

Supplementary Materials and Methods

<https://doi.org/10.1631/jzus.B2400576>

1. The specific procedures for acquiring whole-genome sequencing (WGS) data were conducted following standardized genomic protocols.

Following the standard procedure of the Illumina sequencer and Novogene's (Beijing) standard library preparation, we performed whole-genome sequencing on the 100 samples at $5\times$ depth using the Illumina Novaseq 6000. The Burrows-Wheeler algorithm (Alexander et al., 2009) was used to align the sequences to the human genome g1k_v37. SNP and Indel detection for individual samples was done using bcftools and sentieon (DePristo et al., 2011, McKenna et al., 2010). Single-sample imputation was subsequently conducted using BGI-lowpass-v0.4 (Beijing Genomics Institution, Shenzhen, China) with the REFPANEL_b37_KG reference panel, applying the -map parameter to integrate recombination rate information, applying the -region parameter to delimit targeted genomic intervals, and implementing chromosome splitting with parallel processing to significantly enhance computational efficiency. Finally, we merged the *.vcf files of all samples using bcftools v1.10.2, resulting in a total of 79,556,172 SNPs.

2. f -statistics

The outgroup- f_3 statistics ($f_3(X, Y; \text{outgroup})$) analyzed shared genetic drift between populations X and Y, indicating genetic affinity based on the amount of shared genetic drift. A larger f_3 value suggests more gene flow between X and Y after their separation from the outgroup, indicating closer genetic affinity. The admixture- f_3 statistics ($f_3(X, Y; Z)$) analyzed admixture events in population Z, with significant negative values ($Z \leq -3$), indicating that Z may have formed through admixture between populations related to X and Y.

The f_4 statistics ($f_4(A, B; C, D)$) used A, B and C as study and reference populations, with D as the outgroup (Mbuti in this study). Significant positive values ($Z > 3$) indicated more allele sharing between populations A and C, while significant negative values ($Z < -3$) indicated more allele sharing between populations B and C.

References

- Alexander DH, Novembre J, Lange K, 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19(9):1655-1664.
<https://doi.org/10.1101/gr.094052.109>
- DePristo MA, Banks E, Poplin R, et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491-498.
<https://doi.org/10.1038/ng.806>
- McKenna A, Hanna M, Banks E, et al., 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297-1303.
<https://doi.org/10.1101/gr.107524.110>