Empirical distributions of F_{ST} from large-scale polymorphism data

Eran Elhaik and Aravinda Chakravarti

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

Abstract

Apportionment of human genetic variation has long established that most human variation is within continental populations rather than between them. Local adaptations increase population differentiation, measured with Wright's F_{ST}, thus high-F_{ST} values may be found at closely linked neutral loci where selection acts. Because high-F_{ST} values are unlikely in humans due to short evolutionary times, genetic drift, and migration, they have been frequently used to identify genes undergoing directional or heterotic selection. We used HapMap III data to study human genetic variation and found that only a paucity (12%) of the total genetic variation is distributed between populations of different continents and even lesser (1%) genetic variation is found between populations of the same continent. We also demonstrated that the distribution of F_{ST} varies significantly by allele frequency when divided into non-overlapping groups by allele frequency range. Because the mean allele frequency is a crude indicator of allelic age, these distributions mark the time-dependent change in genetic differentiation. The change in mean F_{ST} of these distributions is linear with changes in allele frequency, indicating the nature of allele frequency dynamics. These results suggest that investigating the extremes of the F_{ST} distribution for each allele frequency group may be more efficient for detection of selection. We demonstrate that such extreme SNPs are more clustered than expected, implying that these genomic regions are likely candidates for natural selection.

Methods

To study the distribution of genetic diversity between HapMap populations (LWK, MKK, YRI; CEU, TSI; CHB, CHD, JPT), we defined a hierarchical population structure of four levels: individuals, populations, continental populations, and total population. Apportionment of genetic variation was calculated using hierarchical *F*-statistics on a common subset of 1.1M autosomal and 32k X-chromosomal SNPs using the HierFstat package.

To study the effect of minor allele frequency (MAF) on the shape of the F_{ST} distribution, SNPs were divided into five allele frequency groups based on their mean MAF. The F_{ST} distribution was then calculated for each allele frequency group with F_{ST} calculated as $\sigma^2_{MAF} / MAF(1-MAF)$.

To study the differences between SNPs with high- and low-F_{ST} values, SNPs from the top >0.005 percentile of each F_{ST} distribution were termed $F_{ST>threshold}$ SNPs and all other SNPs $F_{ST<threshold}$. We tested whether $F_{ST>threshold}$ SNPs are more clustered than $F_{ST<threshold}$ SNPs by calculating the distances between adjacent SNPs for each allele frequency group and using the coefficient of variation to estimate the dispersal of the distance distribution. Because there are fewer $F_{ST>threshold}$ SNPs than $F_{ST<threshold}$ SNPs, we used 10,000 random subsets of $F_{ST<threshold}$ SNPs of equal size. Similarly, we compared the LD patterns between adjacent $F_{ST>threshold}$ and $F_{ST<threshold}$ SNPs.

Figure 1

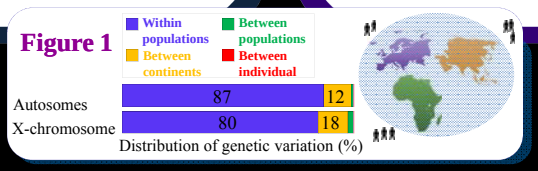
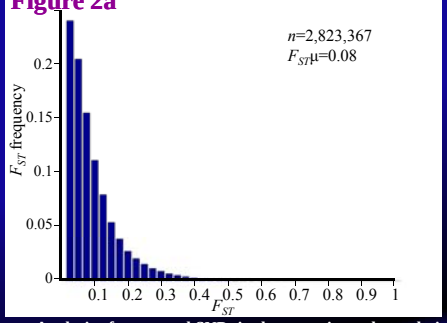
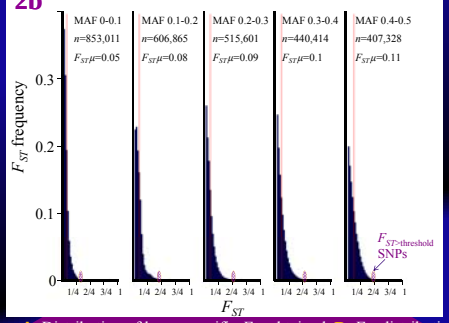


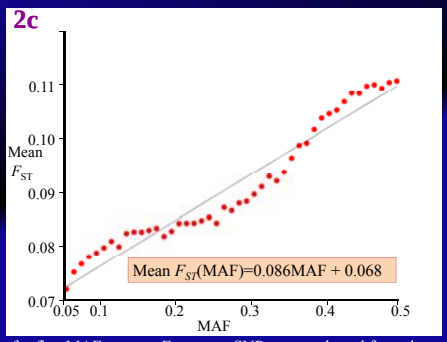
Figure 2a



2b

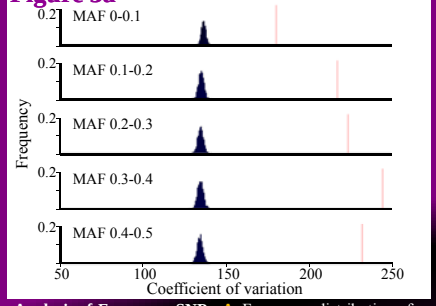


2c

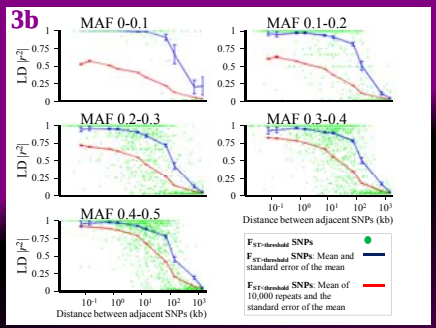


Analysis of autosomal SNPs in three continental populations. A. Distribution of locus-specific F_{ST} obtained. **B.** F_{ST} distributions for five MAF groups. $F_{ST>threshold}$ SNPs were selected from the top >0.005 percentile of each F_{ST} distribution. **C.** The mean F_{ST} plotted for 45 minor allele frequency (MAF) groups (dots) expresses the correlation between the two variables.

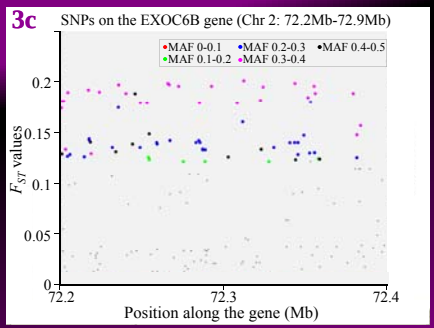
Figure 3a



3b



3c



Analysis of $F_{ST>threshold}$ SNPs. A. Frequency distribution of coefficient of variation calculated between adjacent $F_{ST>threshold}$ SNPs (red) and between random adjacent $F_{ST<threshold}$ SNPs (bars). **B.** LD in European population is plotted as a function of physical distance on a log-scale. $F_{ST>threshold}$ SNPs are marked as green dots. The mean and standard error of the mean r^2 for the $F_{ST>threshold}$ SNPs (blue) and $F_{ST<threshold}$ (red) are presented for different inter-marker distances. **C.** Example of a gene with a high density of $F_{ST>threshold}$ SNPs (color dots) compared with $F_{ST<threshold}$ SNPs (gray dots).

Results

Human populations can be divided into three old-world continents which is further subdivided into populations and individuals (**Figure 1**). Apportionment of human genetic variation using *F*-statistics showed that the great majority of genetic variation in autosomes (87%) exists within populations. Only a paucity of the total genetic variation (13%) is distributed between populations of different continents (12%) and even a lesser amount (1%) between populations within the continent. *F*-statistics were slightly higher in the X-chromosome than in autosomes, as expected from their copy number in the population.

The genomewide F_{ST} distribution (**Figure 2A**) concerned SNPs with dissimilar frequencies and biological properties owing both to the stochastic nature of genetic drift and to differences in the relative importance of the evolutionary region involved. Therefore, an F_{ST} distribution plotted for SNPs with particular minor allele frequency has a unique shape and variance because it describes regions that evolved at the same time and were affected by similar evolutionary mechanisms (**Figure 2B**). The relationship between the mean F_{ST} and the mean MAF is linear (**Figure 2C**). High-F_{ST} SNPs ($F_{ST>threshold}$) obtained from each F_{ST} distribution are significantly more clustered (bootstrap test $p<0.0001$) compared with $F_{ST>threshold}$ (**Figure 3A**). Moreover, these SNPs have a higher LD than expected by chance (**Figure 3B**).

Conclusions

Detecting signatures of natural selection and deciphering their causes can shed light on the evolution of the human genome and have practical implication for the search of loci involved in complex disorders. Unlike genetic drift, selection has a local effect that increases F_{ST} in particular loci due to the hitchhiking effect. Therefore, SNPs with similar minor allele frequencies and high F_{ST} are often targeted when searching for SNPs under selection. Identifying the shape of the F_{ST} distribution is thus critical to the detection of such SNPs.

Using the most complete SNP catalogue we showed that the F_{ST} distribution is approximately exponentially distributed . Moreover, we demonstrated that the F_{ST} distributions vary for different minor allele frequency groups (**Figure 2B**) that are similar in shape to the genome-wide F_{ST} distribution (**Figure 2A**). We showed empirically that the mean F_{ST} of each distribution is linearly correlated with the mean MAF. We identified subsets of high-F_{ST} SNPs from the tails of the F_{ST} distributions of five MAF groups. Because SNPs with similar MAF may share a common origin and demographic history, studying these SNPs is more informative than analyzing high-F_{ST} SNPs based on the genomewide F_{ST} distribution. Moreover, we showed that these SNPs are highly clustered and have a higher LD than expected by chance. These high-F_{ST} SNPs are therefore likely candidates for natural selection.