A

Thesis

Entitled

Determination of Early Human Intercontinental Migration from Genomic IBD segment Flow

By

Joseph S. Mainsah

Submitted to the Graduate Faculty as partial fulfillment of the requirement for the

Master of Science in Biomedical Sciences

_____
Dr. Alexei Federov, Committee Chair

_____
Dr. Sadik Khuder, Committee Member

_____
Dr. Robert Blumenthal, Committee Member

_____
Dr. Amanda Bryant-Friedrich, Dean
College of Graduate Studies

University of Toledo

December 2016

An Abstract of

Determination of Early Human Intercontinental Migration from Genomic IBD segment Flow

By

Joseph S. Mainsah

Submitted to the Graduate faculty as partial fulfillment of the requirements for the Masters of Science in Biomedical Sciences
Bioinformatics: Proteomics & Genomics

The University of Toledo
December 2016

Previous research demonstrated that genomic Identity by Decent (IBD) segments are a means of revealing distant relatedness between individuals from the same or different continents. When these segments are identically shared in one or more loci between individuals from different continents the inference is that these individuals are distantly related. IBD segments are characterized by rare variant clusters (RVCs) of 5 or more very rare genetic variants (vrGVs) with minor allele frequencies >2%. Although RVCs are a means of detecting distant genetic relatedness they cannot determine the directional flow of these IBD segments between continents. My objective was to determine the directional flow of these IBD segments from continents to continent after generations of admixture and to infer human continental migratory history. To perform

this task, I analyzed continental allelic frequency differences of SNPs located within IBD segments coordinates of related pairs of individuals' genomic data. Also, the occurrence of these SNPs in homozygous and heterozygous forms between pairs of individuals when there was a statistically significant difference in the allele frequencies of these SNPs from continent to continent. This experiment comprised individuals from African and Asian populations which shared the least number of IBD segments and African and European populations. Between the African and Asian populations, there was a clear majority (85%) of IBD segments analyzed flowing from Africa to Asia versus 15% from Asia to Africa. As for the IBD segments shared between the African and European populations there was a less definite result with 57% of segments flowing from Africa to Europe and 43% in the opposite direction. Given that the median size of IBD segments found between Asian and African populations was the smallest (54kb) compared to other intercontinental population comparison, there can be a conclusion that the admixture between the African and Asian populations is the most ancient given that with every generational admixture, IBD segments reduce in size from further admixture. These results would tend to imply that there was an ancient migratory event of Africans to Asia with some back migration occurring from Asia to Africa. On the other hand, my data suggest an almost equal degree of migration between African and European populations with a little more migration from Africa to European as opposed to migration from Europe to Africa. With this method of analyzing IBD segment flow it is difficult to determine whether these migrations occurred hundreds of thousands of years ago or more recently because of external issues like recombination and mutation

events which would affect IBD segments size. However, we get a general idea of migration patterns that have taken place between continents.

Acknowledgements

I will like to take this opportunity to thank my adviser Dr. Alexei Fedorov for being an outstanding mentor, a persistent and motivating hand that has installed in me the scientific curiosity to investigate and understand scientific truths. I am also very grateful for my teachers and academic advisors Dr. Robert Blumenthal and Dr. Sadik Khuder who exemplify professionalism and have guided me all through my time pursuing this degree. I am also thankful to Dr. Jyl Matson for her kindness and patience while I did a rotation in her lab and every day since. I greatly appreciate the support and work ethic of JoAnne Gray as she assisted me from my beginning admission into the program and throughout the course of my degree program. Last but not the least I would like to thank my lab colleagues Yuriy Yatskiv, Rajib Dutta, Shuhao Qiu, Patrick Brennan, Sharmistha Chakraborthy and Basil Khuder, I appreciate the collaborative effort in our research projects and I hope the friendship bonds we build last long as we all move on to our respective careers.

I am everyday thankful for my family which includes my mom Florence Mainsah, and my siblings Caroline, Charles, Pamela, Henry, Beri, and Asheri Mainsah, the love and support you have given to me has made me into the man I am today.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CHB……Han Chinese in Beijing, China CHS…….Southern Han Chinese, China

IBD……..Identical by Decent

JPT…….Japanese in Tokyo, Japan

LWK…….Luhya in Webuye, Kenya

RVC…….Rare Variant Cluster

SNP……..Single Nucleotide Polymorphism

TSI…… Toscani in Italy

VCF……..Variant Cell Format
VrGVs…….Very rare genomic variants

YRI……Yuruba in Ibadan, Nigeria

Chapter 1

Intricacies In Arrangement Of Human Common Haplotypes Suggest "Great Admixture" That Created Modern People

Rajib Dutta, Joseph Mainsah, Yuriy Yatskiv, Sharmistha Chakrabortty, Patrick Brennan, Basil Khuder, Shuhao Qiu, Larisa Fedorova, Alexei Fedorov

Program in Bioinformatics and Proteomics/Genomics, University of Toledo, Health Science Campus, OH 43614, USA.

Program in Biomedical Sciences, University of Toledo, Health Science Campus, OH 43614, USA.

Department of Medicine, University of Toledo, Health Science Campus, OH 43614, USA. GEMA-biomics, Ottawa Hills, OH 43606, USA.

.

1.0 My Contributions

I assisted in the writing of Perl scripts which grouped by similarity the human SNP haplotypes of 1092 individual's sequenced data from the 1000 Genome Project. This was followed by manual checking of the output tables to detect possible bugs in the Perl scripts. Also, I did sampling of some segments of the human genome to examine the patterns of frequent SNP haplotypes distributions.

The initial stages of this study involved literature research on similar studies previously published on common haplotypes of the human genome. I could locate a paper by Zhang et al. which discussed the initial discovery of high frequency mismatch SNP haplotypes which they called "Yin" and "Yang" haplotypes. This paper gave us a basic knowledge of what was known about high frequency SNP haplotypes and a base for our experiment which looked for other patterns in high frequency SNP haplotypes. I also participated in brainstorming exercises on problem solving and paths to pursue in the experiment at different points.

1.1 Abstract

Inferring history from genomic sequences is challenging and problematic because chromosomes are mosaics of thousands of small Identical-by-descent (IBD) fragments, each of them having their own unique story. However, the main events in recent evolution might be deciphered from parallel analysis of numerous loci. We computationally studied 5398 segments evenly covering all human autosomes. Common haplotypes built from frequent SNPs that are present in people from various populations have been examined. We demonstrated highly non-random arrangement of alleles in common haplotypes. Abundance of mutually exclusive pairs of common haplotypes that have different alleles at every polymorphic position (so-called Yin/Yang haplotypes) was confirmed in 56% of segments. A novel widely spread category of common haplotypes named Mosaic has been described. Mosaic consists of numerous pieces of Yin/Yang haplotypes and represents an ancestral stage of one of them. Scenarios of possible appearance of large number of frequent human SNPs and their habitual arrangement in Yin/Yang common haplotypes have been evaluated with an advanced genomic simulation algorithm (GEMA). Computer modeling demonstrated that the observed arrangement of 2.9 million frequent SNPs could not originate from a sole stand-alone population. Hence, a "Great Admixture" event has been proposed that can explain peculiarities with frequent SNP distributions. This Great Admixture presumably occurred 100-300 thousand years ago between two ancestral populations that had been separated from each other for about half a million years.

1.3 Introduction

The origin of modern humans has long been a topic of debate and is still an area of active research. The discussion of human evolution has largely progressed around two key models namely the 'out of Africa' versus the 'multi-regional' models. While the most widely accepted 'out of Africa' hypothesis proposes that *Homo sapiens* evolved in Africa before migrating across the world (Armour, et al. 1996; Horai, et al. 1995; Stringer and Andrews 1988; Tattersall 2009), the opposing 'multi-regional' model proposes that intermingling of the various populations evolving in several regions over a long period of time resulted in the emergence of the modern *Homo sapiens* species (Klyosov 2014; Wolpoff 2000). The events leading to the origin of *Homo sapiens* took place long ago, so our direct knowledge of human evolution is limited to several fossils of archaic hominoid individuals discovered in different parts of the world. Researchers have widely used genomic molecular markers such as the mitochondrial DNA (mt-DNA) and the non-recombining region of Y chromosome (NRY) to study different aspects of human evolution. These markers are transmitted uniparentally (mt-DNA maternally and NRY paternally) and thus have their own limitations (Stoneking and Krause 2011). The recent advancement in next generation sequencing (NGS) has made large scale sequencing of human genomes affordable, and has led to a huge amount of genome wide sequencing data from large population cohorts. Modern human genomes preserve and carry signatures of many events in human evolution such as population bottlenecks, migration, admixture, natural selection and genetic drift, and therefore serve as reliably informative resources for elucidating the history of mankind. The 1000 Genomes database includes genetic information of 26 populations belonging to the African, Asian, European and

American ancestry. This comprehensive resource on human genetic variation with diverse populations is ideal for the assessment of humans on a genomic scale. Recently our team computationally processed this database and demonstrated that very rare genetic variants (vrGVs, whose frequencies are less than 0.2%) are valuable markers for deciphering distant human relatedness (Al-Khudhair, et al. 2015; Fedorova, et al. 2016). This examination brought to light the human migration routes and admixture that happened up to ten thousand years ago. However, to reveal more distant events in the history of mankind, genetic variants with higher frequencies should be assessed. Keeping this in mind, here we investigated the distribution of haplotypes built from the most frequent SNPs (whose minor allele frequencies (MAF) are >25%) in people from Africa, America, Asia, and Europe. Surprisingly, intricacies of dynamics of frequent SNPs and their dependence on selection, recombination, and population structure have been investigated in only a handful of papers (Akey, et al. 2002; Altshuler, et al. 2010; Choudhury, et al. 2014; Guthery, et al. 2007; Hinds, et al. 2005; Zhu, et al. 2011). In this paper, we have examined why modern humans have a strikingly large number (2.9 million) of frequent SNPs. Frequent SNPs were studied not individually but in haplotypes – groups of 50 adjacent and closely linked SNPs. Such haplotypes were analyzed in 5398 segments along all autosomes. In a clear majority of human genomic regions, each segment contains a few common haplotypes that are widespread in 10%-90% of people from all continents. This is congruent with the findings of Gabriel and co-authors, who demonstrated that most of the human genome is contained in blocks/segments of substantial size and, within each segment, very few common haplotypes capture a vast majority (~90%) of the chromosomes in each population (Gabriel, et al. 2002).

Intriguingly, common haplotypes very often exist in mutually exclusive pairs. The two individual haplotypes from such a mutually exclusive pair have different alleles practically at every SNP site. Originally, this phenomenon was thoroughly investigated by Zhang with co-authors and they named these mutually exclusive haplotype pairs as "Yin" and "Yang" haplotypes (Zhang, et al. 2003). By analyzing common haplotypes in 62 random genomic loci and 85 gene-coding regions in humans, the Zhang *et al.* study proposed that the Yin/Yang haplotypes are abundant throughout the human genome and are genetic signatures that emerged prior to the African diaspora. We confirmed the widespread distribution of Yin/Yang haplotypes in humans and in addition revealed another widely distributed haplotype pattern, which we named "Mosaic". The Mosaic haplotypes are built from multiple small pieces of Yin/Yang haplotypes. These pieces are ancestral patterns of frequent SNP alleles that have been preserved for hundreds of thousands of years and represent an ancient prototype of Yin or Yang.

All in all, this large-scale bioinformatics examination allowed us to conjecture that modern populations were formed by the admixture of two ancestral lineages that separated from each other around 0.9-0.6 million years ago and re-admixed around 0.3-0.1 million years ago.

1.4 Materials and Methods

Genotype datasets for all the human chromosomes of the 1092 human genomes downloaded from the 1000 genomes ftp site (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/) as Variant Call Format (VCF) files version 4.1 (Abecasis, et al. 2012). This database contains a total of 38.2M SNPs,

3.9M short indels and 14K deletions for all the human chromosomes that have been used in this study. Information about parental haplotypes has been taken directly from Phase 1 of 1000 Genomes Project, since its genomic sequences were entirely "phased". Ancestral/Derived status for every genetic variant was obtained from the "AA=" field inside column 8 of the 1000 Genome VCF files.

For the archaic Neanderthal genome sequence we used Denisovan genomic datasets from the Max Plank Institute for Evolutionary Biology that are available through public ftp site http://cdna.eva.mpg.de/denisova/VCF/hg19_1000g/ (Meyer, et al. 2012). These Denisovan Variant Call Format (VCF) files contained coverage of the genome that is fairly uniform with 99.93% of the 'mappable' positions covered by at least one, 99.43% by at least ten, and 92.93% by at least 20 independent DNA sequences (Meyer, et al. 2012). We computationally processed the Denisovan VCF files with our novel Perl scripts (Denisova_Haplo_Find.pl, Deni_Stat_generator.pl), which are available from the Supplementary file and our web page (http://bpg.utoledo.edu/~afedorov/YinYang.html).

All haplotypes of 1092 individuals were obtained with our pipeline of eight Perl programs (*HaploFind.pl; HapGroupGenerator.pl, MosaicStatGenerator1.pl, MosaicStatGenerator2.pl, MosaicStatGenerator3.pl, YinYangStatExplorer.pl, MosaicStatExplorer.pl, CombineStatsYY_Mos.pl, AncestralHapMatchFinder.pl*). All our programs are available from the Supplementary files and from our web site (http://bpg.utoledo.edu/~afedorov/YinYang.html).

The command lines for programs execution and their instruction manuals are also available from the Supplementary file S2 and from our web site.

The entire dataset of all haplotypes for each 5398 chromosomal segments generated by our programs is available from the Supplementary file S2.

Computational simulations for the distribution and arrangement of SNPs in the population of virtual individuals were performed with our computational resource GEMA (Genome Evolution with Matrix Algorithms), which has been described in Qiu et al. 2014. In these simulations, we varied the size of the population (***N***); the selection pressure (number of offspring per individual ***α***); and the number of recombination events during the gametogenesis in the genomes of virtual individuals (***r***). The program code and instruction manual for GEMA are available from the Supplementary file S2 and from web site (http://bpg.utoledo.edu/~afedorov/YinYang.html). All SNPs generated during GEMA simulations were processed with the pipeline of Perl programs (*GemaBackupA_Process.pl, YinYangGema.pl, GemaSegments.pl, GemaHaplotypes.pl, GemaHapGroupGenerator.pl, Gema_HapGrouping.pl*). Perl scripts for these six programs, command lines for their execution, and their instruction manuals can be found in the Supplementary file S2 and at web site (http://bpg.utoledo.edu/~afedorov/YinYang.html). Statistics. P-values have been calculated using chi-squared test within Microsoft Excel package.

1.5 Results

1.5.1 Common Haplotypes (CHs)

All human autosomes have been divided into 500 Kb segments that are uniformly separated from each other as illustrated in Figure 1. For each chromosomal segment, we studied haplotypes built from 50 adjacent genetic variants occurring with high frequency

in modern humans (which Minor Allele Frequency (MAF) was >25% among 1092 sequenced genomes). Under this consideration, the physical length of haplotypes becomes a variable and depends on the density of frequent genetic variants (GVs) in the locus under investigation. The invariable quantity of 50 frequent GVs in each haplotype allowed us to make a fair comparison of occurrences of haplotypes from different chromosomal locations. In this study, positions of chromosomal segments have not been aligned with positions of genes for the following reasons: i) positions of genes are distributed highly non-randomly along chromosomes; ii) the sizes of genes vary considerably from a few hundred up to two million nucleotides; iii) the beginnings of genes often have elevated GC-composition. Thus, our approach should present an unbiased view on the distribution of haplotypes of frequent genetic variants in the entire human genome.

Each of 1092 individuals from phase 1 of the 1000 Genomes Project is presented by two haplotypes that correspond to two parents of the individual. The presentation of haplotypes of examined individuals is exemplified in Figure 1A. In addition to haplotype data, we extracted the ancestral/derived status for studied frequent genetic variants from the 1000 Genomes Project dataset (Figure 1A). Occurrence of all haplotypes of 1092 individuals have been ranked and examined throughout the human genome as explained in Figure 1C.

For each chromosomal segment, identical haplotypes from different individuals have been combined and ranked per their occurrence among 1092 sequenced individuals. Nearly identical haplotypes (with 0, 1 or 2 allele differences among 50 GVs) have been placed into the same group, which was assigned to the haplotype (“zeros/one’s string”)

with the highest occurrence. These haplotype groups are demonstrated in Figure 1C and are available for each chromosomal segment from the Supplementary Data File S1. When a haplotype group was found 100 or more times among 1092 studied individuals, it was considered as a *Common Haplotype (CH)*. Below we focus our research specifically on CHs because they might have existed in populations for hundreds of thousands of years and remain the same in several people from different populations. Thus, CHs may be of functional importance and their spread among populations and continents may reveal critical events occurred with ancient populations. Distribution of CHs has been examined among all human autosomes in 5398 segments, and these data are shown in the Supplementary Table S1. A subset of Table S1 is shown in Table 1 to illustrate our approach. The basic features of CH occurrence and distribution are illustrated in Figure 2. These data on CHs (number, size, and occurrence) are congruent to the results of Gabriel and co-authors (Gabriel, et al. 2002). Visual examination of CH strings revealed that many segments contain mutually exclusive CHs that differ from each other at practically every polymorphic position. This phenomenon is seen in Figure 2F and in Table 1, where the maximal difference of allelic variants between CHs from the same chromosomal segment is present in column 8. The first comprehensive examination of mutually exclusive CHs of humans was performed by (Zhang, et al. 2003). The authors called these mutually exclusive haplotypes *Yin* and *Yang*, and we will follow their nomenclature here. We made a threshold of 47 or more differences among 50 polymorphic sites (>=92% differences) to name a pair of CHs as Yin and Yang. This threshold was chosen to allow a few sequencing errors and/or occasional "jumping" of a particular SNP from one haplotype into another, which occasionally happens in accordance to the Biased

Gene Conversion (BGC) theory (Duret and Galtier 2009). The example of Yin/Yang CHs are group 1 and 2 in Figure 1C and Yin/Yang strings in Figure 3. All in all, 56% of all segments (or 59% of segments that have two or more CHs) have Yin-Yang pairs of CHs. Since the abundance of Yin and Yang haplotypes was a big surprise to us as well as to Zhang and co-authors (Zhang, et al. 2003), we examined this phenomenon in detail.

1.5.2 Characterization of Yin, Yang, and Mosaic CHs

The highest occurrence of Yin and Yang CHs was detected by Zhang and coauthors (2003) when they considered haplotypes built from high-frequency SNPs (MAF >20%). It dropped to about half when they reduced the MAF threshold to 5%. We also observed the same effect that MAF threshold occurs on the abundance of Yin/Yang pairs for our dataset (see Table 2). In the computations of Yin and Yang CHs by a pipeline of Perl programs, our algorithm assigns "Yin" to the CH with the highest occurrence, and "Yang" to its less frequent mutually exclusive counterpart. In addition to Yin and Yang, we also frequently observed CHs that could be reconstructed from many small pieces of Yin and Yang haplotypes ("zeros/ones" strings), as illustrated in Figure 3. Every haplotype can be reconstructed from fragments of perfectly exclusive Yin and Yang strings, since they contain all possible allelic variants. When a reconstruction is achieved using only two or three large pieces, it can be explained by one or two recombination sites between Yin and Yang respectively. However, we frequently observed situations when reconstruction could only be achieved by combining 10-30 small pieces. We named these CHs, which can only be reconstructed from ≥12 pieces, *Mosaic* haplotypes. The origin of such Mosaic haplotypes was examined below. Our Perl program, *CountMosaics.pl* counts the minimal number of Yin and Yang "pieces"

required to build a Mosaic haplotype. The characteristics of Yin, Yang, Mosaic, and all other CHs are present in the Supplementary Table S2. Their parameters include numbers of ancestral and derived alleles in haplotypes and numbers of Yin-Yang "pieces" required for reconstruction of non-Yin/Yang CHs. An important consequence of this Supplementary Table is that the fraction of derived alleles in Mosaic haplotypes (average 31% of derived alleles) is considerably less than in Yin and Yang (averages of derived alleles 55% and 43% respectively). Moreover, the more "pieces" involved in the construction of Mosaic segments, the less derived alleles they have. For example, when we increased the threshold for Mosaic haplotypes to ≥20 pieces, the percentage of derived alleles in them was reduced to 24%. In addition, in most of the cases, the derived alleles of a Mosaic haplotype predominantly matched only one Yin or Yang haplotype from this mutually exclusive pair. For example, in the Figure 3 ten out of twelve derived alleles of this Mosaic haplotype are found in Yin and only two derived alleles from this Mosaic haplotype are found in the Yang haplotype. Computations of all segments with Yin, Yang, and Mosaic CHs demonstrated that 59% of segments have a majority (80%) of Mosaic derived alleles belonging to one of the CH from Yin/Yang pair (see Figure 4). There is no statistical preference between Yin or Yang for the derived alleles of Mosaic to be matched to. All these observations may have a simple explanation if we assume that a Mosaic haplotype is an ancestral stage for the evolution of one of the Yin or Yang haplotypes. The alternative hypothesis that Mosaic is a product of multiple recombination events between Yin and Yang is not in line with these observations, because in this case one would expect to see a unimodal distribution of derived Mosaic

alleles among Yin and Yang haplotypes (which should be close to "Expected" distribution in the Figure 4).

1.5.3 Continental distribution of Yin, Yang, and Mosaic haplotypes

Distribution of Yin, Yang, and Mosaic haplotypes among continents has been examined and the results are shown in Figure 5. In our computations, we name the most abundant haplotype as Yin and the least abundant as Yang. Figure 5 displays that Yin haplotypes have statistically significant avoidance ($p<4*10^{-6}$, chi-squared test) of the African continent. Yang haplotypes as well have the same trend of minimal occurrence in Africa, though this is not statistically significant ($p=0.27$). Both Yin and Yang are nearly equally abundant in Europe and Asia. At the same time, Mosaic haplotypes are slightly more abundant in Africa than in Europe and Asia. This non-random occurrence among continents strengthens the possibility that Yin and Yang may correspond to two ancestral lineages, as one out of two alternative hypotheses Zhang and co-authors initially suggested (Zhang, et al. 2003). To reconstruct these ancestral lineages, we used Machine Learning approaches such as K-means Clustering and Decision Tree Classifiers to characterize the clusters that may correspond to these hypothetical lineages. Weka (Smith 2016) and Rapid Miner (Mierswa 2006) web computational resources were used for this purpose. Five normalized parameters for Yin and Yang haplotypes for each segment (total haplotype occurrence; the number of derived alleles; percentage of haplotype occurrence in Africa, Asia, and Europe) have been studied. However, despite our repeated attempts, we were unable to obtain any significantly well-separated clusters for these mutually exclusive haplotypes. These results are not shown here; details are provided in the Supplementary Data File S2.

1.5.4 Comparison of Yin, Yang, and Mosaic with ancestral haplotypes.

To evaluate the separation time of Yin, Yang, and Mosaic haplotypes we compared them with the available archaic human genome of one of the Neanderthal lineages ("pinky" Denisovan, (Meyer, et al. 2012)) whose DNA has been perfectly characterized (>30x coverage combined with high-quality reads in "bam" file). Alleles of frequent genetic variants that comprise our studied Yin, Yang, and Mosaic haplotypes of modern humans have been evaluated in the Denisovan genome sequence. Using these alleles, the Denisovan diplotypes have been assembled (parental haplotypes are not phased for this ancestral genome, so the diplotype is the only option). An example of this diplotype is shown in Figure 3. In total, we computationally processed 1720 chromosomal segments that contained Yin, Yang, and Mosaic haplotypes and added to their files corresponding Denisovan diplotypes. This information on Denisovan diplotypes is available from our Supplementary Data File S1. It appears that the analyzed frequent genetic variants in the Denisovan genome were predominantly homozygous (>99%). This phenomenon simplified the comparison of the ancestral diplotypes with modern haplotypes, because in a clear majority of cases a diplotype is the summation of two identical copies of homozygous haplotypes (see Figure 3). The computation of 1720 segments demonstrated that, on average, Denisovan haplotypes have the least number of derived alleles (18.0%), while Mosaic counterparts have 31% derived alleles, Yin – 55% and Yang 43%. Denisovan haplotypes are nearly identical (<=2 differences) to Mosaic haplotypes in 14% of analyzed segments (240 cases), whereas, such similarities with Yin and Yang haplotypes were found in 1.4% (25 cases) and 4.1% (71 cases) segments accordingly. Average allele difference between Denisovan haplotype and human CHs

was also found to be least for the Mosaic haplotypes (12 differences on average), followed by Yang (19 differences on average) and Yin (25 differences on average). All these data indicate that the Denisovan haplotypes are most closely related to Mosaic haplotypes. Since the Denisovan haplotypes contain considerably fewer derived alleles than Mosaics (on average 18% versus 31% of derived alleles respectively), the Neanderthal people must have separated from modern humans earlier than the formation time of Mosaic haplotypes. Taking the separation time of Neanderthals with modern humans between 0.8-0.55 Mya (Meyer, et al. 2012; Prufer, et al. 2014), we estimated that the time of Mosaic haplotypes' formation should be around 0.60-0.25 Mya.

1.5.5 Distribution of ancestral haplotypes among modern humans.

Since the Denisovan haplotypes contain only 18% of derived alleles and 82% of ancestral ones, we were intrigued whether some modern humans still have completely "ancestral" haplotypes built exclusively from ancestral alleles of frequent genetic variants. To answer this question, a 100% ancestral haplotype of the same 50 frequent genetic variants for each of 5398 segments have been deduced and compared with all available haplotypes of 1092 people. We allowed only one or two differences between the real haplotypes and the deduced 100% ancestral one to name them "ancestral". Within 867 out of 5398 segments, ancestral haplotypes were found among modern humans. Within 182 segments we counted less than 5 ancestral haplotypes among all individuals (rare ancestral haplotypes on Figure 5B); in 497 segments, we counted from 5 to 99 ancestral haplotypes (uncommon ancestral haplotypes on Figure 5B); and in 188 segments ancestral haplotypes were common (≥100 occurrences among 1092 people).

The abundance of these ancestral haplotypes among continents have been computed and presented on Figure 5.

Figure 5 reveals that ancestral haplotypes are most abundant in Africa. For 188 segments where ancestral haplotypes are also the common ones (occurred ≥100 times) most them, 184 segments, were observed on all continents and only four predominantly in Africa. However, these 184 "mixed" ancestral haplotypes still have the highest representation in Africa (42%), then in America (21%), Europe (20%), and Asia (17%).

1.5.6 Modeling the origin and abundance of CH using GEMA computer simulations.

Zhang and coauthors (2003) proposed that Yin-Yang haplotypes could arise due to the admixture of two ancient lineages of hominoids well before "Out-of-Africa" exodus or, alternatively, spontaneously from the sole ancestral population. The authors supported the latter hypothesis with computer simulations. However, Zhang et al. used simple simulations that did not consider parameters that notably influence SNP dynamics and linkage. Therefore, to understand the origin of numerous mutually exclusive CHs we performed advanced computer simulations using our GEMA computational resource (Qiu and Fedorov 2015; Qiu, et al. 2014). The GEMA program generates a population of virtual individuals, creates an influx of novel mutations in their genomes and starts multiple cycles of individuals' mating, offspring creations followed by their selection for surviving into the next generation. GEMA simulates dynamics of mutations under conditions close to natural. In these computations, we explored how the following parameters influence the formation of CHs and Yin/Yang pairs: 1) population size [$\boldsymbol{N}$ individuals per generation were changed in different simulations in the following range:

124, 250, 500, 1000, and 2000]; 2) number of meiotic recombination events per gamete (r) [***r*** was either 48 events (average for humans) or 24, 12, or 6 recombination's]; 3) selection pressure [***α*** parameter -- number of offspring per individual, which we changed from 2 (no selection) up to 10, which was the strongest in our experiments]. Other parameters were invariant and we used their default values: 1) flow of novel mutations per gamete [***μ***= 20, which was close to the natural rate of 20-50 novel mutations in human gametes]. 2) Mating schemes: random permanent pairs. 3) Co-dominant effect for ancestral/derived alleles (dominance coefficient: ***h***=0.5). 4) Distribution of mutation effects was ***Experiment-C*** (81% slightly deleterious; 9% beneficial; 10% neutral mutations). The results of our computer simulations are summarized in the Table 3.

In the GEMA simulations we first assessed the distribution of derived alleles by their frequency. A typical picture of such distribution is shown in Figure 6. The highest abundance was always observed for very rare derived alleles and the lowest abundance for nearly-fixed derived alleles. The curve in Figure 6 has the same shape as the real distribution of SNPs occurrence documented for the 1000 Genomes Project (see Extended Data Figure 3 in (Auton, et al. 2015)). In GEMA simulations and also in reality, the influx of novel mutations per generation is in direct proportion to the size of population ***N*** and equals ***2Nμ*** (blue arrow in Figure 6). For the constant size population, approximately 50% of novel mutations transiently exist in a single copy per generation (singletons) and are removed in the next generation or in a few generations after their arrival (bottom red arrow in Figure 6). Also, a considerable fraction of novel mutations exists in a few copies and still will drift away after several generations. Only a very minor fraction of derived mutations will survive and be fixed. The rate of fixed

mutations per generation is $k=2\mu$ according to the Kimura's law, which does not depend on the size of the population N (Kimura 1983). [In several textbooks μ is the number of novel mutations per person and so $k=\mu$.] Therefore, the number of frequent SNPs (MAF >25%) will be between these two extremes ($2N\mu$ and μ) and will grow with the increasing size of the population, approximately as square root of population size, $N^{1/2}$ (see Figure 7A). In 1092 sequenced human genomes, the number of frequent genetic variants (MAF >25%) is considerable and equals 2,944,337. Our simulation experiments with the parameters approximated to nature (α =5; r=48, μ= 20) demonstrated that such high number of frequent SNPs in a sole population is achieved when N is about 25 million (see Table 3 and Figure 7). In these computations, we used the lowest estimations of novel mutations for human gametes μ= 20. If we use the highest evaluation μ= 50, then the size of the population for which number of frequent SNPs is 2.9 million dropped to 10 million people. Since all GEMA simulations gave equal chances for all virtual individuals in mating schemes and the number of offspring was the same for each virtual individual, the size of the population should be equal to the effective size N=N_{eff}. In several independent estimations of N_{eff} for humans, this number is around 10^4 significantly lower than 10^7 (Charlesworth 2009; Hartl 2007; Takahata, et al. 1995). Since everybody agrees that population size of modern humans is much higher than archaic humans, it is unlikely that numerous frequent SNPs arrived from the sole ancestral population, which effective size must be around 10 million. An alternative scenario for the creation of multiple frequent SNPs is the admixture of subpopulations that were separated for hundreds of thousands of years (see Discussion).

We examined the abundance of CHs and mutually exclusive Yin-Yang pairs in the GEMA modeling under different conditions (Table 3). This table demonstrates that selection pressure (***α***), population size (***N***), and meiotic recombination rate (***r***) considerably influence the distribution of SNPs and formation of CHs in populations (see also Figure 7). Many GEMA experiments demonstrate that Yin-Yang CHs are 5-10 times less abundant than in nature (compare Table 2 vs. Table 3). One of the most important parameters that stimulate the creation of abundant CHs and Yin/Yang pairs is the meiotic recombination rate (***r***), which should be low. On the other hand, the increase of the population size (***N***) causes a significant decrease of the abundance of CHs and the Yin/Yang pairs (Figure 7C and Table 3). Due to the limitation of RAM in our Linux workstations, we were unable to increase the size of populations above ***N***=2000 in our computational modeling experiments. However, if we extrapolate the results of our trends in Figure 7A and Table 3, then for the ***N***≥1,000,000 there should be practically no CHs or Yin/Yang pairs. Our computer simulations demonstrate that the observation in modern humans of a high number of frequent SNPs (2.9×10^6), together with an abundance of CHs (85% segments) and Yin/Yang pairs (56% of segments), could not originate from a single homogeneous population.

Table 1- 1: Distribution of Common Haplotypes in the Human Genome

Table 1. Distribution of Common Haplotypes in the Human Genome

CHR	Segment	Starting Position	Haplotype Length (KB)	Total SNPs in Haplotype	# Common Haplotypes	Occurrence of Common Haplotypes	Max Diff Common Haplotypes	# Yin - Yang
CHR_1	1	30923	771	1382	1	166	NA	0
CHR_1	2	808223	44	649	1	348	NA	0
CHR_1	3	1302106	19	298	3	1847	47	2
CHR_1	4	1806647	53	719	4	1499	48	2
CHR_1	5	2302471	49	780	4	702	47	2
CHR_1	6	2802348	36	743	6	1308	43	0
CHR_1	7	3302745	24	439	5	1419	46	0
CHR_1	8	3803755	202	1022	1	121	NA	0
CHR_1	9	4302585	51	874	5	1595	49	2
CHR_1	10	4802513	28	511	6	1478	47	2
CHR_1	11	5302118	6	145	4	1692	46	0
CHR_1	12	5802376	79	1344	4	897	30	0
CHR_1	13	6302510	60	785	3	1278	49	2
CHR_1	14	6802171	24	332	4	1731	48	2
CHR_1	15	7302754	22	385	4	1492	43	0
CHR_1	16	7803891	52	731	7	1286	48	2
CHR_1	17	8304607	47	784	4	1411	49	4
CHR_1	18	8808185	119	1394	5	1592	47	2
CHR_1	19	9302942	34	664	5	1252	40	0
CHR_1	20	9814964	173	2423	3	703	10	0

Table 1 presents segment-wise distribution of Common Haplotypes along the whole human genome. Common haplotypes are defined as those which occur at least 100 times or more in the 1092 individuals. Segment length is the distance between the coordinates of the first and 50th SNPs with frequency >= 0.25. Starting position of each segment has been provided and segment length has been shown in Kb. Haplotype pairs which differ in 47 or more loci (out of 50) has been defined as Yin –Yang haplotypes.

Table 1- 2: Abundance of segments with Yin, Yang and Mosaic CHs

This table presents an overall summary of the investigated chromosomal segments resulting from the analyses performed with different sets of SNPs according to their minor allele frequency (MAF threshold, shown in column 1). While column 2 shows total number of segments obtained in each experiment (see M&M for illustration), columns 3, 4, and 6 presents the number of segments with CHs, Yin/Yang haplotypes and Mosaic haplotypes respectively. Column 5 gives the percentage of total segments having Yin/Yang haplotypes.

MAF threshold	**Total #Seg**	**#Seg with ≥2 CHs**	**#Seg with Yin/Yang**	**%Segs with Yin/Yang**	**# Seg with Mosaic CHs**
>0.1	5425	5259	491	9	228
>0.2	5408	5174	2125	39	1278
>0.25	5398	5097	3024	56	1720
>0.3	5380	4946	3622	67	1574

Table 1- 3: Dynamics and arrangement of SNPs in GEMA simulations

Column four represents the total number of SNPs in the population of virtual individuals. Column five demonstrates the number of frequent SNPs (MAF >25%) in the same modeling population. Column six represents percentages of segments that have one or more CHs (with frequency >=5% in the modeling population), while last column – percentages of segments with Yin/Yang CHs.

Parameters			***R e s u l t s***			
N	***r***	***α***	# SNP x 10^3	#Freq SNP	% seg CHs	% seg Y/Y
124	48	5	50	6070	85	6.9
250	48	2	270	42333	98	12.5
250	48	3	146	16092	82	4.6
250	48	5	103	9644	73	5.5
250	48	10	80	6682	65	4.9
250	24	5	93	7045	96	25.4
250	12	5	82	5255	99	49.2
250	6	5	73	3863	100	69.0
500	48	5	193	12754	49	3.0
500	48	10	160	9109	55	2.6
1000	48	5	407	17724	35	2.0
2000	48	5	802	24897	25	1.1

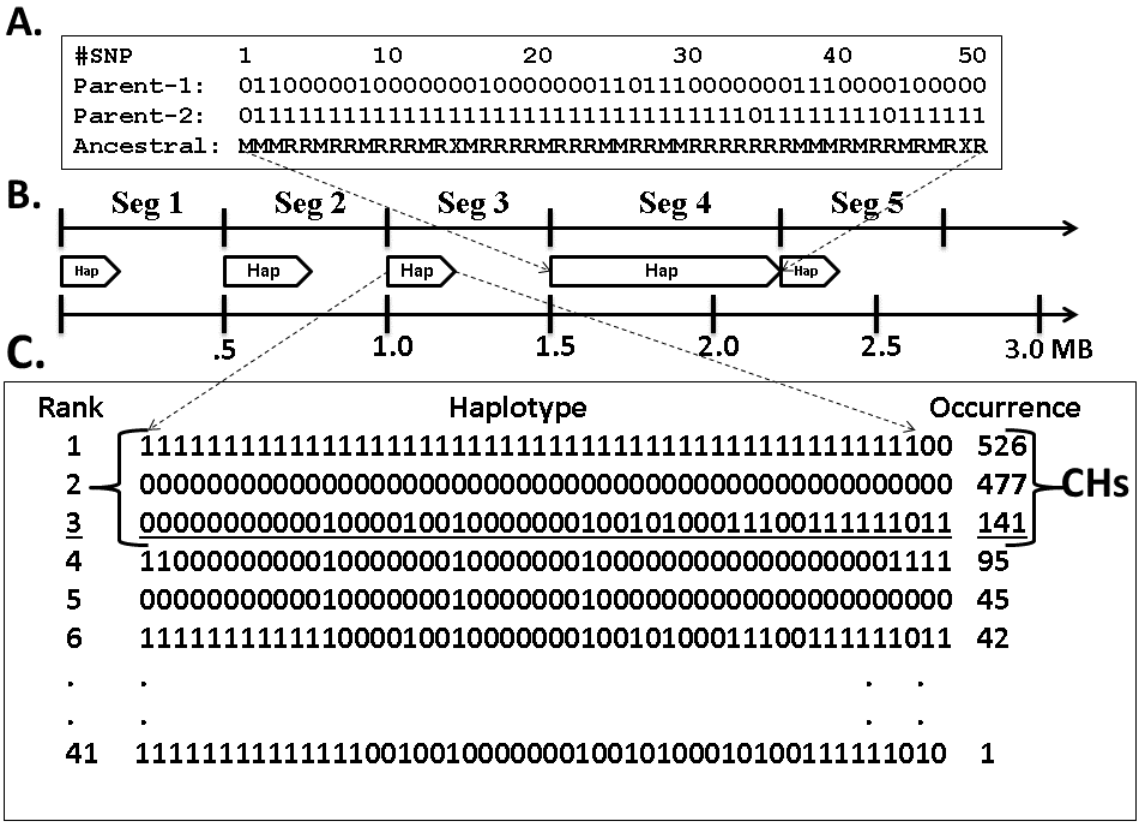


Figure 1- 1: Haplotype construction and characterization

A. Example of two parental haplotypes from segment 12 on chromosome 4 of CEU_NA07357 individual from 1000 Genomes. Following the 1000 Genomes Project, "0" means the presence of a reference allele, while "1" means an alternative allele in the haplotype. Only frequent genetic variants (with minor allele frequency >25%) have been used to construct haplotypes. In the last "Ancestral" line, "R" means that the reference allele is ancestral, "M" means that alternative (mutant) allele is ancestral, and "X" means unknown ancestral/derived status for the genetic variant in the 1000 Genomes dataset. Information about every genetic variant (identifier, location, alleles) and every haplotype are available from the Supplementary Data File S1. **B)** Chromosomes have been divided into segments of equal length (500 Kb). From the beginning of each segment, 50 adjacent high-frequency genetic variants have been selected for

construction of the haplotypes. When less than 50 frequent GV were present inside the segment, this segment was elongated until a full-length haplotype with 50 genetic variants was complete (see Seg 4). **C)** All haplotypes within a segment from 1092 individuals were grouped and ranked by the number of occurrences. Haplotypes that had been counted 100 or more times were named as common haplotypes (CHs). On **1.1 C** three common haplotypes exist and are shown above the solid line.

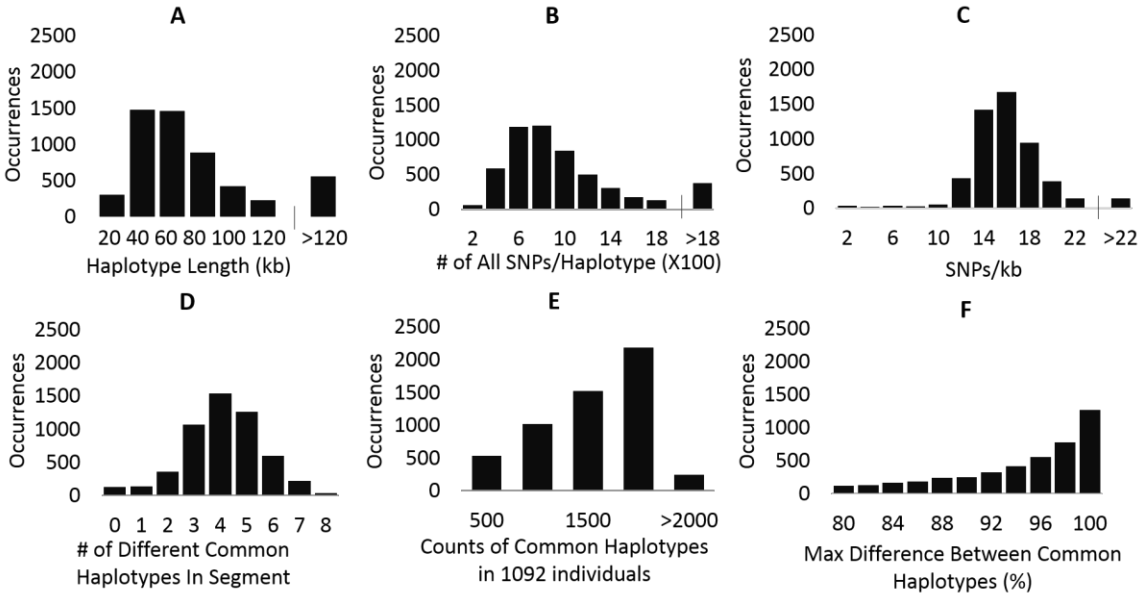


Figure 1- 2: Properties of haplotypes of frequent genetic variants in the human genome

"Occurrences" on the vertical axis represents the number of segments that have specific characteristic shown on the horizontal axis. A) Distribution of haplotype length (kb) in segments. B) Distribution of total number of genetic variants per haplotype. Horizontal axis shown in multiplication by 100. C) Density of all genetic variants in a haplotype. D) Number of CH groups per segment. E) Counts of all CHs in 1092 individuals in a segment. F) Maximum allele differences between all CH groups within a segment.

Figure 1- 3: An example of Yin, Yang, and Mosaic haplotypes and a Denisovan diplotype from the segment 102 of chromosome 1

The alleles that match the human reference genome are shown as "0", while the alternative alleles as "1". The ancestral alleles are shown in black, and the derived ones are shown in red. Blue and yellow highlights demonstrate pieces of Yin (blue) and Yang (yellow) segments from which the Mosaic haplotype can be reconstructed. This Mosaic haplotype is constructed from 14 pieces and has 12 derived alleles. Ten Mosaic derived alleles (83%) match the Yin haplotype and only two Mosaic derived alleles match the Yang haplotype. The Denisovan diplotype is shown in the last row. For the diplotype "0" means that both parental alleles match the human reference genome, "2" means that both alleles match alternative alleles, "1" means that this ancestral Denisovan person is heterozygous at this allele, and "x" means that this allele is unresolved. The heterozygous status for our frequent genetic variants (MAF>25%) is very rare for the Denisovan sequenced individual, so his homozygous diplotypes can be converted to haplotypes by the substitution of "2"s for "1"s.

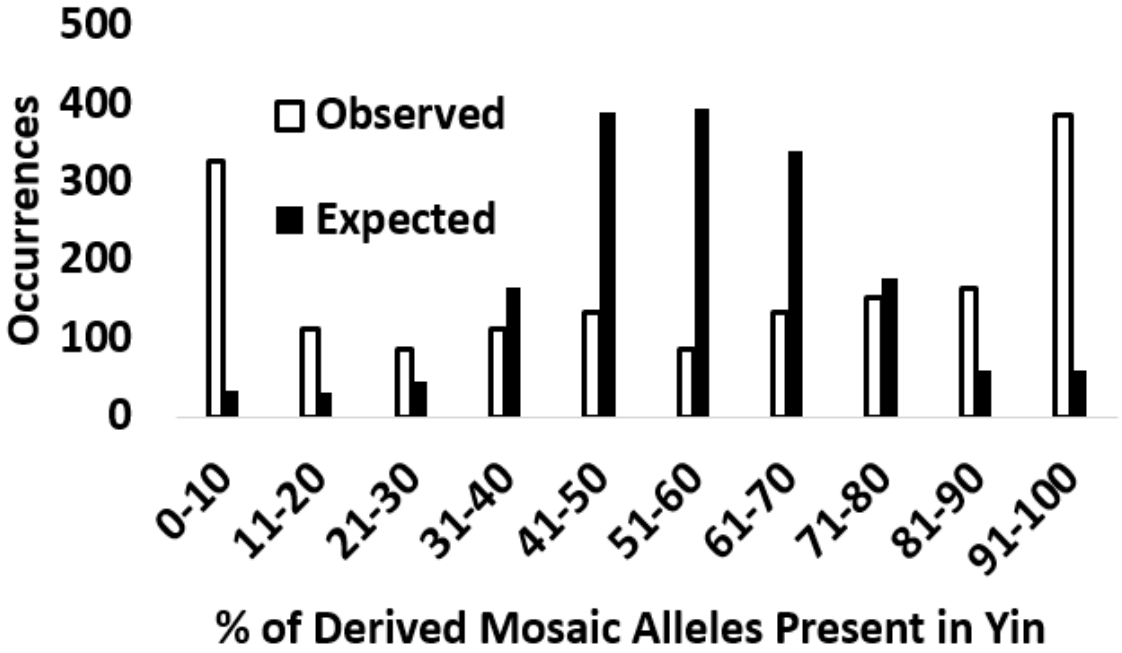


Figure 1- 4: % of Derived Mosaic Alleles Present in Yin

The observed data show 59% of cases where derived mosaic alleles primarily (>= 80%) came from either Yin or Yang. On the other hand, the expected dataset shows a normal distribution of Yin and Yang. The expected dataset was computationally created by randomly assigning each derived mosaic allele to Yin or Yang. In the observed data, 54.6% of mosaic derived alleles came from Yin, while 43.8% of mosaic derived alleles came from Yang. These percentages were used for the random assignment of derived mosaic alleles in the expected dataset.

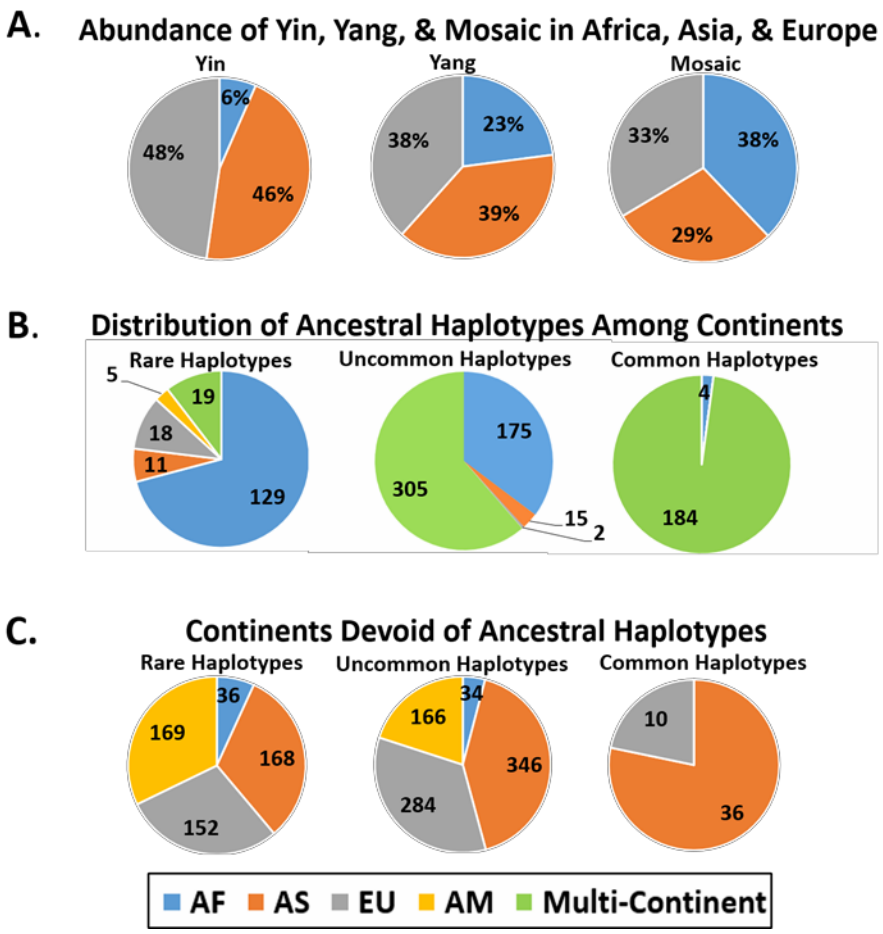


Figure 1- 5: Predominant occurrence of the Common Haplotypes (Yin, Yang, and Mosaic) among the African, Asian, and European populations.

Occurrences of Yin, Yang, and Mosaic haplotypes were computed on each continent and then normalized (see M & M) to account for the uneven population sizes from the different continents. Predominance was determined by the highest normalized occurrence of the respective common haplotype in a segment.

B and C Abundance of ancestral haplotypes in the continents. Rare, uncommon, and common haplotypes were determined by the number of matches to the ancestral haplotype out of 1092 individuals in a segment. Rare was classified as an ancestral haplotype, with only 1-3 matches in the segment, Uncommon was classified as 4-100 matches, and common was >100 matches. **B.** For continent

specificity, uncommon and rare haplotypes were defined as continent-specific if >90% of matches were found in a specific continent. Rare haplotypes were defined as continent-specific if 100% of matches were found in a specific continent. Multi-Continent means there was no continent specificity and matches to the ancestral haplotype were found on two or more continents. **C**. Figure B represents the continents where ancestral haplotypes are absent (shows less than 1% match) for all the three types of haplotypes i.e. Rare, Uncommon, and Common haplotypes.

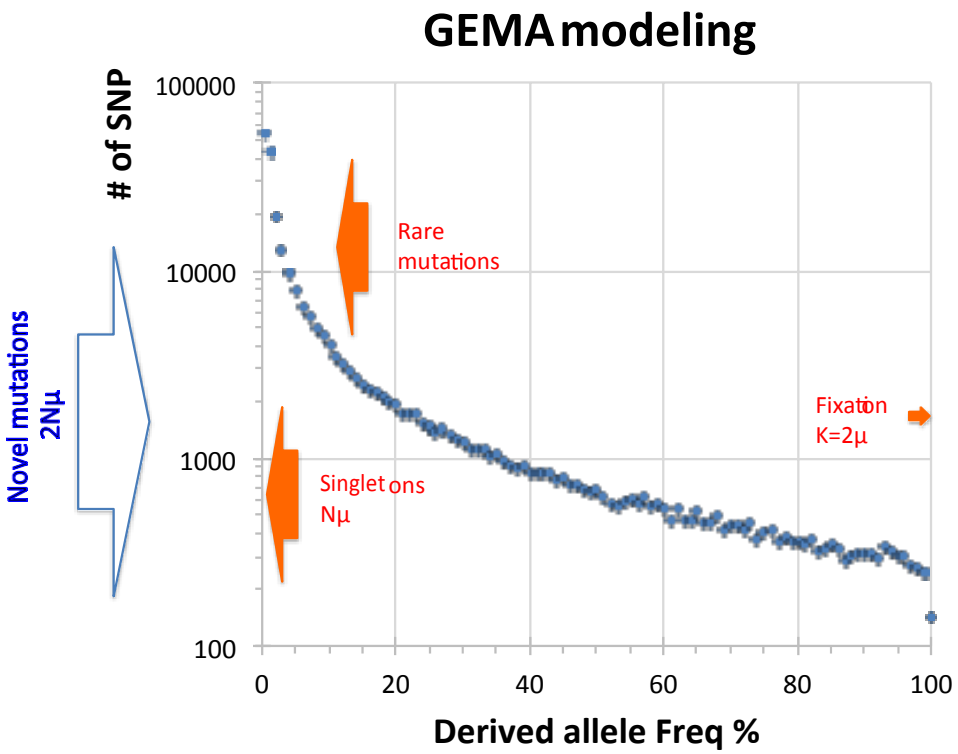


Figure 1- 6: Distribution of derived SNPs in GEMA simulations

Parameters for this computation were the following: M=20; a=2; N=250; r=48; h=0.5. Blue dots represent number of derived SNPs in the range of 1%.

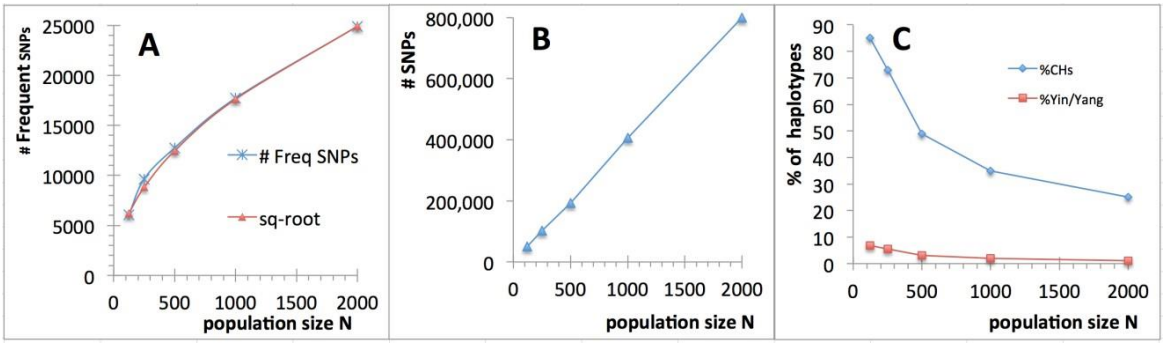


Figure 1- 7: Dependence of SNP number and CH occurrence on the population size (N) in GEMA experiments

A. Frequent SNPs with MAF > 25% are shown as blue stars. Red triangles show a square-root curve $\mathbf{c(N)}^{0.5}$, where **c** is a constant that approximates the GEMA modeling data (c=557). **B.** Number of all SNPs (rare and frequent) in the modeling populations. **C**. Percentage of segments that contain one or more CHs (blue line) and Yin/Yang pairs of CHs (red line).

1.6 Discussion

Humans possess 2,944,337 frequent SNPs (MAF>25%). This number is strikingly large. To get so many frequent SNPs inside an isolated single population, its effective size should be around ten million, as demonstrated by GEMA modeling (see Figure 7 and explanations in the Results). We also demonstrated that in modeling populations with large sizes, the arrival of mutually exclusive Yin/Yang CHs are very rare events. Since 56% of the investigated 5398 human loci have Yin/Yang CHs, special incidents must have happened during recent evolution to create these numerous mutually exclusive CHs. A straightforward possibility for the appearance of numerous Yin/Yang patterns is an admixture of two long-separated populations, which would also explain the observed large number of frequent SNPs.

Let's consider this hypothetical admixture and its consequences. According to Kimura's law, a population has $k=2\mu$ fixed mutations per generation, which does not depend on the population size (Kimura 1983). In humans, the value of k is around 100. To fix a million mutations, 10,000 generations are required, which roughly equals to 250,000 years (we assume 25 years per generation). Thus, after the admixture of two populations of comparable sizes that were separated from each other by 250,000 years, all mutations that had been fixed during their separation should become frequent SNPs. So, this proposed admixture should automatically convert two million recently fixed mutations in both populations into frequent SNPs, which, in addition, should be arranged as Yin/Yang CHs descended from two ancestral populations. (The actual number of frequent SNPs may be a little bit less if we assume that a fraction of the mutations that has been fixed are same

in both populations.) These estimations demonstrate that the observed number and arrangement of $2.9x10^6$ frequent SNPs in humans may have been created by a single "Great Admixture" of two major lineages that had been separated from each other around 300-500 thousand years.

Modern humans are widely spread across the globe and adapted to several diverse environments on different continents. In general, an admixture of different groups of people from different places should be beneficial overall and allow new combinations of various adaptations. For example, a Neanderthal EPAS1 allele is widespread in Tibetans and helps living in high altitudes (Huerta-Sanchez, et al. 2014). Other beneficial examples were recently reviewed in (Haber, et al. 2016). There were multiple well-known admixtures in recent human history, including peopling of New World by Europeans and Africans. Another recently discovered admixture occurred between Neanderthal people and archaic humans (Haber, et al. 2016; Prufer, et al. 2014). Importantly, this latter event did not create Yin/Yang CHs, since the number of Neanderthals was negligible compared to archaic humans and, thus, Neanderthals' recently fixed mutations were predominantly converted into rare SNPs. Presumably, such local admixtures were abundant in the recent human history and created a number of rare SNPs (Mondal, et al. 2016). For the conjectured "Great Admixture" of two ancestral populations named ***A*** and ***B***, their sizes should differ from each other by no more than three times to generate Yin/Yang CHs. Because Yin CHs have strong avoidance of Africa, (see Figure 5) it is reasonable to surmise that one of the ***A*** or ***B*** ancestral lineages should have evolved outside this continent and was a distant relative to the Neanderthals. At the same time, the prevalence of ancestral and Mosaic haplotypes in Africa supports

the possibility that another ancestral lineage had likely developed inside this continent. In our hypothetical scenario, ***A*** and ***B*** ancestral lineages are the primary sources for Yin/Yang CHs. The observed Mosaic CHs may be interpreted as lucky combinations of mutations in one of the ancestral ***A*** or ***B*** populations that have been beneficial to people and, hence, have been preserved for hundreds of thousands of years in the ancestral populations.

Is it possible to estimate the time of the hypothetical "Great Admixture" event? The Denisovan CHs give us a good reference point, which helps the assessment. The analyzed Denisovan CHs still possess 18% of derived frequent alleles present in modern humans, while Yin/Yang pairs share from 0 to 8% of derived alleles. Therefore, separation of two ancestral lineages ***A*** and ***B*** may be prior to the separation of archaic humans with Neanderthals. At first approximation, we place this separation of ***A*** and ***B*** lineages about 900-600 thousand years ago, while their "Great Admixture" may be placed roughly 300-100 thousand years ago.

1.7 Conclusions

Our results support the multi-regional theory of creation of modern people with multiple local admixtures with one “Great Admixture” event that generated most frequent SNPs and abundance of Yin/Yang CHs.

Dynamics and arrangement of genetic variants in modern humans represent very intricate patterns. Multiple parameters including selection pressure, meiotic recombination rates, and size of the population are very important in the analysis of these patterns. Advanced computer simulations, like GEMA, are extremely helpful in understanding SNP abundance and arrangement at the genomic scale.

1.8 References

Abecasis GR, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65. doi: 10.1038/nature11632

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome research 12: 1805-1814. doi: 10.1101/gr.631202

Al-Khudhair A, et al. 2015. Inference of distant genetic relations in humans using "1000 genomes". Genome biology and evolution 7: 481-492. doi: 10.1093/gbe/evv003

Altshuler DM, et al. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58. doi: 10.1038/nature09298

Armour JA, et al. 1996. Minisatellite diversity supports a recent African origin for modern humans. Nature genetics 13: 154-160. doi: 10.1038/ng0696-154

Auton A, et al. 2015. A global reference for human genetic variation. Nature 526: 68-74. doi: 10.1038/nature15393

Charlesworth B 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nature reviews. Genetics 10: 195-205. doi: 10.1038/nrg2526

Choudhury A, et al. 2014. Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. BMC genomics 15: 437. doi: 10.1186/1471-2164-15-437

Durand EY, Eriksson N, McLean CY 2014. Reducing pervasive false-positive identical-by-descent segments detected by large-scale pedigree analysis. Molecular biology and evolution 31: 2212-2222. doi: 10.1093/molbev/msu151

Duret L, Galtier N 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annual review of genomics and human genetics 10: 285-311. doi: 10.1146/annurev-genom-082908-150001

Fedorova L, Qiu S, Dutta R, Fedorov A 2016. Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes. Genome biology and evolution 8: 777-790. doi: 10.1093/gbe/evw034

Gabriel SB, et al. 2002. The structure of haplotype blocks in the human genome. Science 296: 2225-2229. doi: 10.1126/science.1069424

Guthery SL, Salisbury BA, Pungliya MS, Stephens JC, Bamshad M 2007. The structure of common genetic variation in United States populations. American journal of human genetics 81: 1221-1231. doi: 10.1086/522239

Haber M, Mezzavilla M, Xue Y, Tyler-Smith C 2016. Ancient DNA and the rewriting of human history: be sparing with Occam's razor. Genome biology 17: 1. doi: 10.1186/s13059-015-0866-z

Hartl DC, AG. 2007. Principles of population genetics. Sunderland, MA, USA: Sinauer Associates, Inc. Publishers.

Hinds DA, et al. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307: 1072-1079. doi: 10.1126/science.1105436

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proceedings of the National Academy of Sciences of the United States of America 92: 532-536.

Huerta-Sanchez E, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512: 194-197. doi: 10.1038/nature13408

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge, UK: Cambridge University Press.

Klyosov AA 2014. Reconsideration of the "Out of Africa" Concept as Not Having Enough Proof. Advances in Anthropology 4: 18-37.

Meyer M, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222-226. doi: 10.1126/science.1224344

Mierswa I, Wurst, M., Klinkenberg, R., Scholz, M., Euler, T. 2006. Rapid Prototyping for Complex Data Mining Tasks.

Mondal M, et al. 2016. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Nature genetics. doi: 10.1038/ng.3621

Prufer K, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43-49. doi: 10.1038/nature12886

Qiu S, Fedorov A 2015. Maruyama's allelic age revised by whole-genome GEMA simulations. Genomics 105: 282-287. doi: 10.1016/j.ygeno.2015.02.005

Qiu S, et al. 2014. Genome evolution by matrix algorithms: cellular automata approach to population genetics. Genome biology and evolution 6: 988-999. doi: 10.1093/gbe/evu075

Smith TaF, E. 2016. Statistical Genomics: Methods and Protocols. New York, NY, USA: Springer.

Stoneking M, Krause J 2011. Learning about human population history from ancient and modern genomes. Nature reviews. Genetics 12: 603-614. doi: 10.1038/nrg3029

Stringer CB, Andrews P 1988. Genetic and fossil evidence for the origin of modern humans. Science 239: 1263-1268.

Takahata N, Satta Y, Klein J 1995. Divergence time and population size in the lineage leading to modern humans. Theoretical population biology 48: 198-221. doi: 10.1006/tpbi.1995.1026

Tattersall I 2009. Out of Africa: modern human origins special feature: human origins: out of Africa. Proceedings of the National Academy of Sciences of the United States of America 106: 16018-16021. doi: 10.1073/pnas.0903207106

Wolpoff MH, J.; Caspari,R. 2000. Multiregional, Not Multiple Origins. AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 112: 129-136.

Zhang J, Rowe WL, Clark AG, Buetow KH 2003. Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. American journal of human genetics 73: 1073-1081. doi: 10.1086/379154

Zhu Q, et al. 2011. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. American journal of human genetics 88: 458-468. doi: 10.1016/j.ajhg.2011.03.008

Chapter 2

Determination of Early Human Intercontinental Migration from Genomic IBD segment Flow

2.1 Introduction

2.1.1 Increasing Availability of Large Scale Genomic Data

One of the main benefits of GWAS studies like the "1000 Genome Project" has been to provide detailed genetic data illustrating variations in 1092 separate individuals from different populations around the world. With the improvements in next- generation sequencing, we are now able to accomplish large scale genomic data collection at increasingly affordable rates. The availability of the data from these studies for public use encouraging studies in multiple fields of genomics and medicine. The "1000 Genome Project" provides extensive genomic data now being used in many studies which encompass structural genomics, functional genomics pharmacogenomics, population genomics etc. The obtained and constantly growing sequenced data from the 1000 genome project can be located from the project Consortium (Abecasis, et al., 2012).

Genetic studies of Y chromosome present evidence of very early admixture in the human genome, originating during the initial expansion of mankind within and out of Africa given the detected of African ancestry in non-African populations (Cruciani, et al., 2001). However, some recent genetic studies have also identified substantial non- African ancestry in population at the Horn of Africa (HOA), these admixtures are thought to have occurred in the past few 10000 years from admixture between Middle Eastern populations and HOA populations (Hodgson, Mulligan, Al-Meeri, & Raaum, 2013). All this would support the theory that there was some backward migration of nonlinguistic hunter gatherers back to Africa after the initial exit and before any more recent migrations (Henn, et al., 2010). Also, whole sequence data identifies modern Egyptians as the population whose genome information most closely resembles those of non-Africans populations leading to the conclusion that most early human migration took place via the northern route (Pagani, et al., 2015). Variability in the genomic data in different stratified populations, due to generations of genomic events like mutation accumulation and chromosomal recombination, create issues in estimation of population admixture periods and interpretation of distant relatedness between individuals. The advent of genomic computational methods that merge computer sciences and biology create faster and efficient ways of processing large genomic data and using computer programming algorithms created to account for multiple factors causing variability in the human genome. Areas of high chromosomal recombination are still problematic in experiments to trace human ancestry, so a focus on areas of the genome where recombination is minimal has been preferred. For example, a haploid sex

specific portion of the Y chromosome which is paternally transmitted and escapes recombination has been used to study population ancestry (Cruciani, et al., 2001). Also, more recent trends of globalization and increased migration between continents complicates the quest to determine ancient population migration patterns given very few populations unaffected by recent population migration and admixture.

2.1.2 Differences in Allele frequency between Populations

Surveys of databases of human polymorphisms reveal large allele frequency differences between continental regions with as much as 30% of loci showing very large differences between continents (Hofer, Ray, Wegmann, & Excoffier, 2009). As stated before, assessing ancestry is harder at SNP loci with higher recombination rates, and much easier at SNP loci with less recombination but large differences in allele frequency between ancestral continents and other admixed continents (Eyheramendy, Martinez, Manevy, Vial, & Repetto, 2014). Allele frequency at most loci within the human genome differ between populations because of human demographic history and genetic drift (Bansal & Libiger, 2015). With emphasis on demographic history, population migration and subsequent generational admixtures, population ancestry can be determined with the use of Identity by descent (IBD) segments. These are small segments of the genome identically shared by distantly related individuals at one or more loci, IBD segments are characterized by rare variant clusters (RVCs) of 5 or more very rare genetic variants (vrGVs allelic frequency <2%) (Al-Khudhair, et al., 2015). Given pairs of individuals with identical copies RVCs of five or more adjacent vrGVs in the same locations, the

probability of these events occurring at random is 0.002^5, which is less than one in 10^{13}. To further investigate IBD segments, my objective is to determine the direction of flow of these IBD segments between continents, by analysis of continental allele frequency differences of SNPs located with IBD segments in pairs of related individuals. I use the 1000 genome data containing the continental allele frequencies of SNPs in conducting these experiments, as this allele frequency data is available for every continent instead of by regions or different populations within continents. This complicates the analysis of IBD segment flow between individual populations in the same continent or different continents. For example, analysis of unique migrations patterns between a western African population like YRI and an Eastern African population like LWK would be impossible with no allele frequency data available for individual populations. Given the absence of allele frequency data of singular populations and time constraints on computationally deriving these data we are only able to do intercontinental migration pattern analysis. So far other types of computational methods are available in analysis of ancestry and population structure using genomic data including: model based clustering methods like STRUCTURE, FRAPPE, ADMIXTURE, and principal component analysis methods. Given a fixed number of clusters (populations), K, these methods use an unsupervised clustering approach to simultaneously infer the allele frequencies associated with K clusters and estimates the relative contributions of the K clusters to everyone's ancestry (Bansal & Libiger, 2015). Our method of determining ancestral history is similar but it analyses only SNPs located within pairs of RVCs that have a

statistically significant difference in continental allele frequency in different continents where pairs of individuals are from.

As stated earlier even, though evolutionary signatures can be erased by recombination and mutational perturbations, the general geographical distribution of SNPs may show traces of ancestral migrations (JIN, et al., 1999). The use of shared IBD segments does not eliminate some of the problems caused by recombination but, because they are smaller segments of the chromosomal DNA, they are likely to fall outside points of high recombination. Also, because the segments appear identical in individuals from different continents we assume recombination should be less of a disrupting factor given the low chances of exact same recombination events occurring in different continents and leaving these regions identical.

2.1.3 Haplotype Analysis from shared IBD segments

Multiple studies focusing on biallelic sites at non-recombinant portions of the Y chromosome have used binary haplotypes combined with microsatellite polymorphisms to evaluate internal diversities and estimate coalescence ages of binary haplotypes, which are interpreted to determine population history (Luis, et al., 2004). In these studies, haplogroups with varying constituent haplotypes are used as markers given their occurrences in different populations to perform phylogeographic analysis which cluster similar haplotypes and intermediates in deducing haplotype relatedness (Cruciani, et al., 2004). Fifty SNP string Haplotypes help in the examination of long stretches of SNPs located within IBD segments of parallel genomes. However, when

dealing with diploid organisms like mammals, sequencing complications arise in inferring the location of specific alleles in specific copies of the same chromosome. This issue necessitates the understanding of the detailed structure of the genomic chromosome and the application of computational and statistical methods to create phasing in the sequencing process. Phasing determines co-location of alleles in respective chromosome copies and produces maternal and paternal strings of genomic data. Phased DNA sequences opens the avenue of analyzing parental sequences separately for the presence of reference and mutant allele copies of a SNP in a locus. We can determine if SNPs at loci are present in the form of homozygous reference, homozygous mutant allele (two copies of the reference or two copies of the mutant allele) or heterozygous (has a copy of the mutant and the reference). There are many error in sequencing phasing especially at sites of minor allele frequencies like the vrGVs we are studying. Our haplotype strings consist of '0's and '1's representing reference and mutant alleles respective. For this experiment, we are going to have to eliminate the phasing by adding up both parental haplotypes and getting a single string instead of two parental haplotypes

2.2 Objectives

Previously, our team demonstrated that IBD segments produced by population admixture over time are a means of determining distant relatedness. It was discovered that one or more RVCs comprising of 5 or more very rare genetic variants (vrGVs with a frequency of 0.2%) are found to be shared parallel genomic regions of individuals from

different populations. This led to the conclusion that these individuals are distantly related given the low probability that these rare events in the shared clusters occurred randomly (Fedorova, 2016). However, there are a lot of errors in "phasing" the sequencing genome into two parental chromosomes (Sharp, Kretzschmar, Delaneau, & Marchini, 2015). These phasing errors are especially high within vrGVs. Therefore, in Fedorova et al. the authors were unable to determine the direction of exchange of genetic materials between continents. The objective of this experiment is to determine origins of these shared IBD segments found in related individuals from different continents based on the differences in continental allelic frequencies of the frequent SNPs (MAF>20%) contained in shared IBD segments. We would also like to infer broader intercontinental population migration patterns based on the proportional differences in direction of movement of the shared IBD segments within continents.

2.3 Materials and Methods

2.3.1 Data sources

I used data from the 1000 Genomes Project that is available publicly through the ftp site (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/) (Abecasis, et al., 2012). Specifically, Variants Call Format (VCF) files version 4.1 that contained a total of 38.2 million single nucleotide polymorphisms (SNPs), 3.9 million short insertions/deletions, and 14,000 deletions for all the human chromosomes have been used

I also used data from the experiment done to derive and define identity by decent (IBD) between distantly related individuals (Fedorova, 2016). This experiment used the distribution pattern of very rare genetic variants (vrGVs) with minor allele frequencies <2% found in the 1000 Genomes data (Al-Khudhair, et al., 2015) to determine distant relatedness between individuals. Data containing complete lists of vrGVs was obtained computationally for all 1092 individuals from the raw sequenced data at the 1000 genome project site and saved with file names matching individual's identities. For example, a file name TSI_NA20813 would contain a complete list of vrGVs of the individual identified as TSI_NA20813. This experiment uses only vrGV data of African, Asian and European populations consisting the Luhya in West Kenya (LWK) and Yoruba in Ibadan, Nigeria (YRI) from Africa, Japanese in Tokyo (JPT) Chinese Han in Beijing (CHB) and Chinese Han South (CHS) from Asia, and the Toscani in Italy (TSI) from Europe. Also, I obtained information files (Fedorova, 2016) containing list of pairs of individuals from different populations of interest that possessed shared IBD regions. An example of such a file can be seen in Figure 2-1 which shows list of individuals from the LWK and TSI population that have shared RVC segments. These lists are named: info_LWK_CHB, info_LWK_CHS, info_LWK_JPT, info_YRI_CHB, info_YRI_CHS, info_YRI_JPT, info_LWK_TSI and info_YRI_TSI as shown in figure 2-1.

2.3.2 Derivation of IBD Segments

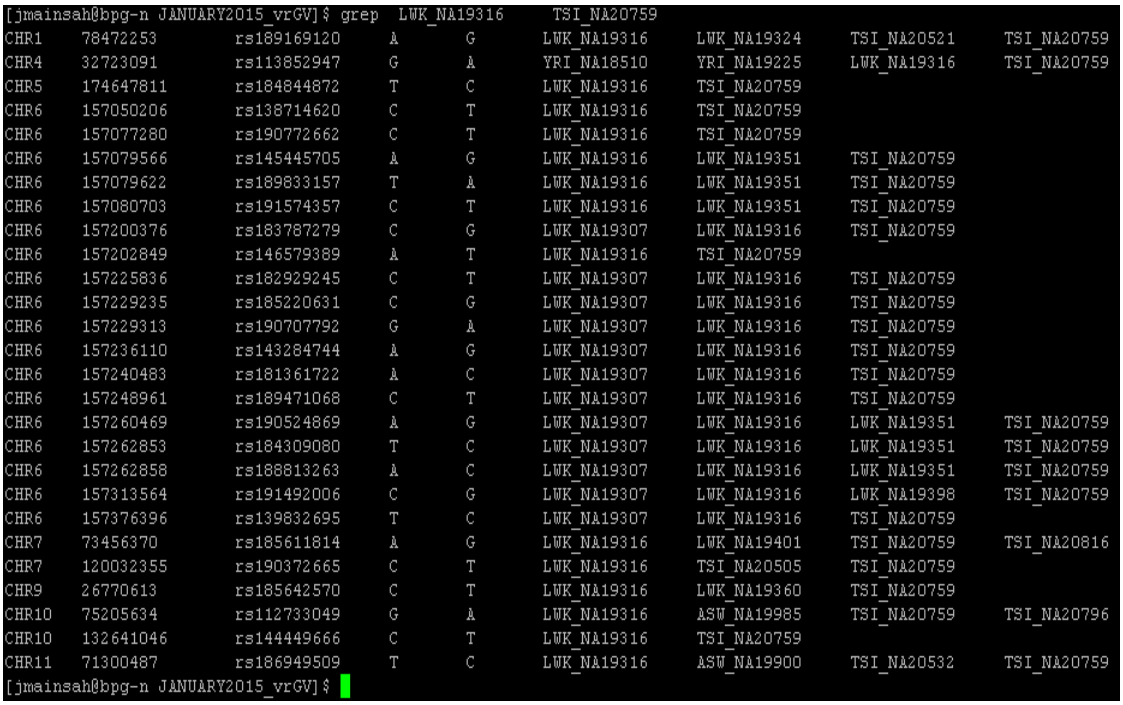
From the info_ files, pairs of individual from different continental populations that have shared RVC segments are listed with other information like chromosome

number from where the IBD segment is found, and number of vrGVs in the RVCs and the size of the RVCs as seen in Figure 2-1. These lists contain the size of the shared RVCs but do not contain the coordinates indicating the starting and ending locations of the IBD segment within chromosomes. To be able to get the exact coordinates of the IBD segments I used the GREP function in UNIX to match the files names like TSI_NA20813 from the info_ list that were shown to list individuals with shared IBD segments. The GREP function compares two file (files named individual's ID) containing vrGV data of two individuals and produces lines that match in the two files. From the matched lines, you can see cluster of vrGVs all located in a region about the size 100kb, these clusters are the IBD segments. From here I get the genomic coordinates and locations of these clusters of SNPs by getting the coordinate of the first and the last SNP from the cluster of SNPs.

CHR9	23593770	rs190097080	G	A	LWK_NA19315	TSI_NA20807	1	29126	6
CHR9	82293130	rs112183430	C	A	LWK_NA19315	TSI_NA20804	1	66562	6
CHR9	82293130	rs112183430	C	A	LWK_NA19315	TSI_NA20542	1	66562	6
CHR19	17950999	rs186380640	C	T	LWK_NA19315	TSI_NA20800	1	191037	6
CHR4	124300428	rs113498578	G	A	LWK_NA19316	TSI_NA20758	1	29378	5
CHR5	107662794	rs182259732	T	C	LWK_NA19316	TSI_NA20766	1	455652	6
CHR5	127218743	rs191158363	G	T	LWK_NA19316	TSI_NA20518	1	217941	8
CHR6	154127983	rs111827359	T	C	LWK_NA19316	TSI_NA20507	1	59770	26
CHR6	156671274	rs185220250	C	T	LWK_NA19316	TSI_NA20759	1	30497	5
CHR14	82644589	rs183344625	G	T	LWK_NA19316	TSI_NA20799	1	343984	6
CHR15	81450505	rs190882570	T	G	LWK_NA19316	TSI_NA20533	1	142659	5
CHR21	19278653	rs140629385	T	C	LWK_NA19316	TSI_NA20801	1	50780	10
CHR1	106034150	rs182500820	T	A	LWK_NA19317	TSI_NA20797	1	93442	7
CHR1	111552611	rs181241023	C	T	LWK_NA19317	TSI_NA20543	1	171548	10
CHR2	154497902	rs190373457	T	A	LWK_NA19317	TSI_NA20785	1	362578	8
CHR3	52605988	rs143628294	A	G	LWK_NA19317	TSI_NA20510	1	116836	6
CHR4	17401593	rs148466099	A	T	LWK_NA19317	TSI_NA20810	1	64166	5
CHR5	87556953	rs182521005	A	G	LWK_NA19317	TSI_NA20529	1	59564	5
CHR5	87556953	rs182521005	A	G	LWK_NA19317	TSI_NA20797	1	59564	5
CHR8	52172454	rs186768008	A	G	LWK_NA19317	TSI_NA20796	1	217471	5
CHR9	127720029	rs142242123	C	T	LWK_NA19317	TSI_NA20586	1	889053	9
CHR10	29421685	rs143015096	A	C	LWK_NA19317	TSI_NA20801	1	67153	8
CHR12	103104322	rs188658827	T	C	LWK_NA19317	TSI_NA20800	1	51939	6
CHR1	46663034	rs181093407	T	C	LWK_NA19318	TSI_NA20515	1	109042	8
CHR1	46663034	rs181093407	T	C	LWK_NA19318	TSI_NA20537	1	109042	8
CHR2	106845886	rs192163009	T	C	LWK_NA19318	TSI_NA20786	1	103215	5
CHR4	29785751	rs188156703	A	G	LWK_NA19318	TSI_NA20522	1	528643	9
CHR5	17696566	rs148508430	A	T	LWK_NA19318	TSI_NA20512	1	44260	11
CHR5	67411715	rs188591447	T	C	LWK_NA19318	TSI_NA20586	1	35891	6
CHR10	21351013	rs150712173	G	C	LWK_NA19318	TSI_NA20524	1	139157	6
CHR14	46530977	rs118033397	C	A	LWK_NA19318	TSI_NA20588	1	374725	5
CHR14	78847878	rs150506091	A	G	LWK_NA19318	TSI_NA20585	1	392890	31

Figure 2- 1: Section of info_LWK_TSI file

Column 6 & 7 are the identities of two individuals with a shared IBD segment, column 1 is the chromosome number of the IBD segment, column 8 is the number of IBD segments shared by the two individuals, column 9 is the size of the IBD segment and column 10 is the number of vrGVs found within this IBD segment



```
[jmainsah@bpg-n JANUARY2015_vrGV]$ grep  LWK_NA19316     TSI_NA20759
CHR1    78472253        rs189169120     A       G       LWK_NA19316     LWK_NA19324     TSI_NA20521     TSI_NA20759
CHR4    32723091        rs113852947     G       A       YRI_NA18510     YRI_NA19225     LWK_NA19316     TSI_NA20759
CHR5    174647811       rs184844872     T       C       LWK_NA19316     TSI_NA20759
CHR6    157050206       rs138714620     C       T       LWK_NA19316     TSI_NA20759
CHR6    157077280       rs190772662     C       T       LWK_NA19316     TSI_NA20759
CHR6    157079566       rs145445705     A       G       LWK_NA19316     LWK_NA19351     TSI_NA20759
CHR6    157079622       rs189833157     T       A       LWK_NA19316     LWK_NA19351     TSI_NA20759
CHR6    157080703       rs191574357     C       T       LWK_NA19316     LWK_NA19351     TSI_NA20759
CHR6    157200376       rs183787279     C       G       LWK_NA19307     LWK_NA19316     TSI_NA20759
CHR6    157202849       rs146579389     A       T       LWK_NA19316     TSI_NA20759
CHR6    157225836       rs182929245     C       T       LWK_NA19307     LWK_NA19316     TSI_NA20759
CHR6    157229235       rs185220631     C       G       LWK_NA19307     LWK_NA19316     TSI_NA20759
CHR6    157229313       rs190707792     G       A       LWK_NA19307     LWK_NA19316     TSI_NA20759
CHR6    157236110       rs143284744     A       G       LWK_NA19307     LWK_NA19316     TSI_NA20759
CHR6    157240483       rs181361722     A       C       LWK_NA19307     LWK_NA19316     TSI_NA20759
CHR6    157248961       rs189471068     C       T       LWK_NA19307     LWK_NA19316     TSI_NA20759
CHR6    157260469       rs190524869     A       G       LWK_NA19307     LWK_NA19316     LWK_NA19351     TSI_NA20759
CHR6    157262853       rs184309080     T       C       LWK_NA19307     LWK_NA19316     LWK_NA19351     TSI_NA20759
CHR6    157262858       rs188813263     A       C       LWK_NA19307     LWK_NA19316     LWK_NA19351     TSI_NA20759
CHR6    157313564       rs191492006     C       G       LWK_NA19307     LWK_NA19316     LWK_NA19398     TSI_NA20759
CHR6    157376396       rs139832695     T       C       LWK_NA19307     LWK_NA19316     TSI_NA20759
CHR7    73456370        rs185611814     A       G       LWK_NA19316     LWK_NA19401     TSI_NA20759     TSI_NA20816
CHR7    120032355       rs190372665     C       T       LWK_NA19316     TSI_NA20505     TSI_NA20759
CHR9    26770613        rs185642570     C       T       LWK_NA19316     LWK_NA19360     TSI_NA20759
CHR10   75205634        rs112733049     G       A       LWK_NA19316     ASW_NA19985     TSI_NA20759     TSI_NA20796
CHR10   132641046       rs144449666     C       T       LWK_NA19316     TSI_NA20759
CHR11   71300487        rs186949509     T       C       LWK_NA19316     ASW_NA19900     TSI_NA20532     TSI_NA20759
[jmainsah@bpg-n JANUARY2015_vrGV]$
```

Figure 2- 2: Matched two files containing vrGV data of two individuals

Above shows a list of pairs of individuals from LWK and TSI that have shared IBD segments. We took the files containing the vrGV data of each pair and used the GREP function in UNIX to get the lines that match in both files. This image shows the matched lines of individuals LWK_NA19316 & TSI_NA20759. From the matched lines, we can get the coordinates of clusters of vrGVs

2.3.3 Derivation of Haplotypes from IBD segments

With the RVC coordinates I ran a Perl program called *Haplo_find_Final_useit.pl* originally written by a lab collogue (Rajib Dutta) which I modified. This program generates haplotype strings of SNPs found within the coordinate regions. It computes the genomic SNP raw data of all 1092 individuals of the 1000 genome project and generates phased parental haplotypes strings of 50 alleles of zeroes and ones with zero being the reference allele and one being the mutant when referenced to the genome assembly GRCh37.p13. The program functions with varying parameters which include chromosome number, desired allele frequency (0.2), desired string length of haplotypes generated (50), the desired threshold occurrence of haplotypes counts (100), beginning coordinate and ending coordinate of the desired IBD region.

The different sizes of the IBD segments creates a challenge when trying to extract haplotypes from these segments. We computationally extract haplotypes from 1092 individuals' sequenced data by inputting specific genomic coordinates of IBD segments within which haplotypes are created. The program creates haplotype strings of 50 SNPs found within these IBD regions but some IBD regions have more than 50 SNPs present within the range while others don't have up to 50 SNPs present in them. The program creates multiple haplotype strings for large IBD segments but if the number of SNPs in an IBD segment is not an exact multiple of 50 then some SNPs at the end of the range will not be formed into haplotypes. Also, if an IBD segment is shorter and does not have up to 50 SNPs then I have to manually change a parameter of the program which

determines the length of the haplotype string. Since it's impossible to exactly estimate the number of SNPs in an IBD segment based on the IBD segment size, some SNPs at the back end are lost during the haplotype building process

The *Haplo_find_Final_useit.pl* program generates files named ALL_STATS_C1-1 the C1 represents the chromosome number 1 and the digit at the end of the file name increases if more files are generated because more than one haplotype string of 50 SNPs is produced with an IBD segment. These files contain information of all the 50 reference SNPs in each haplotypes region including SNP coordinates and IDs, the reference and the mutant allele, allele frequency in all continents etc. The output files also contain varying haplotype strings of all the 1092 individuals and picks identical haplotypes that have an occurrence >100 and list them and their number of occurrences at the bottom of the files. I wrote another program called *AF_STATS.pl* which extracts from the *Haplo_find_Final_useit.pl* program output files only the haplotypes of the two individuals with shared IBD segments from where these haplotypes were produced. This program also extracts the continental allele frequencies of each SNP in the 50-string haplotype. It extracts and writes the allele frequencies of each SNP in the African, Asian and European continents in adjacent columns to the column with the parental haplotype string of 50 SNPs

2.3.4 Elimination of phasing

The next task was to eliminate phasing because the strings of generated haplotypes come in pairs of maternal and paternal strings for everyone. Also, as

previously stated in the introduction, the phased data comes with too many errors so we must minimize the use of phased genomic data. I wrote a Perl program called *add_haps.pl* which simply summed up the maternal and paternal haplotypes for everyone instead of having two haplotype strings of zeroes and ones for each person representing the reference and the mutant allele, I got one haplotype string consisting of zeroes, ones and twos. Zeroes represent homozygous reference allele; ones represent heterozygous allele and twos represent homozygous mutant allele.

2.3.5 Determination of Migration of SNPs between Confinements

Lastly I compare two haplotypes derived from identical shared RVC regions of pairs of individuals from different continents to determine direction of SNP flow within continents. Since the allele frequencies of SNPs for each country or population were not available, I used the allele frequencies of SNP's for entire continents. I compare continental allele frequencies of each haplotype SNPs for pairs of individuals from different continents and if I found a combination of a heterozygous allele with a frequency of <0.10 in continent "A" compared to a homozygous mutant allele with a frequency of >0.10 in continent "B", the inference was that the SNP originated in continent "B" and migrated to continent "A" through generations of admixture, see figure 2-3. This process was completed comparing allele frequency and their actual occurrence in related individuals from all IBD region haplotypes for the LWK-CHB, LWK- CHS, LWK-JPT, YRI-CHB, YRI-CHS, YRI-JPT, LWK-TSI and YRI-TSI populations.

This experiment included all the haplotypes derived from identical shared RVCs regions of individuals from the 3 Asian and 2 African populations and the 2 African populations and 1 European population. Numbers of IBD segments found between pairs population individuals are as follows: 53 pairs from LWK and CHB, 75 pairs from LWK and CHS, 41 pairs from LWK and JPT, 21 pairs from YRI and CHB, 25 pairs from YRI and CHS, and 9 pairs from YRI and JPT. In all there were 224 shared IBD segments between Africa and Asia populations. As for the African and European cases, due to time constraints and a much larger number of shared IBD segments, I could only sample about 228 IBD segments, 100 between YRI and TSI and 128 between LWK and TSI. I found useful data for a total of 30 segments: 17 between the LWK and TSI and 13 between YRI and TSI populations. These were taken randomly from a total of 1515 shared IBD segment found between pairs of individuals from both African populations and TSI, the lone European population we examined. Those data are shown in figure 2-3 and the tables that follow

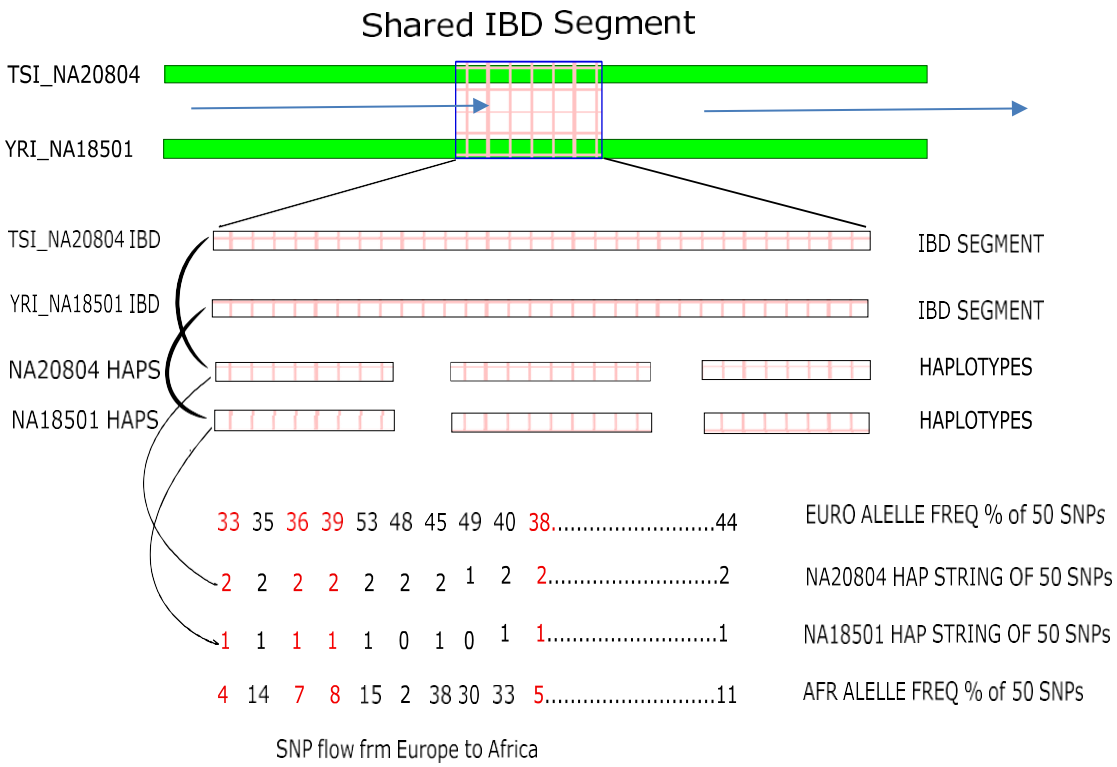


Figure 2- 3: Illustration of SNP flow within a shared IBD segment

The shaded are represents regions of shared rare variant clusters between two distantly related individuals. Below we have two haplotype strings from the RVC regions and the allele frequencies of the various SNPs in different continents

2.4 Results

2.4.1 Shared IBD Segments

We analyzed regions of genomic similarities between pairs of individuals from different continents using DNA sequences available from the "1000 Genome project". Our analysis includes multiple intercontinental sets of pairs of these individuals that shared IBD segments in all 22 chromosomes. Table 2-1 and 2-2 below each illustrates the results of analysis of haplotype strings derived from IBD segments shared by individuals from TSI a European population and YRI an African population. Table 1A shows haplotype strings of two individuals (TSI_NA20804 from Europe and YRI_NA18501 from Africa) with a set of highlighted SNPs that originated in Europe and migrated to Africa, we can deduce this direction of movements by looking at the allele frequencies of SNPs in one population compared to that frequency in the opposing population. On the other hand, Table 1B illustrates our analysis of an example that shows two haplotype strings (TSI_NA20509 from Europe and YRI_NA18516 from Africa) that would indicate that some highlighted SNPs in these haplotype strings originated from Africa and migrated to Europe.

Table 2- 1: Analysis showing flow of SNPs into Africa

Column 1 represents the numbered string of 50 SNPs that make up a haplotype, column 2 represents the chromosome number, column 3 represents the SNP coordinate, column 4 represents the SNP Identity, column 5 represents the allele frequency of that SNP in the European continent, column 6 represents the allele frequency of that SNP in the African continent and columns 7 and 8 represent the strings of 50 SNPs that make up haplotypes from a European and African individual respectively. The parental haplotypes of the individuals TSI_NA20804 and YRI_NA18501 can be referenced from the genome assembly GRCh37.p13 used by the 1000 genome project. From the haplotype strings "0" represents a homozygous reference allele, "1" represents a heterozygous mutant allele and "2" represents the presence of a homozygous mutant allele. The string of "0" "1" and "2" are in columns 7 and 8 in the figure above under individual identity. The blue shading indicates the flow of SNPs from Europe to Africa based on the differences in allele frequency and occurrence of these SNPs in the different individuals

Analysis of haplotype illustrating Migration of SNPs from Europe to Africa							
##	CHR#	SNP_CORD	SNP_ID	EUR_AF	AFR_AF	TSI_NA20804	YRI_NA18501
1	3	6560149	rs62246890	0.29	0.03	1	0
2	3	6563176	rs2034870	0.18	0.54	0	1
3	3	6573711	rs201128309	0.20	0.45	1	2
4	3	6573712	rs56176212	0.22	0.45	1	1
5	3	6577915	rs2220412	0.49	0.67	2	1
6	3	6586574	rs62246893	0.32	0.07	1	0
7	3	6596774	rs9857596	0.19	0.45	1	2
8	3	6598777	rs9864202	0.18	0.45	1	2
9	3	6601604	rs62244291	0.34	0.03	1	0
10	3	6604066	rs4684534	0.37	0.38	2	1
11	3	6605904	rs113888426	0.27	0.18	1	0
12	3	6606566	rs56247305	0.35	0.19	2	1
13	3	6607353	rs62244293	0.34	0.03	1	0
14	3	6607813	rs6790573	0.35	0.19	2	1
15	3	6610068	rs2129907	0.53	0.36	2	1
16	3	6610126	rs6808436	0.33	0.06	2	1
17	3	6610161	rs6802701	0.35	0.14	2	1
18	3	6611701	rs17216153	0.35	0.14	2	1
19	3	6611841	rs17216160	0.37	0.17	2	1
20	3	6612231	rs73020619	0.34	0.06	2	1
21	3	6612345	rs62244294	0.36	0.14	2	1
22	3	6613189	rs6792359	0.53	0.37	2	1
23	3	6614747	rs1909378	0.37	0.35	2	1
24	3	6614780	rs9637445	0.35	0.14	2	1
25	3	6614980	rs155277	0.86	0.73	2	2
26	3	6615370	rs201233776	0.35	0.14	2	1
27	3	6615371	rs150337529	0.35	0.13	2	1
28	3	6616306	rs6765070	0.35	0.13	2	1
29	3	6620310	rs72084002	0.37	0.35	1	1
30	3	6620418	rs60124944	0.34	0.06	1	0
31	3	6621373	rs62244295	0.35	0.19	2	1
32	3	6621821	rs144363035	0.33	0.08	2	1
33	3	6621917	rs9844273	0.49	0.37	2	1
34	3	6622421	rs57903852	0.34	0.05	2	1
35	3	6622849	rs17216167	0.34	0.06	2	1
36	3	6623137	rs59493683	0.35	0.17	2	2
37	3	6623268	rs57351878	0.35	0.13	2	1
38	3	6624353	rs62244297	0.34	0.07	2	1
39	3	6625357	rs17288317	0.34	0.06	2	1
40	3	6626081	rs55853239	0.35	0.14	2	1
41	3	6626230	rs7616876	0.52	0.70	2	2
42	3	6626624	rs7619627	0.52	0.39	2	1
43	3	6627001	rs62244300	0.35	0.13	2	1
44	3	6627177	rs62244301	0.34	0.06	2	1
45	3	6627275	rs62244302	0.34	0.04	1	0
46	3	6627411	rs201978272	0.52	0.43	2	2
47	3	6627414	.	0.34	0.28	1	1
48	3	6627498	rs9838727	0.52	0.37	2	1
49	3	6628526	rs2323808	0.52	0.48	2	2
50	3	6628636	rs1603870	0.34	0.06	2	1

Table 2- 2: Analysis showing flow of SNPs out of Africa

Column 1 represents the numbered string of 50 SNPs that make up a haplotype, column 2 represents the chromosome number, column 3 represents the SNP coordinate, column 4 represents the SNP Identity, column 5 represents the allele frequency of that SNP in the European continent, column 6 represents the allele frequency of that SNP in the African continent and columns 7 and 8 represent the strings of 50 SNPs that make up haplotypes from a European and African individual respectively. The parental haplotypes of the individuals TSI_NA20509 and YRI_NA18516 can be referenced from the genome assembly GRCh37.p13 used by the 1000 genome project. From the haplotype strings "0" represents a homozygous reference allele, "1" represents a heterozygous mutant allele and "2" represents the presence of a homozygous mutant allele. The string of "0" "1" and "2" are in columns 7 and 8 in the figure above under individual identity. The yellow shading indicates the flow of SNPs from Africa to Europe based on the differences in allele frequency and occurrence of these SNPs in the different individuals

Analysis of haplotype illustrating Migration of SNPs from Africa to Europe							
##	CHR#	SNP_CORD	SNP_ID	EUR_AF	AFR_AF	TSI_NA20509	YRI_NA18516
1	17	68711379	rs7226235	0.38	0.45	2	1
2	17	68711729	rs12939443	0.36	0.33	1	0
3	17	68711999	rs7211118	0.36	0.35	1	0
4	17	68712132	rs7215997	0.35	0.34	1	0
5	17	68712791	rs8069488	0.40	0.73	2	1
6	17	68713728	rs5821796	0.38	0.40	2	1
7	17	68714096	rs740674	0.37	0.30	1	0
8	17	68715520	rs8065001	0.09	0.50	1	2
9	17	68715575	rs8065129	0.09	0.50	1	2
10	17	68715988	rs8069523	0.09	0.50	1	2
11	17	68716927	rs9907805	0.09	0.50	1	2
12	17	68717038	rs2109050	0.15	0.04	0	0
13	17	68718188	rs717419	0.47	0.81	2	2
14	17	68718734	rs4793501	0.53	0.82	2	2
15	17	68719000	rs9912421	0.09	0.47	1	2
16	17	68719293	rs8082019	0.09	0.49	1	2
17	17	68721887	rs1024641	0.47	0.79	2	2
18	17	68722140	rs1024642	0.09	0.50	1	2
19	17	68722746	rs1024643	0.10	0.49	1	2
20	17	68723707	rs16976178	0.09	0.49	1	2
21	17	68724036	rs929474	0.62	0.89	2	2
22	17	68724110	rs7213980	0.09	0.46	1	2
23	17	68724495	rs1962801	0.15	0.05	1	1
24	17	68725134	rs1029755	0.09	0.46	2	2
25	17	68726544	rs9912396	0.46	0.80	1	2
26	17	68727070	rs9898253	0.09	0.46	1	2
27	17	68727863	rs10669456	0.46	0.79	2	2
28	17	68727865	rs200009741	0.41	0.67	1	2
29	17	68728276	rs4141187	0.09	0.46	2	2
30	17	68729064	rs8069427	0.09	0.46	1	2
31	17	68731219	rs9907514	0.47	0.81	1	2
32	17	68734558	rs8077813	0.10	0.46	1	1
33	17	68735692	rs7220992	0.58	0.89	1	1
34	17	68736046	rs9891034	0.09	0.50	0	0
35	17	68737418	rs2058177	0.09	0.50	2	2
36	17	68739326	rs9903749	0.37	0.45	0	0
37	17	68739330	rs9909710	0.27	0.20	0	0
38	17	68740369	rs67171306	0.55	0.40	2	2
39	17	68740819	rs7220722	0.74	0.85	2	2
40	17	68744268	rs12451960	0.20	0.27	0	0
41	17	68744362	rs12451982	0.20	0.13	0	0
42	17	68745923	rs8069222	0.59	0.54	2	2
43	17	68746372	rs12936311	0.58	0.54	2	2
44	17	68746415	rs200565362	0.28	0.21	0	0
45	17	68746417	rs35495600	0.53	0.47	1	0
46	17	68746483	rs2367007	0.74	0.81	2	2
47	17	68748959	rs4793318	0.20	0.15	0	0
48	17	68752466	rs142502375	0.21	0.19	0	0
49	17	68755732	rs2109052	0.20	0.12	0	0
50	17	68756931	rs36118145	0.48	0.32	1	0

Table 2- 3: Flow of IBD segments between YRI and TSI populations

Column 1 is the chromosome number, column 2 is the identity of the individual from YRI, column 3 is the identity of the individual from TSI, column 4 indicates with a "+" if the haplotypes analysis of the IBD segment shared by a pair of individual from YRI and TSI showed movement of SNPs from Europe to Africa and lastly column 5 indicates with a "+" if the haplotype analysis of the IBD segment shared by these two individuals shows movement of SNPs from Africa to Europe.

Flow of IBD segments between YRI and TSI populations					
Chr#	YRI	TSI	EU->AF	AF->EU	IBD Segment
17	YRI_NA18516	TSI_NA20509	-	+	68711035-68836465
2	YRI_NA19121	TSI_NA20816	-	+	70003352-70167097
3	YRI_NA18501	TSI_NA20804	+	-	6503228-6686230
13	YRI_NA18489	TSI_NA20517	+	-	21914157-22177066
8	YRI_NA18520	TSI_NA20535	+	-	12535145-12624981
4	YRI_NA19190	TSI_NA20533	+	-	144655658-144755514
10	YRI_NA19121	TSI_NA20828	-	+	32721902-33175261
1	YRI_NA19130	TSI_NA20504	-	+	89189589-89705331
3	YRI_NA19137	TSI_NA20774	+	-	179695533-179758058
12	YRI_NA19152	TSI_NA20502	+	-	73497509-73514238
9	YRI_NA19172	TSI_NA20585	+	-	79667181-79818794
11	YRI_NA19175	TSI_NA20513	-	+	38066832-38436964
6	YRI_NA19200	TSI_NA20796	-	+	120585541-121567012

Table 2- 4: Flow of IBD segments between LWK and TSI populations

Column 1 is the chromosome number, column 2 is the identity of the individual from LWK, column 3 is the identity of the individual from TSI, column 4 indicates with a "+" if the haplotypes analysis of the IBD segment shared by a pair of individual from LWK and TSI showed movement of SNPs from Europe to Africa and lastly column 5 indicates with a "+" if the haplotype analysis of the IBD segment shared by these two individuals shows movement of SNPs from Africa to Europe.

Flow of IBD segments between LWK and TSI populations					
Chr#	LWK	TSI	EU->AF	AF->EU	IBD Region
14	LWK_NA19318	TSI_NA20585	-	+	79273509-79666399
14	LWK_NA19318	TSI_NA20588	-	+	46753884-47128609
3	LWK_NA19328	TSI_NA20765	-	+	5584578-5783892
14	LWK_NA19044	TSI_NA20786	-	+	79377142-80889440
14	LWK_NA19044	TSI_NA20512	-	+	83897021-83942568
20	LWK_NA19473	TSI_NA20768	-	+	46504686-46538466
21	LWK_NA19036	TSI_NA20800	-	+	45372079-45589632
22	LWK_NA19355	TSI_NA20811	-	+	46598399-46801733
8	LWK_NA19028	TSI_NA20504	+	-	53222683-53774741
11	LWK_NA19035	TSI_NA20790	+	-	114584987-114735696
4	LWK_NA19046	TSI_NA20798	+	-	116660331-116937255
12	LWK_NA19374	TSI_NA20508	+	-	121189207-121379814
2	LWK_NA19396	TSI_NA20525	-	+	137446047-137877652
4	LWK_NA19397	TSI_NA20797	-	+	135129272-135249394
7	LWK_NA19398	TSI_NA20797	+	-	8014677-8108633
5	LWK_NA19399	TSI_NA20757	+	-	85013757-85411617
11	LWK_NA19401	TSI_NA20807	-	+	48271045-48482864

Table 2- 5: Flow of IBD segments between LWK and CHS populations

Column 1 is the chromosome number, column 2 is the identity of the individual from LWK, column 3 is the identity of the individual from CHS, column 4 indicates with a "+" if the haplotypes analysis of the IBD segment shared by a pair of individual from LWK and CHS showed movement of SNPs from Asia to Africa and lastly column 5 indicates with a "+" if the haplotype analysis of the IBD segment shared by these two individuals shows movement of SNPs from Africa to Asia.

Flow of IBD segments between LWK and CHS populations				
Chr#	LWK	CHS	AS->AF	AF->AS
3	LWK_NA19036	CHS_HG00707	-	-
3	LWK_NA19036	CHS_HG00449	-	-
1	LWK_NA19041	CHS_HG00614	-	+
6	LWK_NA19310	CHS_HG00525	-	-
11	LWK_NA19312	CHS_HG00689	-	-
8	LWK_NA19313	CHS_HG00479	-	-
11	LWK_NA19313	CHS_HG00611	-	-
10	LWK_NA19324	CHS_HG00651	-	-
10	LWK_NA19324	CHS_HG00443	-	-
6	LWK_NA19327	CHS_HG00500	-	+
6	LWK_NA19327	CHS_HG00446	-	-
13	LWK_NA19331	CHS_HG00404	-	-
2	LWK_NA19332	CHS_HG00707	-	-
2	LWK_NA19334	CHS_HG00707	-	-
13	LWK_NA19334	CHS_HG00404	-	-
3	LWK_NA19355	CHS_HG00584	-	-
3	LWK_NA19355	CHS_HG00577	-	-
1	LWK_NA19359	CHS_HG00614	-	-
8	LWK_NA19359	CHS_HG00418	-	-
17	LWK_NA19360	CHS_HG00537	-	++
8	LWK_NA19371	CHS_HG00418	-	-
2	LWK_NA19372	CHS_HG00684	-	-
14	LWK_NA19373	CHS_HG00590	-	-
3	LWK_NA19374	CHS_HG00458	-	-
6	LWK_NA19374	CHS_HG00457	-	-
6	LWK_NA19376	CHS_HG00500	-	-
6	LWK_NA19376	CHS_HG00446	-	-
13	LWK_NA19379	CHS_HG00684	-	-
13	LWK_NA19379	CHS_HG00436	-	-
10	LWK_NA19381	CHS_HG00566	-	-
10	LWK_NA19381	CHS_HG00557	-	+
10	LWK_NA19382	CHS_HG00566	-	-
10	LWK_NA19382	CHS_HG00557	+	-
1	LWK_NA19384	CHS_HG00427	-	-
1	LWK_NA19384	CHS_HG00418	-	-
12	LWK_NA19390	CHS_HG00692	-	-
17	LWK_NA19393	CHS_HG00537	-	-
7	LWK_NA19399	CHS_HG00684	-	-
7	LWK_NA19399	CHS_HG00699	-	-

11	LWK_NA19404	CHS_HG00584	-	-
4	LWK_NA19436	CHS_HG00684	-	-
2	LWK_NA19438	CHS_HG00657	-	-
10	LWK_NA19439	CHS_HG00651	-	+
10	LWK_NA19439	CHS_HG00443	-	-
2	LWK_NA19443	CHS_HG00707	-	-
11	LWK_NA19443	CHS_HG00437	-	-
13	LWK_NA19443	CHS_HG00404	-	-
3	LWK_NA19445	CHS_HG00650	-	-
12	LWK_NA19446	CHS_HG00692	-	-
18	LWK_NA19446	CHS_HG00672	-	++
2	LWK_NA19448	CHS_HG00406	-	+
20	LWK_NA19448	CHS_HG00708	-	-
2	LWK_NA19449	CHS_HG00704	-	-
2	LWK_NA19449	CHS_HG00442	-	-
14	LWK_NA19449	CHS_HG00464	-	-
18	LWK_NA19449	CHS_HG00592	-	-
5	LWK_NA19455	CHS_HG00524	-	-
5	LWK_NA19455	CHS_HG00512	-	+
6	LWK_NA19455	CHS_HG00592	-	-
8	LWK_NA19455	CHS_HG00449	-	-
6	LWK_NA19457	CHS_HG00566	-	-
11	LWK_NA19457	CHS_HG00611	-	-
14	LWK_NA19462	CHS_HG00464	-	-
5	LWK_NA19466	CHS_HG00683	-	+
14	LWK_NA19467	CHS_HG00418	-	-
6	LWK_NA19468	CHS_HG00590	-	-
21	LWK_NA19469	CHS_HG00705	-	-
13	LWK_NA19470	CHS_HG00404	-	-
3	LWK_NA19472	CHS_HG00650	-	-
8	LWK_NA19472	CHS_HG00533	-	-
11	LWK_NA19472	CHS_HG00437	-	-
5	LWK_NA19473	CHS_HG00683	-	-
4	LWK_NA19474	CHS_HG00684	-	-
21	LWK_NA19474	CHS_HG00705	-	-

Table 2- 6: Flow of IBD segments between LWK and CHB populations

Column 1 is the chromosome number, column 2 is the identity of the individual from LWK, column 3 is the identity of the individual from CHB, column 4 indicates with a "+" if the haplotypes analysis of the IBD segment shared by a pair of individual from LWK and CHB showed movement of SNPs from Asia to Africa and lastly column 5 indicates with a "+" if the haplotype analysis of the IBD segment shared by these two individuals shows movement of SNPs from Africa to Asia.

FLOW IBD segments between LWK and CHB populations				
Chr#	LWK	CHB	AS->AF	AF->AS
12	LWK_NA19028	CHB_NA18602	-	+
6	LWK_NA19310	CHB_NA18543	-	-
6	LWK_NA19313	CHB_NA18553	-	-
2	LWK_NA19316	CHB_NA18740	-	-
1	LWK_NA19317	CHB_NA18536	-	-
2	LWK_NA19318	CHB_NA18637	-	++
9	LWK_NA19321	CHB_NA18606	-	-
1	LWK_NA19324	CHB_NA18566	-	-
3	LWK_NA19324	CHB_NA18592	-	-
13	LWK_NA19324	CHB_NA18567	-	-
12	LWK_NA19328	CHB_NA18545	-	-
4	LWK_NA19331	CHB_NA18549	-	-
6	LWK_NA19331	CHB_NA18553	-	-
7	LWK_NA19331	CHB_NA18597	-	-
13	LWK_NA19331	CHB_NA18616	-	-
7	LWK_NA19334	CHB_NA18597	-	-
13	LWK_NA19334	CHB_NA18616	-	-
2	LWK_NA19350	CHB_NA18618	-	+
12	LWK_NA19350	CHB_NA18549	-	-
1	LWK_NA19355	CHB_NA18536	-	-
10	LWK_NA19355	CHB_NA18595	-	-
17	LWK_NA19355	CHB_NA18610	-	+
1	LWK_NA19359	CHB_NA18639	-	-
10	LWK_NA19359	CHB_NA18595	-	-
2	LWK_NA19372	CHB_NA18550	-	+
7	LWK_NA19373	CHB_NA18560	-	-
7	LWK_NA19374	CHB_NA18560	-	-
6	LWK_NA19376	CHB_NA18606	-	-
6	LWK_NA19377	CHB_NA18574	-	+
4	LWK_NA19379	CHB_NA18541	-	-
6	LWK_NA19379	CHB_NA18574	-	-
12	LWK_NA19390	CHB_NA18647	-	-
10	LWK_NA19395	CHB_NA18647	-	-
1	LWK_NA19435	CHB_NA18637	-	-
3	LWK_NA19437	CHB_NA18535	-	-
6	LWK_NA19437	CHB_NA18641	-	-
3	LWK_NA19439	CHB_NA18553	-	++
4	LWK_NA19440	CHB_NA18541	-	+
13	LWK_NA19443	CHB_NA18616	-	-
6	LWK_NA19446	CHB_NA18626	-	-
1	LWK_NA19448	CHB_NA18645	-	-
13	LWK_NA19449	CHB_NA18567	-	-
4	LWK_NA19457	CHB_NA18542	-	-

2	LWK_NA19463	CHB_NA18618	-	+
3	LWK_NA19463	CHB_NA18553	-	+
4	LWK_NA19467	CHB_NA18558	-	-
8	LWK_NA19469	CHB_NA18574	-	-
8	LWK_NA19469	CHB_NA18564	-	-
13	LWK_NA19470	CHB_NA18616	-	-
12	LWK_NA19472	CHB_NA18740	-	++
6	LWK_NA19473	CHB_NA18612	-	+

Table 2- 7: Flow of IBD segments between LWK and JPT populations

Column 1 is the chromosome number, column 2 is the identity of the individual from LWK, column 3 is the identity of the individual from JPT, column 4 indicates with a "+" if the haplotypes analysis of the IBD segment shared by a pair of individual from LWK and JPT showed movement of SNPs from Asia to Africa and lastly column 5 indicates with a "+" if the haplotype analysis of the IBD segment shared by these two individuals shows movement of SNPs from Africa to Asia.

Flow of IBD segments between LWK and JPT populations				
Chr#	LWK	JPT	AS->AF	AF->AS
12	LWK_NA19028	JPT_NA18973	-	-
6	LWK_NA19307	JPT_NA18963	-	-
8	LWK_NA19311	JPT_NA18949	-	-
8	LWK_NA19311	JPT_NA18957	-	-
8	LWK_NA19313	JPT_NA18999	-	-
3	LWK_NA19324	JPT_NA19065	-	-
12	LWK_NA19328	JPT_NA18985	-	-
7	LWK_NA19331	JPT_NA18947	-	-
1	LWK_NA19332	JPT_NA19000	-	+
2	LWK_NA19332	JPT_NA19083	-	-
2	LWK_NA19334	JPT_NA19083	-	-
7	LWK_NA19334	JPT_NA18947	-	+
12	LWK_NA19347	JPT_NA19058	-	-
12	LWK_NA19350	JPT_NA19077	-	-
1	LWK_NA19355	JPT_NA19000	-	-
2	LWK_NA19374	JPT_NA18952	-	-
10	LWK_NA19375	JPT_NA18968	-	-
10	LWK_NA19375	JPT_NA19063	-	-
4	LWK_NA19381	JPT_NA19088	-	-
1	LWK_NA19384	JPT_NA19000	-	-
10	LWK_NA19385	JPT_NA18977	-	-
2	LWK_NA19393	JPT_NA18966	-	-
5	LWK_NA19394	JPT_NA19002	-	+
18	LWK_NA19394	JPT_NA19058	-	-
10	LWK_NA19395	JPT_NA18977	-	-
5	LWK_NA19404	JPT_NA18939	-	-
13	LWK_NA19429	JPT_NA18957	-	-
2	LWK_NA19443	JPT_NA19083	-	-
5	LWK_NA19443	JPT_NA18973	-	-
2	LWK_NA19449	JPT_NA19056	-	-
4	LWK_NA19457	JPT_NA18952	-	-
4	LWK_NA19457	JPT_NA18976	-	-
6	LWK_NA19457	JPT_NA18984	-	-
14	LWK_NA19467	JPT_NA18956	+	-
14	LWK_NA19467	JPT_NA19070	-	-
11	LWK_NA19469	JPT_NA18977	-	-
20	LWK_NA19473	JPT_NA19084	-	-
2	LWK_NA19474	JPT_NA18952	-	-
11	LWK_NA19474	JPT_NA19080	-	-
11	LWK_NA19474	JPT_NA19059	-	-
11	LWK_NA19474	JPT_NA18943	-	-

Table 2- 8: Flow of IBD segments between YRI and CHB populations

Column 1 is the chromosome number, column 2 is the identity of the individual from YRI, column 3 is the identity of the individual from CHB, column 4 indicates with a "+" if the haplotypes analysis of the IBD segment shared by a pair of individual from YRI and CHB showed movement of SNPs from Asia to Africa and lastly column 5 indicates with a "+" if the haplotype analysis of the IBD segment shared by these two individuals shows movement of SNPs from Africa to Asia.

Flow of IBD segments between YRI and CHB populations				
Chr#	YRI	CHB	AS->AF	AF->AS
11	YRI_NA18498	CHB_NA18596	-	-
4	YRI_NA18502	CHB_NA18549	-	-
8	YRI_NA18502	CHB_NA18628	-	-
6	YRI_NA18510	CHB_NA18749	-	-
13	YRI_NA18511	CHB_NA18567	-	+
11	YRI_NA18516	CHB_NA18596	-	-
17	YRI_NA18516	CHB_NA18573	-	-
18	YRI_NA18517	CHB_NA18616	-	-
6	YRI_NA18858	CHB_NA18532	-	-
1	YRI_NA18910	CHB_NA18634	-	-
1	YRI_NA18933	CHB_NA18634	-	-
3	YRI_NA18933	CHB_NA18627	-	-
6	YRI_NA19093	CHB_NA18527	-	-
6	YRI_NA19118	CHB_NA18553	-	-
1	YRI_NA19129	CHB_NA18637	-	-
6	YRI_NA19130	CHB_NA18606	-	-
13	YRI_NA19131	CHB_NA18567	-	+
9	YRI_NA19137	CHB_NA18606	-	-
6	YRI_NA19149	CHB_NA18527	-	-
8	YRI_NA19150	CHB_NA18602	-	-
8	YRI_NA19150	CHB_NA18559	-	-

Table 2- 9: Flow of IBD segments between YRI and CHS populations

Column 1 is the chromosome number, column 2 is the identity of the individual from YRI, column 3 is the identity of the individual from CHS, column 4 indicates with a "+" if the haplotypes analysis of the IBD segment shared by a pair of individual from YRI and CHS showed movement of SNPs from Asia to Africa and lastly column 5 indicates with a "+" if the haplotype analysis of the IBD segment shared by these two individuals shows movement of SNPs from Africa to Asia.

Flow of IBD segments between YRI and CHS populations				
Chr#	YRI	CHS	AS->AF	AF->AS
11	YRI_NA18508	CHS_HG00611	-	-
11	YRI_NA18522	CHS_HG00590	-	-
2	YRI_NA18861	CHS_HG00671	-	-
2	YRI_NA18868	CHS_HG00534	-	-
7	YRI_NA18916	CHS_HG00684	-	-
7	YRI_NA18916	CHS_HG00699	-	-
9	YRI_NA18923	CHS_HG00671	+	-
6	YRI_NA18924	CHS_HG00590	-	-
17	YRI_NA18924	CHS_HG00451	+	-
7	YRI_NA19102	CHS_HG00651	-	-
4	YRI_NA19129	CHS_HG00556	-	-
11	YRI_NA19129	CHS_HG00584	-	-
6	YRI_NA19130	CHS_HG00500	-	-
6	YRI_NA19130	CHS_HG00446	-	-
4	YRI_NA19149	CHS_HG00428	-	-
4	YRI_NA19149	CHS_HG00427	-	-
11	YRI_NA19160	CHS_HG00689	-	-
2	YRI_NA19175	CHS_HG00534	-	-
4	YRI_NA19185	CHS_HG00422	-	-
1	YRI_NA19190	CHS_HG00427	-	-
1	YRI_NA19190	CHS_HG00418	-	-
14	YRI_NA19190	CHS_HG00590	-	-
1	YRI_NA19223	CHS_HG00705	-	-
18	YRI_NA19223	CHS_HG00651	-	-
18	YRI_NA19223	CHS_HG00672	-	-

Table 2- 10: Flow of IBD segments between YRI and JPT populations

Column 1 is the chromosome number, column 2 is the identity of the individual from YRI, column 3 is the identity of the individual from JPT, column 4 indicates with a "+" if the haplotypes analysis of the IBD segment shared by a pair of individual from YRI and JPT showed movement of SNPs from Asia to Africa and lastly column 5 indicates with a "+" if the haplotype analysis of the IBD segment shared by these two individuals shows movement of SNPs from Africa to Asia.

Flow of IBD segments between YRI and JPT populations				
Chr#	YRI	JPT	AS->AF	AF->AS
11	YRI_NA18486	JPT_NA18994	-	+
11	YRI_NA18486	JPT_NA18948	-	+
8	YRI_NA18502	JPT_NA19072	-	-
16	YRI_NA18516	JPT_NA19059	-	-
2	YRI_NA18861	JPT_NA18980	-	-
5	YRI_NA19099	JPT_NA18950	+	-
13	YRI_NA19171	JPT_NA18989	-	-
13	YRI_NA19171	JPT_NA18951	-	-
5	YRI_NA19198	JPT_NA19002	-	+

2.4.2 Uni-directionality flow of SNPs within shared IBD segments

During the analysis of SNP flow history between pairs we find that once a single SNP in an IBD segment showed movement from one continent to the other we never find another SNP in that same segment that showed an opposite direction of movement. This can be seen in all the tables above that present results of SNP flow between different continents populations. For example, we don't ever see a "+" sign in both SNP movement from Asia to Africa and Africa to Asia in the same IBD region. Though most of the IBD segments did not show any results for origin of SNPs, the ones that showed origin always had one direction only. This is further proof that IBD segments are affected by very little mutation and selection forces which could have potentially destroyed` ancestral history from the genome.

2.4.3 SNP origins and destinations between different continents

The general picture seen when we looked at the sum of direction of SNP movement between African with Asian populations and African with European populations presented different results. When looking at migration of SNPs between African and Asian populations we saw a very clear result that showed that 85% of the IBD segments migrated from Africa to Asia while 15% migrated from Asia to Africa as seen in Figure 2-

5. On the other hand, when looking at the movements of IBD segments between Africa and Europe we found a more mixed picture. There were 53% of SNPs that originated in

Africa and ended up in Europe and 47% of SNPs that originated in Europe that ended up in African populations.

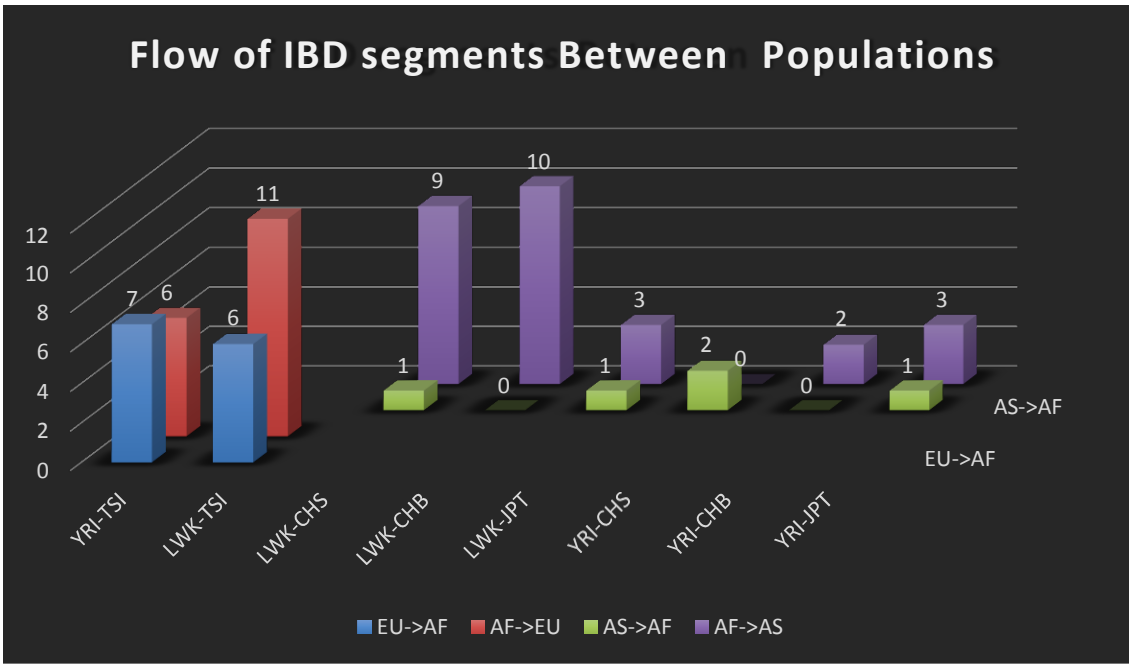


Figure 2- 4: Flow of IBD segments between continental Populations

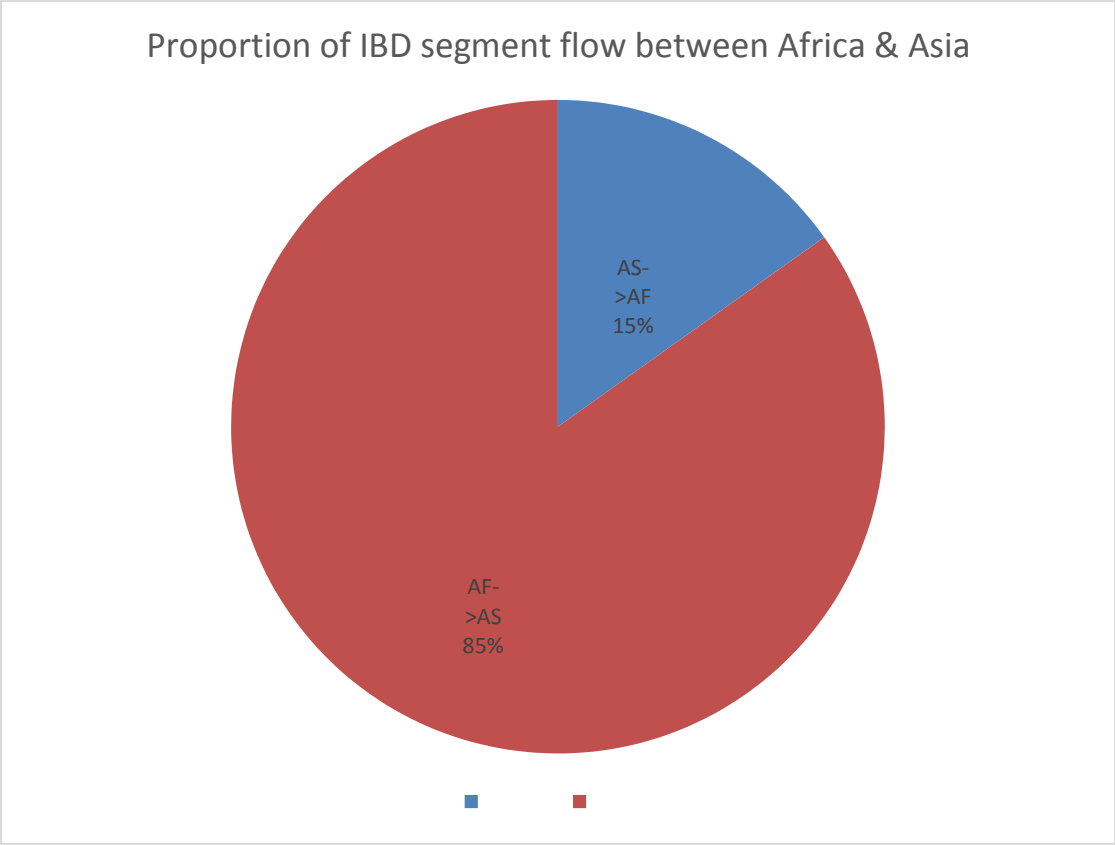


Figure 2- 5: % of IBD segment flow from Asia to Africa Vs Africa to Asia

The red shaded area represents the percentage flow of IBD segment from Africa to Asia while the blue shaded area represents the percentage flow from Asia to Africa. Data is statistically significant using Chi-Square test with the p-value 0.000014 because p-value is less than 0.05

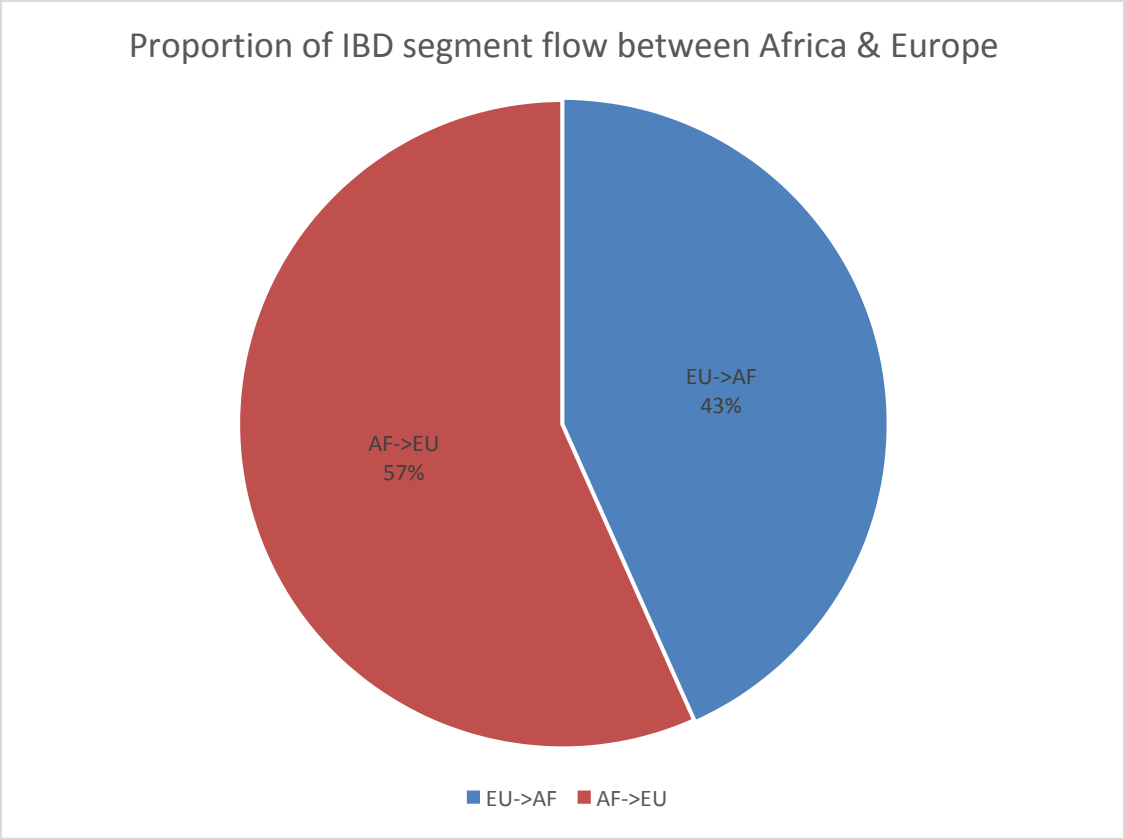


Figure 2- 6: % of IBD segment flow from Europe to Africa Vs Africa to Europe

The red shaded area represents the percentage flow of IBD segments from Africa to Europe while the blue shaded area represents the percentage flow from Europe to Africa. Data is not statistically significant using Chi-Square test with the p-value 0.13 because p-value is greater than 0.05

2.5 Discussions

We previously demonstrated that individuals from different continents possess identical short segments (IBD Segments) in their chromosomal DNA that proves they are distantly related. Due to population migration and admixture there is many genetic variation in populations like Latino and African American populations. Individuals in these populations have unique shared IBD segment with not just one but multiple different populations individuals from which they have ancestry for example the African American south west (ASW) and the Kenyan (LWK) populations have an average of 8.75 IBD segments per person shared with others from the other population (Fedorova, 2016). Trying genetic analyses of early migration patterns from more admixed populations is complex and ineffective using pairs of individuals; more homogenous populations like Asian and African populations are suitable. In previous experiments, we found that Asian populations had the smallest (median size 54kb) and the least number of shared IBD segments with other populations making it the perfect case study in deducing early origin of IBD segment between populations. The shortest sizes of these IBD segments are shared between the Asian and African populations implying they are the oldest admixtures given that with time IBD segments get smaller from subsequent admixture. More recent admixtures tend to show in larger IBD segments as evident between the African and European populations (Al-Khudhair, et al., 2015).

When attempting to determine origins of IBD segments between continents we analyzed the nature of occurrence of alleles from IBD segments found in the genomes of

individuals from different continental populations. The occurrence of SNPs could be homozygous, heterozygous or absent in diploid individuals depending on the frequencies of these alleles in different continents. We found that individuals in continents where SNP likely originated had a homozygous presence of these SNP alleles due to the higher frequency of these alleles in the continent of origin. On the other hand, individuals tended to have a heterozygous presence of the SNP in the continent where the SNP is less frequent and brought from original location by admixing. Per the Hardy-Weinberg proportions, if an allele is found in a continent at a frequency of 10% then the chance of an individual having two copies of this allele is simply the 10% percent chance for each of the two alleles squared is 0.01 which is a 1% chance out of 10,000. This is very evident from our results since these SNPs are mostly present in a heterozygous form in individuals from populations where they are (>10%) less frequent. Further if these same alleles are more frequent (<10%) in the other continent and they are present in a homozygous form in individuals then these alleles probably were transported from the continent where they are more frequent to a continent where they are less frequent (Gautier & Vitalis, 2012). We clearly see most clusters of alleles moving from Africa to Asia in our data from figure 2-5, even though there are several clusters moving from Asia to Africa. When analyzing the IBD segments between the African and European populations the result is a more mixed result because of recent migration and admixture. We see that most IBD segments originated from Africa but there is also substantial number of IBD segments that originated in Europe and end up in Africa. These results follow conventional reasoning from scientific anthropological

and historical evidence of population history given the human expansion out of Africa followed by events that support back migration, the more recent slave trade events and then more recent migration between population's continents (Li, et al., 2008).

A potential study area of IBD segments in determining population migration would be to distinguish the earlier migration directions with the more recent migrations directions. In our experiment, we examined the migration patterns between African European and Asian populations, but we did directional analysis as a whole which gave a mixed result of migration patterns. For example, we see evidence of migration from Africa to Europe and Europe to Africa with no time context. It could be useful to partition IBD segments based on their sizes to see a more detailed picture of time periods when the migrations occurred. Shorter IBD segments will show direction of earlier migration patterns while medium and longer IBD segments will show much later and very recent migration direction patterns between populations. Also, do analyzing all the populations could be useful to get the whole picture of global migration and not just between selected countries. Also, using data from the previous study conducted in chapter one I made the choice of using the major allele frequency (MAF) of > 20% when assembling SNPs from the human genomes into haplotypes. In chapter one we generated haplotypes using SNPs with allele frequencies greater than or equal to: 10%, 20%, 25% and 30% and from here we discovered the highest number of major haplotypes are produced at 20% and 25% allele frequencies.

2.6 Conclusion

Early and subsequent human migrations and admixture contribute to create a mosaic of chromosomal IBD segments in the genome (Watkins, et al., 2012). Within these segments the distribution and proportions of SNPs between different continents illustrates their flow, from continents of more frequency to other continents where they are found with less frequency. Distantly related people from different continents have shared IBD segments in their genomes made of clusters of vrGVs that are transferred between populations with migration and admixture. The SNPs within RVCs that originate from different continents are seen to have continental allelic frequencies >10% and are heterozygous, identical SNPs are homozygous and have continental allelic frequencies <10% in the continent they originate. The migration direction of these SNPs shows significantly more movement from African to Asian populations than the opposite. While there is more movement of SNPs from African to European populations the image is more mixed with a number SNPs showing opposite movement to African populations.

2.7 References

Abecasis, G., Auton, A., Brooks, L., DePristo, M., Durbin, R., Handsaker, R., Consortium, 1. G. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 56-65.

Al-Khudhair, A., Qiu, S., Wyse, M., Chowdhury, S., Cheng, X., Bekbolsynov, D., Fedorov, A. (2015). Inference of Distant Genetic Relations in Humans Using "1000 Genomes". *Genome Biology and Evolution*, 481-492.

Bansal, V., & Libiger, O. (2015). Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. *BMC Bioinformatics*, 1-11.

Cruciani, F., La Fratta, L. F., Santolamazza, P., Sellitto, D., Pascone, R., Moral, P., & Watson, E. (2004). Phylogeographic Analysis of Haplogroup E3b (E-M215) Y Chromosomes Reveals Multiple Migratory Events Within and Out Of Africa. *American Journal of Human Genetics*, 1014-1022.

Cruciani, F., Santolamazza, P., Underhill, P. A., Shen, P., Macauley, V., & Moral, P. (2001). A Back Migration from Asia to Sub-Saharan Africa Is Supported by High-Resolution Analysis of Human Y-Chromosome Haplotypes. *American Journal for Human Genetics*, 1197-1214.

Eyheramendy, S., Martinez, F. I., Manevy, F., Vial, C., & Repetto, G. M. (2014). Genetic structure characterization of Chileans reflects historical immigration patterns. *Nature Communications*, 1-10.

Fedorova, L. S. (2016). Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes. *Genome Biology And Evolution*, 777-790.

Gautier, M., & Vitalis, R. (2012). Inferring Population Histories Using Genome-Wide Allele Frequency Data. *Molecular Biology and Evolution*, 654-668.

Henn, B., Botigue, L., Gravel, S., Wang, W., Brisbin, A., Byrnes, J., . . . Zalloua, P. A. (2010). Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. *PLOS Genetics*, 1-11.

Hodgson, J. A., Mulligan, C. J., Al-Meeri, A., & Raaum, R. L. (2013). Early Back-to-Africa Migration into the Horn of Africa. *PLOS Genetics*, 1-17.

Hofer, T., Ray, N., Wegmann, D., & Excoffier, L. (2009). Large Allele Frequency Differences between Human Continental Groups are more Likely to have Occurred by Drift During range Expansions than by Selection. *Annals of Human Genetics* , 95-108.

Jin, L., Underhill, P. A., Doctor, V., Davis, R. W., Shen, P., Cavalli-Sforza, L., & Oefner, P. J. (1999). Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. *PNAS*, 3796-3800.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., & Cann, H. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science*, 1100-1104.

Luis, J., Rowold, D. J., Regueiro, M., Caeiro, B., Cinnioglu, C., Roseman, C., . . . Herrera, R.

J. (2004). The Levant versus the Horn of Africa: Evidence for Bidirectional Corridors of Human Migrations. *The American Society of Human Genetics*, 532- 544.

Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., & Xue, Y. (2015). Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. *The American Journal of Human Genetics*, 986-991.

Sharp, K., Kretzschmar, W., Delaneau, O., & Marchini, J. (2015). Phasing for medical sequencing using rare variants and large haplotype reference panels. *Oxford Journals*, 1974-1980.

Watkins, W., Xing, J., Huff, C., Witherspoon, D., Zhang, Y., Perego, U., Jorde, L. (2012). Genetic analysis of ancestry, admixture and selection in Bolivian and Totonac populations of the New World. *BMC Genetics*, 13-39.

References

Chapter 1

Abecasis GR, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65. doi: 10.1038/nature11632

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome research 12: 1805-1814. doi: 10.1101/gr.631202

Al-Khudhair A, et al. 2015. Inference of distant genetic relations in humans using "1000 genomes". Genome biology and evolution 7: 481-492. doi: 10.1093/gbe/evv003

Altshuler DM, et al. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58. doi: 10.1038/nature09298

Armour JA, et al. 1996. Minisatellite diversity supports a recent African origin for modern humans. Nature genetics 13: 154-160. doi: 10.1038/ng0696-154

Auton A, et al. 2015. A global reference for human genetic variation. Nature 526: 68-74. doi: 10.1038/nature15393

Charlesworth B 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nature reviews. Genetics 10: 195-205. doi: 10.1038/nrg2526

Choudhury A, et al. 2014. Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. BMC genomics 15: 437. doi: 10.1186/1471-2164-15-437

Durand EY, Eriksson N, McLean CY 2014. Reducing pervasive false-positive identical-by-descent segments detected by large-scale pedigree analysis. Molecular biology and evolution 31: 2212-2222. doi: 10.1093/molbev/msu151

Duret L, Galtier N 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annual review of genomics and human genetics 10: 285-311. doi: 10.1146/annurev-genom-082908-150001

Fedorova L, Qiu S, Dutta R, Fedorov A 2016. Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes. Genome biology and evolution 8: 777-790. doi: 10.1093/gbe/evw034

Gabriel SB, et al. 2002. The structure of haplotype blocks in the human genome. Science 296: 2225-2229. doi: 10.1126/science.1069424

Guthery SL, Salisbury BA, Pungliya MS, Stephens JC, Bamshad M 2007. The structure of common genetic variation in United States populations. American journal of human genetics 81: 1221-1231. doi: 10.1086/522239

Haber M, Mezzavilla M, Xue Y, Tyler-Smith C 2016. Ancient DNA and the rewriting of human history: be sparing with Occam's razor. Genome biology 17: 1. doi: 10.1186/s13059-015-0866-z

Hartl DC, AG. 2007. Principles of population genetics. Sunderland, MA, USA: Sinauer Associates, Inc. Publishers.

Hinds DA, et al. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307: 1072-1079. doi: 10.1126/science.1105436

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proceedings of the National Academy of Sciences of the United States of America 92: 532-536.

Huerta-Sanchez E, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512: 194-197. doi: 10.1038/nature13408

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge, UK: Cambridge University Press.

Klyosov AA 2014. Reconsideration of the "Out of Africa" Concept as Not Having Enough Proof. Advances in Anthropology 4: 18-37.

Meyer M, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222-226. doi: 10.1126/science.1224344

Mierswa I, Wurst, M., Klinkenberg, R., Scholz, M., Euler, T. 2006. Rapid Prototyping for Complex Data Mining Tasks.

Mondal M, et al. 2016. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Nature genetics. doi: 10.1038/ng.3621

Prufer K, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43-49. doi: 10.1038/nature12886

Qiu S, Fedorov A 2015. Maruyama's allelic age revised by whole-genome GEMA

simulations. Genomics 105: 282-287. doi: 10.1016/j.ygeno.2015.02.005

Qiu S, et al. 2014. Genome evolution by matrix algorithms: cellular automata approach to population genetics. Genome biology and evolution 6: 988-999. doi: 10.1093/gbe/evu075

Smith TaF, E. 2016. Statistical Genomics: Methods and Protocols. New York, NY, USA: Springer.

Stoneking M, Krause J 2011. Learning about human population history from ancient and modern genomes. Nature reviews. Genetics 12: 603-614. doi: 10.1038/nrg3029

Stringer CB, Andrews P 1988. Genetic and fossil evidence for the origin of modern humans. Science 239: 1263-1268.

Takahata N, Satta Y, Klein J 1995. Divergence time and population size in the lineage leading to modern humans. Theoretical population biology 48: 198-221. doi: 10.1006/tpbi.1995.1026

Tattersall I 2009. Out of Africa: modern human origins special feature: human origins: out of Africa. Proceedings of the National Academy of Sciences of the United States of America 106: 16018-16021. doi: 10.1073/pnas.0903207106

Wolpoff MH, J.; Caspari,R. 2000. Multiregional, Not Multiple Origins. AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 112: 129-136.

Zhang J, Rowe WL, Clark AG, Buetow KH 2003. Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. American journal of human genetics 73: 1073-1081.

doi: 10.1086/379154

Zhu Q, et al. 2011. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. American journal of human genetics 88: 458-468. doi: 10.1016/j.ajhg.2011.03.008

Abecasis, G., Auton, A., Brooks, L., DePristo, M., Durbin, R., Handsaker, R., Consortium, 1. G. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 56-65.

Al-Khudhair, A., Qiu, S., Wyse, M., Chowdhury, S., Cheng, X., Bekbolsynov, D., Fedorov, A. (2015). Inference of Distant Genetic Relations in Humans Using "1000 Genomes". *Genome Biology and Evolution*, 481-492.

Bansal, V., & Libiger, O. (2015). Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. *BMC Bioinformatics*, 1-11.

Cruciani, F., La Fratta, L. F., Santolamazza, P., Sellitto, D., Pascone, R., Moral, P., & Watson, E. (2004). Phylogeographic Analysis of Haplogroup E3b (E-M215) Y Chromosomes Reveals Multiple Migratory Events Within and Out Of Africa. *American Journal of Human Genetics*, 1014-1022.

Cruciani, F., Santolamazza, P., Underhill, P. A., Shen, P., Macauley, V., & Moral, P. (2001).

A Back Migration from Asia to Sub-Saharan Africa Is Supported by High-

Resolution Analysis of Human Y-Chromosome Haplotypes. *American Journal for Human Genetics*, 1197-1214.

Eyheramendy, S., Martinez, F. I., Manevy, F., Vial, C., & Repetto, G. M. (2014). Genetic structure characterization of Chileans reflects historical immigration patterns. *Nature Communications*, 1-10.

Fedorova, L. S. (2016). Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes. *Genome Biology and Evolution*, 777-790.

Gautier, M., & Vitalis, R. (2012). Inferring Population Histories Using Genome-Wide Allele Frequency Data. *Molecular Biology and Evolution*, 654-668.

Henn, B., Botigue, L., Gravel, S., Wang, W., Brisbin, A., Byrnes, J., Zalloua, P. A. (2010). Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. *PLOS Genetics*, 1-11.

Hodgson, J. A., Mulligan, C. J., Al-Meeri, A., & Raaum, R. L. (2013). Early Back-to-Africa Migration into the Horn of Africa. *PLOS Genetics*, 1-17.

Hofer, T., Ray, N., Wegmann, D., & Excoffier, L. (2009). Large Allele Frequency Differences between Human Continental Groups are more Likely to have Occurred by Drift During range Expansions than by Selection. *Annals of Human Genetics* , 95-108.

Jin, L., Underhill, P. A., Doctor, V., Davis, R. W., Shen, P., Cavalli-Sforza, L., & Oefner, P. J. (1999). Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. *PNAS*, 3796-3800.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., & Cann, H. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science*, 1100-1104.

Luis, J., Rowold, D. J., Regueiro, M., Caeiro, B., Cinnioglu, C., Roseman, C., Herrera, R.J. (2004). The Levant versus the Horn of Africa: Evidence for Bidirectional Corridors of Human Migrations. *The American Society of Human Genetics*, 532-544.

Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., & Xue, Y. (2015). Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. *The American Journal of Human Genetics*, 986-991.

Sharp, K., Kretzschmar, W., Delaneau, O., & Marchini, J. (2015). Phasing for medical sequencing using rare variants and large haplotype reference panels. *Oxford Journals*, 1974-1980.

Watkins, W., Xing, J., Huff, C., Witherspoon, D., Zhang, Y., Perego, U., Jorde, L. (2012). Genetic analysis of ancestry, admixture and selection in Bolivian and Totonac populations of the New World. *BMC Genetics*, 13-39.

Chapter 2

Abecasis, G., Auton, A., Brooks, L., DePristo, M., Durbin, R., Handsaker, R., Consortium, 1. G. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 56-65.

Al-Khudhair, A., Qiu, S., Wyse, M., Chowdhury, S., Cheng, X., Bekbolsynov, D., Fedorov, A. (2015). Inference of Distant Genetic Relations in Humans Using "1000 Genomes". *Genome Biology and Evolution*, 481-492.

Bansal, V., & Libiger, O. (2015). Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. *BMC Bioinformatics*, 1-11.

Cruciani, F., La Fratta, L. F., Santolamazza, P., Sellitto, D., Pascone, R., Moral, P., & Watson, E. (2004). Phylogeographic Analysis of Haplogroup E3b (E-M215) Y Chromosomes Reveals Multiple Migratory Events Within and Out Of Africa. *American Journal of Human Genetics*, 1014-1022.

Cruciani, F., Santolamazza, P., Underhill, P. A., Shen, P., Macauley, V., & Moral, P. (2001). A Back Migration from Asia to Sub-Saharan Africa Is Supported by High-Resolution Analysis of Human Y-Chromosome Haplotypes. *American Journal for Human Genetics*, 1197-1214.

Eyheramendy, S., Martinez, F. I., Manevy, F., Vial, C., & Repetto, G. M. (2014). Genetic structure characterization of Chileans reflects historical immigration patterns. *Nature Communications*, 1-10.

Fedorova, L. S. (2016). Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes. *Genome Biology And Evolution*, 777-790.

Gautier, M., & Vitalis, R. (2012). Inferring Population Histories Using Genome-Wide Allele Frequency Data. *Molecular Biology and Evolution*, 654-668.

Henn, B., Botigue, L., Gravel, S., Wang, W., Brisbin, A., Byrnes, J., . . . Zalloua, P. A. (2010). Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. *PLOS Genetics*, 1-11.

Hodgson, J. A., Mulligan, C. J., Al-Meeri, A., & Raaum, R. L. (2013). Early Back-to-Africa Migration into the Horn of Africa. *PLOS Genetics*, 1-17.

Hofer, T., Ray, N., Wegmann, D., & Excoffier, L. (2009). Large Allele Frequency Differences between Human Continental Groups are more Likely to have Occurred by Drift During range Expansions than by Selection. *Annals of Human Genetics* , 95-108.

Jin, L., Underhill, P. A., Doctor, V., Davis, R. W., Shen, P., Cavalli-Sforza, L., & Oefner, P. J. (1999). Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. *PNAS*, 3796-3800.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., & Cann, H. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science*, 1100-1104.

Luis, J., Rowold, D. J., Regueiro, M., Caeiro, B., Cinnioglu, C., Roseman, C., Herrera, R.J. (2004). The Levant versus the Horn of Africa: Evidence for Bidirectional Corridors of Human Migrations. *The American Society of Human Genetics*, 532- 544.

Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., & Xue, Y. (2015). Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. *The American Journal of Human Genetics*, 986-991.

Sharp, K., Kretzschmar, W., Delaneau, O., & Marchini, J. (2015). Phasing for medical sequencing using rare variants and large haplotype reference panels. *Oxford Journals*, 1974-1980.

Watkins, W., Xing, J., Huff, C., Witherspoon, D., Zhang, Y., Perego, U., Jorde, L. (2012). Genetic analysis of ancestry, admixture and selection in Bolivian and Totonac populations of the New World. *BMC Genetics*, 13-39.