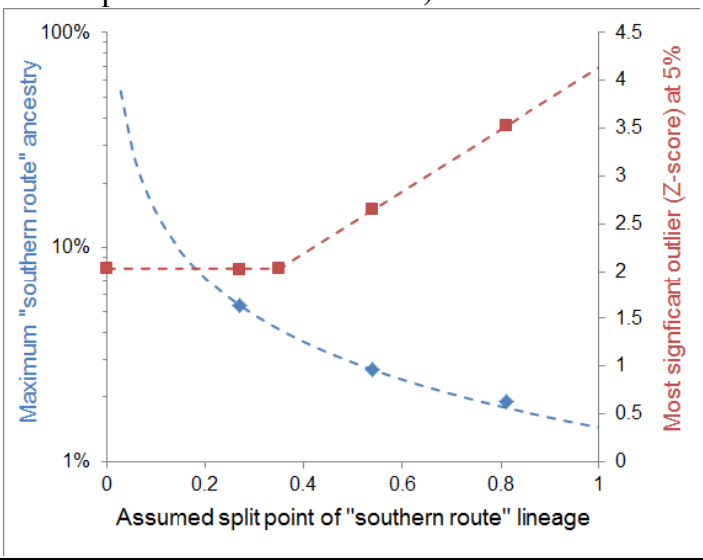
Even though our best-fitting model contained no early dispersal component, we wished to determine the maximum fraction of such ancestry in Australians and New Guineans that would be consistent with the observed drift relationships. Again, we added an early dispersal source mixing into the common ancestor (for simplicity, for this analysis we focused on Australia and New Guinea and omitted Onge from the graph). We fixed the split point of this source at 0.01, 0.02, or 0.03 units above the "non-African" node, where the total length of the branch (from the common ancestor of eastern and western Eurasians back to the common ancestor of Dinka and non-Africans) is 0.037. Previous studies arguing for an early dispersal have placed this split tens of thousands of years before the subsequent Eurasian ancestor [3, 11-12], where calendar-year divergences are roughly proportional to our admixture graph drift units times the effective population size. For example, Rasmussen et al. estimated that Europeans and East Asians split 25-38 kya, Aboriginal Australians 62-75 kya (followed by later gene flow from a lineage related to East Asians), and Africans 81-88 kya [11]. This chronology would place the early dispersal split approximately 35-40 ky earlier than the

eastern/western Eurasian split and approximately two-thirds of the way back to the Eurasian/African split. Thus, our tested points, from ~30-80% of the way back to the Eurasian/African split (in drift units), represent a plausible range of split times (with the estimate from [11] at the high end of the range). Given that non-African population sizes were sharply decreasing over this time period [14-15, 25], the range is shifted downward in terms of calendar years, with the low end on the order of 10-20%, or ~10 ky predating the eastern/western split according to the time scale indicated by our MSMC analyses (Fig. 2; Supplementary Information section 9).

In Fig. 3 inset, we plot the relative likelihood of the graph model (using the multivariate normal approximation from [21]) as a function of the split point and the mixture proportion. The deeper the split, the more divergent the early dispersal ancestry, and hence the less of this ancestry that that can be accommodated while still being consistent with the data. By integrating the area under the curves, we estimate 95% confidence upper bounds of 5.3%, 2.7%, and 1.9% of the deeper ancestry for the 0.01, 0.02, and 0.03 split points, respectively. (By the linearity of *f*-statistics, the bound is expected to be inversely proportional to the split point distance; see Fig. S11.2.)

To complement the full likelihood analysis (Fig. 3 inset), we also computed the absolute Z-score (point estimate divided by standard error) of the most significant outlier in the graph (Figure S11.2). Below a few percent of the early dispersal ancestry, the worst outlier remained as in our best-fitting model, but above this threshold, the worst outlier was f_4(Dinka, Kostenki 14; Dai, New Guinean), which showed too much shared drift in the model between New Guinean and Dinka. The Z-score increased linearly as a function of ancestry proportion, reaching |Z| = 2.5 at roughly 9%, 5%, and 3% for the three split positions.

Figure S11.2: Results of adding putative early dispersal admixture to the graph model, as a function of the position of the early lineage. The position of the early lineage is defined here as its split point along the branch above "Non-African" as a fraction of the total drift, from 0 at the base to 1 at the "Modern" node. In blue is the 95% confidence upper bound of the early dispersal ancestry proportion (dotted line, best-fit with functional form 1/x). In red is the largest outlier *f*-statistic in the graph when the common lineage leading to Australians and New Guineans is assumed to have 5% early dispersal ancestry (dotted line, best-fit piecewise linear function).



Conclusion

Based on patterns of cross-population coalescence and allele frequency correlations, the best-fitting model of Australian, New Guinean and Andamanese history does not involve ancestry from an early-diverging source. Some deep ancestry may be present, but its proportion (beyond that in East Asians) is very likely limited to a few percent.

A caveat is that at present, we are not aware of any ancient DNA data from skeletal remains associated with the earliest putatively modern human archaeological sites in southern Asia or Australia. As a result, our work addresses the specific question of whether present-day Australians, New Guineans, and Andamanese have inherited substantial modern human ancestry that diverged tens of thousands of years earlier than the East Asian/West Eurasian split, rather than the broader question of whether there were any early dispersals at all. In western Eurasia, we know that there existed early modern humans (represented by Ust'-Ishim [26] and Oase 1 [27]) that have no evidence of contributing to present populations. Conversely, even without access to a directly ancestral sample, it has been shown that present-day West Eurasians are descended in part from an early-diverging "Basal Eurasian" lineage [19]. This contrasts with the evidence presented here that present-day Australians, New Guineans, and Andamanese lack an analogous deep ancestry component. Thus, while we cannot rule out the possibility that, in the future, evidence of an early dispersal will be identified from ancient remains, we can confidently conclude that such dispersals did not leave a substantial genetic impact on populations living today.

References

1 Groucutt, H. S. *et al*. Rethinking the dispersal of *Homo sapiens* out of Africa. *Evolutionary Anthropology: Issues, News, and Reviews* **24**, 149-164 (2015).
2 Lahr, M. M. & Foley, R. Multiple dispersals and modern human origins. *Evolutionary Anthropology: Issues, News, and Reviews* **3**, 48-60 (1994).
3 Reyes-Centeno, H. *et al*. Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proceedings of the National Academy of Sciences* **111**, 7248-7253 (2014).
4 Reyes-Centeno, H., Hubbe, M., Hanihara, T., Stringer, C., & Harvati, K. Testing modern human out-of-Africa dispersal models and implications for modern human origins. *Journal of Human Evolution* doi:10.1016/j.jhevol.2015.06.008 (2015).
5 Rasmussen, M. *et al*. The ancestry and affiliations of Kennewick Man. *Nature* **523**, 455-458 (2015).
6 Raghavan, M. *et al*. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* doi: 10.1126/science.aab3884 (2015).
7 O'Connell, J. F. & Allen, J. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *Journal of Archaeological Science* **56**, 73-84 (2015).
8 Clarkson, C. *et al*. The archaeology, chronology and stratigraphy of Madjedbebe (Malakunanja II): A site in northern Australia with early occupation. *Journal of Human Evolution* **83**, 46-64 (2015).
9 Armitage, S. *et al*. The Southern Route "Out of Africa": Evidence for an Early Expansion of Modern Humans into Arabia. *Science* **331**, 453-456 (2011).
10 Blinkhorn J. *et al*. Middle Palaeolithic occupation in the Thar Desert during the Upper Pleistocene: the signature of a modern human exit out of Africa? *Quaternary Science Reviews* **77**, (2013).
11 Rasmussen, M. *et al*. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**, 94-98 (2011).
12 Tassi, F. *et al*. Early modern human dispersal from Africa: genomic evidence for multiple waves of migration. *bioRxiv* preprint doi: http://dx.doi.org/10.1101/022889 (2015).

13 Green, R. E. *et al*. A draft sequence of the Neandertal genome. *Science* **328**, 710-722 (2010).
14 Reich, D. *et al*. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060 (2010).
15 Meyer, M. *et al*. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226 (2012).
16 Prüfer, K. *et al*. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49 (2014).
17 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature genetics* **46**, 919-925 (2014).
18 Seguin-Orlando, A. *et al*. Genomic structure in Europeans dating back at least 36,200 years. *Science* **346**, 1113-1118 (2014).
19 Lazaridis, I. *et al*. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-413 (2014).
20 Pickrell, J. K. *et al*. The genetic prehistory of southern Africa. *Nature Communications* **3**, 1143 (2012).
21 Patterson, N. *et al*. Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).
22 Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354-357 (2014).
23 Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* **8**, e1002967 (2012).
24 Lipson, M. *et al*. Efficient moment-based inference of admixture parameters and sources of gene flow. *Molecular Biology and Evolution* **30**, 1788-1802 (2013).
25 Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496 (2011).
26 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449 (2014)
27 Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* doi:10.1038/nature14558 (2015).

Supplementary Information section 12

No evidence for a shared human selective sweep in the last ~100,000 years

Fernando Racimo*, Heng Li and David Reich*

*To whom correspondence should be addressed: F.R. (fernandoracimo@gmail.com) or D.R. (reich@genetics.med.harvard.edu)

Scan for positive selection

We used the 3P-CLR method[1] to scan the genome for positive selection. We were primarily interested in detecting evidence of selection in the ancestral population of all present-day humans: before the final split of KhoeSan and non-KhoeSan, but after the split from archaic humans.

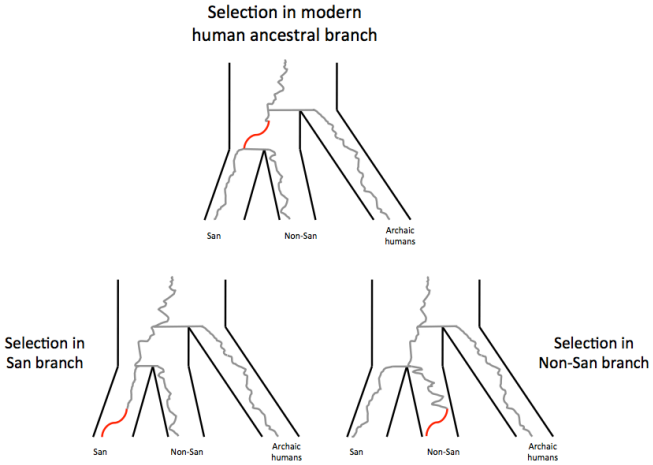


Figure S12.1. Schematic of 3P-CLR scan. We used allele frequency differentiation along a 3-population tree to search for positive selection in the modern human ancestral population, in non-Khoesan, or in Khoesan.

3P-CLR examines patterns of allele frequency differentiation in a tested region of the genome—not just at a single SNP but over an extended window—and computes the approximate composite likelihood for a positive selective sweep, maximized over a range of selection coefficients. Specifically, it compares this likelihood to the approximate composite likelihood of the region under a neutral model of evolution. 3P-CLR can model selection that happened before the split of two populations (after the split from an outgroup), or in either population after the split (Figure S12.1).

To prepare a dataset for 3P-CLR analysis, we restricted to sites for which we have genotype information from two high-coverage archaic humans (the Altai Neanderthal and the Siberian Denisovan), at least 20 African non-KhoeSan individuals, and at least 3 KhoeSan individuals.

To represent the non-KhoeSan, we sampled as many individuals as passed filters at each point in the genome from non-KhoeSan Africans and combined them with an equal number of non-Africans. To obtain as much diversity as possible among the non-Africans, we required that the non-African half of the non-KhoeSan sample be composed in equal parts of individuals from South Asia, West Eurasia, Oceania, East Asia and Central Asia / Siberia. The individuals making up the non-African half of the non-KhoeSan sample were chosen randomly, and if insufficient genotypes from each of these populations were available, the site was not used. As 3P-CLR requires the outgroup sample to be polymorphic, we discarded sites where both archaic humans (Neanderthal and Denisovan) were homozygous derived or where both were homozygous ancestral. We used an African American recombination map[2] to convert

physical distances into genetic distances. We applied 3P-CLR in windows that were 0.25 cM long and that each contained 100 randomly sampled SNPs. Each window was centered on a particular SNP that acted as the candidate beneficial site for that window. We tested a central beneficial SNP every 20 SNPs along the genome.

In Extended Data Fig. 6, we show the 3P-CLR scores along the autosomes for the KhoeSan scan, the non-KhoeSan scan, and the scan in the common ancestral population of modern humans. In Supplementary Data Table 2, we give the numerical results for all windows falling in the 99.9% percentile of the distribution of scores for each scan (we combine contiguous windows that pass this threshold).

(i) Scan in the common ancestral population of all modern humans
In the scan in the common lineage of all modern humans, we do not observe any strong outliers (Extended Data Fig. 6). This result may reflect limited statistical power. Nevertheless, the absence of strong outliers in this scan is of interest because if there had been strong selective sweeps in the common ancestral population of all modern humans since the separation from our archaic human relatives, it is in this scan that such signals would be expected to manifest. The 38 largest peaks we observe (the top 0.1% of peaks in the scan) substantially overlap those recovered earlier using 3P-CLR[1] with 1000 Genomes data[3], when partitioning the daughter branches into African / Non-African, instead of KhoeSan / non-KhoeSan[3].

(ii) Scan on the lineage leading to all modern humans except for KhoeSan
Among the top 5 genes, two are in intergenic regions. The largest signal overlaps *IQCJ-SCHIP1*, an antisense RNA gene that is highly expressed in the brain[4]. The third highest signal overlaps with *MDPZ*, a gene associated with congenital hydrocephalus[5]. The fourth highest signal region overlaps with several genes: *RAP1B, NUP107, MDM2*. *RAP1B* is involved in platelet aggregation[6], *NUP107* plays a role in the assembly of the nuclear pore protein complex[7], and *MDM2* is associated with accelerated formation of tumors[8].

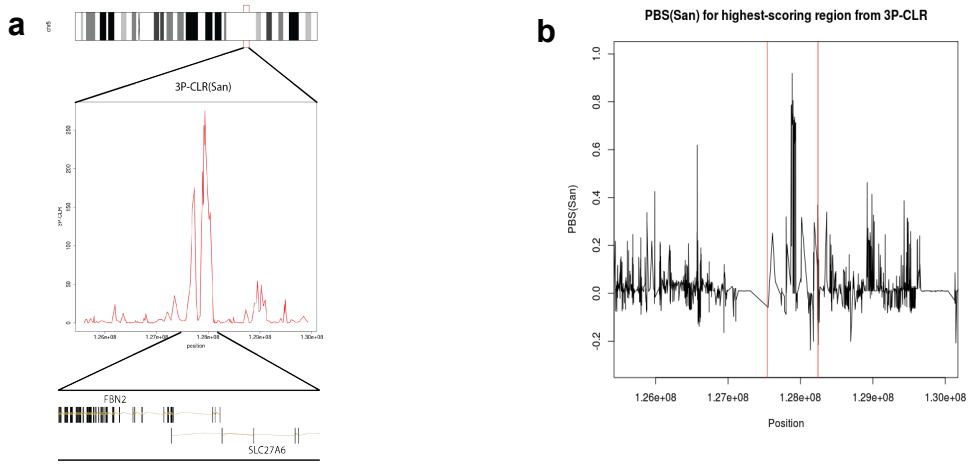
(iii) Scan on the lineage leading to KhoeSan
We observe several strong outliers in the scan leading to the KhoeSan (Extended Data Fig. 6). The largest signal, which is 1.5-times larger than the second highest-scoring signal, corresponds to a region overlapping two genes: *FBN2* and *SLC27A6* (Figure S12.2). We replicate this signal using the SNP-specific PBS statistic[9], which identifies SNPs with locally extended branch lengths exclusive to the San population (Figure S12.2). The signal persists even if we do not condition on the archaic individuals being polymorphic: by computing SNP-wise F_{st}[10-12] between KhoeSan and Non-KhoeSan, we observe a cluster of highly differentiated SNPs in the region identified by 3P-CLR (Figure S12.3).

SLC27A6 codes for a protein that transports long-chain fatty acids across the plasma membrane[13] and may be involved in the uptake of these fatty acids into cardiac myocytes[14]. *FBN2* codes for a fibrillin that is a structural component of connective tissue and regulates elastic fiber assembly[15]. Mutations in this gene have been found to produce congenital contractural arachnodactyly[16-18].

We selected the SNPs with PBS>0.7 in the *FNB2/SLC27A6* region, and used CADD[19] to generate conservation, regulatory and genic annotations at each site. We also

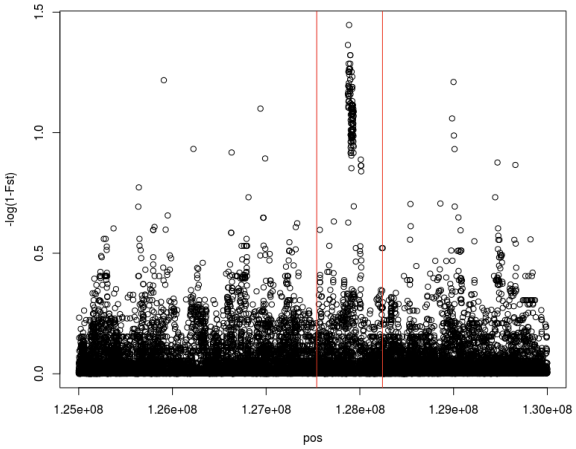
queried GWASdb[20] to check if any of the sites were genome-wide significant GWAS hits, and the GTEx database[21] to check if any of the sites were genome-wide significant cis-eQTLs. We find that none of the sites are GWAS hits or coding changes. Instead, all of them are genome-wide significant cis-eQTLs for *FBN2*, suggesting that the putatively selected haplotype may alter *FBN2* expression in the following tissues: tibial artery, tibial nerve, blood, thyroid, subcutaneous adipose and lung. Among the SNPs with high PBS scores, we find that four that have high (>10) CADD functional disruption scores: rs59567527 (CADD=12.28), rs61375240 (CADD=20.8), rs56766733 (CADD=10.37) and rs6898190 (CADD=10.59).

Figure S12.2. Candidate targets for selection in the Khoesan. a. The region with the highest 3P-CLR score for the Khoesan branch overlaps two genes: *FNB2* and *SLC27A6*. **b.** PBS scores for the Khoesan branch for SNPs in the highest-scoring 3P-CLR region. The red lines denote the boundaries of the candidate region. The statistic was computed on the same sites used for 3P-CLR score.



The SNP with the highest CADD score (rs61375240) is in a GERP conserved element[22] and has high PhastCons conservation scores[23] across primates, mammals and vertebrates. Based on Roadmap Epigenomics (RE) chromatin states[24], the SNP lies in an enhancer region specific to placenta and foreskin keratinocytes.

Figure S12.3. F_{st} between KhoeSan and Non-KhoeSan shows a cluster of highly differentiated SNPs at *FBN2*/*SLC27A6*. We computed F_{st} at each SNP in the region, without requiring the archaic individuals to be polymorphic, as when computing 3P-CLR. The F_{st} values were log-transformed to reflect additive branch lengths. The red lines denote the boundaries of the candidate region identified by 3P-CLR.



The third highest-scoring CADD SNP (rs6898190) lies in an Ensembl promoter flanking regulatory region (ENSR00001289613) that is upstream of *FBN2*. The RE chromatin states indicate that this is a promoter/enhancer region specific to various tissues: placenta, foreskin, astrocytes, breast, bone marrow, skin, muscle, adipose and lung. The latter two are also tissues for which we see that the SNPs are cis-eQTLs, further indicating that this or linked SNPs may alter expression in those tissues.

Scan for loci where the great majority of present humans share a recent ancestor

It has been proposed that the behavioral changes documented in the archaeological record after around 50 kya might have been driven by a new mutation in a gene affecting neurological capacity that swept to high frequency[25]. The evidence from MSMC[26] (Fig. 2a) indicates that the ancestors of some present-day populations were substantially isolated by at least 100 kya, which would be an obstacle to the spread of such a mutation across the ancestors of all populations. However, the MSMC analyses also suggest that gene flow among the great majority of ancestral populations of modern humans continued until around 50 kya (Fig. 2a). Thus, we cannot rule out the possibility that an advantageous mutation that arose in one ancestral population, spread through the others by migration, and then rose in frequency in each population separately under the pressure of natural selection.

To investigate directly whether the genetic data provide evidence of a selective sweep that contributed to all modern humans within this time frame, we took advantage of the fact that the PSMC method produces a posterior decoding—an estimate of the time since the most recent common ancestor (TMRCA) of the two chromosomes that it is comparing—at each position in the genome. This estimate is generated in units of genetic divergence per base pair. To convert this quantity to an estimate in calendar years, we used a previous inference that the average time since the common ancestor of two French chromosomes is 900 ky, based on calibrating to missing evolution in a radiocarbon dated early modern human genome from Siberia (Table S14.6 of ref. [28]).

Concretely, from the PSMC posterior decoding, we obtained the average TMRCA of two chromosomes being compared in a given pair of genomes j at a locus i (t_i^j), and divided this by the average TMRCA for the two autosomes of a French person $\frac{1}{N}\sum_{k=1}^{N} t_k^{French}$ averaged over N loci spaced every 10,000 bp along the genome. This gave us a local estimate of the TMRCA for a test sample j as a fraction of the average autosome-wide TMRCA for a French person. We then multiplied this by the absolute estimate of the TMRCA for French:

$$Local\ estimate\ of\ TMRCA\ for\ sample\ j\ at\ locus\ i = (900\ kya)\frac{t_i^j}{\frac{1}{N}\sum_{k=1}^{N} t_k^{French}}$$

We computed this quantity at equally spaced loci every 10,000 bp along the genome for each of 40 PSMC runs chosen to oversample African genomes. We considered the alternative strategy of analyzing the results of PSMC runs for all SGDP samples, but chose not to do this because with this strategy, >80% of the run would correspond to pairs of non-African genomes. Thus, a selective sweep that occurred in the common ancestral population of all non-Africans, but not in the common ancestor of all modern-humans, could give a significant signal in our scan. The 40 runs were:

24 PSMC runs on Africans: S_Biaka-1, S_Biaka-2, B_Dinka-3, S_Dinka-1, S_Dinka-2, B_Ju_hoan_North-4, S_Ju_hoan_North-1, S_Ju_hoan_North-2, S_Ju_hoan_North-3, S_Khomani_San-1, S_Khomani_San-2, S_Mandenka-1, S_Mandenka-2, S_Masai-1, S_Masai-2, B_Mbuti-4, S_Mbuti-1, S_Mbuti-2, S_Mbuti-3, S_Yoruba-1, S_Yoruba-2, S_Luhya-1, S_Luhya-2, S_Mende-1

8 PSMC runs on non-Africans: S_Dai-1, S_Han-2, B_Australian-3, S_Papuan-1, S_Mala-2, S_French-1, S_Sardinian-1, S_Punjabi-1

8 PSMC runs comparing experimentally two phased genomes from four Africans to a haploid genome of European ancestry CHM1[29]*:* HGDP01029 (Ju_hoan_North), HGDP0456 (Mbuti), HGDP0927 (Yoruba), HGDP1284 (Mandenka)

Extended Data Fig. 7a shows the percent of pairwise coalescent events inferred by the PSMC to be below specified dates. For the 100 kya cutoff, the largest fraction of pairwise PSMC comparisons inferred to be below this threshold—anywhere in the genome—is 68%. This result is difficult to reconcile with the theory that a genetic mutation important to modern human behavior rose to fixation on the human lineage in the last 100 kya. If there was such a locus in the part of the genome we scanned, we would expect a much larger fraction of TMRCAs to be below 100 kya.

A second way to look at this is through the time depth at which the great majority of human genome pairs are inferred to coalesce to a common ancestor based on the PSMC. Extended Data Fig. 7b results for the 80th percentile ("TMRCA80"), the 95th percentile ("TMRCA95"), and the 100th percentile points ("TMRCA100"). The peaks of these distributions are 1300 kya, 1600 kya, and 2000 kya respectively, all far older than the 100 kya cutoff.

A third way to look at this is by studying the extreme low ends of the distribution. As Extended Data Fig. 7c shows, for the 95th percentile point ("TMRCA95"):

- None of the genome has TMRCA95 < 130 kya
- 1/10000th of the genome has TMRCA95 < 300 kya
- 1/1000th of the genome has TMRCA95 < 500 kya
- 1/100th of the genome has TMRCA95 < 810 kya

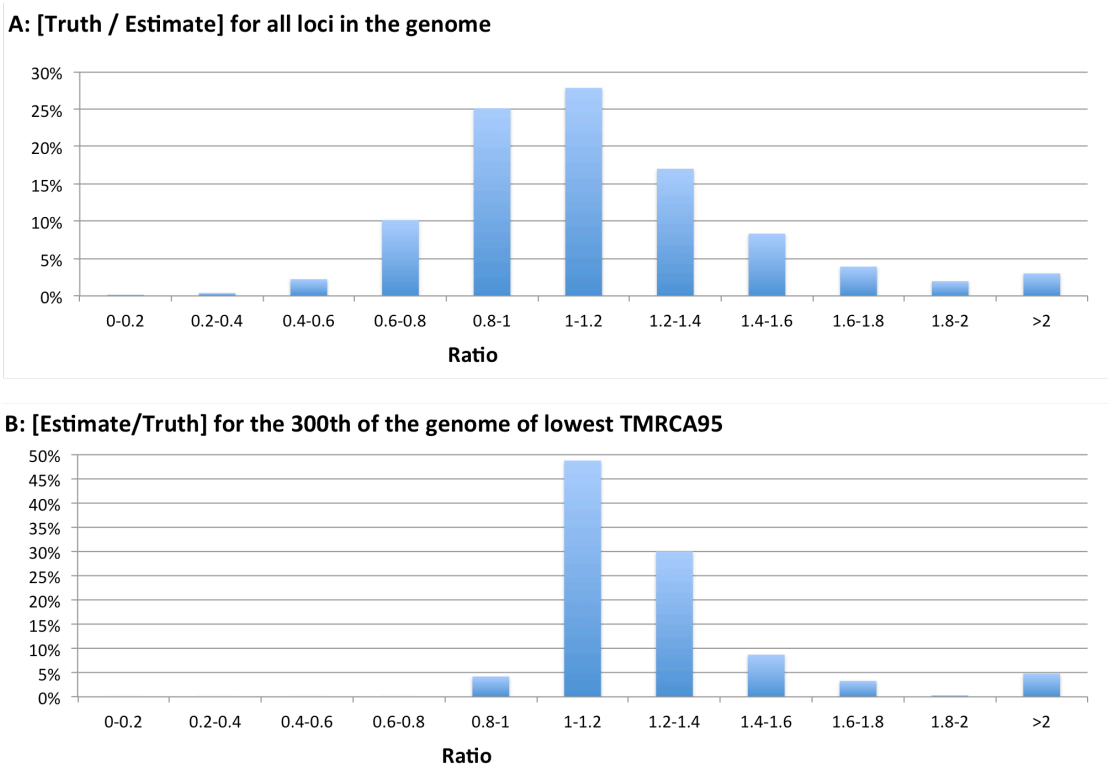
These results highlight how the fraction of the genome that has a prospect of harboring a selective sweep in the common ancestral population <100 kya is minimal.

To investigate the reliability of the PSMC estimates of TMRCA95, we carried out computer simulations of genome-wide data using the software *ms*. We simulated 80 haploid genomes (300 chromosomes of 10 Megabases each) assuming a demographic history that has previously been inferred for the French population[30]. We combined these genomes in 40 pairs, and ran PSMC on each of the resulting diploid genomes.

To analyze these data, we examined the PSMC inference of TMRCA as well as the true TRMCA every 10,000 base pairs along the simulated genomes (a total of 300,000 positions). At each of these locations, we sorted the 40 individuals based both on their inferred TMRCA and their true TMRCA, and recorded the 95th percentile point for each. We recognize that the demographic history of the real samples we are analyzing is substantially different from that of the simulated French. However, the goal of these simulations is only to study how accurately PSMC infers the true TMRCA95.

We generated histograms of the ratio of [Estimated/True] TMRCA95, both for all positions in the simulated genomes (Figure S12.4a), and for the 300th of the genome inferred to have the lowest TMRCA95 (Figure S12.4b). Figure S12.4 suggests that great majority of loci have a ratio of [Estimated/True] TMRCA95 of 0.6-1.8. Thus, our estimated TMRCA is unlikely to be more than a factor of 2 different from the true value. These results imply that it is unlikely that there was a selective sweep common to the great majority of modern humans in the last 50-100 ky, as Extended Data Fig. 7 has negligible density in the range 25-200 ky.

Figure S12.4. Simulations of the PSMC inference of TMRCA95. We present the ratio of the [Truth / Estimate] of the time by which 95% of the 40 simulated genomes coalesce to a common ancestor for (A) all loci, and (B) the 300th of the genome with the lowest true TMRCA95.



We also used the PSMC analysis to specifically examine the 38 largest peaks that emerged in the 3P-CLR scan for selection in the common ancestors of all modern humans. We infer that all these 38 peaks are inferred have TMRCA95 >427 ky. This provides further evidence against these peaks corresponding to sweeps at the time that the archaeological record shows accelerated evidence of behavioral modernity[25].

Table S12.1 also shows the PSMC results at *FOXP2*, which when mutated it cause speech and language pathologies, and which has evidence for natural selection on the lineage leading to modern humans[31]. We estimate that TMRCA50 = 150 ky, far less than the genome average of ~900 ky years, a pattern that might reflect the partial selective sweep previously detected at this locus[31]. However, there is no evidence for a complete sweep since the advent of anatomically modern humans, as TMRCA95 = 1,020 ky, far older than the first known anatomically modern humans (~200 kya).

Table S12.1. Time since most recent common ancestor distribution at *FOXP2*

	Maximum TMRCA	TMRCA95: 95th percentile of TMRCAs	TMRCA80: 80th percentile of TMRCAs	TMRCA50: 50th percentile of TMRCAs
Autosomal mean	3,240 kya	2,350 kya	1,740 kya	900 kya
FOXP2 exon 7	1,890 kya	1,020 kya	540 kya	150 kya

We conclude with two caveats to the PSMC-based analyses. First, our scan only analyzed the autosomes at loci not covered by the *universal mask* of Supplementary Information section 4. Thus, we have not searched for evidence of sweeps on chromosome X, or in repetitive or difficult-to-analyze sections of the genome. Secondly, there is statistical noise in the PSMC posterior decoding, which means that even if there was a locus in which >95% of present-day human sample pairs share a common ancestor <100 kya, it is likely that some would artifactually be inferred to have a substantially older TMRCA. However, as discussed above, the fraction of the genome in which 95% of inferred coalescences below a very lax threshold of 300 kya is minimal (<1/10000th of the genome). Given that our simulations indicate that the TMRCA95 inferences is almost always right within a factor of 0.6-1.8, this gives very little opportunity for <100 ky sweeps in the common ancestors of all modern humans.

References

1 Racimo, F. Testing for Ancient Selection Using Cross-population Allele Frequency Differentiation. *Genetics*, doi:10.1534/genetics.115.178095 (2015).

2 Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170-175, doi:10.1038/nature10336 (2011).

3 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

4 Kwaśnicka-Crawford, D. A., Carson, A. R. & Scherer, S. W. IQCJ-SCHIP1, a novel fusion transcript encoding a calmodulin-binding IQ motif protein. *Biochem Biophys Res Commun* **350**, 890-899, doi:10.1016/j.bbrc.2006.09.136 (2006).

5 Al-Dosari, M. S. *et al.* Mutation in MPDZ causes severe congenital hydrocephalus. *J Med Genet* **50**, 54-58, doi:10.1136/jmedgenet-2012-101294 (2013).

6 Lova, P. *et al.* A selective role for phosphatidylinositol 3,4,5-trisphosphate in the Gi-dependent activation of platelet Rap1B. *J Biol Chem* **278**, 131-138, doi:10.1074/jbc.M204821200 (2003).

7 Boehmer, T., Enninga, J., Dales, S., Blobel, G. & Zhong, H. Depletion of a single nucleoporin, Nup107, prevents the assembly of a subset of nucleoporins into the nuclear pore complex. *Proc Natl Acad Sci U S A* **100**, 981-985, doi:10.1073/pnas.252749899 (2003).

8 Bond, G. L. *et al.* A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell* **119**, 591-602, doi:10.1016/j.cell.2004.11.022 (2004).

9 Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-78, doi:10.1126/science.1190371 (2010).

10 Wright, S. The genetical structure of populations. *Annals of Eugenics* **15**, 323-354 (1949).

11 Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583-589 (1992).

12 Cavalli-Sforza, L. L. Human Diversity. *Proc. 12th Int. Congr. Genet.* **2**, 405-416 (1969).

13 Hirsch, D., Stahl, A. & Lodish, H. F. A family of fatty acid transporters conserved from mycobacterium to man. *Proc Natl Acad Sci U S A* **95**, 8625-8629 (1998).

14 Gimeno, R. E. *et al.* Characterization of a heart-specific fatty acid transport protein. *J Biol Chem* **278**, 16039-16044, doi:10.1074/jbc.M211412200 (2003).

15 Zhang, H. *et al.* Structure and expression of fibrillin-2, a novel microfibrillar component preferentially located in elastic matrices. *J Cell Biol* **124**, 855-863 (1994).

16 Putnam, E. A., Zhang, H., Ramirez, F. & Milewicz, D. M. Fibrillin-2 (FBN2) mutations result in the Marfan-like disorder, congenital contractural arachnodactyly. *Nat Genet* **11**, 456-458, doi:10.1038/ng1295-456 (1995).

17 Wang, M., Clericuzio, C. L. & Godfrey, M. Familial occurrence of typical and severe lethal congenital contractural arachnodactyly caused by missplicing of exon 34 of fibrillin-2. *Am J Hum Genet* **59**, 1027-1034 (1996).

18 Belleh, S. *et al.* Two novel fibrillin-2 mutations in congenital contractural arachnodactyly. *Am J Med Genet* **92**, 7-12 (2000).

19 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

20 Li, M. J. *et al.* GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* **40**, D1047-1054, doi:10.1093/nar/gkr1182 (2012).

21 Consortium, G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110 (2015).

22 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).

23 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

24 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

25 Klein, R. G. & Edgar, B. *The dawn of human culture*. (Wiley, 2002).

26 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature genetics* **46**, 919-925, doi:10.1038/ng.3015 (2014).

27 Steinrücken, M., Kamm, J.A., Song, Y.S. Inference of complex population histories using whole-genome sequences from multiple populations. *biorXiv* **doi:** http://dx.doi.org/10.1101/026591 (2015).

28 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449, doi:10.1038/nature13810 (2014).
29 Steinberg, K. M. *et al.* Single haplotype assembly of the human genome from a hydatidiform mole. *Genome research* **24**, 2066-2076, doi:10.1101/gr.180893.114 (2014).
30 Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496, doi:10.1038/nature10231 (2011).
31 Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**, 869-872, doi:10.1038/nature01025 (2002).