Supplementary figures

Figure S1. Total length of individual specific sequence under different sequence identity threshold. Both the assembled YH (Asian) and NA18507 (African) genomes were aligned against the NCBI human reference genome using Blast to identify individual specific sequence. Unaligned sequences with length less than 100bp were filtered. Different identity thresholds were tried. The total length of individual specific sequences was almost unchanged when setting identity less than 90%, so 90% identity was chosen as the threshold.

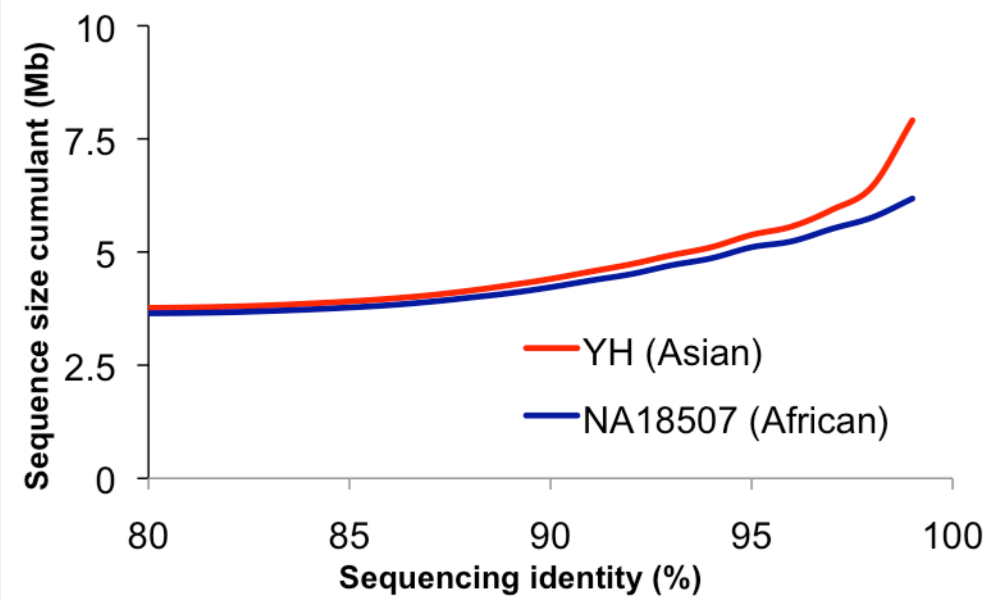


Figure S2. Proportion of novel sequences identified from genome assemblies by subset of full raw data relative to all novel sequences in YH genome (56-fold coverage in total).

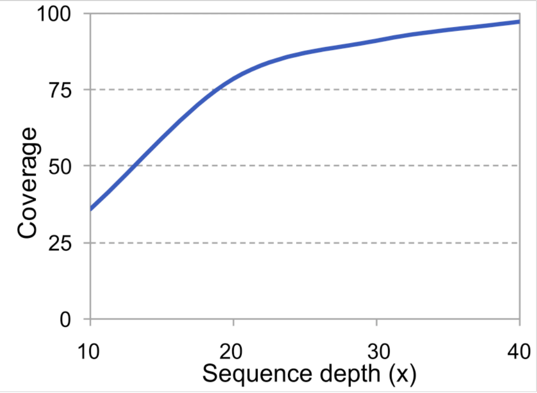


Figure S3. Length distribution of individual specific sequences.

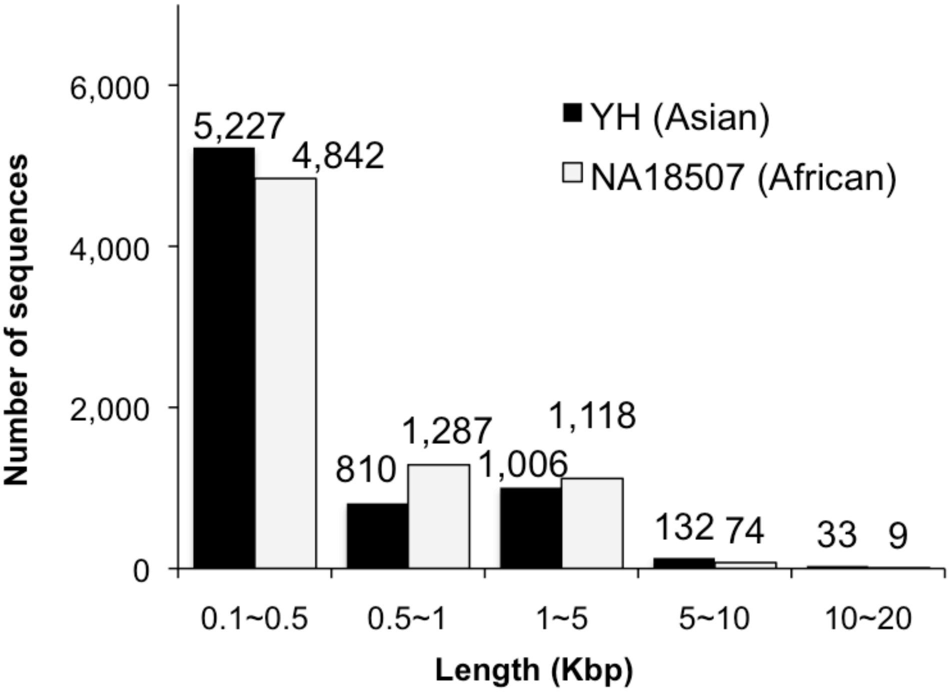


Figure S4. Estimated difference in individual specific sequences between HGDP-CEPH panel sample individuals and NA18507 genome. The estimated difference (*y*-axis) was extrapolated from the 120kb sampling novel sequences identified in NA18507 that were put to PCR validation. The proportion of difference in genotyping sites were extracted from result in previous studies.

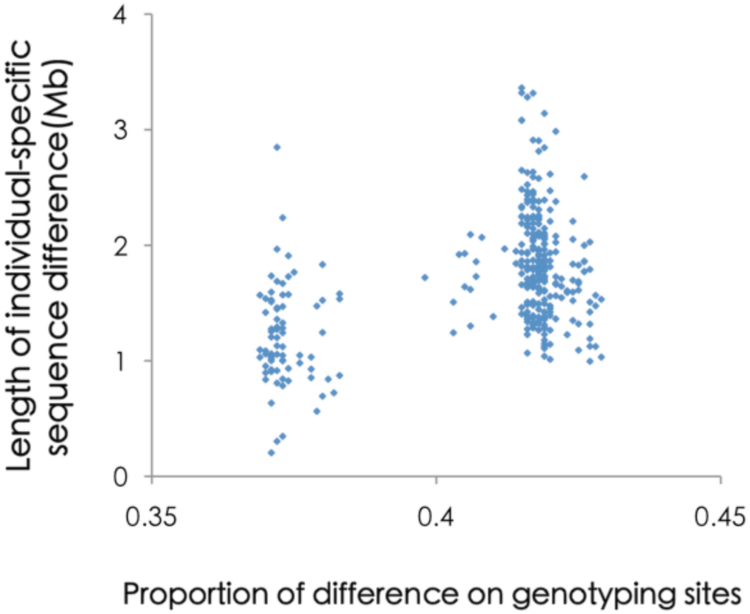


Figure S5. Genome DNA sequence composition difference at different length threshold to claim individual-specific sequences. To be consistent with the threshold we used to identify novel sequences compared with NCBI reference genome, we used the cutoff of 100bp in this study.

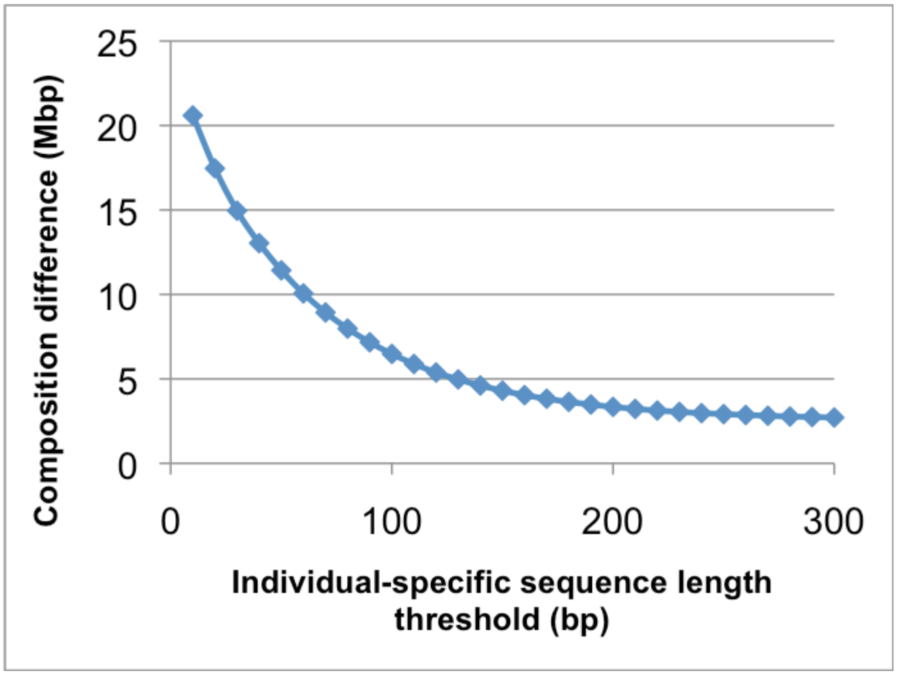


Figure S6. Number of alignment hits in novel sequences of each protein family. Refseq proteins were collected and aligned against the novel sequences using tBlastN (1E-5). Only the best hit was remained on each novel sequence location.

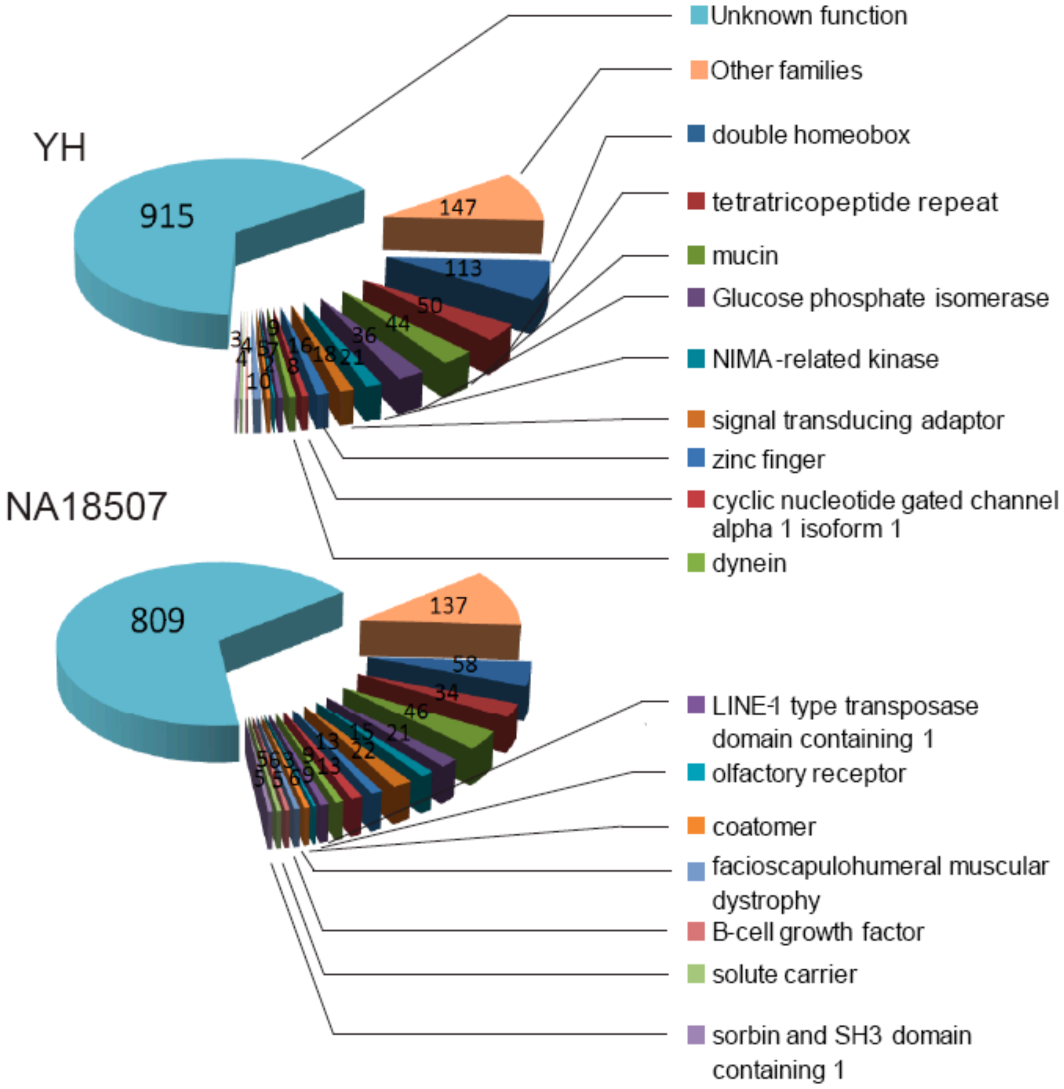
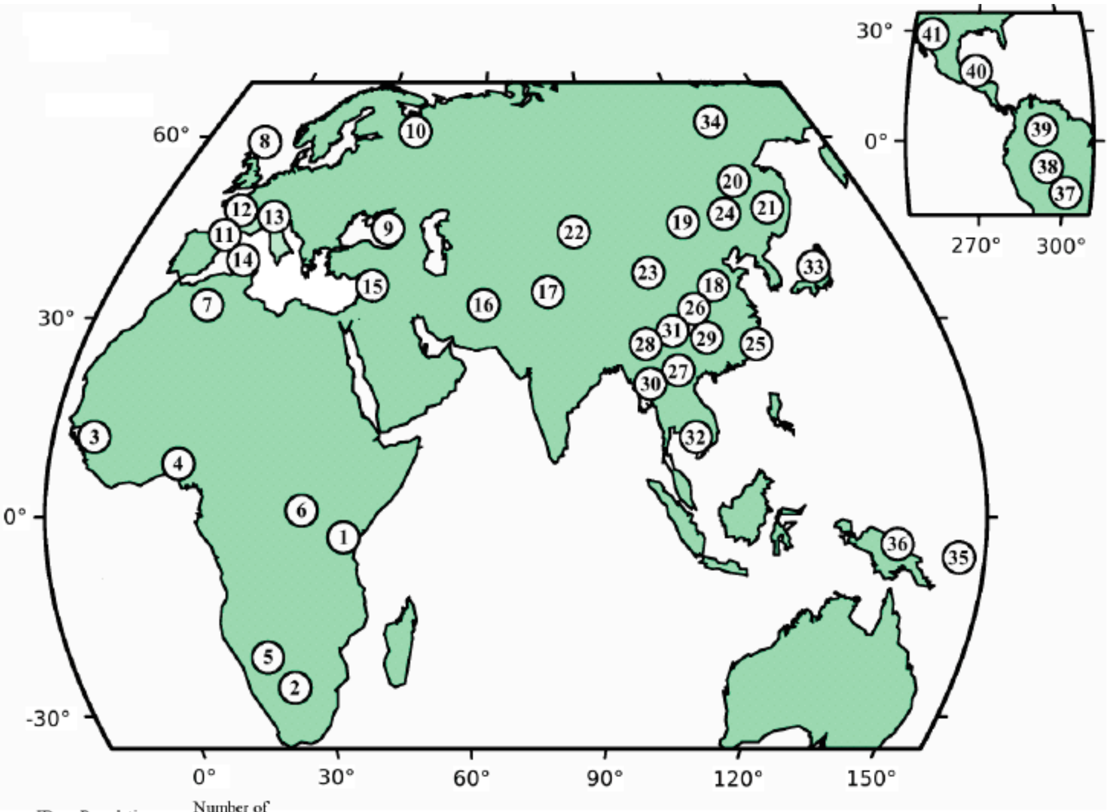


Figure S7. Geographic locations of the studied populations. The DNA samples were provided by CEPH-HGDP (Center for the study of Human Polymorphism-Human Genome Diversity Panel). Number of samples from each population were shown. The populations with only a few samples were merged according to their geographic locations in this analysis. The (12) North Italian and (13) Sardinian were merged as Italian; the (18) Mongolia, (19) Daur, 20) Hezhen, (21) Xibo, (22) Tu and (23) Oroqen were merged as CHN (Northern China) minorities; the (24) She, (25) Tujia, (26) Miaozu, (27) Lahu, (28) Yizu, (29) Dai and (30) Naxi were merged as CHS (Southern China) minorities.



ID	Population	Number of Samples
Africa		
1.	Bantu N.	11
2.	Bantu S.	8
3.	Mandenka	21
4.	Yoruba	23
5.	San	5
6.	Mbuti Pygmy	10
7.	Mozabite	10
Europe		
8.	Orcadian	6
9.	Adygei	15
10.	Russian	15
11.	Basque	15
12.	French	15
Italian		
13.	North Italian	3
14.	Sardinian	15
Middle East		
15.	Druze	5
South Asia		
16.	Balochi	15
17.	Pathan	17
East Asia		
18.	Han	20
CHN minorities		
19.	Mongola	4
20.	Oroqen	2
21.	Hezhen	3
22.	Xibo	4
23.	Tu	2
24.	Daur	1
CHS minorities		
25.	She	3
26.	Tujia	1
27.	Miaozu	4
28.	Lahu	2
29.	Yizu	2
30.	Dai	2
31.	Naxi	3
32.	Cambodian	10
33.	Japanese	10
34.	Yakut	6
Oceania		
35.	Melanesian	10
36.	Papuan	10
America		
37.	Surui	8
38.	Karitiana	10
39.	Colombian	7
40.	Maya	10
41.	Pima	10

Supplementary tables

Table S1. Statistics of newly generated YH (Asian) sequencing reads. The reads were sequenced by Illumina Genome Analyzer (GA) technology.

Library insert size (bp)	**Read legnth**	**Data produced (Gb)**
440	35	2.6
	44	20.5
	75	22.1
2,600	35	1.3
	44	12.3
	75	4.5
6,000	44	12.3
9,600	44	6.9
Total		**82.5**

Table S2. Statistics of anchoring novel sequences on the chromosomes of NCBI reference genome. Novel sequences were anchored according to the alignment of flanking sequence at both ends of chromosomes in the NCBI human reference genome (Build 36.3) (minimal 2 Kb in length and over 90% identity). "Scaffold vs NCBI Ref" refers to anchoring flanking sequences of assembled scaffolds (YH or NA18507); "Venter vs NCBI Ref" refers to anchoring flanking sequences on the Venter genome; "GenBank human clones vs NCBI" refers to anchoring flanking sequences on human clones deposited in GenBank. We categorized the novel sequences as insertions, highly divergent sequences, complex structural variant regions, flanking gaps, and unanchored without sufficient flanking sequences available. Relatively small differences in the size of novel sequences found for the Venter genome and GenBank human clones reflect small differences in YH and NA18507 anchoring sequences.

	Insertion (bp)	High Divergence (bp)	SV (bp)	Flanking gaps in NCBI reference (bp)	Unanchored (bp)
YH Scaffold vs NCBI Ref	834,627	42,324	1,275,015	383,009	2,805,269
Venter vs NCBI Ref	952,891	17,702	1,148,055	511,066	1,187,656
GenBank human clones vs NCBI Ref	459,489	25,269	971,935	207,983	1,032,660
NA18507 Scaffold vs NCBI Ref	514,995	19,127	1,173,131	356,496	3,082,917
Venter vs NCBI Ref	951,373	13,756	1,276,629	589,936	944,743
GenBank human clones vs NCBI Ref	431,173	26,669	936,226	189,819	1,145,836

Table S3. Primer sequences of individual specific sequences for PCR validation and profiling.

(Attached Excel file)

Table S4. S4_A and S4_B are PCR profiling result of individual specific sequences in each DNA sample. S4_C is the info of DNA samples. Each row is one analyzed sequence, and each column is one sample one a 96-well plate. If a sequence exist in a DNA sample, the matrix cell was filled by ‘1’; on the contrary, it was filled by ‘0’.

(Attached Excel file)

Table S5. Frequency of PCR profiled sequences in the African, European and Asian populations.

(Attached Excel file)

Table S6. List of Refseq genes mapped on the YH (Asian) specific sequences. The human Refseq genes that were unalignable to the NCBI reference genome were collected and aligned on the YH specific sequences using blat. Minimal length 100bp and minimal identity 90% was chosen as threshold to defining alignment hits.

(Attached Excel file)

Table S7. List of Refseq genes mapped on the NA18507 (African) specific sequences.

(Attached Excel file)

Table S8. Alignment hits of human proteins on the YH novel sequences. All Refseq human proteins were collected and aligned against the novel sequences using tBlastN (1E-5). On each novel sequence region, only the best alignment hit was picked up.

(Attached Excel file)

Table S9. Alignment hits of human proteins on the NA18507 novel sequences.

(Attached Excel file)

Supplementary data set

Data S1. Novel sequences identified in NA18507 genome.

(Attached txt file)

Data S2. Novel sequences identified in YH genome.

(Attached txt file)

Data S3. *De novo* assembly tool (SOAPdenovo v1.03) used for genome assembly.

(Attached tarball gzipped file, also available on http://soap.genomics.org.cn)

Supplementary discussion

Length of novel sequences

We did not identify any continuous novel sequences longer than, respectively, 19.2Kb and 16.0Kb in YH and NA18507 genome. It's likely because the current genome assemblies are more fragmental than the reference. It may also be the result of our using very stringent criteria to define novel sequences and required that they have less than 90% identity to any region of the reference genome. A manual check, using less stringent criteria and ignoring gaps and small (<300 bp) aligned fragments on the scaffold, allowed us to identify longer novel sequence regions, with longest being 45.5 Kb in YH and 21.0 Kb in NA18507.

Novel sequences frequencies in different populations

The profiled populations included all the major groups spread across the main geographic areas of Africa, Europe, Middle East, Southwest Asia, East Asia, Oceania, and America (Figure S7). Populations with less than 5 samples were merged according to their geographic distance and to previous studies on their genetic relationships[1,2]: the North Italian and Sardinian populations were merged into Italian; the North China minorities (Mongola, Daur, Hezhen, Xibo, Tu, and Oroqen) were grouped as CHN minorities; and South China minorities (She, Tujia, Miaozu, Lahu, Yizu, Dai, and Naxi) were grouped as CHS minorities.

Laboratory quality control

The quality of PCR primers and accuracy of PCR experiments were evaluated by using YH and NA18507 as controls, doing duplicate experiments on one DNA sample and using double non-overlapping PCR primers for one novel sequence. The results showed that only 3 (2.3%) novel sequences that should be present in YH failed to be amplified and that 1 (0.7%) novel sequence failed in NA18507; 13 novel sequences had conflicting amplification outcomes in the duplicated samples (4.0% error rate); and 22 non-overlapping PCR primers had conflicting amplification outcomes (3.8% error rate). In total, we estimated the PCR inaccuracy rate to be lower than 5%. The PCR amplification results are shown in Table S4. Overall, 77.2% of the PCRs (164 sequences in the 351 samples) were positive, and 81 (49.4%) of the 164 novel sequences had over 90% frequency in the 351 samples, which indicated that their absence from the NCBI human reference genome might represent minor genotypes (Table S5).

Phylogenetic and genetic structure analysis

In the neighbor-joining tree (Figure 1a), African populations were clearly separated from all other populations. Within the African populations, we found that phylogenetic relationships were identical to previous SNP genotyping results, which is expected given the high genetic diversity among these populations[3]. The Middle Eastern population (Druze) and South Asian populations (Balochi and Pathan) had novel sequence frequencies that are more similar to each other than to that of the European or East Asian populations. The European population relationships were not as clear; the frequency variation between these populations was only slightly higher than that within a population.

In assessing the genetic components of the populations by the STRUCTURE program [4] (Figure 1b), if we grouped the individuals into 2 clusters (*K*=2), the African populations separated from the others; if we grouped them into 3 clusters (*K*=3), we observed a split between the European/Middle Eastern/South Asian and the East Asian/ Oceanian /Native American populations; at *K*=4, the Native American population split from the East Asian and Oceanian populations; *K*=5 did not provide any further clear splits; at *K*=6, the two Oceanian populations separated from the East Asian populations. Additional *K* increase to 7 or 8 did not provide any further distinction with clear boundaries between groups. More individual samples and novel sequences will likely be required to distinguish more sub-clusters.

The novel sequences also showed distinct frequency clustering in the African, European, Middle East/Southwest Asian, East Asian, Oceanian, and Native American geographic populations (Figure 1a). Group A sequences had a much lower frequency in European and Middle Eastern/South Asian populations compared to the others; group B sequences were rare in Native American populations; group C sequences were rare in Oceanian populations; group D sequences were rare in East Asian populations; group E sequences had very low frequencies in African but very high frequencies in Native American populations; group F sequences had low frequencies in African, European, and Middle Easter/South Asian populations, but high frequencies in East Asian, Oceanian, and Native American populations; group G sequences were common and had little frequency differences across all the populations; group H sequences had high frequencies in African populations, but low frequencies in all other populations; and group I sequences had high frequencies in some of the African and NAN Melanesian populations, but were rare in the others. It suggests that each population has independent gain or lose of sequences after separation from the common ancestor.

A statistical analysis showed that 24 of the novel sequences had a significantly higher frequency in the African populations than in any other population (fisher's exact test, p<0.01), and that 7, 4, 15, 5, and 13 of the novel sequences were more abundant, respectively, in the European, Middle Eastern/South Asia, East Asian, Oceanian, and Native American populations (Table S5).

Novel sequence frequencies change along human migratory paths

More information about the population patterns of novel sequences is available at YH database (http://yh.genomics.org.cn).

References

1 Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381-2385, doi:10.1126/science.1078311
298/5602/2381 [pii] (2002).

2 Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104, doi:319/5866/1100 [pii]
10.1126/science.1153717 (2008).

3 Jakobsson, M. *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998-1003, doi:nature06742 [pii]
10.1038/nature06742 (2008).

4 Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587 (2003).