Supplemental Materials for Reconstructing Genetic History of Siberian and Northeastern European Populations

Emily HM Wong, Andrey Khrunin, Larissa Nichols, Dmitry Pushkarev, Denis Khokhrin, Dmitry Verbenko, Oleg Evgrafov, James Knowles, John Novembre, Svetlana Limborska and Anton Valouev

INVENTORY OF SUPPLEMENTAL MATERIALS

1. Supplemental Figures

2. Supplemental Tables

Supplementary Materials and Methods

Samples

Demographic and genetic data were collected from 110 individuals from 14 distinct indigenous populations from Siberia and Western Russia. Among those, 28 were selected for whole-genome sequencing, and the remaining 82 samples were genotyped using Illumina Omni Express platform.

Whole-genome sequencing data for 32 modern-day humans with well-defined ethnicities (Supp Table 2), 13 ancient humans and 2 hominids – Neanderthal and Denisova (Supp Table 3) were obtained from published studies.

Whole genome sequencing

DNA samples of 28 individuals from Siberia and Western Russia (Table 1) were prepared for sequencing using a standard Illumina TruSeq DNA sample preparation protocol. The samples in this study were sequenced on Illumina HiSeq and X10 instruments. All samples were sequenced to a high coverage of 30x or better.

Sequence read processing

We aligned sequence reads onto the human reference genome (hg19), with the mitochondrial sequence replaced by the revised Cambridge reference sequence (rCRS). The read alignment was performed using Novoalign version 3.00.05 (http://www.novocraft.com) with the following parameters: -H -F STDFQ -S 4000 -s 3 -p 7,10 0.4,2 -t 120 –k. Hard clipping was applied to trailing bases with quality scores equal to or less than 2 (parameter -H), polyclonal filter thresholds were set (parameter -p), and unaligned reads were trimmed at a trimming step size of 3 bp until they aligned or failed the QC tests (-s). The polyclonal filtering accepts two sets of thresholds - the first pair of values (n,t) sets a threshold on how many low quality bases are allowed in the first 20 base pairs of each read. Specifically, if n or up to 20 bases with phred quality score below t for a given read, the read would be flagged as polyclonal and discarded. The second pair of values applies to the entire read rather than just the first 20 bp and is entered as fraction of bases in the read below the threshold. After alignment, reads with mapping quality scores (MAPQ) less than 60 were filtered out by using samtools version 0.1.19[1].

For modern human genomes from other studies, reads were extracted from alignment files (.bam) or Sequence Read Archive (.sra) files. They were then treated using the same procedure as described above.

For all ancient genomes (Supp. Table 3), except for Neanderthal and Denisova, we applied the same procedures except the first and last three bases of each read were removed by using BamUtil version 1.0.12 (http://genome.sph.umich.edu/wiki/BamUtil) to minimize the effects of DNA degradation.

Variant calling

The variants were called using GATK variant caller and standard quality filters were applied. Using Unified Genotyper from Genome Analysis Tool Kit[2] (GATK) version 2.7-1, variants were called from each sample using default settings except for the following parameters: (i) the minimum phred-scaled confidence score threshold at which variants should be called was set to be 0 (-stand_call_conf); (ii) the minimum phred-scaled confidence score threshold at which variants should be emitted was 0 (-stand_emit_conf); (iii) and the minimum base quality required to consider a base for calling was set to be 5 (--min_base_quality_score). For variants on the autosomes, those with genotype quality score (GQ) less than 30 were filtered out. Ti/Tv ratios in the range of 1.9-2.2 are typically observed for whole-genome SNP calls and this metric were used to assess the sample quality.

For the Neanderthal and Denisova genomes, their VCF files were downloaded (from http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/ and http://cdna.eva.mpg.de/neandertal/altai/Denisovan/). Variants with GQ less than 30 or labeled as LowQual in the VCF files were filtered out.

We calculated Ti/Tv values for the four Khanty samples that showed differences in the number of non-dbSNP variants. All four samples had dbSNP Ti/Tv values of 2.03-2.04. The samples with 313 and 319 thousand non-dbSNP variants had Ti/Tv values of 1.78 and 1.79; samples with 92 and 85 thousand non-dbSNP variants had Ti/Tv values of 2.02 and 2.0. The lower Ti/Tv values are observed for two out of four samples, indicating potentially higher rate of false positives. However, the values are still reasonably close to 2.0. The slightly higher rate of false positive SNPs is unlikely to have a profound affect on our results and conclusions. For example, in TreeMix, Y-chromosome and mitochondrial DNA (mtDNA) phylogenetic analyses such SNPs would be accommodated within the terminal edges, but would not affect the clustering of samples.

Only samples from other studies with high quality were selected. However, some samples were sequenced on earlier versions of Illumina sequencers, which may lead to some differences in variant calls due to version-specific biases. Differences in the variant calling pipelines between studies would also lead to differences in variant calls. To further evaluate our variant call quality we compared original variant calls (provided by authors) and our GATK variant calls (Supp. Table 4). We used Ti-to-Tv ratio (Ti/Tv) and genotype concordance to evaluate quality of each set of SNPs.

The genotype concordance was 100% across all evaluated samples, showing good replication of the original study. Our analysis shows that our SNP calls for French, Sardinian, Han, San, Mandenka, Dinka and Yoruba were highly overlapping with original sets of variants. For these genomes, Ti/Tv ratios of our calls were generally better compared to original study SNPs.

Y-DNA SNP assessment

We used individuals exhibiting minimal divergence of Y-DNA haplogroups, such as Sherpa-1, Sherpa-2 and Andea-7, Andean-8. These samples have only a small number of unique SNPs (9, 15, 5, 8), which provides an upper bound approximation on the percentage of false positive Y-DNA SNPs based on the number of Y-DNA derived alleles: 0.7% = 9/1398, 1.1% = 15/1392, 0.4% = 5/1324, 0.6% = 8/1322.

Assignment of Y-DNA haplogroups

Previously published genomes had 100% consistent ISOGG haplogroup assignment compared to original studies.

MtDNA SNP assessment

We analyzed individuals with minimal divergence of their mtDNA haplgroups, such as Sherpa-1 and Sherpa-2, Komi KI-1 and Veps-1, Nenets and Mansi-1. These samples have a handful of unique SNPs, which allows us to provide an upper bound estimate for the false positive rate of SNPs: 0% = 0/36, 3% = 1/37, 0% = 0/31, 4% = 1/28, 4% = 1/24, 3% = 1/38.

TreeMix

We used TreeMix[3] version 1.12 to infer the patterns of population splits and admixture events in our populations. TreeMix takes as input genotype data and uses a statistical framework for building population trees and testing for the presence of gene flow between diverged populations.

We created coverage mask based on a Mansi genome (child) to exclude outlier regions exhibiting low- and high-coverage (i.e. with less than 20 reads or more than 100 reads aligned to the position). These regions are likely to be difficult to be sequenced or represent repeats. As all the genomes were sequenced using Illumina HiSeq platform and Mansi genome has a relatively higher coverage (~44x), we decided to use Mansi as the reference genome for creating a coverage mask.

For all TreeMix analyses, we only used autosomal variants that fall outside of the masked regions. Neanderthal and Denisova were used as the root of the tree (-root). We turned off the sample size correction by specifying –noss. To account for linkage disequilibrium (LD), variants were grouped together in windows of size *n* by using the -k flag. The window size *n* was determined by the number of variants in the analysis. In Pickrell et al.[3], the authors used windows of 500 SNPs for a human dataset with around 125,000 SNPs to account for LD; this corresponds to a window size of approximately 10 Mb. Using this as a guide, we calculated the *n* accordingly for each analysis based on the number of polymorphic sites that were included for TreeMix analysis. To generate the jackknife weight estimate and standard error for the admixture event, we used the flag –se. The 95% confidence interval (C.I.) of the weight estimate for admixture event was calculated as jackknife estimate of weight ± 1.96 x jackknife estimate of standard error, assuming the weight estimates follow a normal distribution.

For all the analyses in this section involving ancient genomes, we sampled reads at genomic positions rather than performing genotype calling. For each of the ancient samples, we randomly picked one read at each site and assigned a single allele to that site according to the allele represented by the read. To avoid errors due to post-mortem deamination, we excluded sites with C-T or G-A transitions from the analysis. Since most ancient genomes we included have low sequence coverage (in some cases < 0.1X), we created an additional coverage mask based on each ancient genome. On top of the coverage mask created from the modern human genome, we masked out regions with zero coverage in the ancient genomes included in the analysis.

To visualize trees and residuals from the model, we used an R script modified based on funcs.R from the TreeMix package. The tree graph was further edited in Adobe Illustrator to improve presentation without making changes to the tree topology or branch positions relative to the X-axis.

Y Chromosome analysis

We filtered out variants with GQ scores less than 5 and variants called as “heterozygous” on the Y chromosome. Heterozygous calls on the Y chromosome are likely a consequence of sequencing errors and mapping errors due to its homology to the X chromosome. We applied the filtering mask used in Poznik et al.[4] to exclude regions with repetitive sequences or homology to the X chromosome. The total length of unique regions of Y Chromosome is 10,446,035 bps, which harbored 12,394 SNPs, which were used to compute the tree.

The haplogroup affiliation of each sample was assigned based on ISOGG Y-DNA SNP index[5]. The haplogroup affiliation was not used to infer the tree topology. Rather, tree topology was determined purely on the basis of SNP sharing between samples. For Y-chromosome tree, we created a character state fasta file composing of alleles from the polymorphic sites and constructed the maximum likelihood tree topology using MEGA version 5.2.2[6] with 500 bootstraps.

For calculating divergence time for each branch point in the tree, we first determined the ancestral alleles of each site based on the pairwise alignment of human reference sequence and chimpanzee sequence. The alignment file (in axt format – chained and netted alignment) was downloaded from UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro4/axtNet/chrY.hg19.panTro4.net.axt.gz). For the sites where human sequence was aligned to the chimp genome, the chimp allele was always assumed to be ancestral allele. Therefore, if the human alternative allele was the same as in chimp, the human reference allele was regarded as derived state. For sites that didn’t have chimp counterparts, the human reference allele was assumed to be the ancestral allele, unless the alternative allele was shared by the haplogroup A clade and individual from immediate subsidiary clades and the alternative allele would thus be assumed to be ancestral allele. For sites that were not alignable to chimp and where haplogroup A clade samples (i.e. the two Sans and one Dinka) had a different allele compared to non-haplogroup A samples, we could not determine whether the mutations occurred along the A-clade branch or along the BT-clade branch.

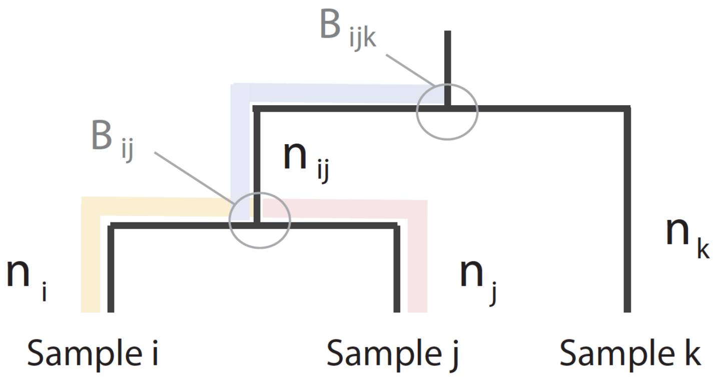
Therefore, for one-half of such sites the derived allele was assigned to the A-clade branch and one half was assigned to the BT-clade branch.

To estimate the branching time (say Bij as indicated in figure 1 below), we counted the number of derived alleles (as assigned above) unique to each branch (ni, nj, nk, and nij). If a derived allele was observed in at least 80% of the samples within a subtree but not in any other samples, the subtree samples without a SNP are likely due to undercalling of that SNP and that allele was considered to be specific to that subtree. If an allele was observed in less than 80% of the samples within a subtree, even if it was not found in any other samples, that allele was considered to be false positive and was not included in any further calculations. It's important to note that we have a lower power to detect variants in Y chromosome compared to autosomes, because Y chromosome is expected to have half the sequencing coverage compared to autosomes.

After assigning the derived mutations to tree branches (including partially supported variants using 80% threshold), we then calculated the mean number of derived mutations per lineage (i.e., $\bar{n} = \frac{n_i+n_j}{2}$). We estimated the branching time T using the following formula:

$$\bar{T} = \frac{\bar{n}}{\mu \cdot L}$$

where μ is the mutation rate of the Y chromosome and L is the total number of nucleotide positions we included in the analysis. We used a mutation rate of 0.82 x 10^{-9} per bp per year as estimated by Poznik et al. (2013)[4].



Supp figure 1: Example of a phylogenetic tree

Under the infinite sites model, accumulation of mutations within Y chromosomal regions follows a Poisson process. Accordingly, the standard error can be estimated using the following formula:

$$T_{SE} = \frac{\sqrt{(\frac{1}{2})^2 \cdot (n_i+n_j)}}{\mu \cdot L}.$$

We then calculated the 95% confidence interval:

95% confidence interval = divergence time estimate ± 1.96 × standard error.

For the Ust'-Ishim Y chromosome, its haplogroup was annotated in the original paper as K(xLT). However, Ust'-Ishim shared one SNP with all individuals specific to N-O branch. This SNP is at position 7,690,182 on the Y chromosome. As listed on ISOGG since 30 March 2014, this SNP may belong to a subgroup upstream from the NO clade. We therefore assigned Ust'-Ishim as a member of N-O clade in Figure 4. We also calculated that Ust'-Ishim's Y chromosome haplogroup had 352 and 335 fewer private SNPs relative to what was expected for a modern-day individual (based on QR and NO clade data respectively). This provided two independent estimates of 41,100 years and 39,200 years for Ust'-Ishim's age using the following formula:

$$Estimated\ age\ of\ Ust'Ishim = \frac{Number\ of\ missing\ mutations}{\mu\ \cdot L}$$

, where number of missing mutations was estimated based on NO clade individuals, or QR clade individuals.

Mitochondrial DNA analysis

We filtered out all the heterozygous variant calls in the mitochondrial genome and excluded sites where there was no coverage in the Mansi (child) genome.

The haplogroup affiliation of each sample was assigned using HaploGrep[7], which is based on the Phylotree database[8]. Samples with haplogroup assignment quality scores less than 0.7 were filtered out from subsequent mtDNA analysis. The maximum likelihood tree was built using MEGA.

Divergence time of each tree branch point was calculated using the same procedure as for the Y chromosome analysis, except for the counting of variants unique to each tree branch. As the mtDNA has a much higher per based mutation rate than Y chromosome, it is reasonable to observe back mutations in mtDNA. A back mutation refers to an instance where more than one mutation events have occurred at a site along the mtDNA lineage. We assigned variants to branches of the phylogenetic tree if they followed the single mutation event model, i.e. assuming only one mutation event occurred along the lineage. For sites incompatible with the phylogeny under this simple model, we assumed that the most parsimonious and probable scenario would be one that involved the least possible mutation events. If the minimum number of possible mutation events was greater than 8, then we would exclude that site. We used the mutation rate of 2.3×10^{-8} per bp per year as estimated by Poznik et al[4], and the total number of bases on the mtDNA we included in the analysis was 14,035 bp.

The multiple sequentially Markovian coalescent (MSMC) analysis

After aligning the reads against hg19, we called the consensus genome sequence for each sample using samtools, and we created masks and called genotypes using

MSMC[9] bamCaller.py. Genotypes were phased using SHAPEIT[10] with the 1,000 Genomes phase 1 haplotypes as the reference panel. The masks and the phased genotypes were then used as input for the MSMC analysis.

To estimate separation time between populations, we ran MSMC on either four or eight haplotypes, i.e. two haplotypes from each of the two populations or four haplotypes from each of the two populations. Specifically, we used the following parameters: --fixedRecombination --skipAmbiguous -p 40*1+15*2. Sites with ambiguous phasing are removed from the analysis (--skipAmbiguous).

Five samples (French, Russian, Mansi, Evenki and Han) were used as reference populations, and their separation times to other populations were computed. For each population analyzed, we reported the median divergence time estimate relative to each of the five reference populations. For the mutation rate and years per generation, we used 1.25 x 10^{-8} per bp per year and 30 years as in Schiffels et al. paper[9].

D-statistic tests

To investigate the relationship between populations of interest and provide statistical support for the admixture events inferred by TreeMix, we performed D-statistic tests in the form of **D (P_1, P_2, P_3, O)** using ADMIXTOOLS[11] release 1.1. ADMIXTOOLS uses the following notation: **D(W,X,Y,Z).** Our parameters correspond to the ADMIXTOOLS parameters in the following way: **O=W, P_1 = Z, P_2 = Y, P_3=X.**

In this test, the null hypothesis is that the tree topology (((P_1, P_2), P_3), O) is correct and there is no gene flow between P_3 and either P_1 or P_2, or any populations related to them. The D-statistic test can be used to evaluate if the data is inconsistent with the null hypothesis, and it is constructed as

$$D = \frac{(nABBA - nBABA)}{(nABBA + nBABA)},$$

where nABBA is the number of sites where P_1 and outgroup share the same allele, when P_2 and P_3 share a different allele (denoted as ABBA sites); and nBABA is the number of sites where P_2 and outgroup share the same allele, when P_1 and P_3 share a different allele (denoted as BABA sites). The standard error for D-statistic is computed using weighted block jackknife[11]. The number of standard errors that this quantity is away from zero forms a Z-score, which is approximately normally distributed. D-statistic is zero under the null hypothesis. A D-statistic that differs significantly from 0 provides evidence either of the tree topology is incorrect and P_3 is not an outgroup to P_1 and P_2, or that there is a gene flow between P_1 and P_3. If the test statistic is significantly less than 0, it suggests that P_3 shares more alleles with P_2 than it does with P_1 (ABBA model), indicating a gene flow between P_2 and P_3. If the D-statistic is significantly greater than 0, it suggests P_3 shares more alleles with P_1 than it does with P_2 (BABA model), therefore suggesting a gene flow between P_1 and P_3.

For all analyses in this section involving ancient genomes, we sampled reads at genomic positions rather than performing genotype calling. For each of the ancient samples, we randomly picked one read at each site and assigned a single allele to that site according to the allele represented by the read. To match the sampling of a single allele (read) at a site for the ancient sample, we sampled a single allele for each of the modern individuals at random based on their GATK-inferred genotype. To avoid errors due to post-mortem deamination, we excluded sites with C-T or G-A transitions.

Principle component analysis (PCA)

To investigate genetic similarities of Siberians with Europeans, Asians and Americans, we performed a principal component analysis (PCA). In the analysis, we included 110 samples from this study (28 were sequenced and 82 were genotyped with microarrays) and 782 samples from seven other studies (all were genotyped using microarrays). A summary of these 892 samples can be found in Supp. Table 1.

Quality control on the samples and variants was performed on each dataset separately by using PLINK release 1.07[12]. Sites with less than 98% genotyping rates and related samples were excluded from the analysis (except for the Mansi trio as we would like to know how they are related to other samples on the PCA plot). Datasets were then merged after quality control. Sites with less than 1% minor allele frequency in the whole data were excluded.

As genotypes of samples were obtained by using different platforms, only autosomal sites where variations were observed across all platforms were included in the PCA. 137,639 sites were included. PCA was performed by using smartpca in the EIGENSOFT package[13] version 4.2, and the results were plotted in R 2.15.2[14].

Note:

The map used in Figure 8 was modified from that downloaded from National Geographic MapMaker tool (http://mapmaker.education.nationalgeographic.com/?ar_a=1&b=1&ls=000000000000).

References

1 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

2 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

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).

4 Poznik, G. D. *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**, 562-565, doi:10.1126/science.1237619 (2013).

5 International Society of Genetic Genealogy (2014). Y-DNA Haplogroup Tree 2014, Version: 9.76, Date: 29 August 2014, http://www.isogg.org/tree/ [Date of access: 5 September 2014].

6 Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution* **28**, 2731-2739, doi:10.1093/molbev/msr121 (2011).

7 Kloss-Brandstatter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human mutation* **32**, 25-32, doi:10.1002/humu.21382 (2011).

8 van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human mutation* **30**, E386-394, doi:10.1002/humu.20921 (2009).

9 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).

10 Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods* **10**, 5-6, doi:10.1038/nmeth.2307 (2013).

11 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093, doi:10.1534/genetics.112.145037 (2012).

12 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575, doi:10.1086/519795 (2007).

13 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).

14 R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. (2012).