9-1-2012

A Tale of Two Haplotypes: The *EDA2R/AR* Intergenic Region is the most Divergent Genomic Segment between Africans and East Asians in the Human Genome

Amanda M. Casto
Department of Genetics, Stanford University, Mail Stop 5120, Stanford, California, 94305, USA, morgan21@stanford.edu

Brenna M. Henn
Department of Genetics, Stanford University, Mail Stop 5120, Stanford, California, 94305, USA

Jeffery M. Kidd
Department of Human Genetics, University of Michigan, 3726A Med Sci II, Ann Arbor, Michigan, 48109, USA

Carlos D. Bustamante
Department of Genetics, Stanford University, Mail Stop 5120, Stanford, California, 94305, USA

Marcus W. Feldman
Department of Biological Sciences, Stanford University, Gilbert Hall 108, Stanford, California, 94305, USA

Revised Version

A Tale of Two Haplotypes: The *EDA2R/AR* Intergenic Region is the most Divergent Genomic Segment between Africans and East Asians in the Human Genome

Research Article

Amanda M. Casto[1], Brenna M. Henn[1], Jeffery M. Kidd[2], Carlos D. Bustamante[1], and Marcus W. Feldman[3]

[1]Department of Genetics, Stanford University, Mail Stop 5120, Stanford, California, 94305, USA
[2]Department of Human Genetics, University of Michigan, 3726A Med Sci II, Ann Arbor, Michigan, 48109, USA
[3]Department of Biological Sciences, Stanford University, Gilbert Hall 108, Stanford, California, 94305, USA

Corresponding Author:
Amanda M. Casto
Department of Genetics
Stanford University
Mail Stop 5120
Stanford, California, 94305
USA
morgan21@stanford.edu

Key Words: Human, Evolution, Selection, Haplotype, X Chromosome

Running Head: Evolutionary History of the *EDA2R/AR* Intergenic Region

Abstract

Single nucleotide polymorphisms (SNPs) with large allele frequency differences between human populations are relatively rare. The longest run of SNPs with an allele frequency difference of one between the Yoruba of Nigeria and the Han Chinese is found on the long arm of the X chromosome in the intergenic region separating the *EDA2R* and *AR* genes. It has been proposed that the unusual allele frequency distributions of these SNPs are the result of a selective sweep affecting African populations that occurred after the Out-of-Africa migration. To investigate the evolutionary history of the *EDA2R/AR* intergenic region, we characterized the haplotype structure of 52 of its highly-differentiated SNPs. Using a publicly-available dataset of 3,000 X chromosomes from 65 human populations, we found that nearly all human X chromosomes carry one of two modal haplotypes for these 52 SNPs. The predominance of two highly divergent haplotypes at this locus was confirmed using a subset of individuals sequenced to high coverage. The first of these haplotypes, the α haplotype, is at high frequencies in most of the African populations surveyed and likely arose prior to the separation of African populations into distinct genetic entities. The second, the β haplotype, is frequent or fixed in all non-African populations and likely arose in East Africa prior to the Out-of-Africa migration. We also observed a small group of rare haplotypes with no clear relationship to the α and β haplotypes. These haplotypes occur at relatively high frequencies in African hunter-gatherer populations, like the San and Mbuti Pygmies. Our analysis indicates that these haplotypes are part of a pool of diverse, ancestral haplotypes that have now been almost entirely replaced by the α and β haplotypes. We suggest that the rise of the α and β haplotypes was the result of the demographic forces that human populations experienced during the formation of modern African populations and the Out-of-Africa migration. However, we also present evidence that this region is the target

of selection in the form of positive selection on the α and β haplotypes and of purifying selection against α/β recombinants.

Introduction

In humans, SNPs with large allele frequency differences between closely related populations are extremely rare [Coop et al. 2009]. There are some examples of SNPs in humans that have an allele frequency difference of one between distantly related populations, such as the Yoruba of Nigeria and the Han Chinese of East Asia [Coop et al. 2009; Casto et al. 2010; Lambert et al. 2010]. These SNPs are also relatively uncommon. There are only 71 such SNPs out of 656,995 SNPs surveyed in the CEPH-HGDP sample set [Casto et al. 2010]. Of these, 68 reside on the X chromosome [Casto et al. 2010]. X-linked SNPs, in general, tend to have higher F_{st} scores than their autosomal counterparts [Casto et al. 2010; Lambert et al. 2010]. This is likely the result of the increased effects of genetic drift on the X chromosome because of its smaller effective population size and the greater exposure of X-linked non-neutral variants to selection in males [Vicoso and Charlesworth 2006; Casto et al. 2010; Lambert et al. 2010].

When a large SNP allele frequency difference exists between African and non-African populations, the derived allele is almost always at high frequency outside of Africa [Lambert et al. 2010]. Again this is likely the result of both genetic drift and selection. The bottlenecks that characterized the Out-of-Africa migration and subsequent founding events gave rare derived alleles the opportunity to reach high frequencies in some populations [Reed and Tishkoff 2006; Li et al. 2008]. This might have been particularly true for X-linked SNPs as there is evidence that the X chromosome passed through a more severe Out-of-Africa bottleneck than did the autosomes [Keinan et al. 2009; Keinan and Reich 2010]. Selection may also have played a role as low frequency derived alleles interacted favorably with new environments outside of Africa

[Hancock et al. 2010]. There are, however, some exceptions. The derived Duffy allele, which is thought to provide protection against malaria, reaches high frequencies in many African populations [Hamblin and DiRienzo 2000]. Additionally, the intergenic region separating the *EDA2R* gene from the androgen receptor (*AR*) on the X chromosome contains a cluster of SNPs with large allele frequency differences between the Yoruba and the Han, a number of which have high derived allele frequencies in Africa [Casto et al. 2010; Lambert et al. 2010]. This run of highly-differentiated SNPs coupled with a reduction in haplotype heterozygosity observed in the Yoruba is suggestive of a selective sweep at this locus in Africans. Lambert et al. [2010] proposed that this sweep occurred after the Out-of-Africa migration as the ancestral alleles of many SNPs in this region are at high frequency in non-Africans despite the derived allele being nearly fixed in many African populations. However, a post Out-of-Africa selective sweep affecting most African populations seems unlikely given the deep genetic substructure in Africa and broad range of environments within which Africans live [Tishkoff et al. 2009; Schuster et al. 2010; Henn et al. 2011].

To investigate these conflicting hypotheses, we conducted a detailed study of the haplotype structure of the *EDA2R*/*AR* intergenic region. Using a composite dataset of 3,000 X chromosomes from many human populations, we found that nearly all of them carried one of two highly divergent haplotypes in this region. This same pattern was observed when we analyzed full sequence data for 26 additional X chromosomes. We discuss how the coexistence of these two haplotypes in some African populations explains the existence of so many SNPs with high derived allele frequencies in Africa and high ancestral allele frequencies in non-Africans without invoking a recent pan-African selective sweep.

Methods

Datasets

We used three SNP datasets in our analysis. The first, the CEPH-HGDP dataset, includes 938 individuals from 51 globally-dispersed populations [Li et al. 2008]. These 938 individuals carry a total of 1261 X chromosomes. The second, the HapMap dataset, contains 1521 X chromosomes from 11 populations that were genotyped as part of the HapMap3 phase of the HapMap project [The International HapMap 3 Consortium 2010]. Appendix Table 1 lists the names and descriptions of the populations in the HapMap dataset; abbreviations for these population names will be used in the text. The third, the Khoisan dataset, includes 91 individuals from three African populations [Henn et al. 2011]. These 91 individuals carry a total of 129 X chromosomes.

We also used sequence data for 26 male individuals from the HapMap3 sample set [Kidd et al. In Review].

Genomic Segment

We focused our analyses on the region of the X chromosome from 66,036,841 to 66,580,944 (NCBI36/hg18 Assembly), which we will subsequently refer to as the "540KB region". This region contains a total of 67 SNPs for which there is genotypic information available from all three of the SNP datasets described above.

Ancestral/Derived Alleles

We determined the identity of the ancestral and derived alleles for some of the SNPs in our region of interest using information from three different sources. The first comprised genotypes for two chimpanzees that were analyzed by Li et al. [2008] alongside the CEPH-HGDP samples. The second consisted of alignments of human, chimpanzee, orangutan,

macaque, and rhesus genome sequences available on the UCSC Genome Browser [Kent et al. 2002]. The third was ancestral/derived allele assignments available from the DBSNP database [Sherry et al. 2001]. Information on all SNPs was not available from every source. When ancestral/derived allele assignments were available from just one of these sources for a given SNP, we used that source's assignments. If ancestral/derived allele assignments were reported in multiple sources for a SNP, we only assigned ancestral/derived alleles to the SNP if all sources agreed which allele was derived and which ancestral.

Networks

Haplotype networks were created using NETWORK version 4.6.0.0 [Bandelt et al. 1999]. We created the networks using the Median Joining Method followed by post-processing with the MP Calculation option. The final versions of the networks displayed as figures were created using the tools found under NETWORK's "Draw Network" menu.

Tajima's D

Tajima's D values and corresponding p-values were calculated using DnaSP [Rozas and Rozas 1995].

Recombination Analysis

We calculated recombination rates for certain sets of HapMap X chromosomes using the *interval* program from the LDhat package [McVean et al. 2004]. We calculated the recombination rate of the β haplotype with itself in the CEU, TSI, and MKK populations (see Appendix Table 1 for full names of populations) using as input the chromosomes from those populations that carried the β haplotype or one of its satellites. The recombination rate of the α haplotype with itself was calculated in the MKK and LWK populations using only the chromosomes that carried the α haplotype or one of its satellites. Overall recombination rates

were calculated in the CEU, TSI, LWK, and MKK populations using all chromosomes except those carrying complex haplotypes. *interval* was run using a block penalty of 5, a total of 1,000,000 iterations for the rjMCMC procedure, and 2,000 updates between samples. The *stat* program was used to summarize the results of *interval*, discarding 50 samples as burn-in. *stat* gave as output an estimate of $\rho = 3N_e r$ between each of the 133 adjacent pairs of SNPs in the 540KB region for all the sample sets described above. For each sample set, we ran *interval* and *stat* 10 times. We then averaged the 1,330 estimates of ρ for the 540KB region for each sample set (10 estimates, one for each *interval* run, for each of the 133 adjacent SNP pairs in the 540kb region). Using a value of 10,000 for N_e, we divided this average ρ by 30,000 to arrive at the estimates of r given in the text.

We also simulated recombination between the α and β haplotypes using a program we wrote in *perl*. Kong et al. [2002] estimated the genetic map distance for the 1MB region containing the 540KB region (Chromosome X 66KB-67KB, hg18 assembly) to be 0.7 cM; this translates into a recombination rate of 7.0×10^{-9} per base pair. We therefore simulated two SNP loci with an interlocus recombination rate of 0.00378, which is equivalent to 540,000 times 7×10^{-9}. That is, the likelihood of recombination between these two loci is equivalent to the likelihood of recombination in a 540KB genomic region with a per base recombination rate of 7×10^{-9}. We began each run at time zero with 30,000 simulated chromosomes, 20,000 carried by diploid females and 10,000 carried by haploid males. At time zero, there was complete linkage between the alleles of the two simulated SNPs such that all chromosomes were either AB or ab. We varied the frequency of the "minor" haplotype from 0.05 to 0.2. These values were chosen to approximate the frequency of the α haplotype in the CEU and TSI (0.12 and 0.16, respectively) and the frequency of the β haplotype in the LWK and MKK (0.04 and 0.21, respectively). The

time since the settlement of Europe by human populations is usually given as ~40,000 years ago [Soares et al. 2010], which is equivalent to 1600 generations at 25 years a generation or 2000 generations for 20 years a generation, while the Out-of-Africa migration is usually thought to have been at least 60,000 years ago, equivalent to 2400 generations at 25 years a generation or 3000 generations at 20 years a generation [McEvoy et al. 2011]. We ran our simulation for 1500, 2000, 2500, or 3000 generations. In each generation, each female randomly mates with a male. Recombination occurs in the female at the rate given above. If recombination occurs in a female carrying an AB and an ab chromosome, Ab and aB chromosomes result. Each mating contributes a male and female offspring to the next generation (that is, this simulation models constant population size and equal fitness of all individuals) with each female passing at random one of her two chromosomes on to each offspring. After the set number of generations, the number of recombinant chromosomes out of 30,000 was counted. For each pair of input variables (frequency of minor haplotype and number of generations), this value was averaged over 100 runs.

Results

The α and β Haplotypes

We began by drawing a median-joining network of the haplotypes formed by the 67 SNPs (see Methods) within the 540KB region for all individuals in our combined SNP dataset. This network, two representations of which are shown in Figure 1, is primarily characterized by two tight clusters of haplotypes, separated by a long internal branch. This type of network has been referred to as a "dumbbell phylogeny" and is typical of sample sets having two subgroups that have been separated for some length of time [Avise 2000]. The long internal branch in our network represents 52 out of the 67 SNPs in the region of interest; these same 52 SNPs have an

allele frequency difference of 1.00 between the Yoruba and the Han Chinese in the HGDP dataset [Li et al. 2008]. To simplify our initial analysis, we proceeded by characterizing the haplotypes defined by these 52 SNPs, which we subsequently call the ”haplotype-defining” SNPs. We named the haplotype shared by all Yoruba X chromosomes the “α” haplotype and the haplotype common to all Han Chinese X chromosomes the “β” haplotype (see Table 1 for haplotype sequences). These two haplotypes account for 93.5% (2,721 of 2,911) of all the X chromosomes in the combined dataset.

The frequencies of these two haplotypes in the individual populations of the combined dataset are shown in Figure 2. Overall, the X chromosomes in our composite dataset illustrate that the α haplotype is fixed in some African populations while the β haplotype is fixed in most East Asian and American populations. Both haplotypes are found at significant frequencies in the populations of the Middle East, Europe, Central Asia, Oceania, and East Africa. Complete frequency data for the α and β haplotypes in all studied populations can be found in Appendix Table 2.

To ensure that the haplotype frequencies discussed above are not the result of phasing error, we also compiled haplotype frequency data separately for chromosomes from males and females. We found haplotype frequencies to be similar when comparing chromosomes from the two genders. In particular, the frequency of the α haplotype was 30.0% and 28.6%, respectively, among chromosomes from females and males and the frequency of the β haplotype was 63.9% and 64.2%, respectively, among chromosomes from females and males. Complete frequency data for the two genders can be found in Appendix Table 3.

The "Satellite", Recombinant, and "Complex" Haplotypes

The combined dataset contains a total of 2,911 X chromosomes; of these, 2,721 (93.5%) carried the α or β haplotype. We separated the haplotypes carried by the remaining 190 chromosomes into three general types. The first type included haplotypes which differed from either the α or β haplotype at four SNPs or less; we called these the "satellite" haplotypes. One satellite haplotype, α(25), was particularly common (see Table 1 for sequence information). It was observed a total of 60 times, almost exclusively on chromosomes from African populations. The second type of non-α, non-β haplotype included those that appeared to be single recombinants of the two major haplotypes. Only 45 X chromosomes in our dataset carried a recombinant haplotype, including 21 that carried the βαSR39 haplotype and 13 that carried the βαSR25 haplotype (see Table 2). We called the third type of non-α, non-β haplotype the "complex" haplotypes. These differed from both of the two major haplotypes at a large number of SNPs, but were not obviously the product of recombination between them. The most common complex haplotype was ComplexLWK1, which was observed 11 times. All complex haplotypes were found on X chromosomes from African populations or from admixed populations with African ancestry, except for two CEU chromosomes with complex haplotypes (ComplexCEU1 and ComplexCEU2). Within individual African populations, the incidence of complex haplotypes ranged from 33% and 20% in the CEPH-HGDP Mbuti and San, respectively to 0% in CEPH-HGDP Biaka Pygmies, Bantu, Mandenka, and Yoruba, and in the Khoisan-speaking Hadza. Haplotype frequencies for all satellite, recombinant, and complex haplotypes in all studied populations are listed in Appendix Table 2.

In an effort to understand the relationship of the complex haplotypes to the α and β haplotypes and to each other, we attempted to identify the ancestral and derived alleles of the SNPs under study; we were able to do this for 50 of the 52 haplotype-defining SNPs (see

Methods). The α haplotype carries a total of 30 ancestral alleles and 20 derived alleles while the β haplotype is composed of 20 ancestral alleles and 30 derived alleles. The number of derived alleles carried by the complex haplotypes ranged from 10 to 28. Ten derived alleles are seen on the ComplexLWK1 haplotype as well as on ComplexMKK2. Only three of the complex haplotypes carried more than 20 derived alleles; these included both the complex haplotypes seen in CEU individuals as well as ComplexMKK3, which differs from the β haplotype at only six SNPs. The possible mutational paths among the complex haplotypes and between the complex haplotypes and the α and β haplotypes are illustrated in a median joining network shown in Appendix Figure 1.

Haplotypes from Full Sequence Data

The SNPs described above were obtained from ascertained array data and are only a subset of the total number of SNPs within the 540KB region. Thus, they may not be representative of the overall genetic variation that exists in this genomic segment. We therefore obtained full X chromosome sequence data from a Complete Genomics dataset [Kidd et al. In Review] for 26 male individuals from the HapMap3 populations (see Methods). Using information from the 1,003 variable sites found between 66,036,841 and 66,580,944 on the X chromosome, we constructed a median joining network illustrating the relationship between the haplotypes carried by all 26 individuals (Figure 3). Within this network, all 26 haplotypes fall into only one of two highly differentiated clusters. Using the 52 haplotype-defining SNPs, we find that all the individuals in the bottom cluster carry the α haplotype haplotype while all the individuals in the top cluster carry the β haplotype. The long-branch separating the two clusters represents a total of 340 sites which are fixed for one allele amongst members of the α cluster and the other allele amongst members of the β cluster. No individuals carrying satellite, complex,

or recombinant haplotypes were observed amongst the 26 individuals for which full-sequence data is available, reinforcing the relative rarity of these haplotypes compared to the α and β haplotypes.

α and β Haplotype Networks

As illustrated in Figure 1, the majority of the individuals in the combined dataset fall into one of the two tight clusters at opposite ends of the haplotype network. Assuming that each of these clusters arose from the expansion of a single "founding" haplotype, the network structure of these two clusters could be informative about the evolution of the α and β haplotypes. To evaluate these two clusters for substructure, we drew one network containing only individuals with the α haplotype or one of its satellites and one network containing only individuals with the β haplotype or one of its satellites; these networks include information from all 67 SNPs genotyped in the combined SNP dataset. These two networks are indeed quite different from one another. The α network (Figure 4A) lacks a single distinct center, as commonly observed haplotypes appear throughout the network. The β haplotype network (Figure 4B), even though it represents many more individuals (1900 in the β network versus 938 in the α network), has many fewer unique haplotypes. It also has a star-like quality not observed in the network depicted in Figure 4A, with the common haplotypes clustered in the center and many rare haplotypes radiating out from this center.

We similarly noted that the structures of the α and β clusters in Figure 3 are somewhat different from one another. The β cluster has a star-like character much like the network depicted in Figure 4B. The α cluster instead has a long internal branch off which individual haplotypes sequentially arise. We calculated Tajima's D values for the α and β clusters in Figure 3 [Tajima 1989; Rozas and Rozas 1995]. The D value for the α cluster is -1.42 ($p > 0.10$) implying that the

expansion of the α haplotype is consistent with a neutral model of constant population growth [Tajima 1989; DiRienzo and Wilson 1991]. The D value for the β cluster is -2.35 ($p < 0.01$) which may indicate that this haplotype has undergone a period of exponential expansion in the past [Tajima 1989; DiRienzo and Wilson 1991].

Recombination Rates

On the 26 chromosomes for which we have full sequence data, there is no evidence of recombination between the α and β haplotypes in the 540KB region. This is also true for the vast majority of the 3,000 chromosomes in our SNP dataset, even though the α and β chromosomes co-segregate in the populations of East Africa, the Middle East, Europe, Central Asia, and Oceania. Given this observation, we carried out the following analyses to try to ascertain whether there was a true deficit of recombinant chromosomes in our sample set given the recombination rate in the region. Kong et al. [2002] estimated the genetic map distance for the 1MB region containing the 540KB region (Chromosome X 66KB-67KB, hg18 assembly) to be 0.7 cM; this translates into a recombination rate of 7.0×10^{-9} per base pair. We used the program LDhat [McVean et al. 2004] to directly estimate an average population recombination rate across the 540KB region using HapMap data from the CEU and TSI populations; we obtained values of 8.86×10^{-13} and 3.59×10^{-12}, respectively, for these two populations. However, we also calculated population recombination rates for the subset of CEU and TSI chromosomes carrying the β haplotype or one of its satellites and obtained values of 1.05×10^{-9} and 8.29×10^{-9}, respectively, estimates much closer to that of Kong et al. [2002]. Kong et al.'s [2002] estimate was made using data from a European population and recombination rates are known to vary among populations [Kong et al. 2010]. However, we did observe a similar trend for the LWK and MKK populations. When we used LDhat to estimate population recombination rates for the

540KB region for all chromosomes from these populations, we obtained values of 9.11×10^{-13} and 7.29×10^{-13}, respectively. We obtained higher estimates for the population recombination rate of the α haplotype with itself in the LWK population (1.71×10^{-12}), the α haplotype with itself in the MKK population (1.61×10^{-12}), and the β haplotype with itself in the MKK population (2.44×10^{-9}). We next ran a simple simulation (see Methods for details) to estimate the number of α/β recombinants we would expect to observe among the CEU, TSI, LWK, and MKK samples given an average recombination rate across the 540KB region of 7.0×10^{-9} per base pair. The results of the simulation are in Table 4. Under a number of different initial parameter values, the simulation consistently estimated that there should be a much larger number of recombinants in the CEU, TSI, LWK, and MKK populations than were actually observed, corroborating our results from the above LDhat analysis (see Tables 3 and 4). Overall, both our investigations of recombination rate in the 540KB region suggest a deficit of α/β recombinants in our sample set.

Discussion

The 540KB region between positions 66,036,841 and 66,580,944 on the long arm of the X chromosome is unique in the human genome. While SNPs with large allele frequency differences between even distantly related populations are relatively rare, this region contains at least 52 SNPs that have an allele frequency difference of one between the Nigerian Yoruba and Han Chinese; analysis of full genome sequence data from this region suggests that the total number of such SNPs may be in the hundreds. We examined the haplotypes formed by 52 of these SNPs and found that nearly all human X chromosomes carry one of two highly differentiated haplotypes, which we called the α and β haplotypes. The majority of the remaining

X chromosomes carry haplotypes that differ from the α or β haplotype at four SNPs or less or are recombinants of the two common haplotypes.

Origin of the α Haplotype

The α haplotype is found in all 12 of the African populations that we surveyed. Its frequency exceeds 50% in all of these populations except the Namibian and South African San and it is the only haplotype found in the CEPH-HGDP Yoruba and the Hadza. Assuming that the *α* haplotype arose just once, its presence in all 12 of these African populations indicates that the haplotype either appeared before the populations of Africa diverged or that it arose in one population and entered others as a result of migration. Under the model of migration, the most likely scenario would be that the α haplotype arose amongst West African, Niger-Congo speaking populations such as the Yoruba and Mandenka in which the frequency of the *α* haplotype is particularly high. These populations have expanded within the last 5,000 years and have moved across the African continent, mixing with and even replacing some previously distinct populations in the process [Beleza et al. 2005; de Filippo et al. 2011]. The Khoisan populations of southern Africa have experienced variable gene flow from Niger-Congo speaking populations within the last 2,000 years. The western African ancestry of South African San individuals varies considerably from person to person. Some individuals have as much as 30-40% while others appear to have none [Henn et al. 2011]. The α haplotype and its satellite haplotypes are prevalent among both admixed and non-admixed South African San individuals. Furthermore, 4 out of 5 male Namibian San individuals, none of whom demonstrate any western African ancestry [Li et al. 2008], carry the α or α(25) haplotype.

A number of studies of uniparental markers have indicated that the San are the most genetically divergent population in Africa [Behar et al. 2008]. Studies of autosomal markers

have come to the same conclusion; a neighbor-joining tree based on autosomal makers [Tishkoff et al. 2009] shows the San branching off first from all other African populations. This deep genetic divide separating the San from the rest of Africa means that if the San and Niger-Congo speaking populations do not share the α haplotype as a result of migration, then the haplotype must pre-date the division of African populations into distinct genetic entities. Since we found that the α haplotype is common amongst South African San individuals regardless of their proportion of western African ancestry, we believe that this is the most likely explanation for the presence of the α haplotype in all 12 of the African populations we surveyed. Additionally, as the α haplotype and its satellite haplotypes are relatively common in all 12 populations, it is reasonable to conclude that they were already prevalent when the San and the remaining African populations separated.

Origin of the β Haplotype

Among Africans, the β haplotype is primarily sequestered in the populations of East Africa, namely the MKK, the LWK, the Kenyan Bantu, the Sandawe, and the South African San. The presence of the β haplotype in the highly divergent San and in other African populations could indicate that the β haplotype is roughly contemporary with the α haplotype. However, on closer inspection, we find that the β haplotype is relatively rare in the San (it was found on only three San X chromosomes) and that it is only seen in individuals who have a significant degree of European admixture. Two further pieces of evidence point toward East Africa as the origin of the β haplotype. First, the β haplotype is completely absent from the Namibian San, the pygmies of Central Africa, and the Niger-Congo speaking populations of West Africa. All of the populations that do carry the β haplotype in Africa (other than the South African San) live in East Africa. Second, the complex haplotype that differs from the β haplotype at only six SNPs

was observed in the MKK, and not amongst the San or Pygmies, who have the highest frequency of complex haplotypes.

Although the frequency of the β haplotype is nearly 20% in the MKK, it attains its highest frequencies in non-African populations. Could this mean that the β haplotype arose outside of Africa and was brought into East Africa by migrants? Although this is a possibility, we believe that the evidence indicates otherwise. First, despite the relatively small number of β haplotypes observed in Africa, these haplotypes are quite diverse. Both of the common haplotypes in Figure 4B are observed in Africa as are several rare haplotypes. Also, common β satellite haplotypes like β(14) and β(30) are observed in Africa as is β(28) which is observed exclusively in the MKK. Second, if the ancestral alleles on the β haplotype were not carried out of Africa on the β haplotype, they must have migrated instead on a complex, proto-β haplotype like ComplexMKK3. However, there is no evidence for the existence of such a haplotype outside of Africa. The only two complex haplotypes observed outside of Africa are two closely related haplotypes observed in the CEU, both of which are less closely related to the β haplotype than is ComplexMKK3 (see Appendix Figure 1). Overall, the haplotype phylogeography of the 540KB region indicates that the β haplotype arose in East Africa and was carried to other continents by the Out-of-Africa migrants.

Out of Africa

The majority of the alleles that make up the α haplotype are ancestral. However, 20 out of the 52 haplotype-defining SNPs have their derived allele fixed in the Yoruba (and at quite high frequencies in most other African populations) and their ancestral allele fixed in the Han. Lambert et al. [2010] suggest that this is the result of a sweep occurring after the Out-of-Africa migration that rapidly increased the allele frequencies of the derived alleles on the α haplotype in

Africa. Thus, they suggest that at the time of the out-of-Africa migration, allele frequencies for SNPs across the 540KB region were quite different from their values today. In particular, the derived alleles that are part of the α haplotype were at either very low frequencies in Africa or had not yet arisen. This would indeed explain why the ancestral alleles for these SNPs were present among the out-of-Africa migrants. However, our analysis indicates that it is not necessary to invoke a pan-African selective sweep that occurred less than 50,000 years ago to explain modern allele frequency distributions for SNPs in the 540KB region. Additionally, the Tajima's D value in this region for individuals carrying the α haplotype is inconsistent with such a sweep having occurred. Rather we suggest that SNP allele and haplotype frequencies in Africa at the time of the Out-of-Africa migration were similar to what they are today. In particular, in East Africa, where the Out-of-Africa migrants originated [Ramachandran et al. 2008], the α haplotype may have been the predominant haplotype while a significant minority of X chromosomes carried the β haplotype. The α and β haplotypes then left Africa together in the Out-of-Africa migrants. The bottleneck engendered by the Out-of-Africa migration and subsequent founding events [Keinan et al. 2009] (possibly along with positive selection; *see below*) were sufficient to result in the β haplotype's prevalence among non-African populations. The α haplotype persisted in non-African populations but as the minority haplotype, establishing itself at low frequencies in the populations of the Middle East, Europe, Central Asia, Siberia and Oceania. Although other scenarios are certainly possible (see Appendix Note 1), we believe that modern SNP allele and haplotype frequencies are most consistent with the above sequence of events.

Complex Haplotypes

Despite the small number of chromosomes carrying complex haplotypes (n = 29 total chromosomes out of nearly 3,000), they exhibited considerable diversity; only three out of the 14 *unique* complex haplotypes were observed on more than one chromosome. Two pieces of evidence suggest that the complex haplotypes are part of a pool of ancestral haplotypes for this genomic region. First, while the α haplotype carries 20 derived alleles and the β haplotype carries 30 derived alleles, only three of the complex haplotypes carry more than 20 derived alleles and the most common complex haplotype, ComplexLWK1, carries only 10 derived alleles. Second, the populations with the highest frequency of complex haplotypes, the Mbuti and the San of Namibia and South Africa, were among the earliest to become genetically distinct in Africa [Knight et al. 2003; Patin et al. 2009; Tishkoff et al. 2009; Henn et al. 2011]. The presence of complex haplotypes at high frequencies in the Mbuti, Namibian San, and South African San (33%, 20%, and 8% respectively) may indicate that when these populations diverged complex haplotypes made up a significant proportion of the haplotypes in the 540KB region despite the paucity of these haplotypes in most of the modern populations that we surveyed. The high degree of diversity found among the complex haplotypes suggests that the 540KB region was once characterized by considerable haplotypic variability and that there has been a precipitous loss of haplotype heterozygosity in this region.

Selection

As we discussed above, the haplotype structure of the 540KB region is inconsistent with a recent pan-African selective sweep. However, there is evidence that selection has been important in the evolution of this genomic region. First, the region is characterized by an extremely high degree of linkage disequilibrium among SNPs, which rapidly breaks down

outside the region. Also, the overall lack of diversity in this region, particularly in populations where the α or β haplotype is fixed, leads to an overall reduction in haplotype heterozygosity relative to other X-linked loci. Low haplotype heterozygosity and reduced genetic diversity are often taken as evidence that positive selection has acted on a locus. Elevated F_{st} scores are also often used to identify genomic regions under selection; the 540KB region has an extremely high average pairwise F_{st} between African and East Asian populations, driven by the large number of SNPs that are fixed for opposite alleles in the two groups. Taken together these genomic features strongly suggest that the 540KB region is not evolving neutrally. Selection at this locus may have taken the form of positive selection on the α haplotype prior to the divergence of African populations, and in some African populations following divergence, leading to the establishment and maintenance of the α haplotype as the dominant haplotype among Africans. For the β haplotype, positive selection may have been important in its establishment in East Africa and in pushing it to high frequencies in non-African populations. Using two different approaches, we also found evidence that there is a deficit of α/β recombinants in populations where the two co-segregate, suggesting negative selection against such recombinants (see Appendix Note 1). Further investigation is necessary to more precisely define the directionality, intensity, and timing of selection, if any, on the 540KB region.

A number of polymorphisms across the 540KB region and in the nearby androgen receptor (*AR*) gene have been associated with androgen alopecia (AGA), also called male pattern baldness [Hillmer et al. 2005; Hillmer et al. 2008a; Hillmer et al. 2008b; Hillmer et al. 2009; Brockschmidt et al. 2010]. The variant with the most significant AGA association (~10^{-25}), rs12558842, is one of the 52 haplotype-defining SNPs [Brockschmidt et al. 2010]. When Brockschmidt et al. [Brockschmidt et al. 2010] corrected for this association in their study of

unrelated individuals, they found that no other polymorphisms remained significantly associated with AGA. The ancestral allele for rs12558842 is the risk allele for AGA and is carried by the β haplotype while the derived allele is protective and is carried by the α haplotype. Brockschmidt et al. [2010] studied Europeans, among whom this SNP is polymorphic due to the co-segregation of the *α* and β haplotypes. Although in modern Western societies AGA might be considered a purely cosmetic trait, the retention of hair on the top of the head may have had significant consequences for fitness in equatorial Africa [Wheeler 1984]. AGA may have also influenced mate preference by signifying maturity and dominance in age-related social hierarchies [Muscarella and Cunningham 1996]. It is notable that phenotypes related to sun-protection [Jablonski and Chaplin 2010] and hair morphology [Fujimoto et al. 2008] have been associated with selective sweeps at other loci in humans. Furthermore, the phenotypic influence of variants in the 540KB region may not be limited to AGA as many trait-associated polymorphisms frequently exhibit pleiotropy [Casto and Feldman 2011]. Indeed, there are a number of putative functional elements within the region (see Appendix Note 2 and Appendix Table 4). Variants in the *EDA2R* and *AR* genes which flank this region also have numerous known phenotypic consequences (Appendix Note 3) and it is possible that mutations in the 540KB region have regulatory effects on these genes (see Appendix Note 4 and Appendix Figure 2). Overall, the elevated F_{st} values, reduction in genetic diversity, extensive linkage disequilibrium, and relative paucity of recombinant haplotypes which characterize the 540KB region, along with the phenotypic connections of mutations within it, suggest that selection had a role in the evolution of the α and β haplotypes.

Conclusions

Our analysis of the haplotype structure of the 540KB region revealed that most SNP haplotypes in this genomic segment are members of one of two highly divergent clusters. One cluster (the α haplotype cluster) contains most of the haplotypes found in Africans and has its origins prior to the divergence of African populations from one another. The second cluster (the β haplotype cluster) is represented in East Africa, but includes the majority of haplotypes found in non-African individuals; this cluster has its origin in East Africa prior to the Out-of-Africa migration. We find that the structure of the haplotype network of the 540KB region is inconsistent with a post-out-of-Africa, pan-African selective sweep affecting this locus as proposed by Lambert et al. [2010]. Selection, however, may still have been important in the evolutionary history of this region. In particular, it may have been responsible for the replacement of the complex haplotypes by the α haplotype and then for the expansion of the β haplotype. In addition, negative selection may be responsible for an apparent deficit of α/β recombinants. If selection has occurred at this locus, it is likely attributable to its association with androgenic alopecia or to other phenotypic effects of the neighboring *EDA2R* and *AR* genes.

Appendix Materials

Appendix Materials are available online at http://www-evo.stanford.edu/pubs.html

Appendix Figure 1: Median Joining Network of Complex Haplotypes.

Appendix Figure 2: AR RNA Expression in HapMap YRI, CEU, and East Asian Individuals.

Appendix Table 1: HapMap Population Abbreviations and Descriptions.

Appendix Table 2: Incidences of Haplotypes for 540KB Region.

Appendix Table 3: Incidences of Haplotypes for 540KB Region for Male and Female Derived X Chromosomes.

Appendix Table 4: Candidate Function Elements in 540KB Region.

Appendix Table 5: Populations in Figure 2A

Appendix Note 1: Co-Segregation of the α and β Haplotypes and the Frequency of Recombinant Haplotypes.

Appendix Note 2: Functional Elements within the 540KB Region.

Appendix Note 3: Functional Elements and Phenotype Associated Variants outside the 540KB Region.

Appendix Note 4: AR Expression and Population.

Acknowledgements

This work was supported by grants (NIH GM28016, NIH 3R01HG003229) from the National

Institutes of Health to MWF and to BMH and CDB, respectively.

Literature Cited

Avise JC. 2000. *Phylogeography: The History and Formation of Species.* Cambridge, Massachusetts: Harvard University Press.

Bandelt H-J, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution 16: 37-48.

Behar DM, Villems R, Soodyall H, et al. (15 co-authors). 2008. The dawn of human matrilineal diversity. American Journal of Human Genetics 82: 1130-1140.

Beleza S, Gusmao L, Amorim A, Carracedo A, Salas A. 2005. The genetic legacy of western Bantu migrations. Human Genetics 117: 366-375.

Brockschmidt FF, Hillmer AM, Eigelshoven S, Hanneken S, Heilmann S, Barth S, Herold C, Becker T, Kruse R, Nothen MM. 2010. Fine mapping of the human *AR/EDA2R* locus in androgenetic alopecia. British Journal of Dermatology 162: 887-908.

Casto AM, Li JZ, Absher D, Myers R, Ramachandran S, Feldman MW. 2010. Characterization of X-Linked SNP genotypic variation in globally distributed human populations. Genome Biology 11: R10.

Casto AM, Feldman MW. 2011. Genome-wide association study SNPs in the human genome diversity project populations: Does selection affect unlinked SNPs with shared trait associations? PLoS Genetics 7: e1001266.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. 2009. The Role of Geography in Human Adaptation. PLoS Genetics 5(6): e1000500.

de Filippo C, Barbieri C, Whitten M, et al. (13 co-authors). 2011. Y-chromosomal variation in sub-Saharan Africa: insights in the history of Niger-Congo groups. Molecular Biology and Evolution 28: 1255-1269.

DiRienzo A, Wilson AC. 1991. The pattern of mitochondrial DNA variation is consistent with an early expansion of the human populations. Proceedings of the National Academy of Sciences 88: 1597-1601.

Fujimoto A, Kimura R, Ohashi J, et al. (15 co-authors). 2008. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. Human Molecular Genetics 17: 835-843.

Hamblin MT, Di Rienzo A. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. American Journal of Human Genetics 66: 1669-1679.

Hancock AM, Witonsky DB, Ehler E, et al. (11 co-authors). 2010. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proceedings of the National Academy of Science 107: 8924-8930.

Henn BM, Gignoux CR, Jobin M, et al. (19 co-authors). 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proceedings of the National Academy of Science 108: 5154-5162.

Hillmer AM, Hanneken S, Ritzmann S, et al. (25 co-authors). 2005. Genetic Variation in the Human Androgen Receptor Gene is the Major Determinant of Common Early-Onset Androgenetic Alopecia. American Journal of Human Genetics 77: 140-148.

Hillmer AM, Brockschmidt FF, Hanneken S, et al. (27 co-authors). 2008a. Susceptibility variants for male-pattern baldness on chromosome 20p11. Nature Genetics, 40: 1279-1281.

Hillmer AM, Flaquer A, Hanneken S, et al. (16 co-authors). 2008b. Genome-wide scan and fine-mapping linkage study of androgenetic alopecia reveals a locus on chromosome 3q26. American Journal of Human Genetics, 82: 737-743.

Hillmer AM, Freudenberg J, Myles S, Herms S, Tang K, Hughes DA, Brockschmidt FF, Ruan Y, Stoneking M, Nothen MM. 2009. Recent positive selection of a human androgen receptor/ectodysplasin A2 receptor haplotype and its relationship to male pattern baldness. Human Genetics 126: 255-264.

Jablonski NG, Chaplin G. 2010. Human skin pigmentation as an adaptation to UV radiation. Proceedings of the National Academy of Sciences 107(Supplement 2):8962-8968.

Jakobsson M, Scholz SW, Scheet P, et al. (24 co-authors). 2008. Genotype, haplotype, and copy-number variation in worldwide human populations. Nature 451: 998-1003.

Keinan A, Mullikin JC, Patterson N, Reich D. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nature Genetics 41: 66-70.

Keinan A, Reich D. 2010. Can a Sex-Biased Human Demography Account for the Reduced Effective Population Size of Chromosome X in Non-Africans? Molecular Biology and Evolution 27: 2312-2321.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Research 12: 996-1006.

Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, Louis D, Ruhlen M, Mountain JL. 2003. African Y chromosome and mtDNA divergence provides insight into the history of click languages. Current Biology, 13: 464-473.

Kong A, Gudbjartsson DF, Sainz J, et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. Nature Genetics 31: 241-247.

Kong A, Thorleifsson G, Gudbjartsson DF, et al. (15 co-authors). 2010. Fine-scale recombination rate differences between sexes, populations, and individuals. Nature 467: 1099-1103.

Lambert CA, Connelly CF, Macdeoy J, Qiu R, Olson MV, Akey JM. 2010. Highly punctuated patterns of population structure on the X chromosome and implications for African evolutionary history. American Journal of Human Genetics 86: 34-44.

Li JZ, Absher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-1104.

McEvoy BP, Powell JE, Goddard ME, Visscher PM. 2011. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Research 21(6): 821-829.

McVean GAT, Myers S, Hunt S, Deloukas P, Bentley D, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304: 581-584.

Muscarella F, Cunningham MR. 1996. The Evolutionary Significance and Social Perception of Male Pattern Baldness and Facial Hair. Ethology and Sociobiology 17: 99-117.

Patin E, Laval G, Barreiro LB, et al. (15 co-authors). 2009. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genetics 5: e1000448

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2008. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences 102: 15942-15947.

Reed FA, Tishkoff SA. 2006. African human diversity, origins, and migrations. Current Opinion in Genetics and Development 16: 597-605.

Rozas J, Rozas R. 1995. DnaSP, DNA sequence polymorphism: An interactive program for estimating population genetics parameters from DNA sequence data. Computer Applications in the Biosciences 11: 621-625.

Schuster SC, Miller W, Ratan A, et al. (48 co-authors). 2010. Complete Khoisan and Bantu genomes from southern Africa. Nature 463: 943-947.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29: 308-311.

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt HJ, Torroni A, Richards MB. 2010. The archaeogenetics of Europe. Current Biology 20: R174-R183.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

The International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.

Tishkoff SA, Reed FA, Friedlaender FR, et al. (25 co-authors). 2009. The genetic structure and history of Africans and African Americans. Science 324: 1035-1044.

Vicoso B, Charlesworth B. 2006. Evolution on the X chromosome: unusual patterns and processes. Nature Reviews Genetics 7: 645-653.

Wheeler PE. 1984. The evolution of bipedality and loss of functional body hair in hominids. Journal of Human Evolution 13: 91-98.

.

Tables

Table 1: Satellite Haplotypes

SNPs	Alternative Alleles	Ancestral Allele	Haplotypes																													
			β	β (1)	β(2)	β(11)	β(14)	β(27)	β(28)	β(30)	β(37)	β(41)	β(1,2)	β(2,3)	β(1,2,3)	β(11,12,13,14)	α	α(3)	α(21)	α(22)	α(25)	α(26)	α(39)	α(41)	α(51)	α(1,2)	α(3,25)	α(4,5)	α(25,39)	α(3,25,39)	α(25,30,31)
rs5919235	A/G	G	A	-	A	A	A	A	A	A	A	A	-	A	-	A	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-
rs7057795	T/C	C	T	T	-	T	T	T	T	T	T	T	-	-	-	T	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-
rs476709	T/C	C	-	-	-	-	-	-	-	-	-	-	-	T	T	-	T	-	T	T	T	T	T	T	T	T	-	T	T	-	T
rs574001	C/A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	-	C	C	C
rs532649	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-
rs485454	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs489099	T/C	C*	T	T	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs531840	T/C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs5919247	T/C	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs4827524	A/G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
rs989345	A/G	G	A	A	A	-	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs1567524	T/C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs5919270	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs1511061	C/T	T	-	-	-	-	C	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs1511060	T/C	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs5918694	C/T	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs4827392	A/C	C	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs5919272	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs5918696	T/C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs4827527	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs938059	A/C	C	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-
rs1988995	C/T	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-
rs984094	G/A	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs1027970	T/C	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs5919325	A/G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	-	A	A	A	A	A	-	A	-	-	-
rs6625150	A/C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	-	A	A	A	A	A	A	A	A	A
rs2335868	T/C	C	-	-	-	-	-	C	-	-	-	-	-	-	-	-	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs1255884	C/A	A	-	-	-	-	-	-	C	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs5918737	T/C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs4827539	T/C	C	T	T	T	T	T	T	T	-	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T
rs5919362	T/G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	T	T	T	T	T	T	T	T	T	T	T	-
rs4601479	A/G	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
rs6625163	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs5919363	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs5965383	T/G	G	T	T	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs2335506	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs2335508	G/A	A	G	G	G	G	G	G	G	G	-	G	G	G	G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs6625174	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs2878642	A/G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	-	A	A	A	A	A	-	-	A
rs2336175	A/G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
rs2497938	T/C	C	T	T	T	T	T	T	T	T	T	-	T	T	T	T	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-
rs2497939	C/A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
rs2223842	A/C	C*	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs2497943	T/G	G	T	T	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs2473897	T/C	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs2497944	A/G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
rs2223841	T/C	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs2473896	T/C	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs2473895	T/C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
rs2207080	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs721451	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-
rs2473891	T/C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

*Ancestral Allele could not be determined for this SNP (see Methods)

Table 2: Recombinant and Complex Haplotypes

SNPs	Alternative Alleles	Ancestral Allele	Haplotypes																				
			αβSR29	αβSR33	αβSR40	α(25)βSR29	βαSR24	βαSR25	βαSR29	ComplexCEU1	ComplexCEU2	ComplexLWK1	ComplexLWK2	ComplexMbuti1	ComplexMbuti2	ComplexMbuti3	ComplexMEX	ComplexMKK1	ComplexMKK2	ComplexMKK3	ComplexSanNB	ComplexSanSA1	ComplexSanSA2
rs5919235	A/G	G	-	-	-	-	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-
rs7057795	T/C	C	-	-	-	-	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-
rs476709	T/C	C	T	T	T	T	-	-	-	-	-	-	T	T	-	T	T	-	T	T	T	T	-
rs574001	C/A	A	C	C	C	C	-	-	-	-	-	-	C	C	C	C	-	-	C	C	-	C	-
rs532649	A/G	G	-	-	-	-	A	A	A	A	A	-	-	-	-	-	-	A	-	-	-	-	-
rs485454	A/G	G	-	-	-	-	A	A	A	A	A	-	-	-	-	-	-	A	-	-	-	-	-
rs489099	T/C	C*	-	-	-	-	T	T	T	T	T	T	-	-	-	-	T	T	-	T	T	-	T
rs531840	T/C	C	T	T	T	T	-	-	-	-	-	-	T	T	T	T	-	-	T	-	-	T	-
rs5919247	T/C	C	-	-	-	-	T	T	T	T	T	T	-	-	-	-	T	T	-	T	T	-	T
rs4827524	A/G	G	A	A	A	A	-	-	-	-	-	-	A	A	A	A	-	-	A	-	-	A	-
rs989345	A/G	G	-	-	-	-	A	A	A	A	A	A	-	-	-	-	A	A	-	A	A	-	A
rs1567524	T/C	C	T	T	T	T	-	-	-	-	-	-	T	T	-	T	-	-	T	-	-	T	-
rs5919270	A/G	G	-	-	-	-	A	A	A	A	A	A	-	-	-	-	A	A	-	A	-	-	-
rs1511061	C/T	T	C	C	C	C	-	-	-	-	-	C	C	C	C	C	C	-	C	C	C	C	C
rs1511060	T/C	C	-	-	-	-	T	T	T	T	T	-	-	-	-	-	-	T	-	-	-	-	-
rs5918694	C/T	T	-	-	-	-	C	C	C	C	C	C	-	-	-	-	C	C	-	C	C	-	C
rs4827392	A/C	C	-	-	-	-	A	A	A	A	A	A	-	-	-	-	A	A	-	A	A	-	A
rs5919272	A/G	G	-	-	-	-	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-
rs5918696	T/C	C	T	T	T	T	-	-	-	-	-	-	T	-	T	-	-	-	T	-	-	T	-
rs4827527	A/G	G	-	-	-	-	A	A	A	A	A	A	-	A	-	A	A	A	-	A	-	-	-
rs938059	A/C	C	-	-	-	-	A	A	A	A	A	A	-	A	-	A	A	A	A	A	A	-	A
rs1988995	C/T	T	-	-	-	-	C	C	C	C	C	C	-	C	-	C	C	C	-	C	C	-	C
rs984094	G/A	A	-	-	-	-	G	G	G	G	G	-	-	-	-	-	-	-	-	-	-	-	-
rs1027970	T/C	C	-	-	-	-	T	T	T	T	T	-	-	-	-	-	-	T	-	-	-	-	-
rs5919325	A/G	G	A	A	A	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-
rs6625150	A/C	C	A	A	A	A	A	A	-	-	-	-	A	-	-	-	-	-	-	-	-	A	-
rs2335868	T/C	C	T	T	T	T	T	T	-	-	T	-	T	-	-	-	-	-	-	-	T	T	T
rs12558842	C/A	A	C	C	C	C	C	C	-	C	C	-	C	-	-	-	-	-	-	-	-	C	-
rs5918737	T/C	C	T	T	T	T	T	T	-	T	T	-	-	-	-	-	-	-	-	-	T	T	T
rs4827539	T/C	C	T	-	-	T	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-
rs5919362	T/G	G	-	T	T	-	T	T	T	T	T	-	-	T	-	-	-	-	-	-	-	-	-
rs4601479	A/G	A	-	G	G	-	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
rs6625163	A/G	G	A	-	-	A	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-
rs5919363	A/G	G	A	A	-	A	-	-	-	A	A	-	A	-	-	A	-	A	-	-	-	-	-
rs5965383	T/G	G	T	T	-	T	-	-	-	-	-	-	T	-	T	T	-	T	-	-	-	T	-
rs2335506	A/G	G	A	A	-	A	-	-	-	A	A	-	A	-	A	A	-	A	-	-	-	A	-
rs2335508	G/A	A	G	G	-	G	-	-	-	G	G	-	-	-	-	-	-	G	-	-	-	-	-
rs6625174	A/G	G	A	A	-	A	-	-	-	-	-	-	A	-	A	A	-	A	-	-	-	A	-
rs2878642	A/G	G	-	-	A	-	A	A	A	-	-	-	-	A	-	-	-	-	-	-	-	-	-
rs2336175	A/G	G	-	-	A	-	A	A	A	A	A	-	-	A	-	-	-	-	-	-	-	-	-
rs2497938	T/C	C	T	T	T	T	-	-	-	-	-	-	T	-	T	T	-	T	-	-	-	T	-
rs2497939	C/A	A	-	-	-	-	C	C	C	C	C	-	-	C	-	-	-	-	-	-	-	-	-
rs2223842	A/C	C*	A	A	A	A	-	-	-	-	-	-	A	-	A	A	-	A	-	-	-	A	-
rs2497943	T/G	G	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-
rs2473897	T/C	C	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-
rs2497944	A/G	G	-	-	-	-	A	A	A	-	-	-	-	A	-	-	A	-	A	A	A	-	A
rs2223841	T/C	C	T	T	T	T	-	-	-	-	-	-	T	-	T	T	-	T	-	-	-	T	-
rs2473896	T/C	C	T	T	T	T	-	-	-	-	-	-	T	-	-	T	-	T	-	-	-	-	-
rs2473895	T/C	C	-	-	-	-	T	T	T	-	-	-	-	T	-	-	-	-	-	-	-	-	-
rs2207080	A/G	G	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rs721451	A/G	G	A	A	A	A	-	-	-	A	A	-	-	-	-	-	-	A	-	-	-	-	-
rs2473891	T/C	C	-	-	-	-	T	T	T	-	T	-	-	T	-	-	-	-	-	-	T	-	T

*Ancestral Allele could not be determined for this SNP (see Methods)

Table 3: Observed Frequencies of α/β Recombinants

Observed		
Population	**Frequency of Alternate Haplotype**	**Frequency of Single Recombinants**
CEU	**0.1188**	**0.0625**
TSI	**0.1570**	**0.0303**
LWK	**0.0435**	**0.0148**
MKK	**0.2081**	**0.0093**

Table 4: Expected Frequencies of α/β Recombinants

Simulation Results			
Frequency of Alternate Haplotype	**Number of Generations**	**Average Frequency of Recombinants**	**Standard Error**
0.05	**1500**	**0.0896**	**0.0035**
0.1	**1500**	**0.1752**	**0.0052**
0.15	**1500**	**0.2495**	**0.0052**
0.2	**1500**	**0.3063**	**0.0056**
0.05	**2000**	**0.1009**	**0.0040**
0.1	**2000**	**0.1839**	**0.0055**
0.15	**2000**	**0.2508**	**0.0063**
0.2	**2000**	**0.3216**	**0.0054**
0.05	**2500**	**0.0927**	**0.0049**
0.1	**2500**	**0.1780**	**0.0065**
0.15	**2500**	**0.2584**	**0.0054**
0.2	**2500**	**0.3143**	**0.0064**
0.05	**3000**	**0.0886**	**0.0053**
0.1	**3000**	**0.1800**	**0.0066**
0.15	**3000**	**0.2548**	**0.0075**
0.2	**3000**	**0.3257**	**0.0064**

Figure Legends

Figure 1: Median Joining Network of CEPH-HGDP Individuals. A) This network illustrates the possible mutational paths among the haplotypes formed by the 67 SNPs found in the 540KB region in the combined dataset. Each node represents an observed haplotype. In this portion of the figure, all nodes are of equal size to better illustrate the structure of the network. The lengths of the edges are proportional to the number of SNPs they represent. B) The same as in A) except that the size of the nodes in this portion of the figure are proportional to the incidence of the haplotype they represent.

Figure 2: Haplotype Counts per Population. A) The vertical bars represent the total number of chromosomes sampled from each CEPH-HGDP population. The blue portions of these bars represent the chromosomes that carry the α haplotype while the red portions represent the chromosomes that carry the β haplotype. The purple portions represent the chromosomes carrying complex haplotypes while the yellow portions represent chromosomes carrying all other haplotypes. The vertical black lines divide the populations into seven continental groups. See Appendix Table 5 for the names of the populations represented in this figure. B) The same as in A) for the HapMap and Khoisan populations. These populations are grouped by continent and dataset. The admixed HapMap populations (ASW, MEX) are also grouped together.

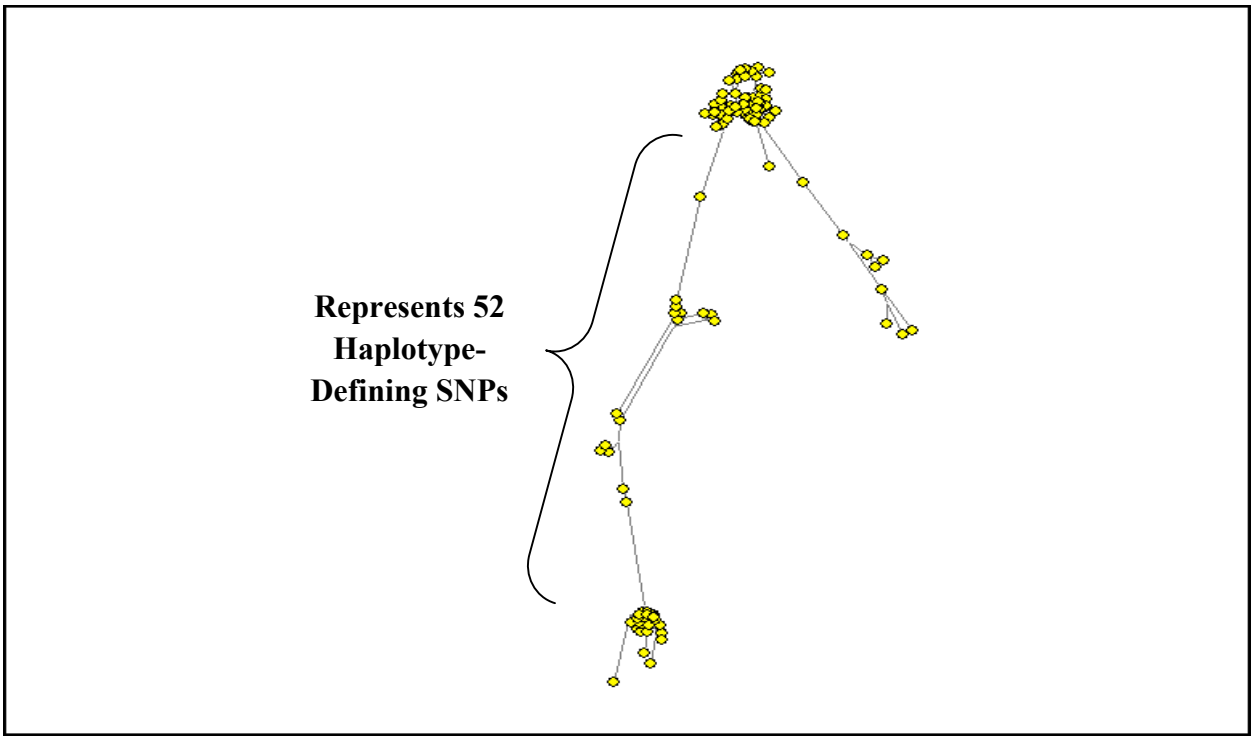
Figure 3: Median Joining Network of Full Sequence Haplotypes. This network illustrates possible mutational paths among haplotypes carried by 26 HapMap X chromosomes. These haplotypes were constructed using high coverage sequence data for the 540KB region. Blue nodes represent chromosomes from African populations, yellow nodes represent chromosomes form European populations, green nodes represent chromosomes from South Asian populations, purple nodes represent chromosomes from Asian populations, and red nodes represent

chromosomes from admixed populations. All of the chromosomes in the bottom cluster carry the α haplotype and all the chromosomes in the top cluster carry the β haplotype. The long branch separating the two clusters represents 340 variable sites.

Figure 4: Median Joining Networks of the α and β Haplotypes. A) This network illustrates possible mutational paths among the haplotypes formed by the 67 SNPs found in the 540KB region in the combined dataset; only chromosomes carrying the α haplotype or one of its satellites are represented in this network. The size of each node is relative to the frequency of its respective haplotype. Each node is color-coded to represent the geographic origin of the chromosomes carrying the haplotype with purple representing chromosomes from European and Asian populations and yellow representing chromosomes from African populations. The lengths of the edges of the network are proportional to the number of SNPs they represent. B) The same as in A) except that only chromosomes carrying the β haplotype or one of its satellites are represented in this network. The color coding of the nodes is the same as in A).

Figures

A.



B.

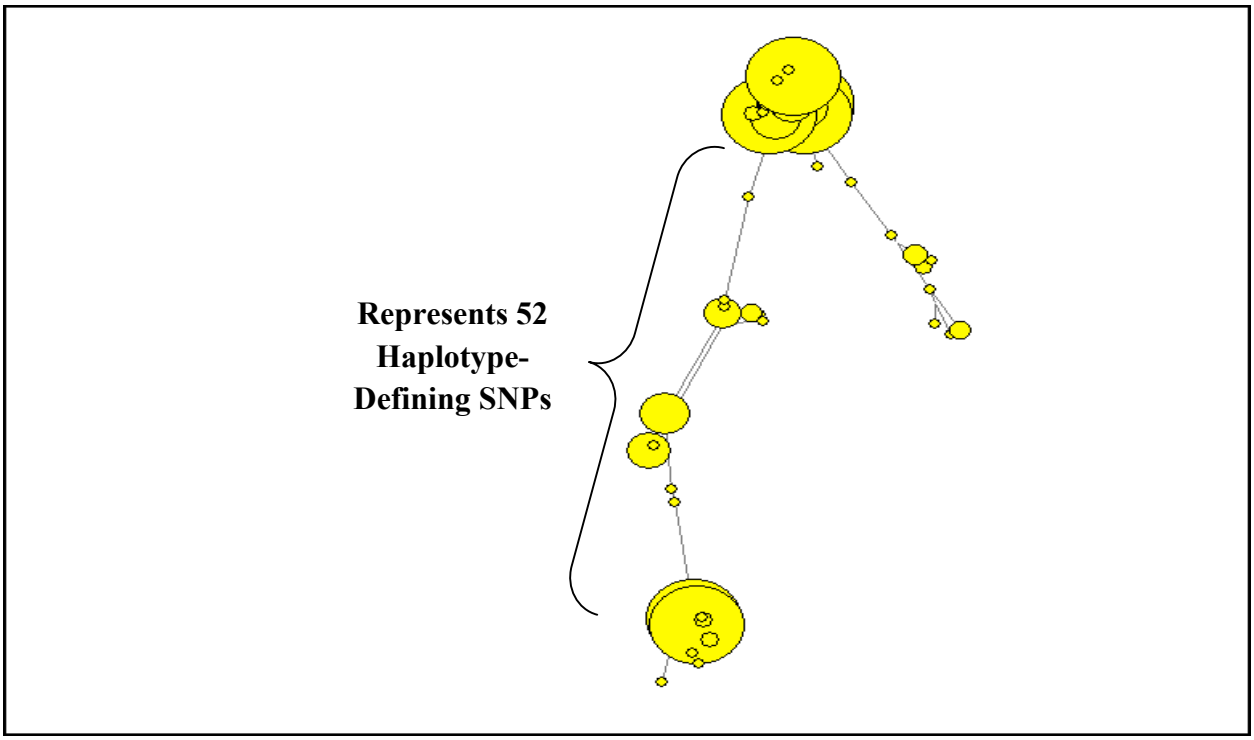
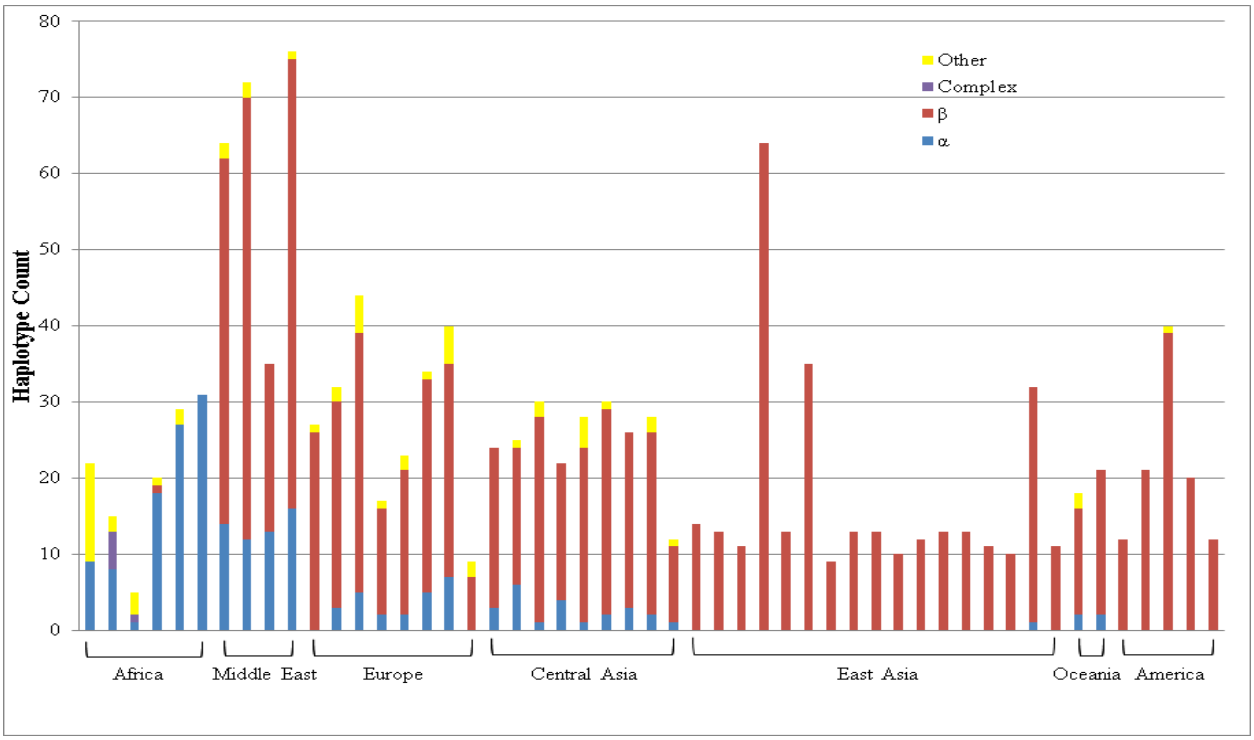


Figure 1

A.



B.

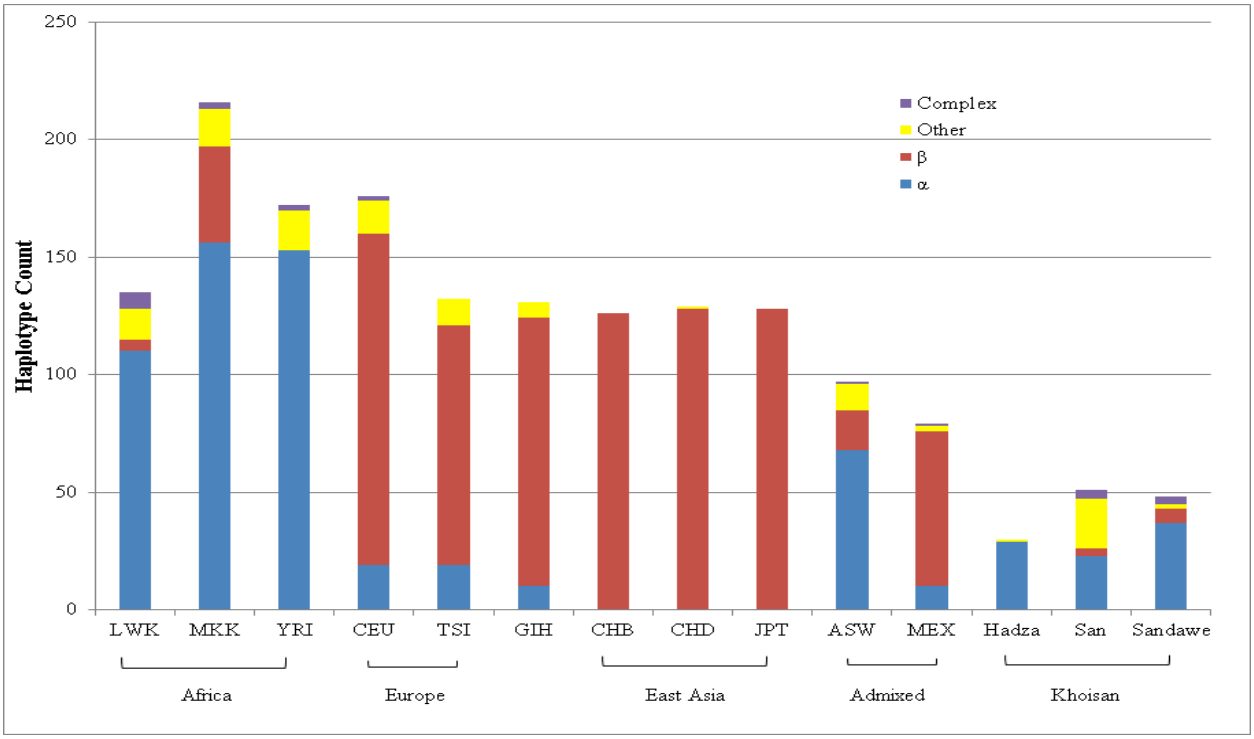


Figure 2

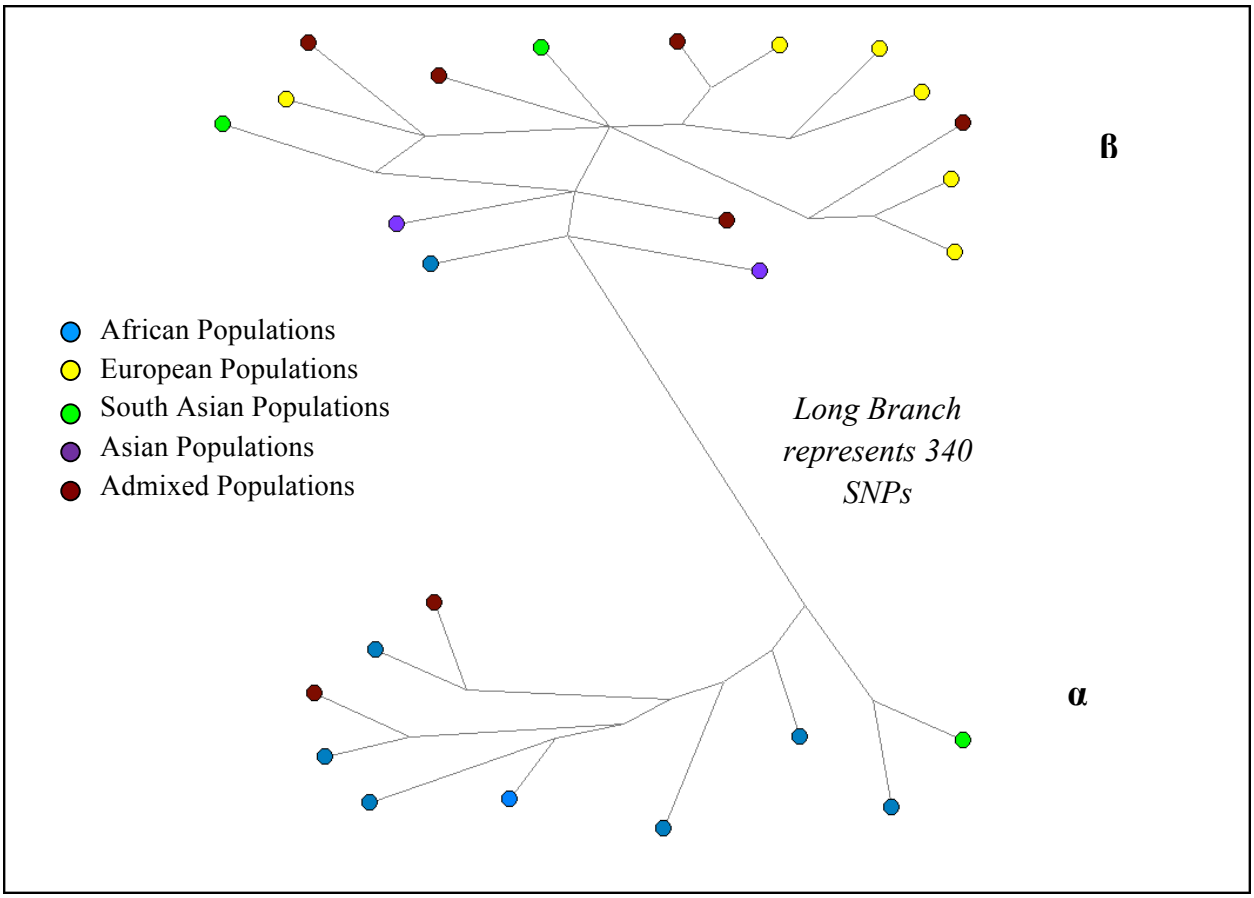
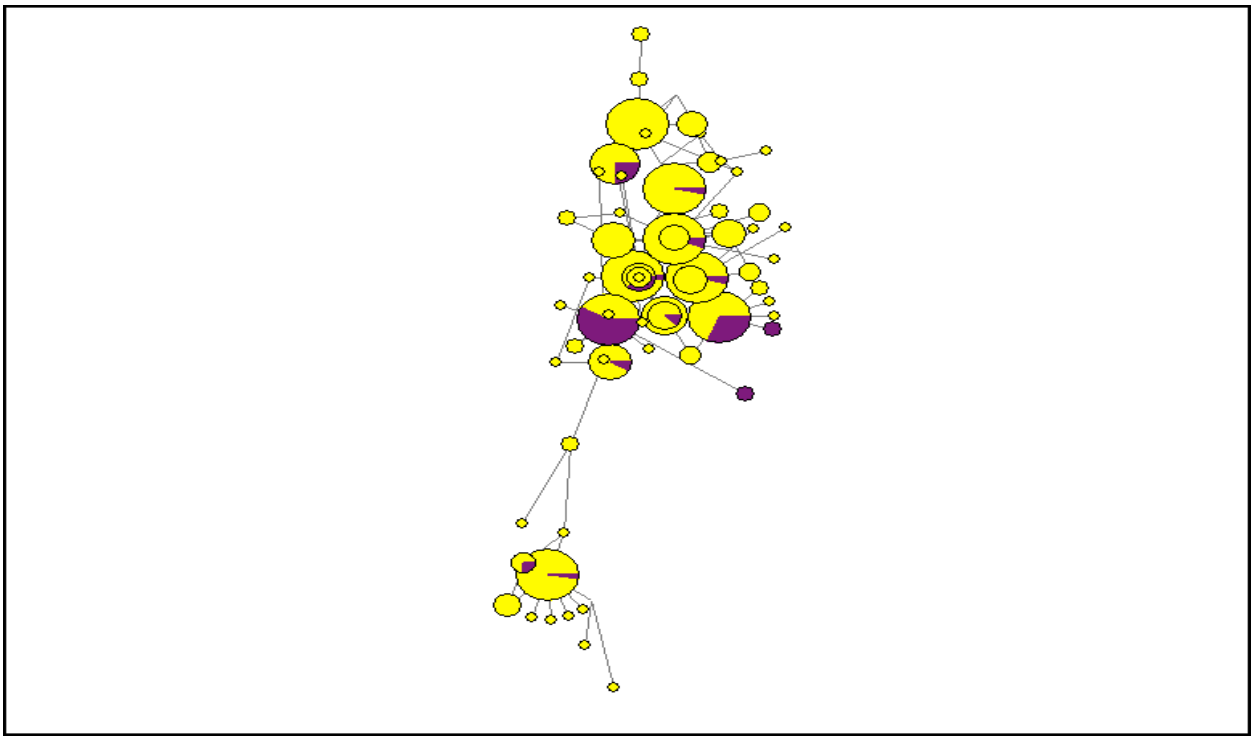


Figure 3

A.



B.

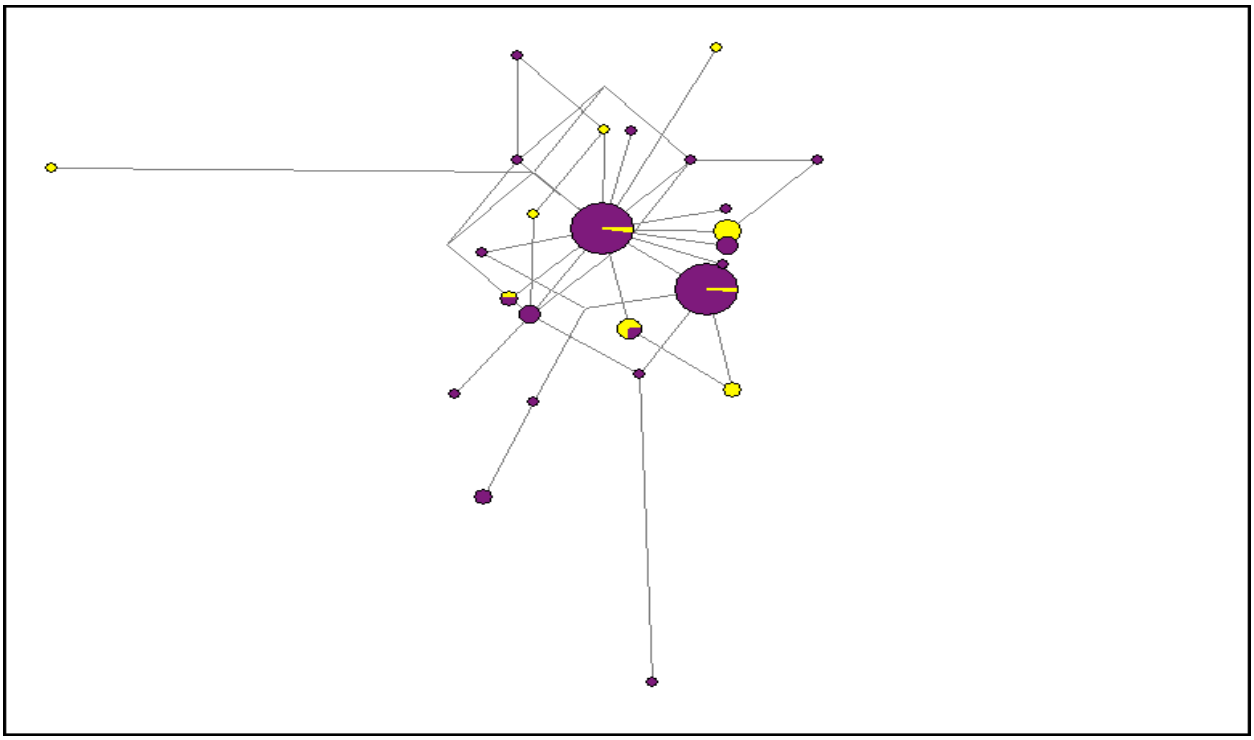


Figure 4