**Supplementary Text**

**Materials and Methods**

**Sample collection**

A total of 60 peripheral blood samples from four populations (including 18 Baltis from Gilgit-Baltistan, 20 Kashmiris from Azad Jammu and Kashmir, 2 Punjabis from [Punjab](https://en.wikipedia.org/wiki/Punjab,_Pakistan) in Pakistan, and 20 Pamiri Tajiks from Gorno-Badakhshan Autonomous Region in Tajikistan) were collected in this study with informed consent. The study protocol was approved by University of Health Sciences, E.N. Pavlovsky Institute of Zoology and Parasitology, Academy of Sciences of Republic of Tajikistan, and the Internal Review Board of Kunming Institute of Zoology, Chinese Academy of Sciences.

**Whole-genome sequencing and SNP calling**

The whole genomes for 60 individuals were sequenced using Illumina Hiseq X10 platform to a mean, per-sample depth of approximately 30X for 150bp paired-end reads. The recently published WGS reads of 20 Persians from Kerman in Iran (Charati, et al. 2019) and 8 Tibetans from Tibet in China (Yang, et al. 2018) were included in joint SNP calling. We aligned the reads for 88 individuals to the human reference genome (hg19) using BWA-MEM (v0.7.12) with default parameter settings (Li 2014). Picard v1.119 (https://broadinstitute.github.io/picard/) was used to marked duplicates, and base quality score recalibration modules under known sites file of human as training data. We performed the joint SNP calling with the HaplotypeCaller module (DePristo, et al. 2011) in the Genome Analysis Toolkit (GATK, version 3.6) (McKenna, et al. 2010). We followed the GATK variant quality score recalibration to finally call and filter the SNPs (Van der Auwera, et al. 2013). To check the batch effect of joint SNP calling potential bias in data merging, we followed the reported strategy to conduct single-sample SNP calling for each of 126 genomes including 60 newly sequenced, 20 Persians from Kerman in Iran (Charati, et al. 2019), 41 Tibetans from Tibet in China (Lu, et al. 2016; Yang, et al. 2018) and 5 Sherpas (Lu, et al. 2016).

The pipeline for this procedure is describrd as following:

Mapping the data to the reference:

bwa mem -K 10000000 -M -Y -t $thread $ref -o $sam.sam$read1$read2 -R "@RG\tID:$sam\tSM:$sam\tPL:$pl"

Converting to BAM, sorting and marking duplicates:

samtools sort -@ $thread -O CRAM --reference $ref -o $sam.sort.cram $sam.sam

sambamba markdup -t $thread ${sam}.sort.bam ${sam}.sort.mkdup.bam

samtools view --reference $ref -O CRAM -o $sam.sort.mkdup.cram ${sam}.sort.mkdup.bam

Base quality score recalibration:

gatk --java-options "-XX:GCTimeLimit=50 -XX:GCHeapFreeLimit=10 -Xms${mem}m -Xmx${mem}m -Djava.io.tmpdir=/tmp" \

BaseRecalibrator -I${sam}.sort.mkdup.bam -R $ref-O ${sam}.grp && \

gatk --java-options "-XX:GCTimeLimit=50 -XX:GCHeapFreeLimit=10 -Xms${mem}m -Xmx${mem}m -Djava.io.tmpdir=/tmp" \

ApplyBQSR -R $ref -I ${sam}.sort.mkdup.cram --bqsr-recal-file ${sam}.grp -O ${sam}.sort.mkdup.recal.cram

call gvcf:

gatk --java-options "-Xms${mem}m -Xmx${mem}m -Djava.io.tmpdir=/tmp" \

HaplotypeCaller -R $ref-I $cram-O$sam.g.vcf.gz --emit-ref-confidence GVCF

joint call:

gatk --java-options "-Xms${mem}m -Xmx${mem}m -Djava.io.tmpdir=/tmp"

CombineGVCFs-V$sam1.g.vcf.gz -V $sam2.g.vcf.gz $region -O Combine.g.vcf.gz-R ${ref}&& \

gatk--java-options '-Xms${mem}m -Xmx${mem}m -Djava.io.tmpdir=/tmp' GenotypeGVCFs \

-V /data/${workdir}/Combine.g.vcf.gz \

-O jont\_out.$region\_block.vcf.gz\

-R ${ref} [-all-sites]

variant quality score recalibration:

gatk VariantRecalibrator -V $vcf\_path\

-O /data/${vcf\_name}.recal \

--tranches-file /data/${vcf\_name}.tranches \

-resource:hapmap,known=false,training=true,truth=true,prior=15 $hampmap \

-resource:omni,known=false,training=true,truth=false,prior=12 $omni \

-resource:dbsnp,known=true,training=false,truth=false,prior=2 $dbsnp \

-an MQRankSum -an ReadPosRankSum -an FS -an MQ -an SOR -an DP \

--rscript-file ${vcf\_name}.R --mode BOTH --max-gaussians 4&& \

gatk ApplyVQSR -V ${vcf\_path} --recal-file /data/${vcf\_name}.recal \

--tranches-file /data/${vcf\_name}.tranches \

--truth-sensitivity-filter-level 99.7 \

--create-output-variant-index true -mode BOTH \

-O /data/${vcf\_name%.vcf.gz}.vqsr.vcf.gz

**Data merging**

1. **WGS dataset of 455 individuals**

We merged our WGS data of 88 individuals with the published dataset of 33 Tibetans (Lu, et al. 2016), 20 Pakistani described in Simons Genome Diversity Project (Mallick, et al. 2016), 103 Han Chinese in Beijing of China, 103 Gujarati Indians in Houston of USA, and 108 Yoruba individuals in Ibadan of Nigeria from the 1000 Genomes Project (The 1000 Genomes Project Consortium. 2015) (supplementary table S2). PLINK v1.9 (Chang, et al. 2015) was used to merge and filter the data. These individuals with a missing rate more than 5% were excluded from the analysis. Only autosomal variants were considered. Then we filtered the SNPs with minor allele frequency <5%, genotyping success rate <95% and/or Hardy–Weinberg equilibrium <0.0001.

1. **WGS dataset of 897 individuals**

We merged the WGS data of 126 genomes generated by single-sample SNP calling with the published genomes from 749 individuals of Human Genome Diversity Project (Bergström, et al. 2020) and 4 Punjabi described in Simons Genome Diversity Project (Mallick, et al. 2016). Then follow the above quality control except minor allele frequency <5%, we filtered the SNPs with minor allele frequency <1%.

**Data phasing**

We phased the datasets of 455 and 879 individuals, respectively. The autosomal chromosomes for all individuals were jointly phased with SHAPEIT v2.r727 (Delaneau, et al. 2013) referring to the panel of the 1000 Genomes Project Phase 3 (The 1000 Genomes Project Consortium. 2015) and the genetic map of HapMap (International HapMap 3 Consortium. 2010).

**Analysis of population structure and ROHs**

As for the dataset of 455 and 879 individuals, we pruned SNPs in strong linkage disequilibrium (LD) pairwise genotypic correlation r2 >0.2 in a window of 50 SNPs (skidding the window by 5 SNPs at a time) using the ‘-indep-pairwise 50 5 0.2’ command in PLINK 1.9, respectively (Chang, et al. 2015). We performed PCA with EIGENSOFT v6.1.4 (Patterson, et al. 2006) and ADMIXTURE (v4.1) (Alexander, et al. 2009) with the pruned data (supplementary table S3). We ran ADMIXTURE twice with different random seeds. As for the merged dataset with ancient DNA data, we also carried out PCA and ADMIXTURE analyses. The results of ancient DNA (not shown) were in agreement with those based on modern data, suggesting the ancient data merging was not biased. Based on the pruned WGS dataset of 455 individuals, ROHs for genomic regions having at least 2 Mb long and 50 SNPs with minor allele frequency above 0.05 were calculated by using PLINK 1.9 with allowance for missing and heterozygous calls (Ceballos, et al. 2018)

**Detection of gene flow**

We computed various f3 and f4 statistics using the qp3Pop and qpDstat program in ADMIXTOOLS (v4.1) (Patterson, et al. 2012) to estimate the relative contribution of ancestral populations to the Balti population. The f3 and f4 statistics were also calculated to check the gene flow from the Tibetan population to ancient DNA samples from South and Central Asia. We used TreeMix v4.1 (Pickrell and Pritchard 2012) to infer population tree and gene flow. The relative degree of gene flow was expressed as IBD sharing statistic between pair of populations (Bai, et al. 2018). The IBD segments were detected by fastIBD (Browning and Browning 2011). The majority of recent IBD segments with lengths greater than 2 cM were identified. To enrich the signal for recent gene flow, IBD segments < 2 cM in length or with a fastIBD score < 10−10 were filtered out.

**Haplotype-based ancestry inference**

We referred to the reported strategy (Busby, et al. 2015) to conducted haplotype-based ancestry estimation. We first applied CHROMOPAINTER (Lawson, et al. 2012) to reconstruct haplotype similarity for the phased dataset of 455 individuals assigned as donors (self-copy). The recombination scaling constant Ne and the mutation parameter θ were jointly estimated for every individual in a subset of chromosomes (1, 5, 8, 9, 11, 13, 17 and 21) with 10 Expectation-Maximization steps. The average Ne and θ values across chromosomes were weighted by chromosome size and then used for subsequent CHROMOPAINTER runs on all autosomes (Ne = 2191.361815 and θ = 0.006759352). The fineSTRUCTURE (Lawson, et al. 2012) was performed based on the similarity of the haplotype copying profiles obtained by CHROMOPAINTER with 1 million MCMC iterations. The MCMC file (.xml) was used to construct the tree-like hierarchy with the option –T to show the population structure.

According to the above population groupings by fineSTRUCTURE, we assigned 433 donors of 14 clusters as and 22 receipts of three clusters (i.e. Balti, Burusho, and Hazara) and re-ran CHROMOPAINTER for subsequent SOURCEFIND and GLOBETROTTER analyses. The Burusho and Hazara were characterized as admixed populations (Lawson, et al. 2012) and were used as for comparison. The SOURCEFIND (Chacon-Duque, et al. 2018) was used to estimate the sub-continental ancestry components related to present-day populations of the Balti, Burusho, and Hazara populations. The SOURCEFIND was performed with 30 independent runs using 200,000 iterations each run. We performed GLOBETROTTER (Hellenthal, et al. 2014) according to the protocol described by Hellenthal et al. (2014). Using the chunk length output and painting samples, we ran GLOBETROTTER with the ‘prop.ind: 1’ and ‘null.ind: 1’ options. We estimated the significance of the estimated admixture date by running 100 bootstrap replicates using the ‘prop.ind: 0’ and ‘bootstrap.date.ind: 1’ options; we considered date estimates between 1 and 400 generations as evidence of admixture. For the Balti population presenting evidence of admixture, we repeated GLOBETROTTER analysis with the ‘null:ind: 0’ option.

We also assigned more populations from China and Pakistan as donors based on the phased WGS dataset of 879 individuals (supplementary table S2 and S3). We ran CHROMOPAINTER, SOURCEFIND, and GLOBETROTTER to infer the genetic sources and admixture scenarios for the Balti population as described above. The results were used to check the batch effect of joint SNP calling as well as potential bias in reference population sampling and data merging for the analyses of the WGS dataset of 455 individuals.

**LD-based dating admixture**

We employed LD-based methods ALDER (Loh, et al. 2013), under the hybrid isolation (HI) model, to date the admixture in the Balti population. The Tibetan and Kashmiri populations were assigned as the proxies of ancestral sources. The unphased allelic data with MAF >0.05 were used. Following the recommendation by ALDER, we start the fit at genetic distance >0.5CM.

**Fitting models of admixture history**

We modeled various admixture graphs for the Balti population with qpAdm (Haak, et al. 2015) and qpGraph (Patterson, et al. 2012) based on the LD pruned WGS dataset of 455 individuals. We used qpAdm to evaluate a set of possible combinations of "left" and “right” populations with N=2 by using “rotation” and estimate admixture proportions as well as the associated standard errors. We fixed the "left" populations of “target-Balti” and “source1-Tibetan”. The “source2” was iteratively sampled from each of populations from Pakistan as well as other regions, disregarding admixed Hazara and Burusho (Lawson, et al. 2012). As for the “right” populations, the remaining populations were grouped as “outgroup”. The admixture proportions as well as the associated standard errors were also calculated with qpAdm. We adopted qpBrute (Leathlobhair, et al. 2018) to explore the space of all possible admixture graphs. A total of 55 graphs without f4 outliers (|Z| < 3) were recorded. We then used the MCMC algorithm implemented in the ADMIXTUREGRAPH R package to calculate the marginal likelihoods and Bayes Factors which were indicated by K values, for the 55 graphs (data were not shown). Among of them, we built the admixture graph for a combination of six populations (i.e. Balti, Tibetan, Kashmiri, Han, Gujarati, and Yoruba) with qpGraph.

**Local ancestry inference**

According to the results of GLOBETROTTER, we assigned the Kashmiri and Tibetan populations as ancestral sources to infer local ancestry in the Balti genomes based on the phased genomic data. PCAdmix v1.0 (Brisbin, et al. 2012) was used with default parameter settings. RFMix v1.5.4 (Maples, et al. 2013) was performed in two-way ancestry models by using a random forest algorithm based on the two steps of expectation maximization (EM = 2). The results of PCAdmix and RFMix were checked by the ancestry specific PCA with the program PCAmask (Moreno-Estrada, et al. 2013), respectively.We used length distribution of the local ancestral tracts under the general admixture model by using *AdmixInfer* v1.04 (Ni, et al. 2016) and multiple-wave admixtures model by using MultiWaver v2.0 (Ni, et al. 2019). AdmixInfer is designed to optimize the parameters of six admixture models of HI (Hybrid Isolation), GA (Gradual Admixture), CGFR (Continuous Gene Flow Recipient) and CGFD (Continuous Gene Flow Donor) via maximum likelihood estimation and to figure out the model best fitting the data (Ni, et al. 2016). MultiWaver, based on the general discrete admixture models, could scan the number of waves of admixture events, and estimate the parameters of admixture models with multi-waves, multi-ancestral populations (Ni, et al. 2018). MultiWaver v2.0 is the updated version to include several continuous admixture models and bootstrapping technique for calculating the confidence interval of genetic dating (Ni, et al. 2019).

**Analysis of X chromosomal data**

We followed the reported strategy (Ongaro, et al. 2019) to analyze X chromosomal data. We conducted joint SNP calling with by HaplotypeCaller module (DePristo, et al. 2011) in GATK v3.6 (McKenna, et al. 2010) for the X chromosomal SNPs of 41 Tibetan individuals (Lu, et al. 2016; Yang, et al. 2018), and then merged with newly sequenced and previously published genome data (Charati, et al. 2019; Mallick, et al. 2016; Genomes Project, et al. 2015). PLINK v1.9 was used to check the sex assignments with --check-sex by calculating X chromosome inbreeding coefficients. We disregarded the heterozygous SNPs in male X chromosomes. We filtered individuals and positions with the threshold of genotyping rate < 95%. A total of 197,966 X chromosomal SNPs of 455 individuals were retained. We performed LD pruning (--indep-pairwise 50 50 0.2) to result in a total of 29,796 X chromosomal. We ran unsupervised ADMIXTURE clustering for X chromosomal SNPs and autosomal datasets with K values = 2 to 10, and then compared the results with those based on LD pruned autosomal SNPs of the WGS dataset of 455 individuals. When K = 5, the Tibetan-specific component was characterized, whereas the rest was assigned as the Kashmiri-specific component. The paired Wilcoxon test with in-house Python script was calculated to test the differences between autosomal and X chromosomal proportions of the Tibetan and Kashmiri-specific components, respectively. To avoid potential bias in chromosomal length and number of markers, we repeated the analyses for chromosomes 7 and 18, respectively. The results were in agreement with those for autosome.

**MtDNA and Y chromosomal variation**

We extracted the mtDNA consensus sequences from the mtDNA SNP calling results with the cutoff value of heteroplasmy as 0.2 (Peng, et al. 2018). We assigned each mtDNA sequences into certain haplogroup with HaploGrep2 (Weissensteiner, et al. 2016) and the haplogrouping results were checked by MitoTool (Fan and Yao 2013). The mtDNA haplogroup nomenclature was referred to PhyloTree (http://phylotree.org/; Build 17) (Van Oven and Kayser 2009). We classified Y chromosomal haplogroup status for male individuals based on the Y-chromosomal SNP calling results by using yHaplo (Poznik 2016), AMY-tree (Van Geystelen, et al. 2013). We also perfromed the haplogrouping based on the Y-chromosomal BAM file with Yleaf (Ralf, et al. 2018). The Y chromosomal haplogroup nomenclature was referred to ISOGG Y-DNA Haplogroup Tree (version 14.164; https://www.isogg.org/tree/index.html).

**Reference**

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19(9):1655-1664.

Bai H, Guo X, Narisu N, Lan T, Wu Q, Xing Y, Zhang Y, Bond SR, Pei Z, Zhang Y, et al. 2018. Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. *Nat Genet.* 50(12):1696-1704.

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, et al. 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science* 367(6484).

Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, Reynolds A, Ostrer H, Mezey J, Bustamante CD. 2012. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum Biol.* 84(4):343.

Browning BL, Browning SR. 2011. A fast, powerful method for detecting identity by descent. *Am J Hum Genet.* 88(2):173-182.

Busby GB, Hellenthal G, Montinaro F, Tofanelli S, Bulayeva K, Rudan I, Zemunik T, Hayward C, Toncheva D, Karachanak-Yankova S, et al. 2015. The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape. *Curr Biol.* 25(19):2518-2526.

Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. 2018. Runs of homozygosity: windows into population history and trait architecture. *Nat Rev Genet.* 19(4):220-234.

Chacon-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuna-Alonzo V, Barquera R, Quinto-Sanchez M, Gomez-Valdes J, Martinez PE, Villamil-Ramirez H, et al. 2018. Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat Commun.* 9(1):1-13.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4(1):s13742-13015-10047-13748.

Charati H, Peng MS, Chen W, Yang XY, Jabbari Ori R, Aghajanpour-Mir M, Esmailizadeh A, Zhang YP. 2019. The evolutionary genetics of lactase persistence in seven ethnic groups across the Iranian plateau. *Hum Genomics.* 13(1):7.

Chaubey G, Ayub Q, Rai N, Prakash S, Mushrif-Tripathy V, Mezzavilla M, Pathak AK, Tamang R, Firasat S, Reidla M, et al. 2017. "Like sugar in milk": reconstructing the genetic history of the Parsi population. *Genome Biol.* 18(1):110.

Delaneau O, Zagury JF, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods.* 10(1):5-6.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43(5):491-498.

Di Cristofaro J, Pennarun E, Mazieres S, Myres NM, Lin AA, Temori SA, Metspalu M, Metspalu E, Witzel M, King RJ, et al. 2013. Afghan Hindu Kush: where Eurasian sub-continent gene flows converge. *PLoS One* 8(10):e76748.

Fan L, Yao YG. 2013. An update to MitoTool: using a new scoring system for faster mtDNA haplogroup determination. *Mitochondrion* 13(4):360-363.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson KJN. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522(7555):207-211.

Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. *Science* 343(6172):747-751.

International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311):52-58.

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet.* 8(1):e1002453.

Leathlobhair MN, Perri AR, Irving-Pease EK, Witt KE, Linderholm A, Haile J, Lebrasseur O, Ameen C, Blick J, Boyko AR, et al. 2018. The evolutionary history of dogs in the Americas. *Science* 361(6397):81-85.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30(20):2843-2851.

Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193:1233-1254.

Lu D, Lou H, Yuan K, Wang X, Wang Y, Zhang C, Lu Y, Yang X, Deng L, Zhou Y, et al. 2016. Ancestral origins and genetic history of Tibetan highlanders. *Am J Hum Genet.* 99(3):580-594.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* 538(7624):201-206.

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet.* 93(2):278-288.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*20(9):1297-1303.

Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, Nelis M, Magi R, Metspalu E, Remm M, et al. 2011. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet.* 89(6):731-744.

Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martínez RJ, Hedges DJ, Morris RW, et al. 2013. Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* 9(11):e1003925.

Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M. 2019. The formation of human populations in South and Central Asia. *Science* 365(6457):eaat7487.

Ni X, Yang X, Guo W, Yuan K, Zhou Y, Ma Z, Xu S. 2016. Length distribution of ancestral tracks under a general admixture model and Its applications in population history Inference. *Sci Rep.* 6:20048.

Ni X, Yuan K, Liu C, Feng Q, Tian L, Ma Z, Xu S. 2019. MultiWaver 2.0: modeling discrete and continuous gene flow to reconstruct complex population admixtures. *Eur J Hum Genet.* 27(1):133.

Ni X, Yuan K, Yang X, Feng Q, Guo W, Ma Z, Xu S. 2018. Inference of multiple-wave admixtures by length distribution of ancestral tracks. *Heredity* 121(1):52-63.

Ongaro L, Scliar MO, Flores R, Raveane A, Marnetto D, Sarno S, Gnecchi-Ruscone GA, Alarcon-Riquelme M, Patin E, Wangkumhang P. 2019. The genomic impact of European colonization of the Americas. *Curr Biol.* 29(23):3974-3986.

Pathak AK, Kadian A, Kushniarevich A, Montinaro F, Mondal M, Ongaro L, Singh M, Kumar P, Rai N, Parik J, et al. 2018. The genetic ancestry of modern Indus valley populations from northwest India. *Am J Hum Genet.* 103(6):918-929.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* 192(3):1065-1093.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2(12):e190.

Peng MS, Xu W, Song JJ, Chen X, Sulaiman X, Cai L, Liu HQ, Wu SF, Gao Y, Abdulloevich NT, et al. 2018. Mitochondrial genomes uncover the maternal history of the Pamir populations. *Eur J Hum Genet.* 26(1):124-136.

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8(11):e1002967.

Poznik GD. 2016. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. *bioRxiv*:088716.

Ralf A, Montiel Gonzalez D, Zhong K, Kayser M. 2018. Yleaf: software for human Y-chromosomal haplogroup inference from next-generation sequencing data. *Mol Biol Evol.* 35(5):1291-1294.

Tätte K, Pagani L, Pathak AK, Kõks S, Duy BH, Ho XD, Sultana GNN, Sharif MI, Asaduzzaman M, Behar DM. 2019. The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. *Sci Rep.* 9(1):3818.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526(7571):68-74.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy‐Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J. 2013. From FastQ data to high‐confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinform.* 43(1):11.10. 11-11.10. 33.

Van Geystelen A, Decorte R, Larmuseau MH. 2013. AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. *BMC Genomics* 14(1):101.

Van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat.* 30(2):E386-394.

Weissensteiner H, Pacher D, Kloss-Brandstatter A, Forer L, Specht G, Bandelt HJ, Kronenberg F, Salas A, Schonherr S. 2016. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* 44(W1):W58-63.

Yang XY, Dai SS, Liu HQ, Peng MS, Zhang YP. 2018. The uncertainty of population relationship and divergence time inferred by the multiple sequentially Markovian coalescent model. *J Hum Genet.* 63(6):775-777.