Genetic Dating and Pattern of Admixture in Modern Human Evolution

Joel Defo (jlxdef001@myuct.ac.za)

23 June 2017

Submitted in partial fulfillment of a research masters degree at UCT South Africa

University of Cape Town (UCT)
Supervised by: Professor Nicola Mulder
Co-supervised by: Dr. Emile Chimusa Rugamika
University of Cape Town, South Africa

Contents

List of Figures

List of Tables

Abstract

Genetic variation is shaped by admixture between populations in an evolutionary process. The mixture dynamic between groups of populations results in a mosaic of chromosomal segments inherited from multiple ancestral populations. The distribution of ancestral chromosomal segments and the recombination breakpoints in an admixed genome provide information about the time of admixture. Studying populations with particular ancestries has become a major interest in population genetics because of medical and evolutionary impacts of the patterns of single nucleotide polymorphisms. It provides a better understanding of the impact of population migrations and helps us uncover interactions between several populations. Most of the research on admixed population dating has focused on a single interaction between two populations using various approaches. Some have extended this to mixing of three populations based on assumptions and approaches which differ from one tool to another. However, the inference of distinct ancestral proportions along the genome of an admixed individual and plausible dates of admixture, still remain a challenge in the case of multi-way admixed populations. This dissertation consists of three research initiatives. First, provide a succinct review of current methods for dating the admixture events. We accomplish this by providing a comprehensive review and comparison of current methods pertinent to date admixture event. Second, we assess various admixture dating tools which estimate the time of admixture between two parental populations. We do so by performing various simulations assuming a particular number of generations and use these to evaluate the tools. Third, we apply the top three assessed methods to some admixed populations from the 1000 Genomes project. Despite MALDER shows improvement and produces reasonable date estimates over other current methods, the results from both simulation and real data suggest that dating ancient admixture events accounting for the effect of other admixtures remains a challenge. Our results suggest the need for developing a new approach to date ancient and complex admixture events in multi-way admixed populations.

Acknowledgements

First and foremost, I would like to thank the Almighty God for giving me the grace that i need to accomplish this project.

- The National Research Foundation and The University of Cape Town for their financial support,
- Many thanks to my supervisor Prof Nicola Mulder who introduce me to the field of bioinformatics and for proposing the project to me and endlessly guided me to accomplish it.
- Great thanks to my co-supervisor Dr Emile Chimusa Rugamika for his unrelenting support, advise and directions,
- I would like to thank my family particularly, my parents Mr and Mrs Simo for their support and encouragement,
- I would like to thank the brethren of the Christian Missionary Fellowship International for their prayers and encouragements,
- And all my colleagues from the CBIO group at the University of Cape Town.

Declaration

I, the undersigned, hereby declare that the work contained in this thesis is my original work, and that any work done by others or by myself previously has been acknowledged and references accordingly.

1. Introduction and Background

1.1 Introduction

The history of human evolution is characterized by the exchange of genetic materials across individuals, resulting in individuals with unique genetic features. This has been the focus of research for many genetics scholars. The existence and history of species in general, and the human population in particular, was ascertained using anthropological, linguistic, and historical approaches. However, with the availability of genetic data and the improvement of computational tools, combined with the appropriate statistical methods, it is now possible to infer the admixture history of human populations. Admixture occurs as a result of interbreeding between two or more previous isolated populations. This interbreeding yields genetic recombination breakpoints and the formation of mixed DNA segments. The chromosome of the descendant displays a pattern of chromosomal segments (or blocks) of different ancestry with different sizes that may provide some information about the time since admixture occurred (Xu et al., 2008; Pugach et al., 2011; Sanderson et al., 2015).

The divergence of genetic ancestry has come about as a result of biogeographical distributions of human populations (Tishkoff and Kidd, 2004; Shriver et al., 2003). This distribution, at the genetic level exhibits a pattern of single nucleotide polymorphism (SNPs) which may have either medical or evolutionary implications. Various methods have been implemented to infer the genetic ancestry either "locally" or "globally". Global ancestry is the average proportion of ancestry across the genome of each ancestral population, and is assigned to each individual; while local ancestry is the estimate of the individual's ancestral proportion of each ancestry at a particular chromosomal location. In addition, local ancestry of an individual includes 0, 1, or 2 copies of alleles derived from each contributing population (Liu et al., 2013; Thornton and Bermejo, 2014).

The most popular tools for admixture analysis include STRUCTURE (Pritchard et al., 2000; Falush et al., 2003), ADMIXTURE (Alexander et al., 2009) and EIGENSTRAT (Price et al., 2006a) which use clustering algorithm to locate admixed population as in relationship with current representative ancestral population. The local ancestry tool, HAPMIX (Price et al., 2009), applies a Hidden Markov Model (HMM) with background Linkage Disequilibrium(LD) to compute a probabilistic estimate of ancestry at each locus. HAPMIX takes as input a recombination map for the regions to be analysed, phased "parental" chromosomes from two reference populations, and "offspring" data from the admixed population being analysed. SABER (Tang et al., 2006), like HAPMIX, uses background LD by considering pairwise allele frequencies when no change of ancestry is inferred, but it doesn't model haplotype structure. LAMP-LD/HAP (Baran et al., 2012) infers locus-specific ancestry in recent admixed populations to output the estimated number of alleles from each ancestry at each locus for each admixed individual. LAMP-LD uses a hierarchical Hidden Markov Model(HMM) combined with a window-based algorithm to represent haplotypes in the population in the case of multi-way admixture. MULTIMIX (Churchhouse and Marchini, 2012) ascertains how ancestry changes along the chromosome by using a multivariate normal model on haplotype probabilities given

ancestry and an HMM. The motivation behind the development of these methods is to (1) assess the migration pattern (Jakobsson et al., 2008; Gravel et al., 2011), (2) to increase the power of association mapping (Pasaniuc et al., 2011), and (3) to enhance admixture mapping for gene-related underling ethnic difference in a particular disease risk and personalized drug therapy applications (Winkler et al., 2010; Seldin et al., 2011; Rodriguez et al., 2013).

It has been shown that when the number of generations increases in admixture processes, the ancestral chromosomal segments from different parental populations are spliced into shorter pieces. The distribution of ancestral chromosomal segments and the recombination breakpoints in an admixed genome provide information about the time of admixture (Pugach et al., 2011; Churchhouse and Marchini, 2012). The study of human history in admixed populations can shed light on the patterns of genetic variation throughout modern human evolution in order to understand the demographics, and adaptive processes of human populations. The study of human genetic variation has evolved over time due to the observation of different traits among individuals in several population groups around the globe. As a result, one needs to understand the dynamics related to the origin of these variations, the evolution process and its consequences in human healthcare.

Several forces have been identified as causing human evolution at the genetic level. These include mutation, migration, genetic drift and natural selection (Hartl and Clark, 1997). These factors cause several changes in the frequency of occurrence of alleles (Gillespie, 2010). The information from these alleles added to computational tools detect the genetic structure of the population to identify regions along the genome responsible for phenotypic traits (Gillespie, 2010; Cho et al., 2009; Hirschhorn and Daly, 2003). The distribution of the length of ancestral blocks from parental populations in admixed populations, which help us to determine the admixture history, have demonstrated implications in finding genes with associations to diseases and drugs response (Cheng et al., 2010), Tuberculosis (Chimusa et al., 2013; Moller and Hoal, 2010), Breast cancer (Fejerman et al., 2009), and Hypertension (Zhu and Cooper, 2007), just to name a few. Studying the distribution of the length of ancestral blocks should provide new insights into the history of species and could shed light on the signature of natural selection and age of mutation in local ancestry at a fine scale. (Jin et al., 2011).

Several methods have been developed to infer the date of admixture, accounting for many factors which include ancestry linkage disequilibrium, ancestral track or haplotype blocks. However, the inconvenience of these methods lies in the fact that either the method is limited to two-way admixture only, or there is no identification of the pattern of ancestry along the chromosome to determine the recombination breakpoints. Xu and colleagues (2008) analysed the pattern of admixture among the Uygurs by making use of the recombination breakpoints to infer the date of admixture, but this method was limited to two-way admixture.

1.2 Motivation and Objectives of the Project

Recent population studies reveal that many world populations are admixed, and the complexity of the admixture become more complex from one generation to another (Loh et al., 2013; Price

et al., 2006b). Both in population genetics and epidemiological studies, an understanding of population mixture, does not only offer the unique opportunities for comprehending the human diversity and evolutionary history, but renders an instructive account of genetic variations between or among population groups with regard to disease susceptibility and drug response. Since the number of ancestral admixture events is reflected in the genome of a founding descendant, we make use of this notion to assess the recent methods that are purported to have accurate estimations in dating the time of admixture events. However, the inference of distinct ancestral proportions along the genome of an admixed individual and plausible dates of admixture, still remains, by some distance, a challenging notion for researchers. In addition, existing methods for estimating the dates of distinct admixture events in admixed populations, are limited to two-way admixed populations and recent admixture events.

This project aimed to

(1) review and provide a succinct account of these methods for dating the admixture events in multi-way admixed populations.

(2) familiarise with computational tools of populations genetics and simulate complex multi-way admixture scenarios to assess and compare current methods for dating the admixture events.

(3) apply the most accurate method to data from the 1000 Genomes project dataset and from the Human Genome Diversity Project.

1.3 Population Genetic Variation

1.3.1 Overview

Genetic variation studies investigate the difference of heritability of the traits between and within organisms (Relethford, 2012; Relethford and Harding, 2001; Hartl and Clark, 1997; Gillespie, 2010). In human populations, it has been estimated that the differences between nucleotides of two unrelated individuals is of the order of 3 million. Particularly in admixed populations, the patterns of genetic variants can shed light on the history of ancestral populations during the evolutionary process and also on the susceptibility to particular diseases (Xu and Jin, 2011; LB Jorde and Bamshad, 2001; McKeigue, 1998; Cavalli-Sforza et al., 1994). The detection of genetic variants is nowadays more commonly achieved using genotyping array or next generation sequencing as means to identify the allelic difference between individuals (Hartl and Clark, 1997; Li et al., 1990). This genetic polymorphism is useful to investigate genetic relationship among populations, admixture, evolutionary process and migration (Hartl and Clark, 1997). Moreover, genetic variation can be used to determine the paternity of a child as well as to identify a suspect in a crime scene with a sufficient amount of alleles (Hartl and Clark, 1997).

In the human genome, we can classify genetic variation into single nucleotide polymorphisms (SNPs) caused by mutations, short insertions and deletions (indels), copy number variation

(CNV), variable number tandem repeats (VNTR), including microsatellite and minisatellite, and epigenetic variation. Genetic variation is generated continuously by mutational processes and is then governed by factors such as gene flow, recombination, genetic drift, and natural selection (Hartl and Clark, 1997; Hamilton, 2009; Gillespie, 2010; Relethford, 2012). Various statistics such as the genetic distance(F_{st}), or the Linkage Disequilibrium coefficient have been developed to measure genetic variability between populations. This variability take its basis on the pattern of allele frequencies between individuals. Below, we review some measures of genetic variation which are widely used in the field of population genetics.

1.3.2 Computational of Genetic Distance (F_{st})

The measurement of genetic variability, has been studied by many scholars with the aim to quantify the variability between populations (Kalinowski, 2002; Masatoshi, 1972; Hartl and Clark, 1997). The genetic distance is a measure which describes genetic differentiation between and within populations. Its estimate plays a major role in population and evolutionary genetic for having wide application in disease association mapping and forensic science (Weir and Hill, 2002; Holsinger and Weir, 2009). Knowing that the genetic variation between populations can shed light on the distribution of alleles, measures have been defined to quantify genetic variability between populations. Although there have been various estimates of the genetic distance (Weir and Cockerham, 1984; Masatoshi, 1986; Hudson et al., 1992; Weir and Hill, 2002), its definition, interpretation and correct estimation have been a subject of debate this recent decades. Li(2008) developed the Wright Fisher F-statistics formula as a genetics distance accounting for mixed population.

According to Li, Let p_1 and p_2 be the allele frequencies of a sub-population of size n_1 and n_2 respectively. Then q_1 = 1-p_1 and q_2 = 1-p_2 are the frequencies of the other alleles (knowing that the human population is diploid). The heterozygous frequency in the two populations will be $2p_1q_1$ and $2p_2q_2$, respectively. We define p^* = kp_1 + (1-k)p_2 and q^* = kq_1 + (1-k)q_2. p^* and q^* are the average allele frequencies of the two alleles and k is the mixing proportion of the two populations (Li, 2008). Hence the heterozygosity in each sub-population denoted H_{sb} is given by:

$$H_{sb} = 2[kp_1q_1 + (1-k)p_2q_2] \tag{1.3.1}$$

and the heterozygote frequency in the combined population denoted H_{wo} will be

$$H_{wo} = 2p^*q^* \tag{1.3.2}$$

Li(2008) proved that $H_{wo} \geq H_{sb}$, therefore the heterozygote frequency in the combined population always increases and varies proportionally with the allele difference $|p_1 - p_2|$. The percentage of increasing the heterozygote frequency by combining sub-populations is given as

follows:

$$\begin{aligned} F_{st} &= \frac{H_{wo} - H_{sb}}{H_{wo}} \\ &= \frac{2k(1-k)(p_1p_2)^2}{H_{wo}} \\ &\approx \frac{2k(1-k)(p_1p_2)^2}{H_{sb}} \\ &= \frac{(p_1p_2)^2}{(\frac{p_1q_1}{1-k}) + (\frac{p_2q_2}{k})} \end{aligned} \tag{1.3.3}$$

This statistic range from 0 (lowest value indicating no differences between the overall population and the sub-populations) to 1 (maximum value indicating the sub-populations are very isolated from each other). In practice, the F_{st} is much less than 1 despite the high differentiation of the sub-populations. The value of F_{st} can be dataset specific or marker-specific; the normal variation between ethnic groups is established when the value of F_{st} is 10^{-1}, while a F_{st}-value between 10^{-4} and 10^{-3} describes variation between regions of an isolated population (Li, 2008). Recently, Bathia and colleagues (2013) came across the choice of F_{st} estimator for the purpose of comparing populations using a series of bi-allelic SNPs. They recommended the estimator of genetic distance based on the work of Hudson and colleague(1992), as H accounts for heterozygosity, and for its non-sensitivity to sample size. This estimate is defined as follows:

$$F_{st} = 1 - \frac{H_w}{H_b} \tag{1.3.4}$$

$$= \frac{(p_1 - p_2)^2(\frac{p_1(1-p_1)}{n_1-1}\frac{p_2(1-p_2)}{n_2-1})}{p_1(1-p_2) + p_2(1-p_1)} \tag{1.3.5}$$

n_i and p_i are respectively the sample size and the sample allele frequency of the population, H_w is the mean number of differences within populations and H_b is the mean number of difference between populations (Bhatia et al., 2013). These are the statistics used in the EIGENSOFT package to measure genetic distance (Bhatia et al., 2013) and have been applied to investigate the structure of the Europeans population as illustrated in table (1.1).

	Sp	Fr	Be	UK	Sw	No	Ge	Ro	Cz	Sl	Hu	Po	Ru	CEU	CHB	JPT
Fr	0.0008															
Be	0.0015	0.0002														
UK	0.0024	0.0006	0.0005													
Sw	0.0047	0.0023	0.0018	0.0013												
No	0.0047	0.0024	0.0019	0.0014	0.0010											
Ge	0.0025	0.0008	0.0005	0.0006	0.0011	0.0016										
Ro	0.0023	0.0017	0.0018	0.0028	0.0041	0.0044	0.0016									
Cz	0.0033	0.0016	0.0013	0.0014	0.0016	0.0024	0.0003	0.0016								
Sl	0.0034	0.0017	0.0015	0.0017	0.0019	0.0026	0.0005	0.0014	0.0001							
Hu	0.0030	0.0015	0.0013	0.0016	0.0020	0.0026	0.0004	0.0011	0.0001	0.0001						
Po	0.0053	0.0032	0.0028	0.0027	0.0023	0.0034	0.0012	0.0028	0.0004	0.0004	0.0006					
Ru	0.0059	0.0037	0.0034	0.0032	0.0025	0.0036	0.0016	0.0030	0.0008	0.0007	0.0009	0.0003				
CEU	0.0026	0.0008	0.0005	0.0002	0.0011	0.0012	0.0006	0.0028	0.0014	0.0016	0.0016	0.0026	0.0031			
CHB	0.1096	0.1094	0.1093	0.1096	0.1073	0.1081	0.1085	0.1047	0.1080	0.1069	0.1058	0.1086	0.1036	0.1095		
JPT	0.1118	0.1116	0.1114	0.1117	0.1095	0.1103	0.1107	0.1068	0.1102	0.1091	0.1079	0.1108	0.1057	0.1117	0.0069	
YRI	0.1460	0.1493	0.1496	0.1513	0.1524	0.1531	0.1502	0.1463	0.1503	0.1498	0.1490	0.1520	0.1504	0.1510	0.1901	0.1918

Table 1.1: Fst statistics calculated between each pair of countries taken from the paper of Heat al.(2008). Spain (Sp), France (Fr), Belgium (Be), Sweden (Sw), Norway (No), Germany (Ge), Romania (Ro), Czech (Cz),Slovakia (Sl), Hungary (Hu), Poland (Po), Russia (Ru), and the four HapMap cohorts CEU, CHB, JPT and YRI

1.3.3 Relevance of Human Linkage Disequilibrium

The term linkage disequilibrium (LD) was initially introduced in 1940 to designate the degree of non-random gametic association of alleles at different loci (Pritchard and Przeworski, 2001; Mueller, 2004; Slatkin, 2008). LD can also be defined as the correlation between neighbouring alleles which descend from common ancestral chromosomes (Reich et al., 2001). LD has proven to play a major role in reconstructing historical events as a result of long-range migrations and mixture between populations (Pfaff et al., 2001; Loh et al., 2013). LD is useful to detect causal variants that underlie common and complex diseases particularly in admixed populations (McKeigue, 1998; Zhu et al., 2004; Patterson et al., 2004; Dries, 2009). For instance, suppose two bi-allelic loci A and B with their two respective alleles (A_1, A_2) and (B_1, B_2), there are four possible haplotypes (or gametes) in the population A_1B_1, A_1B_2, A_2B_1, and A_2B_2. The expected gamete frequency of the haplotype AB is $p_A p_B$; but if the observed population frequency of the haplotype block AB is different to $p_A p_B$, then we conclude that the alleles A and B are in LD (Hartl and Clark, 1997; Mueller, 2004), otherwise they are noted to be in linkage equilibrium. In the latter state, the genotypes at the two loci are independent of each other; this is the concept similar to the Hardy-Weinberg equilibrium law. Let u_1, u_2, v_1 and v_2 be the respective allele frequencies of the alleles A_1, A_2, B_1 and B_2, and p_{11}, p_{12}, p_{21} and p_{22} be the actual gametic frequencies. Assuming non-random mating the observed gametic frequencies are defined as:

$$\begin{aligned} D_{A_2B_2} &= p_{22}u_2v_2 \\ D_{A_2B_1} &= p_{21}u_2v_1 \\ D_{A_1B_1} &= p_{11}u_1v_1 \\ D_{A_1B_2} &= p_{12}u_1v_2 \end{aligned} \tag{1.3.6}$$

We will consider this notation throughout the chapter. Equation (1.3.6) is for a specific pair of alleles and does not depend on the case of the other alleles, meaning each allele has his own quantity of LD. As both loci are bi-allelic, the constraint of the opposite sign and coupling phase

are considered; meaning $D_{A_1B_1}$ = -$D_{A_1B_2}$ = -$D_{A_1A_2}$ = $D_{A_2B_2}$ (Iewontin, 1974; Slatkin, 2008).

With the increased availability of dense genome-wide data of single nucleotide polymorphisms (SNPs), the study of human evolution has led scholars to ascertain the degree and the level of LD for the purpose of the association mapping between unlinked loci (McKeigue, 1998; Chakraborty and Weiss, 1998; Weiss and Clark, 2002). Such studies have important applications in medical population genetics, particularly in mapping susceptible genes for complex diseases (McKeigue, 2005; Hirschhorn and Daly, 2003). In addition, LD has range of applications including evolutionary history and demographic processes to locate mutations responsible for a particular phenotype. The extent of its presence within a region is likely to vary in inverse relation to the local recombination rate within that region(Reich et al., 2001; Mueller, 2004). In such a dynamic, the allele frequencies between sub-populations display significant differences due to factors such as selection, genetic drift, mutations, or population admixture throughout the generations (Lonjou et al., 2003; Hartl and Clark, 1997). Several methods have been proposed to measure the amount of linkage disequilibrium either based on the frequency of the gametes, or based on the information from genotype data. Here we define the methods which are widely used. Lewontin (1974) suggested a quantity of LD which scale the value between -1 and 1 as follows:

$$D' = \begin{cases} \frac{D}{\min\{u_i(1-v_i),(1-u_1)v_i\}}, & (D \geq 0) \\ \frac{D}{\min\{u_iv_i,(1-u_i)(1-v_i)\}}, & (D < 0) \end{cases} \tag{1.3.7}$$

Hill and Robertson (1968), quantify the LD by using the correlation approach:

$$r^2 = Corr(A,B)^2 = \frac{D^2}{u_1u_2v_1v_2} \tag{1.3.8}$$

LD measures are descriptive statistics and its quantity could not indicate whether there is significant statistical association between alleles in haplotypes. To address this, one can perform a chi-square test, Fisher exact test or likelihood ratio test to ascertain the significance of the LD between loci. Assuming the existence of a historical relationship between alleles at two closely located loci, the gene-specific pattern of linkage disequilibrium will influence the presence of the trait in the human population and this occurrence will decay gradually in the population by recombination during meiosis (Pfaff et al., 2001; Reich et al., 2001; Pritchard and Przeworski, 2001). Therefore, the relative allele distributions of an unknown gene and that of a very nearby marker will be non-random, or in other words, the two are in linkage disequilibrium. Linkage disequilibrium-based genetic association studies offer a potentially powerful approach for mapping causal genes (Pritchard and Przeworski, 2001).

1.4 Genomic Admixture

Genomic admixture has become an interesting subject of research in biology for its implication in population history reconstruction and disease-gene mutations. Admixture comes into existence when two or more genetically distant populations interbreed (Patterson et al., 2012).

The resulting admixed population generates a pattern of genetic variation which provides some information concerning the contribution and the distribution of ancestry along the admixed chromosome, including the number and time of the admixture events. Its also provides mutations that arises from the mixture, and possibly the past natural selection. Therefore the genetic trace of the admixed population needs to be investigated in order to determine what has happened in the past and how we can project genetic patterns in the future population (Slatkin, 2008).

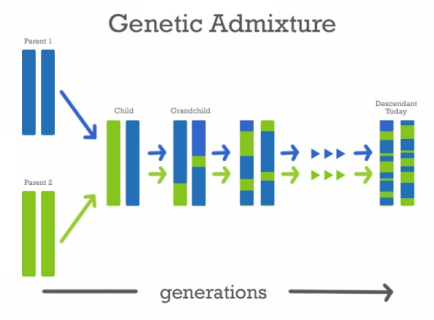

Figure 1.1: genetic admixture process (Saltarin, 2014)

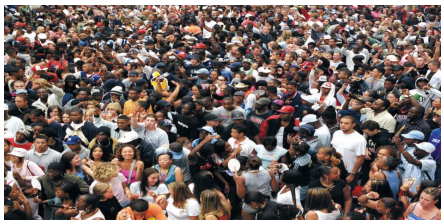

Figure 1.2: genetic variation in human population (Picture issue from the paper of Evan Birney and Nicole Soranzo(2015))

1.4.1 Population Structure

Population genetic structure refers to the heterogeneity in allele frequency among populations caused by limited gene flow (Dunpanloup et al., 2002; Excoffier et al., 2009). It can be estimated using hierarchical analysis of molecular variance (Pritchard et al., 2000). The population structure can provide valuable information regarding the migration patterns, natural selection, gene flow, and genetic drift of the concerned populations. More importantly, it can exhibit significant knowledge on human ancestral history. Since the number of admixture blocks reflects previous ancestral recombination events, the genome of an admixed individual contains information on the ancestries and past interbreeding events. The analysis of population structure falls into two different approaches which are widely used in population genetics: probabilistic-based models and principal component analysis (PCA). Here, we describe the different models used to study population structure. Both models can be expressed under a general model as follow: Let G be the genotype matrix expressed in terms of two low-rank matrices. Assuming that G is the genotype matrix at p SNPs for n individuals with p taking values in 0, 1, or 2 copies of the reference allele, then the model of population structure can be defined by

$$E[G] = \alpha \times G \tag{1.4.1}$$

or by expansion

$$E[G_{ij}] = \sum_{k=1}^{K} \alpha_{i,k} F_{k,j} \tag{1.4.2}$$

Here, α is a $n \times K$ matrix and F is a $K \times p$ matrix. K is the proposed number of ancestral populations, $\alpha_{i,k}$ is the admixture proportion of individual i in population k, $F_{k,j}$ is the allele frequency of the reference allele in population k and $E[G_{i,j}]$ is the expected gamete frequency in the admixed individual (Engelhardt and Stephen, 2010). Under the probabilistic-based model, Pritchard et al. (2000) developed a Bayesian modelling approach, which uses unlinked genotypes to infer population substructure and implemented it in the software STRUCTURE. STRUCTURE uses a bayesian approach via a Markov Chain Monte Carlo (MCMC) method to infer the presence of distinct populations, assigns individuals to populations, computes individual ancestry proportions, and estimates ancestral population allele frequencies in admixed populations. Later, Falush et al. (2003) extended the method to accommodate linked markers. To apply the above formula, we assume in this model that $G_{i,j}$ follows a binomial probability distribution with parameters 2 and $a_{i,j}$, i.e., $G_{i,j} \sim B(2, a_{i,j})$ where :

$$a_{i,j} = \sum_{k=1}^{K} \alpha_{i,k} P_{k,j} \tag{1.4.3}$$

$P_{k,j}$ being the allele frequency of the reference allele in population k and $\alpha_{i,k}$ defined previously. It follows that

$$E[G_{ij}] = \sum_{k=1}^{K} 2\alpha_{i,k} P_{k,j} \tag{1.4.4}$$

Similar to the model in STRUCTURE, Alexander et al.(2009) developed a method based on maximum likelihood with the same approach of Pritchard and colleagues under the assumption of Hardy-Weinberg equilibrium

$$L(Q,F) = \sum_{i} \sum_{j} \left(g_{i,j} \ln \sum_{\mathrm{k}} \mathrm{q_{i,k} f_{k,j}} + (2 - g_{i,j}) \ln \sum_{\mathrm{k}} \mathrm{q_{i,k}} (1 - \mathrm{f_{k,j}}) \right) \tag{1.4.5}$$

L is the value of the maximum likelihood, which depends on the Q matrix proportion of contribution of ancestry, and F the allele frequencies of each individual, modelled as the mixture of the q fractions of ancestral population at the allelic frequency f at locus j of individual i in the genotype matrix g. Since these methods provide information on the parameter of interest, they can also be computationally costly when it comes to large datasets with many different populations. Recently, ChromoPainter and fineSTRUCTURE developed by Lawson and colleagues (Lawson et al., 2012) make use of the haplotype structure to infer the Personal Components(PCs) and population structure respectively. FineSTRUCTURE and ChromoPainter make use of the Li and Stephens algorithm (Li and Stephens, 2003) to cluster admixed individuals identified as "recipients" that have a similar genetic make-up to other sampled individuals and identified as "donors" within the ancestral populations. This process, called "chromosome painting", assumes that chunks of every admixed individual provide independent information about ancestry and that for every individual haplotype in the admixed population, at each locus there exists one or more closest haplotype relative in the sample ancestral population that closely matches them (Lawson et al., 2012; Hellenthal et al., 2014). The model aims to partition the dataset into K groups with indiscernible genetic ancestry, and utilizes a

Bayesian approach combined to the Markov Chain Monte Carlo(MCMC) model to generate the coancestry matrix. Mathematically, let K be the number of clusters where all the donor populations will be placed, we define a the recipient population and b the donor population with their respective total number of individuals n_a and n_b. The total number of chunks in population a that come from population b is given as follows:

$$x_{ab} = \sum_{i \in a, j \in b} x_{ij} \tag{1.4.6}$$

where i, j are individuals in each population. For each individuals i, j in population a,b respectively, the probability that a single chunk is donated to individual i from individual j is $\frac{P_{ab}}{\hat{n_b}}$ where $\hat{n_b} = n_b$ if $a \neq b$ and $\hat{n_b} = n_b$-1 if not. Since the chunks are independent, therefore the overall likelihood of the coancestry matrix X given the distribution of the chunks P is defined as follows:

$$p(X|P) = \prod_{n_a, n_b = 1}^{K} \left(\frac{P_{ab}}{\hat{n_b}} \right)^{x_{ab}} \tag{1.4.7}$$

P_{ab} is the coancestry matrix that gives the proportion of chunks from any individual in population a that come from population b with 1$\leq$ a, b$\leq$ K. P_{ab} follows the Diriclet probability distribution with parameter β_{ab}, i.e., $P_{ab} \sim$ D(β_{ab}) in which β_{ab} represents the rate of chunks of population a coming from population b as given below:

$$\beta_{ab} = \begin{cases} \frac{1-F}{F} V_b \frac{N}{N-1} & if\ a \neq b \\ (1+\sigma) \frac{1-F}{F} V_b \frac{N}{N-1} \frac{n_a - 1}{n_a} & if\ a = b \end{cases} \tag{1.4.8}$$

(1+σ) is the population growth, $\frac{1-F}{F}$ is the shared variance analogous to the correlated allele frequency defined in the paper of Falush et al.(2003) and N is a sample number of individuals that will help us for adjustment because individuals do not act as donors to themselves.

The principal components approach is used to derive a 2 or more dimensional scatter plot of individuals such that the genetic distances among individuals may be reflected by the geometrical distances among individual genotypes (Price et al., 2006a; Patterson et al., 2006). Principal Component Analysis (PCA) aims to project the individuals into a low-dimensional subspace with orthogonal axes in such a manner that there may be genetic similarities among them in their projected locations (Engelhardt and Stephen, 2010). The principal component reduces the shape of data to clarify the link between breeding materials into explainable fewer dimensions and to make new variables. These new variables are pictured as different non correlating groups. It is expected that the first few axes will explain a large sum of the variations captured by the genotypes (Price et al., 2006a; Aremu, 2011). Initially, PCA was used to summarize allele-frequency data collected from worldwide populations of humans (Cavalli-Sforza et al., 1994). Cluster and principal component analysis can be jointly used to explain the variations in breeding materials as well as in genetic diversity studies. In 2006, Price and colleagues developed a method of PCA which accounts for continuous population stratification in association studies from the genotype data (Price et al., 2006a; Novembre and

Stephens, 2008). The method has been implemented in the software package EIGENSTRAT and is called the Eigenstrat method. The results of PC projections can be interpreted in various area including continuous demographic processes, geographical isolation and admixture (McVean, 2009).

To elucidate the PCA, we have said previously that the general model is based on equation (1.4.1) , but here G_{ij} follow the Normal Distribution with mean $\alpha_{ij}\ F_{ij}$ and variance Ψ^{-1}, i.e. $G_{ij} \sim \mathrm{N}(\alpha_{ij}\ F_{ij};\Psi^{-1})$ where Ψ^{-1} is the residual variance of the distribution $\alpha \times F$. Further, we maximize the model with respect to the parameters α, F and β subject to the constraint which includes the K columns of α verifying $\alpha^T\alpha$ diagonal and K columns of F verifying $F^TF = I_n$ with I_n the Identity matrix of order n. The columns α and the rows F provide the principal components (PCs) and the corresponding PC entries (Price et al., 2006a; Patterson et al., 2006; Engelhardt and Stephen, 2010; Ma and Amos, 2012).

1.4.2 Local Ancestry Inference

During the admixture process, after generations, the chromosomes of the descendant of the admixed individual are broken into chromosomal segments (or blocks) of different ancestry with different sizes; this may provide some information about the time of the admixture occurrence and the points of ancestry along the genome of the admixed offspring (Pugach et al., 2011). As stated ealier above, local ancestry regions of an individual include 0, 1, or 2 copies of alleles derived from each contributing population (Liu et al., 2013; Thornton and Bermejo, 2014). It helps to investigate the presence of recent selection and to analyse the pattern of variation in recombination rates (Baran et al., 2012; Bhatia et al., 2011), as well as to map genes that show ethnic differences in disease risk (Pasaniuc et al., 2011; Chimusa et al., 2013). In local ancestry inference, each chromosome in an individuals genome exhibits a combination of segments that originate from different ancestral populations and the goal is to find the ancestral population of origin at each position.

Price et al.(2009) have developed a method incorporated into HAPMIX that applies a Hidden Markov Model (HMM) to compute a probabilistic estimate of ancestry at each locus; HAPMIX takes as input a recombination map for the regions to be analysed, phased "parental" chromosomes from two reference populations, and "offspring" data from the admixed population being analysed. However, the running time of HAPMIX is long and its conception has been designed solely for a mixture of two references populations. Baran et al. (2012) implemented a method incorporated into LAMP-LD which infers local ancestry by leveraging the haplotype structure of the ancestral populations. The method employs a window-based process followed by a Hidden Markov Model (HMM) with fixed-size state space and no recombinations. LAMP-LD has the advantage that it runs faster than the previous HMM-based methods such as HAPMIX, and it can infer ancestry in a multi-way admixture case (2, 3 or 5 ancestries). WINPOP (Pasaniuc et al., 2009) modifies the original LAMP framework and uses a refined model of recombination events and an efficient dynamic programming algorithm to improve local ancestry inference for situations where ancestral populations are closely related.

Let A= (Q, σ, e) where Q is the set of states composed of disjointed sets Q_i i $\in 0, 1, \ldots, L$ at

SNPs i, with initial state Q_0=a_0, σ is the transition probability function, and e is the emission probability function. We define L as the length of the window(set of SNPs). As a result, there are S× L states in the model with e the probability of emission of each reference or minor allele. If we define S, the number of a fixed state space, then all the Q_i have the condition that $|Q_i| = |Q_j|$ = S for i, j in $0, 1, \ldots, L$, i≠ j. $\sigma(a, a')$ represents the transition probability function from state a at SNP j to state a' at SNP (j+1) such that $\sum_{a'} \sigma_j$(a, a') = 1, then e_j(a,0) = 1- e_j(a, 1). Considering H= $h_1h_2h_3 \ldots h_n$ the observed haplotype, then the probability to observe H given the model A is defined below:

$$P(H|A) = \sum_f \sigma_0(a_0, f_1) \times e_1(f_1, h_1) \prod_{i=2}^{L} S_i(f_{i-1}, f_i)e_i(f_i, H_i) \tag{1.4.9}$$

Where the sum can be taken across all paths of state f= $f_1f_2f_3 \ldots f_n$ (Baran et al., 2012). From the above equation (1.4.9) we define, within the genome non-overlapping windows w= [i, i+L) setting over SNPs i to $i + L$. The model assumes that there are no crossovers that change ancestry occurring within the window and we constrain all crossovers to occur at the boundary of two consecutive windows. If we define S_w =(M_{1w}, M_{2w}) the pairs of ancestry states for each window w across the genome, then we will apply the HMM of this ancestral population separately across the genome with $\binom{K}{2}$ possible states corresponding to the pair of ancestries S_w. Therefore, each state S_w emits a probability G_w by sum(M_{1w}, M_{2w})P$(H_{1w}|M_{1w})$P$(H_{2w}|M_{2w})$ where P$(H_{1w}|M_{1w})$ is the probability of emitting the haplotype segment H_{1w} under the Hidden Markov Model for ancestry M_1 based on equation (1.4.9) with a pair of haplotype (H_{1w}, H_{2w}) compatible with G_w. Then, the transition probability from state (M_{1w}, M_{2w}) to state $(M_{1w'}, M_{2w'})$ where w' =$[i + L, i + 2 \times L)$ is given as follows:

$$P(M_w, M'_w) = \begin{cases} \theta = 10^{-8} \times B & \text{if unordered ancestry pairs } (M_{1w}, M_{2w}) \text{ and } (M_{1w}, M_{2w}) \text{ differ by one ancestry} \\ \theta^2 & \text{if both ancestries differ} \\ 1 - 2\theta - 3\theta^2 & \text{if the respective ancestry pairs are the same} \end{cases} \tag{1.4.10}$$

B is the length in base-pairs between windows.

1.5 Genome-wide Data: Current Challenges and Opportunities

The amount of genome-wide data has increased massively over recent years with rapid advances in sequencing technologies. Through this data, we can anticipate new ancestry inference as well as refining new inference methods in order to take advantage of all the information found within the chromosome. Studying populations with mixed ancestry has become useful to identify complex traits, but it is important to consider the current challenges that genome-wide data is facing and some opportunities to remedy the situation (Padhukasahasram, 2014). One of the

current challenges is the access to the maximum amount of information within the genome to better infer the ancestry (Medina-Gomez et al., 2015). Moreover, since a decade, genome-wide association studies mainly focused on populations of European ancestry descend. According to Medina-Gomez and colleagues (2015), among the 1734 GWAS papers indexed in the GWAS catalogue, 66 included individuals from European ancestry, 34 included Non-Europeans only (most of these carried out in Asian populations), and 12 included both Europeans and Non-European individuals. Given the fact that Africa is the home land of the human species, the continent is characterized by high levels of haplotype diversity and low levels of LD, and this has both advantages and disadvantages from a statistical genetics perspective. Despite the fact that it is possible to map causal variants with efficient tools, screening the genome using current SNP genotyping approaches is challenging for disease associations due to the low level of LD. With the inclusion of African genetic data, one can address study design in the conduct of GWAS. However, interpreting data in African populations can be challenging when it comes to ascertaining the markers due to decrease levels of tag-SNP transferability and genotype imputation error (Peprah et al., 2015; Tishkoff and Verrelli, 2003). These challenges, open opportunities to involve more multi-ethnic populations, particularly African to increase the power in association in studies and the understanding of human variation.

1.6 Overview of Genomic Dating

Studies of admixed populations have become of increasing interest for population geneticists because of the inference of population history and past demographic processes. A key parameter of interest to quantify in this field is the date of the mixture between two or more populations. One of the features of human evolution is the migrations of populations from one place to another which yields human mixture with other populations. During the mixture process, the chromosomes of the new generation contain continuous blocks inherited from parental populations that will breakdown through successive generations. Methods of ancestry inference explicitly infer these blocks, so that, based on the distribution of ancestry block length, we can determine the time of admixture. However, a major limitation is that pinpointing ancestry along the genome of a complex multi-way admixed population such as the South African Coloured population is currently an unsolved problem (Chimusa et al., 2013). Existing methods may attain high accuracy on average but may suffer from spurious deviations in average local ancestry at particular regions (e.g. regions in which the modelled ancestral population is unusually different from the true ancestral population due to the historical action of natural selection). These spurious deviations may lead to bias in dating admixture event. Addition for case-control studies, these deviations would be present in both affected and unaffected individuals, and would lead to spurious mapping of genes underlying ethnic difference in disease risk.

Recent work has shown the utility to date admixture events through LD statistics, by fitting an exponential decay for the admixture LD with the genetic distance (Moorjani et al., 2011; Patterson et al., 2012; Loh et al., 2013). Others made use of the haplotype approach by assigning haplotype segments of the admixed individual to those that closely match the

ancestral populations, and plot the genetic distance against a measure of how often a pair of haplotype chunks separated by this distance come from each respective donor in the ancestral population. They consequently fit an exponential distribution with unknown rate λ, which is the estimate of the date of admixture(Lawson et al., 2012; Hellenthal et al., 2014).

Some scholars use a PCA-based approach to obtain the block-like admixture signal across each chromosome for each parental population and through the width of the admixture block, the time since admixture is inferred by comparison with the obtained frequency estimate from simulated data(Pugach et al., 2011; Sanderson et al., 2015). The details of the aforementioned approach will be discussed in the next chapter. Xu and colleagues analysed the genetic admixture in the Uygurs, using the breakpoint recombination approach (Xu et al., 2008) to estimate the date of admixture, assuming a two-way single point admixture. Recently, Jin and colleagues (2012) derived a method describing the admixture dynamics in order to infer distinct admixture events by taking into account distinct admixture models: the Hybrid Isolated model (HI), the Gradual Admixture model (GA) and the Continuous Gene Flow model (CGF). Later, they inferred a theoretical distribution of ancestral tracks under HI and GA models (Jin et al., 2014; Ni et al., 2016).They suggested a method that describes the ancestry history dealing with multiple ancestral populations and multiples waves of admixtures by exhibiting the length of ancestral tracks and using the Akaike information criterion or the likelihood ratio test to select the best admixture model developed by Jin and colleagues previously. For this, they estimated the date of admixture. However, the capability to infer population history is greatly influenced by the kind of input data we have for admixture dating inference; so it is important to use an appropriate method depending on the pattern and the dynamic of the admixture to infer ancestry (Padhukasahasram, 2014). The next chapter describes some of the existing methods. the outline of the project is presented as follow:

In chapter 2, we review various admixture dating methods with mathematical description. The object of chapter 3 will be to evaluate some of these tools of admixture dating which include ROLLOFF, ALDER, stepPCO and GLOBETROTTER using 2-way single point, 3-way single point and 3-way multi-point approach taking data from the HapMap project phase 3. We will perform various simulations based on a pipeline created for this project. In chapter 4, we will apply the best methods into real data which include Africans Americans, Mexicans Americans and Luhya from the 1000 Genome Project Phase 3. Finally, chapter 5 will focus on the conclusion and the discussion of the project.

2. Models of Dating Admixture Events

2.1 Overview

With the advent of genome sequencing and the development of computational tools, various methods to date the admixture events have been developed. These methods are utilizing the information from the genome of current populations and are presenting the equivalent ancient populations known to be involved in the admixture processes. In this chapter, we describe some of the existing methods of dating the admixture events based on various approaches which include the haplotype-based (Hellenthal et al., 2014), the linkage disequilibrium-based (Moorjani et al., 2011; Loh, 2013), the principal component Analysis(PCA)-based (Pugach et al., 2011) and the ancestry block-size distribution(or tract length)-based (Xu et al., 2008; Gravel, 2012).

2.2 The Principal Component-based Method

Principal Component Analysis (PCA) is a tool or a statistical technique which detects and adjusts population stratification on genome-wide analysis scale (Price et al., 2006a; Patterson et al., 2006). PCA simplifies complex data by detecting new variables (principal components) which are linear combination of the original variables in a multidimensional data set and cluster individuals into groups reflecting their genetic heritage. If data are well standardized, those principal components are the directions whereby the sample population shows the greatest variation. This technique has been used to illustrate population-specific variation that may have risen as a result of varying frequencies of minor alleles in genetically distant ancestries (Patterson et al., 2006; Pritchard et al., 2000) as well as to model ancestry difference between cases and controls (Ringner, 2008). Application of PCA also include the analysis of microarray data in search of outlier genes (Price et al., 2006b) as well as the analysis of other types of expression data (Ringner, 2008).

The PCA-method for dating admixture is decomposed unto two analytic parts. First, it applies stepwise principal component analysis (stepPCO) to pick up the block-like admixture signal across each chromosome of each individual in the admixed population in order to determine the structure of the population. This is done in such a way that the result is introduced into a new method that estimates the date of admixture. StepPCO (Pugach et al., 2011), generates the above block-like signal and computes the Discrete Wavelet Transform (DWT) to estimate frequencies (width) and position as sum of waves within each signal. The dominant frequency, which is an indirect estimate of the average width of the admixture block, is compared to the obtained one, and generated from data simulation using the admixture rate observed in the empirical data (Pugach et al., 2011; Sanderson et al., 2015). However, this method could not be extended to multi-way admixture. Nevertheless, one should note that the utilization of phased data in this method derive haplotypes with significant switch errors at the level of the entire chromosome (Pugach et al., 2011).

The value of the dominant frequency could be underestimated if the ancestral populations are very close to each other (Pugach et al., 2011). Mathematically, let **e** be an individual chromosome, **e** is defined by the collection of coordinates $(e_i)_{i=1}^{N}$ where N is the total number of SNPs in the chromosome **e**. Each e_i takes the value -1 or 1 if the SNP is homozygote and 0 if heterozygote. Let E be a real N-dimensional vector space with canonical basis. We define **w**, a sliding window along each chromosome referring to the contiguous sub-range of SNPs as follows:

$$\mathbf{w} = \{w_{\text{first}}, w_{\text{first}} + 1, \ldots, w_{\text{last}} - 1, w_{\text{last}}\}, \tag{2.2.1}$$

. If A is an axis of the space E spanned with a non-zero vector with positive components $\vec{u} = (a_i)_{i \in w}$, given **w** and **e**, the measurement of relation between them with respect to the axis A is defined as follows:

$$M_{\mathbf{w}}(\mathbf{e}) = \frac{\sum_{i \in \mathbf{w}} a_i e_i}{\sum_{i \in \mathbf{w}} |a_i|}, \tag{2.2.2}$$

Note that the choice of the spanning vector $\vec{u}$ does not influence the resulting value of M_w.

Given x, a particular physical position along the chromosome, we say that w is centered at a point x with width l if it encompasses all the SNPs that are situated within the distance $\frac{l}{2}$ from x as follows:

$$\mathbf{w}_l(x) = \left\{ w : |x - p_w| \leq \frac{1}{2} \right\}; \text{where } p_w \text{ is the physical position of the SNPs w.} \tag{2.2.3}$$

Consider an admixed population R with two ancestral population P and Q, we define the principal axis A1 spanned by a non-zero vector with coordinate $(a_i)_{i \in w}$, let denote PC1 the first principal component with its coordinates for the parental populations P and Q. We find the average value of SNPs within each window using the coefficients of the first principal component PC1 as weights. We normalize it so that the ancestral populations may correspond to values with means of 1 and -1 respectively. The stepPCO signal of an individual chromosome e from the admixed population R is a vector of measurement defined by

$$\mathfrak{S} = (M_{w_k}(\mathbf{e}))_{k=1}^{K} \tag{2.2.4}$$

where K is the number of windows (sufficiently large as a power of 2) and w_k is the window centred at x_k. The segment's size of the chromosome r which corresponds to the bin that belongs to one of the ancestral populations is given by each component of the vector $\mathfrak{S}$.

The second part of the stepPCO method deals with the wavelet transform analysis which provides a decomposition of the data in terms of location along the genome on the x-axis and wavelet scale on the y-axis. Consider a 2^L-dimensional vector space $V = \mathbb{R}^{2^L}$ in which we define a scalar product:

$$\langle (u_i), (v_i) \rangle = \sum_{i=1}^{2^L} u_i v_i \tag{2.2.5}$$

The wavelet (ω_l, p) which is normal to the system $2^L - 1$ vectors in V, is an orthogonal basis of the vector space V(V is indexed by the level $l = 1, 2, \ldots, L$ and the position $p = 1, 2, \ldots, 2^{l-1}$). The coefficient of the wavelet is defined as follows:

$$(\omega_l, p)_k = \begin{cases} 1, & \text{if } (p-1)2^{L-l+1} + 1 \leq k \leq 2^{L-l+1} + 2^{L-l}, \\ -1, & \text{if } (p-1)2^{L-l+1} + 2^{L-l} + 1 \leq k \leq p2^{L-l+1}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2.6}$$

For $\gamma = (\gamma_i)_{i=1}^{2^L}$ a discrete time signal, we define filters which its wavelet have to pass for efficient evaluation to obtain at each step the wavelet coefficient at each level and a down-sampled signal for the next evaluation. This filter is defined by:

$$\begin{cases} \text{For } k = 1, \ldots, 2^{L-1} \begin{cases} \gamma'_k = \frac{1}{2}(\gamma_{2k} + \gamma_{2k-1}), & \text{low pass filter,} \\ wt_{L,k}(\gamma) = \frac{1}{2}(\gamma_{2k} + \gamma_{2k-1}), & \text{high pass filter,} \end{cases} \\ \text{For } k = 1, \ldots, 2^{L-2} \begin{cases} \gamma''_k = \frac{1}{2}(\gamma'_{2k} + \gamma'_{2k-1}), & \text{low pass filter,} \\ wt_{L-1,k}(\gamma) = \frac{1}{2}(\gamma'_{2k} + \gamma'_{2k-1}), & \text{high pass filter,} \end{cases} \end{cases} \tag{2.2.7}$$

After filtering, the signal γ will be written as a linear combination of the wavelet coefficient plus an additional value called $(\gamma_{average})$ which corresponds to the aggregate value of γ as written below:

$$\gamma = \gamma_{average} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \sum_{l,k} wt_{l,k}(\gamma)\omega_{l,k}, \tag{2.2.8}$$

The above equation correspond to the inverse wavelet transform iwt given by:

$$\gamma = \text{iwt}(wt_{l,k}, \gamma_{average}) \tag{2.2.9}$$

A wavelet scale that has high frequency values, exhibits the noise of the signal and needs to be removed completely by considering a collection $(t_{l,k})$ of the threshold level, one for each of the wavelets in the wavelet decomposition. A threshold filter $\mathfrak{T}$ is defined as follows

$$\widetilde{wt} = \mathfrak{T}(wt), \quad \text{where} \tag{2.2.10}$$

$$\widetilde{wt}_{l,k} = \begin{cases} 0 & \text{if } |wt_{l,k}| \leq t_{l,k}, \\ wt_{l,k} & \text{otherwise.} \end{cases} \tag{2.2.11}$$

Having the discrete signal γ and its corresponding wavelet coefficient $wt_{l,k}$, one can compute its wavelet summary by averaging its wavelet coefficient at each level as given below:

$$s_l = \frac{\sum_{k=1}^{2^l} |wt_{l,k}|}{2^l}, \quad l = 1, 2, \ldots, L. \tag{2.2.12}$$

and henceforth deduct the wavelet "center" which is the average over all the scales of all the wavelets as follows :

$$C(\gamma) = \frac{\sum_{l=1}^{L}(l \times s_l)}{\sum_{l=1}^{L} s_l}. \tag{2.2.13}$$

This wavelet "center" is the indirect measure of the average width of admixture blocks. Practically, the wavelet scale value depends on the length of the chromosome and stepPCO's manual suggests a maximum scale of 7, 6 and 5 for chromosomes 1 to 5, 6 to 20 and 21 to 22, respectively (Pugach et al., 2011).

While Pugach and colleagues (2011) introduced the Discrete Wavelet Transform (DWT), Sanderson et al.(2015) extended the method of Pugach by introducing the Maximal Overlap Discrete Wavelet Transform (MODWT), which compute the Average block size metric (ABS). This measure estimates the time of an admixture event using the local ancestry along the genome. This method improves the statistical power to differentiate between admixture processes. Compared to stepPCO, the method has the advantage of fewer running time, but offers reduced uncertainty in model parameter estimates. In contrast to stepPCO, to compute the wavelet summary, it does not make use of the scale as the threshold, and it is limited only to two-way admixture (Sanderson et al., 2015).

2.3 Linkage Disequilibrium-based Method

During the admixture process, gene flow between ancestral populations in the admixed population displays linkage disequilibrium (or allelic correlation) relative to the ancestral population and the rate of decay of the linkage disequilibrium(LD) depends on the proportion of admixture, the recombination rate and the time since admixture occurred. Various methods have used this approach to estimate this time.

Moorjani and colleagues (2011) developed ROLLOFF (Moorjani et al., 2011) which computes the weighted correlation between a pair of SNPs that reveals their allele frequency differentiation in the ancestral populations. The method obtains an estimate of the time since admixture happened by exploring the change in this correlation with increasing the genetic distance among these markers. Further, ROLLOFF fits an exponential distribution to the decay of the correlation by least-squares.

Mathematically, given diploid SNPs x_1 and x_2 with their respective alleles frequencies $w(x_1)$ and $w(x_2)$ separated by a distance d Morgans, We define the allele frequency difference between two ancestral populations as a weight function $w(x_1, x_2)$. Assuming that they are in LD, let define $z(x_1, x_2)$ the covariance between SNPs x_1 and x_2, as we will explain below. We then compute the linkage disequilibrium statistic. This covariance is expected to be positive, is correlated to the product weight function $w(x_1) \cdot w(x_2)$; and the correlation coefficient, and is proportional to e^{-nd} where n is the rate of decay interpreted as the admixture date in generations Moorjani et al. (2013).

In the presence of no missing data in the individual genotype:

1. The method first computes the Pearson correlation ρ for the diploid genotype at SNPs x_1 and x_2.

2. Let N be the size of the population genotypes when there is no missing data; the covariance is given by

$$z(x_1, x_2) = \sqrt{N\rho}, \quad \text{where } N \geq 4, \tag{2.3.1}$$

   If we succeed to "prune" ρ to fall within $[-0.9, 0.9]$ then we apply the Fisher z-transform to tail the behaviour of z by :

$$z(x_1, x_2) = \frac{\sqrt{N-3}}{2} \log\left(\frac{1+\rho}{1-\rho}\right), \tag{2.3.2}$$

3. The method computes the measure of correlation between the pair of SNPs x_1 and x_2 as follows :

$$R(d) = \frac{\sum_{|x_1-x_2|=d} w(x_1)w(x_2)z(x_1, x_2)}{\left[\sum_{|x_1-x_2|=d}(w(x_1)w(x_2))^2 \sum_{|x_1-x_2|=d}(z(x_1, x_2))^2\right]^{1/2}}, \tag{2.3.3}$$

   The value of $R(d)$ is the measure of LD due to population mixture (Moorjani, 2013).

After computing $R(d)$ between all pairs of SNPs x_1 and x_2, we obtain the date through the fitting of an exponential decay with an affine term as defined below :

$$R(d) \approx R_0 e^{-nd} + c, \tag{2.3.4}$$

and n is the estimated number of generations since admixture (Patterson et al., 2012; Loh et al., 2013). This method has been incorporated into ROLLOFF in the ADMIXTOOL package developed by Moorjani and colleagues.

After this method, Loh et al.(2013) extended the work of ROLLOFF and developed ALDER. ALDER computes a new LD statistic but maintain the same methodology as ROLLOFF. Given the same approach as defined in ROLLOFF, the new LD statistic is defined as follows :

$$A(d) = \frac{\sum_{x_1,x_2 \in S(d)} w(x_1)w(x_2)z(x_1, x_2)}{|S(d)|}, \tag{2.3.5}$$

where $S(d)$ is the set of all pairs of SNPs given by

$$S(d) = \left[(x_1, x_2) : -d + x < x_1 - x_2 < d - x\right], \tag{2.3.6}$$

where x is a discretization parameter (Loh et al., 2013; Moorjani et al., 2011, 2013).

This new LD statistic (equation (2.3.5)) removes the bias in the estimated date of admixture, and makes the computation more flexible such that the admixture proportion in ALDER is related to the amplitude in the exponential decay fitting (Loh et al., 2013; Moorjani, 2013).

All these methods compute the estimate but are limited to two-way admixture. Pickrell et al. (2014), extended the method of ALDER to create a method that addresses multiple admixture events (as a mixture exponential decay) in the population's history using background linkage disequilibrium. The fit performs well when the model found older admixture events, but for recent admixture, the exponential decay fitting curves of the data becomes poor (Pickrell and Reich, 2014). Theoretically, the model curve is defined by :

$$a_{ij}(d) = K_{ij} + \sum_{k=1}^{n} \left(m_{ijk} \exp\left(-t_k \dot{d}\right)\right) + e_{ij}(d) \tag{2.3.7}$$

where a_{ij} is the weighted LD statistic between each pair of population i and j, K_{ij} is the affine term estimated for each pair of populations, C_{ijk} is the amplitude of the k^{th} exponential term for populations i and j, t_k is the estimated number of generations at the k^{th} admixture event, e_{ij} is the error in fitting curve between populations i and j following the normal distribution and d is the genetic distance (Pickrell and Reich, 2014).

2.4 The Haplotype-based Method

The haplotype-based method makes use of the haplotypes, which are more informative for ancestry than individual markers.The haplotype method came as a result of one of the weakness of the linkage disequilibrium method which does not really capture all information about ancestry in data from genome sequencing. Price et al. (2009), therefore developed HAPMIX to estimate the time since admixture occurred. The method is based on the number of calculated ancestry transitions, that is the number of breakpoints. HAPMIX determines whether the admixed descendant has 0, 1 or 2 of alleles of a particular ancestry at a given locus by viewing each haplotype block as a representative sample from the predefined ancestry and computes the likelihood of the haplotype of the admixed individual coming from one reference population versus others.

Using a Hidden Markov Model(HMM) assumption for a given state, the likelihood of an observed allele in an admixed individual genome is given by :

$$p^*_{ijk}(s) = \begin{cases} \theta_i \delta(t_{jk} = 0) + (1 - \theta_i)\delta(t_{jk} = 1) & \text{if } i = j \\ \theta_3 \delta(t_{jk} = 0) + (1 - \theta_3)\delta(t_{jk} = 1) & \text{if } i \neq j \end{cases} \tag{2.4.1}$$

where $\theta_i,\ i = 1, 2, 3$, are the mutation parameters, s is the offspring chromosomal haplotype site, t_{jk} denote if at genotyped SNPs, an individual k from the offspring s copy a segment of

genome from a reference population j, and δ is the probability for a copy of ancestry genotype segment of an individual k to have a single pair of haplotypes (Price et al., 2009). The point estimate λ of the number of generations is given by:

$$\hat{\lambda} = \frac{N}{4\mu(1-\mu) \times \vartheta} \tag{2.4.2}$$

where ϑ is the total Morgan length, μ the proportion of admixture, and N the ancestry transition, which is the observed number of breakpoints (Price et al., 2009). One should note that HAPMIX uses locus-specific ancestry to date the admixture, and the above formula of the estimate of the date of admixture depends on the number of ancestry transitions based on a Hidden Markov model(HMM). However, Pugach and colleagues (2011) criticized this technique of inferring the population history based on the number of breakpoints, emphasizing that the formula for estimating the date of admixture from the HAPMIX method should expect that the date increases linearly as the number of breakpoints increases, which is in contrast to the results from studies of the admixture in the Mozabites (Pugach et al., 2011). Nevertheless, this technique stabilizes the number of breakpoints leading to underestimate admixture dates. Furthermore, in the case of closed-related population, there is a need to have enough power to reliably assign chromosomal segments to an ancestral population by defining large genomic windows, which correspondingly reduces detection of closely-spaced breakpoints (Pugach et al., 2011).

Later, Hellenthal et al.(2014) developed the software GLOBETROTTER based on the works of Lawson et al.(2012), which also uses haplotype-based method to infer the time since admixture occurred. This approach considers each individual in a sample as a recipient, whose chromosomes are reconstructed using chunks of DNA donated by the other individuals. The first part of this method (CHROMOPAINTER) uses a Hidden Markov Model to break down the chromosomes of each individual from the admixed population into "chunks". In addition, based on similarity, it assigns each chunk to a single individual from one of the ancestral populations.

In the second part, the GLOBETROTTER method assigns haplotype segments of the admixed populations to different ancestral populations, and identifies and infers the admixture which employs the co-distribution of such segments from different ancestral populations (Lawson et al., 2012; Hellenthal et al., 2014). For each pair of ancestral populations, an exponential decay distribution curve, which quantifies each genetic distance, determine how often a pair of haplotype chunks, separated by the genetic distance, comes from each pair of populations.

The decay rates of these curves are utilized to ascertain whether or not an admixture event occurred and to infer its time. Meanwhile the proportion contribution of the ancestries are determined by the amplitude of the curve. In the case of the evidence of admixture, GLOBETROTTER further examines whether the data fits a single exponential decay (i.e. one-point admixture event), or a combination of exponential decays (i.e., several admixture events or continuous admixture over a longer period). To describe this, let λ be the rate of decay assuming a single time point mixture model where two populations interbreed at a particular point of time λ defined above. By the time admixture occurred, for each pair of

ancestral groups (M,N) separated by a genetic distance d, the coancestry vector $v_{MN}(d)$ from the CHROMOPAINTER output is labelled and weighted to a new coancestry $\Psi_{MN}(d)$ and from there we fit the observed coancestry curve defined by :

$$\Psi_{MN}(d) = \tau_{MN} + \sigma_{MN} \times e^{-\lambda d} \tag{2.4.3}$$

λ is interpreted as the date of admixture derived from the fitting of the coancestry curve. The parameters $\tau_{\hat{MN}}$, $\sigma_{\hat{MN}}$ and $\hat{\lambda}$ can be estimated trough the minimization of the square errors sum

$$\sum_{(M,N)} \sum_{d} (\Psi_{MN}(d) - \tau_{MN} - \sigma_{MN} \times e^{-\lambda d})^2 \tag{2.4.4}$$

GLOBETROTTER also suggests two important clusters of admixture, each may be composed from several populations, which together represent the genetic structure of the ancestral population. The exchange of genes can be distinguished through multi-marker haplotype data by changing migration rate over time. However, the difficulty to infer older dates of admixture resides in the fact that haplotype-based method needs to model background LD properly to avoid biased estimates (Moorjani, 2013; Pool and Nielsen, 2009).

2.5 Ancestry Block-size Distribution Method (Track Length)

The local ancestry approaches analyses the chromosomes in the admixed individual aiming to identify blocks of ancestry inherited directly from each parental population. As recombination breaks down ancestry blocks through successive generations, it is possible to infer the time of admixture from the track length distribution (Tang et al., 2006; Sankararaman et al., 2008; Price et al., 2009; Lawson et al., 2012; Loh et al., 2013).

Inferring the history of populations using the ancestral block was primarily developed by Pool and Nielsen (2009), in order to determine the distribution of ancestral blocks on a the basis of a single point of admixture. The method assumes that migrant tracks do not recombine together and irrealistically, ignores the effect of the end of the chromosome. After a number of generations t from the advent of admixture, the distribution of the track length follows an exponential distribution with means $\frac{1}{t}$ given by the equation below :

$f(x,t) = te^{-tx}$, t is the number of generations and x is the length of an admixture tract. (2.5.1)

Gravel (2012) applied the same study to multiple ancestral populations, but these methods demand accurate identification of the boundaries of admixture segments, which is not always available for fitting the distribution of ancestral segments. Jin and colleagues (2012) developed a method to unravel the admixture dynamics in order to infer distinct admixture events by taking into account distinct admixture models (Figure (2.1)) including the Hybrid Isolated

model (HI), Gradual Admixture model (GA) and the Continuous Gene Flow model (CGF). They later inferred the theoretical distributions of ancestral tracks under the HI and GA models (Jin et al., 2012, 2014).

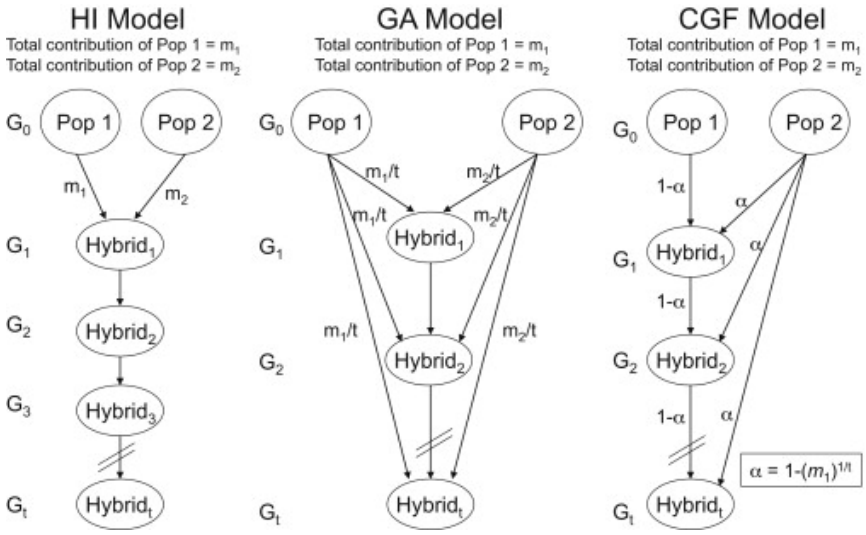

Figure 2.1: Different Admixture Model from the paper of Jin et al. (2012).

Jin et al.(2014) improved the equation suggested by Pool and Nielsen by including the population proportion from the ancestral population under a two-way admixture process considering the Hybrid isolation(HI) model and the gradual admixture(GA) model. According to Jin et al.(2014), if two populations A and B interbred at a particular time in the past, the recombination broke down the chromosome into different pieces or tracts of distinct parental populations. The probability that a given ancestral segment from population A could recombine with those from the same parental population was m and the probability that a given ancestral segment from population A could recombine with those from population B is (1-m) (Gravel, 2012; Jin et al., 2014; Ni et al., 2016). The distribution of the length of ancestral chromosomal segments (tracts) after T generations under the HI model follows an exponential distribution with means $\frac{1}{(1-m)T}$ as given below:

$$f(x,t) = (1-m)Te^{-(1-m)Tx}. \tag{2.5.2}$$

Under the GA model, when a chromosome from population A contributes a proportion of admixture m (the contribution of the population B will be $(1-m)$), t generations ago $(1 \leq t \leq T)$, the relative gene flow at each generation from population A will be $\frac{m}{t}$ (respectively $\frac{1-m}{t}$ for population B). Assuming that the chromosome ends of the population are ignored, therefore we expect the chromosome of population A during the gene flow to split into $(1-m)t$ pieces per unit length

$$f(x,t) = \frac{\int_0^T (1-m)te^{-(1-m)tx}(1-m)tdt}{\int_0^T (1-m)tdt}. \tag{2.5.3}$$

To infer the date of admixture, one needs to quantify the genetic contribution of each ancestry and the pattern of ancestral chromosomal segments using available software such as HAPMIX. However limitations arise especially with the assumption of infinite length of chromosomes in the theoretical framework and also in the fact that the method deals with two-way admixture (Jin et al., 2014). Ni et al.(2016) extended the method of Jin and colleagues and suggested one that describes the ancestry history for multiple ancestral populations and multiple waves of admixture. The method identifies the length of ancestral tracks and uses the Akaike information criterion or the likelihood ratio test to select the best admixture model developed by Jin and colleagues previously, and consequently estimate the date of admixture. However, its limitation lies in how to identify the correct model in the cases of recent admixture and minor ancestral proportions. Ni and colleagues described the model distribution of ancestral tracts based on the continuous gene flow (CGF) model whereby the ancestral populations can contribute either in one pulse or in a continuous pulse. This model can be subdivided into two cases:

1. The continuous gene flow donor (CGFD) model for one pulse contribution.
2. The continuous gene flow recipient (CGFR) model for the continuous contribution of one parental population.

For the CGFR model, assuming that the gene flow descends from population B, the ancestry proportion for the parental population is given by $m^{\frac{1}{T}}$ for population A at the first generation and 0 for the rest of the generations and for population B, the ancestral contribution is given by $1 - m^{\frac{1}{T}}$ for $2 \leq t \leq T$. Then, the distribution of ancestral tracts for the two ancestral populations is given by :

$$f_1(x,t) = \left(T - \frac{(1-m)m^{\frac{1}{T}}}{1-m^{\frac{1}{T}}}\right) exp\left(-x\left(T - \frac{(1-m)m^{\frac{1}{T}}}{1-m^{\frac{1}{T}}}\right)\right) \text{ for population A .} \quad (2.5.4)$$

$$f_2(x,t) = \frac{\sum_{t=1}^{T} m^{\frac{-t}{T}}\left(m^{\frac{t}{T}} - m^{\frac{(T+1)}{T}}\right)^2 exp\left(-x\left(\frac{m^{\frac{t}{T}} - m^{\frac{(T+1)}{T}}}{1-m^{\frac{1}{T}}}\right)\right)}{\sum_{t=1}^{T}(1-m^{\frac{(T+1-t)}{T}})(1-m^{\frac{1}{T}})} \text{ for population B .} \quad (2.5.5)$$

To derive the distribution of the ancestral tract for the continuous gene flow donor (CGFD) model, one needs to replace m by $1-m$ in the equation (2.5.5) to yield those for population B and A respectively (Ni et al., 2016).

To estimate the time T, we first need to infer the ancestral tracts and calculate the estimate of admixture proportion $\hat{m}$ by dividing the total length of tracks from population A by the total length of tracks assuming a model of admixture. After that, we use the Akaike information Criterion (AIC) to select the optimal model of the admixture process which include the HI, GA, CGFD and CGFR models, and then maximize the likelihoods of the optimal model to estimate the optimal time $\hat{T}$ (Ni et al., 2016). This method is incorporated in the software Admixinfer developed by Ni and colleagues.

After reviewing various available methods from the literature, in the next chapter, we will assess some of the methods of admixture dating, which include ROLLOFF, ALDER, stepPCO and GLOBETROTTER.

3. Assessment of Different Admixture Dating Methods

3.1 Introduction

With the increased availability of high-throughput genetic data, several studies have revealed that many world populations are admixed as a result of migrations pattern and genetic drift (Pickrell and Pritchard, 2012; Price et al., 2007; Patterson et al., 2012; Baran et al., 2012). As described in previous chapter, some researchers have attempted to develop admixture dating methods to estimate the admixture events based on either 2-way or 3-way admixture scenarios (Pugach et al., 2011; Moorjani et al., 2013; Gravel et al., 2011; Hellenthal et al., 2014; Jin et al., 2012), but very few studies have focused on the assessment of these methods, especially for admixture events beyond 15000 years and complex multi-way admixed populations. In this chapter, through simulations, we assess various admixture dating methods, which include ALDER, ROLLOFF, stepPCO and GLOBETROTTER based on time of admixture events under the simulation of 2 and 3-way admixture scenarios. This assessment will allow us to compare the estimate from each tool versus the true time of admixture events in our simulated data. The result of this chapter will also allows to identify the best method.

3.2 Data Description and Simulation Framework

3.2.1 Simulation Framework

Table 3.1: Data used for Simulations.

Label	Number of Samples	Description	Source
CEU	105	Utah residents with Northern and Western European ancestry from the CEPH collection	HapMap phase 3
YRI	203	Yoruba in Ibadan, Nigeria	HapMap phase 3
CHB	137	Han Chinese in Beijing, China	HapMap phase 3

We simulated a genome of 600 individuals of mixed ancestry coming from the populations highlighted from Table (3.1). To generate these individuals, our simulation framework uses $2n$ ancestral haplotypes where n is the minimum sample size among the parental populations. We independently expanded each ancestral populations to a total size of $n = 600$ plus its original size with 278972 Single Nucleotides Polymorphisms (SNPs) in common, using the model of exponential growth following the expansion model from Rogers and Harpendings(Chimusa et al., 2013). The model is implemented using three parameters $a_0 = 2 * P_0 * \mu$, $a_1 = 2 * P_1 * \mu$

and $\lambda = 2 * \mu *$ t where P_0 is the population size of an initial population assuming to grow exponentially to a new population size P_1 at a time t generation back from present. The mutation rate μ , is the per-generation probability that a mutation strikes a random nucleotide along the genome.

3.2.2 Data Description

We independently simulated 140 individuals with single point 2-way and 3-way admixture scenarios which include 2-way and 3-way from the expanded data.from developed in the previous section. We split the expanded data into two parts, the first part were used to simulate admixture events, and the rest were utilized to assess the admixture dating methods. The above simulation was based on different number of generations (N) since admixture occurred which include N = 5, 20, 50, 100, 200, 450, 600 and 800, we simulated a single point admixture with specific proportions of contribution from the ancestral populations in a 2-way (CEU and YRI) and 3-way (CEU, YRI and CHB) single point scenario (Table (3.2)). However, in real human populations, admixture typically occurs continuously over an extended period of time. We deduced the average number of breakpoints and the genotypes of the admixed population from the model of expansion described in the previous section. For the 2-way admixed scenario, we merged the admixed genotypes data with the ancestral populations and we applied existing admixture dating methods which include ROLLOFF, ALDER, stepPCO and GLOBETROTTER to determine the date of admixture. In the 3-way single point admixture scenario, ROLLOFF, ALDER and stepPCO can only use two ancestral panels, we considered all possible pairwise combination that we merged with the admixed population and then we applied these methods. For GLOBETROTTER, we combined all the ancestral panels with the generated admixed population from the simulation before applying the method.

Table 3.2: Ancestry proportions for Single-Point Admixture Scenario (N is the number of generations.)

3-way Admixture CEU YRI CHB				
140	admixed	YRI	CHB	CEU
1	0	0.40	0.30	0.30
N	1	0.0	0.0	0.0

(a) 3-way Single point Admixture simulation Scenario.

2-way Admixture CEU-YRI			
140	admixed	YRI	CEU
1	0	0.50	0.50
N	1	0.0	0.0

(b) 2-way Single-point Admixture simulation Scenario.

3.3 Simulation Results from Different Dating Admixture Event Models

3.3.1 Assessing Dating Admixture Using ROLLOFF

The protocol to estimate the date of admixture has been developed in the previous chapter 2 section 2.3. ROLLOFF also computes an approximately normally distributed standard error by carrying out Weighted Jackknife analysis, this enables us to compute the 95% confidence interval of the estimate(Moorjani et al., 2013).

Table 3.3: 2-way and 3-way Single-point Admixture Results for ROLLOFF. Values in the table are in number of generation (and its ± 95% CI standard error) taken one generation is 35 years.

2-way ROLLOFF Results CEU-YRI (p-value $\leq$ 0.001)								
True date	5	20	50	100	200	450	600	800
Expected date	1.3 ± 0.05	12.8 ± 0.2	34.6 ± 0.5	74.5 ± 1.1	166.8 ± 3.3	437.8 ± 28.6	634.3 ± 84.3	836 ± 303.2
3-way ROLLOFF Results (p-value $\leq$ 0.001)								
Expected CEU-YRI	1.7 ± 0.05	12 ± 0.17	37 ± 0.6	76.8 ± 1.3	175.5 ± 4.4	454 ± 36.1	576 ± 99	758 ± 348
Expected CEU-CHB	1.8 ± 0.06	13 ± 0.2	37 ± 0.7	76 ± 1.8	178 ± 7	427 ± 50.8	568 ± 143	914 ± 717
Expected YRI-CHB	1.8 ± 0.06	12 ± 0.2	36 ± 0.7	75 ± 1.4	172.5 ± 4.4	463 ± 36	538 ± 78	1190 ± 612

In the 2-way admixture scenario, the results shows that ROLLOFF's date underestimate the true date of admixture (see Table (3.3)) for recent admixture events. For older generations (for 600 and 800 generations) in the 2-way admixture scenario, the results overestimate the true. Moreover, the exponential decay fitting curve tends to become like an L-shape for older dates as a result of the fact that the weighted LD for the pairs of SNPs tend to 0 when the number of generations increases (Figure (3.1)). Particularly, at generation 800, the fitting curves become inconsistent with the data. Although the estimates of the date for each generation are significant, the range of the 95% confidence interval grows wider for older dates. This shows a lot of variance in the estimate of the simulation's date for older admixture events. In the 3-way admixture scenario, the same pattern occurs as in the 2-way admixture scenario for each pair-wise comparison except that the simulation's result overestimates the true date at generations 450 and 800 (Table (3.3)).

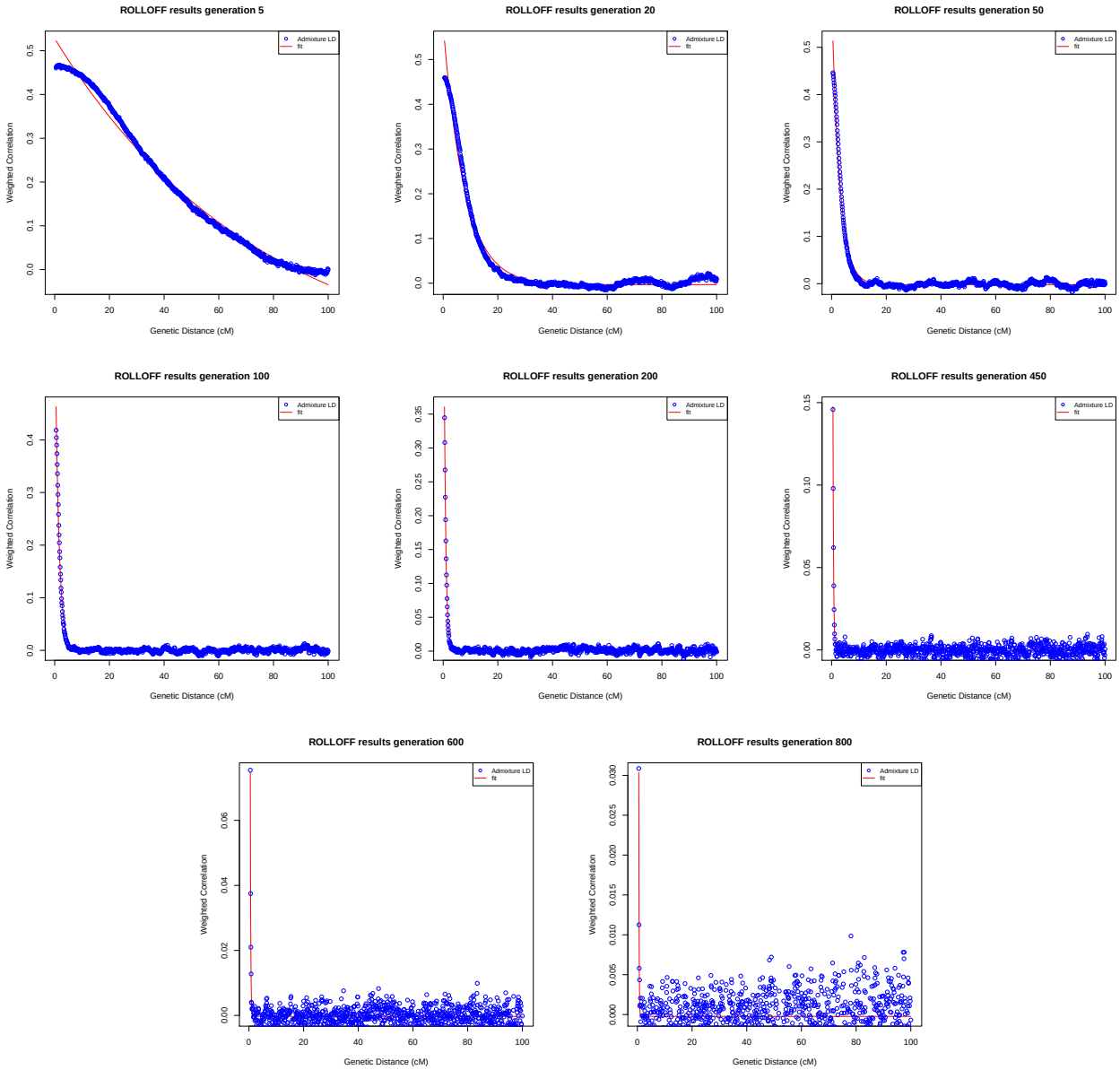

Figure 3.1: 2-way dating admixture event from the simulation of the 140 admixed individuals, based on CEU and YRI as reference ancestral populations using ROLLOFF. Here the plots represent the weighted correlation LD as a function of the genetic distance for all generations. The fitting curve to the data show a pattern of an exponential decay distribution through which the date of admixture is generated. The x-axis represents the genetic distance and the y-axis is the weighted linkage disequilibrium between pair of SNPs. The Plots are based on the simulation of number of generations from 5 to 800.

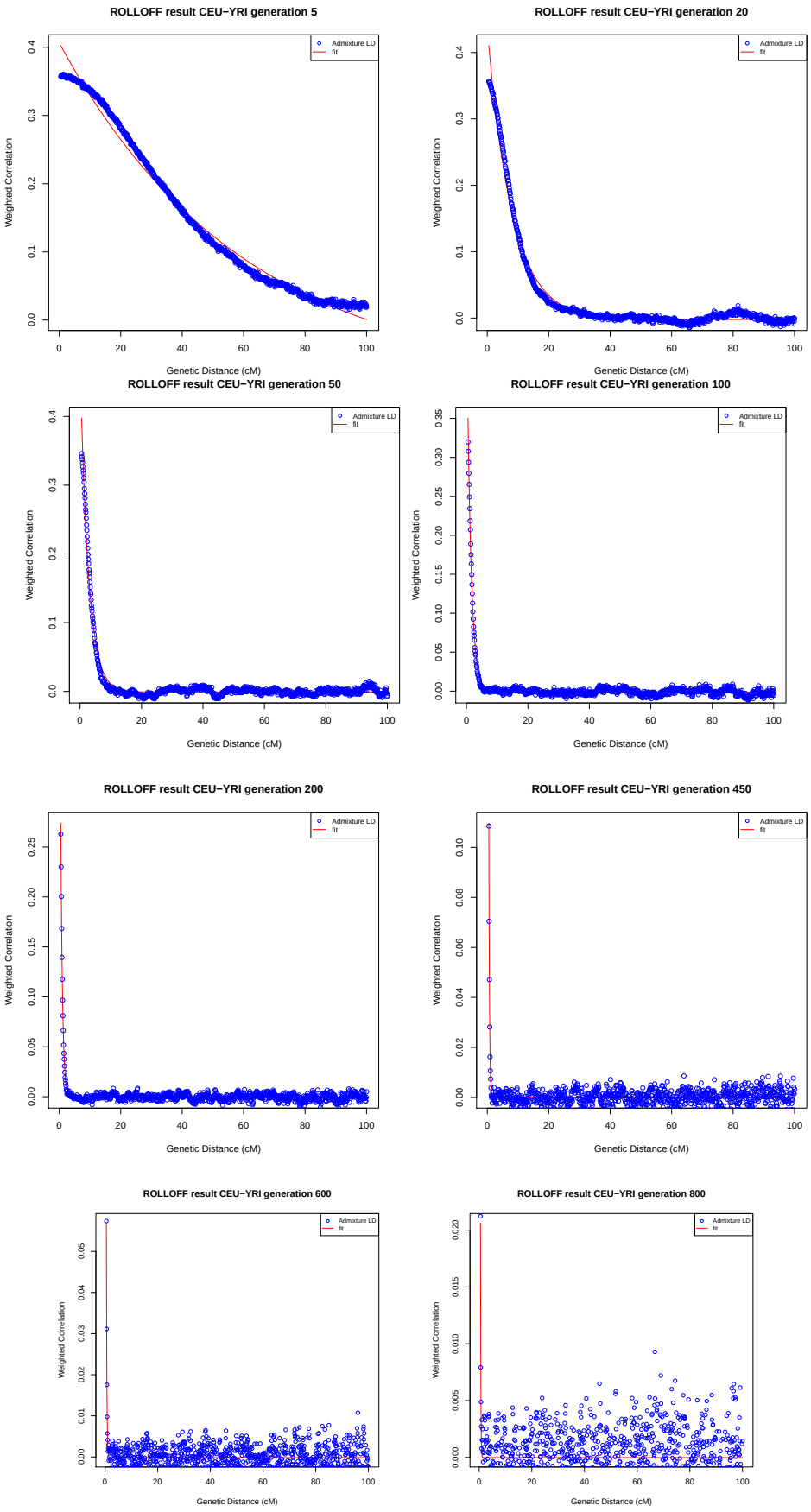

Figure 3.2: 3-way dating admixture event from the simulation of the 140 admixed individuals, based on CEU and YRI as reference ancestral populations using ROLLOFF. Here the plots represent the weighted correlation LD as a function of the genetic distance for all generations. The fitting curve to the data show a pattern of an exponential decay distribution through which the date of admixture is generated. The x-axis represents the genetic distance and the y-axis is the weighted linkage disequilibrium between pair of SNPs. The Plots are based on the simulation of number of generations from 5 to 800.

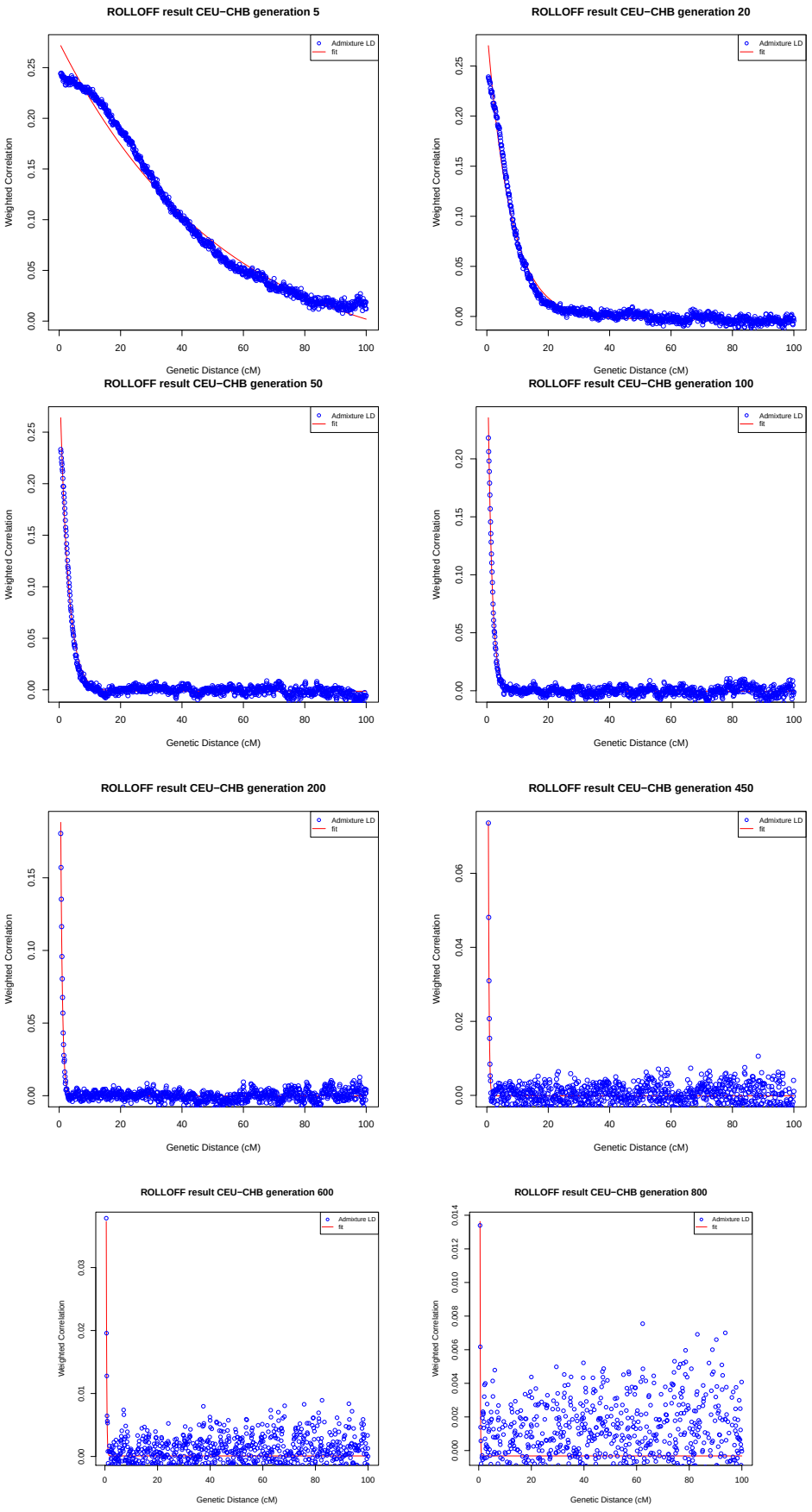

Figure 3.3: 3-way dating admixture event from the simulation of the 140 admixed individuals, based on CEU and CHB as reference ancestral populations using ROLLOFF. Here the plots represent the weighted correlation LD as a function of the genetic distance for all generations. The fitting curve to the data show a pattern of an exponential decay distribution through which the date of admixture is generated. The x-axis represents the genetic distance and the y-axis is the weighted linkage disequilibrium between pair of SNPs. The Plots are based on the simulation of number of generations from 5 to 800.

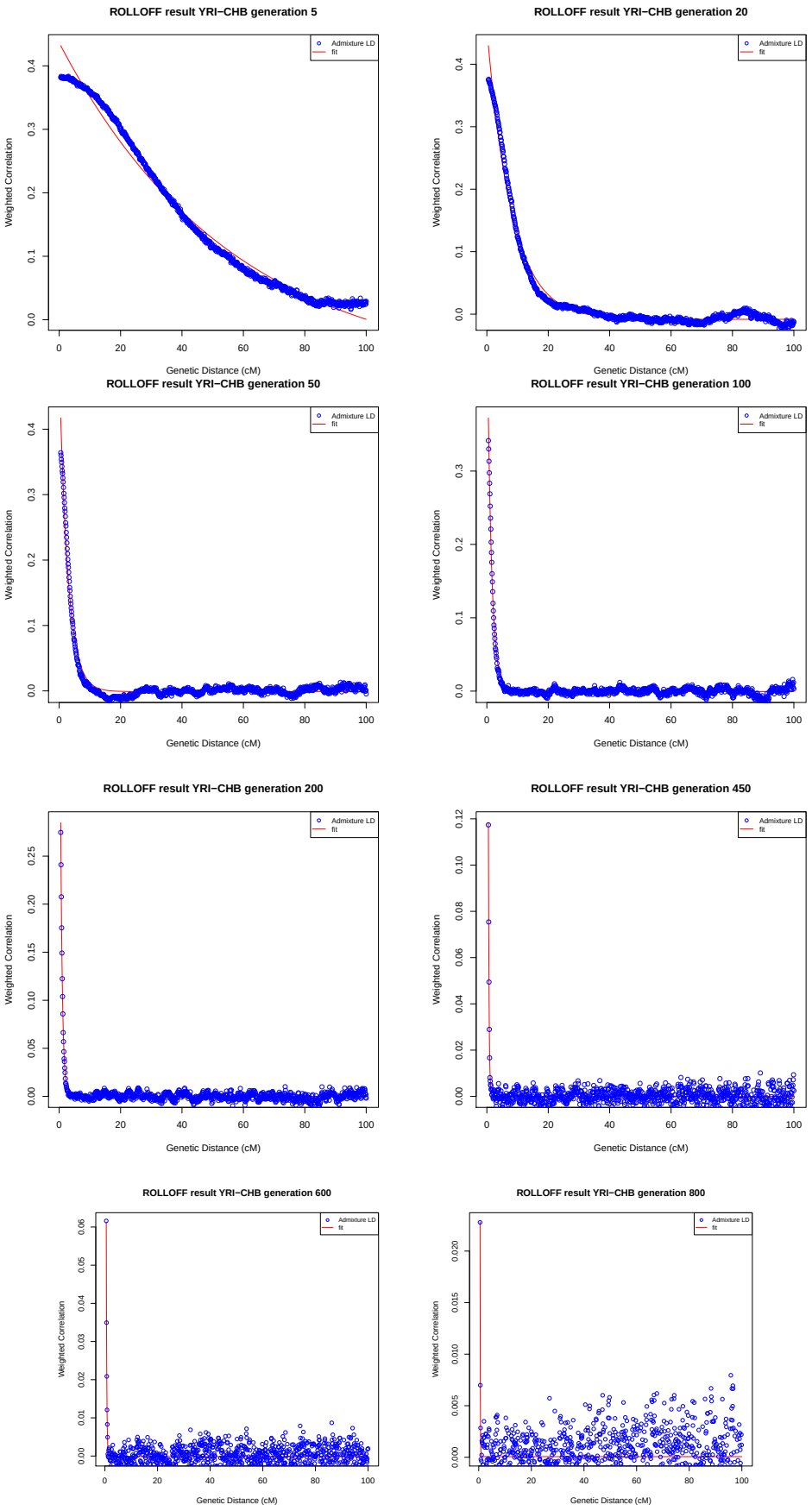

Figure 3.4: 3-way dating admixture event from the simulation of the 140 admixed individuals, based on YRI and CHB as reference ancestral populations using ROLLOFF.Here the plots represent the weighted correlation LD as a function of the genetic distance for all generations. The fitting curve to the data show a pattern of an exponential decay distribution through which the date of admixture is generated. The x-axis represents the genetic distance and the y-axis is the weighted linkage disequilibrium between pair of SNPs. The plots are based on the simulation of the number of generations from 5 to 800.

3.3.2 Assessing Dating Admixture Using ALDER

With regard to ALDER, we applied the same protocol as in the previous section. ALDER (Loh et al., 2013) computed two-reference weighted LD curves using pairs of references, one from each group between CEU and YRI for the 2-way admixed scenario and the other one for all combined pair-wise ancestral populations (CEU and YRI, CEU and CHB and YRI and CHB) for the 3-way admixture scenario. ALDER also fit an exponential decay curve to each LD curve, starting from 0.5 centiMorgans. The results of 2-way are shown in Table (3.4),the results for 3-way single point are presented in Table (3.5). We observe, in contrast to ROLLOFF, that the date inferred by ALDER tended to be closer to the true date in both scenarios. Despite this, ALDER failed to estimate the date of admixture for generation more than 450 between YRI and CHB and more greater than 600 for the rest of the pair-wise comparison. In 2-way admixture, the expected date overestimates the true date starting from N = 450 generations onwards, while for 3-way admixture, the pattern varies depending on the pairwise comparison, even at older generations. Moreover, ALDER failed to estimate the date of admixture in the 3-way scenario for generations 600 and 800. The range of the 95% confidence interval is very narrow for the estimates for shorter durations of admixture and wider as the number of generations increases; but still we observe a large increase in the range especially for older generations (200 and beyond). These results suggests how precise and reliable the expected date of admixture is for recent generations. However, for older generations, ALDER was unable to provide estimates for generations beyond 450 in some cases and we expect the 95% confidence interval to be wider due the pattern of LD within the admixed population for older generations. In addition, the amplitude of the fitting curves for each admixture scenario are similar for each generation More precisely, the amplitudes of the curves in the 3-way admixture scenario differ depending on the pairwise combination.

Table 3.4: 2-way ALDER Simulations Results between CEU and YRI. The results are in number of generation, taken one generation is 35 years.

2-way ALDER Results CEU and YRI.			
True Generation	Estimated Date ± 95% CI	p-value*	Amplitude 95% CI
5	4.1 ± 0.1	3.4e-317	0.00123442 ± 0.00001958
20	19.1 ± 0.4	0	0.00125058 ± 0.00002082
50	48.9 ± 0.6	0	0.00124592 ± 0.00002275
100	98.5 ± 1.2	0	0.00124626 ± 0.00002147
200	196 ± 3.3	0	0.00123781 ± 0.00002696
450	454.4 ± 21	9e-19	0.00132645 ± 0.0001499
600	NA	NA	NA
800	NA	NA	NA

Table 3.5: 3-way ALDER Simulations results between pairwises CEU and YRI, CEU and CHB and YRI and CHB. The results are in number of generation (and its $\pm$ 95% CI), taken one generation is 35 years.

True Date	Estimated Date	Amplitude	p-value*
ALDER Results pairwise CEU and YRI			
5	4 $\pm$ 0.1	0.00090409 $\pm$ 0.00001453	9.6e-217
20	18.5 $\pm$ 0.3	0.00091517 $\pm$ 0.00001726	0
50	51.1 $\pm$ 0.9	0.00094270 $\pm$ 0.00001722	0
100	100.7 $\pm$ 1.4	0.00092757 $\pm$ 0.00001541	0
200	205 $\pm$ 3.24	0.00096398 $\pm$ 0.00002220	0
450	466.9 $\pm$ 25.6	0.00104206 $\pm$ 0.00013037	1.3e-15
600	NA	NA	NA
800	NA	NA	NA
ALDER Results pairwise CEU and CHB			
5	4.2 $\pm$ 0.1	0.00046996 $\pm$ 0.00001136	0
20	19.6 $\pm$ 0.5	0.00047163 $\pm$ 0.00001237	4e-318
50	52.06 $\pm$ 0.9	0.00048693 $\pm$ 0.00001144	0
100	100.2 $\pm$ 1.25	0.00047885 $\pm$ 0.00001245	9.9e-324
200	208.8 $\pm$ 3.2	0.00051884 $\pm$ 0.0000191	1.9e-162
450	445.50 $\pm$ 30.67	0.00048948 $\pm$ 0.00007563	9.7e-11
600	NA	NA	NA
800	NA	NA	NA
ALDER Results pairwise YRI and CHB			
5	4.1 $\pm$ 0.1	0.00104693 $\pm$ 0.00001368	0
20	19.1 $\pm$ 0.3	0.00105166 $\pm$ 0.00001517	0
50	49.5 $\pm$ 0.7	0.00105471 $\pm$ 0.00001596	0
100	98.6 $\pm$ 1.7	0.00105608 $\pm$ 0.00001751	0
200	202.3 $\pm$ 4.5	0.00106654 $\pm$ 0.00003311	1.3e-227
600	NA	NA	NA
800	NA	NA	NA

* p-value taking value equal 0 means the exponent is less than (-1000)

3.3.3 Assessing Dating Admixture Using stepPCO

In order to date the admixture events using stepPCO, we applied the following pipeline:

(1) We split the merged genotype data into 22 chromosomes and converted each split genotype from each chromosome into stepPCO format.

(2) We generated the genetic map file from each chromosome by interpolation using the genome-wide recombination rate, estimated as part of the HapMap Project B_{37}.

(3) For each chromosome, we ran stepPCO (Pugach et al., 2011) which generates the wavelet transform center of each individual in the real data.

(4) We deduced the admixture rate that was later implemented in a forward simulation for the purpose to bring about the wavelet transform frequencies at each time point.

(5) In the forward simulation, we chose the recombination rate to be 2.78 per chromosome per generation for all chromosomes which is the recombination rate observed for human chromosome 1 (Pugach et al., 2011).

(6) We ran 100 simulations for each of the migration parameters and for each simulation, we sampled 100 chromosomes at exponential growing time points.

(7) Depending on the number of generations N (Table (3.2)), we varied the growth rate and the effective population size interval to obtain the wavelet transform frequencies at each generation time point.

(8) To reduce the noise of the wavelet in the simulated data, we normalized the distribution of the wavelet by the length of the chromosome (in the simulated data, we used the length of the chromosome in the HaPMap Project B_{37}) by subtracting the log of the chromosome length, which would correspond to the threshold.

(9) As the wavelet transform is deduced for each chromosome, we applied the average frequency wavelets over all the 22 chromosomes in the real and the simulated data.

(10) The date of admixture is deduced by matching the dominant frequency in the observed data to its correspondent in the simulated data.

Table 3.6: 2-way stepPCO Simulations Results between CEU and YRI. The results are in number of generation (and its $\pm$ 95% CI), taken one generation is 35 years.

2-way stepPCO Results CEU and YRI.			
True Date	Dominant Frequency	Estimated Date	95% CI
5	1.95	2	[0,13]
20	2.83	10	[3,45]
50	3.42	29	[11,268]
100	3.88	56	[24,218]
200	4.129	90	[41,808]
450	4.65	203	[95,1231]
600	4.87	345	[92,1667]
800	5.06	609	[177,2943]

Table 3.7: 3-way stepPCO Simulations Results of the pairwises CEU and YRI, CEU and CHB, and YRI and CHB. The results are in number of generation (and its $\pm$ 95% CI), taken one generation is 35 years.

3-way stepPCO Results pairwise CEU and YRI			
True Date	Dominant Frequency	Estimated Date	95% CI
5	2	2	[0, 13]
20	2.78	9	[3,38]
50	3.39	27	[12,164]
100	4	70	[33, 285]
200	4.34	90	[47, 381]
450	4.77	312	[123, 1914]
3-way stepPCO Results pairwise CEU and CHB			
5	1.99	2	[0,16]
20	2.85	11	[4,42]
50	3.5	32	[13,268]
100	4.1	89	[24,1981]
200	4.38	110	[47,431]
450	4.79	364	[73,1824]
3-way stepPCO Results pairwise YRI and CHB			
5	1.92	2	[0, 11]
20	3	13	[5, 70]
50	3.6	39	[16, 190]
100	3.9	56	[24, 285]
200	4.42	110	[41, 361]
450	4.75	364	[84, 1914]

The simulation results showed that in both admixture scenarios the expected number of generations underestimates the true number of generation as the number of generations increases (see Table (3.6) and Table (3.7)). We also found that the dominant frequency of the wavelet increases as generations increase, which would correspond to an abundance of low frequency wavelets (that is wider ancestry blocks) for recent generations to high frequency wavelets indicating narrow ancestry blocks for older generations (Pugach et al., 2011). Moreover, none of the expected results from the simulation overestimated the true generation. We note that the 95% confidence interval is very wide, which suggests a great level of variance in the data. This behaviour could be due to our large sample size because we have simulated 140 individuals, while a sample size of only 10 individuals is needed in order to get accurate estimate with narrow confidence interval (Pugach et al., 2011). Besides this, computational process also influence the value of the estimate.

3.3.4 Assessing Dating Admixture Using GLOBETROTTER

In order to generate the date of admixture using GLOBETROTTER, we applied the following pipeline:

(1) All genotype data were split into chromosomes and phased together with SHAPEITv2 (Delaneau et al., 2012). Haplotype "painting" with Chromopainter v2 (Lawson et al., 2012) was done on the high density SNP dataset, defining each cluster of populations as target or donor/surrogate.

(2) Mutational rates and Ne (sample size) parameters were first estimated with an Estimation Maximization (EM) algorithm by running Chromopainter v2 on all 22 autosomes for the entire dataset with 10 iterations (Lawson et al., 2012).

(3) The weighted average of these parameters, according to the SNP coverage of each chromosomes and the number of individuals, were then used to compute the chromosomal painting.

(4) Each cluster of populations was successively identified as a target and the others as surrogates. The painted chromosomes obtained for each cluster were used in GLOBETROTTER v1.0 (Hellenthal et al., 2014) to estimate the proportion and the dates of the potential admixture events characterizing them.

(5) Coancestry curves were estimated with and without standardization with a NULL individual, and consistency between each of the estimated parameters was checked.

(6) 100 bootstrap re-sampling was done to estimate the p-value of the admixture events and the 95% confidence interval for the obtained dates. The 'best-guess' scenario given by GLOBETROTTER v1.0 (Hellenthal et al., 2014) was considered for the target population.

Fit.quality.1event and fit.quality.2events correspond respectively to the fit of a single admixture event and the fit of the first two principal components capturing the admixture events. MaxScore.2events corresponds to the additional R^2 explained by adding a second date versus assuming only a single date of admixture ($M \geq 0.35$ to infer multiple dates events). Assuming single admixture event, if ancestries E and F associate with the same admixed population, for example, whenever E = F the fitted curve will have negative slope, as seen for the pairwise CEU versus CEU plot, YRI versus YRI plot and CHB versus CHB plot. If a positive slope is seen, as for the CEU versus YRI, CEU versus CHB and YRI versus CHB plots, this implies these populations contribute to the admixed population.

Table 3.8: Accuracy of the different admixture dating methods.

ERROR	ALDER	ROLLOFF	stepPCO	GLOBETROTTER
Root-Mean-Square-Error(RMSE)	2.6	19.2	112.2	158
Bias	0.655	16.2	72.5	93.8

Table 3.9: Accuracy of the different admixture dating methods based on the generations ≤ 100

ERROR	ALDER	ROLLOFF	stepPCO	GLOBETROTTER
Root-Mean-Square-Error(RMSE)	1.1	15.4	24.9	26.1
Bias	1.1	13	19.5	19.1

The results shows that the expected date underestimated the true date in all the admixture scenarios. Moreover, we noticed that for generations greater than 100, the expected date underestimated the true date more significantly and the range of the 95% confidence interval grows wider as the estimated generation increases. We plotted the graph (Figure (3.8)) which assessed the admixture dating methods for 2-way admixture scenario. We realized that all the methods provide estimates for all the generations N except for ALDER which failed to compute the estimates taking two ancestral populations as references especially for older generations (N $=$ 600 and 800). Also, the estimates provided by GLOBETROTTER are questionable due to the big difference between the true and the expected generation for older generations; this is also due to the fact that as GLOBETROTTER is haplotype-based the method performs poorly with less than 300,000 SNPs (Hellenthal et al., 2014). Nevertheless, the estimates determined by ALDER and GLOBETROTTER for recent generations provides powerful confidence intervals close to the date. This shows how precise the estimates given by these two methods are. In contrast, stepPCO was able to consistently provide good estimates for older generations in a shorter time, but with large variance compared to GLOBETROTTER. We compute the Root-mean-square error (RMSE) (Chai and Draxler, 2014; Walther and Moore, 2005) between the true and estimated values and bias (difference between sums of true and estimated values divided by the number of observations) up to 450 generations to compare the accuracy of all the method-based inferences. This comparative table (Table (3.8)) shows that the ALDER method recovers true dates with less bias than the other methods under default settings.

Table 3.10: Result of 2-way single-point simulation of 140 admixed samples based on GLOBETROTTER using both CEU and YRI as reference ancestral populations. The results are in number of generation (and its $\pm$ 95% CI), taken one generation is 35 years.

		Result of 2-way based on GLOBETROTTER using both CEU and YRI.			
True date	Estimated date	maxR2fit.1date	Estimated date event 2	maxScore.2events	95% CI for First Event
5	2.4	0.999	-	-	[2.2, 2.6]
20	13.25	0.999	-	-	[12.7, 13.8]
50	30.5	0.9986	-	-	[29.5, 31.7]
100	52.2	0.997	-	-	[50, 54]
200	77	0.9895	-	-	[72.5, 80.3]
450	86.8	0.959	193.1	0.073	[75.9, 89.6]
600	74.6	0.917	78.6	0.001	[62, 78]
800	83.1	0.879	92.8	0.04	[63, 84]

Table 3.11: 3-way single-point simulation of 140 admixed samples based on GLOBETROTTER using CEU, YRI and CHB as reference ancestral populations. The results are in number of generation (and its $\pm$ 95% CI), taken one generation is 35 years. GLOBETROTTER estimated the number of generation since admixture happened using all the three reference populations in contrast to previous methods that used pairs-wise in case of multi-admixed samples.

3-way GLOBETROTTER results between CEU and YRI and CHB.							
True Date (in Gen.)	Estimated Date event 1(in Gen.)	maxR2fit.1date	Estimated date event 2	maxScore.2events	fit.quality.1event	fit.quality.2events	95% CI for Event 1
5	2.7	0.999	2.9	-	0.87	1	[2.4, 2.8]
20	13.3	0.999	14.7	0.26	0.859	1	[12.9, 13.9]
50	30.5	0.999	35.0	-	0.835	1	[29.8, 31.2]
100	52.6	0.998	53.5	-	0.80	1	[51, 54]
200	81	0.992	84	0.00076	0.736	1	[79, 83]
450	96.7	0.96	133.5	0.06	0.77	1	[87, 99]
600	88.2	0.94	582.5	0.024	0.75	1	[76.7, 86.7]
800	81.7	0.88	87.76	0.03	0.74	1	[72, 82]

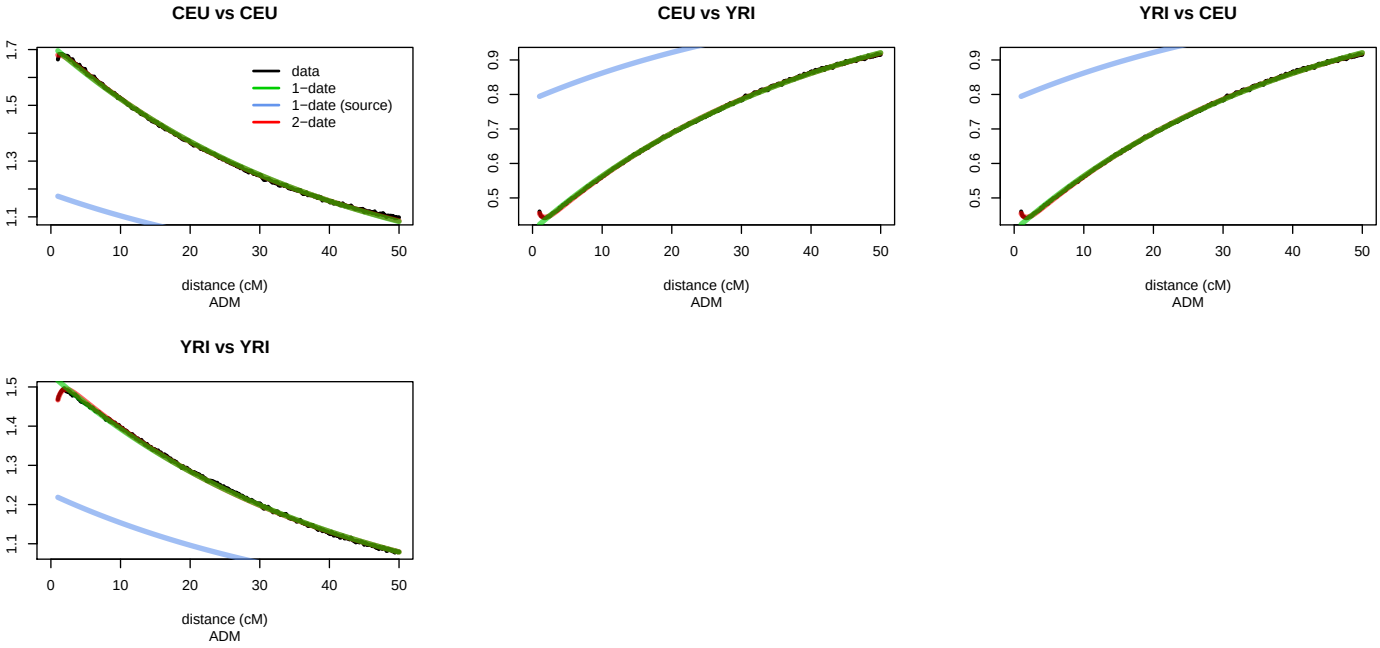

Figure 3.5: 2-way results for dating admixture event with GLOBETROTTER for generation 5. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chunks either from CEU, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from CEU and YRI donors(Y-axis), at varying genetic distances(x-axis). The pattern is the same for all the generations in the 2-way approach. The figures CEU vs YRI or YRI vs CEU shows a negative slope while the remaining shows a positive slope.

3.3.5 Assessing Dating Admixture Using MALDER

Motivation

The discovery of multi-way admixed populations has raised the problem of how to ascertain the admixture dating process in a sequential manner. ALDER, by using pairwise admixture events failed to estimate the in-between admixture time where the admixture process become sequential with more than 2 ancestral populations. Pickrell (2014) has developed a method(MALDER) which is an extension of ALDER which infers multiple admixture events. In this section, we propose an ALDER-based method which make use of ALDER by inferring multiple admixture events in such a way that the date of admixture between 2 populations accounts for the effect of the admixture of other ancestral populations.

Conceptual Framework

Let A, B, and C be three ancestral populations and D the admixed population of A, B and C. We use ALDER to compute the Weighted linkage disequilibrium for all possible pairwise populations. Given r_{AB}, r_{AC}, and r_{BC} the weighted linkage disequilibrium computed from ALDER output of all pair of SNPs between populations A and B, A and C, and B and C respectively in the admixed population D, we compute the Multiple Weighted Correlation Coefficient defined below:

$$R_{AC,AB-BC} = \sqrt{\frac{r_{AC}^2 + r_{AB}^2 - 2r_{AC}r_{AB}r_{BC}}{1 - r_{BC}^2}} \tag{3.3.1}$$

The above statistic assume that the pattern of linkage disequilibrium of all pairs of SNPs for all possible pairwise ancestry A-B and B-C in the admixed population D has an effect on the pattern of LD of the admixture between A and C in the admixed population D. After computing the multiple Weighted correlation Coefficient, we fitted the output and we plotted it as a function of the genetic distance using an exponential decay. This generates the date of admixture between A and C, accounting for the effect of the admixtures A-B and B-C, or to a sum of two exponentials in order to infer two admixture dates which exhibit the date between A and C and an admixture event which has occurred either between A and B or between B and C:

$$R_{AC,AB-BC} = \begin{cases} a_0 + a_1 \exp(-g_1 \frac{dist}{100}) & \text{assuming one admixture event} \\ a_0 + a_1 \exp(-g_1 \frac{dist}{100}) + a_2 \exp(-g_2 \frac{dist}{100}) & \text{assuming two admixture events} \end{cases} \tag{3.3.2}$$

g_1 and g_2 are the number of generations and $dist$ the genetic distance in centiMorgan, a_0 is the affine term and a_1 and a_2 the amplitude of the multiple weighted LD curve.

Data Description and Simulation Framework

We simulated 140 individuals in a 3-way multi-point admixture process with ancestry coming from Europeans (CEU 165 samples), Yoruba (YRI 203 samples), and Chinese (CHB 137

samples). The admixed populations were built based on the pattern developed similarly at the previous section. We divided the data into two parts, the first part were used to simulate admixture events, and the remainder were utilized to assess the admixture dating methods. In our case we assume two admixture processes at generations N_0 and $N_1 = \frac{N_0}{2}$ taking $N_0 = 5$, 10, 20, 50, 70, 100, 200, 300, 450(Table (3.12)). We specified the proportions of contribution from the ancestral populations and we generated the admixed population using our simulator mentioned in the previous section. We merged the admixed genotype data with the ancestral populations and we applied our proposed methods for testing. We assessed the proposed model using MALDER to compare the dates of admixture events.

We also simulated 3-way multipoint admixture whereby we assumed two admixture process at generations N_0 and $N_1 = \frac{N_0}{2}$ taking $N_0 = 5$, 10, 20, 50, 70, 100, 200, 300, 450(Table (3.12)). We specified the proportions of contribution from the ancestral populations and we generated the admixed population using our simulator described in the previous section. We merged the admixed genotype data with the ancestral populations and we applied our proposed methods for testing. We assessed the ALDER-based method using MALDER to compare the dates of admixture events.

Table 3.12: 3-way Multi-point Admixture Scenario (N_0 and N_1 are the number of generations.)

3-way Admixture	CEU	YRI	CHB
140	admixed	CEU	YRI	CHB
1	0	0.3	0.0	0.7
N_1	0.6	0.0	0.4	0.0
N_0	1	0.0	0.0	0.0

Table 3.13: Simulation Results for 3-way multi-point admixture.

Estimate results simulation between CEU, YRI and CHB									
Date assuming One event		Dates assuming two events							
Estimated		1st Event Admixture				2nd Event Admixture			
Estimate	p-value	True	MALDER	Estimate	p-value Estimate	True	MALDER	Estimate	p-value Estimate
3±0.02	$\leq 2e-16$	2	3±0.5	3±0.2	$\leq 2e-16$	5	5±1	11±3.7	$4.8e-09$
8±0.02	$\leq 2e-16$	5	8±0.3	3±6.7	0.3	10	9±0.2	7±3.86	0.000854
18±0.06	$\leq 2e-16$	10	17±1	3±2.5	0.04	20	30±11	18±0.1	$\leq 2e-16$
48±0.17	$\leq 2e-16$	25	46±1	11±8	0.01	50	220±145	47±0.4	$\leq 2e-16$
70±0.3	$\leq 2e-16$	35	69±0.96	23±5.8	$2.77e-11$	70	30400±50861	67±1	$\leq 2e-16$
100± 0.6	$\leq 2e-16$	50	99±1	65±40	0.0016	100	137±214	90±15	$\leq 2e-16$
202.6±2	$\leq 2e-16$	100	1.52±222	17.6±19	0.07	200	198±96	200±2.8	$\leq 2e-16$
316±4.76	$\leq 2e-16$	150	14±12	65±65	0.05	300	308±6	306±10	$\leq 2e-16$
494±29	$\leq 2e-16$	225	90±49	6.5±6	0.04	450	510±26	473±28	$\leq 2e-16$

The results showed that the new method is able to estimate, in the context of previous admixture events distinct admixture dates and to generate the fitting curve associated with the pattern of LD in the admixed population(Table (3.13)). By assuming one admixture event, the estimated date of admixture from our approach tends to be closer to the true date N_0. We observed that the expected date of admixture underestimated the true date with a small range; The model showed accuracy especially for true dates between 70 and 100 generations though

the expected date is an overestimate of the true one for generations greater than 200. In addition, assuming two admixture events, the results showed that the second event admixture from the new method underestimates the true date for generations varying between 10 and 200. It overestimates it when the true number of generations is equal to 5, 300 or 450 compared to the first event, where the estimated date is an underestimate of the true date for all generations except for generation N_0 equal to 100. The results also showed that the estimated dates are all significant for second admixture events, but less significant for the first admixture event. The only estimate where non-significance is observed is when the value of the true generation for the first event is 5. The results from MALDER overestimate the true generation overall for $N_0 \leq$ 100 except for younger generation where N_0 is equal to 5 or 10. For N_0 greater than 100, the range between N_1 and the MALDER's estimate is very wide with larger 95% confidence interval. MALDER produce a poor estimate for the second admixture event for true generation equal to 70; This shows that MALDER can produce unreliable estimate. The results of the estimate from the ALDER-based method that we propose are overall, closer to the true generation compared to the results from MALDER except in older generations. Our results also shows that the fitting curve of the multi-weighted LD is consistent with the data for all the generations taken for the simulation either by assuming one admixture event of two admixture events(Figure (3.7)). As the number of generation increases, the fitting curve to the data tends to take the form of L-shape. We computed the Root-mean-square error (RMSE) between the true and estimated values and bias (sum of absolute difference between true and estimated values divided by the number of observations) up to 450 generations to compare the accuracy of all the method-based inferences. This comparative table (Table (3.14)) shows that our approach to estimate distinct admixture events recovers true dates with less bias than MALDER assuming two admixture events under default settings. Moreover, we compared the 2nd event admixture date estimate between MALDER and the ALDER-based approach and the plot shows that the ALDER-based approach shows closer values to the true date compared to MALDER (Figure (3.6)).

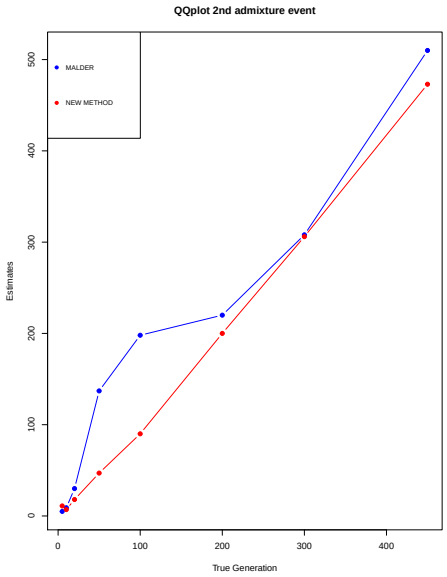

Figure 3.6: 2nd event assessment graph

Table 3.14: Accuracy of the different admixture dating methods. The Table displays the Root-Mean-Square-Error(RMSE) generation less or greater than N_0, where N_0 is the number of generation

PERFORMANCE	1st admixture event		2nd admixture event	
DATING METHOD	MALDER	ALDER-based	MALDER	ALDER-based
RMSE	74.9	83.2	7.2*	1.05*
Bias	53.8	48.5	32*	5.9*
RMSE generation less than $N_0 = 70$	8.02	8.1	21.2	0.9
RMSE generation greater than $N_0 = 70$	7.5	7.2	8,8	3.2
Bias generation less than $N_0 = 70$	8	6	45.25	3.5
Bias generation greater than $N_0 = 70$	104.6	100.2	26.75	9.75

* In the calculation of these statistics, i excluded $N_0 = 70$
due to the inconsistency of the second admixture event estimate

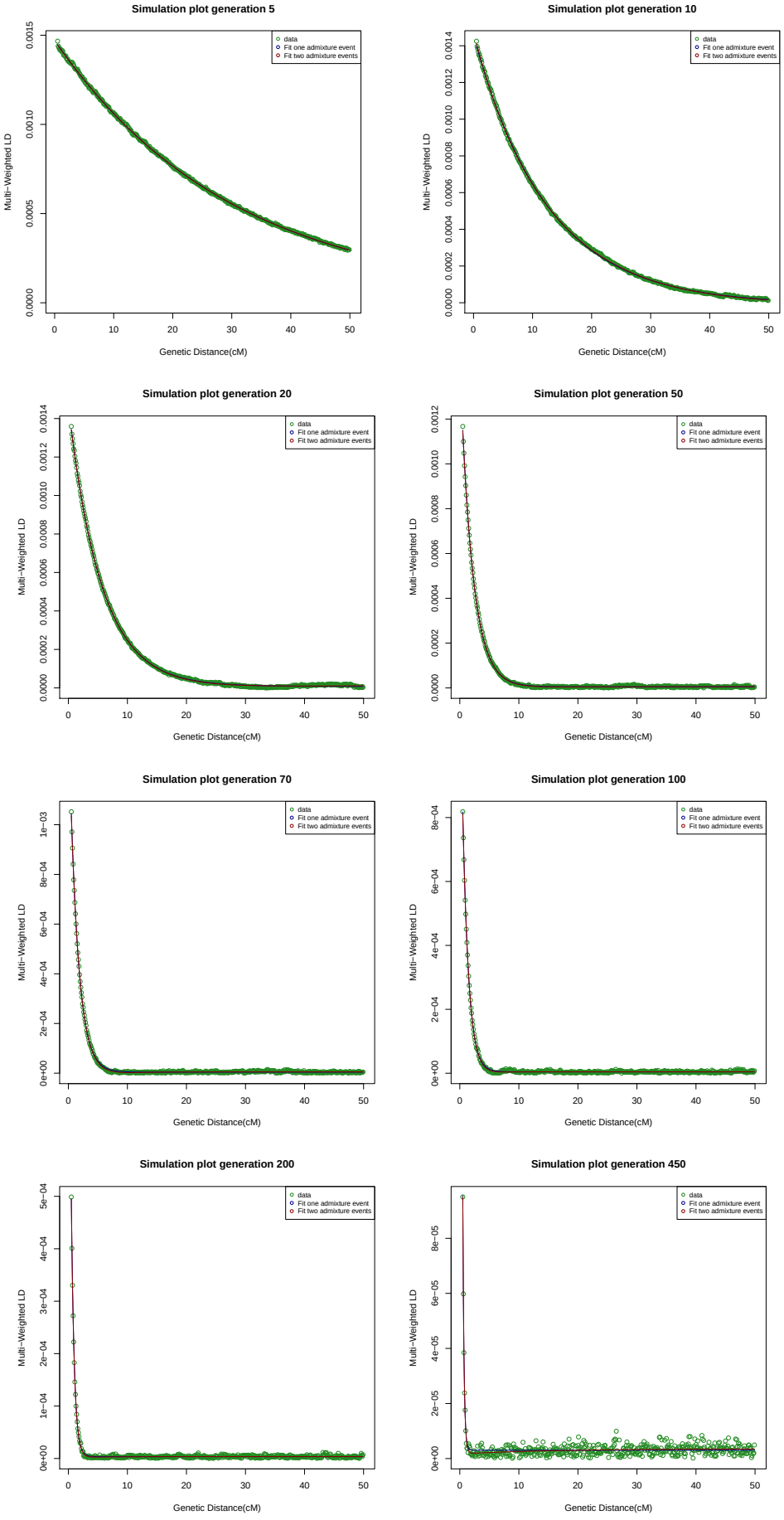

Figure 3.7: 3-way Dating Admixture between CEU and YRI and CHB results using the new method. We plot the different fitting curves of the multi-weighted correlation coefficient to the genetic distance for generation 5, 10, 20, 50, 70, 100, 200 and 450. The figures shows that the fitting curves are consistent with the data for all generations either assuming one event of two events of admixture.

3.4 Summary

We have assessed various admixture dating methods which include ROLLOFF, stepPCO, ALDER and GLOBETROTTER for 2-way admixture approach and we showed that ALDER is the best method which infer the date of admixture for 2-way approach. Also, we have developed a method using ALDER as a baseline to infer distinct admixture events accounting for the effect of other admixtures in a 3-way multi-point scenario. We have assessed this method by performing various simulations using data from the HapMap Project phase 3 (Consortium, 2007) and we have compared the results using MALDER .

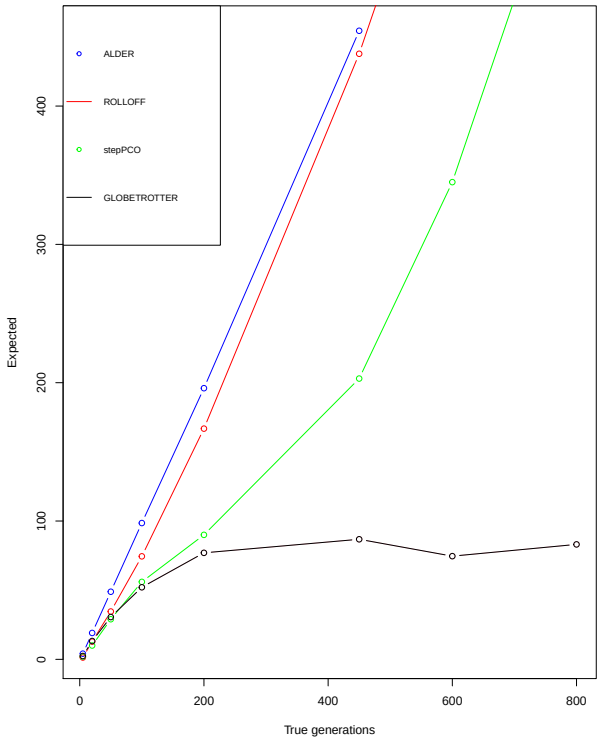

Figure 3.8: 2-way Assessment Graph.

We showed that in the simulation results that the aproach to estimate distinct admixture events perform well for inferring second admixture event. In the next chapter, we will apply ALDER and stepPCO in real data which include African Americans, Mexican Americans and Luhyan in order to estimate the date of admixture. We will also apply our method and compare the results of our method with the outcome of MALDER.

4. Application of Admixture Dating Methods to Real Data

4.1 Data Description

We applied admixture dating methods on real genotypes data which include African American, Mexicans Americans and Luhyan who are admixed in order to estimate the date of admixture. We will be using ALDER and stepPCO in order to estimate pairwise admixture events and we will also apply the proposed ALDER-based approach (chapter 3, section (3.3.5)) and compare it with the results of MALDER. We considered genotype data from 1000 Genomes Project phase 3 (Consortium, 2015) including Americans of African Ancestry in South West USA (ASW, n = 54); Mexican Ancestry from Los Angeles USA(MXL, n = 63);Individuals of Northern and Western Europeans Ancestry living in Utah(CEU, n = 99); Yoruban individuals from Ibadan, Nigeria (YRI, n = 108); Han Chinese individuals from Beijing, China (CHB, n = 103); Luhyan individuals from Webuye, Kenya (LWK, n = 97). We also included Native American samples (NAT) selected from Mao and colleagues (Mao et al., 2007) which comprised 616,568 SNPs and 43 individuals from a combined group of populations including Nahua (n=10), Maya (n=6), Quechua (n=2), and Aymara (n=25). We conducted quality control keeping bi-allelic SNPs in all autosomal chromosomes. We pruned these using a window of 1500 variants and a shift of 150 variants between windows, with an r^2 cut-off of 0.2 and we exclude SNPs with high-LD and non-autosomal regions from the pruned dataset. We merged the above populations with the 1000 Genomes and with the Native American populations, and removed ambiguous SNPs whose strand orientation could not be determined (that is G/C and A/T SNPs). The final dataset comprised 567 individuals and 2,282,325 SNPs for the analysis.

4.2 Principal Component Analysis

We carried Principal Component Analysis (PCA) using the "smartpca" program of EIGENSOFT(Patterson et al., 2006) and we generated graphical overviews. The PCA results indicated a signal of admixture in ASW, MXL and LWK (Figure (4.1) and 4.2). The first component (PC1) separates LWK, YRI and one part of ASW from the other groups, while the second component (PC2) shows CEU, and CHB at both ends with MXL and NAT in the middle. The third component (PC3) shows ASW and MXL being distributed toward the middle of the NAT cline more or less in the middle between CEU and CHB.The African Americans are distributed toward the middle of the YRI; while the Mexican Americans are on the native American cline which suggests a recent gene flow of the native Americans and the Yoruban into the Mexican American population and African American population respectively, taking CEU as one of the source populations. NAT and YRI are the farthest populations and LWK looks to be very closer to YRI than any other populations. This is confirmed by the estimate of the genetic distance between them (Table (4.1)). LWK and YRI are in the same cline with CEU which

presupposes the contribution of CEU and YRI to the Luhyan population. Finally, the PCA results suggest that ASW and MXL are 3-way admixed taking CEU, YRI and NAT as ancestral populations. Moreover, the PCA plot suggests that LWK is also admixed with a more gene flow of YRI with CEU and CHB as ancestral populations.

Table 4.1: Fst statistics calculated between each pair of population.

	CEU	YRI	CHB	LWK	MXL	ASW	NAT
CEU	0.00						
YRI	0.124	0.00					
CHB	0.081	0.142	0.00				
LWK	0.115	0.005	0.135	0.00			
MXL	0.018	0.113	0.047	0.105	0.00		
ASW	0.063	0.005	0.096	0.006	0.056	0.00	
NAT	0.136	0.202	0.109	0.195	0.037	0.145	0.00

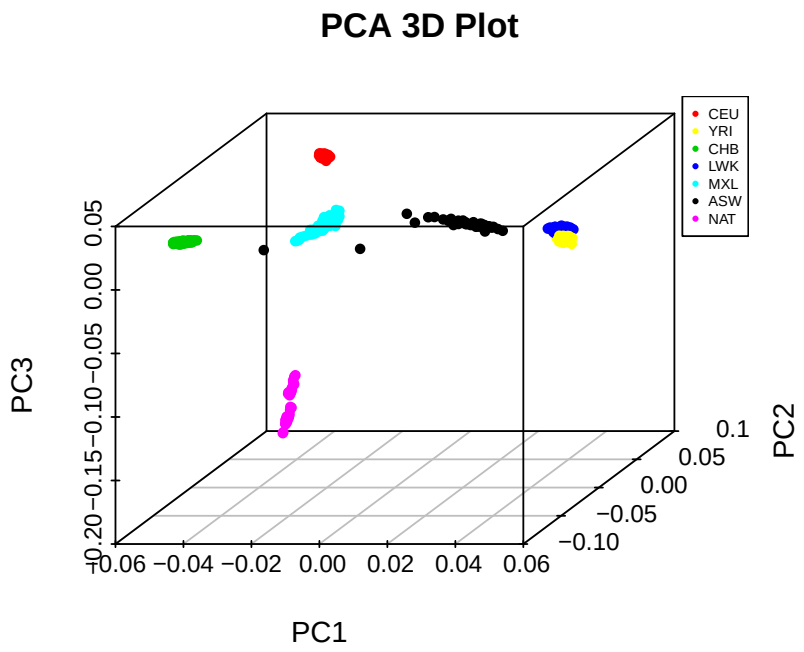

Figure 4.1: Three Dimensions PCA results. The plot shows the different cluster populations. We observe that the MXL cluster is locate at the middle of the triangular combination CEU-YRI-NAT; The ASW cluster is in the same cline between CEU and YRI but more directed to the YRI cluster.

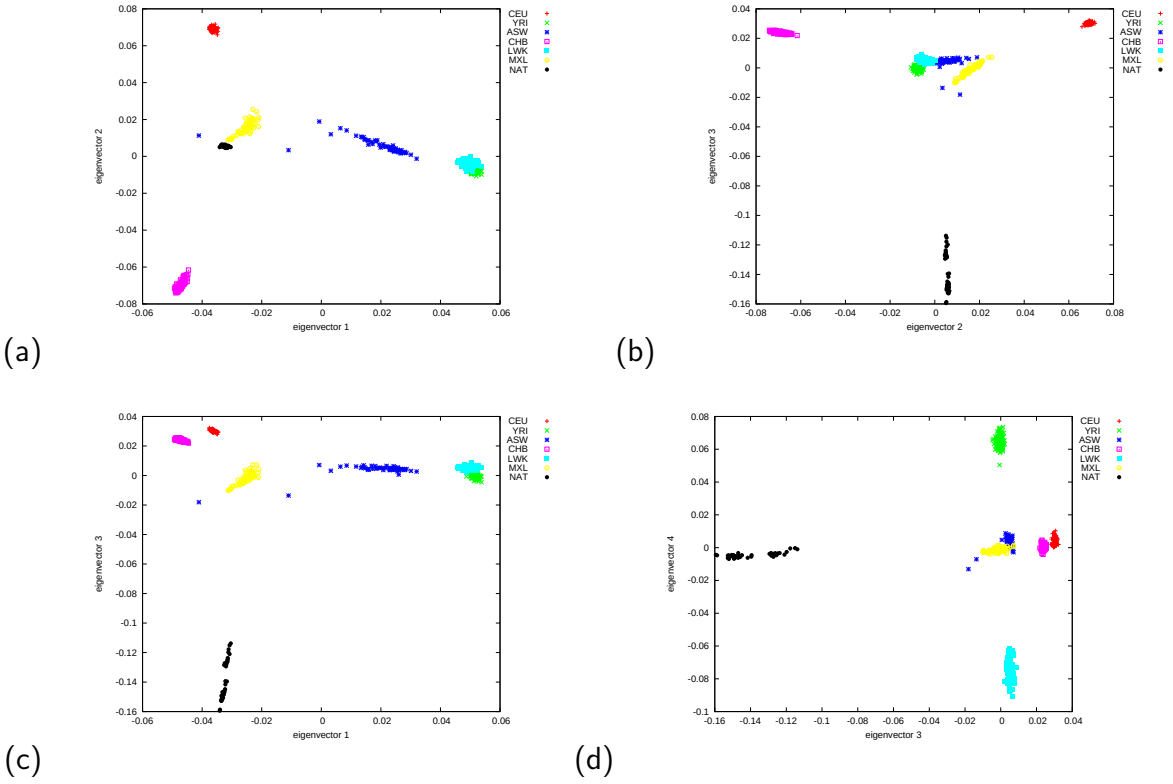

Figure 4.2: Figures (a), (b), (c), (d) represent PCA plot Results in two dimensions: Figure (a) shows that ASW and MXL are in the cline with YRI and NAT respectively, wich suggest that ASW and MXL are admixed taking CEU,YRI and NAT as parental populations. LWK is the closest population of YRI and is in the same cline with CEU and CHB which suggests the contribution of CHB, CEU and CHB in the LWK with more gene flow of YRI in LWK.

4.3 F_3 statistics and Admixture

Previous studies have identified African Americans (ASW), Mexicans Americans (MXL) and luhyan (LWK) populations as admixed populations (Murray et al., 2010; Salzano and Sans, 2014; Bryc et al., 2015; Mathias et al., 2016; Dobon et al., 2015). Here, we computed the F_3 statistics to detect admixture between two populations using the $qp3Pop$ program from the ADMIXTOOLS software (Patterson et al., 2012). We performed every possible combination of source populations for all pair-wise population taking ASW, MXL and LWK as target populations. A negative value of F_3(A; B, C) implies that population A (target population) comes from an admixture event between the two ancestral-like populations B and C (source populations). The only situation where this test will not detect admixture is when the target population suffered a high-degree of population specific drift after the admixture event (Dobon et al., 2015). For each F_3 estimate, we kept the results with a significantly negative value of the F_3 statistic. The F_3 statistics are provided in Table (4.2).

Based on the above information, we performed supervised analysis implemented in the

Table 4.2: F_3 statistics results

Source 1	Source 2	Target	F_3	std.error	Z	SNPs
CEU	YRI	ASW	-0.017069	0.000221	-77.147	28880
CEU	LWK	ASW	-0.012446	0.000227	-54.753	29140
CEU	NAT	MXL	-0.021667	0.000791	-27.407	2961
CEU	YRI	LWK	-0.002077	0.000221	-9.412	34038
YRI	MXL	ASW	-0.014939	0.000221	-67.460	23767
YRI	NAT	ASW	-0.015489	0.000455	-34.019	8832
YRI	CHB	LWK	-0.001669	0.000255	-6.538	31832
YRI	MXL	LWK	-0.001563	0.000233	-6.698	27340
YRI	NAT	MXL	-0.019546	0.001501	-13.026	2936
LWK	MXL	ASW	-0.011256	0.000236	-47.711	24200
LWK	NAT	ASW	-0.011745	0.000525	-22.364	8815
LWK	NAT	MXL	-0.021406	0.001494	-14.330	2994

ADMIXTURE software (Alexander et al., 2009; Alexander and Lange, 2011) with a reference panel which includes CEU, YRI, NAT and CHB, with target populations constituted of ASW and MXL for K = 3 and LWK for K = 3. Genome-wide estimates shows average ancestry proportions of African Americans as 75.9% of Yoruba, 21.4% of Europeans, 2.7% of Native Americans, which is concordant with previous studies about the little genetic contribution of Native Americans in the African Americans population (Bryc et al., 2015; Martin et al., 2016). Moreover, the results show that more than 50% of Native Americans have contributed to the Mexican Americans and the large contribution of the Yorubans to the Luhyan population confirmed the previous results in the PCA analyses (Figure (4.1)). This confirms the results of Bryc and colleagues who highlighted the same pattern among Mexican Americans located in the South Western region of the USA (Bryc et al., 2015).

Table 4.3: Genome-wide ancestry estimates in African Americans, Mexican Americans and Luhyan populations; Admixture with mean percentages ± standard deviations

Populations	CEU	YRI	NAT	CHB
ASW	21.4±14.4	75.9±17.7	2.7±8.7	nan
MXL	47.6±23.1	10.8±6.5	41.6±18.4	nan
LWK	3.0±2.4	95.3±2.2	nan	1.7±1.9

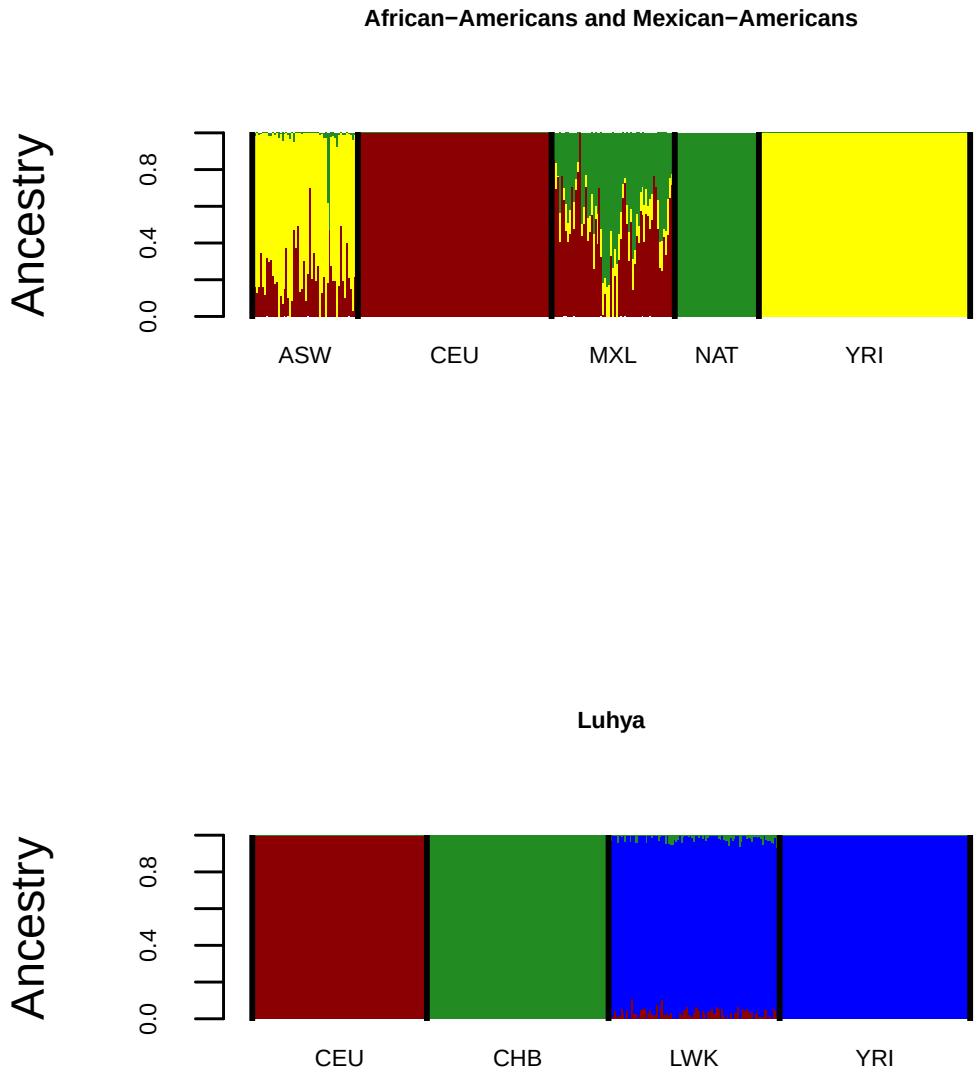

Figure 4.3: Supervised Admixture plots among the African Americans, the Mexican Americans and the Luhyan populations.

4.4 Application of Dating Methods to Real Data

Here, we assume 29 years per generation.

4.4.1 ALDER and stepPCO

To estimate the time of admixture events, we applied pairwise admixture using two methods: the linkage disequilibrium-based method ALDER with minimum genetic distance equal to 5 cM (Loh et al., 2013) and the wavelet transform method stepPCO (Pugach et al., 2011) and we compared the results of both methods for the real data. Knowing that both methods are based on two-way admixture, we performed a first analysis using ALDER, whereby we extracted the list of SNPs used for admixture dating using parental populations. With that list of SNPs, we performed a second run of admixture dating using both methods to estimate the date of admixture for all pair-wise (Yunusbayev et al., 2015). Interestingly, both results confirmed the initial Native Americans and Europeans admixture and subsequent African admixture among the African Americans. ALDER and stepPCO detected admixture events in African American population earlier between the Native and Europeans at around 15 to 16 generations ago meaning about 435 to 464 years ago; following by the mixture combination between Yoruba, and Europeans populations and Yoruba and Native American populations, respectively, around 6 generations ago to 7 generations ago. These findings are consistent with previous studies and genealogical records (Pugach et al., 2011; Jin et al., 2014; Bryc et al., 2015; Ni et al., 2016). Historical studies highlight the evidence for European/Native Americans miscegenation further back, before the 17th century (Sandefur and Trudy, 1986; Sollors, 2000), but the combination of wars and particularly diseases, drastically reduced the size of the Native American population. It was during that time that Europeans began the search for a labour force in West Africa to work in the sugar cane and cotton fields (Sandefur and Trudy, 1986; Sollors, 2000). The migration of West African Bantu to America during the slave trade occurred along side the miscegenation between Europeans and West Africans later on, around the eighteenth century, and also between Africans and Native Americans during that period. In the Mexican Americans, ALDER and stepPCO detected admixture events between Europeans and Yoruba between 18 and 20 generations or between 522 years to 580 years ago. This was followed by Europeans and Native Americans 8 generations ago but the mixture event between Yoruba and Native Americans is estimated to be 13 generations ago. This corroborates historical records which highlight the beginning of the slave trade from the European Spaniards in West Africa. The settlement of the Spaniards in America introduced the admixture between the Native Americans and the Africans (who where mostly from Bantu origins) and the migration of West Africans to America. ALDER and stepPCO estimated the admixture event in the Luhyan between Chinese and Yorubans at around 33 to 34 generations ago followed by the admixture events between Europeans and Yorubans 32 to 33 generations ago Table (4.4). These estimates dates are in agreement with historical records assuming 29 years per generations which date the interactions between these populations during the trade market in eastern Africa from 11th to the 14th century. The rising of the Sung Dynasty in China has played a key role in the trade market between Bantu East Africa and China as a whole (Beaujard, 2007), which has favoured the

admixture between these two populations (Beaujard, 2007). Meanwhile, the Europeans were already involved in the trade route via the Suez canal to Asia. This trade has developed a serious dynamic of migration and exchange behaviour Europe-Africa-Asia which gave rise to the existence of the admixture between them. By the end of the 14th century, the pattern of migration and trade declined due to a major climatic deterioration, which suggests that the gene flow in the Luhya was more commonly with the Bantu Africans populations within the African continent. This supports the results showing the greatest proportion of the Yoruba population in the admixed Luhyan (Beaujard, 2007).

Table 4.4: ALDER and stepPCO Dating results in African Americans, Mexican American and the Luhya.

pop1	pop2	admixed	ALDER estimate	stepPCO estimate	95% CI
CEU	YRI	ASW	7±0.7	4	[0;22]
CEU	NAT	ASW	15.75±3.37	15	[5;62]
YRI	NAT	ASW	6.47±1.33	6	[0;27]
CEU	YRI	MXL	17.6±1.73	20	[9;89]
CEU	NAT	MXL	8±3.17	8	[3;46]
YRI	NAT	MXL	13.31±5.40	13	[3;29]
CEU	YRI	LWK	32±8.7	33	[9;158]
YRI	CHB	LWK	33±10	34	[10;121]

4.4.2 Application of ALDER-based Method and Comparison with MALDER

The results of MALDER in the African Americans indicated an admixture event between Europeans (CEU) and Native Americans(NAT) and between Yoruba(YRI) and Native Americans(NAT) one generation ago following by and older admixture 15 generations ago (435 years ago). The latter estimate confirms the estimate that we found in the previous chapter related to the 2-way admixture in the African Americans population. Interestingly, MALDER's results estimated the date of admixture in the Mexican Americans between the pair-wises CEU-NAT, CEU-YRI, YRI-NAT at 2 generations ago followed by the older admixture event 23 generations ago (667 years ago).This confirms the results from previous studies on this population (Jin et al., 2014; Bryc et al., 2015). In the Luhya, MALDER estimated the admixture date between CEU-YRI and between YRI-CHB at 8 generations ago (232 years ago) following by an older date 55 generations ago(1595 years ago) with a large variance.

In the Luhya, the ALDER-based method, assuming one admixture event, inferred the date of admixture between CEU and CHB, accounting for the effect of other admixtures, as 89 generations ago (p-value $= 1.89e-8$). Based on the result of table (4.3), the $F3$-statistic between CEU and CHB, CEU and YRI, and YRI and CHB are negative, this implies a signal of admixture in the Luhya. But this signal might not necessary means that there exists a gene flow from western Europeans or Chinese into the Kenyan Luhya population or these populations are the true admixing populations (Patterson et al., 2012). These populations can be regarded as

merely proxies for the non-sub-Saharan ancestry present in the Luhya population. Many studies suggested, based on oral history that the Luhya tribes migrated from Egypt though historians generally believe the Luhya tribes migrated from West-Central Africa alongside other Bantu tribes (Hodgson et al., 2014; Joubert et al., 2010). Moreover, the study led by Gurdasani et al.(2015) suggested that "a large proportion of differentiation observed among African populations could be due to Eurasian admixture, rather than adaptation to selective force". Based on that, we investigated to use CEU and CHB as non-africans population as proxy for the Luhya population (Gurdasani et al., 2015). On the other hand, assuming two admixtures events, our model generated an older admixture event 87 generations ago (2523 years ago) with a non-significant recent admixture event 17 generations ago; Henceforth, we can conclude using our model that the admixture date between CEU and CHB, accounting for the effect of other admixtures, could have occurred 89 generations ago. We also notice that the admixture dating curve between CEU and CHB are exactly the same. Because we do not expect migration to have happened instantaneously, we hypothesize that a population equally genetically equidistant to both ancestors CEU and CHB contributed some ancestry to the Luhya populations. Many scholars highlighted the contribution of Cushitic or Nilotic-speakers in the formation of the Luhya population and the Cushitic and Nilotic-speakers represents a "back to Africa" migration from the Near Eastern region of Asia. There is no bioinformatics studies at this moment which confirm that Europeans and Chinese has contributed directly in the formation of the Luhya which is true because the Luhya population is the consequence of a recent migration in Kenya with a time still under debate (Hodgson et al., 2014). Nevertheless, there have been historical account of the presence of Chinese and Europeans Byzantine in contact with the kingdom of Akxum in Ethiopia. This interaction, accompanied with different migrations, could have probably begot the Kenya Luhya population (Al-Radi, 1990; Wolbert, 2002).

In the same way, the ALDER-based method also significantly inferred the admixture date between CEU and YRI, assuming one event as 97 generations ago (p-value $= 8e-7$). Assuming two events, the model was able to detect a non-insignificant date 5 generations ago and a significant event 68 generations ago (p-value $= 8e-5$). In addition to that, our model was able to estimate a significant date of admixture between YRI and CHB assuming one admixture event, 175 generations ago (p-value $= 0.05$). And assuming two events, the model detected a recent date not enough significant as 6 generations ago(p-value $= 0.35$) and an older date which is significant, 52 generations ago. There is no mention of interaction between the Chinese and the Bantu 175 generations ago, therefore it is likely that the interaction between Yoruba and Chinese may have occurred 52 generations ago. Although the formation about the Luhya is sparse in the literature, we can confirm this estimate based on the historical records highlighting the slave trade market and different migration which occurred in the first century between the Roman Empire, the Han Chinese kingdom and the African Bantu migration within Africa via the Great Lakes in the first century (Beaujard, 2007; Al-Radi, 1990).

In the Mexican American population, accounting for the effect of the other admixtures events, our model infers the date of admixture between Europeans and Native Americans, assuming one admixture event, 16 generations ago. Assuming two events, our suggested approach, detected one event occurring 12 generations ago (p-value $= 0.07$) and a non-significant event 100 generations ago. Moreover, the model detected an admixture event between CEU and YRI and between YRI and NAT assuming one admixture event respectively 31 generations ago (p-value

$= 0.000361$) and 27 generations ago (p-value $= 2.04e-5$), which is closer to the second admixture event generated by MALDER (33 generations). More interestingly, assuming two admixtures events, the model estimated an admixture event occurring between CEU and YRI and between YRI and NAT respectively 37 generations ago. One thing that we also noticed is that the fitting curve (Figure (4.5)) is inconsistent to the data and also the fact that the estimate does not corroborate any historical event of admixture with Europeans, Native Americans and Yoruba as ancestral populations.This led us to conclude that our approach failed to estimate the admixture events in the Mexican Americans. Therefore, the assumption of the effect weighted LD is not valid in the admixture process in the Mexican Americans.

Table 4.5: Result from dating admixture event in the Luhya, African Americans and Mexican American .

	LUHYA					
	Assuming one event		Assuming two events			
Scenario	Time 1	p-value	Time 1	p-value	Time 2	p-value
CEU-YRI	97±37.8	8e-7	5±11.7	0.5	68±33.7	8e-5
CEU-CHB	89 ±30	1.89e-8	17±904	0.9	87±66.7	0.001
YRI-CHB	175 ± 178.5	0.05	6±12	0.36	52±55	0.06
MALDER	n.a		8±18		55± 234	
	Africans Americans					
	Assuming one event		Assuming two events			
CEU-NAT	15±7	6.51e-05	17±15.5	0.03	99.7±802	0.8
CEU-YRI	9±2	3.4e-12	4±12	0.5	22±50	0.4
YRI-NAT	10±5	0.000171	11±8	0.008	110±886.5	0.8
MALDER	n.a		1±0.5		15±4	
	Mexicans Americans					
	Assuming one event		Assuming two events			
CEU-NAT	16±8	0.000105	12±12.7	0.06	100±344	0.6
CEU-YRI	31±17	0.000361	0.01±37.4	0.999	37±38	0.05
YRI-NAT	27±12	2e-5	31±19	0.00178	327±1683.6	0.7
MALDER	n.a		1±1		22± 12	

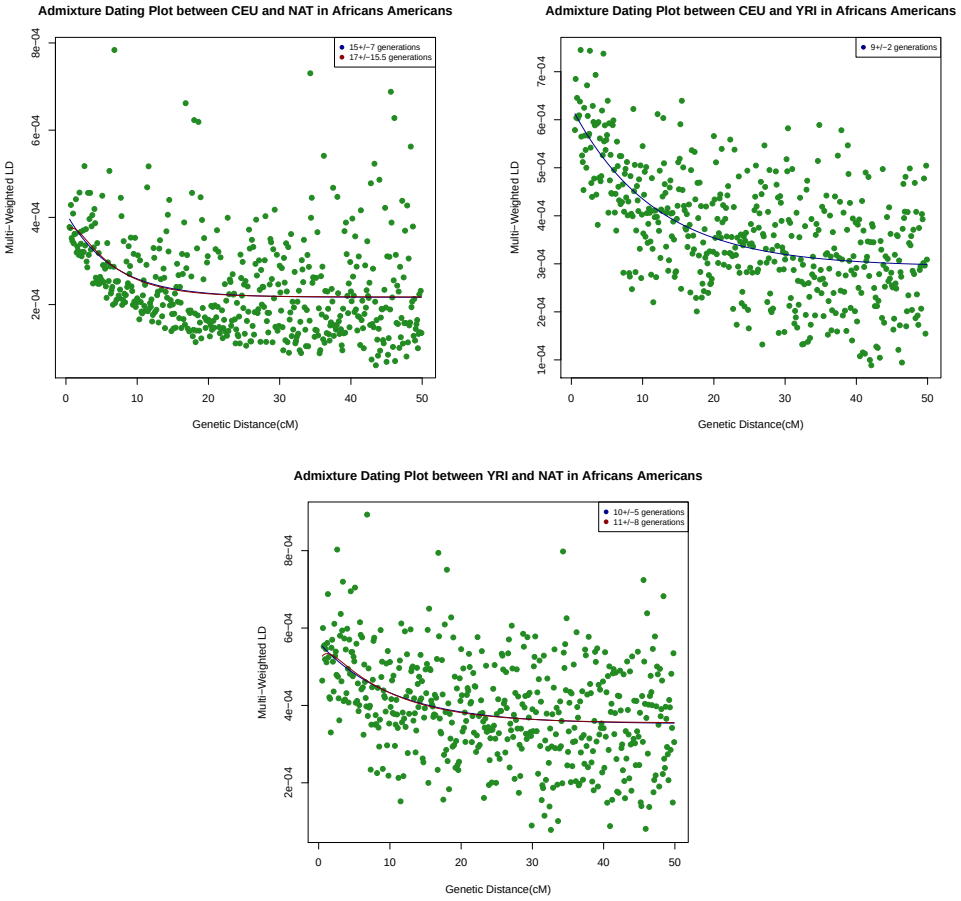

Figure 4.4: 3-way Dating admixture in Africans Americans using the ALDER-based Method. We computed weighted LD using ALDER for every pairwise population, then we computed the Multi-Weighted Correlation coefficient at each pair of SNPs accounting for the effect of the other Weighted LD and we estimate the date of admixture by fitting data with either an exponential or a sum of two exponentials. The fitting curve is consistent with the data for all the pairwises CEU-YRI, CEY-NAT and YRI-NAT.

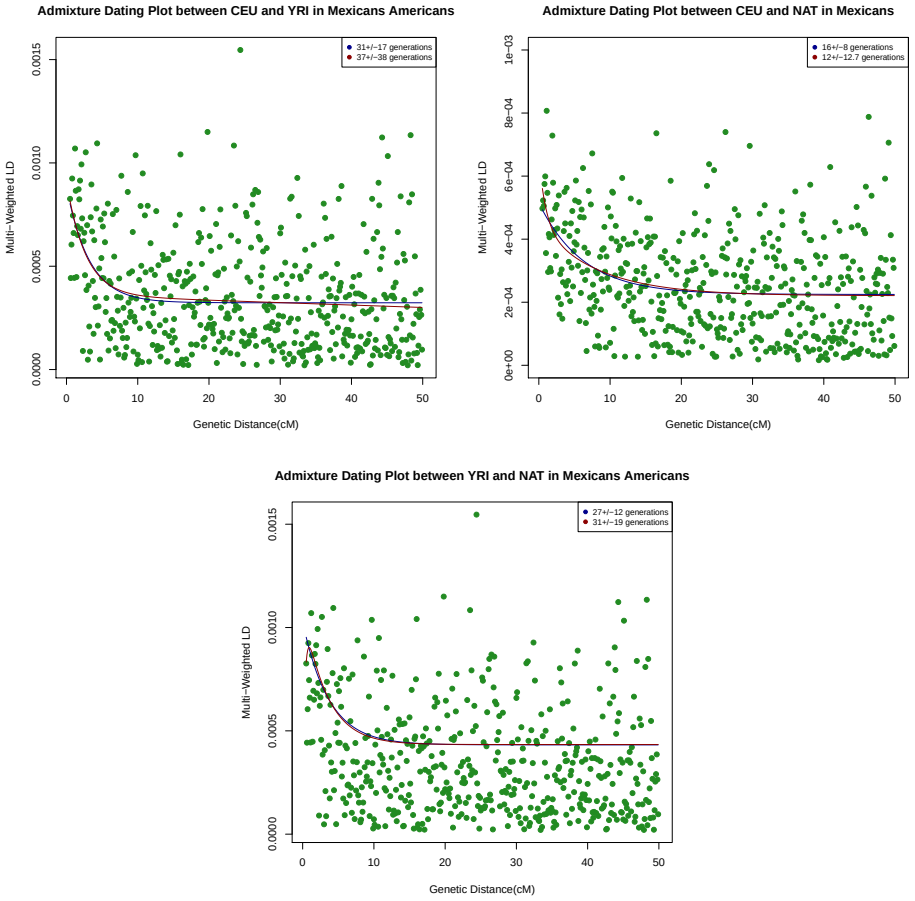

Figure 4.5: Dating admixture in Mexican Americans using the ALDER-based Method. We compute weighted LD using ALDER for every pairwise population, then we compute the Multi- Weighted Correlation coefficient of at each pair of SNPs accounting for the effect of the other Weighted LD and to estimate the date, we fit either an exponential or a sum of two exponential in order to infer either one admixture events of two admixture events. The fitting curve is inconsistent with the data for all the pairwises CEU-NAT, CEU-YRI and YRI-NAT.

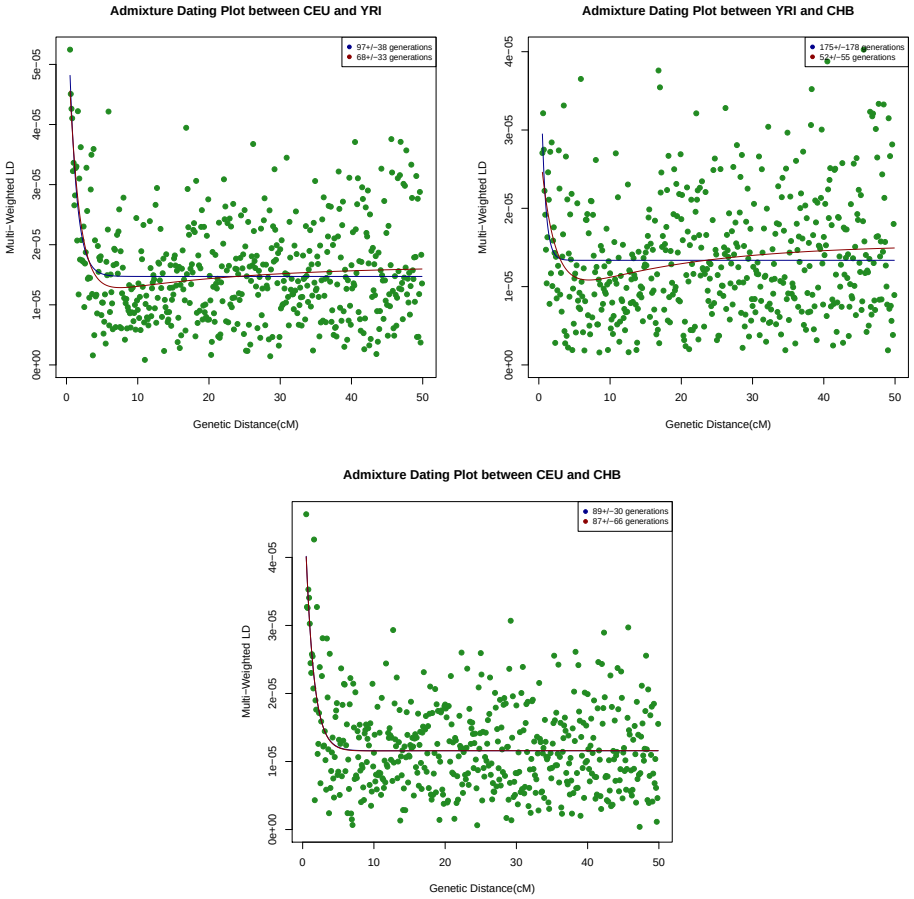

Figure 4.6: Dating admixture in the Luhyans using MALDER. We fit either an exponential or a sum of two exponentials with affine term in order to infer either one admixture event or two admixture events. We observe that the fitting curve is consistent with the data for all the pairwises CEU-YRI, CEU-CHB and YRI-CHB.

In the African Americans, the model estimated the date of admixture between Europeans and Native Americans and between Yorubans and Native Americans, respectively, at 15 generations (435 years ago) and 10 generations ago (290 years ago), assuming one admixture event. Assuming two admixture events, the model detected an admixture event respectively 17 generations ago (493 years ago) and 11 generations ago (319 years ago) which confirms the results that we found earlier in the previous section concerning a prior admixture between Europeans and Native Americans and the subsequent involvement of Africans in the formation of the African Americans. By estimating admixture date between every pairwise group in the African American population, we observed a decay of Linkage Disequilibrium with genetic distance for all the pairwise(Figure (4.4)). The curves fits the data and by fitting an exponential function, we were able to estimate the date of admixture. We noticed that the curves fits the data either for one admixture event of for two admixture events except for the pairwise CEU-YRI who was able to estimate one admixture event, the second assumption for two admixture events were giving non-significant dates (Table (4.5)). In addition, the model discovered and event between Europeans and the Yorubans 9 generations ago (261 years ago) but could not detect

any significant event of admixture assuming two events; which suggest that the main admixture process in the Africans Americans has been significantly characterized by the mixture and gene flow between Europeans and Yorubans accounting for the effect of other admixtures. This estimate also confirms previous studies related to the Africans Americans populations during the slave trade in the eighteen century (Jin et al., 2014; Bryc et al., 2015; Pugach et al., 2011).

Discussion and Conclusion

Different methods have been developed to estimate the date of admixture events and some were tested using available data either from the HapMap or from the 1000 Genomes Project. These data have become a reference panel or proxy panel for testing methods of admixture dating. Recently, Hodgson et al.(2014) attempted to study the back to Africa migration at the Horn of Africa. They compared admixture dates using a Linkage Disequilibrium approach (ROLLOFF and ALDER) and found that earlier episodes of admixture are largely masked by more recent admixture events. Additionally, Busby and colleagues (Busby et al., 2016a) studied the admixture into and within sub-Saharan Africa by comparing the LD-based method (ALDER/MALDER) and the Haplotye-based approach (GLOBETROTTER) and found that MALDER analyses display evidence for deep Eurasian and some hunter-gatherer ancestry across Africa, while GLOBETROTTER analysis provides clarity on the composition of the admixture sources, as well as the timing of events and their impact on different population groups. Moreover, most of the admixture dating methods struggle to accurately estimate the date of admixture events in more than 3-way admixture cases. In addition, many methods find it difficult to infer the date of admixture for older admixture events due to either the pattern of linkage disequilibrium or the distribution of ancestral block(or ancestral tracts). but one should also investigate the population structure of these panel data which could themselves be admixed, consequently making the reference panel data inaccurate.Moreover, most of the admixture dating methods struggle to accurately estimate the date of admixture events in more than 3-way admixture cases. In addition, many methods find it difficult to infer the date of admixture for older admixture events due to either the pattern of linkage disequilibrium or the distribution of ancestral block(or ancestral tracts). Here,we have assessed a variety of methods by performing simulations using less than 300,000 SNPs, therefore there is a need to evaluate the GLOBETROTTER method taking into account the number of SNPs. These methods rely on either accurate local ancestry information or global ancestry inference, and when the reference populations are highly divergent from the true mixing populations or when the ancestry tracts are short, in the case of ancient admixture, it becomes harder to produce accurate estimates.

In this dissertation, we systematically reviewed current approaches to date admixture events. We have detailed the benefit and limitation of each methods. We conducted an assessment of the methods using their implemented tool though the simulation of admixture events that mimicked real admixture scenarios and various times since the event occurred. Our results indicate that most current tools to date admixture have some limitation in capturing ancient admixture events and multi-way admixture scenarios. For example, GLOBETROTTER, which is currently the method widely used because of the variety of informations that can be detected, has been proven to have limitations in our simulation tests, firstly since it has been designed for recent admixture events and secondly, because it only performs well for number of SNPs greater that 300,000 (Hellenthal et al., 2014). Our results showed that, using less than 300,000 SNPs, GLOBETROTTER estimates become inconsistent using data from the HapMap Project. Therefore, one needs to account for different datasets and the total number of SNPs used for the evaluation. In addition, we identified three tools, StepPCO, MALDER and ALDER which

produce reasonable date estimates close to the simulated true dates. Our simulations test using the data from HapMap Project demonstrated that ALDER performs better for 2-way admixture scenario compared to other tools. However, ALDER can only produce estimates for generations less than 450. stepPCO was able to go beyond 450 generations but the range of the 95% confidence interval became quite wide as we increased the number of generations in our simulations. Additionally, we applied StepPCO, MALDER and ALDER to a real dataset of admixed populations from the 1000 Genomes project. Though MALDER shows improvement and produces reasonable date estimates compared to current methods, the results from both simulation and real data suggest that dating admixture events occurring more than 5000 years ago accounting for the effect of other admixtures remains a challenge. Our findings show that the improvement of ALDER for admixture dating accounting for the effect of other admixtures has helped us to identify distinct admixture events and trace the admixture process among the African Americans, Mexican Americans and the Luhya populations, which MALDER was unable to do.In this dissertation, we simulated, with a particular pipeline, admixed populations using ancestral panels which include CEU, YRI and CHB from the HapMap Project. Therefore, we can be confident that using other ancestral panels in our same dataset and applying the same pipeline for admixture dating assessment will produce the same result. However, one should apply this assessment to different datasets in order to increase our confidence in the use of various admixture dating tools.

While more appropriate and updated genome-wide panel data are required to enhance the accuracy of the results, the pinpointing of ancestry along the genome of a multi-way admixed individual is important to ensure correct ancestry breakpoints to enable the estimation of admixture events. The opportunity to develop and improve statistical models for dating multi-way admixture events for admixture dates greater than 5000 years is required in order to improve our understanding of human demography and movement.Improved methods of ancestry inference based on new and large datasets, particularly from African populations known to have high diversity and admixture will also increase our understanding of human movement and the implication in adaptation and consequently health. Therefore, we need to develop a more powerful tool for admixture mapping and admixture dating.

Further investigations into admixture dating should aim to improve accuracy with minimum variance of the date of admixture in a more complex scenario using reliable data for the description of the admixture patterns in any admixed population. Future work should be to identify the exact and the closest populations that have contributed to the mixture when it comes to testing the method for admixture dating and for futures GWAS studies. Moreover,technological advances have made the genome sequencing of the large amount of individuals possible and effective. This provides an opportunity to make inferences of demographic parameters based on ancestral block distribution and linkage disequilibrium patterns. There is a need to update the reference panels available for a better inference of ancestry and particularly for admixture dating and the opportunity to develop computational statistical model for multi-admixture event as well as to estimate the date of admixture for ancient admixture.

Appendix

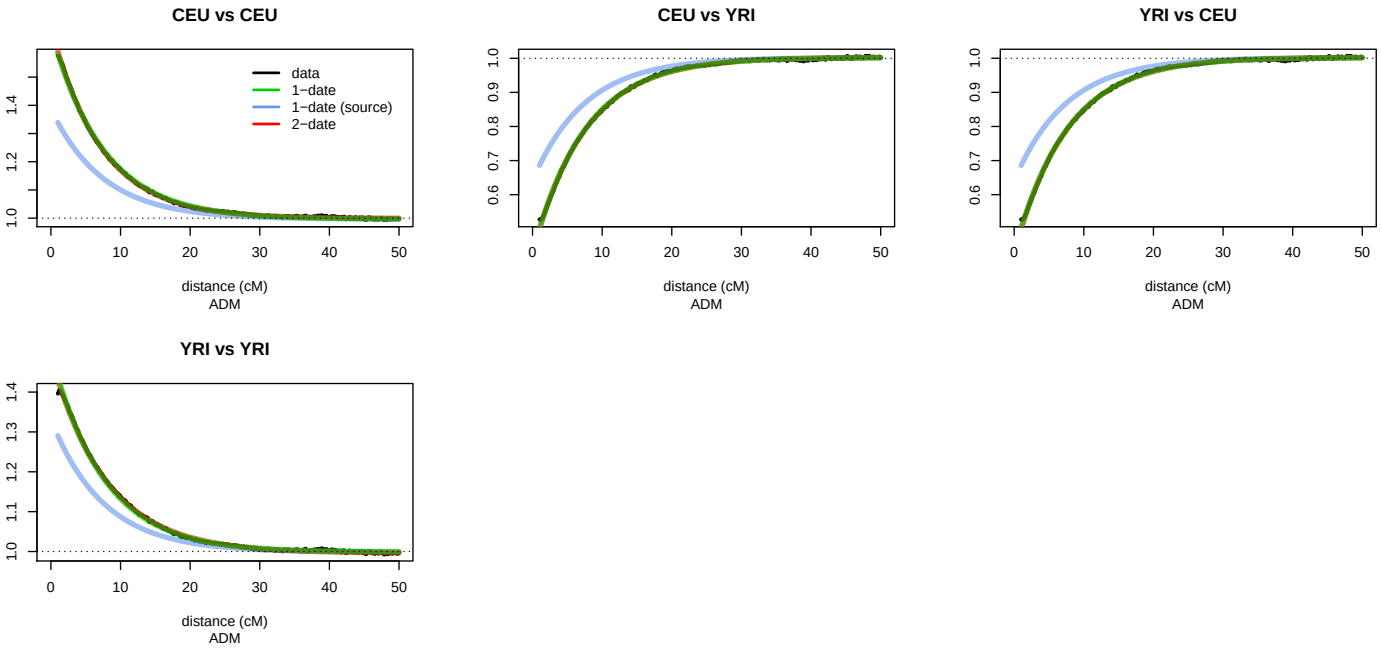

Figure 4.7: 2-way Dating Admixture results with GLOBETROTTER for generation 20. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chunks either from CEU, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from CEU and YRI donors(Y-axis), at varying genetic distances(x-axis). The pattern is the same for all the generations in the 2-way approach. The figures CEU vs YRI or YRI vs CEU shows a negative slope while the remaining shows a positive slope.

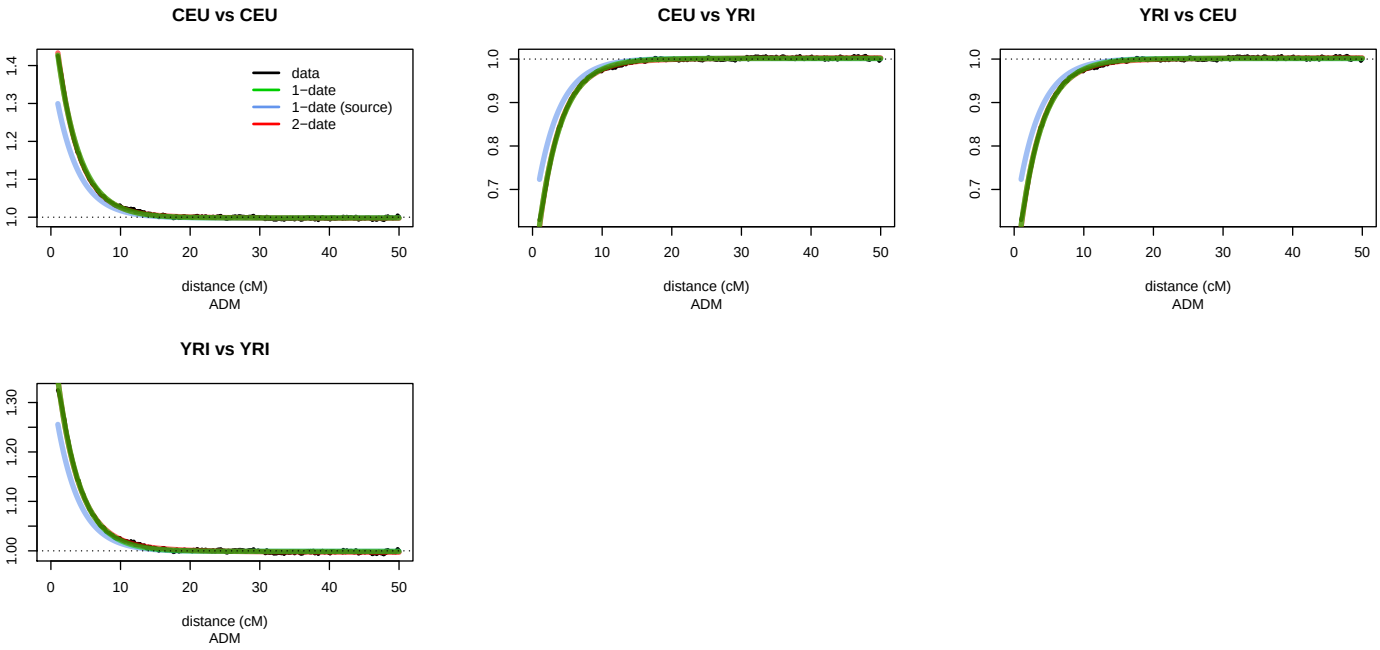

Figure 4.8: 2-way Dating Admixture results with GLOBETROTTER for generation 50. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chunks either from CEU, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from CEU and YRI donors(Y-axis), at varying genetic distances(x-axis). The pattern is the same for all the generations in the 2-way approach. The figures CEU vs YRI or YRI vs CEU shows a negative slope while the remaining shows a positive slope.

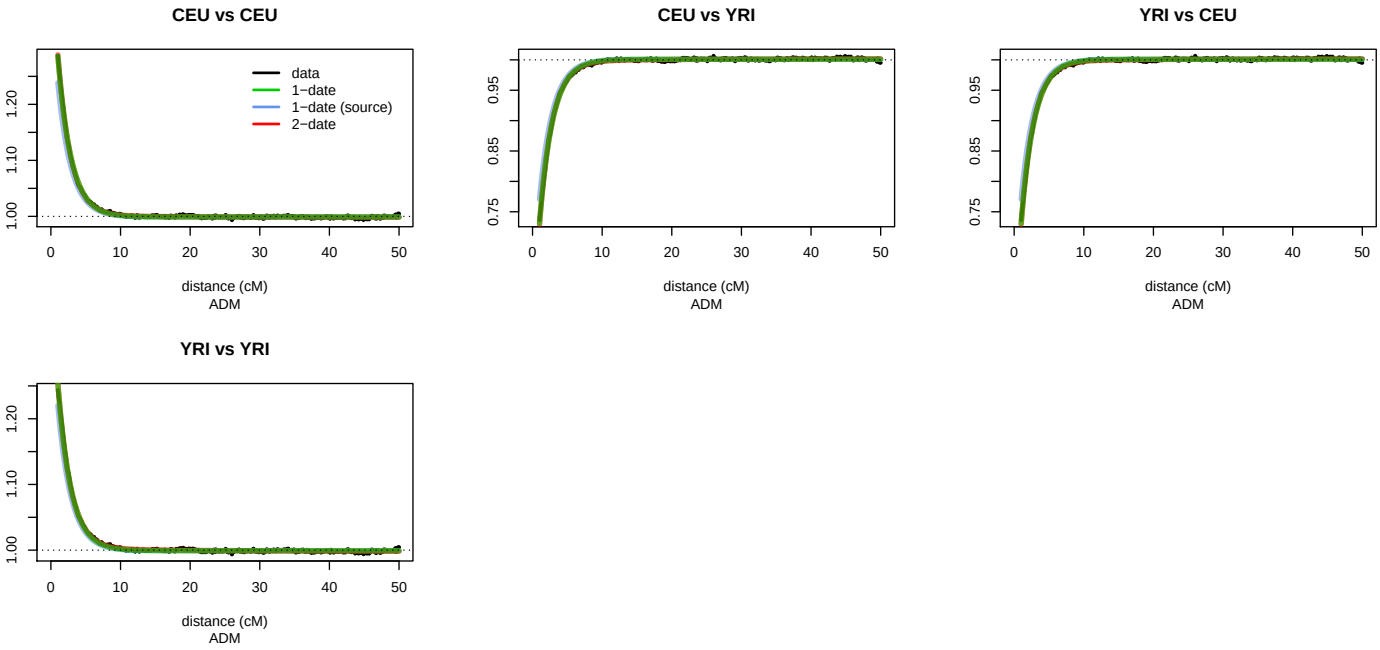

Figure 4.9: 2-way Dating Admixture results with GLOBETROTTER for generation 100. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chunks either from CEU, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from CEU and YRI donors(Y-axis), at varying genetic distances(x-axis). The pattern is the same for all the generations in the 2-way approach. The figures CEU vs YRI or YRI vs CEU shows a negative slope while the remaining shows a positive slope.

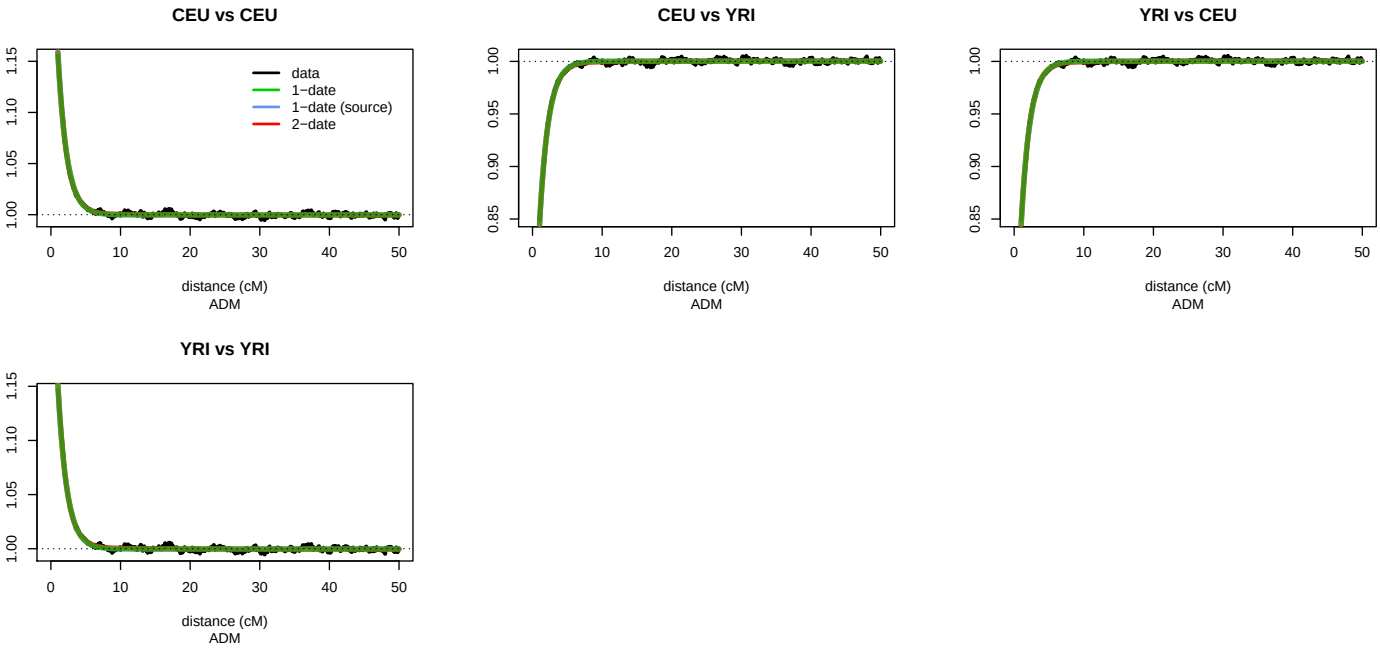

Figure 4.10: 2-way Dating Admixture results with GLOBETROTTER for generation 200. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chunks either from CEU, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from CEU and YRI donors(Y-axis), at varying genetic distances(x-axis). The pattern is the same for all the generations in the 2-way approach. The figures CEU vs YRI or YRI vs CEU shows a negative slope while the remaining shows a positive slope.

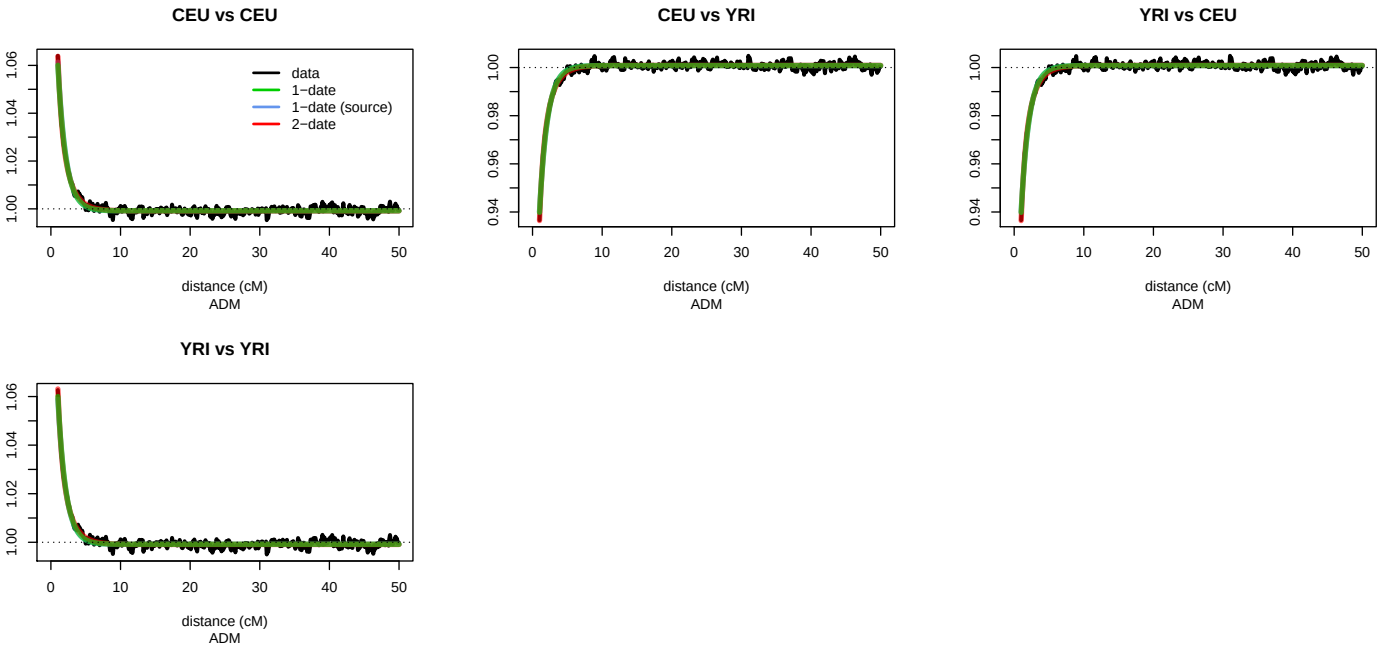

Figure 4.11: 2-way Dating Admixture results with GLOBETROTTER for generation 450. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chunks either from CEU, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from CEU and YRI donors(Y-axis), at varying genetic distances(x-axis). The pattern is the same for all the generations in the 2-way approach. The figures CEU vs YRI or YRI vs CEU shows a negative slope while the remaining shows a positive slope.

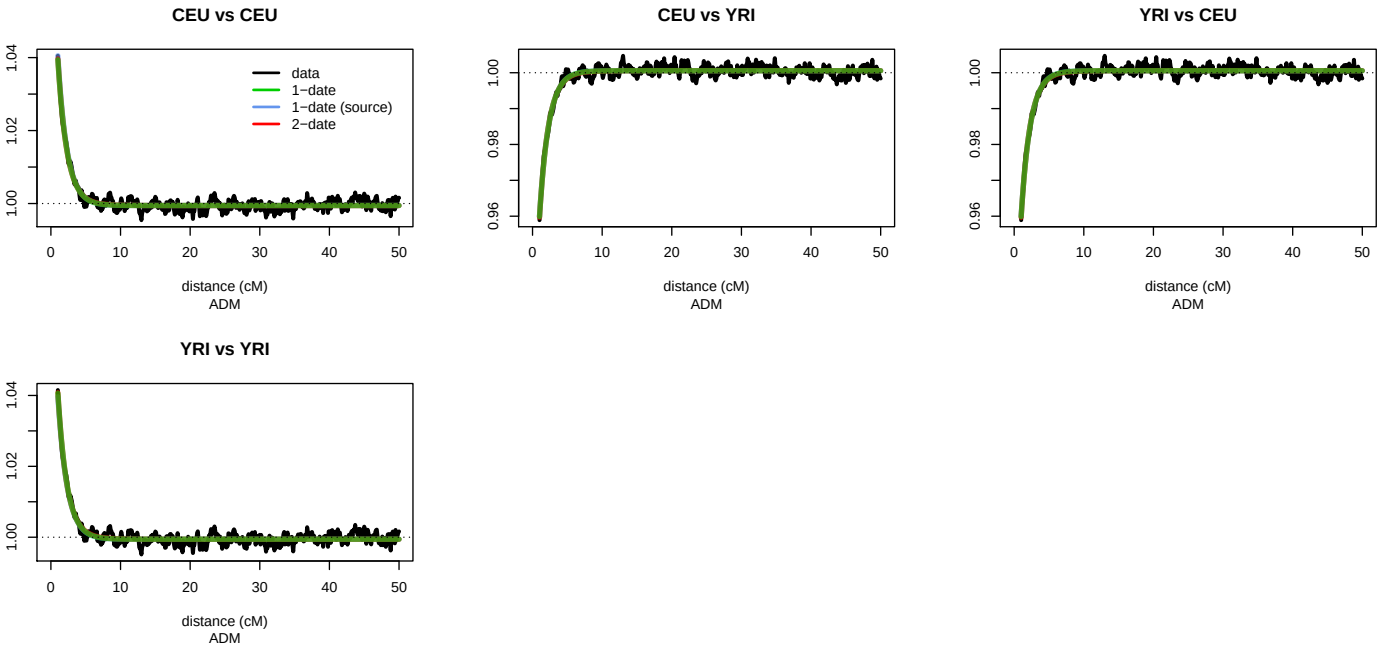

Figure 4.12: 2-way Dating Admixture results with GLOBETROTTER for generation 600. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chunks either from CEU, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from CEU and YRI donors(Y-axis), at varying genetic distances(x-axis). The pattern is the same for all the generations in the 2-way approach. The figures CEU vs YRI or YRI vs CEU shows a negative slope while the remaining shows a positive slope.

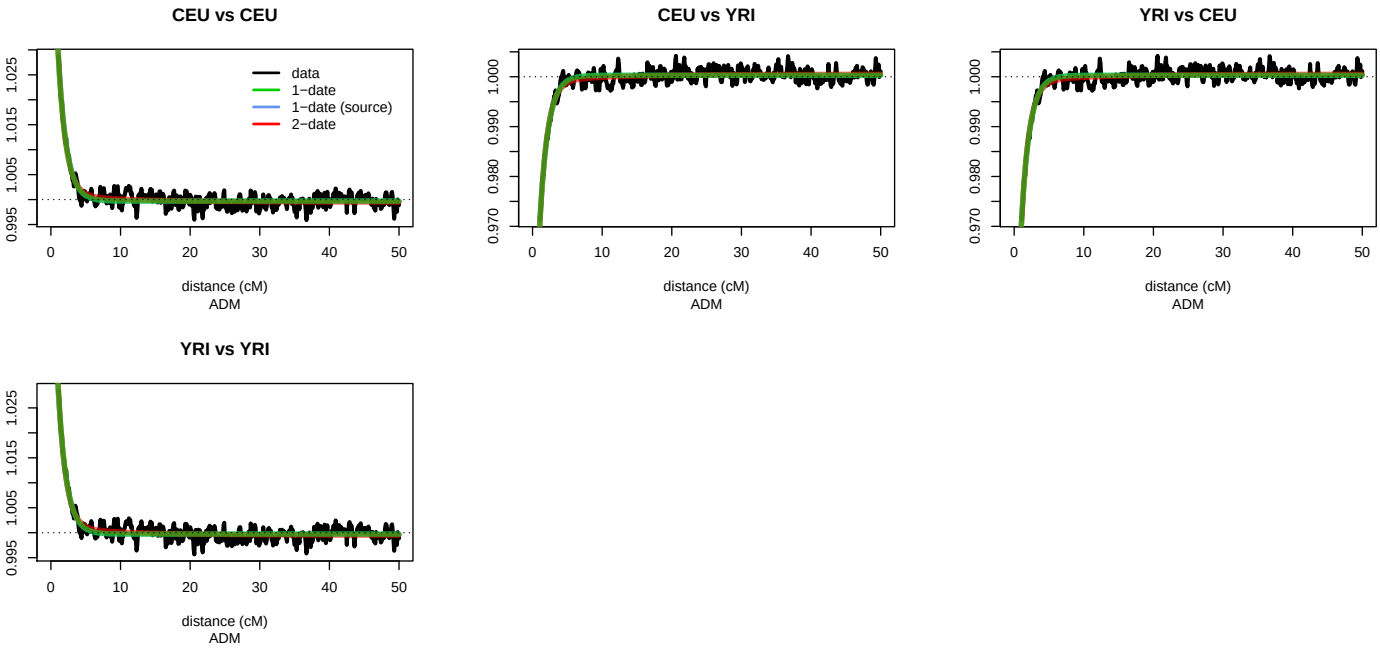

Figure 4.13: 2-way Dating Admixture results with GLOBETROTTER for generation 800. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chunks either from CEU, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from CEU and YRI donors(Y-axis), at varying genetic distances(x-axis). The pattern is the same for all the generations in the 2-way approach. The figures CEU vs YRI or YRI vs CEU shows a negative slope while the remaining shows a positive slope.

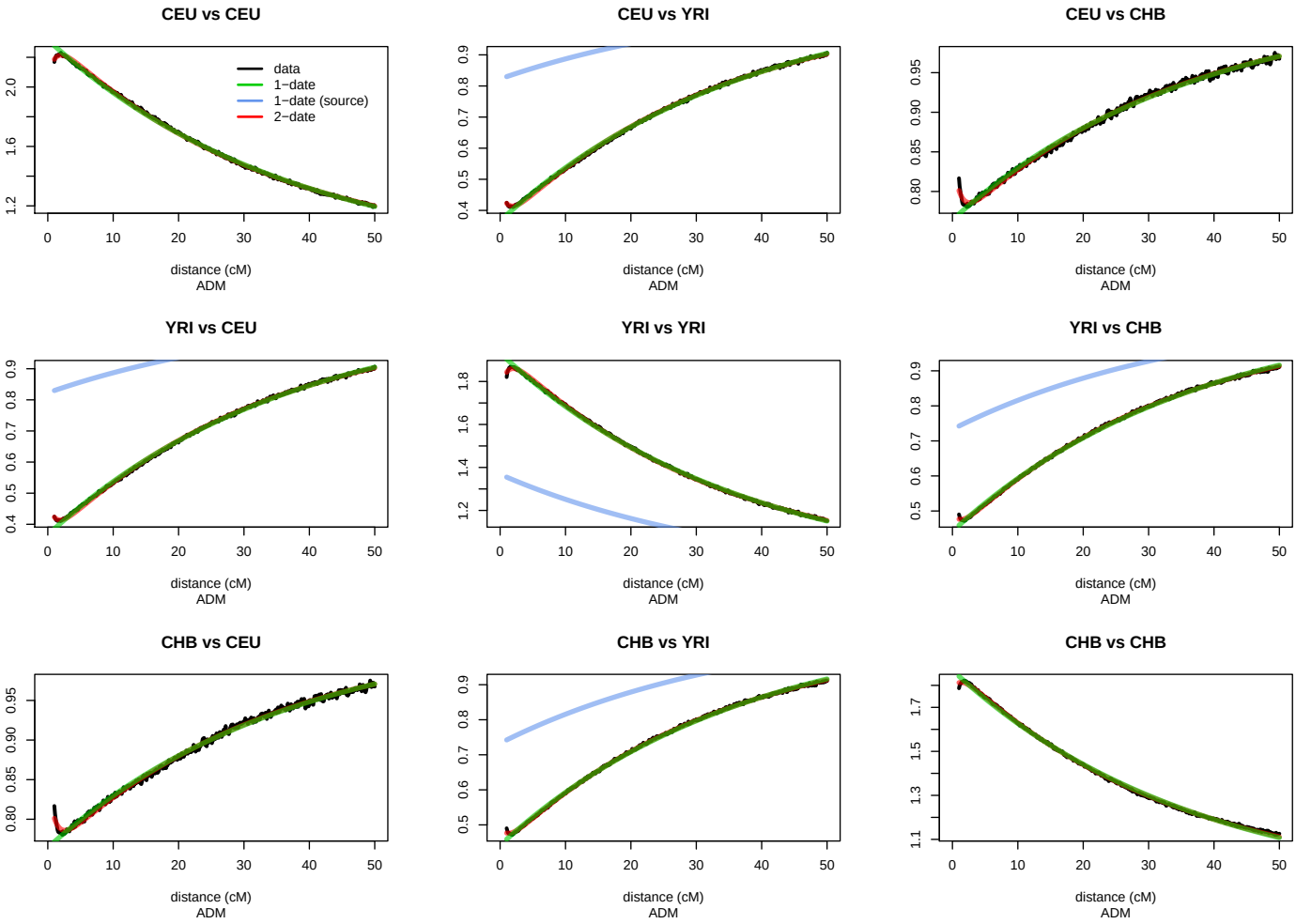

Figure 4.14: 3-way Dating Admixture results with GLOBETROTTER for generation 5. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chuncks either from CEU, or from CHB, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from the combination pairwise between CEU, YRI and CHB donors(Y-axis), at varying genetic distances(x-axis).All the figures in the diagonal shows a negative slope while the remaining shows a positive slope.

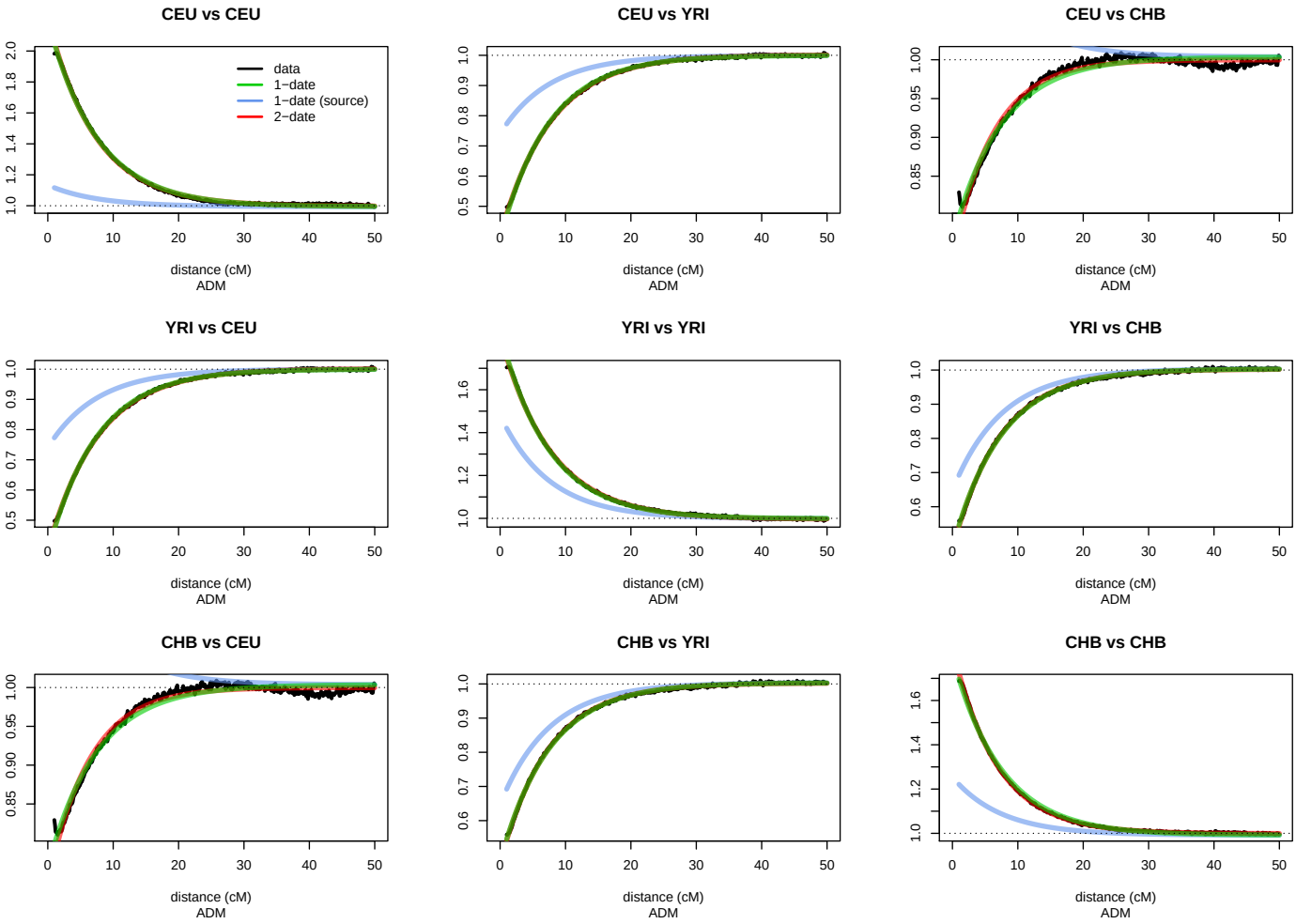

Figure 4.15: 3-way Dating Admixture results with GLOBETROTTER for generation 20. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chuncks either from CEU, or from CHB, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from the combination pairwise between CEU, YRI and CHB donors(Y-axis), at varying genetic distances(x-axis).All the figures in the diagonal shows a negative slope while the remaining shows a positive slope.

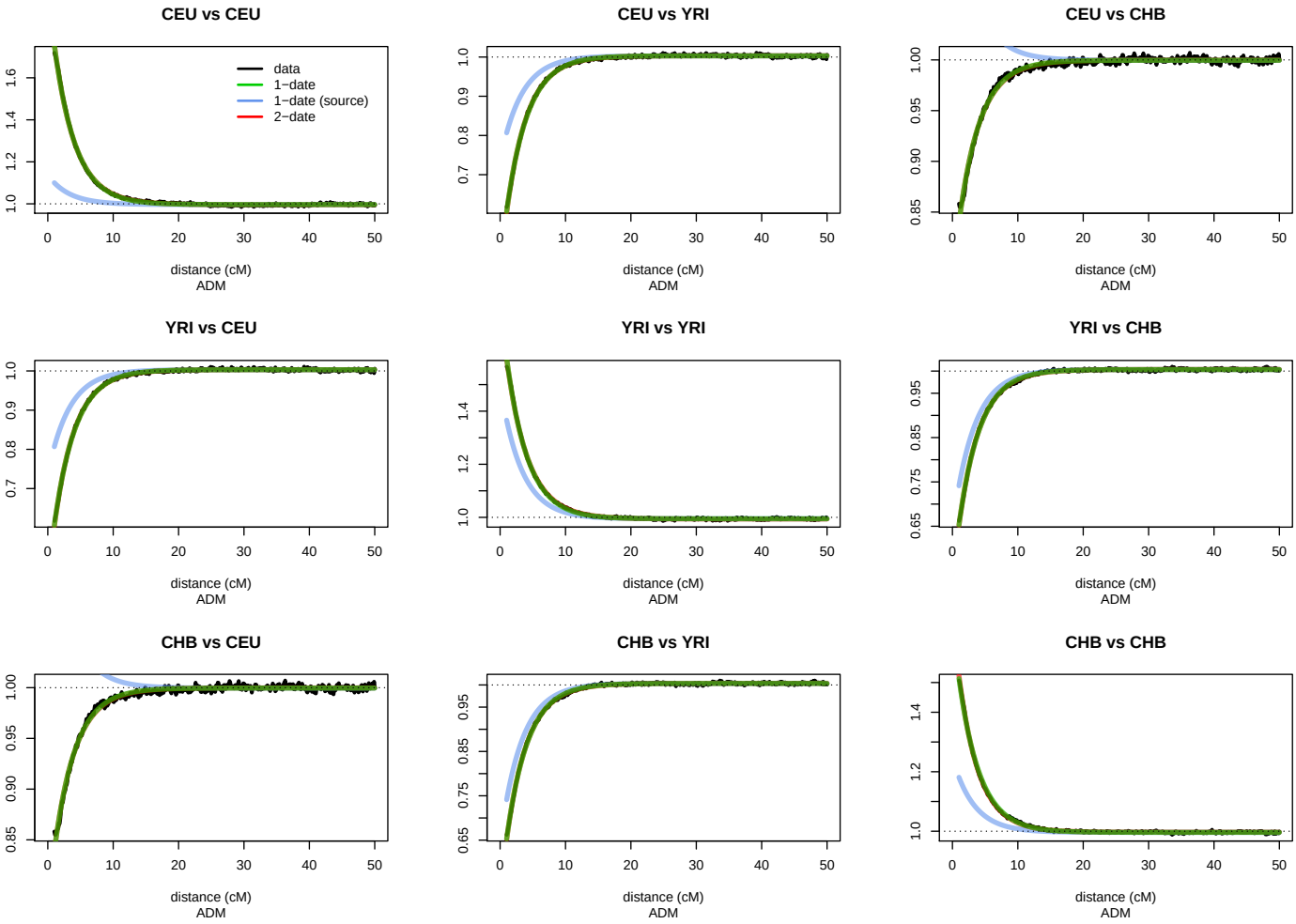

Figure 4.16: 3-way Dating Admixture results with GLOBETROTTER for generation 50. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chuncks either from CEU, or from CHB, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from the combination pairwise between CEU, YRI and CHB donors(Y-axis), at varying genetic distances(x-axis).All the figures in the diagonal shows a negative slope while the remaining shows a positive slope.

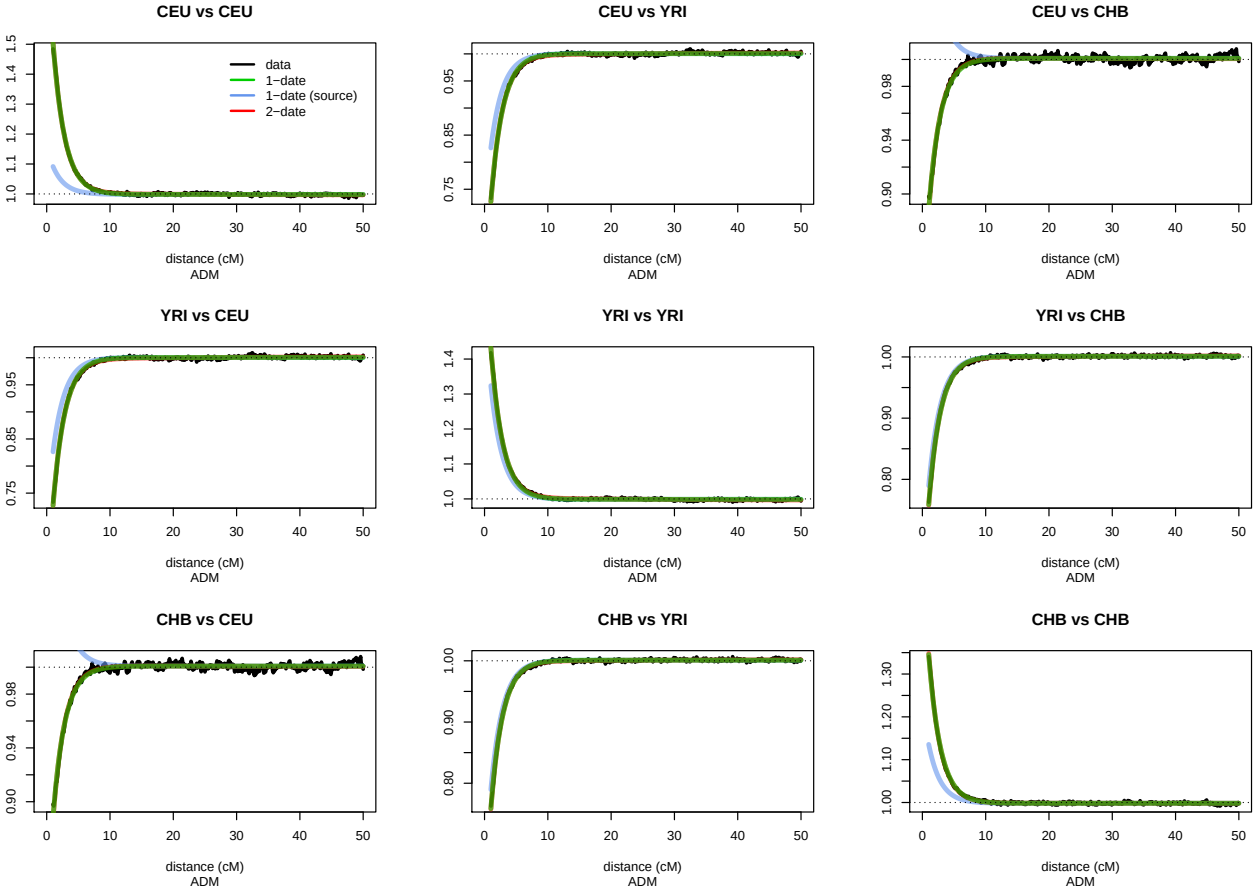

Figure 4.17: 3-way Dating Admixture results with GLOBETROTTER for generation 100. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chuncks either from CEU, or from CHB, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from the combination pairwise between CEU, YRI and CHB donors(Y-axis), at varying genetic distances(x-axis).All the figures in the diagonal shows a negative slope while the remaining shows a positive slope.

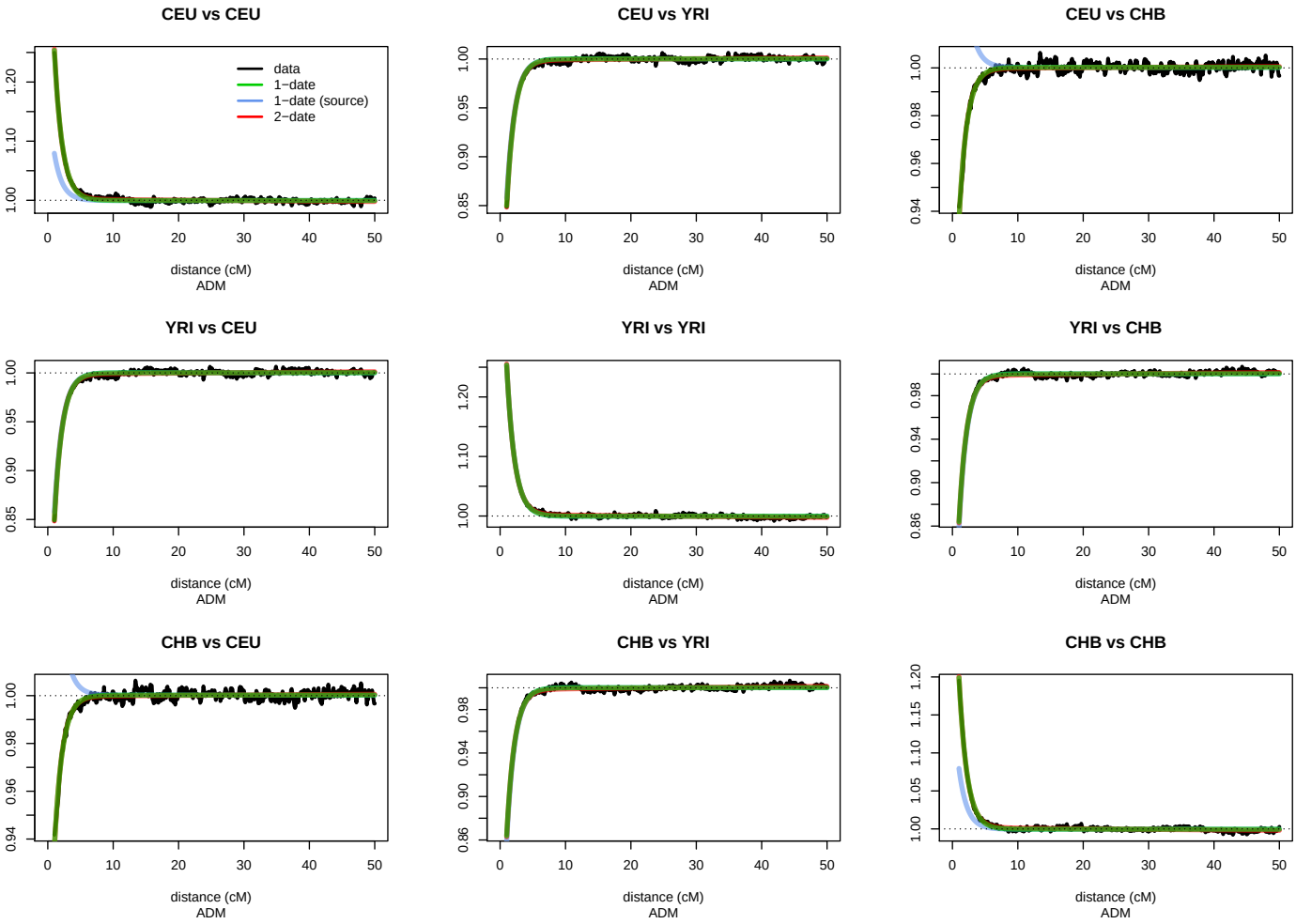

Figure 4.18: 3-way Dating Admixture results with GLOBETROTTER for generation 200. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chuncks either from CEU, or from CHB, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from the combination pairwise between CEU, YRI and CHB donors(Y-axis), at varying genetic distances(x-axis).All the figures in the diagonal shows a negative slope while the remaining shows a positive slope.

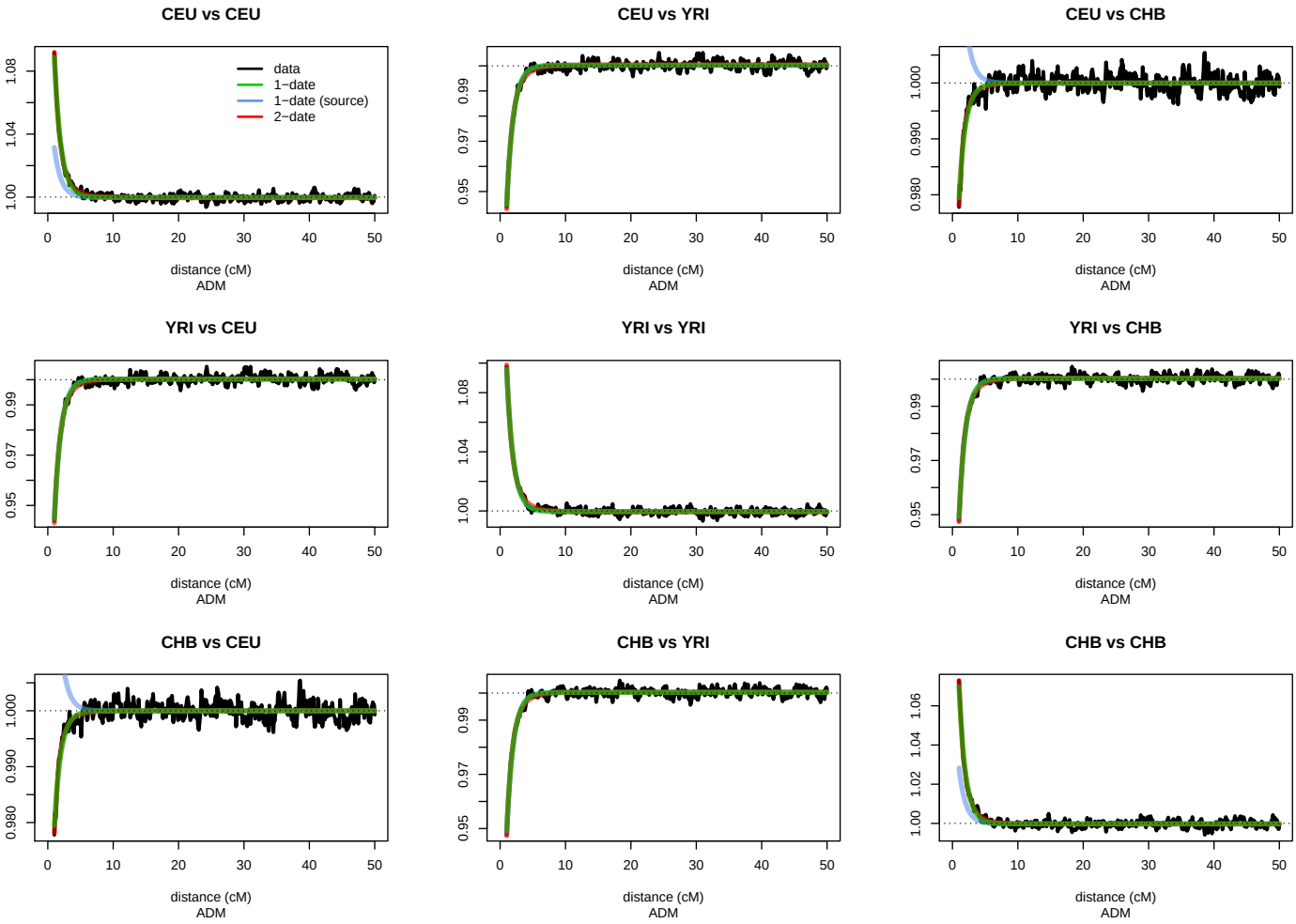

Figure 4.19: 3-way Dating Admixture results with GLOBETROTTER for generation 450. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chuncks either from CEU, or from CHB, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from the combination pairwise between CEU, YRI and CHB donors(Y-axis), at varying genetic distances(x-axis).All the figures in the diagonal shows a negative slope while the remaining shows a positive slope.

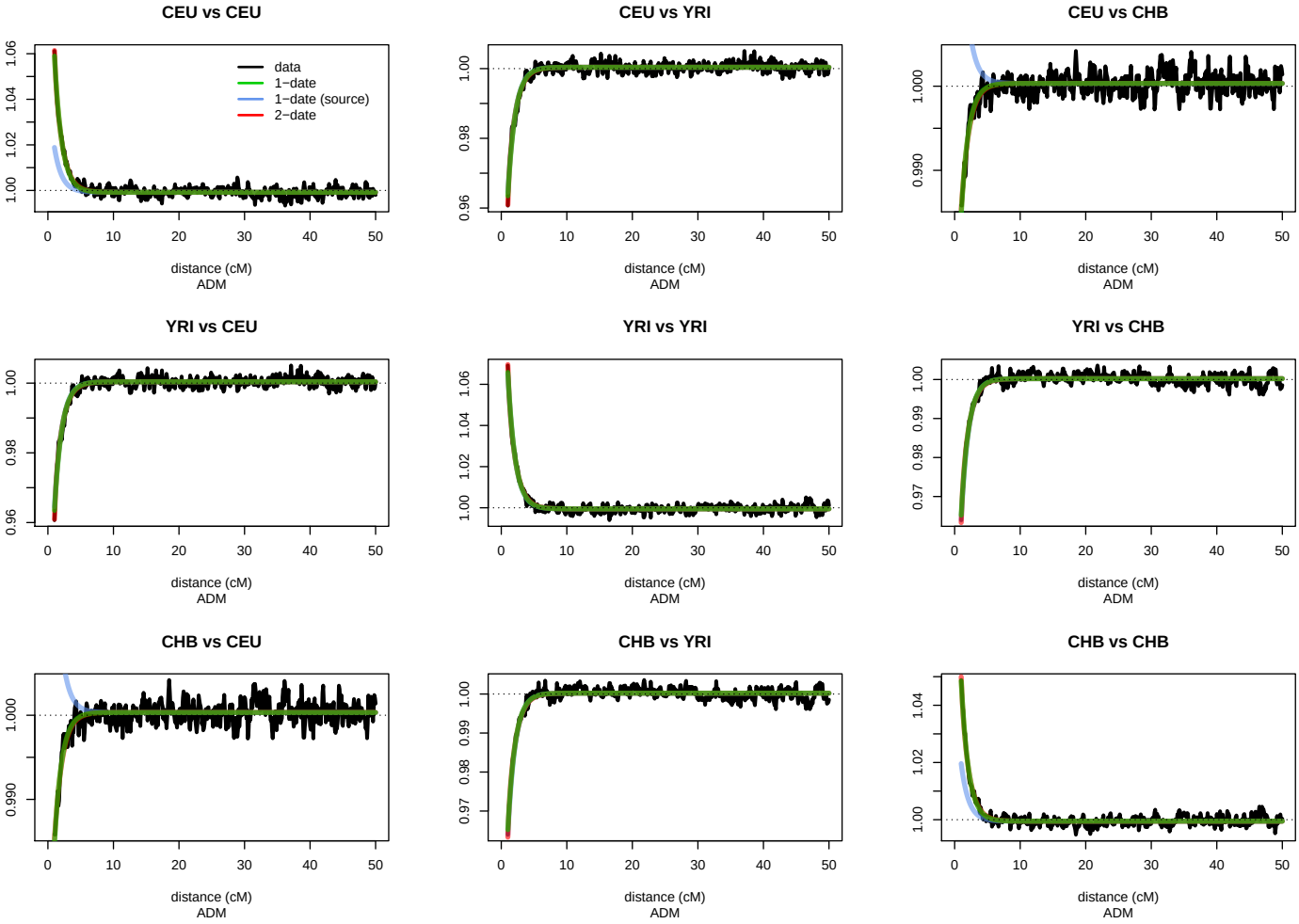

Figure 4.20: 3-way Dating Admixture results with GLOBETROTTER for generation 600. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chuncks either from CEU, or from CHB, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from the combination pairwise between CEU, YRI and CHB donors(Y-axis), at varying genetic distances(x-axis).All the figures in the diagonal shows a negative slope while the remaining shows a positive slope.

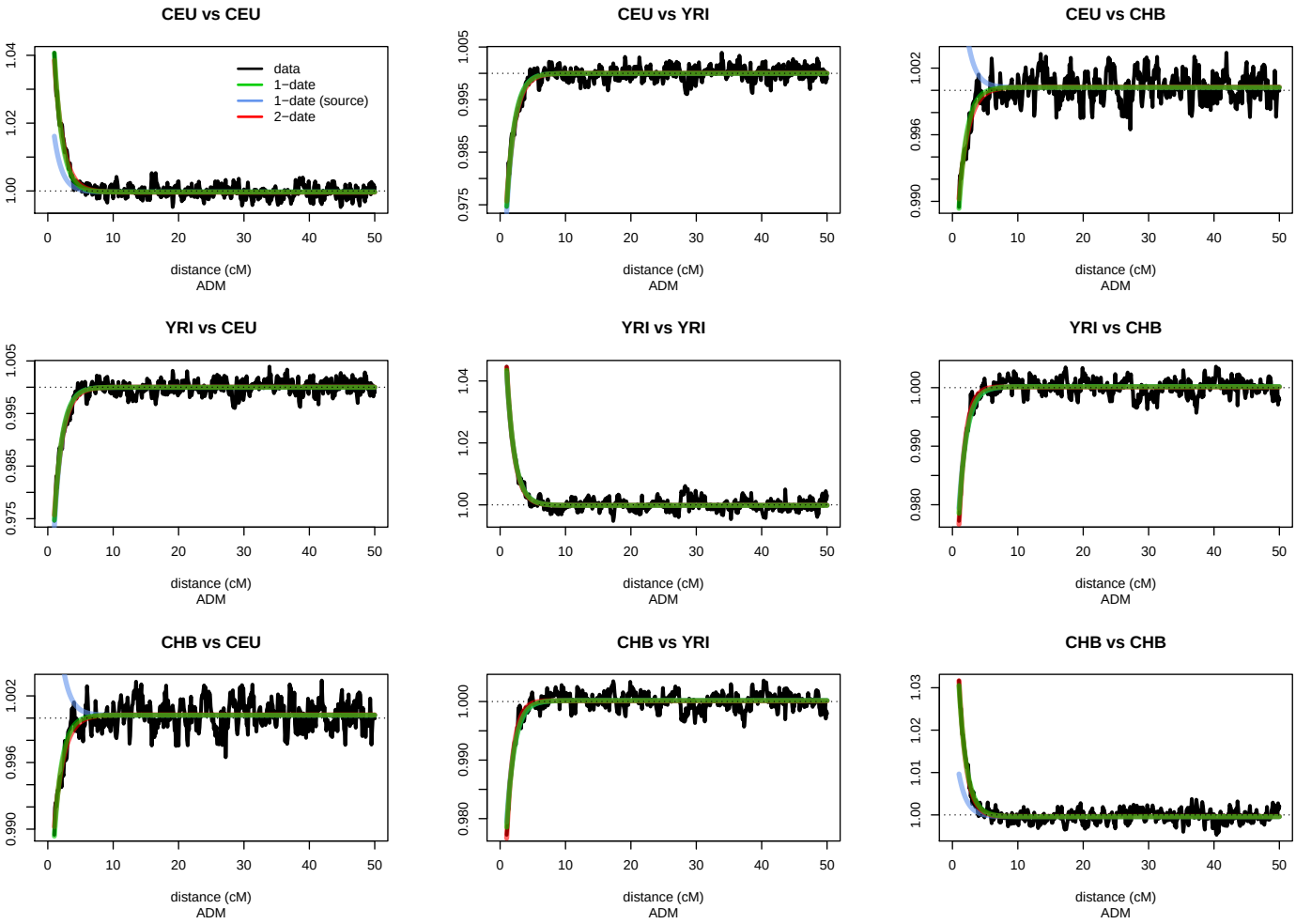

Figure 4.21: 3-way Dating Admixture results with GLOBETROTTER for generation 800. We used CHROMOPAINTER to identify the chunks of DNA inside the genome of each admixed individuals that are most closely related to each donor groups CEU and YRI. GLOBETROTTER jointly fits an exponential distribution to the decay curves for all pairwise combinations of donor groups and determines the single best fitting rate, hence determining the most likely single admixture event and estimating the date it occurred. The curves closely fit an exponential decay (green line) with a rate which is the number of generations. The coancestry curves (black line) shows relative probability of jointly copying two chuncks either from CEU, or from CHB, or from YRI. The negative slope for the curve suggests that these donors contributed to different sides of an admixture event. The positive slope, in the other hand, shows that donors contribute to the same side of the admixture event. The plots shows the relative probability of jointly copying two chunks from the combination pairwise between CEU, YRI and CHB donors(Y-axis), at varying genetic distances(x-axis).All the figures in the diagonal shows a negative slope while the remaining shows a positive slope.

References

S Al-Radi. Brief history of the east african coast. *In The Architecture of Housing, edited by Robert Powell. Singapore: Concept Media/Aga Khan Award for Architecture*, 1990.

DH Alexander and K Lange. Enhancements to the admixture algorithm for individual ancestry estimation. *BMC Bioinformatics. 18; 12():246*, 2011.

DH Alexander, J Novembre, and L Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research 19: 1655-1664*, 2009.

CO Aremu. Exploring statistical tools in measuring genetic diversity for crop improvement. *In: Caliskan M , ed. Genetic diversity in plants . InTech. doi: 10.5772/34950*, 2011.

Y Baran, P Bogdan, S Sankararaman, G Dara, C Gignoux, C Celeste, W Torgerson, R Chapela, G JeanFord, CP Avila, J Rodriguez-Santana, EG Burchard, and E Eran. Fast and accurate inference of local ancestry in latino populations. *Bioinformatics. 28(10): 1359-1367*, 2012.

P Beaujard. East africa, the comoros islands and madagascar before the sixteenth century : on a neglected part of the world system. *Azania : The journal of the British Institute of History and Archaeology in East Africa, Routledge (imprime)/ Taylor & Francis Online (en ligne), 42:15-35*, 2007.

G Bhatia, N Patterson, B Pasaniuc, NA Zaitlen, G Genovese, S Pollack, S Mallick, S Myers, A Tandon, C Spencer, CD Palmer, AA Adeyemo, EL Akylbekova, LA Cupples, J Divers, M Fornage, WHL Kao, L Lange, M Li, S Musani, JC Mychaleckyj, A Ogunniyi, G Papanicolaou, CN Rotimi, JI Rotter, JI Ruczinski, B Salako, DS Siscovick, BO Tayo, Q Yang, S McCarroll, P Sabeti, G Lettre, P De Jager, J Hirschhorn, R Cooper, D Reich, JG Wilson, and AL Price. Genome-wide comparison of african-ancestry populations from care and other cohorts reveals signals of natural selection. *The American Journal of Human Genetics, 89(3), 368-381*, 2011.

G Bhatia, N Patterson, S Sankararaman, and AL Price. Estimating and interpreting fst: The impact of rare variants. *Genome Research, 23(9):1514-1521*, 2013.

E Birney and N Soranzo. Human genomics: The end of the start for population sequencing. *Nature. 526(7571):52-3.*, 2015.

K Bryc, EY Durand, JM Macpherson, D Reich, and JL Mountain. The genetic ancestry of african americans, latinos, and european americans across the united states. *Am J Hum Genet 96(1): 3753*, 2015.

G Busby, G Band, M Jallow, E Bougama, and et al. Admixture into and within sub-saharan africa. *Elife 5: pii: e15266*, 2016a.

GBJ Busby, G Hellenthal, F Montinaro, S Tofanelli, K Bulayeva, I Rudan, T Zemunik, C Hayward, D Toncheva, S Karachanak-Yankova, D Nesheva, P Anagnostou, F Cali, F Brisighelli, V Romano, G Lefranc, C Buresi, J Ben Chibani, A Haj-Khelil, S Denden,

R Ploski, P Krajewski, T Hervig, T Moen, RJ Herrera, JF Wilson, S Myers, and C Capelli. The role of recent admixture in forming the contemporary west eurasian genomic landscape. *Current Biology 25(19), 2518-2526, DOI: http://dx.doi.org/10.1016/j.cub.2015.08.007*, 2016b.

LL Cavalli-Sforza, P Menozzi, and A Piazza. *The history and geography of human genes*. Princeton university press, 1994.

T Chai and RR Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *Geosci. Model Dev., 7, 1247-1250*, 2014.

R Chakraborty and KM Weiss. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Nat l.Acad.Science. 85, 9119-9123*, 1998.

R Cheng, JE Lim, KE Samocha, G Sokoloff, M Abney, AD Skol, and AA Palmer. Genome-wide association studies and the problem of relatedness among advanced intercross lines and other highly recombinant populations. *Genetics, 185(3), 10331044*, 2010.

ER Chimusa, NA Zaitlen, M Daya, M Mller, PD van Helden, NJ Mulder, and EG Hoal. Genome-wide association study of ancestry-specific tb risk in the south african coloured population. *Hum Mol Genet. 23(3):796-809*, 2013.

YS Cho, MJ Go, YJ Kim, JY Heo, JH Oh, HJ Ban, D Yoon, MH Lee, DJ Kim, M Park, SH Cha, JW Kim, BG Han, H Min, Y Ahn, MS Park, HR Han, HY Jang, Cho EY, JE Lee, NH Cho, C Shin, T Park, JW Park, JK Lee, L Cardon, G Clarke, MI McCarthy, JY Lee, JK Lee, B Oh, and HL Kim. A large-scale genome-wide association study of asian populations uncovers genetic factors influencing eight quantitative traits. *Nature Genet. 41(5):527-534*, 2009.

C Churchhouse and J Marchini. Multiway admixture deconvolution using phased or unphased ancestral panels. *Genet. Epidemiology. 37(1): 1-12*, 2012.

DF Conrad, M Jakobsson, G Coop, X Wen, JD Wall, NA Rosenberg, and JK Pritchard. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet. 38(11): 1251-1260*, 2010.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature 526:68-74. doi:10.1038/nature15393.*, 2015.

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature. 449(7164): 851-861*, 2007.

O Delaneau, J Marchini, and JF Zagury. A linear complexity phasing method for thousands of genome. *Nat Methods 9: 179-181. doi: 10.1038/nmeth.1785*, 2012.

B Dobon, HY Hassan, H Laayouni, P Luisi, I Ricao-Ponce, A Zhernakova, C Wijmenga, H Tahir, D Comas, MG Netea, and J Bertranpetit. The genetics of east african populations: a nilo-saharan component in the african genetic landscape. *Sci. Rep. 5, 9996; doi: 10.1038/srep09996*, 2015.

Daniel L Dries. Genetic ancestry, population admixture, and the genetic epidemiology of complex disease. *Circ Cardiovasc Genet. 2: 540-543*, 2009.

I Dunpanloup, S Schneider, and L Excoffier. A simulated annealing approach to define the genetic structure of populations. *Mol Ecol 11: 25712581*, 2002.

BE Engelhardt and M Stephen. Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. *PLoS Genet 6(9), e1001117, doi:10.1371/journal.pgen.1001117*, 2010.

L Excoffier, T Hofer, and M Foll. Detecting loci under selection in a hierarchically structured population. *Heredity 103, 285298; doi:10.1038/hdy.2009.74*, 2009.

D Falush, A Stephens, and JK Pritchard. Inference of population structure: Extensions to linked loci and correlated allele frequencies. *Am J Hum Genet. 164, 1567-1587*, 2003.

L Fejerman, CA Haiman, D Reich, A Tandon, EM John, SA Ingles, CB Ambrosone, CH Bovbjerg, LH Jandorf, W Davis, G Ciupak, AS Whittemore, MF Press, G Ursin, L Bernstein, S Huntsman, BE Henderson, E Ziv, and ML Freedman. An admixture scan in 1,484 african american women with breast cancer. *Cancer Epidemiol Biomarkers Prev 18(11):31103117*, 2009.

GH Gillespie. *Population genetics: A Concise Guide,second edtition*. Baltimore: Johns Hopkins University Press, 2010.

S Gravel. Population genetics models of local ancestry. *Genetics 191(2): 607-619*, 2012.

S Gravel, BM Henn, RN Gutenkunst, AR Indap, GT Marth, AG Clark, F Yu, RA Gibbs, The 1000 Genomes Project, and CD Bustamante. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences 108(29): 11983-11988*, 2011.

D Gurdasani, T Carstensen, F Tekola-Ayele, L Pagani, I Tachmazidou, K Hatzikotoulas, S Karthikeyan, L Iles, MO Pollard, A Choudhury, GRS Ritchie, Y Xue, J Asimit, RN Nsubuga, EH Young, C Pomilla, K Kivinen, K Rockett, A Kamali, AP Doumatey, G Asiki, J Seeley, F Sisay-Joof, M Jallow, S Tollman, E Mekonnen, R Ekong, T Oljira, N Bradman, K Bojang, M Ramsay, A Adeyemo, E Bekele, A Motala, SA Norris, F Pirie, P Kaleebu, D Kwiatkowski, C Tyler-Smith, C Rotimi, E Zeggini, and MS Sandhu. The african genome variation project shapes medical genetics in africa. *Nature, 517(7534), pages 327332. http://doi.org/10.1038/nature13997*, 2015.

MB Hamilton. *Population genetics*. West Sussex, Wiley-Blackwell, 2009.

DL Hartl and AG Clark. *Principles of Population Genetics*. 3rd edn. Sinauer Associates, Inc, Sunderland, MA, 1997.

G Hellenthal, A Auton, and D Falush. Inferring human colonization history using a copying model. *PLoS Genet. 4(5), e1000078*, 2008.

G Hellenthal, GBJ Dusby, G Band, JF Wilson, C Capelli, D Falush, and D S Myers. A genetic atlas of human admixture history. *Science, 343 (6172), 747 DOI: 10.1126/science.1243518*, 2014.

JN Hirschhorn and MJ Daly. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet. 6(2), 95-108*, 2003.

JA Hodgson, CJ Mulligan, A Al-Meeri, and RL Raaum. Early back-to-africa migration into the horn of africa. *PLoS Genet 10(6): e1004393*, 2014.

KE Holsinger and BS Weir. Genetics in geographically structured populations: defining, estimating and interpreting fst. *EEB Articles. 22.http://digitalcommons.uconn.edu/eeb_articles/22*, 2009.

R Hudson, M Slatkin, and W Maddison. Estimation of levels of gene flow from dna sequence data. *Genetics, 132(2):583-589*, 1992.

M Jakobsson, SW Scholz, P Scheet, JR Gibbs, JM VanLiere, H-C Fung, ZA Szpiech, JH Degnan, K Wang, R Guerreiro, JM Bras, JC Schymick, DG Hernandez, BJ Traynor, J Simon-Sanchez, M Matarin, A Britton, J van de Leemput, I Rafferty, M Bucan, HM Cann, JA Hardy, NA Rosenberg, and AB Singleton. Genotype, haplotype and copy-number variation in worldwide human population. *Nature 451(7181), 998-1003*, 2008.

W Jin, S Xu, H Wang, Y Yu, Y Shen, B Wu, and L Jin. Genome-wide detection of natural selection in african americans pre- and post-admixture. *Genome Res, 22, 519527*, 2011.

W Jin, S Wang, H Wang, L Jin, and S Xu. Exploring population admixture dynamics via empirical and simulated genome-wide distribution of ancestral chromosomal segments. *Am J Hum Genet, 91(5): 849862*, 2012.

W Jin, R Li, Y Zhou, and S Xu. Distribution of ancestral chromosomal segments in admixed genomes and its implications for inferring population history and admixture mapping. *European Journal of Human Genetics, 22(7), 930-937, doi:10.1038/ejhg.2013.265*, 2014.

BR Joubert, KE North, Y Wang, V Mwapasa, N Franceschini, SR Meshnick, EM Lange, and the NIAID Center for HIV/AIDS Vaccine Immunology. Comparison of genome-wide variation between malawians and african ancestry hapmap populations. *Journal of Human Genetics 55,366 374*, 2010.

ST Kalinowski. Evolutionary and statistical properties of genetic distances. *Molecular Ecology 11:1263-1273*, 2002.

Joshua Eng Sin Kueh. *THE MANILA CHINESE: COMMUNITY, TRADE AND EMPIRE, C.1570 C. 1770*. PhD thesis, Georgetown University, 2014: 212 pages; 3636414.

DJ Lawson, G Hellenthal, S Myers, and D Falush. Inference of population structure using dense haplotype data. *PLoS Genetics, 8(1), e1002453*, 2012.

WS Watkins LB Jorde and MJ Bamshad. Population genomics: a bridge from evolutionary history to genetic medicine. *Hum. Mol. Genet. 10(20), 2199-2207*, 2001.

RC lewontin. *The Genetic Basis of Evolutionary Change*. Columbia University Press,New York and London, 1974.

H Li, X Cui, and N Arnheim. Direct electrophoretic detection of the allelic state of single dna molecules in human sperm by using the polymerase chain reaction. *Proceedings of the National Academy of Sciences of the United States of America, 87(12), 45804584*, 1990.

N Li and M Stephens. Modelling linkage disequilibrium, and identifying recombination hotspots using snp data. *Genetics. 165(4): 2213-2233*, 2003.

W Li. Three lectures on case-control genetic association analysis. *Brief Bioinform, vol. 9, no. 1, pp. 113, 2008*, 2008.

Z Lin and R Altman. Finding haplotype tagging snps by use of principal components analysis. *Am Soc of Hum Genet. 75(5), 850-61*, 2004.

Yushi Liu, Toru Nyunoya, Shuguang Leng, Steven A. Belinsky, Yohannes Tesfaigzi, and Shannon Bruse. Softwares and methods for estimating genetic ancestry in human populations. *Human Genomics 7(1), 1-7*, 2013.

P-R Loh, M Lipson, N Patterson, P Moorjani, JK Pickrell, D Reich, and B Berger. Inferring admixture histories of human populations using linkage disequilibrium. *Am J Genet. 193(4): 12331254*, 2013.

Po-ru Loh. *Algorithms for Genomics and Genetics: Compression-Accelerated Search and Admixture Analysis*. PhD thesis, Massachusetts Institute of Technology, 2013.

C Lonjou, W Zang, and A Collins. Linkage disequilibrium in human populations. *Proceedings of the National Academy of Sciences of the United States of America, 100(10): 69-74*, 2003.

J Ma and CI Amos. Principal components analysis of population admixture. *PLoS One. 7(7):e40115 doi: 10.1371/journal.pone.0040115*, 2012.

X Mao, AW Bigham, R Mei, G Gutierrez, KM Weiss, TD Brutsaert, F Leon-Velarde, LG Moore, E Vargas, PM McKeigue, MD Shriver, and EJ Parra. A genomewide admixture mapping panel for hispanic/latino populations. *The American Journal of Human Genetics 80(6): 1171-1178*, 2007.

Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *bioRxiv*, 2016. doi: 10.1101/070797. URL http://biorxiv.org/content/early/2016/11/24/070797.

N Masatoshi. Genetic distance between populations. *The American Naturalist .Vol. 106, No. 949, pp. 283-292*, 1972.

N Masatoshi. Definition and estimation of fixation indices. *Evolution 40: 643645*, 1986.

RA Mathias, MA Tauband CR Gignoux, W Fu, S Musharoff, TD O'Connor, C Vergara, DG Torgerson, M Pino-Yanes, SS Shringarpure, L Huang, N Rafaels, MP Boorgula, HR Johnston, VE Ortega, AM Levin, W Song, R Torres, B Padhukasahasram, C Eng, DA Mejia-Mejia, T Ferguson, , ZS Qin, AF Scott, M Yazdanbakhsh, JG Wilson, J Marrugo, LA Lange, R Kumar, PC Avila, LK Williams, H Watson, LB Ware, C Olopade, O Olopade, R Oliveira, C Ober, DL Nicolae, D Meyers, A Mayorga, J Knight-Madden, T Hartert, NN Hansel, MG Foreman, JG Ford, MU Faruque, GM Dunston, L Caraballo, EG Burchard, E Bleecker, MI Araujo, EF Herrera-Paz, K Gietzen, WE Grus, M Bamshad, CD Bustamante, EE Kenny, RD Hernandez, TH Beaty, I Ruczinski, J Akey, and KC Barnes. A continuum of admixture in the western hemisphere revealed by the african diaspora genome. *Nat. Commun. 7:12522 doi: 10.1038/ncomms12522*, 2016.

Paul McKeigue. Prospects for admixture mapping of complex traits. *Am J Hum Genet. 76(1), 1-7*, 2005.

PM McKeigue. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *American Journal of Human Genetics, 63(1), 241251.*, 1998.

G McVean. A genealogical interpretation of principal components analysis. *PLoS Genet 5(10): e1000686. doi:10.1371/journal.pgen.1000686*, 2009.

C Medina-Gomez, JF Felix K Estrada, MJ Peters, L Herrera, CJ Kruithof, L Duijts, A Hofman, CM van Duijn, AG Uitterlinden, VW Jaddoe, and F Rivadeneira. Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the generation r study. *European Journal of Epidemiology, 30(4): 317330*, 2015.

M Moller and EG Hoal. Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. *Tuberculosis. 90(2): 71-83*, 2010.

P Moorjani. *Genetic Study of Population Mixture and Its Role in Human History*. PhD thesis, Harvard University, 2013.

P Moorjani, N Patterson, JN Hirschhorn, A Keinan, L Hao, G Atzmon, E Burns, H Ostrer, AL Price, and David Reich. The history of african gene flow into southern europeans, levantines, and jews. *PLoS Genet. 7(4), e1001373*, 2011.

P Moorjani, K Thangaraj, N Patterson, M Lipson, P-R Loh, P Govindaraj, D Reich, and L Singh. Genetic evidence for recent population mixture in india. *Am J Hum Genet 93(3), 422-438*, 2013.

JC Mueller. Linkage disequilibrium from different scales and applications. *Briefings in Bioinformatics, 5, 355-364*, 2004.

T Murray, TH Beaty, RA Mathias, N Rafaels, AV Grant, MU Faruque, HR Watson, I Ruczinski, GM Dunston, and KC Barnes. African and non-african admixture components in african

americans and an african caribbean population. *Genet Epidemiol. 34(6):10.1002/gepi.20512*, 2010.

X Ni, X Yang, W Guo, K Yuan, Y Zhou, Z Ma, and S Xu. Length distribution of ancestral tracks under a general admixture model and its applications in population history inference. *Scientific Reports, 6, 20048, doi:10.1038/srep20048*, 2016.

J Novembre and M Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet 40(5):646-649*, 2008.

B Padhukasahasram. Inferring ancestry from population genomic data and its applications. *Frontiers in Genetics, 5: 204*, 2014.

B Pasaniuc, S Sankararaman, G Kimmel, and E Halperin. Inference of locus-specific ancestry in closely related population. *Bioinformatics. 25(12): i213-i221*, 2009.

B Pasaniuc, NA Zaitlen, G Lettre, G Chen, A Tandon, W Linda Kao, I Ruczinski, M Fornage, D Siscovick, X Zhu, E Larkin, L Lange, A Cupples, Q Yang, E Akylbekova, S Musani, J Divers, J Mychaleckyj, M Li, G Papanicolaou, R Millikan, C Ambrosone, E John, L Bernstein, W Zheng, J Hu, R Ziegler, S Nyante, E Bandera, S Ingles, M Press, S Chanock, S Deming, J Rodriguez-Gil, C Palmer, S Buxbaum, L Ekunwe, J Hirschhorn, B Henderson, S Myers, C Haiman, D Reich, N Patterson, J Wilson, and AL Price. Enhanced statistical tests for gwas in admixed populations: Assessment using african americans from care and a breast cancer consortium. *PLoS Genet.7(4), e1001371*, 2011.

N Patterson, N Hattangadi, and B Lane. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet. 74(5), 979-1000*, 2004.

N Patterson, AL Price, and D Reich. Population structure and eigenanlysis. *PLoS Genet. 2(12), e190*, 2006.

NJ Patterson, P Moorjani, Y Luo, S Mallick, N Rohland, Y Zhan, T Genschoreck, T Webster, and D Reich. Ancient admixture in human history. *Genetics 92(3):1065-1093*, 2012.

E Peprah, H Xu, F Tekola-Ayele, and CD Royal. Genome-wide association studies in africans and african americans: Expanding the framework of the genomics of human traits and disease. *Public Health Genomics, 18(1), 40-51*, 2015.

CL Pfaff, EJ Parra, C Bonilla, K Hiester, PM McKeigue, MI Kamboh, RG Hutchinson, RE Ferrell, E Boerwinkle, and MD Shriver. *Population Structure in Admixed Populations: Effect of Admixture Dynamics on the Pattern of Linkage Disequilibrium*. Am. J. Hum. Genet. 68, 198207, 2001.

JK Pickrell and JK Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet 8(11): e1002967.doi:10.1371*, 2012.

JK Pickrell and D Reich. Toward a new history and geography of human genes informed by ancient dna. *Trends Genet, 30(9):377389*, 2014.

J Pool and R Nielsen. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics 181(2), 711-719*, 2009.

AL Price, NJ Patterson, R Plenge, M Weinblatt, N Shadick, and D Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet. 38(8): 904-909*, 2006a.

AL Price, NJ Patterson, RM Plenge, ME Weinblatt, NA Shadick, and D Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet. 38(8):904-909*, 2006b.

AL Price, N Patterson, and Y Fuli. A genomewide admixture map for latino populations. *Am J Hum Genet. 80(6), 1024-1036*, 2007.

AL Price, A Tandon, N Patterson, K Barnes, N Rafaels, I Ruczinski, T Beaty, R Mathias, D Reich, and S Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *Plos Genet. 6(5):1-18*, 2009.

JK Pritchard and M Przeworski. Linkage disequilibrium in humans: models and data. *Am J Hum Genet. 69:114. doi: 10.1086/321275*, 2001.

JK Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics, 155(2), 945959*, 2000.

I Pugach, R Matveyev, A Wollstein, M Kayser, and M Stoneking. Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biology. 12(2): R19*, 2011.

I Pugach, F Delfin, E Gunnarsdttir, M Kayser, and M Stoneking. Genome-wide data substantiate holocene gene flow from india to australia. *Proceedings of the National Academy of Sciences 110(5): 1803-1808, 10.1073/pnas.1211927110*, 2013.

S Purcell, B Neale, K Todd-Brown, L Thomas, MA Ferreira, D Bender, J Maller, P Sklar, de Bakker, MJ Daly, and PC Sham. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet. 81(3): 559-575*, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

D Reich, M Cargill, S Bolk, J Ireland, PC Sabeti, DJ Richter, T Lavery, R Kouyoumjiam, SF Farhadian, R Ward, and ES Lander. Linkage disequilibrium in the human genome. *Nature, 411, 199204*, 2001.

JH Relethford. *Human Population genetics*. John Wiley and Sons, Inc, 2012.

JH Relethford and RM Harding. *Population Genetics of Modern Human Evolution*. John Wiley and Sons, Ltd, 2001.

Markus Ringner. What is principal component analysis? *Nat Biotech, 26(3):303-304*, 2008.

JM Rodriguez, S Bercovici, M Elmore, and S Batzoglou. Ancestry inference in complex admixtures via variable-length markov chain linkage models. *Journal of Computational Biology, 20(3), 199211. http://doi.org/10.1089/cmb.2012.0088*, 2013.

Alex Saltarin. Dna used to create genetic atlas of human history. http://www.techtimes.com/articles/3457/20140214/dna-used-to-create-genetic-atlas-of-human-history.htm, june 2014.

FM Salzano and M Sans. Interethnic admixture and the evolution of latin american populations. *Genetics and Molecular Biology, 37(1, Suppl. 1): 151-170*, 2014.

GD Sandefur and McK Trudy. American indian intermarriage. *Social Science Research 15 (December, 1986):347-371*, 1986.

J Sanderson, H Sudoyo, TM Karafet, MF Hammer, and MP Cox. Reconstructing past admixture processes from local genomic ancestry using wavelet transformation. *Genetics. 200(2): 469481*, 2015.

S Sankararaman, S Sridhar, G Kimmel, and E Halperin. Estimating local ancestry in admixed populations. *Hum Genet 82, 290303*, 2008.

MF Seldin, B Pasaniuc, and AL Price. New approaches to disease mapping in admixed populations. *Nat Rev Genet. 36, S21-S27*, 2011.

MD Shriver, EJ Parra, S Dios, C Bonilla, H Norton, C Jovel, C Pfaff, C Jones, A Massac, N Cameron, A Baron, T Jackson, G Argyropoulos, L Jin, CJ Hoggart, PM McKeigue, and RA Kittles. Skin pigmentation, biogeographical ancestry and admixture mapping. *Human Genetics 112(4): 387-399*, 2003.

M Slatkin. Linkage disequilibrium understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics 9(6): 477485. doi:10.1038/nrg2361*, 2008.

Werner Sollors. *Interracialism: Black-White Intermarriage in American History, Literature, and Law*. Oxford University Press, 2000.

H Tang, M Coram, P Wang, X Zhu, and N Risch. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet. 79, 1-12*, 2006.

TA Thornton and JL Bermejo. Applications to genetic association analysis for admixed populations. *Genetic Epidemiology, 38(01), S5S12*, 2014.

SA Tishkoff and KK Kidd. Implications of biogeography of human populations for 'race' and medicine. *Nat Rev Genet. 36, S21-S27*, 2004.

SA Tishkoff and BC Verrelli. Patterns of human genetic diversity: Implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet. 4:293-340*, 2003.

SA Tishkoff, FD Reed, F Friendlaender, C Ehret, and AL Ranciaro. The genetic structure and history of africans and african americans. *Sciences. 324, 1035-1044*, 2009.

BA Walther and JL Moore. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators with a litterature of estimators performance. *ECOGRAPHY 28: 815-829*, 2005.

BS Weir and CC Cockerham. Estimating f-statistics for the analysis of population structure. *Evolution. 38, 1358-1370*, 1984.

BS Weir and WG Hill. Estimating f-statistics. *Annu Rev Genet 36:721750*, 2002.

KM Weiss and AG Clark. Linkage disequilibrium and mapping of human traits. *Trends in Genetics. 18(1):1924*, 2002.

CA Winkler, GW Nelson, and MW Smith. Admixture mapping comes of age. *Annu.Rev.Genomics Hum. Genet. 11, 65-89*, 2010.

GC Smidt Wolbert. A chinese in the nubian and abyssinian kingdoms (8th century). *Chroniques ymnites, URL : http://cy.revues.org/33 ; DOI : 10.4000/cy.33*, 2002.

S Xu and L Jin. Chromosomewide haplotype sharing: A measure integrating recombination information to reconstruct the phylogeny of human populations. *Annals of Human Genetics, 75(6), 694-706*, 2011.

S Xu, W Huang, J Qian, and L Jin. Analysis of genomic admixture in uyghur and its implication in mapping strategy. *The American Society of Human Genetics 82(4):883-894*, 2008.

B Yunusbayev, M Metspalu, E Metspalu, A Valeev, S Litvinov, R Valiev, V Akhmetova, E Balanovska, O Balanovsky, S Turdikulova, D Dalimova, P Nymadawa, A Bahmanimehr, H Sahakyan, K Tambets, S Fedorova, N Barashkov, I Khidiyatova, E Mihailov, R Khusainova, L Damba, M Derenko, B Malyarchuk, L Osipova, M Voevoda, L Yepiskoposyan, T Kivisild, E Khusnutdinova, and R Villems. The genetic legacy of the expansion of turkic-speaking nomads across eurasia. *PLoS Genet 11(4): 1-24, 10.1371/journal.pgen.1005068*, 2015.

X Zhu and RS Cooper. Admixture mapping provides evidence of association of the vnn1 gene with hypertension. *PLoS One 2(11): e1244. doi:10.1371/journal.pone.0001244*, 2007.

X Zhu, RS Cooper, and RC Elston. Linkage analysis of a complex disease through use of admixed populations. *Am J Hum Genet.74, 11361153*, 2004.