

## Supplementary files

Supplementary Figure S1: Graphical representations and example population allele frequencies of the top 40 unique patterns

Supplementary Figure S2: The loci distributions of 39 linear patterns along the 22 autosomes.

Supplementary Figure S3: The mappings from 38 linear allele frequency patterns of four continental populations to 80 phylogenetic tree topologies and the branches harboring allele frequency changes. Members within the same parentheses correspond to a node in the linear pattern. Continental populations were coded as 1 for AFR, 2 for EUR, 3 for SAS, and 4 for EAS, respectively. “+” denotes allele frequency change

Supplementary Figure S4: Hierarchical bi-clustering of genes according to their distributions of allele frequency pattern loci stratified into five levels. Each row denotes a gene, and each column denotes an allele frequency pattern. Each cell denotes the normalized or unnormalized SNP count for each pattern in each gene. The vertical color bars on the left annotate the dominant allele frequency patterns and cluster identities of the sorted genes.

Supplementary Table S1: Significant enrichments of ClinVar phenotype-associated loci among the linear allele frequency patterns. Each row reports the pattern index, ID and disease of the ClinVar phenotype, chromosome number and coordinates of the hot spot, the counts of associated loci in ClinVar, the counts of the associated loci within the hot spot,

the counts of the associated loci assigned to the pattern, the counts of the hot spot associated loci assigned to the pattern, the Poisson p-value pertaining to the loci density on the hot spot, and the enrichment (hyper-geometric) p-value of the designated pattern in the associated loci.

Supplementary Table S2: The Gene cluster members displayed in Supplementary Figure S4. Each sheet reports the clustering memberships at each level. Each row denotes a gene with the gene name displayed in column 4. Columns 1-3 denote the cluster index, total number of SNPs on the gene, and the dominant pattern indices of the SNPs. The remaining columns report the unnormalized (level 1) and normalized (other levels) frequencies of SNPs assigned to each of the 39 patterns.

Supplementary Table S3: Top-ranking enrichment of GO terms among the gene clusters. Each row denotes an enrichment record. Columns 1-3 denote the level, cluster index, and dominant pattern index of the enrichment record. Column 4 reports the enriched gene set. Column 5 reports the overlapped genes of the gene set members and cluster members. Column 6 reports the enrichment p-value. Columns 7-9 show the chromosome numbers, occurrence frequencies and fractions of the target loci on the reported genes.

Supplementary Table S4: Assessment of the effect of neutral evolution on the observed population allele frequency patterns. (A) Deviation of the allele frequencies from the Hardy-Weinberg Equilibrium. For each continental population (a row), we report the numbers of SNPs whose allele frequencies are beyond (column 2) and within (column 3) the HW Equilibrium, and the fraction of non-equilibrium SNPs. A SNP is within the HW

Equilibrium if its population allele frequency is within the two standard deviation confidence interval of the HW Equilibrium curve. (B) Occurrence numbers and fractions of the top-ranking allele frequency patterns from the 1000G data and from synthetic data simulated by a neutral evolution model. Each row denotes a pattern. Columns 1-2 report the ranks and the undirected graphs of the patterns. Columns 3-4 report the occurrence frequencies and fractions of these patterns in the 1000G data. Columns 6-7 report the means and standard deviations of the occurrence frequencies in the simulated data. Columns 8-9 report the means and standard deviations of the occurrence fractions in the simulated data. Column 5 reports the directions and p-values of the deviations from the neutral evolution model: a positive direction denotes that the empirical occurrence fraction is greater than that of the simulated data. (C) Robustness assessment of the comparison results between empirical and simulated data on chromosome 1. We report the means and standard deviations of the occurrence frequencies and fractions from the simulations of two model parameter configurations: default parameter values and a “small population model” by reducing the initial population sizes by five folds. The directions and p-values of the deviations from the two model settings are also reported.

Supplementary Text S1: Detailed description of the algorithms.

Supplementary File S1: The inferred allele frequency patterns of 78,122,254 loci in the 1000G data.

Supplementary File S2: The source codes of the programs for inferring the SNP allele frequency patterns and using these patterns for local ancestry inference, and toy example data for test running these programs.