Supplementary Materials for

Modern human origins: multiregional evolution of autosomes and East Asia origin of Y and mtDNA

## Dejian Yuan, Xiaoyun Lei, Yuanyuan Gui, Zuobin Zhu, Dapeng Wang, Jun Yu, and Shi Huang\*

## \*Corresponding author. E-mail: [huangshi@sklmg.edu.cn](mailto:huangshi@sklmg.edu.cn)

## This PDF file includes:

## Supplementary text

## Tables S1 to S10

## Figures S1 to S14

## References

SUPPLEMENTARY TEXT:

TABLE OF CONTENTS

S1. Genetic diversity patterns in different types of SNPs in different human groups in the 1000 genomes project …………………………………………..………………………………… 3

S2. Conserved slow evolving sequences still have neutral variations…………………………3

S3. Known positively selected genes are fast evolving………………………………………....4

S5. Genetic distance between Neanderthal Y chromosome and modern Y haplotypes……………………………………………………………………………………………4

S5. Admixture analysis……………………………………………………………………………...4

S6. References………………………………………………………………………………………5

Supplementary Tables……………………………….…….........................................................7

Supplementary Figures……………………………….…….......................................................9

Supplementary Information 1

Genetic diversity patterns in different types of SNPs in different human groups

In addition to SNP types examined in detail in the main text, we also examined the pairwise genetic distance (PGD) pattern of the splicing SNPs that are also expected to be functional (Supplementary Figure S1 A-B). To examine whether the cut-off point in protein conservation in selecting the slow evolving proteins was appropriate, we verified that using nonsyn SNPs located within the next set of just slightly less conserved proteins (346 autosomal proteins with 800-1102 aa in length and with identity between human and monkey >99% but <100%) produced PGD patterns similar to the functional stop codon or splicing SNPs (Supplementary Figure S1 C-D). We further showed that synonymous SNPs within the slow set of proteins gave patterns similar to the stop codon or splicing SNPs but unlike the nonsyn SNPs within the same set of slow evolving proteins (Supplementary Figure S1 E-F).

Supplementary Information 2

Conserved slow evolving sequences still have neutral variations

As we aim to use slow evolving or highly conserved genes to do phylogenetic analysis, it is important to show that they still have neutral sequences that are not under natural selection. We determined whether they have fewer overlap sites by examining, of those different positions between human and monkey, how many are also different between any pair of the three species human, monkey and mouse ([Huang, 2010](#_ENREF_3)). As discussed before, such overlap sites would be rare if mutations occur in a neutral fashion following the infinite sites model of the Neutral theory rather than cluster at either positively selected sites or sites left untouched/neutral by negative selection. We examined the informative genes in the set of slow evolving proteins as described in Table S1 that have orthologs in the species concerned and show less than 5 bp gaps in any pair of alignments. Of 31 informative proteins, the average overlap ratio is 9.6% (the number of overlap positions divided by the number of amino acid differences between human and monkey). In contrast, a randomly chosen set of 63 fast evolving genes (~93% identity between human and monkey), the average overlap ratio is 26.1% (2.73 fold greater than that of slow evolving genes, P < 0.01). These results confirm that highly conserved proteins have fewer overlap sites, and variants in them do follow the Neutral theory ([Huang, 2012](#_ENREF_4)).

Supplementary Information 3

Known positively selected genes are fast evolving

We next looked at whether known positively selected genes are more likely to be fast evolving. We did not find any of the slow genes present in the set of 56 positively selected genes identified by Nielson et al ([Nielsen et al., 2005](#_ENREF_6)) but found two (WDR7 and KIAA1429) in the set of 154 human genes identified by Bakewell et al, which does overlap with the Nielson et al set ([Bakewell et al., 2007](#_ENREF_2)). Among the list of ~14000 genes with informative identity scores between human and monkey, 24 and 87 were found in the 56 and 154 sets, respectively. These 111 genes have 88.6% identity between human and monkey with average length 601 aa. The genome average of ~14000 genes are 87.8% and 560 aa. Thus, positively selected genes have mutation rates similar to the genome average, which is much faster than the slow evolving set of genes identified here.

Supplementary Information 4

Genetic distance between Neanderthal Y chromosome and modern Y haplotypes.

We merged the Y chromosome SNPs of the 1000 genomes project with the Y chr sequence of a ~49,000-year-old Neanderthal from El Sidron of Spain and obtained informative genotype data for a total of 15 SNPs ([Mendez et al., 2016](#_ENREF_5)). We then calculated the average distance to the Neanderthal for each major haplogroup and the results are presented in Supplementary Figure S3.

Supplementary Information 5

Admixture analysis

ADMIXTURE implements a model-based estimation of ancestry in unrelated individuals, which may detect relatively recent admixture events in a population ([Alexander et al., 2009](#_ENREF_1)). Admixture v1.3.0 was used to determine the ancestral population components. Input data were prepared using the same procedure as for the PCA. To explore ancestral populations, we used K=2–12. The cross validation (CV) method, implemented in ADMIXTURE, was used to estimate the best k value. All parameters were set to default. Despite the low number of SNPs in our dataset, the data was pruned for LD as ADMIXTURE generally assumes unlinked loci. PLINK was used to calculate an LD (r2) score for each pair of SNPs in a window of 200 SNPs, and one SNP from the pair was excluded if r2 > 0.4. This LD pruning removed 125 SNPs from a total of 15435 SNPs and the LD pruned dataset was used for ADMIXTURE analysis.

As any software must have assumptions which may or may not realistic, we tested whether the ADMIXTURE software can produce consistent results. We studied which group in the 1kGP is most related to the Mbuti group here consisting of 4 genomes from SGDP. As shown in Supplementary Figure S13, Mbuti was found related to Africans at K <7, and most related to LWK at K=7 or 8. But unexpectedly, at K=10 or 12, Mbuti was found most related to YRI. We also tested the affinity of the ancient Ust’-Ishim to specific groups in the 1kGP but again failed to obtain consistent results (Supplementary Figure S14). Depending on the K values selected, Ust’-Ishim could be mostly related to non-African (K =2), CHS (K=3 or 8), CEU (K=4 or 6), GIH (K=5, 9, or 11), or LWK (K=7).

Therefore, the ADMIXTURE software failed to produce consistent results. Given such, it may be premature to use the software and we have instead used other methods with few uncertain assumptions as described in the main text.

Supplementary Information 6

References:

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

Bakewell, M.A., Shi, P., and Zhang, J. (2007). More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci U S A *104*, 7489-7494.

Huang, S. (2010). The overlap feature of the genetic equidistance result, a fundamental biological phenomenon overlooked for nearly half of a century. Biological Theory *5*, 40-52.

Huang, S. (2012). Primate phylogeny: molecular evidence for a pongid clade excluding humans and a prosimian clade containing tarsiers. Sci China Life Sci *55*, 709-725.

Mendez, F.L., Poznik, G.D., Castellano, S., and Bustamante, C.D. (2016). The Divergence of Neandertal and Modern Human Y Chromosomes. Am J Hum Genet *98*, 728-734.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J.*, et al.* (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol *3*, e170.

Poznik, G.D., Henn, B.M., Yee, M.C., Sliwerska, E., Euskirchen, G.M., Lin, A.A., Snyder, M., Quintana-Murci, L., Kidd, J.M., Underhill, P.A.*, et al.* (2013). Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science *341*, 562-565.

Pugach, I., Delfin, F., Gunnarsdottir, E., Kayser, M., and Stoneking, M. (2013). Genome-wide data substantiate Holocene gene flow from India to Australia. Proc Natl Acad Sci U S A *110*, 1803-1808.

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T.*, et al.* (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. Science *334*, 94-98.

Supplementary Tables:

## Table S1. A randomly selected set of 255K SNPs from 1kGP. See separate file 1.

## Table S2. Slow SNPs set: nonsyn SNPs in slow evolving proteins from 1kGP, including 423 genes > 304 amino acid in length with 100% identity and 178 genes > 1102 amino acid in length with 99% identity between monkey and human.

## See separate file 2.

## Table S3. Syn and Nonsyn SNPs in slow evolving proteins from 1kGP.

## See separate file 3.

## Table S4. Nonsyn SNPs in 346 autosomal proteins with 800-1102 aa in length with identity between human and monkey >99% but <100%.

## See separate file 4.

## Table S5. Overlap ratio in slow and fast evolving proteins.

## See separate file 5.

## Table S6. 34 genes used for mutation rate calculation.

## See separate file 6.

## Table S7. Mutations inconsistent with the existing Y phylogeny as reported in Poznik et al. (2013) ([Poznik et al., 2013](#_ENREF_7)).

## .

## Table S8. Y and mtDNA haplotypes of the 1000 genomes samples.

## See separate file 7.

## Table S9. SNPs defining major haplogroups among 58251 cleanly called SNPs and their genotypes in the 1000 genomes.

## See separate file 8.

## Table S10. Slow mtDNA SNPs in 1kGP and four archaic humans.

## See separate file 9.

## Supplementary Figures:

## 

## Figure S1. Genetic diversity patterns as measured by different types of SNPs. Average pairwise genetic distances (PGD), either hom or total, within each of the five groups in the 1000 genomes are shown for splicing SNPs (A and B), nonsyn SNPs in 346 autosomal proteins (800-1102 aa in length with identity between human and monkey >99% but <100%) (C and D), and syn SNPs in slow evolving proteins (including 423 genes > 304 amino acid in length with 100% identity and 178 genes > 1102 amino acid in length with 99% identity between monkey and human) (E and F). Known heavily admixed groups such as ASW and ACB in the African group or CLM and PUR in the American group were excluded in the analysis. Data are means with standard deviation.

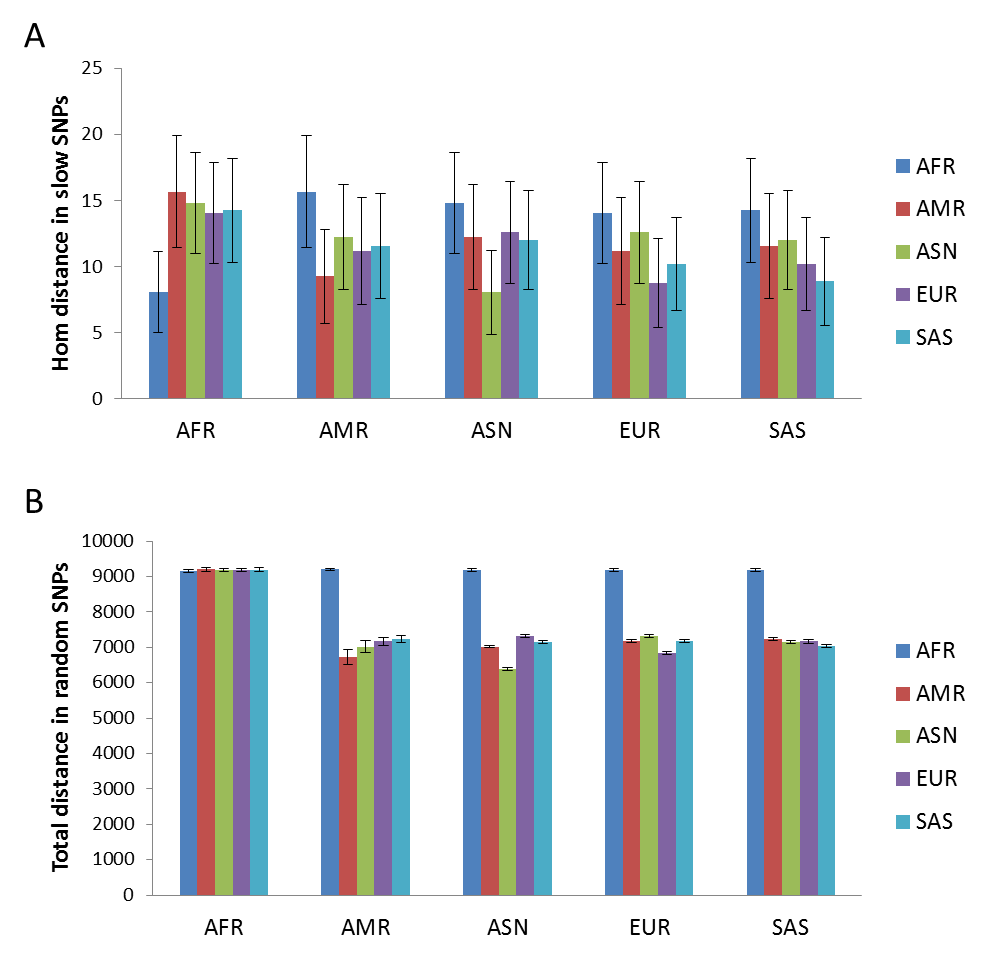


Figure S2. Genetic distance between human groups as measured by different types of SNPs. A. Genetic distance (Hom) measured by slow SNPs. B. Genetic distance (Total) measured by fast SNPs (255K randomly selected SNPs). Data are means with standard deviation.

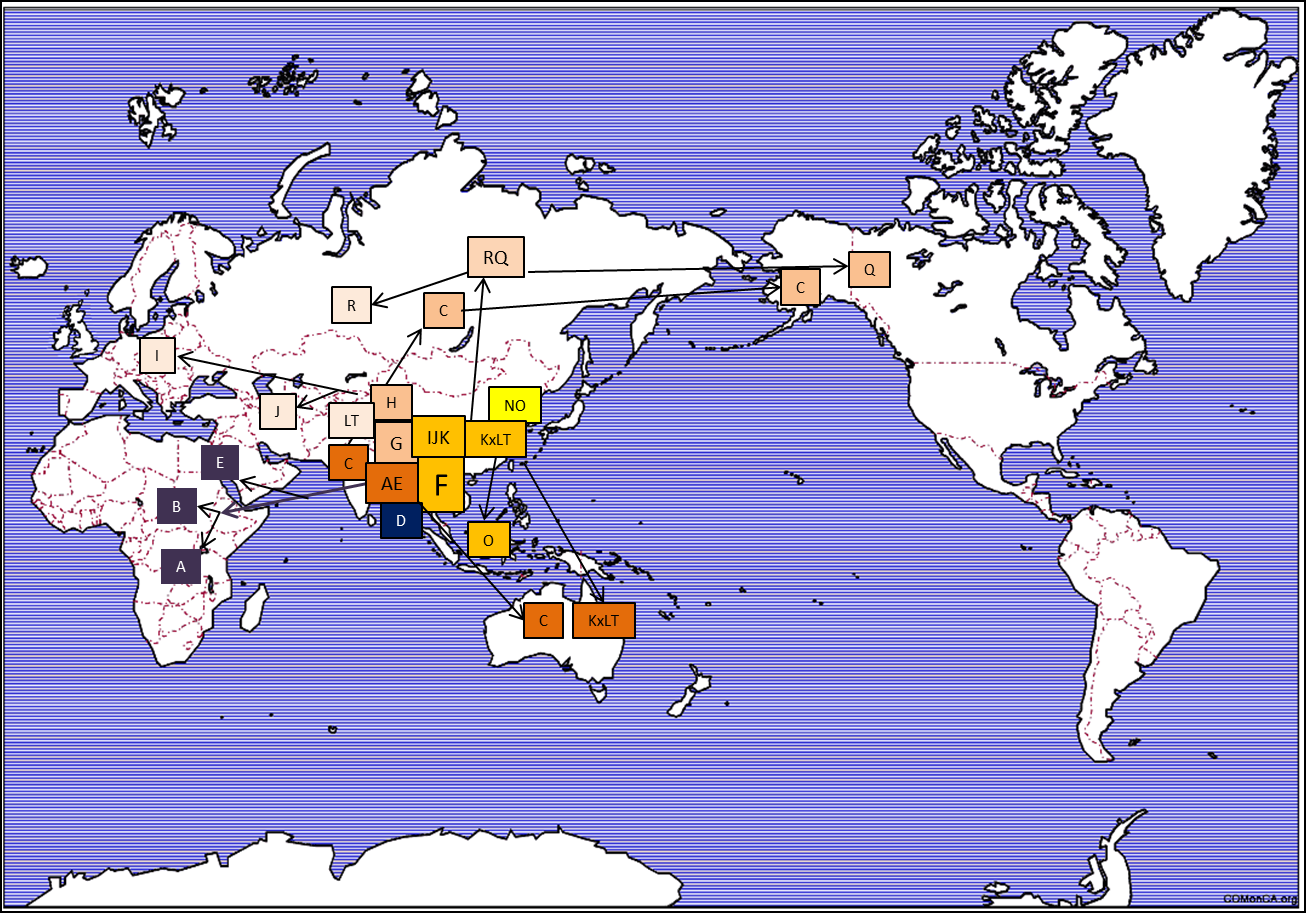


Figure S3. Y chromosome migration map. The location and direction of migration of major Y haplotypes are shown on a world map.

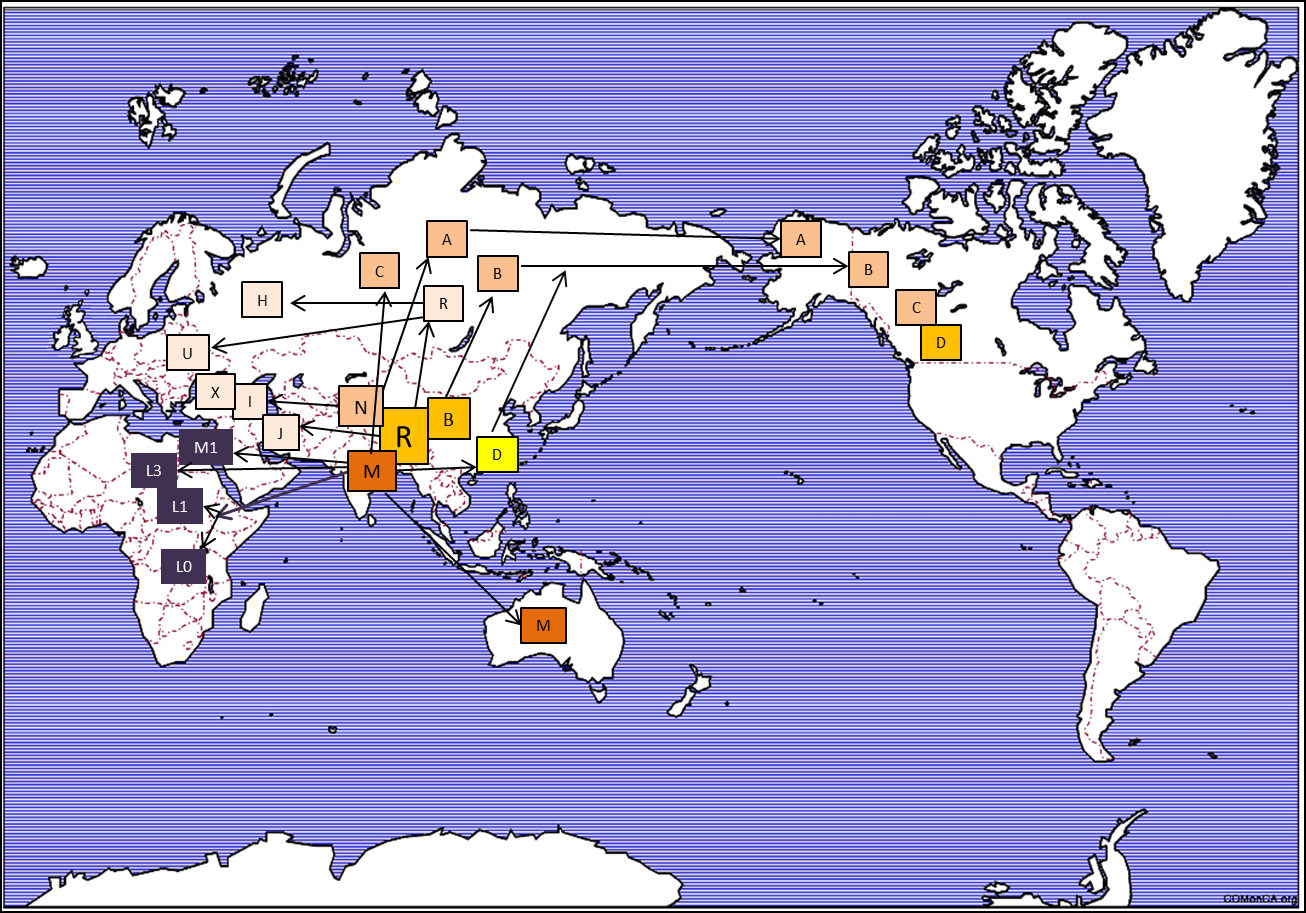


Figure S4. mtDNA migration map. The location and direction of migration of major mtDNA haplotypes are shown on a world map.

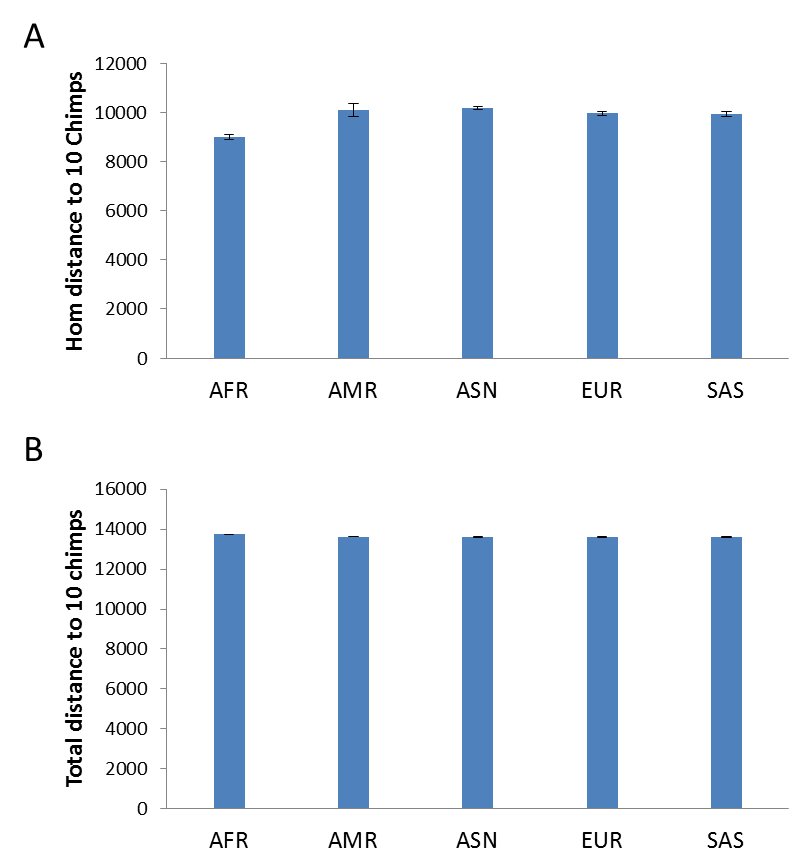


Figure S5. Genetic distance between chimpanzees and different human groups in fast SNPs. Shown is the hom (A) or total distance (B) between 10 chimpanzees and the 5 human groups of the 1000 genomes as measured by the randomly selected set of 255K SNPs as defined in the main text. Data are means with standard deviation.

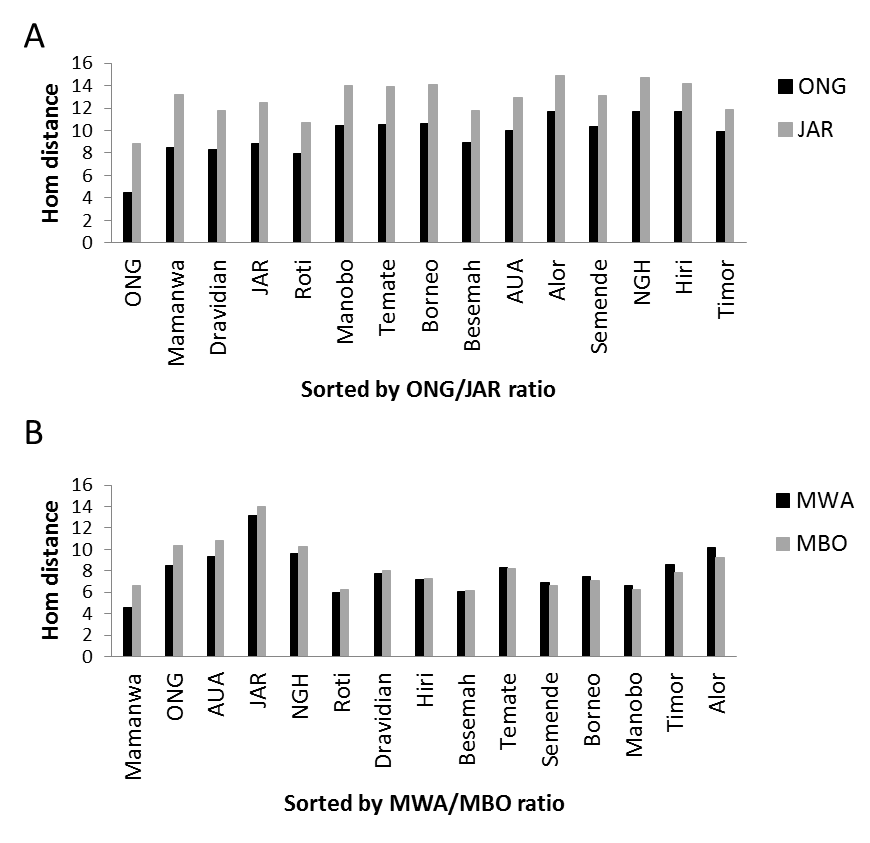


Figure S6. Relationship between different Negrito groups in South Asia. A. Autosomal distances to ONG or JAR are shown for each of the Indian and South Asian island groups as sampled by Pugach et al (2013) ([Pugach et al., 2013](#_ENREF_8)). The groups were ordered based ONG/JAR distance ratio from small to large. B. Autosomal distances to Mamanwa (MWA) or Manobo (MBO) are shown for ONG, JAR, and each of the Indian and South Asian island groups as sampled by Pugach et al (2013)([Pugach et al., 2013](#_ENREF_8)). The groups were ordered from small to large based MWA/MBO distance ratio.

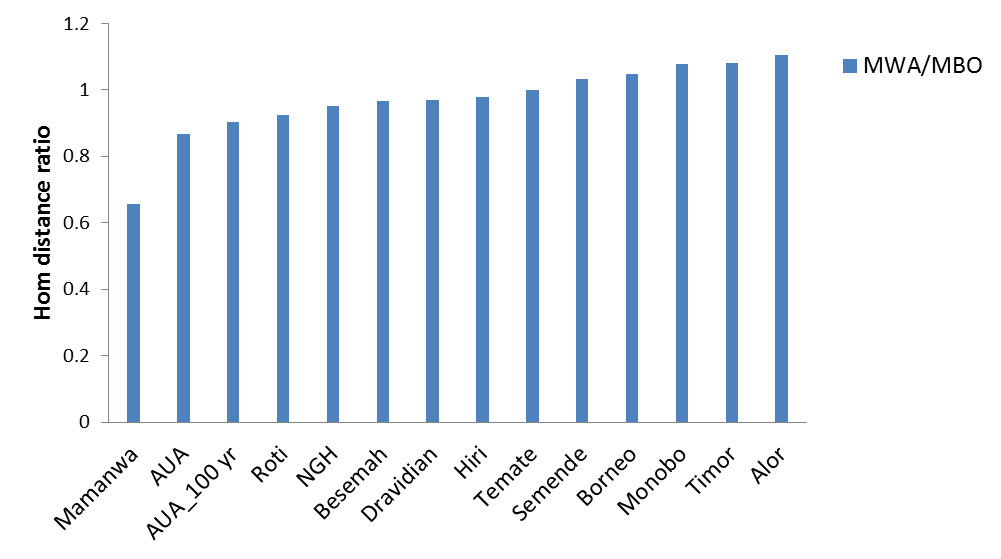


Figure S7. Relationship between Aboriginal Australians (AUA) and Negritos. Distance to Mamanwa (MWA) or Manobo (MBO) are shown for a 100 year old AUA genome (AUA\_100 yr) as described by Rasmussen et al. (2011) ([Rasmussen et al., 2011](#_ENREF_9)) and each of the Indian and South Asian island groups that also include AUA as sampled by Pugach et al (2013) ([Pugach et al., 2013](#_ENREF_8)). The groups were ordered from small to large based MWA/MBO distance ratio.

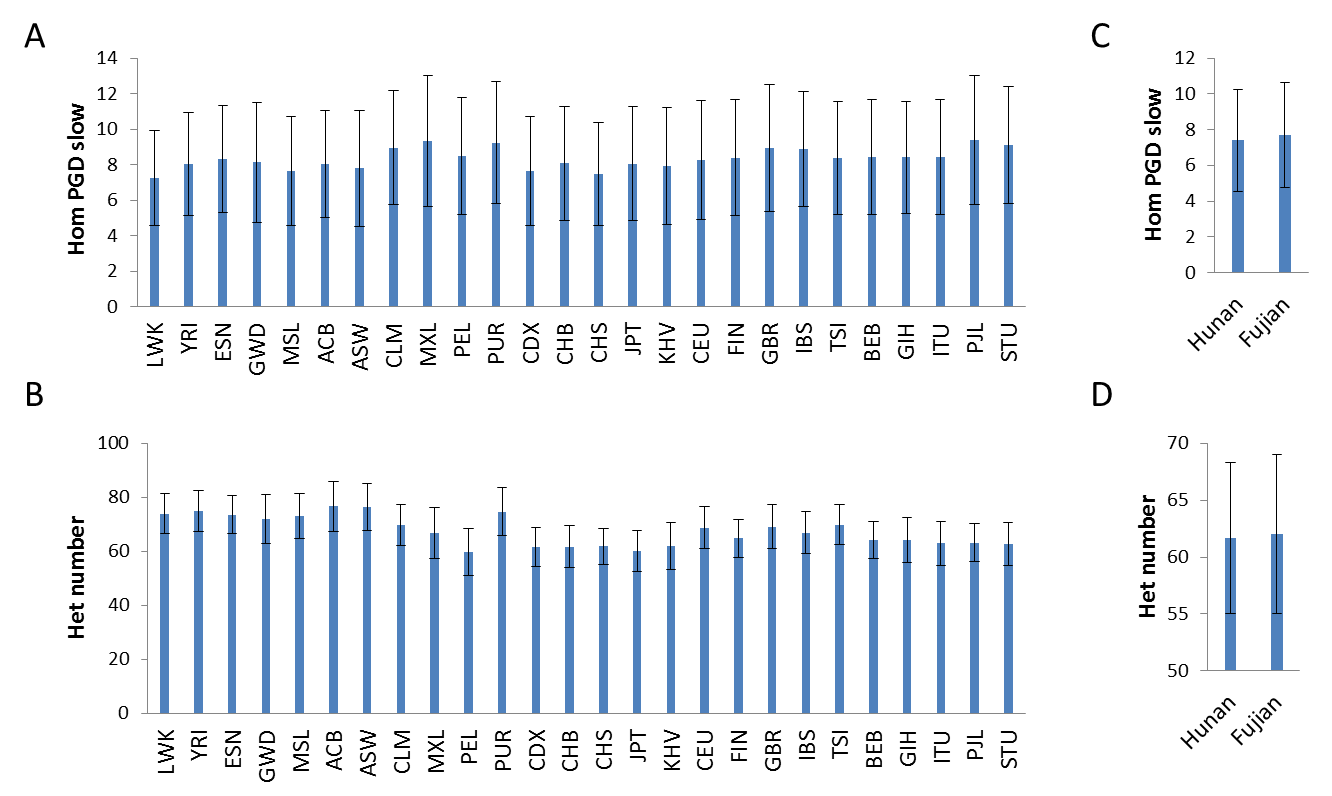


Figure S8. Genetic diversity levels of different human groups in slow SNPs. A. Average pairwise genetic distance (Hom) in slow SNPs within each of the 25 groups of the 1000 genomes project. B. Average number of heterozygous sites in slow SNPs for each of the 25 groups. C. Average pairwise genetic distance (Hom) in slow SNPs within Hunan and Fujian group. D. Average number of heterozygous sites in slow SNPs for Hunan and Fujian group. Error bars indicate standard deviations. Data are means with standard deviation.

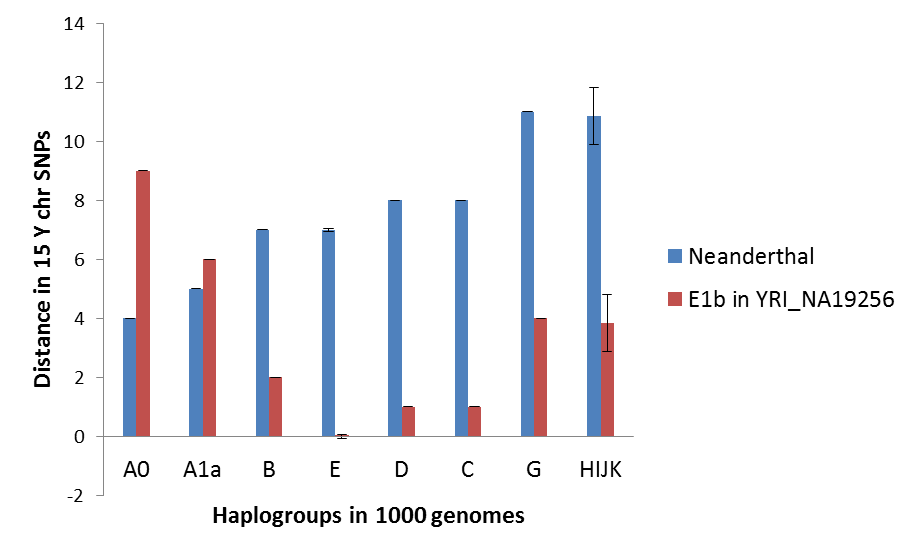


Figure S9. Neanderthal Y chromosome distance to major haplotypes in 1000 genomes. Shown are the genetic distances in Y chr between haplotypes in the 1000 genomes and the Neanderthal from El Sidron of Spain or a randomly picked YRI individual with E1b haplotype. Data are means with standard deviation.

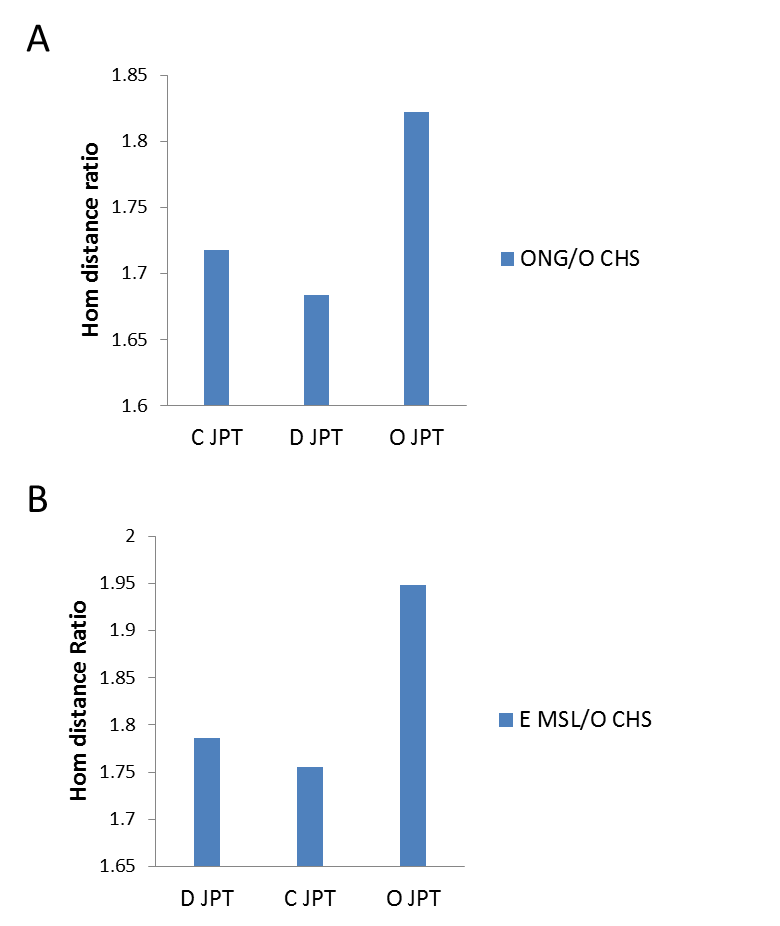


Figure S10. Autosomal genetic distance in slow SNPs between Japanese and groups with African type Y haplotypes. A. Ratio of autosomal distance to Onge versus CHS with O haplotype for JPT individuals with C, D, or O haplotype. B. Ratio of autosomal distance to MSL with E haplotype versus CHS for Japanese individuals with C, D, or O haplotype.

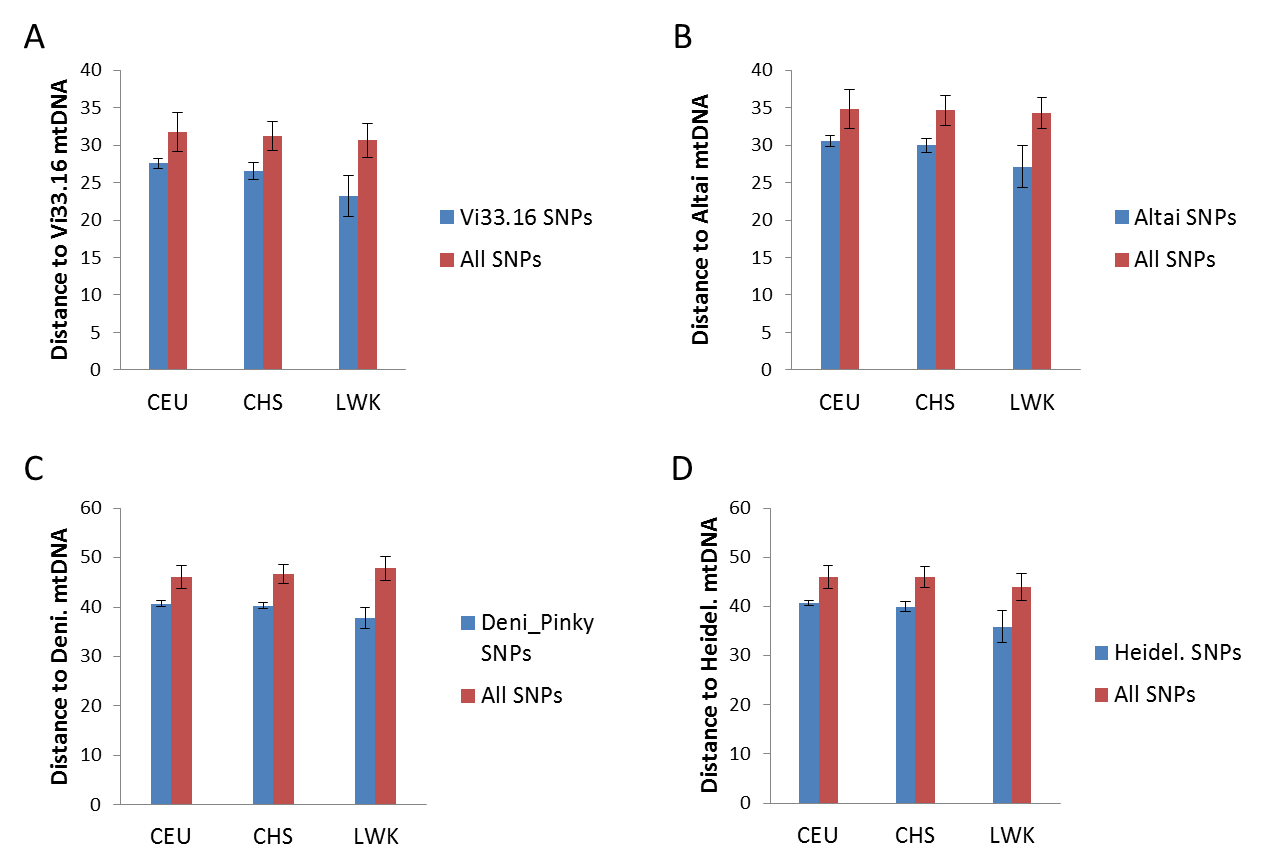


Figure S11. Distance between archaic and modern mtDNA in slow mtDNA SNPs. Shown are the average genetic distances as measured by using either all slow mtDNA SNPs or slow SNPs found in archaic human Vi33.16 (A), Altai (B), Denisovan (C), and Heidelbergensis (D). Data are means with standard deviation.

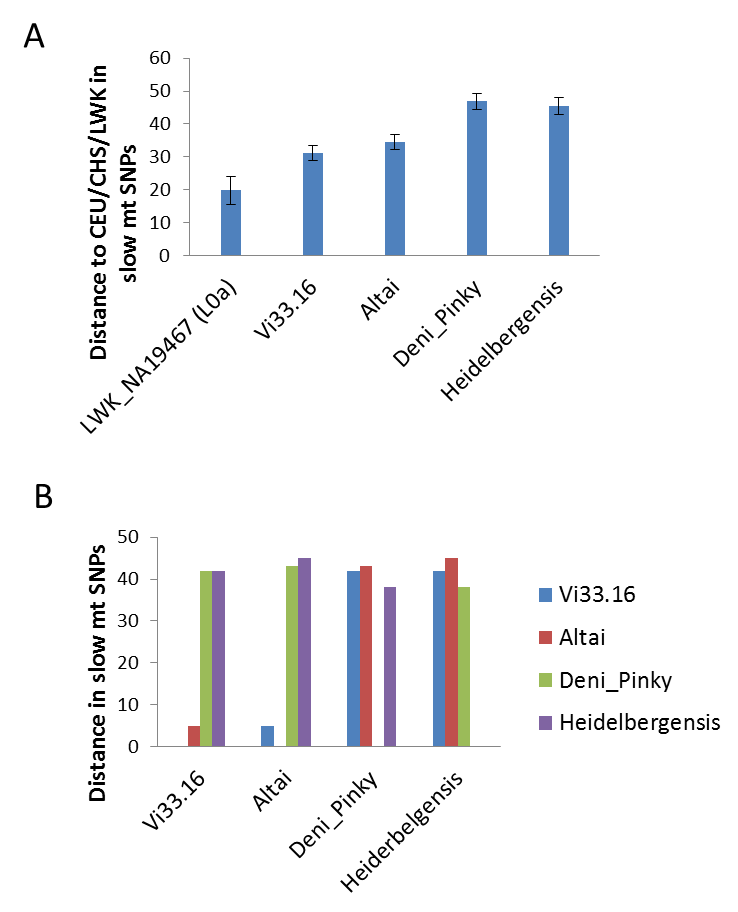
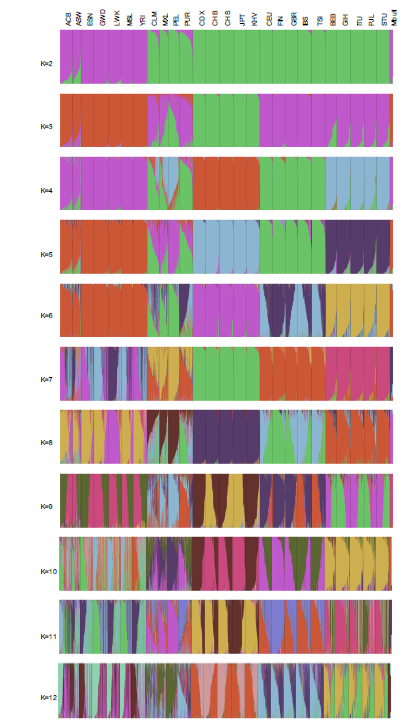


Figure S12. Archaic mtDNAs as outgroup to modern mtDNAs. A. Genetic distance in slow mtDNA SNPs between modern humans (including LWK, CHS, and CEU) and a randomly selected modern human with L0a haplotype (LWK\_NA19467), or four archaic mtDNAs as shown. B. Genetic distance in slow mtDNA SNPs among four archaic mtDNAs. Data are means with standard deviation.

A



B

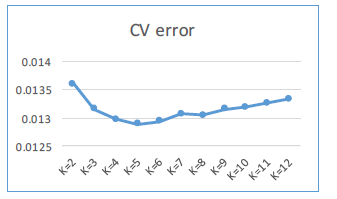
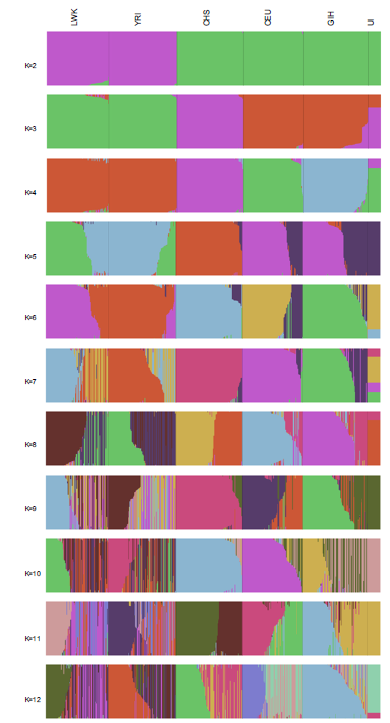


Figure S13. ADMIXTURE plots for K=2 to 12 on Mbuti and 1kGP. ADMIXTURE analysis was done using LD pruned slow SNPs data on Mbuti (4 individuals) and the 1kGP. Plots for K=2 to 12 are shown in A, and values of cross validation (CV) are shown in B.

A



B

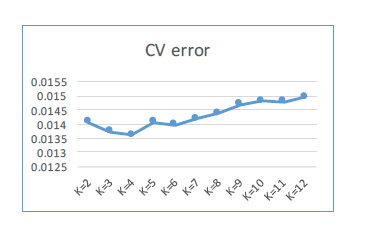


Figure S14. ADMIXTURE plots for K=2 to 12 on Ust’-Ishim and selected groups from 1kGP. ADMIXTURE analysis was done using LD pruned slow SNPs data on Ust’-Ishim (UI) and 5 selected groups from the 1kGP. Plots for K=2 to 12 are shown in A, and values of cross validation (CV) are shown in B.