Investigating local ancestry inference models in mixed ancestry individual genomes

Ephifania Geza

Supervisor: Dr. Gaston K. Mazandu
Co-supervisors: Prof. Emile R. Chimusa and Prof. Nicola J. Mulder

September 7, 2022

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the Department of Integrative Biomedical Sciences, at the Faculty of Health Sciences, University of Cape Town

Declaration

I, Ephifania Geza, hereby declare that the work contained in this thesis is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other University. I authorise the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate Date: September 7, 2022

Abstract

Owing to historical events including the slave trade, agricultural interests, colonialism, and political and/or economical instability, most modern humans are a mosaic of segments originating from different populations. They result from the interbreeding of two or more previously isolated populations, leading to admixture. Known admixed populations include the mixed ancestry of South Africa, Latin Americans and African Americans. Admixed individuals play important roles in understanding population history, disease aetiology, and personal genomics. Accordingly, efforts have been made to understand the genetic composition of such individuals, yielding several models that infer the ancestry of every chromosomal segment in admixed individuals (local ancestry). However, new research questions emerged concerning model statistical and biological parameters, as well as the performance of these models across admixed datasets. This elicited the need for examining existing local ancestry inference models in order to identify and tackle critical issues of these models, which is the main goal of this thesis. We achieve this in four steps, constituting the main contributions of this PhD project: (1) Qualitative assessment of existing models through a systematic review; (2) Building a unified framework integrating existing models for inferring and assessing local ancestry estimates; (3) Quantitative assessment of existing methods within the same framework; and (4) Proposing a model extension to account for natural selection and the origin of modern humans to improve the accuracy of local ancestry estimates.

Firstly, we assess models using published results on different datasets and performance measures, to orient modellers and software developers on the future trends in local ancestry inference. Secondly, to address the challenges identified in (1) including model complexity reflected in the distinct inputs each model requires and outputs formats, we design a unified framework, referred to as FRANC, to manipulate tool-specific inputs, deconvolve ancestry and standardise outputs, to ease the inference process and pave the way for model assessment. Thirdly, using FRANC, we assess the performance of eight state-of-the-art models on simulated admixed population datasets involving three and five ancestral populations. LAMP-LD and LOTER performed better than the other six tested models on admixed populations involving five an-

cestral populations while RFMIX, WINPOP, ELAI and LAMP-LD were comparable in admixed datasets involving three populations. Performance was evaluated based on performance measures borrowed from the machine learning confusion matrix. Finally, we noted that it may be more practical to extend existing models to incorporate more realistic biological assumptions. Hence, we propose a nonparametric hidden Markov model, that adjusts an existing model mSPECTRUM to account for natural selection and state-persistence when deconvolving local ancestry, which should improve the accuracy of estimates. Similarly to mSPECTRUM, this acknowledges the two common hypotheses on the origin of modern humans, making it comparable to mSPECTRUM which has been shown to be competitive with HAPMIX, a benchmark for two-way admixtures. Therefore, these four are a good contribution to admixture analysis of populations.

. . . To my late supervisor, Dr. Gaston Kuzamunu Mazandu . . .

Acknowledgements

I am forever grateful to the God of hosts and my Saviour for accomplishing the great work He started in me and watching over His word until it was effected.

Secondly, I appreciate my research supervisor, the late Dr. Gaston K. Mazandu for believing in me, introducing me to the Bioinformatics field, his patience, inspiring drive, thoughtful criticisms, analytical and computational skills. May his dear soul continue to rest in the Light. My co-supervisors Profs. Emile R. Chimusa and Nicola Mulder for introducing me to population genetics, biology and helping me throughout my time at the University of Cape Town (UCT). Most importantly, I acknowledge my funders the Organization for Women in Science for the Developing World (OWSD) and Swedish International Development Cooperation Agency (Sida).

I acknowledge the love and moral support of my family: my dad, Mr Lawrence T. Geza, for confessing the accomplishment of this PhD during my primary education level; mom, for giving me the values that shaped me to a woman I am today; brothers, Tafadzwa Geza for providing a shoulder to lean on and teaching me to be kind, Terrence and Peter for teaching me to be considerate, tolerant and patient in life and Mary Geza, a sister, mother, friend, mentor and caretaker. You are special, I love you all.

The moral support of Prof. Edward Chiyaka and his wife, Prof. Emmanuel Adabor, Mr. and Mrs. Mutemaringa and Ms Olina Ngwenya boosted my confidence to keep moving. I acknowledge Prof. Collet Dandara's counselling and words of encouragement; God bless you. Finally, my indebtedness gratitude goes to my colleagues and staff members at the Computational Biology group (CBIO), from the UCT and the African Institute for Mathematical Sciences throughout all centres. Special thanks to: Dr. Nana A. Mbroh, Noella Aganze, Rhoda Mahamah, Irene Kyomugisha, Fumnilayo L. Makinde, Jacquiline W. Mugo, Samar Alsheikh and Onias Manjangaya who proof read my work most of the times. I appreciate your support! Essential thanks goes to my husband to be, for his incredible love and support during my thesis writing.

Contents

List of Figures

List of Tables

Publications

The University authorises publication of the whole or part of the thesis depending on the project outputs and agreement between the candidate and supervisor(s). In the next sections we outline the publications (1) Directly related to this thesis and (2) Not directly related to this thesis.

Publications directly related to this thesis

We published some papers and submitted others for publication as outlined below.

1 Geza, E., Mugo, J., Mulder, N. J., Wonkam, A., Chimusa, E. R., Mazandu, G. K. (2018). A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. Briefings in Bioinformatics bby044.

2. Geza, E., Mulder, N. J., Chimusa, E. R., Mazandu, G. K. (2019). FRANC: A unified framework for multi-way local ancestry deconvolution with high density SNP data. Briefings in Bioinformatics.

3. Mazandu, G. K., Geza, E., Seuneu, M., Chimusa, E. R. (2019). Orienting Future Trends in Local Ancestry Deconvolution Models to Optimally Decipher Admixed Individual Genome Variations. In: Samadikuchaksaraei, A., Seifi, M. (Eds.), Bioinformatics Tools for Detection and Clinical Interpretation of Genomic Variations. IntechOpen, Rijeka, Ch. 3.

Publications indirectly related to this thesis

Although not directly related to the thesis during this PhD training, we also published the following:

1. Mugo, J. W., Geza, E., Defo, J., Elsheikh, S. S., Mazandu, G. K., Mulder, N. J., Chimusa, E. R. (2017). A multi-scenario genome-wide medical population genetics simulation framework. Bioinformatics.

2. Yocgo, R.E., Geza, E., Chimusa, E.R. and Mazandu, G.K. (2017). A post-gene silencing bioinformatics protocol for plant-defence gene validation and underlying process identification: case study of the Arabidopsis thaliana NPR1. BMC Plant Biology.

3. Mazandu, G.K., Kyomugisha, I., Geza, E., Tchamga, M., Bah, B. and Chimusa, E.R. (2019). Designing data driven learning algorithms: a necessity to ensure effective post-genomic medicine and biomedical research. In the book: Machine Learning in Medicine and Biology.

Chapter 1

Introduction

1.1 Overview

The origins of modern humans is still a topic of significant interest for many researchers. There are two most common hypotheses on the origin of modern humans: the multi-regional and the recent African origin [1, 9, 10]. Both hypotheses agree that modern humans originated in Africa [11] and migrated to different parts of the world (first out of Africa migration). The multi-regional hypothesis states that *homoerectus* in different geographical regions evolved to archaic humans including Neanderthal and Denisovans nearly 1.0–1.3 million years ago [9, 10]. The archaic humans finally developed into modern humans and appeared in Africa nearly 200 thousand years ago (kya) [9]. Despite the possibility of being extinct, Neanderthal occupied some parts of Europe and West Asia almost 30 kya. Contrarily, the recent African hypothesis submits that *homoerectus* who remained in Africa after the first migrations out of Africa developed into *homosapiens*. These *homosapiens* then developed into modern humans almost 200 kya and, subsequently migrated to different parts of the world nearly 80–120 kya, replacing the other co-existing archaic humans, including Neanderthal in Asia and Europe [9, 10, 12]. Hence, the recent African hypothesis is also called the African replacement hypothesis [13]. **Figure 1.1** summaries these two most common hypotheses on the origin of modern human.

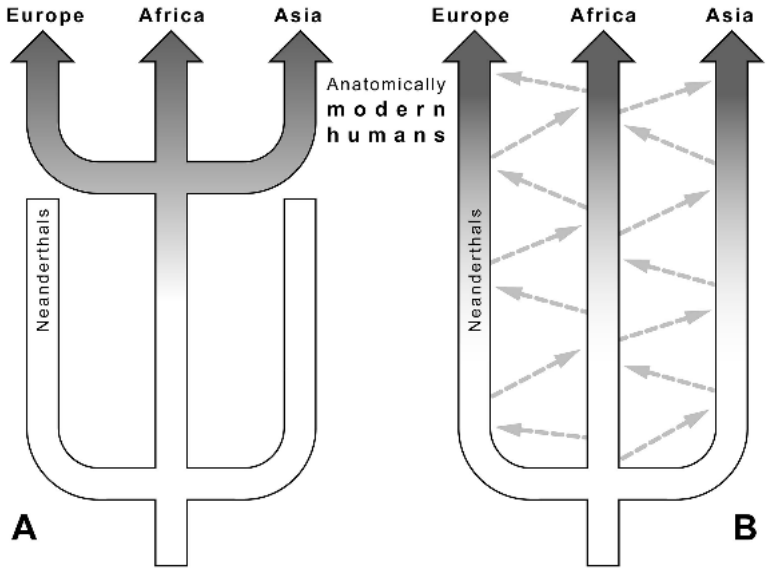


Figure 1.1: Two most common hypotheses on the origins of modern humans: (A) the recent African versus (B) the multi-regional, adapted from Murray et al. [1]. Dashed arrows represent the possible matings between different archaic human groups and darker continuous arrows represent how close the archaic human groups get to the modern humans.

The patterns of variations observed in modern humans is compatible with both hypotheses. However, it is believed that the number of individuals that have contributed to modern human numbers, referred to as effective population size, slightly support the multi-regional hypothesis. All individuals are assumed to originate from a population of 10 000 individuals [14]. This value is too small to facilitate the movement of individuals between three continents at a density that would maintain the gene flows [13]. On the other hand, the patterns largely support the recent African hypothesis [13]. For example, the high diversity observed in the African populations concur with the recent African hypothesis, assuming less drift in such populations [13]. More so, the large genetic distances between the Sub-Saharan Africans and the Europeans or the Asians also concur with the out of Africa migrations[1]. Irrespective of the popularity of the recent African hypothesis, several studies revealed that some non-African populations share genetic material with archaic humans. This is the case of modern Eurasians and Melanesians sharing 1–6% genetic material with archaic humans [9, 15, 16, 17]. One possible explanation for this is the possibility of gene flow between non-Africans and archaic humans [15, 17]. Thus, regardless of the hypothesis, modern humans developed through the exchange of genetic materials across individuals [18]. Could this therefore explain the individual differences in physical appearance, disease susceptibility, drug response or generally

[1] "Out of Africa migrations" tells of the movement of humans from Africa to different parts of the world.

the different phenotypes[2] within or between populations? The next section discusses the variations that exist among population individuals.

1.2 Human variations

Observable phenotypes differ within or between populations [20]. Such observable phenotypes include physical appearance, disease susceptibility, drug response, etc. It has been shown that observable phenotypes result from the genes each individual carries, their environment[3] [21] and epigenetic[4] factors [22]. Patterns of individual differences are important for understanding population history and diversity, investigating population structure and are pertinent to human health [23]. For example, recently, Baud et al. [24] demonstrated the influence of human variations on different human phenotypes. It was shown that social genetic effects (SGEs) contribute almost 29% of the variations in the organism's phenotypes. Therefore, an individual's genetic make-up may contribute to the health or the disease risk of their peers [24]. This is consistent with the peer drinking, postulating that an individual's alcohol consumption may increase the risk of drinking of their peers. This section discusses human variations in detail, from forms of variations describing populations, methods used in detecting the variation patterns, factors influencing these variations to approaches which can be used to capture such variations.

1.2.1 Origins of human variations

In humans, most phenotypic variations are due to sequence variations [25]. Previously, it has been shown that the patterns and nature of human genetic variations result from natural variations in individual human genomes, the effects of genetic drift, the mixing of previously isolated populations (admixture) or gene flow, mutation and natural selection [9, 26, 27, 28]. Human genetic variations refer to the differences that are seen in the human Deoxyribonucleic acid (DNA) sequences due to polymorphisms [12, 29, 30]. Human individuals contain information in the form of sequences coded in letters A, C, G and T in the DNA [29], called bases or nucleotides [22], which make up individual genomes. The human genome is made-up of almost "3.1 billion base pairs" and about 20 000 genes [12, 25]. An offspring inherits two different sets of information (genome) from each parent [22, 29]. The inherited information further interacts

[2] A phenotype is a distinguishable feature of an individual [19].

[3] Environment is the individual's surroundings including but not limited to altitude, social factors and life style [21].

[4] Epigenetic factors are contributed by changes in gene expression [22].

with the environment leading to observed differences in disease susceptibility, drug response, physical appearance and other phenotypes present in humans [31]. Inheriting the same base pair from both parents (e.g. G and G) is called the homozygous state, while inheriting different bases from the two parents (e.g. G and A) is called the heterozygous state [32, 33]. Despite the observed individual differences, it is surprising to note that generally, two individuals, either related or unrelated are 99.5%–99.9% similar in DNA sequences [12, 32, 34, 35, 36]. This could be due to the sharing of the most common recent ancestor [12], concurring with the recent African hypothesis on modern human origins [18]. The 0.5%–0.1% variation separates humans, for which 93%–95% is observed within a population, while 7%–5% is observed between populations [13, 37]. As a result, humans are rarely differentiated by race [13, 34], but rather by the differences in DNA sequences.

DNA sequence variations occur due to nucleotide insertions, deletions, substitutions, differences in number of tandem repeats (NTRs) and in copy number of genomic segments or a combination of these [25, 30, 32]. When DNA sequence variations are observed in most individuals within a population, particularly at a frequency greater than 1%, they are called DNA polymorphisms [30]. Insertions, deletions and substitutions of nucleotides, and differences in sequence repeats and number of genomic segments yield to different forms/classes of DNA polymorphisms. These DNA polymorphism classes include but are not limited to restriction fragment length polymorphisms (RFLPs), structural variations (SVs), tandem repeats (TRs) and single nucleotide polymorphisms (SNPs) [25, 30]. RFLP is a variable length piece of the DNA fragment that is cut from a large molecule by a restriction enzyme protein (or restriction endonuclease) [30]. RFLPs are mostly used for discovering genes associated with Mendelian diseases [30]. Such diseases result from a mutation at a single genetic locus. Although it may be single base pair (nucleotide) site [12, 29, 32], a genetic locus can span many nucleotide sites [32].

On the other hand, structural variations involve the alteration of large DNA segments [30]. Examples of structural variations include insertion-deletion (indels), inversions, transpositions, translocations and copy number variants (CNVs) [30]. CNVs are alterations in the copy counts of specific DNA regions as a result of either deletions or duplications to reduce or increase the copy counts [30]. Contrastingly, tandem repeats include variable number tandem repeats, such as mini-satellites and short tandem repeats (STRs) such as micro-satellites. From their name, variable number tandem repeats occur when a DNA pattern shows up again after a variable length of base pairs, normally between 10 to 100 base pairs (bps), while STRs occur when a DNA pattern recurs after a variable length of 1 to 6 base pairs [32, 38]. Since they tend to be informative in differentiating paternal alleles due to high heterozygosity, tandem

repeats are useful in forensic and clinical studies [30, 38].

Unlike all the other forms of genetic polymorphisms that involve more than one variant, a SNP is a genomic position where individuals have two or more alleles at more than 1% of the population [12, 39, 40]. An allele is a variant form of a gene [29, 32]. In a population, alleles are either more or less frequent. A more frequent (common) allele is known as the major allele, while a less common ($< 5\%$) is the minor allele. Therefore, the minor allele frequency (MAF) is the proportion of the less frequent allele [12]. A SNP can have two to four possible alleles [32]. When two alleles are possible, SNPs are said to be bi-allelic, and when more than two alleles are possible, they are multi-allelic. SNPs are the most widely used form of variation [12, 30, 32, 41] and have been successfully used in identifying disease risk genes. This is because SNPs occur in large numbers, are easy to genotype, and may be genetic determinants which characterise populations in cases when they are unique to a population [12, 30, 41].

1.2.2 Major factors contributing to human variations

Observed human DNA sequence differences result from the effects of the following forces: mutation, recombination, non-random mating, genetic drift, natural selection and gene flow or population admixture [9]. Among these, mutation, recombination and gene flow introduce variation in a population while, non-random mating and genetic drift, in most cases, tend to reduce variation within a population, but increase between population variations [32].

If genetic variations occur in less than 1% of the population, they are rare variants or mutations [39, 40]. Mutations are changes in nucleotides that are heritable [32, 42]. They either change a single letter (point mutation) where a base pair is substituted or deleted, or change multiple letters of the DNA sequences [32]. Mutations are caused by errors in DNA replication, radiation, mutagenic toxins, structural methylation, or viral activity [12, 43]. In human populations, mutations introduce variations through the creation of new variants [32]. If the introduced variant/s is/are in the coding region, they yield to changes in an amino acid and such a mutation is non-synonymous or missense. Contrariwise, a synonymous mutation does not change the amino acid [32].

Generally, recombination breaks down chromosomal segments that are usually inherited together during the meiosis process. Thus, low recombinant areas tend to have many alleles inherited together (high genetic linkage), while high recombinant areas (hot spots) have low genetic linkage [44]. Similarly to mutation, recombination introduces variation within popula-

tions. However, this is not in the form of new variants rather, variation is introduced as new haplotypes[5] [32], which perhaps may change protein functions. In short, recombination shapes the linkage disequilibrium patterns [45] and facilitates the natural selection processes [44, 46]. Linkage disequilibrium (LD) is when the relationship of alleles between two or more loci is predictable (see **Section 1.2.3.3**, for details), and natural selection is the process by which individuals with heritable advantageous traits adapt to their environment and produce more offspring [32, 47]. Most interestingly, natural selection acts on existing variations that are either introduced by recombination or mutation [46], we provide more details on natural selection in **Section 1.4**.

Population individuals' mating patterns largely contribute to human variations. Often, partners mate according to specific values; this is called non-random mating [48], and at times they mate in a random manner. Non-random mating is also called preferential mating, which can be endogamy (inbreeding) or exogamy (outbreeding) [32]. The former occurs when individuals breed within their subgroup, in the case of different population groups, and the latter occurs when individuals breed according to the specific phenotypes [32]. In the case of specific phenotypes, individuals either breed with those with dissimilar phenotypes (negative-assortative/disassortative preferences) or those with similar phenotypes (assortative preferences) [32, 49, 50]. Also, individuals might breed with respect to their location, that is, isolation by distance [32, 51]. Inbreeding and assortative mating increases homozygosity, reducing within population genetic variation [32, 49]. Given inbreeding, homozygosity is increased via the inbreeding coefficient [32]. One possible explanation is, due to being close relatives, mating individuals have at least one common ancestor, as such, the probability that an allele carried by each of them originates from the same DNA molecule is non-zero [32]. It should be noted that the increase in homozygosity due to inbreeding only changes the genotype but not the allele frequencies. In contrast, disassortative mating tends to maintain genetic variation in a population by increasing heterozygosity [49].

Similarly to inbreeding, genetic drift reduces variation within a population and increases variation amongst different populations [32]. Contrarily to inbreeding, genetic drift involves random variations of allele frequencies in different generations [32]. In other words, drift determines the possibility of a new mutation staying in the population for selection to act on it or be removed completely [32, 52]. It occurs in two forms, a population bottleneck or founder effects [32]. A population bottleneck may result from cataclysmic natural events such as earthquakes and famines, migration or failure to adjust to the environment [32, 53]. Although bottlenecks may

[5] A haplotype is a sequence made from combining individual alleles at distinct loci given an individual genome [32, 33].

reduce the variations within a population, they may be advantageous when inbreeding removes deleterious alleles from the population, often called purifying selection [32]. From the modern human origins, one can note that different populations are a product of migration. As such, another case of genetic drift occurs when a small number of population individuals migrate to a new area (founder effects) [32]. It is important to mention that the immigrants (individuals that migrated) may interbreed with the natives in the new area. If the populations involved are genetically distinct, the interbreeding process is called population admixture. This is another factor which contributes to human variations and, we provide more details on gene flow or population admixture in the next section.

1.2.2.1 Population Admixture

The admixture process changes the allele frequency of the recipient population, specifically when the donors (immigrants) and recipients have different gene frequencies [32]. As a result, by introducing new alleles, migration or gene flow increases genetic variation within a population and reduces variation between populations [32]. Admixture can also be considered as a form of preferential (non-random) mating or exogamy. This thesis focuses on admixed populations. Consider two ethnically distinct populations, denoted P_1' and P_2', in different geographical regions, R_1 and R_2, respectively. Let a proportion of P_1', denoted by P_1, migrate from R_1 to R_2 and interbreed with a proportion of P_2', denoted by P_2, to form a third population O. P_1 and P_2 are known as ancestral or parental populations of the admixed population O. Thus, population admixture is the interbreeding of two or more genetically distinct populations for some number of generations.

Now, assume two opposite sex individuals originate from two populations, P_1 and P_2. In the first generation, their offspring inherits one chromosome from every parent. In the second generation, the genetic material from the first generation is further broken by recombination yielding mosaic segments. **Figure 1.2** illustrates the population admixture process for two previously isolated populations, in this case P_1 and P_2. From the graph, it can be noted that the length of ancestral segments of admixed population individuals decrease with an increase in time since the admixture process, which ranges between one and G generations. A generation is approximately 20-30 years which is when the current offspring replaces the immediate previous offspring [9, 32, 54, 55]. The admixture process is either formed by a single or multiple of waves. When all populations contribute in the first generation and no further gene flow is received from the ancestral populations thereafter, it is called the hybrid isolation (HI) or single-point/wave admixture model. While the admixture process that is formed by multiple

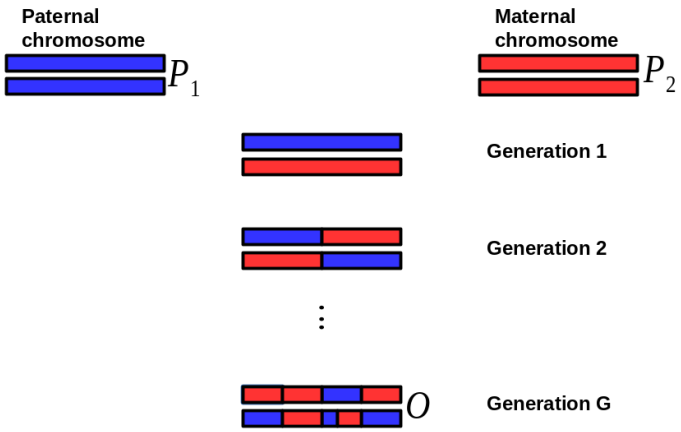


Figure 1.2: Illustrating an admixed individual genome O, formed by the interbreeding of two previously isolated populations P_1 and P_2, G generations ago.

waves is called the continuous gene flow (CGF) or gradual admixture (GA) [6, 7, 9]. Here, parental populations are expected to continue mating with admixed individuals [6, 7, 9].

Previous studies have highlighted that most modern humans are admixed [56, 57]. **Figure 1.3**, represents a partial worldwide map showing population migration patterns and admixed population distribution [2] supporting the belief that most modern humans are admixed. These

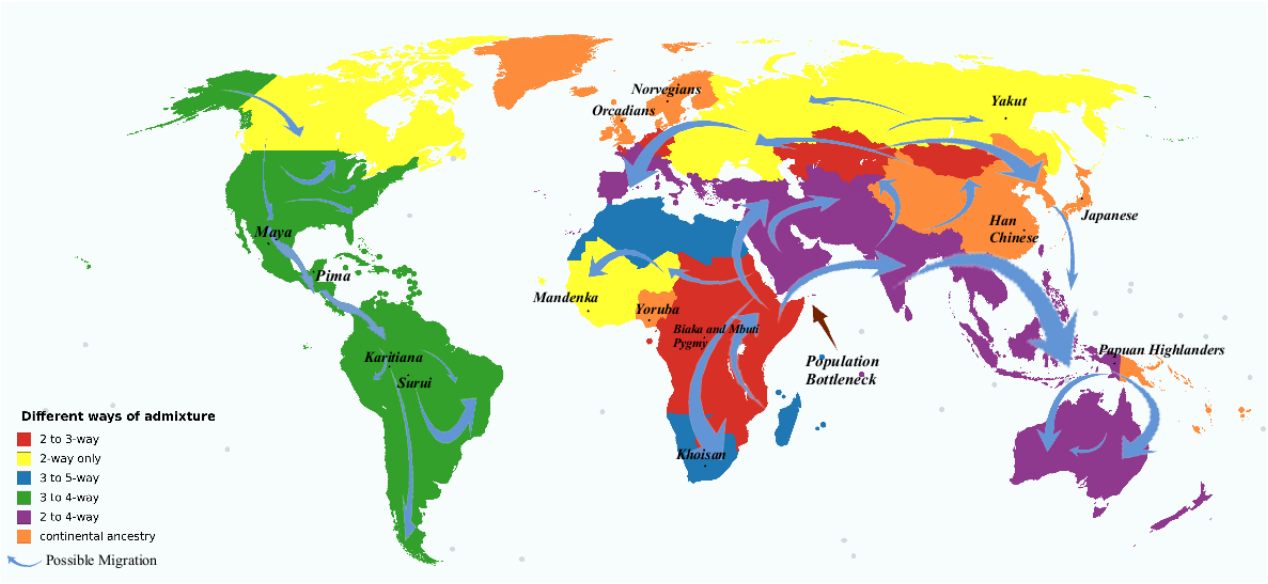


Figure 1.3: A partial worldwide admixture painting map exhibiting migration patterns that took place between and within continents, adapted from Mazandu et al. [2]. Such patterns yielded to two-, three-, four- and five-way admixed populations and were identified from population structure research articles published between 2008 and 2018.

migration patterns could have been a result of slave trade, colonialism, change in agricul-

tural interests and recently, economical and political instability [12, 58]. Examples of admixed populations include: the African and Latin Americans, the Uyghurs and the South Africans of mixed ancestry often called the South African Coloureds (SAC) [32, 59]. The African Americans are two-way admixed with approximately 20% European and 80% African [60, 61] ancestry proportions. The Uyghurs are also a two-way admixed population consisting of 60% European and 40% East Asian ancestries [62]. On the other hand, Latin Americans are three-way admixtures, and have varied ancestry proportions. In particular, Mexicans and Puerto Ricans contain $34\% - 64\%$ and $12 - 15\%$ of the Native American ancestry, respectively [63]. The South Africans of mixed ancestry are five-way admixed individuals whose ancestors are: the European (about 16%), the isiXhosa (about 33%), the Gujarati Indian (about 13%), the Khoisan (about 31%) and East Asian (about 7%) populations [64, 65]. Populations that are produced when three or more ancestral populations interbreed are called multi-way admixed populations. This is the case of the Latin Americans and the South Africans of mixed ancestry. A multi-way admixture is complex when it is a mixture of closely and distantly related ancestral populations, e.g., the South Africans of mixed ancestry [32, 66, 67].

1.2.3 Basic approaches to capture human variations

1.2.3.1 Allele frequency

Given a particular variant form of a gene (allele) [12], we can determine the frequency of its occurrence in a population and this is called the allele frequency [32]. Generally, considering haploid organisms, the frequency of an allele is the fraction of individuals with that allele out of the total population individuals. Given the gene frequencies, allele frequencies are estimated by adding half the frequency of heterozygotes to the frequency of homozygotes [32]. Consider a diploid organism, where the two possible alleles are A and G, let the frequency/proportion of allele G be p. Now, p is given by the proportion of homozygote individuals (f_{GG}) added to half the proportion of heterozygote individuals (f_{AG}). That is,

$$p = f_{GG} + 1/2 f_{AG}$$

Yet, the frequency of A (the alternative allele), is denoted by q, given by

$$\begin{aligned} q &= f_{AA} + 1/2 f_{AG} \\ &= 1 - p. \end{aligned}$$

In the case of a multi-allelic SNP, the frequency of one of the alleles, say Z, is

$$p = f_{ZZ} + \frac{1}{2}\sum_{\substack{u=1 \\ u \neq Z}}^{z} f_{Zu} \tag{1.2.1}$$

where $u = 1, \ldots, z$ are other allelic forms of a gene in the population where $u \in \{A, C, G, T\}$ and $u \neq Z$. Accruing to drift, natural selection, mutation and gene flow, the frequency of a particular allele may vary in different populations [32, 68]. However, it was shown that mutation alone may not yield to rapid changes in allele frequency [32].

As highlighted above, given the allele frequencies, one can determine genotype frequencies. In population genetics, this is made possible by the Hardy Weinberg equilibrium law. This law states that, given an infinitely large population where individuals mate randomly, male and female allele frequencies are equal, and, in the absence of natural selection, mutation and gene flow then the frequency of alleles remains constant over a period of time [12, 32].

1.2.3.2 Genetic distance

Different statistics are used to quantify genetic variation within or between populations. Commonly used are three F statistics, which are derived from the inbreeding coefficient, F. These are related as follows

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST}). \tag{1.2.2}$$

Thus,

$$F_{ST} = 1 - \frac{(1 - F_{IT})}{(1 - F_{IS})};$$

so that given two populations, F_{ST} contrasts the heterozygosity of the sub-population to total population (or is a measure of differentiation between sub-populations), F_{IS} contrasts the heterozygosity of an individual to a sub-population (measures the deviation of alleles from local panmixia) and F_{IT} contrasts the heterozygosity of an individual to the total population [8, 69, 70]. F_{ST} in Equation (1.2.2) can also be regarded as a measure of gene frequencies [71, 72],

such that

$$\begin{aligned} F_{ST} &= \frac{\sigma_{\bar{p}}^2}{\bar{p}(1-\bar{p})} \\ &= \frac{\sum_k (p_k-\bar{p})^2/(K-1)}{\bar{p}(1-\bar{p})} \end{aligned} \tag{1.2.3}$$

where p_k is the allele frequency of population k, c_k is size of population k, $\bar{p}$ is the mean allele frequency, $\bar{p}(1-\bar{p})$ is the variance between allele frequencies and $\sigma_{\bar{p}}^2$ describes how allele frequency varies among populations [73].

1.2.3.3 Linkage disequilibrium (LD)

As mentioned earlier, LD is the predictable relationship between alleles at two or more loci [51]. It occurs when the frequency of a particular genotype differs from that of the product of individual allele frequencies that make up the genotype [51]. A genotype of an individual at a genomic position (t) is an unordered pair of alleles. For a diploid individual ι, the genotype at t, $G_{\iota t} \in \{0,1,2\}$ corresponds to the number of minor (or non-reference) alleles [74]. Although LD can be evaluated on more than two loci, the most common LD measures evaluate LD on two loci. Now, assume two bi-allelic SNPs (loci): 1 and 2 with alleles C/c and A/a, respectively. Regardless of the order of appearance, there are four possible allele combinations at these SNPs, CA, Ca, cA and ca. We denote the allele frequencies of C and A by f_C and f_A, respectively, and, the haplotype frequency of CA by f_{CA}, therefore, the linkage disequilibrium between SNPs 1 and 2 is

$$\begin{aligned} D_{CA} &= D \\ &= f_{CA} - f_C f_A \end{aligned} \tag{1.2.4}$$

However, Lewontin [75] normalised Equation (1.2.4), defining LD as

$$D' = \begin{cases} \frac{D}{\min\{f_C(1-f_A), f_A(1-f_C)\}} & \text{if } D < 0 \\ \frac{D}{\min\{f_C f_A, (1-f_A)(1-f_C)\}} & \text{if } D > 0 \end{cases} \tag{1.2.5}$$

Since D' is affected by population size, another measure, the correlation coefficient (r^2) is more preferable when measuring LD between two loci [76, 77]. It is defined by

$$r^2 = \frac{D^2}{f_C(1-f_C)f_A(1-f_A)} \tag{1.2.6}$$

Equations (1.2.4), (1.2.5) and (1.2.6) are only applicable when there are two alleles at a SNP.

LD sources are divided into three: mixture, admixture and background LD [78]. Mixture LD emanates from the differences in ancestry between individuals [4, 78]. It can occur even between unlinked SNPs. On the other hand, admixture LD occurs during the admixture process and causes ancestry at nearby SNPs to be inherited together in the local chromosomes [4]. Admixture LD forms large haplotype blocks, and decays slowly but, is weak at short distances. On the contrary, background LD results from population history, such as genetic drift and population bottlenecks. It occurs within ancestral populations and decays rapidly [4]. Due to its strength at short distances it yields to short haplotype blocks [4, 78]. Generally, LD provides information on population bottlenecks, genetic drift, natural selection, mutation rate and genetic recombination rate. Therefore, it is useful for the understanding of human history and disease gene mapping. It is either evaluated at particular genomic regions or for the whole genome [79]. Whole genome LD provides insights in understanding breeding systems, genetic drift, population admixture and population bottlenecks [79]. On the other hand, genomic region LD highlights the effects of gene frequency factors, such as natural selection, recombination and mutation.

We note that LD is not precisely essential for genetic ancestry inference. Nonetheless, when SNPs in LD are discarded, local ancestry estimates accuracy deteriorates and when linked and unlinked SNPs are used without accounting for LD, it may result in noise and large systematic biases [4, 80]. This is due to the complexity of LD patterns in the source and admixed populations [4, 80]. Also, due to admixture LD, the distribution of global and local ancestry proportions in an admixed individual genome may reveal some genomic regions that may be of medical or evolutionary significance [4, 18]. As such, we provide more details on genetic ancestry in the next section.

1.3 Genetic ancestry

1.3.1 Overview

In order to understand admixed populations, ancestry inference is the most commonly used approach. It helps in determining: (1) The best proxy populations that interbred forming the admixed population (2) The number of source populations that interbred to form the admixed individuals (3) The average genome-wide ancestry proportions each source population contributed to each admixed individual (4) The ancestral origins of every chromosomal segment

at every SNP [81]. Points (1) to (3) are determined mostly at a global level and are referred to as global ancestry inference, while point (4) refers to local ancestry inference or local ancestry deconvolution [4, 81]. Ancestry proportions vary globally and locally within and between admixed populations [63]. Hence, the differences in observed individual traits. For instance, the prevalence of asthma is 14.6% among African Americans, 4.2% among Asian Americans, 4.8% among Mexicans and 19.6% among Puerto Ricans [63]. Ancestry estimates play important roles in the understanding of population demographics and history, discovering the genetic basis of diseases and in personalising medicines [82].

1.3.2 Global ancestry inference

Estimating the genome-wide average ancestry proportion contributed by each ancestral population dates back to the year 2000 [81, 83, 84]. Currently, existing global ancestry inference models are divided into two: (1) Models that depend on model parameters to estimate ancestry (model-based) and (2) Models that use algorithmic approaches (non-model-based).

Model-based approaches are more popular and in addition to estimating the global ancestry proportion, they also estimate ancestral population allele frequency [85, 86]. Model-based methods include STRUCTURE [87], FRAPPE [88] and ADMIXTURE [85]. STRUCTURE is one of the most popular software. Its first version assumes linkage equilibrium while its extension STRUCTURE V2 [78] accounts for linkage disequilibrium. Although ADMIXTURE is faster [81], in terms of estimate accuracy, STRUCTURE is as good as ADMIXTURE [85]. Similarly to STRUCTURE, FRAPPE [88] is a likelihood-based model, but, unlike ADMIXTURE and STRUCTURE, FRAPPE does not require the user to provide rules on how an optimum number of populations is chosen [85]. FRAPPE runs faster than STRUCTURE, but slower than ADMIXTURE [81, 86] which handles large datasets. Therefore, ADMIXTURE is faster than both FRAPPE and STRUCTURE [86]. Its accuracy is comparable to STRUCTURE and better than FRAPPE [86]. Another method is sparse non-negative matrix factorization (sNMF) [89] that is based on maximum likelihood and is used for global ancestry inference. sNMF is faster than ADMIXTURE [86]. We note that model-based methods (STRUCTURE, FRAPPE and ADMIXTURE) do not model LD. This could be because the impact of unmodelled LD is perhaps greater in local ancestry than it is in global ancestry estimates.

Conversely, non-model-based approaches are most popular for grouping population individuals according to their subgroups and visualising them [81]. Unlike model-based, non-model-based methods are difficult to interpret [86] and therefore less popular [85]. They include the principal component analysis (PCA)-based methods such as EIGENSTRAT [90] and smartpca. It

is important to note that traditional PCA-based models might inaccurately assign individuals to their corresponding sub-populations, especially when sub-populations are closely related and when one of the sub-populations is genetically distinct [81]. However, recent improvements of these nonparametric models, such as iterative pruning principal component analysis (ipPCA) [91] may overcome this individual sub-population map accuracy. We describe the general idea behind the PCA-based approaches in **Section 2.3.3**. Since local ancestry inference investigates admixed populations on a fine scale [86], the next section introduces local ancestry inference.

1.3.3 Local ancestry inference

Local ancestry, unlike global ancestry inference, yields finer details on the population history and demographics by estimating the number of chromosomal segments inherited from a particular population at every position along an admixed genome [81, 82, 92]. Several models have been proposed to deconvolve local ancestry (over 20 models). These include but are not limited to ANCESTRYMAP [93], ADMIXMAP [94], HAPAA [95], LAMP [96], HAPMIX [92], LAMP-LD [80], ELAI [97], LOTER [56], etc. Refer to **Chapter 2** for more details on these models, and particularly **Figure 2.1** for a snapshot providing a global summary of these models.

1.3.4 Current issues in genetic ancestry

As aforementioned, individuals vary within or between populations. These variations are useful in understanding population history through investigating the structure and diversity of populations and in medical studies. Several datasets exist on human variations, these include (1) The Human Genome Diversity Project (HGDP), which avails a set of over 1000 DNA samples [98] from 53 indigenous populations across the world; (2) The Human Genome Project (HGP), which maps and identifies human genes [99, 100]; (3) The International Haplotype Map (HapMap), which provides a sequence of 269 individuals from four populations in addition to the haplotype map that describes the common human variation patterns [101]. More recently, the 1000 Genomes Project (1000 GP) provided information on 2504 individuals from twenty six (26) nations [39, 102]. These, as well as similar projects are aimed at facilitating the understanding of human diseases, evolution, population history and demographics. However, despite the availability of datasets which can capture variations in populations, and the research that have been conducted to improve the accuracy of estimates, previous studies show that local ancestry estimates are inaccurate [4, 5, 66, 103, 104]. Hence, the need to further

investigate the existing local ancestry inference models.

1.4 Natural selection

Natural selection is a well explored relationship between fitness and other organism attributes within a population in a specific environment [47]. It usually targets particular regions and is not a random event. Selected loci or regions are important, especially if the locus (region) represents or is linked to a gene. Natural selection acts on both non-synonymous and synonymous SNPs. Although they do not alter amino acids, synonymous SNPs influence mRNA splicing, stability, structure and protein folding [105], which may significantly affect protein functions, change the response of cells to therapeutic targets or even provide details on why response to drug treatment differs among individuals [106]. As such, identifying selected regions facilitate the understanding of population history, human evolution processes and identifying biologically important genes, such as genes with protective effect against diseases [107], climate (or altitude adaptations) [108], cultural differences [108, 109, 110], and genes that may provide insights into some traits that could be useful in animal husbandry and crop farming [108].

Within a population, natural selection either reduces (the case of directional or purifying selection) or maintains (the case of balancing selection) variation [32]. When a more fit variant increases in frequency due to being favoured for, it is called positive/directional selection [111], and when a less fit variant is removed from the population, it is negative or purifying selection. We note that since they are more deleterious than they are advantageous, most random mutations lead to negative selection [110]. Now, continuously removing the deleterious mutations from the population yields to background selection (a special case of negative selection) [110]. Regardless of the resulting selection type (background, negative or positive selection), when the frequency of a given variant reaches 100%, it is referred to as fixation [111].

Natural selection is relevant in both medical and population genetics studies. As such, several studies have been conducted to identify selection targets in different species or populations [112, 113]. Since the natural selection process may yield to at least one of the following signals: loss of genetic diversity, skewed allele frequency, unexpected substitution ratio, extended haplotype homozygosity and elevated linkage disequilibrium [110], methods that identify selected regions aim to identify at least one of these signals. Existing methods are grouped into two: micro- and macro-evolution methods [110]. Macro-evolution methods detect major evolutionary changes which could happen between species [110]. These methods mostly use exome sequence data. For example, the McDonald-Kreitman test (MKT)

and the Hudson-Kreitman-Aguade (HKA) test [110]. Contrarily, micro-evolution methods detect changes within species using whole sequence data. They include linkage disequilibrium-based, genetic differentiation-based and frequency-based tests [110]. Since *de novo* mutations tend to create a surplus of rare alleles, frequency-based tests (Tajima D [114], and Fay and Wu [115] tests) assume that a beneficial mutation increases in frequency with the nearby derived alleles [110]. Different from frequency-based tests, linkage disequilibrium-based tests assume extended haplotype homozygosity across a haplotype containing the selected allele. For example, the integrated haplotype score (iHS), and the Rsb [110] tests. Finally, genetic differentiation-based tests assume an increased genetic difference between populations as a result of population-specific selection, yielding differences in allele frequencies. Of the several genetic differentiation-based models, F_{ST} which measures the population variances within or between populations (**Section 1.2.3.2**), is the most popular [110].

In admixed populations, natural selection occurs either pre-admixture or post-admixture. Pre-admixture selection occurs in the previously isolated populations after migration but before the admixture process. As a result, it rarely impacts admixed genomes [112]. Contrastingly, post-admixture selection occurs during the admixture process and is currently not well understood [116]. Apart from the macro- and micro-selection detection methods, in admixed individuals, the local ancestry estimates may be used to identify regions under selection [113, 117]. Here, an admixed individual genome is scanned to identify regions with excessive or reduced ancestry components, often termed the deviations in local ancestry [112, 118]. Unlike post-admixture selection or post-admixture genetic drift that target particular regions, genetic drift in ancestral populations, systematic biases (genotyping errors) and using reference ancestral populations that are not genetically similar to the actual ancestral populations affects the whole genome equally [80, 112, 116, 119, 120]. To illustrate the detection of selection (pre- and post-admixture) regions based on local ancestry estimates, Jin et al. [112] compared the distribution of ancestry in the reconstructed ancestral genome of an admixed individual to the source populations and admixed genomes. We demonstrate pre- and post-admixture selection for a population that mimics the South Africans of mixed ancestry in **Figure 1.4**.

1.5 Importance and applications of local ancestry inference

Over the past twenty years, various applications of local ancestry inference estimates have been suggested. These include biomedical applications for instance, identifying regions under

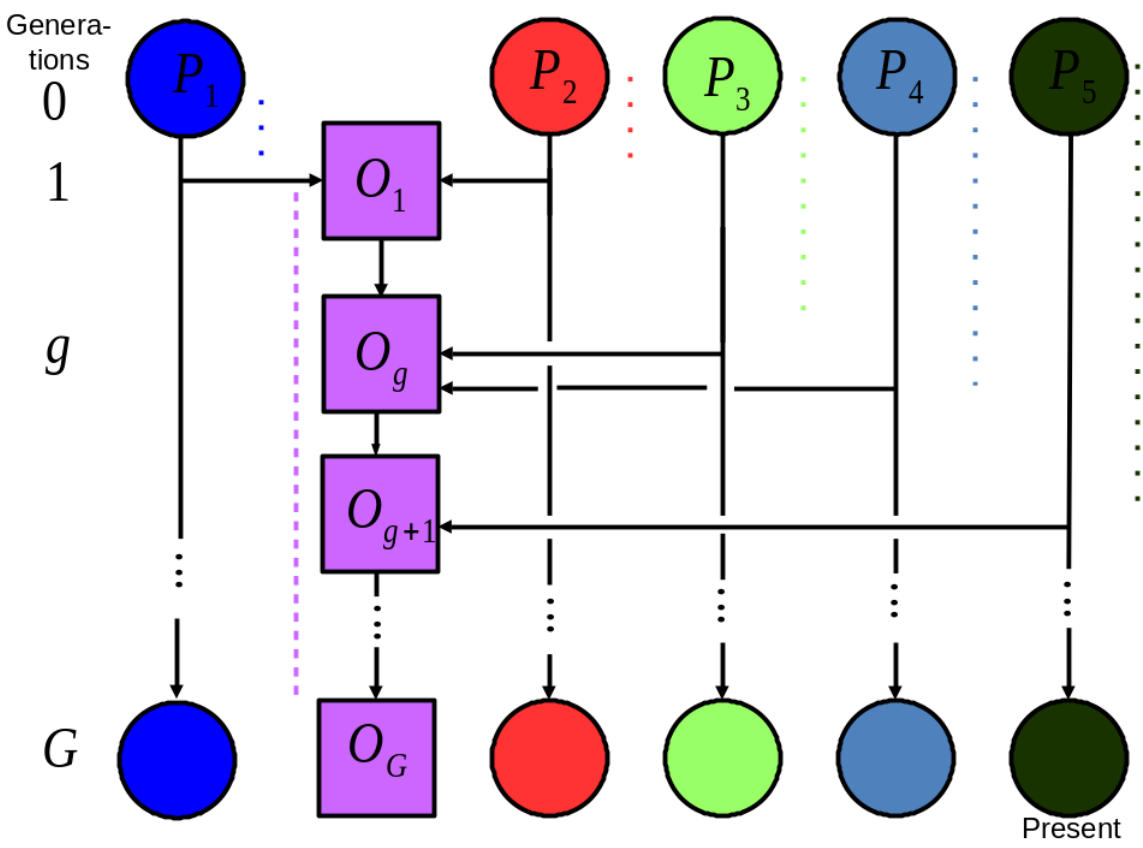


Figure 1.4: An illustration of pre- and post-admixture selection in a five-way admixed population. Arrows represent gene flow, dotted lines represent pre-admixture selection, where the colour corresponds to pre-admixture selection in an ancestral population P_k, $1 \leq k \leq 5$, highlighted in that colour, and a dashed line in the admixed population colour represents post-admixture selection. The admixed population at generation g is denoted by O_z, $0 \leq z \leq G$.

selection [121], determining the alleles responsible for the differences observed in disease risk and response to drugs among different population groups [4], tracking down sequences that might be absent in reference genomes [122], and providing some clues into the genetic history and demographics of populations [64, 123].

In order to understand population history and demographics in admixed individuals, the information on break points, admixture LD and the length of ancestral tracts is used [86]. Since it is inversely related to the time since admixture, ancestral tracts length provide insights into the history of admixed populations; particularly the cultures, the shared genealogy, the population bottlenecks and the human languages. For example, an admixture of the Khoisan and the Bantu speakers, the Xhosa of South Africa [124] inherited the click sound from the Khoisan (KHS) population. An example of studies that use the local ancestry estimates to understand population history is that by Omberg et al. [125] which used estimates to un-

derstand the genome-wide admixture patterns of the Qatar (a peninsular Arab) population. This population was decomposed into three main components: Arab-Qataris, an admixture of the middle easterners and Bedouin or Hadar-Arab ancestry; the Persian-Qataris, an admixture of Europeans, Asians (that are not from the middle east) and sub Saharan-Africans; and finally the African-Qataris, an admixture of Africans, Persians and a few middle easterners [125]. On the other hand, the use of local ancestry estimates in admixture dating was demonstrated in Galaverni et al. [126] where PCADMIX, a PCA-based local ancestry inference model was used to date the admixture of Italian wolves. PCADMIX estimated an admixture time of 19 generations. According to the history of the Italian wolves, this estimate supports the hypothesis of re-expansion of the Italian wolf population as opposed to the population bottleneck hypothesis. Thus, in comparison to the estimates from existing dating admixture models including ALDER [127], PCADMIX performed better. And, based on this study, one can conclude that given admixtures that are formed through continuous gene flow (in this case re-expansion is an instance of continuous gene flow) PCADMIX may perform better than existing admixture dating models. The history of a population also includes knowing the place where an admixture event took place. As such, Xue et al. [128] used the RFMIX [129] model to show that Southern Europeans are the major contributors of the European component in the Ashkenazi Jews [128].

Another local ancestry estimate application is to detect selection target regions. As highlighted in **Section 1.4**, selection target regions may facilitate the understanding of human evolution, the identification of genes that are biologically important and those that may be useful in understanding some human traits [110]. For example, Jin et al. [112] identified six regions that harbour disease genes that are highly associated with two traits: prostate cancer and hypertension. In addition to shading insights into the understanding of these two conditions in the concerned populations, this study also facilitates the understanding of the evolution of the current populations. On the other hand, Deng et al. [113] provided insights into human adaptations, where two regions with excessive European ancestry and one with excessive African ancestry were identified.

Furthermore, local ancestry estimates are also critical for understanding the genetic basis of the disease [130]. Unlike with the effects of the surroundings (environment) and/or life-style, it is challenging to control the genetic basis of a disease. However, admixture analysis is used to understand complex traits through genome-wide association studies (GWAS) [131]. To uncover the genetic basis of the disease, GWAS scans the whole genome for SNPs of a given population individuals to identify genetic variations related to a particular disease [131]. Newly found associations are then used in detecting, treating and preventing diseases. GWAS has

successfully identified genetic variations associated with complex diseases[6] in homogeneous populations including the Europeans residing in a homogeneous environment [132]. However, chances are high that the information contained in admixed individuals might be directly related to the variations seen in phenotypes and drug responses [119]. As such, admixed populations are important in understanding diseases that are not well expressed in homogeneous populations, such as sleeping sickness [132, 133]. However, conducting GWAS without accounting for population stratification at global and/or local scales in disease-affected and unaffected admixed individuals may yield false positive results [134]. Hence, the need to correct for population structure in GWAS using global and local ancestry [135].

In addition, local ancestry estimates are relevant in admixture mapping. Admixture mapping is a gene mapping method that identifies genetic disease related loci in admixed populations [63, 136]. It scans the admixed genome for regions with excessive or reduced ancestry proportions [63, 93, 137]. This is because individuals differ in the global ancestry, the variations in ancestral history or the frequency of an allele at a locus. Now, given non-polymorphic loci in ancestral populations, admixed individuals tend to be more diverse. Since they are used as covariates in ancestry phenotype associations, local ancestry estimates are important in admixture mapping. The recent developments in identifying disease related loci (or variants) combine the genotype and local ancestry within the framework of single marker tests [138]. The combination is more powerful than the independent GWAS or admixture mapping [132, 138, 139, 140, 141].

Local ancestry estimates also play important roles in understanding how genes interact with the environment [4]. For example, in personalising medicines where possible interactions between genes and drugs (environment) is key [142, 143]. Since local ancestry estimates may be useful in detecting regions that are resistant or susceptible to disease, it provides insights into the diseases and drugs that are suitable for specific populations. This minimises the costs of the trial-and-error method of prescription, which in turn reduces health costs and assists individuals to manage their health [144, 145, 146]. As an illustration, Yang et al. [57] used local ancestry estimates to show how different ethnic groups associated with the relapse of acute lymphoblastic leukaemia.

Finally, local ancestry is vital in localising missing sequences in the human genome. In fact, the human genome sequences and physical maps are important in interpreting data in human genetic variation and genome biology [122]. However, this is impacted by the presence of some inaccessible regions where sequences are missing in the genome [122, 147]. To localise these

[6]Complex diseases occur when multiple genes interacts with an individual's living standards and their environment.

sequences in human genomes, ancestry linkage disequilibrium that depends on local ancestry estimates across the genome, is used. For example, Genovese et al. [122] used local ancestry estimates of 242 Latino individuals to localise 569 scaffolds with 20 Mbps missing sequences in the human reference genome.

1.6 Thesis rationale and motivation

Although mutation and recombination play important roles in DNA sequence differences among modern humans, human migrations have a significant role in shaping genetic diversities in human populations yielding population admixtures [33]. As mentioned in **Section 1.1**, modern humans originated from Africa almost 200 kya and later spread world-wide through migration [148]. Nevertheless, modern humans differ in physical appearance, disease susceptibility and disease treatment response. Inheritance of the traits from their genetic ancestors, together with the environment may contribute to the adaptation and modification of the inherited traits. Therefore, analysing the ancestry of every chromosomal segment of individuals formed through the interbreeding of two or more previously isolated populations (local ancestry) may provide insights into population history and demographics by studying genetics processes including, natural selection [118], recombination [149] and migration [60]. Admixed individual chromosomes are a mosaic of segments originating from each of the source populations. Local ancestry is essential in identifying new disease genes [94, 103, 150, 151] for personal genomics [57], specifically in developing personalised medicines. However, the usefulness of local ancestry estimates entirely depends on their accuracy.

Even though many local ancestry inference models exist [80, 92, 95, 96, 97, 129, 152, 153, 154, 155, 156, 157, 158, 159], a major challenge is the performance of these models in complex multi-way admixtures, including the Latin Americans and South Africans of mixed ancestry [64, 66, 119] that have both continental and sub-continental ancestries. Several studies have shown that deviations in local ancestry exist in such populations. This could be due to the complexity in the number of populations that make up these populations. For instance, South Africans of mixed ancestry are a mixture of the Bantu-speaking Africans, Khoisans, Europeans, East Asians and Indians [64]. A portion of the existing models are efficient in two-way admixtures, e.g., HAPMIX [92], while, others are more accurate in continental ancestry admixtures, the case of LAMP [96].

Moreover, most existing models have been evaluated on three-way admixtures, Latin Americans in particular, and rarely on populations that mimic the South Africans of greater multi-

way mixed ancestry. Examples of such models include LAMP-LD [80], MULTIMIX [152] and ELAI [97]. Recently, to characterise local ancestry accuracy in the real data of Latin Americans, Pasaniuc et al. [119] used 489 nuclear families from the Mainland USA, Puerto Rico and Mexico and 3204 unrelated Latin Americans from the Multi-ethnic Cohort study. Unlike in analyses that are based on simulated data where errors may be associated with the simulations, this study is based on real data, hence, the reported errors are due to the Mendelian inheritance of local ancestry in families [119].

Although existing models may attain high accuracy on average, they tend to falsely deviate in average local ancestry at some regions [66]. Therefore, this study investigates local ancestry inference models to provide insights into the implementation of practical tools or models to improve the accuracy of local ancestry estimates.

1.6.1 Significance of the study

Through its important role in understanding the genetic basis of diseases, local ancestry inference associates with the burden (direct and indirect costs) of some of the non-communicable diseases (NCDs). These are also known as chronic diseases and cause more than 70% of the total deaths in the world [3]. Their burden is expected to keep increasing in both developed and developing countries [160]. Currently, the effects of NCDs in Sub-Saharan Africa are beyond their global average [161]. NCDs include complex diseases such as cancer, diabetes and cardiovascular diseases. **Figure 1.5** shows the ten leading causes of death globally in 2019 [3]. Seven of these are NCDs, that is, Alzheimer disease, Chronic obstructive pulmonary diseases, Diabetes mellitus, Ischaemic heart disease, Kidney diseases, Stroke, and Trachea, bronchus and lung cancers. These contribute 44% to the total death causes globally (55 415 773 deaths) [3]. Apart from causing death, NCDs may affect an individual's quality of life. For example, it can lead to role reversal between parents and children leading to school drop-outs, reduction in productivity and sometimes depression. Since local ancestry inference also plays important roles in understanding complex diseases and personalising medicines, it is therefore a first step in understanding the location and origin of potential genetic links to the NCDs. Hence contributing to the achievement of the health-related sustainable development goals (SDGs).

Although different directions have been explored when deconvolving the ancestry of every admixed segment including the modelling of linkage disequilibrium within or between populations, different studies have shown that existing models are inaccurate in some regions on average [66, 103, 119]. However, these studies are all based on different datasets, performance

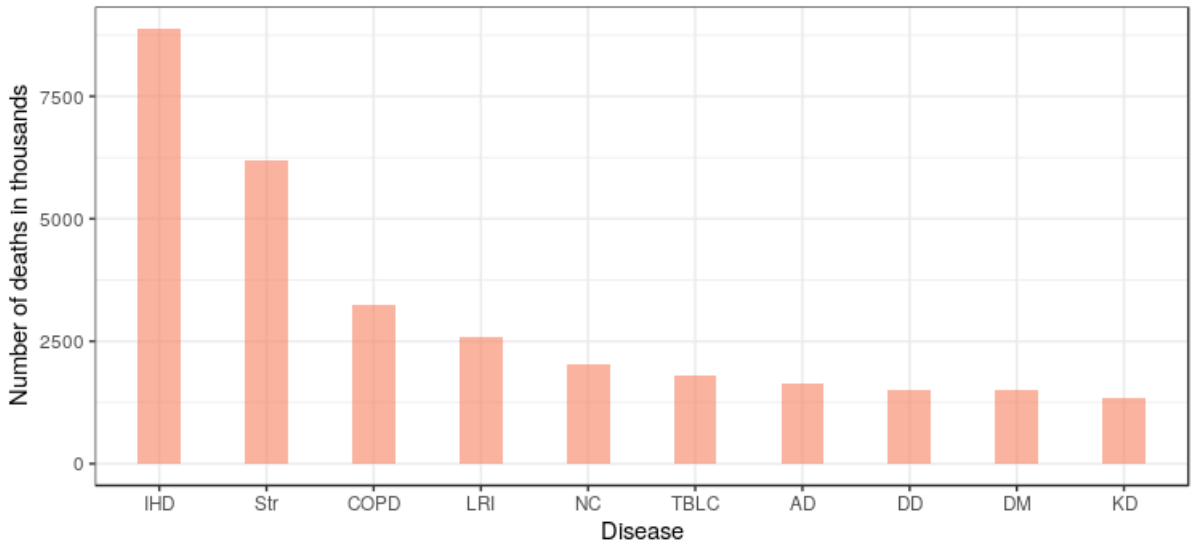


Figure 1.5: Illustrating the burden of NCDs: The ten leading underlying causes of death globally in 2019. AD–Alzheimer disease, COPD–Chronic obstructive pulmonary diseases, DD–Diarrhoeal diseases, DM–Diabetes mellitus, IHD–Ischaemic heart disease, KD–Kidney diseases, LRI–Lower respiratory infections, NC–Neonatal conditions, Str–Stroke, TBLC–Trachea, bronchus, lung cancers. Source: Organization [3].

measures and assess models in different frameworks. To address this, a unified framework that uses standard input files to manipulate tool specific inputs, deconvolve ancestry and standardise outputs to user-specified formats has been designed and developed. This framework facilitates proper admixture and/or association mapping to identify genomic regions underlying ethnic differences in disease risk. More so, it enables the implementation of the most appropriate local ancestry inference methods for the dataset under study. Further, using the developed framework, a comprehensive survey of existing models was performed to orient software developers and modellers on the future trends on the local ancestry inference problem. Finally, although this was less significant in our assessment results, previous studies have highlighted that the deviations in local ancestry that occur at particular regions may be due to natural selection. As a result, we recommend that further studies consider selection in local ancestry inference.

1.6.2 Purpose of the study

The overall aim of this study is to investigate models that estimate ancestry along the genome of admixed individuals, so as to identify and tackle critical issues of such models. This, we achieve through the following objectives:

1. To develop a Python framework that integrates existing multi-way local ancestry inference models to facilitate the inference process and pave the way for model assessment and implementation.

2. To investigate whether the local ancestry deconvolution problem is solved or not using existing models.

3. To provide some model recommendations in order to address modelling challenges that exist in local ancestry inference.

1.6.3 Thesis organization

This section provides an overview and structure of each chapter. Generally, this study aims to investigate and improve on the models that infer local ancestry in individuals formed when three or more genetically distinct populations interbreed for some G generations (multi-way admixtures). The study is relevant for understanding population history, demographics, disease aetiology [4], and personal genomics (see **Section 1.5**).

Chapter 2 qualitatively assesses existing local ancestry inference models systematically, based on previous studies. This is achieved by first providing a theoretical overview of models, dividing them into two major categories depending on their ability or inability to account for linkage disequilibrium (LD). For each model, considering the chronological ordering, model strengths and weaknesses, and biological and/or statistical parameters are highlighted. Secondly, the mathematical or statistical approaches of these models are discussed in detail. Thirdly, challenges and opportunities in local ancestry deconvolution are highlighted and, finally, we give a summary of the chapter findings.

Chapter 3 introduces a unified framework that integrates eight existing state-of-the-art models. It manipulates input files, deconvolves local ancestry and standardises outputs to a format that can ease estimate applications. Firstly, the information on the framework implementation is given, and secondly, the inputs and the standard output formats are discussed in detail. Further, we illustrate how to run the framework, and, then discuss other related frameworks, and possible improvements to this framework. Finally, we conclude with the chapter summary.

Chapter 4 quantitatively evaluates existing models on different admixture scenarios within the same framework, based on the same performance measures. The chapter is structured as follows, we (1) Describe the pre-simulation procedures, that is, the ancestral populations used, the quality control procedures and the scenarios the chapter evaluates; (2) Describe

the simulation framework we use and how each of the admixture scenario is simulated; (3) Provide details of the evaluation process, highlighting some of the performance measures that have been previously used, describe the ones we use and how we determine them; (4) Discuss chapter findings and compare them to the qualitative evaluation in **Chapter 2** (highlighted in **Section 4.4**); and finally, (5) Provide the chapter summary.

Finally, **Chapter 5** concludes the thesis. We provide a general discussion, conclusions and finally provide some future recommendations. Since several studies have shown that existing models fail in particular regions, we suggest a new avenue of accounting for natural selection and population relationships when deconvolving ancestry in multi-way admixtures. The assumptions and mathematics of the model are presented along with a qualitative comparison with the existing methods.

1.6.4 Overview of scientific contributions

The review in **Chapter 2** is different from existing reviews that describe model relationships, similarities and differences in mathematical structure and the estimate applications. We believe that our comprehensive exploration of the different model assumptions and statistical properties, and the results from existing assessments provide a valuable resource to orientate users, modellers and software developers on the future trends in local ancestry inference, including providing useful information on the implementation of practical tools. Most of **Chapter 2** contents have been published in the peer reviewed paper by Geza et al. [4] and Mazandu et al. [2].

In **Section 2.4**, we noted that each model requires unique inputs to deconvolve local ancestry and produces distinct outputs. This may hinder the local ancestry inference process and its downstream applications. Therefore, **Chapter 3** focuses on tackling this issue and paves the way for model assessment, so that the best model can be selected for different dataset and ancestry estimate applications. We describe the implementation of a unified framework, FRANC [5], integrating existing models, facilitating the assessment of three or more models and allowing integration of new and future local ancestry inference models. Most of **Chapter 3** contents are reported in the peer reviewed paper by Geza et al. [5].

Existing models have been evaluated contextually and some by a single study, making it difficult to identify the most appropriate local ancestry inference methods for use and adaptation [4]. Until now, no study has evaluated four or more models on complex multi-way human admixtures [4]. Therefore, **Chapter 4** provides insights into the performance of each of the

eight state-of-the-art models given (1) Single- and multiple-wave admixtures; (2) Recent and ancient admixtures; (3) Disease-affected admixtures with some SNPs under selection; (4) An underestimate of admixture generations; (5) Unbalanced representation of reference ancestral panels; and (6) Real data on complex multi-way admixtures.

Finally, it is believed that capturing appropriate biological parameters of an admixed population may improve the accuracy of local ancestry estimates [80, 112, 116, 119, 120]. Since current models do not account for natural selection when deconvolving local ancestry in multi-way admixtures yet previous studies pointed out that deviations in local ancestry at particular regions may be due to selection, we discuss and provide recommendations on accounting for selection in local ancestry deconvolution in **Chapter 5**.

1.7 Notation, symbols and terminology

In this thesis, unless otherwise specified the notations and symbols are as defined in **Table 1.1**. Also, unless specified, in subsequent chapters, we

- Assume SNPs/markers are bi-allelic.
- Refer to traditional hidden Markov models as HMMs.
- Denote an indicator function by $I(b = b')$, defined as

$$I(b = b') = \begin{cases} 1 & \text{if } b = b' \\ 0 & \text{otherwise} \end{cases}, \tag{1.7.1}$$

  and $I(b \neq b') = 1 - I(b = b')$.

Table 1.1: Thesis notation, symbols and their definitions

Symbol/ Notation	Definition
G	a measure of time and $G \approx 25$ years,
K	number of unmixed previously isolated populations that mixed G generations ago to form an admixed population,
k	source population index where $k \in \{1,2,\ldots,K\}$,
n_k	sample size of population k
n	sample size of all the K populations, where $n = \sum_{k=1}^{K} n_k$,
N	sample size of the admixed population,
ι	individual index, where, $\iota \in \{1,2,\ldots,n_k\}$ for source population k and $\iota \in \{1,\ldots,N\}$ for admixed individuals,
T	number of polymorphic loci at which the source and admixed samples are typed at,
t	marker or locus index, where $t \in \{1,\ldots,T\}$,
$q_{\iota k}$	ancestry proportion contributed by population k to individual ι,
$\mathbf{q}_\iota$	vector of ancestry proportion for individual ι, i.e, $q_\iota = (q_{\iota 1}, q_{\iota 2}, \ldots, q_{\iota K})$, $\sum_{k=1}^{K} q_{\iota k} = 1$,
p_{tka}	frequency of allele $a = 0,1$ in source population k at t,
h	haplotype, and $h \in \{0,1\}$,
$Y_{\iota t}$	observed variable for individual ι at a SNP t,
$\mathbf{Y}_\iota$	observed sequence for individual ι for the T loci, such that $\mathbf{Y}_\iota = (Y_{\iota 1}, \ldots, Y_{\iota T})$,
$x_{\iota t}$	locus-specific ancestry for individual ι, where $x_{\iota t} \in \{1,\ldots,K\}$,
$\mathbf{x}_\iota$	hidden sequence that generated the observed sequence $\mathbf{Y}_\iota$ for individual ι, where $\mathbf{x}_\iota = (x_{\iota 1}, \ldots, x_{\iota T})$,
x_{-t}	all x_r except for $r = t$, i.e., $x_{-t} = (x_1, x_2, \ldots, x_{t-1}, x_{t+1}, x_{t+2}, \ldots, x_T)$,
$x_{1:T}$	is a sequence from x_1 up-to and including x_T, such that $x_{1:T} = (x_1, x_2, \ldots, x_T)$,
r_t	recombination rate between $t-1$ and t,
r	rate at which breakpoints occur,
d_t	physical distance between $t-1$ and t,
S	defines probabilities of moving from one state to another, denoted by a matrix S,
s_1	defines transition probabilities at a SNP 1,
E	emission probability model,
λ	hidden Markov model parameters such that $\lambda = \{s_1, S, E\}$,

Chapter 2

Dissecting local ancestry deconvolution in the human genome

2.1 Introduction

Recent advances in high-throughput technologies promoted the availableness of huge amounts of genomic datasets to the public [4]. To make well-informed decisions about the health and genetic make-up of populations, such datasets require the use of suitable techniques. As discussed in **Chapter 1**, the genetic variations observed in individual DNA sequences originate from population migration and genetic inheritance processes (including mutation, admixture and recombination) [4, 148]. Since individuals are characterised by migration, most modern humans are of mixed ancestry [56, 57]. The admixture process produces genetic recombination break points, mixed DNA segments, and yields to the formation of genetic variation [4]. This therefore facilitated the need to understand the dynamics related to the origins of such differences, and the way evolution took place and its repercussions in human health [4]. More so, exploring populations formed from the interbreeding of two or more genetically distinct populations may provide important clues in the genetic structure of diseases [4, 103, 132, 150]. Currently, a combination of advanced high-throughput sequencing, genotyping technologies, and appropriate computational and statistical techniques have played important roles in understanding the history of admixed populations [4]. This includes the implementation of several local ancestry deconvolution models. We note that, since 2003, over 20 models have been introduced [4], some of which are based on the same statistical techniques. Nonetheless, models have improved to account for ancient and recent admixtures, closely and distantly related ancestral populations, subcontinental and continental ancestry, and two- and multi-way

admixtures. However, the accuracy of existing models remains questionable, given that most model evaluation studies involve two- or three-way admixtures [56, 80, 156, 162] (**Table B.1**).

This chapter reviews existing multi-way local ancestry deconvolution models. Although some review studies exist, they mainly describe local ancestry inference-associated models, their relationship with one another and the different local ancestry estimates applications [81, 82]. Furthermore, existing reviews give the mathematical structure of some of the existing models, with an emphasis on model similarities and dissimilarities [4], for example Gompert and Buerkle [83]. Contrarily, this chapter dissects approaches for local ancestry deconvolution in admixed individual genomes, summarises the evolutionary factors existing models account for, explores the different statistical/mathematical properties [4], and the biological assumptions and assesses models based on existing evaluation studies. This will help researchers when selecting the most appropriate models given the specific admixed population and local ancestry estimates application [4]. Moreover, it may aid software developers to pinpoint existing challenges and advances in local ancestry inference. Thus, according information that maybe suitable for developing reliable integrative tools which take into account current medical and population genetics demands [4]. Further, it may be helpful in determining the future trends in local ancestry, that is, whether to develop new or extend existing models. In short, this chapter acts as a reference manual on local ancestry models [4] and highlights the biological practicality of existing models given the applications of the inferred estimates. As mentioned in **Chapter 1**, this chapter is based on Geza et al. [4].

2.2 Overview of local ancestry deconvolution

Genetic ancestry inference models have contributed to a better understanding of human variations and traits. As mentioned in **Chapter 1** genetic ancestry inference is either global or local. In 2001, the first and most popular, genetic ancestry inference model, STRUCTURE [87], which estimates ancestry at the global level, was proposed. Since STRUCTURE does not provide finer details about population history and cannot clarify the genetic basis of the diseases locally, local ancestry deconvolution emerged three years later [78]. To date, several models (over 20) and local ancestry estimates applications (see, **Section 1.5**) exist. Nevertheless, the choice of the statistical and/or mathematical models and the complexity of linkage disequilibrium (LD) patterns significantly affect local ancestry estimates [4]. Depending on their ability or inability to account for linkage disequilibrium, local ancestry deconvolution models are categorised into two types, LD-based or non-LD-based models [4]. **Figure 2.1** shows the evolution of the two categories with time and the mathematical and statistical advances.

LD-based models include STRUCTURE V2 [78], ANCESTRYMAP [93], ADMIXMAP [94], SABER [157], HAPAA [95], SWITCH [163], HAPMIX [92], GEDI-ADMIX [164], mSPECTRUM [165], CHROMOPAINTER [155], SUPPORTMIX [125], PCADMIX [166], LAMP-LD [80], MULTIMIX [152], SEQMIX [167], ALLOY [158], and ELAI [97]. On the other hand, non-LD-based models include LAMP [96], WINPOP [156], RFMIX [129], EILA [159] and LOTER [56]. We discuss the two local ancestry deconvolution model categories in detail in subsequent sections.

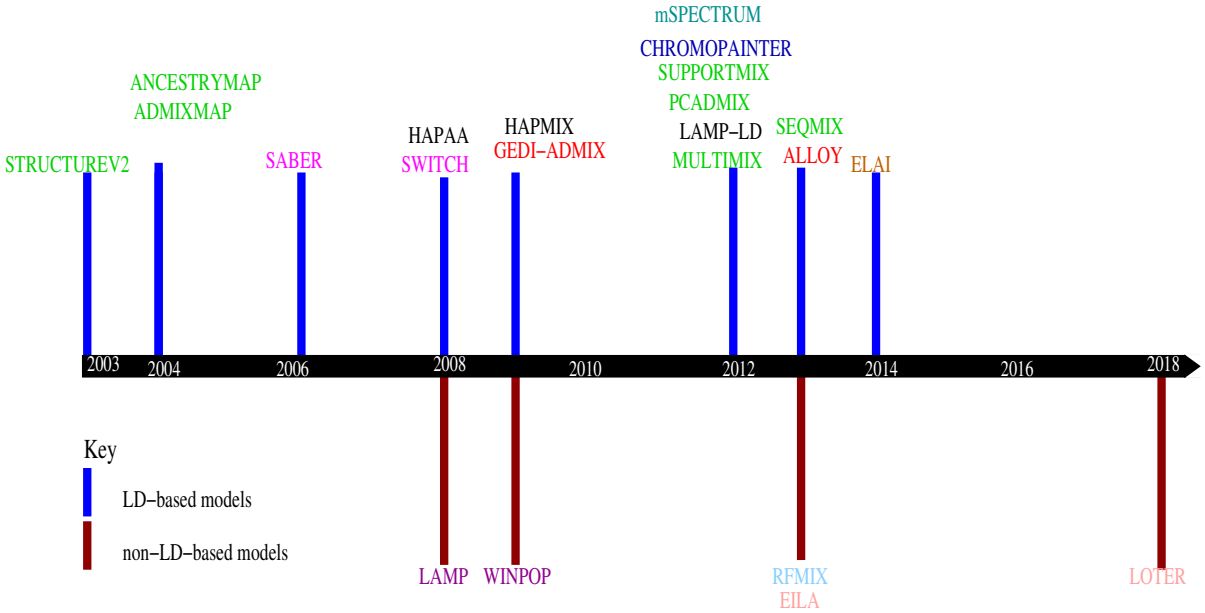


Figure 2.1: A pictorial representation of local ancestry inference models developed between 2003 and 2018: source Geza et al. [4]. The colours on model names indicate the mathematical or statistical approaches each model is based on. For example, STRUCTURE V2, ANCESTRYMAP, ADMIXMAP, MULTIMIX, PCADMIX, SEQMIX are based on traditional hidden Markov models, SABER and SWITCH are based on Markov hidden Markov model, HAPAA, HAPMIX and LAMP-LD are based on hierarchical hidden Markov model and LAMP and WINPOP are based on the "LAMP" framework.

2.2.1 LD-based models

LD-based models may account for either admixture linkage disequilibrium (LD) only or both admixture and background LD. Models that account for admixture LD only are known as admixture LD-based approaches, whereas those that account for the two are referred to as admixture-background LD-based approaches [4]. As mentioned in **Section 1.2.3.3**, generally, linkage disequilibrium is less important in the local ancestry inference process. Nonetheless, pruning loci or SNPs in LD may yield spurious estimates, whereas using all SNPs without

modelling admixture and background LD may yield noise and/or systematic biases [4, 80]. As such, modelling linkage disequilibrium explicitly may improve the accuracy of local ancestry estimates and, most importantly, increases power in association mapping [4, 80]. Now, to explicitly describe a phenomenon, assumptions should be clearly stated. As a result, LD-based models assume ancestry at consecutive SNPs or windows follow a Markov chain process [4, 103]. Thus, to infer local ancestry along an admixed individual genome, LD-based models use the traditional (or standard) hidden Markov models and their extensions [78, 157].

Unless specified, we refer to the "traditional hidden Markov models" as HMMs. These models assume recombination occurs at every generation such that the number of recombinations in a given genomic interval are Poisson distributed [4] with rate $r(r_t, G)$, where, r_t is the recombination rate, and G is the time since the admixture process started [4, 80, 83, 93, 167]. Adapted from Maples et al. [129], **Figure 2.2** summarises the connection between LD types and local ancestry deconvolution. Here, X_t and Y_t are as defined in **Section 1.7**. We note that models that account for mixture LD only, including the original STRUCTURE [87] are graphically represented by excluding all horizontal (dotted, dashed and solid) arrows, while admixture LD models (STRUCTURE V2, ANCESTRYMAP, ADMIXMAP, PCADMIX and SUPPORTMIX) are represented by excluding horizontal solid arrows between observed variables. Sohn et al. [165] highlighted that unlike mSPECTRUM, most admixture LD-based models do not account for mixture LD. Finally, admixture-background LD-based models are graphically represented by excluding all parents nodes and edges before the hidden ancestry.

Assuming individuals are haploid and independent, the ancestry at SNP t, $X_t \in \{1, \ldots, K\}$ is such that:

$$\begin{cases} P(X_t = k|\mathbf{q}) & = q_k & \text{for } t = 1, \\ P(X_t = k|X_{t-1} = k', \mathbf{q}, r(r_t, G)) & = I(k' = k)\exp(-d_t r(r_t, G)) + (1 - \exp(-d_t r(r_t, G))q_k & \text{for } 1 < t \leq T, \end{cases} \tag{2.2.1}$$

where $r(r_t, G)$ is the mean number of recombinations between SNP $t-1$ and t, q_k, $\mathbf{q}$, r_t and G are as defined in **Table 1.1**, while $I(k' = k)$ is as defined in Equation (1.7.1) and $(1 - \exp(-d_t r_t G))$ is the probability of a recombination between SNP t and $t-1$ [93, 151].

2.2.1.1 Admixture LD-based models

Admixture LD-based models estimate local ancestry by using ancestry informative markers (AIMs), which are markers whose allele frequency significantly differs among popula-

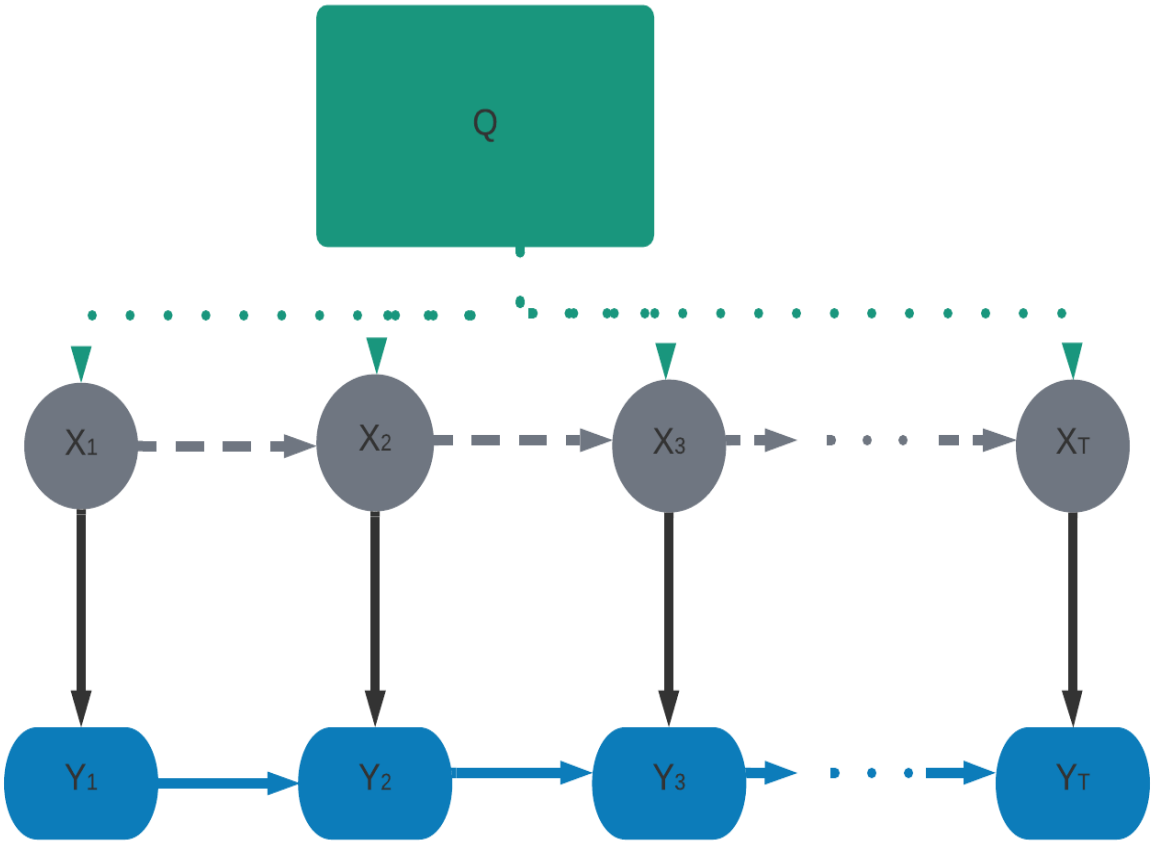


Figure 2.2: The relationship between LD types in ancestry deconvolution. Circle nodes represent the unobserved sequence, rectangular-like blue nodes represent the observed sequence (observations). Arrows represent dependence relationship. Dotted arrows represent mixture LD, dashed arrows represent admixture LD and bold arrows between observations represent background LD.

tions [4]. Admixture LD-based models infer local ancestry using the HMMs integrated with Markov Chain Monte Carlo (MCMC) as a way of dealing with uncertainty in model parameters [4, 81, 150, 151]. They assume that the generations since admixture, the genome-wide average ancestry and the allele frequencies of involved populations are known [4]. Practically, this may not always be the case [78, 93, 94]. Furthermore, inference in most admixture LD-based models is based on the Bayesian framework and they presume unrelated markers [83]. These models use Markov chains to account for admixture LD only. They were developed to improve admixture mapping [4]. As pointed out in **Section 2.2.1**, these models include ANCESTRYMAP [93], ADMIXMAP [94] and STRUCTURE V2 [78]. They are often called early methods. STRUCTURE V2 and ADMIXMAP are multi-way admixtures while ANCESTRYMAP is two-way. The observation probability model for these early methods depends on

the allele frequency of ancestral population at a SNP [4]. According to Sundquist et al. [95], the observation probability model for early methods is defined by

$$P(Y_{\iota t} = z | X_{\iota t} = k) \;=\; \frac{1}{2n_k} \sum_{k=1}^{n_k} \sum_{h=0}^{1} I(a_{\iota hkt} = z), \tag{2.2.2}$$

where $X_{\iota t}$, $Y_{\iota t}$, n_k and h are as defined in **Table 1.1**, and $I(.)$ is an indicator function for $a_{\iota hkt} \in \{A, C, G, T, -\}$ which is the allele of individual ι emitted by haplotype h which was copied from population k at SNP t, and $-$ represents a missing allele.

Early methods improved the application of local ancestry estimates in admixture mapping given small datasets [4]. However, advances in technologies and the progress made in computational methods have availed huge amounts of publicly accessible high density SNP datasets, making early methods inadequate in local ancestry deconvolution. Compared to the AIMs, high density SNPs are more powerful in inferring local ancestry, particularly when the aim is to incorporate SNP and admixture association signals [4, 103]. However, in HMMs SNPs are presumed independent yet high dense datasets do not meet this assumption therefore creating noise and systematic biases, resulting from not accounting for background LD [4]. As such, to reduce the noise and systematic biases, state-of-the-art models evolved [2, 103] including SUPPORTMIX [125], PCADMIX [166] and SEQMIX [167]. Modelling admixture LD only, SUPPORTMIX [125] combines support vector machines (SVMs) and HMMs [2]. SUPPORTMIX emerged in 2012 and is faster than most models proposed before it. More so, it accounts for scarce or extinct reference ancestral populations when deconvoluting ancestry in admixtures formed by three or more source populations [4]. To deconvolve ancestry, SUPPORTMIX trains admixed individuals on a number of source populations that is even greater than that of the ones that mate [4]. Although SUPPORTMIX uses haplotype information (thus, it is haplotype-based), it is faster than early methods that use AIMs and are non-haplotype-based [4]. This is consistent with the fact that SVMs are flexible and can easily handle huge datasets [125].

PCADMIX [166] was introduced in 2012, and similarly to SUPPORTMIX, it subdivides the entire admixed individual genome into windows of contiguous SNPs. It leverages principal component analysis (PCA) from ancestral individuals haplotypes to account for admixture LD using the HMMs [4]. Comparable to SUPPORTMIX, PCADMIX is fast and haplotype-based. Nonetheless, the haplotype information is not always available such that haplotype inference models which often introduce phase switch errors [4] are used to restore the phase information [168, 169]. However, SUPPORTMIX and PCADMIX do not model such errors. Henceforth, in 2013, SEQMIX [167], an HMM-based model which uses exome data was pro-

posed. Exome sequence data is advantageous in that it may detect rare variants that could potentially contain information not contained in other dense SNP datasets [170].

Even though populations are genetically related due to the shared most recent common ancestor (MRCA), most existing models including those discussed so far assume ancestral populations are independent of each other [4]. This is not the case of mSPECTRUM [165], a multi-way ancestry deconvolution method which assumes all source populations originate from a pool of common hypothetical founders. In this case, admixed individuals are a mosaic of segments of hypothetical founder haplotypes. mSPECTRUM extends and generalises HAPMIX by characterising each ancestral population using infinite HMMs (iHMMs) [83]. It models recombination and mutation events as transition and emission probabilities, respectively. Although it requires more computational time than HAPMIX, it has been shown that it performs better than HAPMIX in inferring local ancestry even with few ancestral individuals [83, 165]. However, HMM-based methods including mSPECTRUM do not model background LD [171]. Hence, to reduce noise and systematic biases due to unmodelled background LD, admixture-background LD models emerged [80, 103].

2.2.1.2 Admixture-background LD models

As highlighted in **Section 2.2.1.1**, dense SNP data violate the independence of SNPs assumption. Thus, modelling admixture LD only is inadequate when dealing with such datasets. To account for admixture and background LD, HMMs are extended to Markov hidden Markov models (MHMMs), hierarchical hidden Markov models (HHMMs), factorial hidden Markov models (FHMMs), layered hidden Markov models (layered HMMs) and other multivariate statistics approaches, such as the integration of multivariate normal (MVN) distribution with the HMMs [4]. We note that most admixture-background LD models use haplotype information, for instance, SABER [157], SWITCH [163], HAPAA [95], HAPMIX [92] and ALLOY [158]. The first HMMs extension to MHMMs was implemented in the SABER model in 2006, followed by the SWITCH model of 2008.

In contrast to modelling correlations in local ancestry only as in Equation (2.3.1), correlations in allelic states within ancestry blocks are also modelled. Here, if ancestry transitions between SNPs $t-1$ and t, the emission model is a function of the distribution of alleles at SNP t and $t-1$ [81, 82]. On the other hand, if there is no switch in ancestry between SNP t and $t-1$, we have the emission model as in Equation (2.2.2). The effects of the large parameter set and the inappropriately modelled background LD in the SABER model motivated the proposal of the SWITCH model [81, 92, 103]. Unlike SABER which conditions MHMMs on ancestry state, the

SWITCH MHMMs are conditioned on recombination even if the recombination may not yield an ancestry switch. Similarly to early methods, the occurrence of recombination is determined by the generations since admixture, the physical distance and the recombination rate between SNP $t-1$ and SNP t. However, following the recommendations of authors (richer MHMMs other than the pairwise model in SWITCH and SABER [163]), hierarchical hidden Markov models (HHMMs) were introduced. These consists of the large- and small-scale HMMs [4].

In local ancestry deconvolution, HAPAA [95], a multi-way admixture model pioneered HHMM-based models. It prunes admixed segments that fail to meet some defined threshold [4, 95]. Unlike previously described models, HAPAA accounts for phase switch errors in admixed samples [95]. Nonetheless, HHMMs are challenged by spurious ancestry switches [4], which tend to increase when ancestral populations are: small in size, not close to the ones that mated and closely related [4, 80]. Spurious switches result from challenges in distinguishing haplotypes from different ancestral populations locally [4]. Since, HAPAA does not address the spurious switches problem, in 2009, another HHMM-based method, HAPMIX [92], was proposed.

Unlike HAPAA, HAPMIX assumes unknown phase information for admixed individuals, but does haplotype inference internally. HAPMIX accounts for phase switch errors by allowing miscopying from wrong ancestral populations. However, due to a large number of parameters it estimates, the computational time of HAPMIX increases with the square of the total ancestral haplotypes [4, 92]. As a result, although it outperforms most existing models, HAPMIX is only applicable in two-way admixtures [4, 92]. Thus, the lack of applicability to multi-way admixtures of HAPMIX, the spurious ancestry switches in HAPAA and other HMM-related issues raised opportunities for the development of models based on other HMMs extensions, such as the factorial HMMs, and the two layer HMMs. One such model is a genotype imputation-based and admixture-background LD-based model, GEDI-ADMX [164]. In order to impute genotypes, it trains factorial HMMs (FHMMs) [169, 172, 173] on ancestral haplotypes [4]. Unlike the HHMMs, GEDI-ADMX uses a fixed state HMMs and happens to be the first ancestry deconvolution model that imputes missing genotypes [164].

Local ancestry inference estimates are useful to enrich individuals with the knowledge about their recent family genealogies [174]. To facilitate such studies, CHROMOPAINTER [155] accounts for genealogical information in local ancestry inference. Similarly to STRUCTURE, a clustering approach, CHROMOPAINTER leverages ancestral haplotypes by using haplotype data to extract appropriate genealogical information and represents it in the co-ancestry matrix [4, 83]. Since it accounts for the genealogical processes and the structure of individual genomes, the co-ancestry matrix reveals the genetic similarity among individuals [4]. We

note that, its elements are approximated using the Li and Stephens [45] HMMs. CHROMOPAINTER deconvolves local ancestry in either linked or unlinked SNPs. However, its accuracy deteriorates given an admixture of sub-continental ancestries (i.e., population groups belonging to one continent, such as South Asians and East Asians [175]). More so, it doesn't account for familial relationships, admixture information or genetic drift [4]. Therefore, LAMP-LD/LAMP-HAP [80] which leverages the background LD structure as in HAPAA and HAPMIX using the Li and Stephens [45] model, was introduced [4].

Generally, in multi-way admixtures, LAMP-LD/LAMP-HAP performs better than the other two HMMM-based models introduced before it (HAPAA and HAPMIX). Firstly, LAMP-LD/LAMP-HAP reduces the computational time by compressing haplotype information in each ancestral population [4]. Secondly, since ancestral haplotypes are used in parameter estimation, it is less prone to model parameters misspecification [4]. Finally, it addresses the ancestry switch problem experienced in HHMMs by using a window-based approach. Nonetheless, the assumption of constant ancestry within every window in LAMP-LD/LAMP-HAP is not always practical [4]. This is also true for the assumption of known ancestral haplotypes. This is because methods that are used to obtain phase information are costly and are not always accurate. For example, molecular haplotyping and analysis of family trios is costly [33], while statistical techniques that reconstruct haplotypes (phase information) are not always accurate (phase uncertainty) [4]. As a result, to cater for the scarce haplotypes while accounting for admixture LD, Churchhouse and Marchini [152] introduced MULTIMIX in 2013. MULTIMIX trains model parameters using the MCMC, the expectation maximisation or the classification expectation maximisation algorithms. Similarly to PCADMIX and LAMP-LD/LAMP-HAP, MULTIMIX is window-based, but uses multivariate normal (MVN) distribution to model emission probability. Its ancestry switch model is as in Equation (2.2.1).

In a race towards reducing the noise and systematic biases leading to improved estimates by accurately modelling LD [4], ALLOY [158], was proposed. ALLOY uses FHMMs with non-stationary Markov chains whose memory varies, known as non-homogeneous variable length Markov chains (VLMC) [176]. Through FHMMs, ALLOY captures the inheritance process of both admixed maternal and admixed paternal haplotypes. Most interestingly, accuracy in the ALLOY model improves as it leverages the haplotype structure in the compound state.

Despite all the model developments in local ancestry inference, evaluation studies highlighted that spurious deviations in local ancestry estimates are still limiting admixture mapping [2, 66, 103]. As such, to avoid inaccuracies that may result from pre-specifying biological parameters, that is, recombination rates, ELAI [97], was proposed. It is a two-layer HMM which estimates recombination rates during local ancestry inference and models LD between and

within ancestral populations (i.e., between haplotype groups) [4]. It is a multi-way local ancestry deconvolution tool that plays an important role in inferring recombination rates. Just like MULTIMIX, ELAI requires either phased or unphased population data (ancestral and admixed), hence limiting phase uncertainty and scarcity of ancestral and admixed haplotypes. Accordingly, ELAI performs well even when high-quality haplotypes are not available [4]. Its accuracy is exceptional in recent admixtures where ancestry segments are long [97]. ELAI consists of two statistical parameters: the upper cluster and the lower cluster. Fixing these two statistical parameters highlights the relationship between ELAI and other models in genetics. For example, given a single ancestral/source population (equivalent to an upper cluster of 1), ELAI reduces to a haplotype inference model, fastPHASE [177]. Also, ELAI extends STRUCTURE from unlinked SNPs to densely linked SNPs [97]. In other words, STRUCTURE is an ELAI model that assumes SNPs are independent, the lower and upper cluster are equal, and the lower cluster descends from the upper cluster [4, 97]. Even though the lower and upper clusters can be equal, Guan [97] recommended the lower cluster to be five times the upper cluster which is normally represented by the number of ancestral populations.

To avoid requiring information on which donors relate to which ancestry while accounting for admixture and background LD, MOSAIC [171] was proposed in 2019. Identical to ELAI, it is a two-layer HMMs model, modelling ancestry and copied haplotype switches. Since these switches occur on the first admixture point, they occur at a faster rate than ancestry switches. Instead of dividing the chromosome into windows, MOSAIC inflicts an even grid (approximately 60 grid points per centiMorgan) on recombination along the admixed genome [171]. This speeds up the algorithm and simplifies the inference, as recombination only occurs between consecutive grid points [171]. Additionally, the running time of the MOSAIC model is not affected by the denseness of the SNP data. Generally, MOSAIC is a combination of two models, HAPMIX and GLOBETROTTER [178] (a dating admixture model) [171]. Consistent with HAPMIX, it subsumes admixture directly into the HMMs, but unlike HAPMIX, it is a multi-way admixture model. Identical to GLOBETROTTER, MOSAIC uses the data to infer the relationship between admixed and ancestral individuals [171]. Contrastingly to GLOBETROTTER which fits a mixture model to the ancestry results without using HMMs, MOSAIC increases the accuracy of estimates by establishing the connection between admixed and ancestral individuals directly into the HMMs [171].

2.2.2 Non-LD-based models

This section discusses models that neither account for background nor admixture LD. Some of these models discard SNPs in LD, e.g., LAMP [96] and WINPOP [156], while others include linked and unlinked SNPs without modelling LD [4], e.g., EILA [159] and RFMIX [129].

Since MHMMs are complex, LAMP [96], a window-based model, emerged in 2008. Using the majority vote for every SNP, LAMP infers the most probable ancestry by combining the naive Bayes and "the iterative conditional modes (a clustering algorithm)" [4]. Unlike HMM-based models, it is fast, robust and works with or without ancestral genotypes, particularly in two-way admixtures [4]. In addition to admixed genotypes, LAMP requires ancestral allele frequencies to infer ancestry in multi-way admixtures. It is worth mentioning that by neglecting the LD, LAMP improves in computational speed, but, at the expense of estimate accuracy [4]. We also note that the assumption of no recombination within every window in the LAMP model compromises estimates accuracy in closely related populations [82, 156]. Hence, the modification of LAMP to WINPOP [156], a dynamic programming algorithm. Contrary to LAMP, WINPOP assumes one or more recombination events in every window and the size of windows vary according to the genetic similarity among the involved source populations [4]. These two modifications improve the performance of WINPOP in closely related populations [4, 156].

In order to tackle certain issues in local ancestry that negatively affect the inference power, EILA [159], a multivariate statistics-based method, was proposed. The three issue are:

1. Difficulties in determining how similar admixed SNPs are to the ancestral populations given genotypes have three values at a SNP which are 0, 1 or 2.
2. Difficulties in identifying transition points for ancestral blocks within an individual genome [159]. Such points provide insights into the admixture dates which are rarely known.
3. Reduction in inference power due to the independence of SNPs assumption which possibly causes the exclusion of potential informative SNPs/markers.

Now, EILA addresses the three issues using the following three steps:

1. Instead of using the raw genotype data values, it assigns a quantitative score in the range of 0 to 1 to admixed genotypes. The score quantifies the similarity between admixed SNPs and ancestral populations [4, 159].
2. It identifies transition points for ancestral blocks within an admixed genome by using fused quantile regression (QR).

3. It infers ancestry of each admixed segment using a K-means classifier on admixed genotypes for both linked and unlinked SNPs, overcoming the independence of SNPs challenge.

As sequence data availability increased, in 2013, a discriminative approach (which estimates the posterior probability without using the joint probability), RFMIX [129], was introduced. Unlike generative approaches that rely on information in ancestral individuals, RFMIX is limited to information in admixed individuals [4], which may be lucrative when reference ancestral populations are scarce, for example, Native Americans [129]. To estimate local ancestry, RFMIX uses conditional random fields parametrised with random forests [129]. As shown by the several studies that have deconvoluted local ancestry using RFMIX for different applications [128, 175, 179, 180, 181], the RFMIX model might have improved and/or stabilised local ancestry estimates in different multi-way admixtures. This could be due to its ability to model phase switch errors [4].

However, models developed up-to now assume organisms are diploids, and biological parameters, including recombination rates (except for ELAI), and admixture generations are known, which is not always the case. Consequently, in the race to deconvolute local ancestry in a variety of species without requiring biological or statistical parameters (e.g., admixture date, recombination, mutation rates, window size and number of hidden states), LOTER [56], was proposed in 2018. It is a non-probabilistic model that is formulated from an optimisation problem [4]. Nonetheless, LOTER uses the copying model of Li and Stephens [45]. Although this works for two-way admixtures only, we note that the current version of LOTER models phase switch errors. Furthermore, in comparison to RFMIX, LOTER performs better in ancient admixtures [56].

2.3 Modelling local ancestry deconvolution

This section explores different mathematical and statistical models used for inferring ancestry along admixed genomes. These models are either generative or discriminative or both. Unlike discriminative approaches, generative approaches are easier to fit and therefore most common. They learn the joint probability distribution and use the Bayes theorem to predict the posterior [4]. Generative approaches include hidden Markov models (HMMs) and their extensions, such as, FHMMs, MHMMs, HHMMs, layered HMMs and infinite HMMs, and principal component analysis (PCA) [62]. Contrastingly, to construct the decision boundary and estimate parameters directly from the data, discriminative models learn the conditional

distribution [4]. They include conditional random fields and random forests used in parametrising these conditional random fields in RFMIX, and support vector machines (SVMs) used in SUPPORTMIX.

2.3.1 Finite space hidden Markov models and extensions

A Markov chain process is a stochastic model that explains successive events in which an immediate past state captures the entire history of all other past states (Markov chain assumption) [182]. The number of previous states that each current state depends on is the order of a chain. For example, the Markov chain assumption defines a first order chain. An HMM extends the Markov chain process (or a single stochastic model) [183] to a doubly stochastic model whose observations (observed sequence) are generated by a discrete-valued Markov process [183, 184]. This process is unobserved or invisible to the observer, hence the name "hidden". In order to recover the hidden state sequence, the observed state sequence is used [183]. HMMs are most commonly used in modelling time series and sequential data in both machine learning and statistics [185, 186]. They have been successful in applications including, languages, finance and genomics [78, 87, 92, 93, 97, 184]. The data which can be described by HMMs is generally in the form $Y_{1:T} = (Y_1, \ldots, Y_T)$. An observation at t (where t can be time given time series data or a SNP position in the case of genomic data), denoted by Y_t (discrete or real valued), is generated by a discrete valued hidden state at t, $x_t \in \{1, \ldots, K\}$, such that

$$P(Y_t|x_t, x_{-t}, Y_{1:t-1}, Y_{t+1:T}) \;=\; P(Y_t|x_t) \tag{2.3.1}$$

under the independence assumption, where $x_{-t} = \{x_1, x_2, \ldots, x_{t-1}, x_{t+1}, x_{t+2}, \ldots, x_T\}$ is a set of all states except the one at t. Also, the hidden state sequence is a Markov chain such that

$$P(x_t|x_{-t}, Y_{1:T}) \;=\; P(x_t|x_{t-1}) \tag{2.3.2}$$

under the Markov or 'memoryless' assumption [4]. It is important to note that Equations (2.3.1) and (2.3.2) are the two basic assumptions of the HMMs. Using these equations, three parameters characterise HMMs, these are:

(a). Transition probability model–which describes the probability of transitioning from state k' at time/position $t-1$ to state k at time/position t where $k, k' \in \{1, \ldots, K\}$. Herein,

we denote the transition probabilities by a matrix S with elements given by

$$s_{k'k} = P(x_t = k|x_{t-1} = k') \tag{2.3.3}$$

where $0 \leq s_{k'k} \leq 1$ and $\sum_k s_{k'k} = 1$.

(b). Initial probability model–which is a special form of the transition probabilities and defines the probability of occurrence of a given hidden state, k at time/position $t = 1$, and hence does not have a previous state. The initial probability model is a vector with entries denoted by

$$s_1 = P(x_1 = k). \tag{2.3.4}$$

(c). Emission probability model–which describes the observation model for each state k at t. It is denoted by E and the probabilities are

$$\begin{aligned} e_k(Y_t) &= P(Y_t|x_t = k) \\ &= F(\theta_{x_t}) \end{aligned} \tag{2.3.5}$$

where θ_{x_t} parametrises the emission model distribution F for an observation Y_t given the hidden state at t, denoted as $Y_t|x_t \sim F(\theta_{x_t})$. An emission probability model is either discrete or continuous, if discrete the probability model is a matrix.

The above parameters are denoted by $\lambda = \{s_1, S, E\}$. Based on λ and the relationship between variables in an HMM, the joint probability of the observed and hidden state sequence, provided the independence and Markov assumptions hold is

$$\begin{aligned} P(Y_{1:T}, x_{1:T}|\lambda) &= P(x_1 = k)\prod_{t=2}^{T} P(x_t = k|x_{t-1} = k')\prod_{t=1}^{T} P(Y_t|x_t = k) \\ &= \overbrace{s_1}^{\text{(b)}}\prod_{t=2}^{T}\overbrace{s_{k'k}}^{\text{(a)}}\prod_{t=1}^{T}\overbrace{e_k(Y_t)}^{\text{(c)}} \end{aligned} \tag{2.3.6}$$

Using Equation (2.3.6) and the Bayes theorem, the hidden state sequence is restored as follows

$$\begin{aligned} P(x_{1:T}|\lambda) &= \frac{s_1 e_k(Y_1) \prod_{t=2}^{T} s_{k'k} e_k(Y_t)}{P(Y_{1:T}|\lambda)} \\ &\propto s_1 e_k(Y_1) \prod_{t=2}^{T} s_{k'k} e_k(Y_t) \end{aligned} \tag{2.3.7}$$

In local ancestry inference, most existing models assume the probability of an ancestry k at the first SNP position, $t = 1$ is given by the genome-wide ancestry proportion inherited from population k by the admixed individuals often called ancestry proportions, q_k, i.e.,

$$s_1 = q_k. \tag{2.3.8}$$

These ancestry proportions are either known or unknown, if unknown, they are inferred. Generally, existing Bayesian-based models assume q_k is Dirichlet distributed with parameters $\alpha_1, \ldots, \alpha_K$, that is, $(q_1, q_2, \ldots, q_K) \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$ [78, 94]. We note that, in reality, ancestry proportions vary among same population individuals. Along the admixed individual genome, the way the ancestry switches is determined by the occurrence or non-occurrence of recombination between SNP $t-1$ and t. Hence, given ancestral population k' at SNP $t-1$, the probability of ancestral population k at SNP t is as in Equation (2.2.1). Since admixed genotypes (observed variables) are generated by the ancestry at that SNP (unobserved variables), the probability of observing an admixed genotype depends on the ancestral allele frequency at t and is as in Equation (2.2.2).

Despite their ability to capture some important properties of a system, the two HMMs assumptions are inadequate when solving many real life problems. This is the case with biological data where often the independence assumption is violated [4] and other systems where an observed sequence is generated by more than one hidden state sequence. Thus, HMMs are extended to make Markov models more practical in many real life problems. Subsequent sections discuss the HMM extensions.

2.3.1.1 Accounting for dependencies between observations

A simple HMMs extension is when observations are not drawn from a conditionally independent fixed distribution, but are a Markov chain. Hence, the name Markov hidden Markov models

(MHMMs). Thus, Equation (2.3.1), changes to

$$P(Y_t|x_t, x_{-t}, Y_{1:T}) \;=\; P(Y_t|x_t, Y_{t-1}).$$

In the local ancestry inference context, MHMMs are implemented in the SABER and SWITCH models. In observed genetic data, MHMM or autoregressive HMM (ARHMM) model dependencies that may exist between consecutive markers (or SNPs) [187]. This may provide insights into the variation that may exist between the SNPs. Let τ_k be the expected size of the ancestral blocks copied from each population k at individual level. Given a hybrid isolation (HI) admixture process, $\tau_k = G$ (generations since admixture), such that, the initial and transition probabilities are as in Equation (2.2.1). However, when ancestral populations contribute in different generations: continuous gene flow (CGF), G_k represents the admixture time since ancestral population $k \in \{1, \ldots, K\}$ contributed to the admixture process. Assuming $K \geq 3$, SABER extends Equation (2.2.1) to Equation (2.3.9) [157]:

$$P(x_t = k|x_{t-1} = k', \mathbf{q}, r) \;=\; \exp(-d_t S), \tag{2.3.9}$$

where $S = (s_{k'k})_{1 \leq k', k \leq K}$ is a matrix representing the transition rate from ancestral population k' at SNP $t-1$ to k at t, defined by

$$s_{k'k} \;=\; \begin{cases} q_{k'} \left(\dfrac{G_{k'}^2}{\sum\limits_{\ell=1}^{K} q_\ell G_\ell} \right) - G_{k'} & \text{for } k' = k, \\ q_k \left(\dfrac{G_{k'} G_k}{\sum\limits_{\ell=1}^{K} q_\ell G_\ell} \right) & \text{otherwise,} \end{cases}$$

while its emission model depends on the alleles observed at $t-1$ and t. If these originate from the same population, SABER's emission model is given by Equation (2.2.2), otherwise, it depends on the joint allele frequency at SNP $t-1$ and t. Since SWITCH is closely related to SABER, marginalizing the transition model of SWITCH yields Equation (2.3.9). The emission model of SWITCH is similar to that of SABER, except that in SABER, recombinations that do not result in an ancestry switch are ignored [4, 163].

2.3.1.2 Hierarchical hidden Markov models

In contrast to MHMMs, hierarchical hidden Markov models (HHMMs) consist of HMMs at different levels. HHMMs capture the correlation in observations over long distances via the

higher levels of the hierarchy [188]. In short, HHMMs account for stochastic processes of different scales at different levels. They consist of two hidden states: those that emit (production/child) and those that do not emit (internal/high level/non-emitting) observations. Internal states activate either production or internal states. Three models use HHMMs to infer local ancestry: a two-way admixture, HAPMIX and two multi-way admixture models, HAPAA and LAMP-LD. Since several studies showed that LAMP-LD outperforms HAPAA, we only provide the mathematical details of LAMP-LD, a window-based framework. It separates an admixed genome into non-intersecting windows, $w = [t, t+M)$ where M is the count of SNPs in each window consisting of SNPs from t to the SNP we have after adding M SNPs to SNP t [4]. Given a pair of ancestries within window w, $\mathbf{x}^w = (x_1^w, x_2^w)$ that generates a genotype block in that window Y^w, the probability of observing a genotype within the w^{th} window is

$$P(Y^w|\mathbf{x}^w = (x_1^w, x_2^w)) = \sum_{H_1^w, H_2^w} P(H_1^w|x_1^w) P(H_2^w|x_2^w)$$

where $P(H_r^w|x_r^w)$ for $r \in \{1,2\}$, is the probability that, within a window w, a haplotype segment H_r has been generated by ancestry x_r and the haplotype pair (H_1^w, H_2^w) matches the observed genotype, Y^w. As aforementioned, LAMP-LD estimates the HMMs parameters from ancestral populations provided. Its top level HMMs have $\binom{K}{2}$ states which are the local ancestries of window w, (x_1^w, x_2^w) generating Y^w. Ancestry switches from $(x_1^{w'}, x_2^{w'})$ in window $w-1$ to (x_1^w, x_2^w) in w, with probability given by

$$p\left(\mathbf{x}^w = (x_1^w, x_2^w)|\mathbf{x}^{w-1} = (x_1^{w'}, x_2^{w'})\right) = \begin{cases} \rho & \text{if } x_r^{w'} \neq x_r^w, \text{ for at least one } r \\ \rho^2 & \text{if } x_r^{w'} \neq x_r^w, \ \forall\ r \\ 1 - 2\rho - 3\rho^2 & \text{if } x_r^{w'} = x_r^w \end{cases}$$

where $r \in \{1,2\}$(assumming bi-allelic SNPs), $\rho = D \times 10^{-8}$, D measures the distance between $w' = [t, t+M)$ and $w = [t+M, t+2\times M]$ (in base pairs), where M is the count of SNPs in every window [4, 80].

Given small sample sizes of training sets from ancestral haplotypes, HHMMs have large parameter space and usually suffer from over-fitting [189]. In order to make use of the few (available) training data (that is, genotyped ancestral individuals) while reducing the parameter space and increasing robustness to over-fitting [189], layered hidden Markov models (LHMMs) emerged.

2.3.1.3 Layered hidden Markov models

Layered hidden Markov models (LHMMs) train each layer independently and use the inferential results of the previous layer to train the current layer [189]. LHMMs are less powerful when modelling temporal data [189]. In the local ancestry inference framework, ELAI [97] and MOSAIC [171] are the two LHMM-based models. ELAI consists of two HMM layers (often called clusters): the upper and lower layer. They are responsible for learning the structure of local haplotypes [97], and they both account for the admixture and background LD in admixed individuals.

For each individual $\iota \in \{1,\ldots,2N\}$, let $x_{\iota t} \in \{1,\ldots,K\}$ and $Y_{\iota t} \in \{1,\ldots,L\}$, be the latent state of the upper and the lower clusters at SNP t, and $p_{t\ell}$ be the allele frequency of the lower cluster at SNP t. The transitions between ancestral populations (upper layer switch probabilities) and within ancestral populations (lower layer cluster switch probabilities) occur with frequency u_t and v_t, respectively [97]. Assume $H_{\iota t}$ is the observed haplotype of individual ι from a lower layer cluster at a SNP t. Now, the probability of transitions between hidden states is given by

$$P(x_{\iota t}=k, Y_{\iota t}=\ell | x_{\iota t-1}=k', Y_{\iota t-1}=\ell') \;=\; \begin{cases} q_{\iota k}\beta_{tk\ell} & \text{for } t=1 \\ u_t q_{\iota k}\beta_{tk\ell} + \bar{u}_t v_t \beta_{tk\ell} I(k=k') + \bar{u}_t \bar{v}_t I(k=k') I(\ell=\ell') & \text{for } t \geq 2 \end{cases} \tag{2.3.10}$$

where $k', k \in \{1,\ldots,K\}$, $\ell', \ell \in \{1,\ldots,L\}$, $\bar{a} = (1-a)$, for $a \in \{v_t, u_t\}$, $q_{\iota k}$ is the probability that individual ι jumps into the upper cluster k assuming the jump occurs (it is an equivalent of ancestry proportions), and $\beta_{tk\ell}$ is the probability of individual ι jumping to the lower cluster assuming the jump occurs when the upper cluster is k, thus, β is a $T \times K \times L$ tensor shared by individuals [97]. The following assumptions are made for Equation (2.3.10): given a transition occurs,

- The upper cluster of SNP $t-1$ is independent to the upper cluster at t. The same applies for the lower cluster (Y_{t-1}) and (Y_t). Since it reduces the count of parameters, this assumption simplifies computations. The same assumption has also been employed in the Li and Stephens [45] model.
- The upper cluster only takes values according to the ancestry proportions of an admixed individual, $q_{\iota k}$.
- If a switch occurs and $x_{t-1} = k'$, then the lower cluster takes values according to β, allowing variations in LD across SNPs.

- Whenever the upper cluster switches, the lower cluster also switches. On the other hand, the lower cluster might switch without the upper cluster switching, promoting upper layer specific LD [97].

The emission model in the main HMMs is defined by

$$p(H_{\iota t}|x_{\iota t},Y_{\iota t},\zeta) \;=\; p(H_{\iota t}|Y_{\iota t},\zeta) \;=\; \begin{cases} p_{tY_{\iota t}} & \text{if } H_{\iota t}=1 \\ 1-p_{tY_{\iota t}} & \text{if } H_{\iota t}=0, \\ 1 & \text{if } H_{\iota t} \text{ is missing,} \end{cases}$$

where $\zeta=\{p,q_{\iota k},\beta_{tk\ell},u_t,v_t\}$ is a set of parameters in the HMMs. Therefore, given $2N$ haplotypes, the data likelihood is given by

$$p(H_1,\ldots,H_{2N},x_1,\ldots,x_{2N},Y_1,\ldots,Y_{2N}) \;=\; \prod_{\iota=1}^{2N}\prod_{t=1}^{T} p(H_{\iota t}|Y_{\iota t},\zeta)P(x_{\iota t},Y_{\iota t}|\zeta)$$

Assuming that the main and ancillary HMMs are independent and ancestral haplotypes, Z, have been observed, $W_{zt}\in\{1,\ldots,K\}$ is the population at SNP t that haplotype z inherited, i.e., the upper cluster. The probability of switching in the ancillary HMMs is given by

$$p(W_{zt}=k|W_{z(t-1)}=k') \;=\; \begin{cases} a_{kz} & \text{for } t=1 \\ \rho_t a_{k'z}+(1-\rho_t)I(k'=k) & \text{for } t\geq 2, \end{cases}$$

where a_{kz} is the probability that the upper cluster is k. Since inference is done independently in layers, there is no relationship between jump probabilities (ρ) in the ancillary HMMs and the main HMMs. The emission model of the ancillary HMMs is defined by

$$p\left(p_{t\ell}|W_{zt},\zeta\right) \;=\; \text{Beta}\left(p_{t\ell};F\eta_{tW_{zt}},F(1-\eta_{tW_{zt}})\right)$$

where, $\zeta=\{\eta,a_{kz},\rho\}$, $\text{Beta}(x;b_1,b_2)$ is the Beta distribution with parameters b_1 and b_2, and F models population divergence, i.e., F_{ST}. In this case F stabilises $p_{t\ell}$ estimates and is set to 1.

Similarly to mSPECTRUM [165], every ancestry switch leads to a haplotype switch in MOSAIC [171]. Assume $x_{\iota t}$ is the ancestry of individual ι at grid point t, $s^{\iota}_{\ell k}$ is the probability that a switch occurred from ancestry ℓ to k between grid points t and $t-1$ in admixed individual ι, where these probabilities yield individual-specific matrices S^{ι}, the sum of row entries are not required to be 1. $\mu_{\mathfrak{p}k}$ is the probability of selecting from ancestral panel $\mathfrak{p}$ when the local

ancestry is k, such that $\sum_k \mu_{\mathfrak{p}k} = 1$, and ρ (a scalar), is the probability of recombination within ancestry in the absence of an ancestry switch. Now, the probability of transition from ancestry, haplotype pair (ℓ, h'_d) to (k, h_d), where h_d is the contributing haplotype h in ancestral panel d for admixed individual ι is

$$P(x_{\iota t} = (k,h_d) | x_{\iota(t-1)} = (\ell, h'_d)) = \begin{cases} s^{\iota}_{\ell k} \frac{\mu_{\mathfrak{p}k}}{N_{\mathfrak{p}}} & \text{if } \ell \neq k \\ \left(\vec{s}^{\iota}_{\ell.}\rho + s^{\iota}_{\ell\ell}\right) \frac{\mu_{\mathfrak{p}k}}{N_{\mathfrak{p}}} & \text{if } \ell = k,\ h_d \neq h'_d \\ \left(\vec{s}^{\iota}_{\ell.}\rho + s^{\iota}_{\ell\ell}\right) \frac{\mu_{\mathfrak{p}k}}{N_{\mathfrak{p}}} + \vec{s}^{\iota}_{\ell.}(1-\rho) & \text{if } \ell = k,\ h_d = h'_d \end{cases} \tag{2.3.11}$$

where $\vec{s}^{\iota}_{\ell.} = (1 - s^{\iota}_{\ell.})$, $s^{\iota}_{\ell.} = \sum_k s^{\iota}_{\ell k}$ and $N_{\mathfrak{p}}$ counts the contributing haplotypes in ancestral panel $\mathfrak{p}$. It is worth mentioning that, S^{ι} is specific to admixed individual ι thus, admixed individuals might have different ancestry proportions and admixture models (continuous gene flow or hybrid isolation), while, μ and ρ are the same for all admixed population individuals. As a result, in the case of panmixia post-admixture or ancient admixture events, it is expected to have similar estimates of S^{ι} for all individuals.

Given bi-allelic SNPs and admixed individual information denoted by Y_{th} for contributing haplotype h at SNP t, the emission probability is

$$\theta(1 - Y_{th}) + (1 - \theta)Y_{th} \tag{2.3.12}$$

where

$$Y_{th} = \begin{cases} 1 & \text{if contributing haplotype } h \text{ has SNP 1 at } t \\ 0 & \text{otherwise} \end{cases},$$

θ is the probability of having an error in the allele of a locally copied haplotype and that of a copying haplotype, similar to miscopying, in HAPMIX [171].

2.3.1.4 Factorial hidden Markov models

Apart from dividing the problem into smaller time series patterns or units, the observation sequence can be generated by two or more Markov chains. This is the case for factorial hidden Markov models (FHMMs). They capture the complex statistical structures in the observed sequences [190], hence, they are most useful for SNP data where the independence assumption of HMM is violated. Let x_t^f be the state of a factor $f \in \{1, \ldots, F\}$ at position t. In the local

ancestry inference context, t is the SNP position and we assume two independent factors yielding chains: $\mathbf{x}^1 = (x_1^1, \ldots, x_T^1)$ and $\mathbf{x}^2 = (x_1^2, \ldots, x_T^2)$ that generate the observation sequence $\mathbf{Y} = Y_{1:T} = (Y_1, \ldots, Y_T)$. This is the idea behind ALLOY [158], the most popular FHMM-based model. It assumes a set of ancestral haplotypes share a common local structure through a state space of haplotype clusters A_t and that $a_t \in A_t$ is a particular ancestral haplotype at SNP t with ancestry $anc(a_t)$. Each ancestral haplotype represents an allele, $z_t \in \{0,1\}$ at SNP t. For the T SNPs of an admixed individual, the hidden haplotypes are $x^{m,p} = (x_1^{m,p}, x_2^{m,p}, \ldots, x_T^{m,p})$ where $(x_t^m, x_t^p) \in A_t$ are maternal and paternal haplotypes, respectively, while $(Y_1, \ldots, Y_T)$ is the observed minor allele count such that $Y_t \in \{0,1,2\}$. Recombination between consecutive SNPs ($t-1$ and t) occurs with probability $1 - \exp(-r_t(G-1)d_t)$, and does not occur with probability $\exp(-r_t(G-1)d_t)$, where G is the time since admixture in generations, d_t and r_t is the genetic distance and the recombination rate between SNP $t-1$ and t, respectively. Therefore, the probability of switching ancestry is given by

$$P(x_t|x_{t-1}) \;=\; \exp(-r_t(G-1)d_t)P_{anc(x_t)}(x_t|x_{t-1}) + (1-\exp(-r_t(G-1)d_t))\,P_{anc(x_t)}(x_t) \tag{2.3.13}$$

where $P_{anc(x_t)}(x_t)$ is the intra-population haplotype cluster prior at t and $P_{anc(x_t)}(x_t|x_{t-1})$ models within ancestral population transitions while accounting for LD. The prior probability of the ancestral haplotype cluster is defined by

$$P(x_t) \;=\; q_{anc(x_t)}P_{anc(x_t)}(x_t) \tag{2.3.14}$$

where $q_{anc(x_t)}$ is the ancestry proportion of the ancestral haplotype at SNP t, and $P_{anc(x_t)}(x_t)$ is as defined before. For each population, a directed acyclic graph is formed with an edge at SNP t, for cluster c, denoted by e_t^c [4]. The number of haplotypes in a population sample passing through cluster c are the weights at SNP t defined by w_t^c. Assume s_t^c is the source and τ_t^c is the target node for the edge of cluster c at SNP t (e_t^c). The prior probabilities in Equation (2.3.14), and transition probabilities in Equation (2.3.13) for each specific population are defined by

$$P_{anc}(x_t = a_t^{anc,c}) \;=\; w_t^c / \sum_{c'} w_t^{c'}, \text{ and}$$

$$P_{anc}(x_t = a_t^{anc,c} | x_{t-1} = a_{t-1}^{anc,c'}) \;=\; \begin{cases} \dfrac{w_t^c}{\sum_{c' \text{ s.t } \tau_{t-1}^{c'} = s_t^c} w_{t-1}^{c'}} & \text{if } \tau_{t-1}^{c'} = s_t^c \\ 0 & \text{otherwise} \end{cases},$$

where $a_t^{anc,c}$ is the haplotype cluster c of ancestral population anc at SNP t. Meanwhile, the emission probability of the ALLOY model is given by

$$P(Y_t|x_t^m = a_t, x_t^p = a_t') = \begin{cases} 1-2\varepsilon & z_t + z_t' = Y_t \\ \varepsilon & \text{otherwise} \end{cases},$$

where ε is the genotyping error rate.

2.3.2 Infinite hidden Markov models

Although finite space HMMs (**Section 2.3.1**) have been successfully applied in solving many real life problems, they are not so appropriate for characterising complex and real datasets, where the cardinality of the latent or hidden states may be difficult to specify before learning [185, 186, 191, 192]. Pre-specifying the number of latent states is often called fixing parameters in finite space HMMs (standard HMMs, MHMMs, HMMMs, LHMMs and FHMMs), and this complicates accounting for model complexities. Fortunately, Bayesian nonparametrics (BNPs) is an alternative as it allows hidden states and their cardinality to be inferred from the data [185, 186, 191, 192]. In short, BNPs extend finite space HMMs to infinite space HMMs whose cardinality is unknown called infinite hidden Markov models (iHMMs). Although Teh et al. [193] presented a theoretical foundation of iHMM, they were first proposed by Beal et al. [192]. Similar to the finite space HMMs, iHMMs have three HMM parameters, corresponding to Equations (2.3.3), (2.3.4) and (2.3.5). However, unlike finite space HMMs, the transition probability matrix in iHMMs consist of infinite dimensional row vectors. To define such an infinite dimensional matrix, the hierarchical Dirichlet process (HDP) is used, a reason why iHMMs are called hierarchical Dirichlet process hidden Markov models (HDP-HMMs). We provide more details on HDP in **Section 2.3.2.1**.

2.3.2.1 Hierarchical Dirichlet process

Consider a survey conducted on individual dish preferences in the Southern African Development Community (SADC) region: among South Africans, Namibians, Malawians, Tswanas from Botswana, Zimbabweans, Malagasy people, Congolese from the Democratic Republic of Congo (DRC) and Mauritians. When modelling dish preferences in the SADC region, it is practical for nationals of different countries (groups) to prefer the same dish. For example, although some might prefer rice, most Zimbabweans, Malawians, Namibians, Congolese and South Africans prefer pap. The hierarchical Dirichlet process (HDP) is most useful for

modelling data from multiple groups where each group characteristics can be captured by a Dirichlet process (DP)[1] [33, 191, 193, 194]. Therefore, the HDP captures the similarities and differences between and within groups, yet allowing the unknown count of unobserved components to be shared among multiple groups. This is the case when modelling haplotype inheritance in different population groups that have common origins (founders). Here, individual haplotypes are similar and unique in certain ways [33, 165].

Assuming K populations with common origins, for each population $k \in \{1,2,\ldots,K\}$, define a DP, Q_k such that all the DPs are coupled through a base measure, Q_0 which is also a DP. That is,

$$Q_k \sim \text{DP}(\alpha_2, Q_0), \qquad Q_0 \sim \text{DP}(\alpha_1, H),$$

where α_2 and α_1 are the scaling parameters of Q_k and Q_0, respectively and H is the base measure and mean of Q_0. Now, given a set of haplotypes (observations) in each population k, $y_{k1}, y_{k2}, \ldots, y_{kN_k}$, the generative model of the HDP mixture model is

$$\begin{aligned}
Q_0 &= \sum_{\ell=1}^{\infty} \beta_\ell \delta_{\theta_\ell}, \quad \beta \sim \text{GEM}(\alpha_1), \quad \theta_\ell \sim H, \quad \ell = 1,2,\ldots, \\
Q_k &= \sum_{t=1}^{\infty} s_{kt}^* \delta_{\theta_{kt}^*}, \quad s_k^* \sim \text{GEM}(\alpha_2), \quad \theta_{kt}^* \sim Q_0, \quad t = 1,2,\ldots \quad k = 1,2,\ldots,K, \\
\theta'_{ku} &\sim Q_k, \quad y_{ku} \sim F(\theta'_{ku}), \quad u = 1,2,\ldots,N_k.
\end{aligned}$$

The Chinese restaurant franchise

Taking the different groups in the HDP as restaurants in a franchise, the HDP has a culinary metaphor known as the Chinese restaurant franchise (CRF). CRF extends the Chinese restaurant process (see **Section A.1.4.2**) to include more than one restaurant, say K groups of restaurants all trading under the same franchise. Each restaurant has an unknown number of tables and dishes that are not globally unique. A dish can be served at multiple tables in the same restaurant. We note that customer assignment to restaurants depend on their restaurant. Therefore, each restaurant describes a Chinese restaurant process (CRP). When a customer (y_{ku}) enters restaurant k, they sit at table $t_{ku} \sim s_k^*$. The table then chooses a dish $\theta_{kt}^* \sim Q_0$ or rather dish indicator $\ell_{kt} \sim \beta$. Therefore, customer y_{ku} eats dish $\theta'_{ku} = \theta^*_{kt_{ku}} = \theta_{\ell_{kt_{ku}}}$ [191, 193, 194, 195]. **Figure 2.3** depicts a Chinese restaurant franchise,

[1] see, **Appendix A** for the definition of a Dirichlet process (DP)

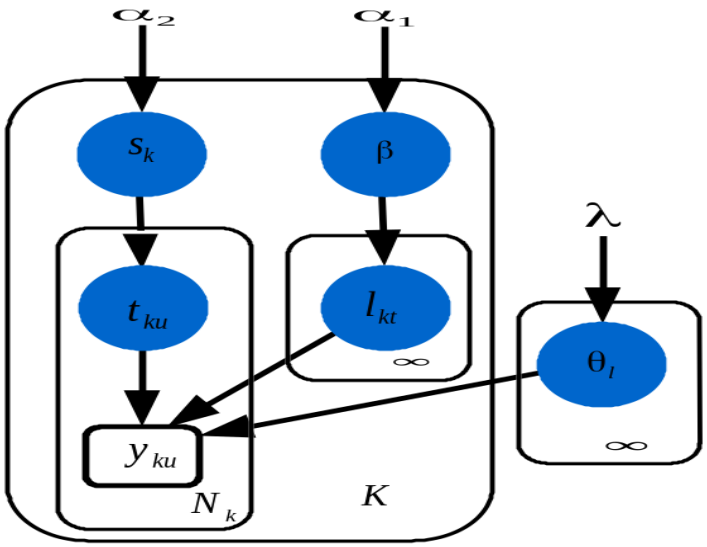


Figure 2.3: The graphical model of the Chinese restaurant franchise in Equation (2.3.15).

whose generative model is given in Equation (2.3.15).

$$\begin{aligned} \ell_{kt}|\beta &\sim \beta, & \beta &\sim \text{GEM}(\alpha_1), \\ t_{ku}|s_k^* &\sim s_k^* & s_k^* &\sim \text{GEM}(\alpha_2), \\ \theta_\ell &\sim H & y_{ku} &\sim F(\theta_{\ell_{kt_{ku}}}). \end{aligned} \tag{2.3.15}$$

Assume n_{kt}^* counts customers seated at table t in restaurant k and $m_{k\ell}$ counts the tables serving dish ℓ in restaurant k. Integrating out the stick breaking measures, β and s_k^* yields the conditional probabilities of table and dish assignments

$$\begin{aligned} p(t_{ku}|t_{k1},t_{k2},\ldots,t_{ku-1},\alpha_2) &\propto \alpha_2 I(t_{ku},\tilde{t}) + \sum_{t=1}^{T_k} n_{kt}^* I(t_{ku},t) \\ p(\ell_{kt}|\underline{\ell}_1,\underline{\ell}_2,\ldots,\underline{\ell}_{k-1},\ell_{k1},\ldots,\ell_{kt-1},\alpha_1) &\propto \alpha_1 I(\ell_{kt},\tilde{\ell}) + \sum_{\ell=1}^{L} m_{\cdot\ell} I(\ell_{kt},\ell) \end{aligned} \tag{2.3.16}$$

where L is the count of distinct dishes that have been served so far in the franchise, T_k is the count of tables represented so far in restaurant k excluding the table assignment of the u^{th} customer, $\underline{\ell}_k = \{\ell_{k1},\ell_{k2},\ldots,\ell_{kT_k}\}$, $\tilde{t}$ and $\tilde{\ell}$ represents the unrepresented table and dish, respectively, $I(a,b)$ is the indicator variable and $m_{\cdot\ell} = \sum_{k=1}^{K} m_{k\ell}$ is the count of all tables serving dish ℓ in the franchise.

As mentioned earlier, in a restaurant, a dish can be served by multiple tables, that is, $\theta_{kt}^* = \theta_\ell$

(or same ℓ_{kt} for different tables) [194, 195]. Thus, we can express Q_k as a function of θ_ℓ as follows

$$Q_k = \sum_{\ell=1}^{\infty} s_{k\ell}\delta_{\theta_\ell}, \qquad s_k \sim \text{DP}(\alpha_2, \beta), \qquad \beta \sim \text{GEM}(\alpha_1), \qquad \theta_\ell \sim H \tag{2.3.17}$$

where

$$s_{k\ell} = v_{k\ell}\prod_{m=1}^{\ell-1}(1-v_{km}), \qquad v_{k\ell}|\alpha_2, \beta_1, \ldots, \beta_\ell \sim \text{Beta}\left(\alpha_2\beta_\ell, \alpha_2\left(1-\sum_{m=1}^{\ell}\beta_m\right)\right),$$

and, unlike s_k^* a restaurant-specific density over tables, s_k is over dishes, such that,

$$s_{k\ell} = \sum_{t|\ell_{kt}=\ell} s_{kt}^*.$$

Now, assuming z_{ku} is an indicator variable for the dish observation y_{ku} eats, then, $z_{ku} = l_{kt_{ku}}$ [195]. Thus, we have another HDP representation as follows

$$\begin{aligned} \beta &\sim \text{GEM}(\alpha_1), & s_k &\sim \text{DP}(\alpha_2, \beta), & z_{ku} &\sim s_k \\ \theta_\ell &\sim H, & y_{ku} &\sim F(\theta_{z_{ku}}) \end{aligned}$$

2.3.2.2 Hierarchical Dirichlet process hidden Markov models (HDP-HMMs)

Since it is build from multiple DPs (infinite dimensional vectors), the HDP is a suitable prior for the transition probability density in infinite state-space HMMs. Its hierarchical structure allows the sharing of statistical strength in this case the sparse state-space [193, 196]. Let z_t be the Markov chain at time t and s_ℓ be the transition density for state ℓ. Therefore, $z_t \sim s_{z_{t-1}}$. This means $z_{t-1} = \ell$ determines the cluster that observation y_t associates with [195], hence, we write $z_t \sim s_\ell$. Given a hidden state sequence $\{z_t\}_{t=1}^T \in \{1, 2, \ldots\}$ and observed sequence $\{y\}_{t=1}^T$, the HDP-HMM is described as

$$\begin{aligned} z_t &\sim s_{z_{t-1}}, & s_\ell &\sim \text{DP}(\alpha_2, \beta), & \beta &\sim \text{GEM}(\alpha_1), \\ \theta_k &\sim H, & y_t &\sim F(\theta_{z_t}). \end{aligned} \tag{2.3.18}$$

Assume z_t and z_{t+1} are the parent and child, respectively. In the HDP-HMMs context, we assume a one-to-one correspondence between indices ku (as described in the CRF, **Section 2.3.2.1**) and t, and that every t have an index match in ku (bijective mapping g :

$ku \rightarrow t$) [195]. Unlike in HDP, where a customer is pre-assigned to a restaurant based on their group, in HDP-HMMs, a parent (z_t) enters a restaurant according to its parent (grandparent) $z_{t-1} = k$,

$$\underset{z_{t-1}}{\text{grandparent}} \longrightarrow \underset{z_t}{\text{parent}} \longrightarrow \underset{z_{t+1}}{\text{child}}.$$

This parent chooses a table $t_{ku} \sim s^*_k$ and is served dish $\ell_{kt} \sim \beta$. Thus, a dish index the parent eats, $\ell_{kt_{ku}} = z_{ku} = z_t$ governs their likelihood dish θ_{z_t} and the restaurant the child z_{t+1} visits [195].

2.3.2.3 Application of HDP in local ancestry inference

Assume K previously isolated populations originating from a pool of common founder haplotypes mate to form an admixed population G generations ago [165]. At a genomic position (SNP) t, the observed admixed allele is generated by an unknown founder that has been copied from an unobserved ancestral population [165]. Therefore, ancestry switch is governed by the transition in copying from a founder at SNP $t-1$ to a founder at SNP t, and the occurrence of recombination between $t-1$ and t of an individual [165]. This is similar to part of speech tagging using the joint labelling of tags and chunks [197]. Denote the founder haplotype by $\ell \in \{1,2,\ldots\}$, a founder allele at SNP t by $F_{\ell t} \in \{0,1\}$, where 0 is the minor and 1 is the major allele. Also, let $x^1_{\iota t}$ be the variable that selects for the founder haplotype of admixed individual ι at t, $x^2_{\iota t}$ be the ancestry of individual ι at SNP t, $\nu^k_\ell = P(x^1_{\iota 1} = \ell, x^2_{\iota 1} = k)$ be the initial and background probability of founder ℓ whose population of origin is k, and $s^k_{\ell'\ell}$ be the probability of transitioning from copying from founder $\ell' \in \{1,2,\ldots\}$ at SNP $t-1$ to founder ℓ at SNP t in population k. The initial probability is given by

$$\begin{aligned} P(x^1_{\iota 1} = \ell, x^2_{\iota 1} = k) &= P(x^1_{\iota 1} = \ell) P(x^2_{\iota 1} = k) \\ &= \nu^k_\ell q_{\iota k} \end{aligned} \tag{2.3.19}$$

On the other hand, founder haplotypes and ancestries switch according to

$$P\left(x^1_{\iota t} = \ell, x^2_{\iota t} = k | x^1_{\iota t-1} = \ell', x^2_{\iota t-1} = k', \omega\right) = \begin{cases} e_1 e_2 + \bar{e_1} e_2 s^k_{\ell'\ell} + e_1 \bar{e_2} q_{\iota k} \nu^k_\ell & \text{if } \ell' = \ell \text{ and } k' = k \\ \bar{e_1} e_2 s^k_{\ell'\ell} + e_1 \bar{e_2} q_{\iota k} \nu^k_\ell & \text{if } \ell' \neq \ell \text{ and } k' = k \\ e_1 \bar{e_2} q_{\iota k} \nu^k_\ell & \text{if } \ell' \neq \ell \text{ and } k' \neq k \end{cases} \tag{2.3.20}$$

where $\omega = \{v_\ell^k, s_{\ell'\ell}^k, r_t, d_t\}$, $e_1 = \exp(-r_t d_t \tau_k)$ is the probability of no recombination occurring in individuals belonging to ancestral population k between SNP $t-1$ and t, $\bar{e_1} = 1-\exp(-r_t d_t \tau_k)$ is the probability of having at least one recombination event in the ancestral population k individual, $e_2 = \exp(-r_t d_t G)$ is the probability of no recombination occurring in the admixed individual between SNP $t-1$ and t, $\bar{e_1} = 1-\exp(-r_t d_t G)$ is the probability of the occurrence of at least one recombination event in the admixed individual. The initial and background probability and the probability of transitioning from state ℓ' to any other state is distributed as follows

$$v_\ell^k \sim \text{DP}(\alpha_2, \beta), \qquad s_{\ell'}|\alpha_2, \beta \sim \text{DP}(\alpha_2, \beta),$$

where $\beta|\alpha_1 \sim \text{GEM}(\alpha_1)$ is the base measure of all the Dirichlet processes, α_1 and α_2 are scaling parameters of the global and group DPs. They are hyper-parameters and have an inverse Gamma prior. On the other hand, $q_{\iota k}$, r_t, d_t and G are as defined in **Section 1.7**, and τ_k is a scaling parameter for recombination in each ancestral population k. The probability of observing admixed individual ι allele at a SNP t is given by

$$P(y_{\iota t}|F_{\ell t}, \eta_\ell) = \eta_\ell I(y_{\iota t} \neq F_{\ell t}) + (1-\eta_\ell) I(y_{\iota t} = F_{\ell t}) \tag{2.3.21}$$

where η_ℓ is a Beta distributed parameter that governs the mutation process from founder haplotype ℓ to individual ι.

Although beam sampling (a variant of the forward-backward) mixes slowly in HDP-HMMs with state persistence[2]

2.3.3 Principal component analysis

Owing to its computational tractability and ability to handle large datasets, principal component analysis (PCA) is among the most widely used approaches in statistical genetics [90, 148, 166, 199, 200]. Its specific applications include correcting for population stratification in association studies [90, 200], inferring shared genetic ancestry [90], understanding population history and migrations [166, 199] and cluster analysis in sub-populations [199]. During population clustering, individual clusters represent genetic populations. If the clustering problem involves admixed individuals, it is expected that they cluster in between the ancestral individuals, to show that they do not belong to one genetic population [166, 199]. PCA decomposes

[2]State persistent is when it is most probable for the system to remain in the same state and less probable to visit new states [198]. [195], it is the inference algorithm used by mSPECTRUM [165].

the variance and covariance matrix to reduce dimensionality [199, 201]. In other words, principal components explain the differences among individuals. In the reduced dimension space, individual locations reflect their genetic similarity [155]. Consider N admixed individuals genotyped at $T > N$ SNPs, let $\mathscr{G} = (\mathscr{G}_{\imath t})_{1\leq \imath \leq N, 1\leq t\leq T}$ be a matrix of genotypes with individuals in rows and SNPs in columns, where $\mathscr{G}_{\imath t} \in \{0,1,2\}$ counts the reference alleles for individual $\imath$ at SNP t. Therefore, the column mean is described by

$$\mu_t \;=\; \frac{1}{N}\sum_{\imath=1}^{N}\mathscr{G}_{\imath t}.$$

As highlighted in **Section 1.2.2**, genetic drift may change the SNP frequency. It is noted that drift usually occur at a rate that is proportional to $\hat{p}_t(1-\hat{p}_t)$ in every generation [200]. Therefore, it is appropriate to correct and normalise each $\mathscr{G}_{\imath t}$ as follows

$$\tilde{\mathscr{G}}_{\imath t} \;=\; \frac{(\mathscr{G}_{\imath t}-\mu_t)}{\sqrt{\hat{p}_t(1-\hat{p}_t)}} \tag{2.3.22}$$

where $\hat{p}_t$ estimates the allele frequency at SNP t and $\hat{p}_t(1-\hat{p}_t)$ estimates the variance. Based on unpublished data, we note that the normalisation in Equation (2.3.22) may improve results in simulated SNP data and perhaps enhance visibility of known structure in real SNP data [200]. But, this is not the case for micro-satellite data that does not need normalisation except correction [90, 200]. Now, using Equation (2.3.22), the genetic related matrix (GRM) for all possible individuals is derived as follows

$$\Psi \;=\; \frac{1}{T}\tilde{\mathscr{G}}'\tilde{\mathscr{G}}$$

where $\tilde{\mathscr{G}}$ is as defined in Equation (2.3.22) and $\tilde{\mathscr{G}}'$ is the transpose of $\tilde{\mathscr{G}}$. This means entries $(\psi_{\imath j})_{1\leq \imath, j\leq N}$ are given by

$$\psi_{\imath j} \;=\; \frac{1}{T}\sum_{t=1}^{T}\frac{(\mathscr{G}_{\imath t}-\mu_t)'(\mathscr{G}_{jt}-\mu_t)}{\hat{p}_t(1-\hat{p}_t)},$$

and they measure the average genetic similarity between individuals $\imath$ and j in a given population. Assume Ψ has $M \leq N$ eigenvalues, given by $(\lambda_m)_{1\leq m\leq M}$, such that $\lambda_1 > \lambda_2 > \cdots > \lambda_M$ are the eigenvalues from the largest to the smallest. Since Ψ is a Wishart matrix which is a generalisation of Chi-squared distributions, we set

$$\mu_{MT} \;=\; \frac{\left(\sqrt{T-1}+\sqrt{M}\right)^2}{T}$$

and

$$\sigma_{MT} = \left(\frac{\left(\sqrt{T-1}+\sqrt{M}\right)}{T}\right)\left(\frac{1}{\left(\sqrt{T-1}+\frac{1}{\sqrt{M}}\right)}\right)^{1/3}$$

to be the mean and standard deviation of the Gaussian distribution used for the rectangular matrix $\mathscr{G}$ [200]. Now, correcting and normalizing the largest eigenvalue, we have

$$\nearrow_1 = \frac{(\lambda_1 - \mu_{MT})}{\sigma_{MT}}$$

which is approximately Tracy-Widom distributed. The inference of population or individual relationships is influenced by the sample size and the Ψ matrix [90, 202]. Given T (number of SNPs under study) $> N$ (number of individuals under study), the rank of Ψ is $M-1$ and its eigenvalues are from the $(M-1)\times(M-1)$ Wishart matrix [200]. Thus, there are two unknown parameters: the theoretical statistical parameter modelling the Wishart Ψ, denoted by $\mathfrak{T}'$ and the variance of the normal distribution used for the entries of the rectangular matrix $\mathscr{G}$, denoted by $\hat{\sigma}^2$. The two are estimated as follows

$$\mathfrak{T}' = \frac{(M+1)\left(\sum_{m=1}^{M}\lambda_m\right)^2}{\left((M-1)\sum_{m=1}^{M}\lambda_m^2\right)-\left(\sum_{m=1}^{M}\lambda_m\right)^2}$$

and

$$\hat{\sigma}^2 = \frac{\left(\sum_{m=1}^{M}\lambda_m\right)}{(M-1)\mathfrak{T}'}$$

2.3.4 Quantile regression

Quantile regression was first applied in biology by Eilers and De Menezes [203] on array Comparative Genomic Hybridisation (CGH) data for smoothing. Recently, it has been applied by Yang et al. [159] in the EILA model, to detect ancestry breakpoints in admixed individuals based on the similarity between admixed and ancestral population SNPs. Assume A and B are two ancestral populations that interbred to form two-way admixed individuals. As a result, alleles at a SNP are either copied from both A and B or A only or B only [4]. EILA applies fused quantile regression in order to estimate a series of terms z_{it}, that optimises the numerical

score $0 \leq n_{\iota t} \leq 1$, thus, solving the following optimisation problem:

$$\min_{z_{\iota t}} \left(\sum_{t=1}^{T} \left| n_{\iota t} - z_{\iota t} \right| + \lambda \sum_{t=2}^{T} \left| z_{\iota t} - z_{\iota t-1} \right| \right) \tag{2.3.23}$$

where λ is a tuning parameter and $n_{\iota t}$ is the probability that a genotype of admixed individual ι at t, $\mathscr{G}_{\iota t} \in \{0,1,2\}$ descends from ancestral population A. Relation (2.3.23) terms measure the extent to which $z_{\iota t}$ fit $n_{\iota t}$, and the penalty, respectively [203]. The measure of penalty discourages change in $z_{\iota t}$. Thus, a small λ produces an insignificant penalty effect, while a large λ produces a significant penalty, indicating $z_{\iota t}$ is not a best fit of $n_{\iota t}$, so that $z_{\iota t} - z_{\iota t-1} \approx 0$ [159]. Although the Euclidean norm ($\|\cdot\|_2$) could replace the Manhattan norm ($\|\cdot\|_1$) in Equation (2.3.23), Eilers and De Menezes [203] showed that $\|\cdot\|_1$ improves the results as it flattens plateaus leading to sudden jumps. In the local ancestry inference context, sudden jumps and plateaus represent breakpoints [4]. For example, in the two-way admixture illustration, plateaus indicate SNPs descended from population A, B or the two within a given region [159].

2.3.5 Support vector machines (SVMs)

Support vector machines (SVMs) solve classification problems consisting of high dimensional datasets [183, 204]. Due to their flexibility to various data types, SVMs have been widely used in computational biology [205]. In the local ancestry inference framework, SVMs allow admixed individuals to be trained on ancestral populations that may be greater or equal in number to the ones that mixed, i.e., $K' \geq K$ [4]. It is assumed that admixed and ancestral population individuals are genotyped on the same number of SNPs (T) and the K' training classifiers include the K ancestral populations. For every ancestral and admixed haploid genome ι, let $\mathbf{y}_\iota \in \{-1,1\}$ denote a vector of dimension T, where -1 and 1 represent the minor and major allele, respectively. Also, we associate each ancestral haploid genome with $x_\iota \in \{1,2,\ldots,K'\}$, to identify the population of the haploid genome [125]. Now, divide $\mathbf{y}_\iota$ into windows $\mathbf{y}_\iota^j$, such that the j^{th} window of haploid genome ι consist of w consecutive SNPs for $j = 1,\cdots,T/w$. Therefore, for every window of the admixed haploid ι, we aim to determine the optimal label, x_ι^j. To address this, first, for every window $\mathbf{y}_\iota^j$, we classify the labels x_ι^j in each admixed genome using SVMs. Secondly, we apply the outputs of SVMs in HMMs for smoothing the population assignments across windows [125]. Now, to determine the decision boundary that

separates classes or populations, we solve the following optimisation problem

$$\begin{aligned} &\min_{\mathbf{w},b}\left(1/2\left|\left|\mathbf{w}\right|\right|^2+C\sum_{\iota=1}^{w}\varepsilon_\iota\right), \\ \text{subject to: } &x_\iota^j(\mathbf{w}\cdot\mathbf{y}_\iota^j+b)\geq 1-\varepsilon_\iota, \text{ and} \\ &\varepsilon_\iota\geq 0 \end{aligned}$$

where b (a real number) is the threshold and $\mathbf{w}$ (a vector) defines a $w-1$ dimensional hyperplane. The hyperplane is subject to the penalty of classifying admixed individual ι incorrectly, proportional to the slack variables ε_ι, thus, C is a constant [4, 125]. The population of every haploid genome ι in window j, x_ι^j is identified based on the decision boundary: $g(\mathbf{y}_\iota^j)=x_\iota^j(\mathbf{w}\cdot\mathbf{y}_\iota^j+b)$. If $K'=2$, the inference problem is a binary classification such that, $x_\iota^j\in\{-1,1\}\equiv\{1,2\}$. However, if $K'>2$, then we have a multi-class classification problem resulting in testing ${}^{K'(K'-1)}/_2$ classifiers. The optimal classifier is given by the most common ancestry assignment.

2.3.5.1 HMMs smoothing and support vector machines

The outputs of SVMs are fed into the HMMs consisting of K' hidden and outputs states. Since at every generation the genetic material breaks and rejoins, recombination points follow a Poisson process with a mean proportional to the product of the admixture generations G and the genetic distance between consecutive windows (d_w in Morgans). Therefore, Omberg et al. [125] defined the probability of transition between hidden states, p_w as follows

$$p_w = \frac{1-\exp(-Gd_w)}{(K'-1)},$$

and, modelled emission probabilities as follows

$$P(x_\iota^j|j)=\begin{cases} p & \text{if state is unobserved;} \\ \frac{(1-p)}{(K'-1)}, & \text{if any other state} \end{cases}$$

where p measures how successful the SVM is based on the cross validation process [125].

2.3.6 Conditional random fields

Generative approaches model the posterior probability by modelling the joint distribution between inputs and outputs. However, it has been shown that in classification problems, generative approaches are limited given large dimensional inputs and features with complex dependencies [206]. Such limitations are addressed by discriminative approaches such as logistic regression and conditional random fields. In the local ancestry deconvolution framework, conditional random fields are used by the RFMIX [129] model. RFMIX divides the admixed individual genome into W contiguous windows, each of length d centiMorgans (cM), based on the SNP genetic position [4]. Given $2N_1$ admixed and ancestral haplotypes, where N_1 is the count of individuals, let K be the count of contributing populations, H be a $2N_1 \times W$ matrix of haplotypes, then a sequence of alleles $H_{\iota w} = (H_{\iota w}^{(1)}, H_{\iota w}^{(2)}, \ldots, H_{\iota w}^{(s_w)})$ is the ι^{th} haplotype in window w consisting of s_w SNPs. The local ancestry of these haplotypes is a $2N_1 \times W$ matrix denoted by X, where $x_{\iota w}$ is the local ancestry of the ι^{th} haplotype in window w. We define the haplotype structure and local ancestry of the entire haplotype by $H_{\iota *}$ and $x_{\iota *}$, respectively. Now, the probability of local ancestry given the haplotype structure is calculated by applying a linear conditional random field parametrised with random forests as follows

$$P(x_{\iota *}|H_{\iota *}) = \frac{1}{Z(H_{\iota *})} \exp\left(\sum_{w=1}^{W} \sum_{k=1}^{K} \sum_{h \in H_w} \Theta_{wkh}^{X} I(x_{\iota w} = k) I(H_{\iota w} = h) + \sum_{p=1}^{W-1} \sum_{\ell=1}^{K} \sum_{m=1}^{K} \Theta_{p\ell m}^{Z} I(x_{\iota p} = \ell) I(H_{\iota(p+1)} = m) \right) \tag{2.3.24}$$

where

$$Z(H_{\iota *}) = \sum_{H_{\iota *}} \exp\left(\sum_{w=1}^{W} \sum_{k=1}^{K} \sum_{h \in H_w} \Theta_{wkh}^{X} I(x_{\iota w} = k) I(H_{\iota w} = h) + \sum_{p=1}^{W-1} \sum_{\ell=1}^{K} \sum_{m=1}^{K} \Theta_{p\ell m}^{Z} I(x_{\iota p} = \ell) I(H_{\iota(p+1)} = m) \right)$$

is a normalisation, $I(a = b)$ is the indicator function and the parameter $\Theta^X = \ln(P(x_{\iota w} = k | H_{\iota w} = h))$ is learnt when random forests are trained on source panels. On the other hand, $\Theta^Z = \ln(P(x_{\iota p} = \ell, H_{\iota(p+1)} = m))$ is obtained from the admixture model as in Equation (2.2.1) implemented in Falush et al. [78]. The inference algorithms for linear conditional random fields are analogous to HMMs. Hence, the viterbi or forward-backward algorithms are used to obtain the maximum posterior estimation or smoothe [129].

2.3.7 Non-probabilistic, dynamic programming approach

As aforementioned in **Section 2.2**, most existing local ancestry deconvolution models rely on biological and statistical parameters. In addition, they are designed to deconvolve ancestry in humans. However, this is not the case for LOTER [56], a non-probabilistic optimisation model. Due to the improvements in haplotype inference algorithms, LOTER assumes haplotypes are known. Hence, given a count of source population k individuals n_k, we have a total of $2n = 2\sum_{k=1}^{K} n_k$ ancestral haplotypes. Now, we consider a sequence of ancestral haplotypes $(H_1, \ldots, H_{2n})$ with the i^{th} haplotype at SNP t denoted by H_{it}, where $H_i \in \{0,1\}$. Denote an admixed haplotype by h which is estimated from a sequence of haplotype labels $(c_1, c_2, \ldots, c_T)$. An admixed haplotype at t, h_t is a copy of source population k haplotype (H_{it}), if the haplotype label $c_t = k$. The aim is to identify a haplotype label sequence $(c_1, c_2, \ldots, c_T)$, that minimises $f(c_1, \ldots, c_T)$. That is,

$$\min_{(c_1,c_2,\ldots,c_T)} f(c_1,\ldots,c_T) \;=\; \min_{(c_1,c_2,\ldots,c_T)} \left(\sum_{t=1}^{T} |h_t - H_{c_t t}| + \lambda \sum_{t=1}^{T-1} I(c_t \neq c_{t-1}) \right) \tag{2.3.25}$$

where λ is a tuning parameter,

$$|h_t - H_{c_t t}| = \begin{cases} 1 & \text{if the admixed haplotype is not copied from } H_{c_t t} \\ 0 & \text{otherwise} \end{cases}$$

is the loss function and $(c_1, \ldots, c_T) \in \{1, 2, \ldots, 2n\}^T$. Therefore, the first term on the right hand side of Equation (2.3.25) sums up the loss function over all SNPs and the second term is proportional to the count of transitions between ancestral haplotypes [56]. Although Equations (2.3.23) and (2.3.25) are optimisation problems, the former uses quantile regression while the latter uses dynamic programming to solve the problem involving T SNPs based on an optimal solution of $T-1$ SNPs. At SNP t and $t-1$, an admixed haplotype can either copy the same ancestry, so that

$$f(c_1, \ldots, c_t) = f(c_1, \ldots, c_{t-1}) + |h_t - H_{c_{t-1} t}|,$$

or different ancestries, so that

$$f(c_1, \ldots, c_t) = f(c_1, \ldots, c_{t-1}) + |h_t - H_{c_t t}| + \lambda.$$

By determining the shortest path from the beginning to the end of the chromosome, we obtain the optimal sequence of labels $(c_1, c_2, \ldots, c_t)$ and hence local ancestry given that each parental haplotype is part of one of the ancestral populations [56].

2.4 Current challenges and opportunities in local ancestry inference

Regardless of the existence of several studies that aim to improve the local ancestry inference estimates, several challenges still exist [4]. These challenges heavily affect the accuracy of estimates, and hence further related applications, and have been highlighted in several evaluation studies. Overall, previous studies question the accuracy of the local ancestry estimates in multi-way admixtures (**Tables B.1** and **B.2**). For instance, using three local ancestry deconvolution models, Chen et al. [104] highlighted the inaccuracies of local ancestry estimates in multi-way admixtures. In this study, the local ancestry estimates of the two leading models (both LD-based: LAMP-LD and MULTIMIX) were different on nearly 20% of the studied SNPs (**Table B.1**). This might have emanated from the biological or statistical parameters of each model [4]. Both models require the window size to infer ancestry. Nonetheless, the performance of LAMP-LD did not change given different window sizes. The window sizes tested comprised of 50, 75, 100 and 150 SNPs. Contrastingly, as the window length decreases, the performance of MULTIMIX improved [4, 152]. Furthermore, another separate study, based on a different dataset and models showed that inaccuracies in local ancestry estimates limit admixture mapping [66]. But still, the two models used in this study belong to different categories (**Section 2.2**). LAMP-LD is an admixture-background LD-based model (**Section 2.2.1.2**) which uses ancestral haplotype information, while WINPOP is a non-LD-based model (**Section 2.2.2**) which uses ancestral allele frequencies to infer local ancestry [4].

Previously, deviations in local ancestry have been shown to occur due to inaccuracies in local ancestry inference which could be a result of the signals of recent (post-admixture) selection [9, 116, 119, 120], the disease, calling the true ancestry incorrectly or errors in genotyping [66, 104]. Inaccurate local ancestry estimates have affected admixed population studies [4] and admixture mapping application [66]. Considering current local ancestry inference models, existing challenges can be summarised as follows:

1. To estimate local ancestry, most existing models require biological or statistical parameters, however, these are not always accurate if provided [4].

2. Most existing models suppose that the information about the contributing populations is known [4]. This is not always the case since some populations have few individuals that have been sampled.

3. Most existing models assume markers are unlinked, which is often not true.

4. Most existing models have been designed for recent admixtures. Thus, in general, may not perform well in ancient admixtures.

5. Genotypes take three values only (0, 1 or 2), which may reduce the power in local ancestry inference, as genotypes do not provide information on how close an admixed segment may be to a particular ancestral population at a given SNP.

6. Most existing models are HMM-based, and so they tend to have a large parameter set.

7. Existing models are complex, and their complexities are revealed in the unique inputs and outputs of each model. This challenges the testing of different models and hence hinders the local ancestry inference process and its applications.

8. Existing models are benchmarked for three-way admixtures.

9. Although the two most popular hypotheses on modern human origins agree that all populations have common origins, almost all existing models assume ancestral populations are independent.

10. Although it was shown that existing models fail to accurately infer ancestry at particular regions which might be targets of natural selection, currently, in multi-way admixtures, none of the existing models account for natural selection in local ancestry.

Among these challenges, some have been partly addressed, while others are yet to be addressed. This is the case for the 1^{st} and 4^{th} challenges which have been recently addressed by LOTER. It requires neither biological nor statistical parameters. It has been shown to perform well in ancient admixtures (or admixed populations that started the mating process over 150 generations ago) [56]. However, LOTER is challenged by large computational time due to the optimisation problem [4]. To address the 2^{nd} challenge, SWITCH can estimate local ancestry given the genotype data of one of the ancestral populations. This is done by estimating the allele frequency of the other ancestral population based on the genotypic information of the known population [4]. Nonetheless, this is limited to two-way admixtures. Contrastingly, to address the 2^{nd} challenge, RFMIX uses the information contained in the admixed individuals. Nevertheless, ancestral population information is still required in order to estimate local ancestry.

In order to address the 5^{th} challenge, the EILA model uses a numerical score that takes values between 0 and 1. The score measures the distance between the admixed and ancestral SNPs. Also, the same model use all the genotyped SNPs, relaxing the linkage equilibrium assumption listed in the 3^{rd} challenge. Unfortunately, this model has only been tested for up-to three-way admixtures (see the 8^{th} challenge), and has not been recently applied. Unlike other HMM-based models, the SWITCH model attempted to address the 6^{th} challenge by making use of the LAMP model to initialise parameters in HMMs. However, possibly due to the complexity of modern admixtures and the pairwise HMMs used in SWITCH, the SWITCH model has become obsolete, and thus has not been recently applied (see **Tables B.1** and **B.2**). So far, challenges 8–10 are yet to be addressed.

In this thesis, we tackle the 8^{th}, 9^{th} and 10^{th} challenges. Due to model complexities, existing models may fail to facilitate the local ancestry inference process. Therefore, in **Chapter 3** we develop a framework that manipulates tool-specific inputs, deconvolves local ancestry using existing multi-way admixture models and standardises output results [5]. This facilitates the use of existing models in local ancestry deconvolution and estimate applications, and paves the way for model assessment prior to application, handling the 7^{th} challenge related to model complexities. It is important to note that this framework also addresses the 8^{th} challenge making it possible to estimate ancestry in admixed populations formed by an arbitrary number of ancestral populations, where applicable. In **Chapter 4**, we assess existing models in different admixture scenarios including the five-way admixed populations. Lastly, although mSPECTRUM addresses the 9^{th} challenge it neither accounts for state persistence nor post-admixture selection when inferring local ancestry in multi-way admixed individuals. Based on this, to address challenges 9 and 10, we suggest a modification of mSPECTRUM to account for post-admixture selection and possible relationships that may exist between modern humans in admixed genomes in **Chapter 5**.

2.5 Summary

Over the past two decades, local ancestry inference has been the topic of interest for many researchers [4, 56, 78, 80, 97]. We note that over 20 multi-way local ancestry deconvolution models or tools exist (see **Figure 2.1** and **Table B.3** for a snapshot of the existing multi-way admixture models). In this chapter, we gave both a theoretical and mathematical overview of the local ancestry deconvolution models, including the assumptions of each model. Existing models have not fully facilitated admixture analyses in multi-faceted admixed populations due to the complexity of the admixture dynamics [4, 18]. An illustration of the multi-faceted

admixed population is the mixed ancestry of South Africa, a complex multi-way admixture characterised by the complex admixture dynamics caused by South Africa's colonisation history and its geographical location with regards to slave trade routes [18]. Also, populations with some individuals that could have experienced post-admixture selection is another case of multi-faceted admixture [4, 66]. However, evaluations of local ancestry models have been contextual, based on different performance measures, datasets and frameworks [5] (**Table B.1**). Thus, we design a framework that paves the way for local ancestry inference model assessment in **Chapter 3**.

Chapter 3

Integrating multi-way local ancestry inference tools

3.1 Introduction

Local ancestry deconvolution estimates the number of alleles inherited from a particular population at every chromosomal locus [207]. Local ancestry deconvolution models have been introduced to handle challenges that previous models faced [5]. As such, model performance can be affected by a dataset and factors related to the main modelling effort. This means that for a given dataset, various models should be tested to identify the one that best fits it. However, each existing model has been implemented independently, requiring unique inputs and producing outputs in formats which may not be comparable to the model outputs of the other. As a result, existing models do not facilitate performance evaluation analyses [5].

Considering the model complexities, one requires both analytical and computational skills to fully understand existing local ancestry inference models, manipulate model inputs and analyse outputs [5]. However, such skills are not always available for many biomedical researchers. Hence, the issues associated with manipulating inputs and standardising outputs of existing local ancestry deconvolution tools have negatively affected the local ancestry inference process [5]. This highlights the need for a tool to integrate existing tools, easing the local ancestry inference process and paving the way for selecting an appropriate model based on testing different models [5].

In this chapter, we introduce a tool which integrates local ancestry models to analyse populations formed by the admixture of three or more previously isolated populations; we name

the tool FRANC [5]. It "manipulates inputs, infers ancestry and standardises outputs to user-specified format for eight existing multi-way admixture models" [5]. Apart from easing the local ancestry process through allowing users to choose output formats appropriate for the intended application studies, FRANC also enables researchers to select an optimal model for their dataset as it can analyse data using more than one model at a time [5].

3.2 Implementing the FRANC interface

We implement FRANC version 19.1 in `Python` ≥ 2.7 on a Linux operating system (OS) [5]. Although we have not tested it on other operating systems, given all the necessary conditions to run it (**Section 3.2.2**), we believe FRANC can run on any computer and OS; thus, it is adaptable [5]. Currently, FRANC integrates the eight most commonly used tools, but is easy to expand to incorporate other existing local ancestry tools [5]. Using a one line command FRANC runs on a local machine or cluster using high performance computing facilities. Although this should be done with appropriate citation of the framework and its components, from http://web.cbio.uct.ac.za/ITGOM/franc, one can freely "copy, distribute, display and make unrestricted non-commercial use of FRANC under the GNU General Public License (https://www.gnu.org/licenses/gpl-3.0.en.html)" [5].

3.2.1 FRANC interface integrated tools

Since most modern human populations are admixed (**Section 2.1**), analysing admixture patterns has become fundamental in genomics research including biomedical applications [56]. Given admixed populations, local ancestry inference models have played important roles in identifying and characterising the ancestral populations of disease risk-related SNPs [208]. Considering there are over 20 local ancestry models, users have a wider choice. Nonetheless, each tool is generally available as individual scripts with unique inputs and output files [5]. To solve this, FRANC [5] incorporates WINPOP [156], LAMP-LD/LAMP-HAP [80], PCADMIX [166], CHROMOPAINTER [155], SUPPORTMIX [125], RFMIX [129], ELAI [97] and LOTER [56]. **Table 3.1** summarises these tools, listed from the oldest to the most recent. Considering the categories discussed in **Section 2.2**, FRANC integrates five LD-based models: LAMP-LD/LAMP-HAP, ELAI, PCADMIX, CHROMOPAINTER and SUPPORTMIX, and three non-LD-based models: WINPOP, RFMIX and LOTER [5].

It is important to note that tools were included based on their popularity in recent estimate

applications, and previously reported exceptional performance. **Table B.2** shows a partial list of recent studies conducted between 2015 and 2019 using the tools, and **Table B.1** lists some of the studies that evaluate existing models. RFMIX, LAMP-LD and ELAI are among the most widely used models [171]. RFMIX is popular in disease mapping studies such as the genome-wide association and admixture mapping [175, 179, 181]. On the other hand, ELAI is the most popular in detecting selection signal applications [209, 210]. Although Guan [97] reported similar performance of ELAI and LAMP-LD in simulated three-way admixtures formed 20 and 100 generations ago, ELAI has been shown to perform better than LAMP-LD in more recent admixtures (< 20 generations). This was also recently reported by Pierron et al. [162]. However, its performance is yet to be evaluated in cases of more ancient admixtures, such as Uyghurs with ~ 129 generations [62].

3.2.2 FRANC software requirements

In addition to Python, FRANC has other requirements that depend on the choice of tool under consideration. For example, CHROMOPAINTER requires the `zlib` software library, which can be installed by:

```
sudo apt-get install zlib1g-dev
```

and the `Perl` programming language (which is also required by LAMP-LD). LOTER on the other hand requires the `scikit allel` package, while SUPPORTMIX requires the `libpng12.so.0` package and RFMIX requires the `R` programming language. Upon executing the FRANC command and before proceeding with analysis, we use a screen message to prompt the user to check if they have all the software requirements. The non-system requirements for FRANCE include the study population name prefix, the tools to analyse, the working directory and the parameter file. This file is edited according to the user's specifications. We assume organisms are diploids and that the population data is in/convertible to the standardised PLINK bed/bim format (see **Section 3.3.2**). More so, we assume duplicate individuals, individuals with high missing rate and, SNPs with high missing rate, minor allele frequency < 0.05, monomorphic and heterozygous have been removed. To manipulate inputs, standardise outputs and visualise results, PLINK 1.9 [211] and/or Python is used, while EAGLE v2.3 [212] is used for phasing (haplotype inferencing). **Figure 3.1** represents the overall work flow of the FRANC scheme. It consists of two main components: the user and input processing interface [5]. Basically, the parameter file provides the input to the user interface for processing where tool-specific inputs are generated and the local ancestry inference process takes place [5]. Although every tool incorporated produces ancestry estimates in their tool-specific output format, estimates are converted to the user-specified output format [5].

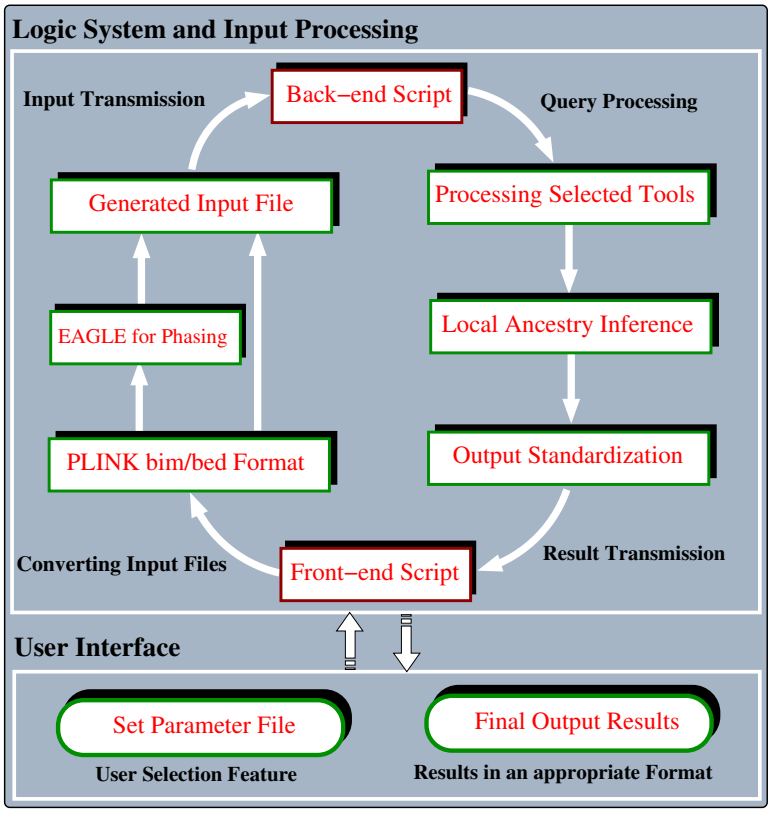


Figure 3.1: Overall work flow of the FRANC scheme.

3.2.3 FRANC usage

As mentioned before, FRANC is free and can be downloaded from http://web.cbio.uct.ac.za/ITGOM/franc. After downloading, the user should extract the files using the Linux command:

tar –xzf franc.tar.gz.

Upon extracting the zipped files, the FRANC scheme consists of three main folders: the *"franc_interface"* which contains interface modules (including `franc.py` and `convertall.py`) and the parameter file; the *"franc_util"* which contains all text files required for running FRANC, that is, *"configs"* and different software implemented in FRANC, *"soft"*; and finally, the *"test_data"* which contains a two-way admixture simulated dataset for testing FRANC as shown in **Figure 3.2**.

To explore the FRANC interface, the user should change to the "franc_interface" directory, using

cd franc/franc_interface

in the Linux terminal. This directory is where the framework is run. For help on running the framework inside the "franc_interface" directory, execute:

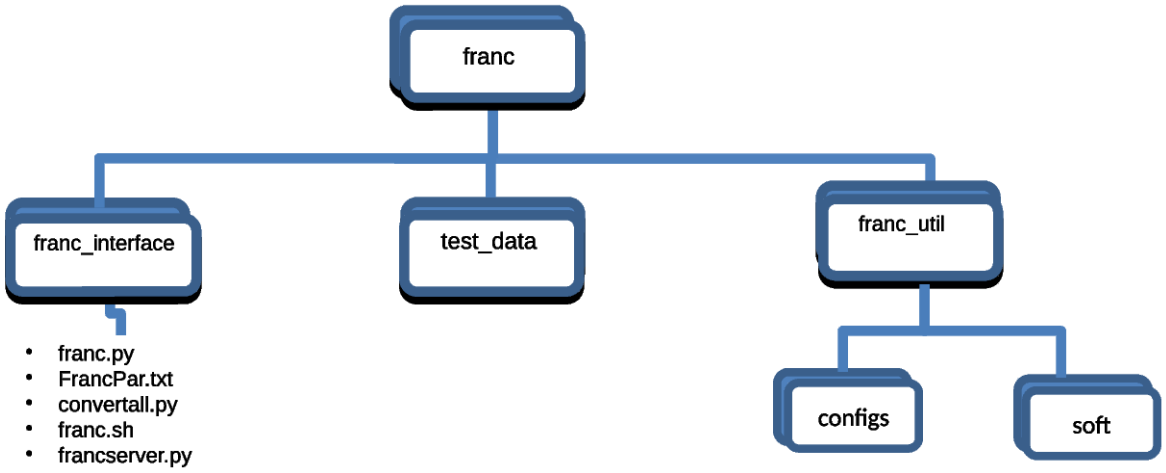


Figure 3.2: Directory structure within FRANC scheme after downloading.

python franc.py -h,

which will display the following on the screen:

```
usage: franc.py [-h] -p FILE [-d FILE] -t FILE -a FILE [-o FILE] [-m MODE]
-f FILE.
```

On this message, the tag -h, --help will show the above help message and exit while other tags display on the screen as in **Table 3.2**. In order to view a short version of the FRANC interface documentation on their screen, the user should use the command:

python setup.py --long-description .

Table 3.2: Tags shown by querying the help command in FRANC.

-p FILE, --par FILE	parameter file containing population-specific and tool-specific information (default: None)
-d FILE, --dir FILE	input data directory (default: current working directory)
-t FILE, --tool FILE	choice of tool(s) to analyse (default: None)
-a FILE, --admix FILE	admixed population name prefix (default: None)
-o FILE, --out FILE	output name prefix (default: None)
-m MODE, --mode MODE	platform in use, that is, desktop or server(default: local)
-f FILE, --outformat FILE	user specified output format (default: None)

3.3 FRANC parameter inputs, running and result output

Just like any other software, FRANC requires some inputs to analyse data. The parameter file and the population data are compulsory when deconvoluting ancestry within the framework. Other required inputs depend on the tool under consideration, these include the genetic (recombination) maps and the reference genetic map. In order to run FRANC, the user should set the parameter file and specify all other input files specific to the analysed samples in one folder. This folder is just like the "test_data" in **Figure 3.2**, and is referred to as *"infolder"*, see **Figure 3.3**. In addition to the population data, the genetic maps and the reference genetic maps, "infolder" contains a folder named after the study population name prefix which is created after executing the one line Linux command, and specified by -a tag. Below are more details on the FRANC inputs.

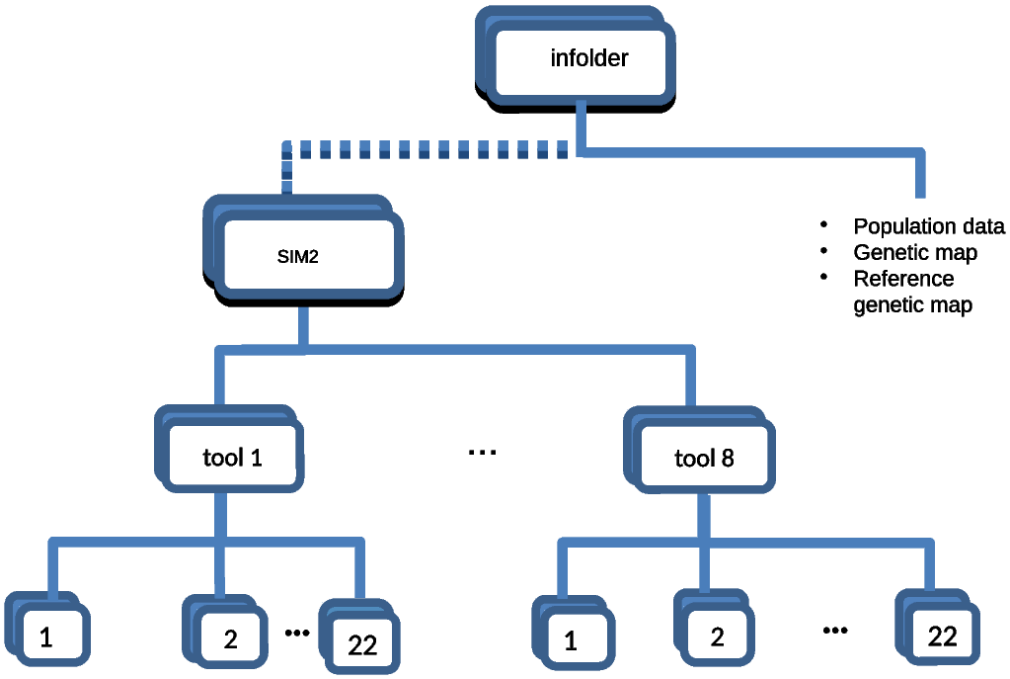


Figure 3.3: The FRANC representation of "infolder" given the user input files are not in the "franc_interface" directory. The -a tag (admixed population name prefix), in this case "SIM2" is created after running FRANC to store the results.

3.3.1 Setting the parameter file

In addition to the parameters specified on the command line (**Table 3.2**), FRANC tool- and population-specific parameters are provided in the parameter file, *FrancPar.txt*. This file is contained in the "franc_interface" directory (**Figure 3.2**). The order of this file should be maintained, and edited according to user specifications. A new parameter file with the same structure can be created, and a path to this file provided after the `-p` tag in the command line when running FRANC. It should be noted that each integrated tool within the FRANC framework requires this file to execute, more details on the contents of `FrancPar.txt` are given in **Table 3.3**. However, not all options in this table are required for running each tool. Nonetheless, each tool requires the ancestral population name prefix (`anc_pop`) and the chromosome number (`CHR`) parameters. The user should provide all ancestral population name prefixes separated by colon, and chromosomes separated by comma. If the user intends to analyse SUPPORTMIX, PCADMIX, or CHROMOPAINTER, they need to provide genetic map files corresponding to each chromosome. Also, if they intend to analyse WINPOP, they are required to provide the ancestral proportions (`anc_prop`), recombination rate (`recombrate`), LD cut-off (`ldcutoff`), `offset` and the admixture generations (`G`) parameters. By default, tool-specific options/parameters are set to default in the FRANC scheme and in the `FrancPar.txt`, before editing, since the tools have been shown to yield optimal results using these parameters.

3.3.2 Population data

Assuming organisms are diploids, the population data is required in the PLINK [211] bed/bim standard format (**Section 3.2**). This consists of three files per population (ancestral or admixed) with suffix `.bim`, `.bed` and `.fam`. For example, given the population parameters, `-a SIM2` and `CHR = 19` in the parameter file, then admixed population data files are `SIM2.19.bim`, `SIM2.19.bed` and `SIM2.19.fam`. It is assumed that data quality control have been performed on the population data files, including, the removal of duplicate individuals and SNPs, monomorphic heterozygous markers, and SNPs and individuals with high missingness. The bed/bim format is advantageous in that it requires less memory and is obtainable easily from the PLINK ped/map files by the command:

--make-bed .

More so, the ped/map files are easy to manipulate from the standard OXFORD format using EIGENSTRAT [90]. The OXFORD format consists of three files per population, that is, `SIM2.19.snp`, `SIM2.19.geno` and `SIM2.19.ind` given the above population parameters.

Table 3.3: Other FRANC parameters edited according to user specifications in the parameter file (`FrancPar.txt`).

1. anc_pop:	a colon separated parameter of ancestral populations for a given admixed population which are required for each run .
2. anc_prop:	average proportion of ancestry contributed to admixed individuals by each ancestral population. A comma separated parameter, required for WINPOP.
3. CHR:	chromosome/s to be analysed. Required for all tools, since each tool runs each chromosome separately.
4. G:	number of generations since the population started mixing. Required for WINPOP, FRANC, SUPPORTMIX and ELAI.
5. gen_map:	as described under the genetic map, **Section 3.3.3**.
6. recombrate:	fixed recombination rate according to units of distance in the bim file, this is required for WINPOP.
7. offset:	determines the overlap between adjacent windows for WINPOP.
8. ldcutoff:	determines the threshold for markers to be considered as independent thus it determines the number of markers to retain after screening for WINPOP.
9. phased:	whether to run ELAI on phased (YES) or unphased data (NO).
10. PopPhased:	determines whether to correct for phase switch errors. Recommended for unrelated individuals in RFMIX.
11. parallel:	determines whether to run FRANC or LOTER in the parallel mode.
12. w:	window size, edited according to user specifications, for SUPPORTMIX, RFMIX, and PCADMIX.

3.3.2.1 The binary (bed) file

This is a file containing information on individual genotypes. It can only be viewed by a Unix xxd command, hence, the first 5 lines of SIM2.19.bed are viewed by

xxd -b SIM2.19.bed — head -5

which should display the following contents on the computer screen:

```
0000000: 01101100 00011011 00000001 11111110 11111011 10111010  l.....
0000006: 11111010 11111111 11111110 10101111 11111110 11111111  ......
000000c: 10111111 11101111 10101010 11111010 11111011 00111111  .....?
```

```
0000012: 11111111 11101111 11111011 11111111 11111011 11111110  ......
0000018: 11111110 10101111 10111110 10111110 00101111 11001111  ..../.
```

where the first three columns after the colon represent the magic number, the SNP-major and the individual major, while the remaining columns contain the genotype information.

3.3.2.2 The bim file

This is a file containing information on the SNPs under study. Each SNP occupies one row and six (6) columns. The columns of this file are: the SNP chromosome number, SNP ID (marker ID), genetic distance, genetic position, allele 1 and allele 2. Below we give the SNP information of the SIM2 population genotyped on five (5) SNPs from chromosome 19.

19	rs8105536	0	212033	A	G
19	rs11084928	0	228776	A	G
19	rs4897940	0	252639	G	A
19	rs4897941	0	252668	C	G
19	rs11883060	0	286055	C	A

3.3.2.3 The family (fam) file

This is a file containing information on individuals, that is, individual family identity (ID), individual ID, paternal ID, maternal ID, sex and disease status. Corresponding to the above `.bim` file is the `.fam` file on three individuals.

1	SIM1	0	0	1	1
2	SIM2	0	0	2	1
3	SIM3	0	0	1	1

3.3.3 Genetic map (recombination map) file

The recombination map file consists of three columns: the physical position, the recombination rate and the genetic position. It is directly downloaded from the HapMap website (https://ftp.ncbi.nlm.nih.gov/hapmap/recombination/) and requires no further processing. As stressed in **Section 3.3.1**, the user should provide one genetic map file for each chromosome to deconvolute ancestry with SUPPORTMIX, PCADMIX, or CHROMOPAINTER. It should be saved in the "infolder" directory as shown in **Figure 3.3**. The file should

be named as `genetic_map_chr#.txt`, where # is the chromosome number. For instance, `genetic_map_chr19.txt`, if analysing chromosome 19. The first six lines of the genetic map file are given by

Position	COMBINED_rate (cM/Mb)	Genetic_Map (cM)
253938	0.2214599891	0
256859	0.2213076426	0.0006464396240346
259772	0.2213078499	0.0012911093907933
260970	0.2215185766	0.0015564886455601
261033	0.2216183478	0.0015704506014715

3.3.4 Reference genetic maps

Apart from WINPOP and ELAI (unphased option), all integrated FRANC tools require the reference genetic map file. It is used for haplotype inference using the EAGLE [212] software. The map is downloadable from https://data.broadinstitute.org/alkesgroup/Eagle/downloads/tables/ and should be stored in "infolder" using the name `genetic_map_phase.txt.gz`. It is noteworthy that users should download the correct reference genetic map build that matches their population data.

3.3.5 Other parameter inputs

Except for the user defined inputs, FRANC comes with some software related inputs, such as those in the "franc_util" directory. These include "configs" and "soft" directories (**Figure 3.2**). The "configs" directory contains text files (with suffix `.txt`) that are useful in manipulating tool-specific inputs for WINPOP, LAMP-LD, RFMIX and SUPPORTMIX. It also contains python files which might be useful in converting files from the OXFORD to the PLINK ped format. The order and structure of such files including, `configRfmix.txt`, `configPedGeno.txt`, `configsvm.txt`, `par_winpop.txt` should be maintained. While, `configGenoPed.py` and `configGenoPed.txt` have to be modified according to the user specifications of input data files and ancestral populations in order to convert from the OXFORD to the PLINK format. Also accompanying the FRANC scheme is `franc.sh`, contained in the "franc_interface" directory (**Figure 3.2**). `franc.sh` is a shell script edited according to the server details in the event that the user intends to perform the analysis using high-performance computing (`-m server` option).

3.3.6 Running FRANC

After setting the parameter file and specifying user-defined inputs in the "infolder" including population data (**Section 3.3.2**), recombination maps and reference genetic map (for tools that require these), the following one line command runs FRANC:
python franc.py -p FrancPar.txt -d infolder -t tool -a admixed_pop_prefix_name -o outfile_prefix -m local -f output_format.

When running two or more tools provided after the -t tag, separate the tools by a colon. For example, to deconvolute ancestry with ELAI, WINPOP and LOTER in the framework, we have -t elai:winpop:loter. Running the above FRANC command will show an interactive interface orienting users of the systems, platforms and/or library requirements of the framework. If all system requirements are satisfied, a 1 should be entered to proceed, else a 2 should be entered to exit. If the -m tag is not specified in the command line, FRANC runs locally. Therefore, users running the framework using the high-performance computing platform should always specify the -m server tag after providing the server details in the `franc.sh` file in the "franc_interface" directory (**Figure 3.2**). It is worth mentioning that modules required in running the selected tools are loaded either in this shell script (`franc.sh`) or in a separate file, in this case, `config.bash`, contained in the "infolder" directory, executed in the `franc.sh` file using the following `bash` command:

. infolder/config.bash

3.3.7 FRANC outputs

Upon executing the FRANC command in **Section 3.3.6**, the FRANC output directory named after the admixed population name prefix ("admixed_pop_prefix_name"), is created within "infolder" to store the results from the tool(s) under study (**Figure 3.3**). If the "admixed population name prefix" directory exists, results are added onto its contents. This directory will contain sub-directories named after the tool(s) under consideration (names are in lower cases). Each tool sub-directory will contain the chromosomes under consideration sub-directories, specified in the parameter file (**Figure 3.3**). FRANC outputs depend greatly on user-defined specifications. Apart from the tool-specific output format, FRANC eases the local ancestry process and potential local ancestry estimate applications by standardising results to the RFMIX, LOTER, WINPOP and LAIT [213] output formats (see **Section 3.5.1**, for more explanation on the standard output formats for FRANC) [5]. The final local ancestry results in the user-specified format (specified by -f tag) are stored in their respective chromosomes,

under the `output name prefix` specified by the `-o` tag in the command line, with suffix containing the `CHR`, `-t`, and `-f.txt`. As aforementioned, `CHR` is specified in the parameter file. **Figure 3.4** shows the FRANC directory representation when "infolder" is the "franc" interface. Now, to analyse individuals belonging to SIM2, the two-way admixed population on chromosome 19 using RFMIX when results are standardised to the WINPOP format, and output name prefix specified by AA, yields results in `infolder/SIM2/rfmix/19/`, under the name `AA.19.rfmixwinpop.txt`.

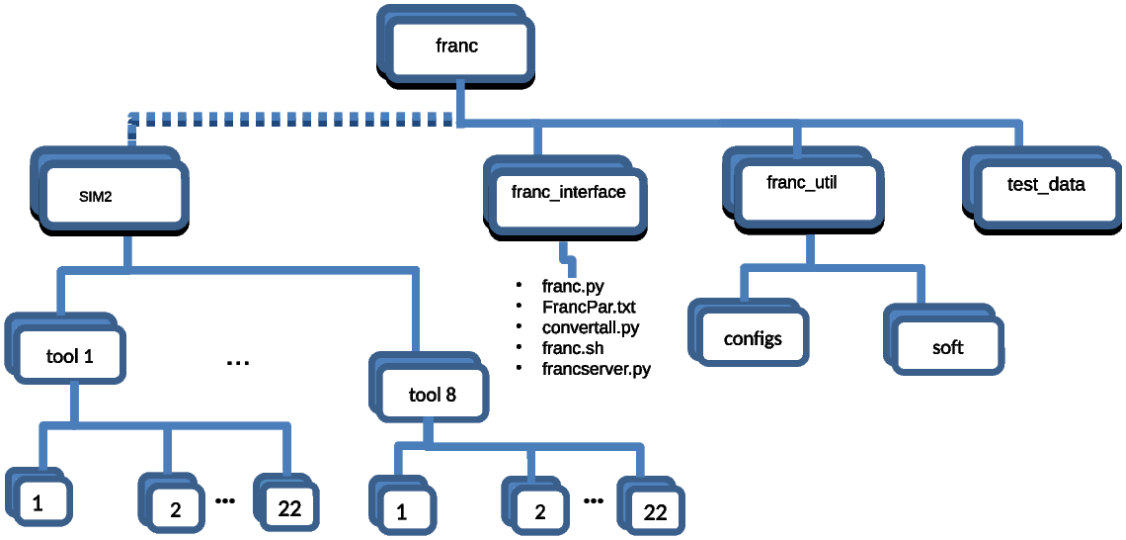


Figure 3.4: The FRANC folder representation when the user-defined inputs are in the "franc" interface directory, that is, `-d` tag is the "franc" directory, assuming all user-defined inputs are moved to the "franc" directory. The `admix_name_prefix`, "SIM2" is created in the "franc" directory upon running FRANC, to store the output results of the executed tool(s).

3.4 Illustrating the eight tools integrated in FRANC

We illustrate how to implement the eight tools integrated in FRANC by analysing simulated data on two admixed populations that result from the interbreeding of three and five previously isolated populations. Each of the simulations is a hybrid isolation or single-point (SP) admixture model consisting of 200 individuals. As mentioned in **Chapter 1**, a hybrid isolation admixture model assumes all ancestral populations interbreed in a single generation and mating only occurs among admixed individuals thereafter [208]. We give the details of the previously isolated populations used in the simulation of ancestral populations, and the quality control procedures later in **Section 4.2.2**.

3.4.1 Ancestral and admixed simulations

The simulation process requires the haplotype information to be known. As a result, after quality control, haplotype inference and the simulation process are carried out as in **Section 4.2.5**. Now, to simulate the 200 admixed individuals, 400 individuals were randomly selected from expanded populations of 800 individuals. Although more details are given **Table 4.5**, the two admixed populations were generated as follows:

- SIM3G3 - an ancient complex three-way admixture (multi-way) that mimics Latin Americans. The mating process of this population started 600 generations ago. It is complex since the ancestry segments are shorter making ancestry deconvolution difficult.
- SIM5G1 - a recent five-way admixture (multi-way), mimicking the mixed ancestry of South Africa who mixed 15 generations ago. Two of the contributing populations are African (sub-continental): Bantu speaking (YRI) and Khoisan population groups. Thus, it is complex.

3.4.2 Setting the parameter file and running the framework

Chromosome 6 contains the human form of the major histocompatibility complex (MHC) termed the human leukocyte antigens (HLA) [214], which harbours over 200 genes that are associated with diseases and plays important roles in organ transplant [214]. The HLA region is known to bear natural selection signals [215], which have been shown to yield deviations in local ancestry [66, 112, 118]. As such, we believe the accuracy of local ancestry models in this region may be biased; hence, we test the FRANC framework on chromosome 6 of SIM3G3 and SIM5G1. Firstly, we set the parameter file with ancestral populations and proportions of each admixed population (SIM3G3 and SIM5G1). Also we set the following parameters in the parameter file: RFMIX: PopPhased, -w 0.2, -n 5, G; PCADMIX: -w 20, -prune 0; SUPPORTMIX: -w 20, G; WINPOP: ldcutoff=0.1, offset=0.2, recombination=10^{-8}, G; CHROMOPAINTER V2: -ip -b and ELAI: G, -C K, -c 5×K; where G is the admixture generations and w is the window size [5].

Secondly, we estimate local ancestry using the eight tools integrated in the framework using high-performance computing platform from the University of Cape Town's ICTS High Performance Computing team, http://hpc.uct.ac.za/, and the Centre for High Performance Computing (CHPC), South Africa, https://chpc.ac.za/ [5]. In the integrated framework, the number of HMM states for LAMP-LD is fixed to a default value of 50. Each admixed popula-

tion is analysed separately by a single command. For example, to estimate the local ancestry for SIM3G3 individuals using all FRANC integrated tools, we use the command:
python franc.py -d /mnt/lustre/egeza/FRANC/SIM5/SINGLEPOINT/PLAIN/G1/ -t winpop:pcadmix:rfmix:supportmix:elai:lampld:loter:chromopainter -p FrancPar.txt -a SIM3G3 -f winpop -o SIM3G3.

3.5 Results

In genomics and biomedical research, it is important to identify and characterise the local ancestry of SNPs that confer increased disease susceptibility to specific traits or phenotypes [208]. Today, over 20 local ancestry inference tools exist [2, 4]. However, these require inputs that are not always easy to manipulate, hence, making it challenging for the genomic community to access them. We test the FRANC interface on ancient three-way (**Figure 3.5**) and recent five-way admixed individuals using simulated data (**Figure 3.6**). The plots show the actual and predicted ancestry copies of every chromosomal segment of admixed individuals. Based on these plots, we note that the most appropriate model to infer local ancestry in the three-way admixed individual formed 600 generations ago was LAMPL-LD, while for the five-way admixed individual that was formed recently it was LAMP-LD and ELAI. A detailed evaluation of such tools is provided and discussed in **Chapter 4**.

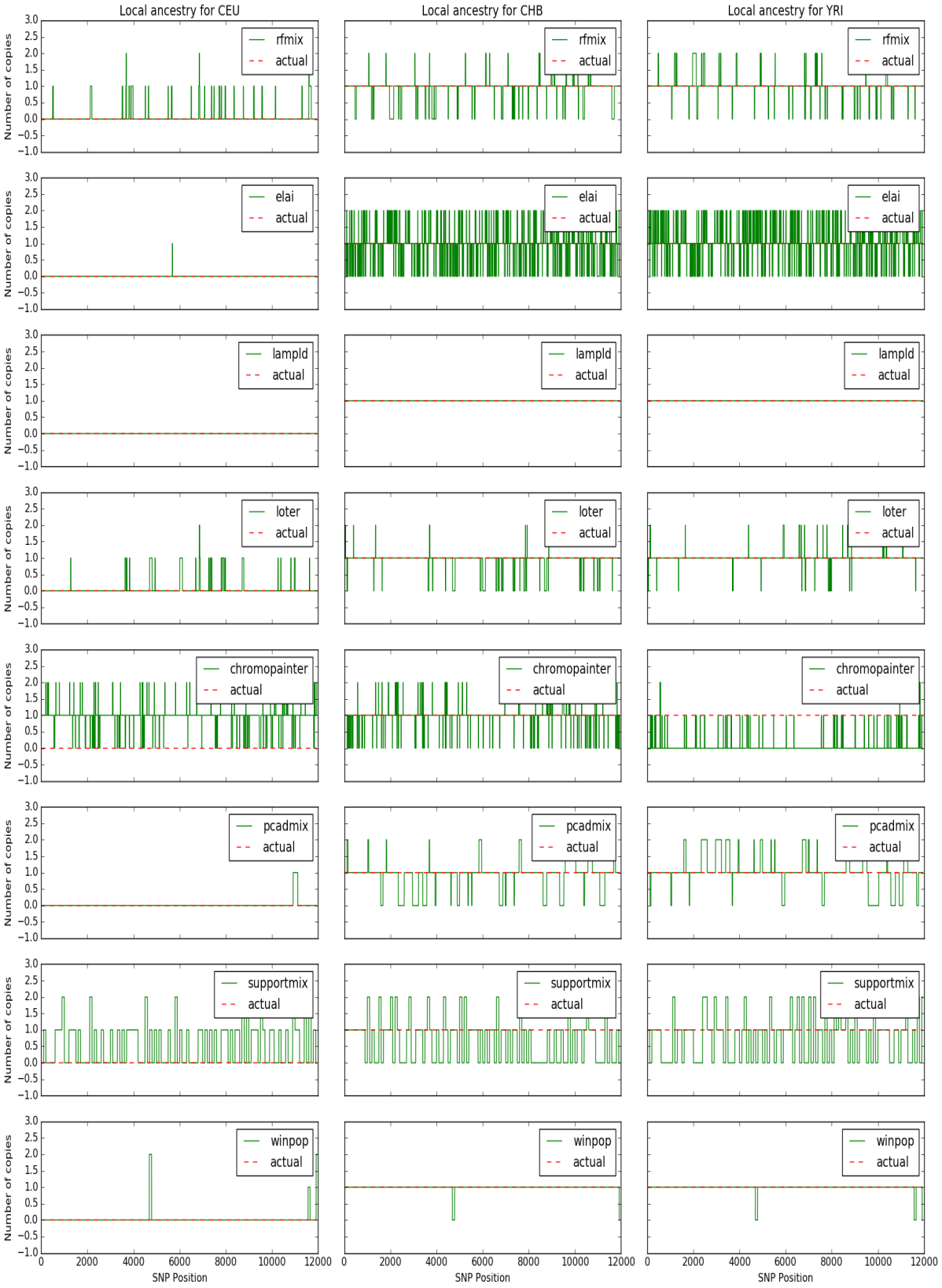


Figure 3.5: Count of CEU, CHB and YRI ancestry copies (column-wise) as approximated by RFMIX, ELAI, LAMP-LD, LOTER, CHROMOPAINTER, PCADMIX, SUPPORTMIX and WINPOP (row-wise), respectively for a SIM3G3 individual (Source: Geza et al. [5]).

3.5.1 FRANC output conversion

As discussed previously, FRANC gives users different choices of output formats which eases further estimate applications. It also opens the door for evaluating a broad range of existing multi-way local ancestry models, thus selecting a more suitable model based on a given application. Hence, the choice of the output format depends on the user's needs. As such, one can infer ancestry using a given tool yet the output is obtained in a another tool format specified according to the user's needs. We also incorporate the LAIT standard output format within the framework in case users would like this output format given two- and three-way admixtures only [5]. **Table 3.4** summarises the standard output formats which are obtainable upon running the eight state-of-the-art/leading edge tools within the FRANC framework. The diagonal of the table is always possible since the framework first produces tool-specific outputs and standardises them to user-specific output.

It is noteworthy that, regardless of the executed tool, it is always possible to obtain the WINPOP [156] output format. WINPOP local ancestry estimates correspond to the fraction of alleles passed down from a contributing population at every SNP. Its estimates are 0.0, 0.5 or 1.0, and such results can not be standardised into RFMIX or LOTER output formats, since, the ancestry information of every haplotype at every SNP is unknown. On a separate note, ELAI is run with either phased or unphased data (**Section 3.3.4**). Its local ancestry estimates are dosages, which correspond to the count of alleles from a particular population at a SNP. Dosage values range between 0 and 2 for unphased data, and between 0 and 1 for phased data. As a result, only ELAI results under the phased option (in the parameter file) can be standardised into the RFMIX and LOTER formats. If instead the unphased option is specified, outputs can only be obtained in the WINPOP or LAIT format.

Since the outputs format might not facilitate potential local ancestry estimate applications, it is less important to convert the ancestry of a haplotype at every SNP to the ancestry of a haplotype within every window. Hence, outputs cannot be standardised to SUPPORTMIX, PCADMIX, and LAMP-LD. Similarly, the CHROMOPAINTER output format is not part of the standard format because estimates are probabilities of inheriting a particular ancestry at a SNP. These are usually in a file which contains haplotypes in rows, each haplotype occupying $T+1$ rows, where T counts all analysed SNPs. This format requires large memory and sound computational skills to decode local ancestry estimates.

Table 3.1: The eight leading edge local ancestry tools within the FRANC framework. Columns denote the: name of the tool, ancestral and admixed population data format, input files required, statistical/biological parameters required, the ability (✓) or inability (✗) of the tool to model LD, and the tool reference paper. frq and configs stands for frequencies and configurations.

Tool	Ancestral (anc) data	Admixed (admix) data	Files required	Biological/statistical parameters	A/c LD	Reference paper
WINPOP	allele frq	genotypes	anc frq, admix genotypes position and config file	anc proportions, LD cutoff admixture generations, recombination rate	✗	[156]
CHROMOPAINTER	haplotypes	haplotypes	merged anc and admix haplotypes, population label genetic map, donor list	mutation and recombination rate	✓	[155]
SUPPORTMIX	haplotypes	haplotypes	anc and admix haplotypes, genetic map, and config file	window length and admixture generations	✓	[125]
PCADMIX	haplotypes	haplotypes	anc and admix haplotypes, genetic map	window size	✓	[166]
LAMP-LD	haplotypes	genotypes	anc haplotypes, admix genotypes and position	window size and hidden states number	✓	[80]
RFMIX	haplotypes	haplotypes	anc and admix haplotypes and genetic map	window size and admixture generations	✗	[129]
ELAI	genotypes/ haplotypes	genotypes/ haplotypes	anc and admix genotypes/ haplotypes	lower and upper cluster admixture generations	✓	[97]
LOTER	haplotypes	haplotypes	anc and admix haplotypes	-	✗	[56]

Source Geza et al. [5]

Figure 3.6: Count of CEU, CHB, GIH, KHS and YRI ancestry copies (column-wise) as approximated by RFMIX, ELAI, LAMP-LD, LOTER, CHROMOPAINTER, PCADMIX, SUPPORTMIX and WINPOP (row-wise), respectively for a SIM5G1 individual (Source: Geza et al. [5]).

Table 3.4: FRANC output conversion table. Entries ✓and ✗, represents the possibility and impossibility of standardising the outputs of the tool in row to the output format of the tool in column, respectively.

TOOL	RFMIX	LOTER	PCADMIX	SUPPORTMIX	CHROMOPAINTER	WINPOP	LAMP-LD	ELAI	LAIT
RFMIX	✓	✓	✗	✗	✗	✓	✗	✗	✓
LOTER	✓	✓	✗	✗	✗	✓	✗	✗	✓
PCADMIX	✓	✓	✓	✗	✗	✓	✗	✗	✓
SUPPORTMIX	✓	✓	✗	✓	✗	✓	✗	✗	✓
CHROMOPAINTER	✓	✓	✗	✗	✓	✓	✗	✗	✓
WINPOP	✗	✗	✗	✗	✗	✓	✗	✗	✓
LAMP-LD	✓	✓	✗	✗	✗	✓	✓	✗	✓
ELAI	✓	✓	✗	✗	✗	✓	✗	✓	✓

3.6 Discussion

We stress that the FRANC framework neither address model discrepancies including those resulting from relaxing certain model assumptions nor shortfalls, but highlights other tool that perform better than a given model for a given admixture scenario. For example, using FRANC to deconvolve local ancestry with WINPOP does not improve estimates accuracy given distant ancestral populations, but provides alternatives including LAMP-LD. Furthermore, it does not address the challenge of deviations in average local ancestry at particular regions which could be due to unaccounted for natural selection.

3.6.1 FRANC vs existing local ancestry inference frameworks

We present an existing suchlike scheme, LAIT [213], local ancestry inference toolkit, which integrates three multi-way models: LAMP, LAMP-LD/LAMP-HAP, ELAI, and HAPMIX, a two-way admixture model. Similarly to FRANC, LAIT requires the standardised PLINK bim/bed format to run every tool individually, but outputs results in a single standardised LAIT format. However, it is important to note that, identical to the LAIT software, FRANC standardises results to the LAIT format for two- and three-way admixed populations only thus promoting further estimates application and lay the foundations for tool assessment. Unlike LAIT, FRANC is benchmarked on complex five-way admixed individuals mimicking the mixed ancestry of South Africa population. We compare and contrast LAIT and FRANC in **Table** 3.5.

Table 3.5: Comparing FRANC to the LAIT toolkit.

Description	**LAIT**	**FRANC**
Inputs format	PLINK bim/bed	PLINK bim/bed
Tools incorporated	LAMP, LAMP-LD, ELAI and HAPMIX	WINPOP, LAMP-LD, ELAI, SUPPORTMIX PCADMIX, RFMIX, LOTER and CHROMOPAINTER
Outputs format	LAIT standardised	LAIT standardised, WINPOP, RFMIX and LOTER

Currently, FRANC depends on the two- and three-way admixture analysis `Perl` implemented LAIT [213] tool. This may be insufficient considering modern humans such as the SAC population which is five-way. As such, future FRANC versions will generalise the LAIT output format to accommodate more than three-way admixtures.

3.6.2 Pinpointing tools currently not integrated in FRANC

Although the FRANC framework integrates state-of-the-art tools, we inform users about other multi-way local ancestry deconvolution tools that use high density SNPs data (described in **Chapter 2**), yet are not part of the current version, FRANC v19.1. These are: EILA [159], ALLOY [158], mSPECTRUM [165] and MOSAIC [171]. As highlighted in **Chapter 2**, mSPECTRUM is a Bayesian nonparametric model which, in contrast to other existing models, assumes ancestral populations are not independent. mSPECTRUM outperformed HAPMIX, a leading edge two-way admixture model, however, mSPECTRUM is not easily accessible to the genomics community as it is not available online. Although EILA is based on a sophisticated technique, quantile regression, that tends to relax the regression slope assumption [216], in addition to being benchmarked for three-way admixtures, it has been rare in recent applications (**Table B.2**). This is also the case for ALLOY which accounts for background LD using non-homogeneous variable length Markov chains. Contrastingly, MULTIMIX has recently been applied and evaluated [102, 104], however, when we tested it on multi-way admixtures, it resulted in a segmentation fault in most chromosomes. Hence, its exclusion from the current FRANC version. Fortunately, the framework includes ELAI which, just like MULTIMIX, is flexible on data type availability.

FRANC also excludes MOSAIC a more recent model/tool whose accuracy is comparable with or exceeds the leading existing tools [171]. Similarly to SUPPORTMIX, MOSAIC requires labelled donor population data (for example, the Luhya in Webuye, Kenya and Gambians in Western Divisions, Gambia, and Utah Residents with Northern and Western European Ancestry), but does not require knowledge on which donor relates to which ancestry (that is, which one is African: AFR or European: CEU). Unlike other existing tools, MOSAIC also produces results on the genetic distance between ancestral groups and each panel and the co-ancestry curve, providing insights into admixture dates [171]. However, MOSAIC could not be included as it was proposed during the time of writing after the implementation of FRANC. Future versions would incorporate these tools.

In order to keep the framework up to date with local ancestry inference applications, the tool will be updated every six months. Currently, when analysing more than one tool, FRANC sequentially runs each chromosome per tool. Thus, we ought to fully utilise the high performance computing (server) option by parallelising the analysis and incorporating the pipeline in a workflow management system: nextflow to allow different chromosomes to run at the same time using different allocated resources (CPUs).

3.7 Summary

Although a handful of local ancestry inference models exist, previous studies have shown that existing models fail to fully facilitate the local ancestry inference process and estimate applications. This results from the complexities in inputs and outputs produced by each model. Therefore, to fill this gap, we have developed a portable, flexible and easy to use `Python` framework, which integrates eight state-of-the-art local ancestry inference tools. The framework is expandable to allow for the inclusion of all current and future models, for example, incorporating MOSAIC [171]. This paves the way for model assessment since users can select a tool appropriate to their dataset and/or their potential estimate applications.

Chapter 4

Assessing multi-way admixture models within a unified framework

4.1 Introduction

To date, models have improved to account for recent and ancient admixtures, continental and sub-continental ancestry, and to allow ancestry deconvolution without statistical or biological parameters. Nonetheless, previous studies have reported the deviations that exist in local ancestry at particular regions [4, 9, 66, 104]. This limits the local ancestry estimates applications [4, 66, 104]. Due to the different model assumptions, numerous ancestry estimates applications and diverse population admixture scenarios, as illustrated by studies in **Table B.1**, a model cannot perform well across all datasets [5]. Of particular concern is the fact that existing evaluations are based on different datasets and performance measures. Also, recently proposed models (like LOTER) should be replicated, while models designed based on the same assumptions and applications should be tested within the same framework [4]. Currently, no single study has tested over five models on all autosomal chromosomes given different admixture scenarios. This makes it challenging to guide users on choosing appropriate models for different admixture dynamics [4, 5].

As highlighted by Pasaniuc et al. [119], models should be compared if they fall within the same category; this can be, LD-based or non-LD-based, window-based or non-window-based. As an illustration, Pasaniuc et al. [119] tested three LD-based models: LAMP-LD, PCADMIX, and ALLOY. Whereas Chen et al. [104] compared two LD- and window-based models, LAMP-LD and MULTIMIX. In both studies, LAMP-LD performed better than other models.

In accordance with Pasaniuc et al. [119], Zheng-Bradley and Flicek [102] tested three LD-based models: ELAI, LAMP-LD and MULTIMIX on Mexican children; ELAI performed best followed by LAMP-LD. More recently, ELAI performed better than PCADMIX when applied to an admixed population from Madagascar [162]. In support of the Pasaniuc et al. [119] idea, Yang et al. [159] tested two non-LD-based models: EILA and LAMP on admixtures involving distant ancestral populations. Most recently, Dias-Alves et al. [56], compared two non-LD-based models that require haplotype information to estimate local ancestry, LOTER and RFMIX. While Maples et al. [129] did not conform to Pasaniuc et al. [119] recommendations by comparing RFMIX (a non-LD-based model) to two LD-based models: LAMP-LD and SUPPORTMIX.

This chapter assesses existing multi-way local ancestry inference models proposed between 2003 and 2018. Model selection was based on recent recommendations (including Geza et al. [4] and Baran et al. [80]) and application studies. We use the FRANC framework designed in **Chapter 3** to evaluate WINPOP, PCADMIX, SUPPORTMIX, LAMP-LD, CHROMOPAINTER, RFMIX, ELAI and LOTER. Unlike previous evaluations, we test LD-based and non-LD-based, window-based and non-window-based, and haplotype-based and non-haplotype-based models. Furthermore, the chapter replicates evaluation studies for the LOTER model. We assess the models based on performance measures calculated from the categories of a confusion matrix [217, 218, 219, 220, 221]. Instead of just considering the model accuracy (accuracy metric) and ability to predict ancestry correctly (recall), following the ideas of Chicco and Jurman [218], we also consider the extent to which the model estimates agree with the ground truth using the Mathews correlation coefficient (MCC) and the Kappa performance measures. We believe this provides insights into determining if the local ancestry inference problem is solved in admixed populations that result from the interbreeding of three or more genetically distinct populations. Therefore, the chapter focuses on multi-way admixed individuals that is, the three- and five-way admixed populations. Furthermore, we also apply state-of-the-art models to the real data of the South Africans of mixed ancestry.

4.2 Materials and methods

4.2.1 Ancestral population data description

In order to estimate local ancestry in admixed individuals, existing models rely on ancestral and admixed individual information [56, 78, 80, 97, 129, 152]. For this reason, the reference

ancestral populations used for the inference process should be genetically close to the true ancestral populations of a given admixed population [64, 222]. This section provides the details of the ancestral populations used in simulating different admixed populations and testing existing models on the different admixtures. Although ancestral population donors may be the same, the genome-wide ancestry proportions differ between admixed population individuals; this is the case of Latin Americans [63, 116, 223], a mixture of Europeans, Native Americans and Africans [60, 224]. In this chapter, the ancestral population panels for all the simulated three-way admixtures are individuals whose ancestors include: the northern and western Europeans collected from the Centre d'Etude du Polymorphism Humain (CEPH) represented as CEU; the Han Chinese from Beijing, China represented as CHB and the Yoruba individuals from Ibadan, Nigeria represented as YRI. Based on our earlier discussions (**Section 1.2.2.1**), the mixed ancestry of South Africa is a mixture of five previously isolated populations, the European, the isiXhosa, the Gujarati Indian, the Khoisan and the East Asian. Therefore, for five-way admixed populations we use CEU, CHB, YRI, the Gujarati Indians from Houston (GIH), Texas, USA, and the Khoisan (KHS) individuals from Namibia as proxy ancestral populations. These individuals are used to generate mixed ancestry individuals and assess models on simulated data for three- and five-way (mimicking the mixed ancestry of South Africa, the SAC) admixed populations, and to assess models on the SAC Tuberculosis real data. We use the reference ancestral panels described in **Table 4.1**. This data was used as proxy ancestral populations in Chimusa et al. [18, 66].

Table 4.1: Reference ancestral populations for generating test and train datasets.

Population	Description	Number of individuals	Source
CEU	Northern and western Europeans residing in Utah	165	HapMap3
CHB	Han Chinese in Beijing, China	137	HapMap3
GIH	Gujarati Indians in Houston, Texas, USA	101	HapMap3
KHS	Khoisan individuals from Namibia	24	HGDP
YRI	Yoruba individuals from Ibadan, Nigeria	203	HapMap3

Table 4.2 shows the genetic distances between the reference ancestral panels (Wright's Fst) based on the Weir and Cockerham [8] method. We used all the SNPs in the data, and quality checking was done in PLINK, using the following QC parameters: `--geno 0.1 --mind 0.1 --hwe 0.0000001 --nofounder --allow-no-sex`.

Table 4.2: Estimates of pairwise genetic distances (the Wright's Fst according to [8]) between the reference populations panels

	CEU	CHB	GIH	KHS	YRI
CEU	0.00				
CHB	0.11	0.00			
GIH	0.04	0.07	0.00		
KHS	0.43	0.44	0.42	0.00	
YRI	0.15	0.18	0.14	0.49	0.00

4.2.2 Data quality control procedures

Previous studies have shown that, in addition to post-admixture selection and genetic drift, genotyping errors or systematic biases may cause deviations in local ancestry estimates [113]. Genotyping errors are due to the inclusion of poor quality DNA samples, the variations in DNA sequences, human error and biochemical and equipment artefacts [225, 226]. These errors may bias allele and/or genotype frequencies [227], yielding spurious local ancestry estimates [113] and, impacting the type 1 error and power in association studies [225, 227, 228, 229]. Nevertheless, genotyping errors can be controlled by quality control [225, 226, 229]. Quality checks include, removal of individuals and SNPs with a high missing rate, duplicated individuals, monomorphic, multi-allelic or low minor allele frequency (MAF < 0.05) SNPs and those that deviate from the Hard-Weinberg equilibrium [225, 227, 228]. In this thesis, we use PLINK 1.9 [211] for quality control and to extract common SNPs. The following quality control parameters are used: --geno 0.02 to remove the SNPs with high missingness, --biallelic-only strict to remove monomorphic and multi-allelic SNPs, --maf 0.05 to remove all SNPs with MAF $<$ 0.05 and --mind 0.02 to remove individuals with missingness rate beyond 0.02. **Table 4.3** lists the size of each reference ancestral population panel before and after the quality control procedure, including the information about the remaining SNPs that were common to all populations. It should be noted that three and five individuals were removed from the YRI and the CEU population, respectively, due to high missing rates of genotypes. After quality control, 179 621 SNPs were common to all reference ancestral panels.

4.2.3 Admixture patterns tested

Motivated by the challenges highlighted in **Section 2.4**, we assess the performance of existing models:

Table 4.3: Ancestral population panels before and after the quality control process.

Population	Number of individuals		Number of SNPs		
	Before QC	After QC	Before QC	After QC	Common SNPs
CEU	165	160	272 796	245 148	179 621
CHB	137	137		238 527	
GIH	101	101		250 420	
KHS	24	24		239 194	
YRI	203	200		239 290	

1. On single-point (single-wave) and multi-point (multi-wave) admixtures.
2. In recent and ancient admixtures.
3. As the population complexity increases, that is from three- to five-way.
4. On disease-affected admixed individuals with some SNPs under the effects of post-admixture selection.
5. Given incorrect biological parameters, in this case admixture generations.
6. On skewed and balanced ancestral population panels.

The following sections discuss in detail the simulation process and the admixed populations considered.

4.2.4 The simulation framework

Although several genome-wide simulation frameworks exist, including SimuPOP [230] (a forward-time simulator) and SIMCOAL2 [231] (a coalescence approach), we adopt a re-sampling approach that has been implemented in FractalSIM [208] and a recently published forward-time simulator implemented in AdmixSim 2 [232]. After data cleaning, to simulate the different admixed populations, haplotype inference is done for all population panels given in **Table 4.3** using EAGLE V2.3 [168]. When evaluating local ancestry inference models, the set of individuals used in generating the admixed samples are different from the reference ancestral population panels that are used to assess the models in the generated admixed samples [64, 80, 156, 158, 233]. In this study, we generate 200 admixed individuals and 100 reference ancestral population individuals for each simulated admixture scenario. In the FractalSIM model, the ancestral populations in **Table 4.3** mate among themselves first to a sample size of 800 individuals each following a demographic and population growth model.

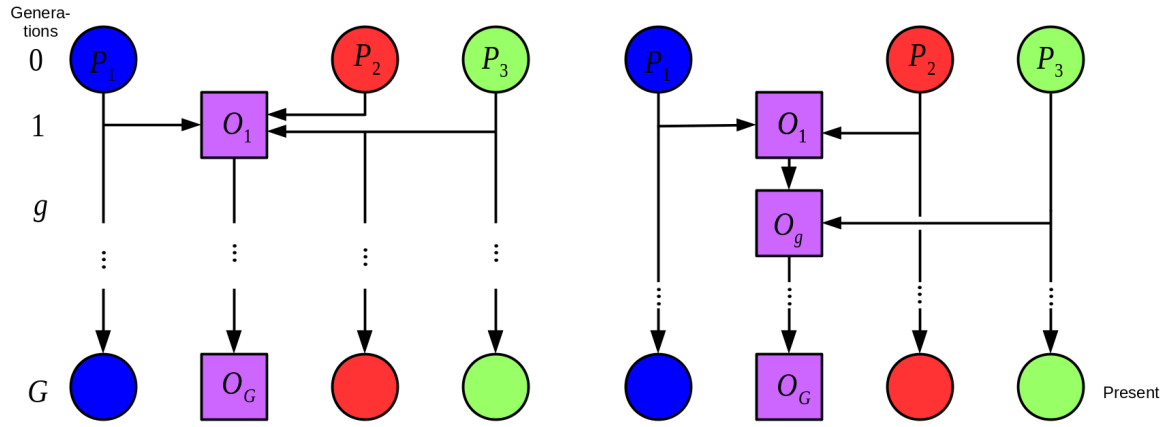
This represents isolated growth of ancestral populations before the admixture process [208]. Whereas, in the AdmixSim 2, at each generation, say g, parallel to the admixture process is the mating of individuals belonging to same population group. From these generated individuals, a random sample is picked for a specified number to mate at generation $g+1$.

Mimicking recombination and the mutation process of reference ancestral haplotypes during mating in the isolated growth process, new haplotypes are considered as mosaic copies of reference haplotypes [208]. In the FractalSIM model, the isolated growth process is conducted in three steps: (1) The state transitions are determined bearing in mind that the copying state is dependant on the effective population size, the current sample space, the physical distance and the recombination rate between the consecutive SNPs; (2) The actual copying state is determined. In the absence of pre-admixture selection, each segment of the copying state is randomly sampled from a uniform distribution; finally, (3) The genotypes that match the actual copying states are reproduced in simulated segment regions. The third step depends on the mutation process, in which the expected number of mutations are used to select mutant SNPs randomly [208]. Copying starts with mutation on selected SNPs; after being simulated a haplotype becomes part of the sample space [208].

4.2.5 Simulating admixed population individuals

We assume Poisson distributed break points between the discrete chromosomal segments with a rate proportional to the genetic distance [208]. Over and above the ancestral population haplotypes, the admixture generations, the information on recombination and the number of admixed individuals to simulate, the ancestry proportion contributed by each ancestral population [208] is required. Furthermore, a constant admixture process subsequent to the founding of the admixed population is assumed [37]. An admixture process can be formed by a single admixture event (hybrid isolation/single-point/single-wave admixture model) or multiple admixture events (multiple-wave admixture model) [208]. An admixture is single-point or single-wave if all the contributing ancestral populations interbreed in the first generation with mating only occurring between admixed individuals without further contributions from the ancestral populations in subsequent generations [9, 171, 208]. Contrarily, multiple admixture events occur when ancestral populations keep contributing to the admixture process after the first generation [9, 208]. Multi-wave can be gradual (ga) when each of the ancestral populations keep contributing to the admixture in each generation or continuous gene flow (cgf) when a portion of the ancestral populations contribute to the admixture at each generation. **Figure 4.1** illustrates single-point and multi-wave (a simple case) admixture models

given a three-way admixed population formed G generations ago. Admixed individuals can be



(a) Hybrid-isolation admixture model given three ancestral populations

(b) Multi-wave (or cgf) admixture model given three ancestral populations

Figure 4.1: Admixture models involving three ancestral populations, P_1, P_2 and P_3: (a) A single-wave/point (hybrid isolation) admixture model. (b) An extension of the Pickrell et al. [6] multi-wave admixture model to allow for continuous gene flow (cgf) from ancestral populations up-to a given number of generations g, for $1 < g < G$ [7]. Circles represent ancestral populations (P_k), $k = 1,2,3$, squares represent the admixed population at generation z, (O_z), $1 \leq z \leq G$. Arrows represent the direction of gene flow.

disease-affected or -unaffected. To simulate disease-affected admixed individuals, the isolated growth process should allow one of the ancestral populations to grow while having the effects of the diseases (pre-admixture disease) [208]. The resulting populations are the ancestral populations to use when simulating diseased-affected and/or -unaffected admixed individuals. Further, to mimick real populations where natural selection may occur during the admixture process (post-admixture selection), admixed individuals are simulated with some SNPs/regions under the effects post-admixture selection [208].

Admixed individual segments are simulated by sampling haplotypes from each source population according to the probability given by the ancestry proportion contributions [18, 208]. For each admixed individual, ancestry segments are copied based on the admixture generations and the recombination rate. The copied ancestry is recorded, representing the true ancestry [18, 208].

AdmixSim 2 simulates hybrid isolation, gradual admixture and continuous gene flow. More so, it can account for complex admixture scenarios including *denovo* mutation sites, recombination events and selection which can be ancestry-specific or a combination of alleles [232]. For

this study, we test the constant recombination events and positive and negative selection. **Table 4.4** provides details on the chromosome, SNP position and ID, the selection type that was simulated and the gene closest to the region under selection. We give details of the populations, ancestry proportions, generations and population sizes of each of the involved populations in **Section**. It is important to note that, AdmixSim 2 does not account for diseased individuals.

Table 4.4: AdmixSim2 parameters and post-admixture selection chromosomes and SNP positions.

Chromo-	SNP Position	SNP-ID	Selection type	Nearest gene
1	46406461	rs3855959	Positive	FAAH
2	227413405	rs4673171	Positive	RHBDD1
2	227567352	rs10210997	Positive	RHBDD1
10	579228	rs12252141	Negative and Positive	
20	217690	rs926617	Positive	EDM2

4.2.5.1 Healthy, single-wave multi-way admixtures at different generations

A single generation is approximately 20-30 years [9, 32, 54, 55]. If generations since admixture are 20 or less, the admixture event is recent, and, if generations are above 100, the event is ancient [56, 97]. This section considers healthy individuals simulated by a single-wave admixture event. As a result, for disease-unaffected admixed populations, ancestral populations were expanded as in **Section 4.2.4**. Admixed individuals were formed 15, 100 or 600 generations ago, in the absence of pre-admixture disease, pre- and post-admixture selection. The simulated admixed populations all mate according to the ancestry proportions given in **Table 4.5**.

Table 4.5: Healthy, multi-way admixed populations simulated by a single-wave event.

	Population name	Admixture generations	Ancestry proportions				
			CEU	CHB	GIH	KHS	YRI
Three-way	SIM3G1	15	0.65	0.28	-	-	0.07
	SIM3G2	100	0.65	0.28	-	-	0.07
	SIM3G3	600	0.65	0.28	-	-	0.07
Five-way	SIM5G1	15	0.18	0.07	0.13	0.32	0.30
	SIM5G2	100	0.18	0.07	0.13	0.32	0.30
	SIM5G3	600	0.18	0.07	0.13	0.32	0.30

4.2.5.2 Healthy, multiple-wave multi-way admixtures at different generations

Considering an admixed population formed by the mixing of three or more previously isolated populations, a single-wave admixture model is less practical than the multiple-wave (or continuous gene flow) [171]. Previous studies showed that given multiple-wave or continuous gene flow admixture, it is challenging to estimate the generations since admixture [171, 178]. However, we are yet to be establish if this is the case when inferring local ancestry. Therefore, we test the performance of existing models in two-event admixtures formed recently (15 generations) and 100 generations ago. The simulation process involves mimicking isolated growth in parental populations as described in **Section 4.2.4**. Secondly, ancestry segments of the specified admixed individuals are simulated by identifying breakpoints based on the Markov chain process identical to Equation (2.2.1) [78]. Similarly to single-wave admixtures, the ancestry of every chromosomal segment is drawn independently from the ancestry proportions [208]. Since this involves two-wave admixture events, in a K-way admixture, the first wave involves K-1 ancestral populations, ending in the fifth (5^{th}) and fiftieth (50^{th}) generation for admixtures simulated under 15 and 100 generations, respectively. Afterwards, the second wave which involves the contribution from the remaining ancestral population takes place and, finally, isolated growth among admixed individuals occur. For three-way admixtures, the admixture process is as shown in **Figure 4.1b**. We provide details of the admixture time, the contributing populations and ancestry proportions at each wave in **Table 4.6**. Admixed populations in this section are all disease-unaffected (healthy).

Table 4.6: Healthy, multi-way admixed populations simulated in two admixture waves.

	Population name	**Admixture generations**	**Proportions contributed**					
			Admx	CEU	CHB	GIH	KHS	YRI
Three-way	SIM3MG1	5	0.00	0.65	0.35	-	-	0.00
		15	0.89	0.00	0.00	-	-	0.11
	SIM3MG2	50	0.00	0.65	0.35	-	-	0.00
		100	0.89	0.00	0.00	-	-	0.11
Five-way	SIM5MG1	5	0.0	0.20	0.1	0.15	0.55	0.0
		15	0.68	0.0	0.0	0.0	0.0	0.32
	SIM5MG2	50	0.00	0.20	0.10	0.15	0.55	0.00
		100	0.68	0.00	0.00	0.00	0.00	0.32

4.2.5.3 Disease-affected and/or -unaffected individuals with SNPs under selection

Irrespective of the admixture model, single- or multiple-wave, we also consider disease-affected and/or -unaffected individuals with some of their SNPs under selection. As highlighted in **Section 1.5** admixed populations play important roles in understanding diseases, personalising medicines and identifying genes of biological relevance. It has been reported that the effects of the disease and post-admixture selection may yield deviations in local ancestry [9, 66, 119, 120, 116]. As a result, this section evaluates existing models in disease-affected and -unaffected admixed individuals with some SNPs/regions under selection.

Similarly to healthy individual simulations, the reference panels experience isolated growth before the admixture as described in **Section 4.2.4**. We noted that, to mimic the case-control causal model during the isolated growth process one or more source populations is/are simulated according to the causal disease model [208]. In our case, the CHB individuals were allowed to mate among themselves under the causal disease model while all other reference populations mate among themselves under the null disease model.

Previous studies have shown that signals of recent selection in admixed populations (post-admixture selection) yield to deviations in local ancestry [9, 113, 116, 166, 234, 235, 236, 237]. As a result, we do not simulate pre-admixture selection. To obtain admixed individual genotypes with SNPs under post-admixture selection and their true ancestral segments, we sample a third of the simulated admixed individuals and use a specified relative fitness to mimick a population bottleneck scenario [208]. Now, for admixed individuals to inherit the disease, haplotypes for healthy individuals are sampled from healthy segments of the disease-unaffected populations; whereas haplotypes for diseased segments are sampled from the disease-affected population (CHB). **Table 4.7** provides information on the diseased and/or selected chromosome and SNPs (position and identity). It is worth mentioning that all defined disease-risk SNPs in **Table 4.7** must be passed onto the admixed individuals. On the other hand, **Table 4.8** summarises the simulated disease-affected cases, highlighting the number of contributing population (that is, three or five). If the population has suffix "S1NS", it is a single-wave admixture, while suffix "M1NS", is a two-wave admixture representing multi-point admixture. Also shown is the admixed population name, generations since admixture and ancestry proportion each population contributed to the admixture process (admixed: admix or ancestral populations: CEU, CHB, YRI, GIH and KHS). Similarly to **Section 4.2.5.1** and **4.2.5.2**, each admixture simulation consists of 200 individuals simulated as described in **Section 4.2.4**. Since the set of ancestral individuals used for generating admixed individuals should be different from the one for inferring ancestry in these admixed individuals, we set aside 100 randomly

Table 4.7: Disease-affected and post-admixture selection chromosomes and SNP positions. Diseased and selected SNPs are highlighted by "D" and "S", respectively in column 1.

Disease/ selection status	Chromo-some	SNP position	SNP ID	Fitness			Disease risk
				s0	s1	s2	
D	1	2408485	rs7512269	✗	✗	✗	1.5,2.44
D	1	3502201	rs2794327	✗	✗	✗	1.8,3.24
S	1	46406461	rs3855959	1	0.5	0.2	✗
S	2	241750403	rs4675991	1	0.625	1	✗
S	9	182129	rs3008131	1	6.5	12.8	✗
D	10	579228	rs12252141	✗	✗	✗	2.3,3.69
S	14	19495751	rs10141075	1	2.2	4.2	✗
D	20	14863051	rs2423854	✗	✗	✗	2.3,3.69
S	20	217690	rs926617	1	5.6	10.1	✗

Table 4.8: Disease-affected, multi-way admixed populations with diseased and selected SNPs.

	Population name	Admixture generations	Proportions contributed					
			Admx	CEU	CHB	GIH	KHS	YRI
Three-way	SIM3S1NS	15	0.00	0.65	0.28	-	-	0.07
	SIM3M1NS	5	0.00	0.65	0.35	-	-	0.00
		15	0.89	0.00	0.00	-	-	0.11
Five-way	SIM5S1NS	15	0.00	0.18	0.07	0.13	0.32	0.3
	SIM5M1NS	5	0.00	0.20	0.10	0.15	0.55	0.00
		15	0.68	0.00	0.00	0.00	0.00	0.32

selected ancestral individuals which are not part of the 400 ancestral individuals used in the admixture simulation process.

4.2.6 Measuring the performance of local ancestry models

Although several studies evaluated local ancestry inference models, they have been contextual (see **Table B.1**), each based on different datasets and performance measures [5]. Measures differ depending on whether the data is simulated or real. One way to measure the performance of a model in simulated data is by considering the correlation coefficient between the approximated and the actual ancestry [66, 92], the higher the correlation, the better the model. Other ways are the mean absolute deviation between the estimated and actual ancestry [97] where the best model is the one with the least mean absolute deviation, and the accuracy performance measure [217] which in the local ancestry inference context is the ability of a model

to correctly estimate the number of allele copies contributed by every ancestral population at every SNP. The number of copies are 0, if the segment is not from that particular population, or 1 or 2, if one or both copies are from that particular population [2]. Furthermore, performance has been evaluated based on the rate at which segments from a particular ancestry are correctly (sensitivity or true positive rate [217]) or wrongly classified [66].

Now, given real (or simulated) data of disease-affected (cases) and -unaffected (controls) individuals, model performance is evaluated by comparing the deviations that may exist in the local ancestry of cases and controls [66], and/or comparing the global ancestry estimated by models [104] including ADMIXTURE [85] to the local ancestry at the global level. Contrariwise, in family datasets models are tested by assessing the Mendelian inconsistencies in local ancestry [119].

4.2.7 Simulated admixtures and local ancestry models within a unified framework

We note that given a dataset, performance measures may rank models differently. This is consistent with what was reported by Chicco and Jurman [218] on ranking the performance of five models in genomic data based on F_1 score, the accuracy and the Mathews correlation coefficient (MCC). Following this and the urge to utilise the popular confusion matrix metrics, we evaluate the predictive performance of local ancestry models on simulated K-way (classes) admixed population data. The metrics are determined class-wise (including the accuracy (ACC), the true positive rate (TPR) or recall [66, 221], the MCC [238] and the receiver operating characteristic (ROC) graph [219, 221] (here represented as TPR and FPR values)) or overall (including the overall accuracy (OACC), the overall MCC (OMCC) [239], the Cohen's Kappa (κ) [240] and the F_1 micro-average denoted as "F_1-micro, herein). We provide more details of these performance measures in subsequent sections.

4.2.7.1 The confusion matrix

Assume the ground truth (true local ancestry) of simulated admixed individuals formed from the mating of K$\geq$ 2 ancestral populations and the predicted (estimated) local ancestry (from different models) is known. Consider each local ancestry model as a classifier or classification algorithm. The aim is to evaluate the classifier performance. We note that most local ancestry inference models are discrete classifiers as they have only a single class label, the ancestry. As such, a single confusion matrix or contingency table, whose rows and columns represent the

ground truth (actual ancestry) and the predictions made by the classifier, is constructed [219]. **Table 4.9** is a confusion matrix, which we denote by **CM**. Let CM_{kl} be the row k, column

Table 4.9: A K-class confusion matrix

		Estimated						**Totals**
		1	2	3	4	⋯	K	
Ground truth	1	CM_{11}	CM_{12}	CM_{13}	CM_{14}	⋯	CM_{1K}	$CM_{1\bullet}$
	2	CM_{21}	CM_{22}	CM_{23}	CM_{24}	⋯	CM_{2K}	$CM_{2\bullet}$
	3	CM_{31}	CM_{32}	CM_{33}	CM_{34}	⋯	CM_{3K}	$CM_{3\bullet}$
	4	CM_{41}	CM_{42}	CM_{43}	CM_{44}	⋯	CM_{4K}	$CM_{4\bullet}$
	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
	K	CM_{K1}	CM_{K2}	CM_{K3}	CM_{K4}	⋯	CM_{KK}	$CM_{K\bullet}$
	Totals	$CM_{\bullet 1}$	$CM_{\bullet 2}$	$CM_{\bullet 3}$	$CM_{\bullet 4}$	⋯	$CM_{\bullet K}$	$CM_{\bullet\bullet}$

l entry representing the count of the true k segments that are estimated as l segments, k,l $= 1,\ldots K$; $CM_{k\bullet} = \sum_{l=1}^{K} CM_{kl}$ be the total count of true k segments (sum of row k); $CM_{\bullet l} = \sum_{k=1}^{K} CM_{kl}$ be the total number of l predictions (the column sum); and

$$CM_{\bullet\bullet} = \sum_{k,l} CM_{kl} = \sum_{k} CM_{k\bullet} = \sum_{l} CM_{\bullet l}$$

be the total segments considered. **Table 4.9** entries are divided into four, that is,

- The true positives of class (ancestry) k (TP_k) - are counts of ancestral population k segments that are correctly estimated, k $= 1,2,\ldots,K$. In **Table 4.9**, $TP_k = CM_{kk}$.
- The false negatives of class (ancestry) k (FN_k) - are counts of population k segments that are predicted/classified as l segments:

$$\begin{aligned} FN_k &= \sum_{\substack{l=1 \\ l\neq k}}^{K} CM_{kl} \\ &= CM_{k\bullet} - CM_{kk}. \end{aligned}$$

  Thus, the false negatives of class k consists of all row k entries except for the k=l entry.
- The false positives of class k (FP_k) - are counts of non-k segments that are predicted

as k segments:

$$\begin{aligned} FP_k &= \sum_{\substack{l=1 \\ l \neq k}}^{K} CM_{lk} \\ &= CM_{\bullet k} - CM_{kk}, \end{aligned}$$

Thus, the false positives of class k consists of all column k entries except for entry k=l.

- The true negatives of class k (TN_k) - are counts of non-k segments that are predicted as non-k:

$$\begin{aligned} TN_k &= \sum_{\substack{k,l \\ l \neq k}}^{K} CM_{kl} \\ &= CM_{\bullet\bullet} - (CM_{k\bullet} + CM_{\bullet k}). \end{aligned}$$

Thus, TN_k is made up of all entries that are not in row k and column l, yielding $(K-1)^2$ entries.

4.2.7.2 Evaluating the performance of a classifier in classes

As previously highlighted, confusion matrix metrics either assess classifier performance in every class or globally. In this chapter, we call metrics for every class confusion matrix class metrics, while those measuring global performance, we call confusion matrix overall metrics. Now, four metrics correspond to the four divisions of the confusion matrix entries in **Section 4.2.7.1**. These are the true positive rate (TPR) or recall, false negative rate (FNR), false positive rate (FPR) and true negative rate (TNR) or specificity [217, 218, 221, 241]. However, for this analysis, we only use the TPR and FPR. As such, we give the definitions of these two only. Based on the definitions given in **Section 4.2.7.1**, the TPR for ancestry k is

$$\begin{aligned} TPR_k &= \frac{TP_k}{TP_k + FN_k} \\ &= \frac{CM_{kk}}{\sum_{l=1} CM_{kl}}, \end{aligned} \tag{4.2.1}$$

Basically, the TPR is the proportion of correctly estimated k segments that are actually k, and $0 \leq TPR \leq 1$. The best model should correctly identify ancestry segments [219] and thus, TPR values should be high. In the local ancestry framework, Chimusa et al. [66] defined the

rate of estimating an ancestry k as l as

$$\frac{\mathscr{K}_l}{\mathscr{L}_k} \tag{4.2.2}$$

where $\mathscr{K}_l$ counts alleles estimated as ancestry l when the ground truth ancestry is k and $\mathscr{L}_k$ is the number of alleles whose ground truth ancestry is k. Therefore if l=k, Equation (4.2.2) gives the rate of correctly estimating a particular ancestry that is, the true positive rate (TPR) of ancestry k in Equation (4.2.1). Otherwise, if l≠k we have the miscall rate [66].

On the contrary, the FPR highlights the extent to which a model wrongly labels a class positively. It is defined by

$$\mathrm{FPR}_k = \frac{\mathrm{FP}_k}{\mathrm{FP}_k + \mathrm{TN}_k}. \tag{4.2.3}$$

However, to clearly visualise the stability of the gains to the losses of a classifier we use the receiver operating characteristic (ROC) [219, 221]. It graphs the TPR against the FPR [219, 221]. Since our classifiers are discrete we obtain a single point on the ROC space from the single confusion matrix [219]. Nonetheless, the resulting ROC graph still consists of the four important corners that are useful for interpretations. The point where both the false and true positive rates are zero, (0,0) indicating that the classifier failed to detect any segment of the given class (positives) [219, 221], and (1,1) a point showing that the classifier classified all segments to one class k, while other classes have zero counts [219, 221]. Joining (0,0) to (1,1) gives a line of no-discrimination [242], which indicates estimates are random guesses [219, 242]. If a classifier performs well, its ROC curve (or point) should be above the no-discrimination line, otherwise it fails to perform well. Now, point (0,1) indicates that the classifier correctly classified all k and non-k ancestry segments [221], and finally (1,0) indicates that the classification algorithm misclassified all the k and non-k ancestry segments. Generally, a classifier A is better than B, if it is North-west of B [219].

Another most commonly used confusion matrix class metric is the accuracy [218, 221]. It assesses the ability of a classifier to correctly classify both positives and negatives of a given class [221]. Hence, it is a ratio of the true positives and the true negatives to the total number of the considered samples (segments) [217]. Let the accuracy (ACC) of ancestry (class) k be ACC_k, then

$$\mathrm{ACC}_k = \frac{\mathrm{TP}_k + \mathrm{TN}_k}{\mathrm{TP}_k + \mathrm{FN}_k + \mathrm{TN}_k + \mathrm{FP}_k}. \tag{4.2.4}$$

When there are more samples (segments) in one class than there are in other classes (class imbalance), the ACC may be misleading [218, 221, 242]. We note that, TPR, ROC and ACC fail to account for how well the classifier does on negatives [218, 221], thus, we also use the Mathews coefficient correlation (MCC) metric. It measures how correlated the estimates are to the ground truth [218, 242]. MCC has been successfully applied in many scientific fields, including genomics [218, 242]. Calculated from the confusion matrix, the MCC for ancestry (class) k is given by

$$\mathrm{MCC_k} = \frac{(\mathrm{TP_k} \times \mathrm{TN_k}) - (\mathrm{FP_k} \times \mathrm{FN_k})}{\sqrt{(\mathrm{TP_k} + \mathrm{FP_k})(\mathrm{TP_k} + \mathrm{FN_k})(\mathrm{TN_k} + \mathrm{FN_k})}}$$

where, $-1 \leq \mathrm{MCC_k} \leq 1$, -1 indicates bad, 0 average random and 1 perfect prediction [218].

4.2.7.3 Evaluating the overall performance of a classifier

A classifier may do well in one class, but, perform badly in other classes. As a result, it is always good to assess how good a classifier performs overall. Therefore, in addition to class metrics, this chapter also assesses performance by four overall metrics: the overall ACC (OACC), overall MCC (OMCC), the Cohen's Kappa (κ) and the F_1 micro-average score which we denote as F_1-micro.

OACC determines the extent to which a classifier manages to correctly classify classes in general [241]. Using **Table 4.9**, we define OACC as follows

$$\mathrm{OACC} = \frac{\sum_{k=1}^{K} \mathrm{CM_{kk}}}{\mathrm{CM_{\bullet\bullet}}} \tag{4.2.5}$$

where, $\mathrm{CM_{\bullet\bullet}}$ is the totality of all the confusion matrix entries.

Also, consider a metric that does not only focus on the positives, a generalization of the $\mathrm{MCC_k}$, referred to as the overall MCC, denoted by OMCC. Gorodkin [239] showed that given a K-class confusion matrix **CM**, the OMCC can be expressed as

$$\mathrm{OMCC} = \frac{(\mathrm{CM_{\bullet\bullet}})\mathrm{Tr}(\mathbf{CM}) - \sum_{k,l} \widetilde{\mathrm{CM}}_k \widehat{\mathrm{CM}}_l}{\sqrt{(\mathrm{CM_{\bullet\bullet}})^2 - \sum_{k,l} \widetilde{\mathrm{CM}}_k (\widehat{\mathrm{CM}}_l)^T} \sqrt{(\mathrm{CM_{\bullet\bullet}})^2 - \sum_{k,l} (\widetilde{\mathrm{CM}}_k)^T \widehat{\mathrm{CM}}_l}}$$

where $\widetilde{\mathrm{CM}}_k$ and $\widehat{\mathrm{CM}}_l$ are the k^{th} and l^{th} row and column of **CM**, $\mathrm{CM_{\bullet\bullet}}$ is the sum of all entries of **CM**, and CM_k^T is the transpose of vector $\mathrm{CM_k}$.

Also, it is good to determine the extent to which the predicted values (estimates) agree with the ground truth, taking into account the agreement by chance. For this, the Kappa (κ) [240] statistic is normally used. Based on the confusion matrix, κ is calculated as follows

$$\kappa = \frac{\text{OACC} - \text{ORACC}}{1 - \text{ORACC}} \tag{4.2.6}$$

where OACC is the overall accuracy as in Equation (4.2.5), ORACC is the overall random accuracy given by

$$\text{ORACC} = \frac{1}{(\text{CM}_{\bullet\bullet})^2} \sum_{\text{k}=1}^{K} (\text{TP}_\text{k} + \text{FP}_\text{k})(\text{TP}_\text{k} + \text{FN}_\text{k})$$

where $\text{CM}_{\bullet\bullet}$ is as previously defined. We note that $0 \leq \kappa \leq 1$ [243].

F_1-micro score gives equal weight to the precision which is the rate of correct positive predictions (a specific column entries) and the sensitivity (TPR) [244]. The F_1-micro score micro-average (which we will refer as F_1-micro) for a given classification algorithm is defined by

$$F_1\text{-micro} = \frac{2\text{PS}}{\text{P} + \text{S}} \tag{4.2.7}$$

where $\text{S} = \frac{\sum_\text{k}^K \text{TP}_\text{k}}{\sum_\text{k}^K (\text{TP}_\text{k} + \text{FN}_\text{k})}$ and $\text{P} = \frac{\sum_\text{k}^K \text{TP}_\text{k}}{\sum_\text{k}^K (\text{TP}_\text{k} + \text{FP}_\text{k})}$ are the micro-average TPR and micro-average precision, respectively [244].

All the confusion matrix metrics are determined by the PyCM multi-class confusion matrix Python package [241]. In addition to the confusion matrix metrics, we also assess performance based on the deviations in ancestry, we provide details in the next section.

4.2.7.4 Deviations in ancestry

Apart from the confusion matrix metrics, given the ancestry of N individuals in the form of a $KN \times T$ matrix, define the mean absolute error, often termed the mean absolute deviation Guan [97] for the model by

$$\frac{1}{NKT} \sum_{\ell=1}^{KN} \sum_{t=1}^{T} |\hat{x}_{\ell t} - x_{\ell t}|,$$

where, K is the count of ancestral populations, T is the number of SNPs under study, $\hat{x}$ is the estimated ancestry proportion of individual $\imath$ copied from population ℓ at SNP t and x is the ground truth (ancestry proportion of individual $\imath$ as produced during the simulations). The

mean absolute deviation is calculated overall for all classifiers (models). The lower the mean absolute error, the better the estimates are.

Furthermore, given case-control datasets, we also use the deviations in local ancestry between the disease-affected and -unaffected individuals to evaluate model performance [66]. Usual deviations in ancestry yields a linear relationship in case and control deviations [66].

4.2.8 Local ancestry models, inaccurate dates and skewed ancestral sizes

As previously mentioned, models are designed under different assumptions and their accuracy varies from one population to another, justifying the "no free lunch" theorem [5]. Since the information they require is not always accurate when provided [4], the accuracy of estimates may deteriorate (**Section 2.4**). For example, it has been shown that the choice of reference ancestral panels affects estimate accuracy [64]. Consequently, in addition to evaluating existing models on simulated cases (**Tables 4.5**, **4.6** and **4.8**), we test existing models using an underestimate of generations since admixture. In particular, we assess the performance of ELAI [97], RFMIX [129], SUPPORTMIX [125] and WINPOP [156]) on a simulated healthy, three-way admixture formed 15 generations ago (SIM3G1) using 10 generations. Also, motivated by challenge 2 in **Section 2.4**, we evaluate the effects of skewed reference ancestral panels on the accuracy of local ancestry models. We used the following ancestral population sizes, CEU=140, CHB=100, GIH=80, KHS=20 and YRI=160 to infer local ancestry in a recent five-way admixture (SIM5G1).

4.2.9 Application of local ancestry models in real data

Attributable to historical events including slave trade, colonialism, political and economic instability, the human population of South Africa is diversified [18]. Different population groups exist, including the Khoisan people who are one of the oldest human population groups, and were the natives of South Africa. Almost 600 years ago, these natives interbred with other African populations including the Bantu speakers, and some non-African populations including the Europeans, the Malaysians and the Chinese [18]. The resulting population is often referred to as the mixed ancestry of South Africa or "South African coloured" (SAC). We believe that local ancestry might provide important clues in drug discovery and personalising medicines. Hence, our interest to test the two best models as determined from the five-way admixture simulations on the SAC, specifically, the residents of the Western Cape region of South Africa.

This dataset consists of 733 case-control individuals (91 controls and 642 cases) on 272 796 SNPs, and has been reported elsewhere including Chimusa et al. [18, 66]. To infer local ancestry using the two selected models, the ancestral populations are chosen according to previous findings [18, 64, 65].

4.3 Results

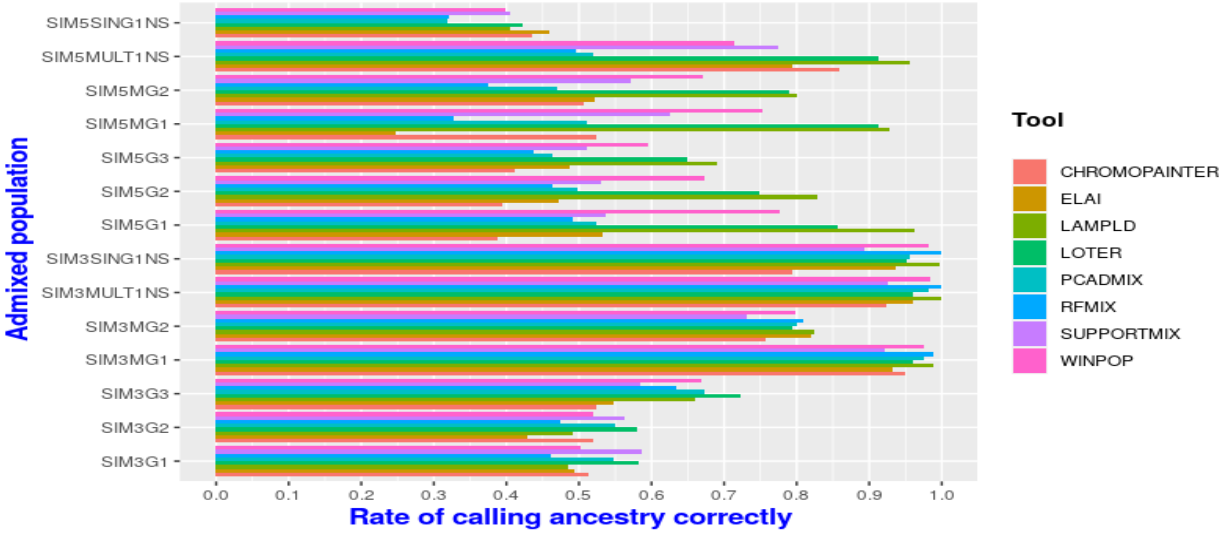
This section discusses results as obtained based on the different performance measures highlighted in **Section 4.2.7**. These are, class metrics including the accuracy (ACC), the rate of calling a given ancestry correctly (TPR), the Mathews correlation coefficient (MCC) and the receiver operating characteristics (ROC), which we represent by a combination of TPR and false positive rate (FPR), and overall metrics including the OACC, the OMCC, the Kappa (κ), the F_1-micro and the mean absolute deviation (MAD). For all admixed simulations, we assess the performance of models on predicting each ancestry (class) and overall. We use the first three letters of the model name to represent the longer model name in tables. For example, "SUP" represents SUPPORTMIX and "WIN" represents WINPOP.

4.3.1 Existing models on simulated data: correct admixture dates and equal ancestral population sizes

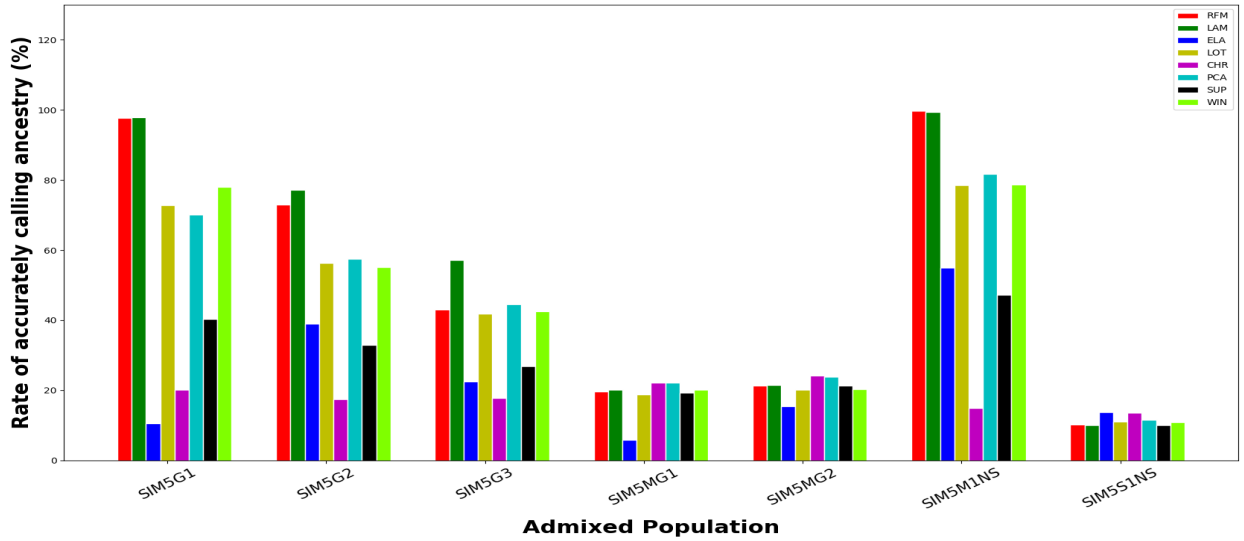
In this section, we discuss the performance of existing models on simulated data assessed based on confusion matrix metrics when local ancestry is inferred using correct admixture dates and equal ancestral population sizes. Each subsection presents ancestry-specific (class) model performance and concludes with a paragraph on the global (overall) performance. **Figures 4.2**, **4.3** and **4.4** represent the measure of model performance based on the rate of calling a particular ancestry correctly (TPR), the accuracy (ACC) and the Mathews correlation coefficient (MCC).

(a) True positive rate when identifying CEU

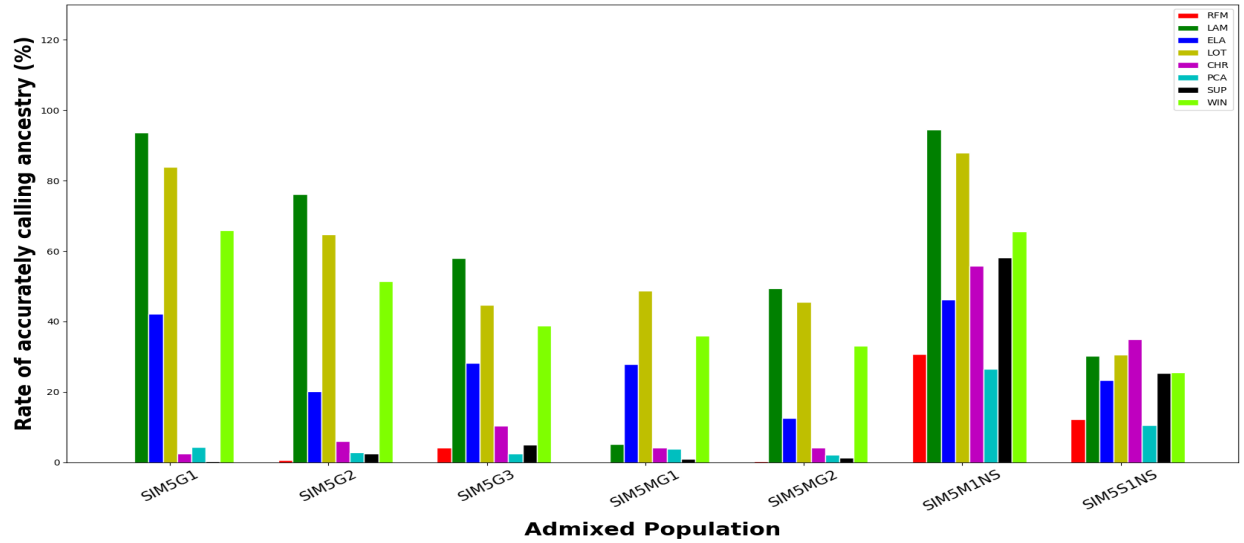
(b) True positive rate when identifying CHB



(c) True positive rate when identifying YRI

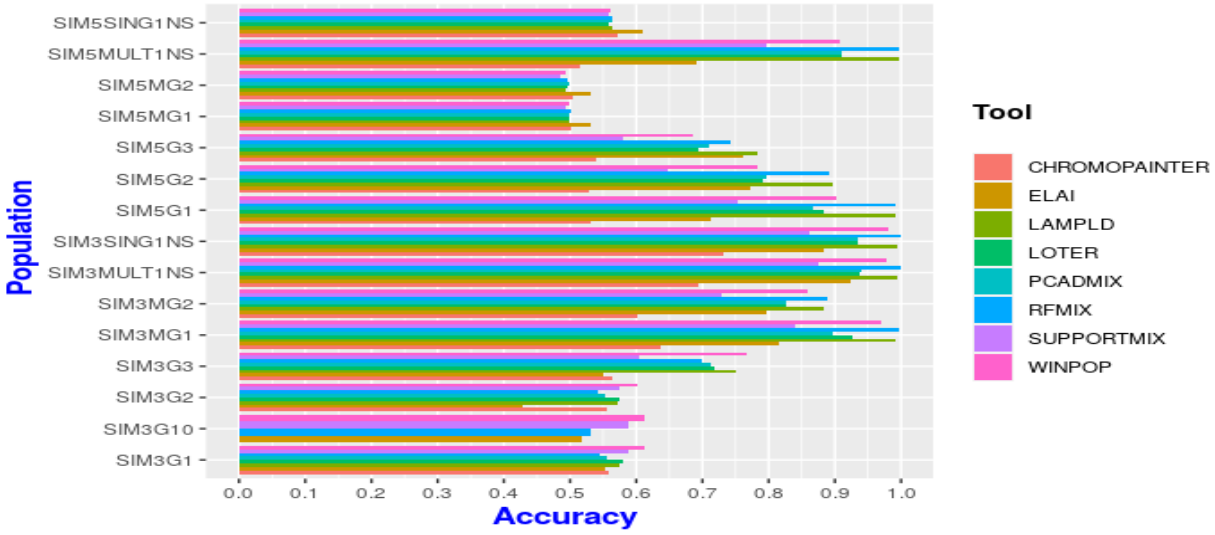


(d) True positive rate when identifying GIH

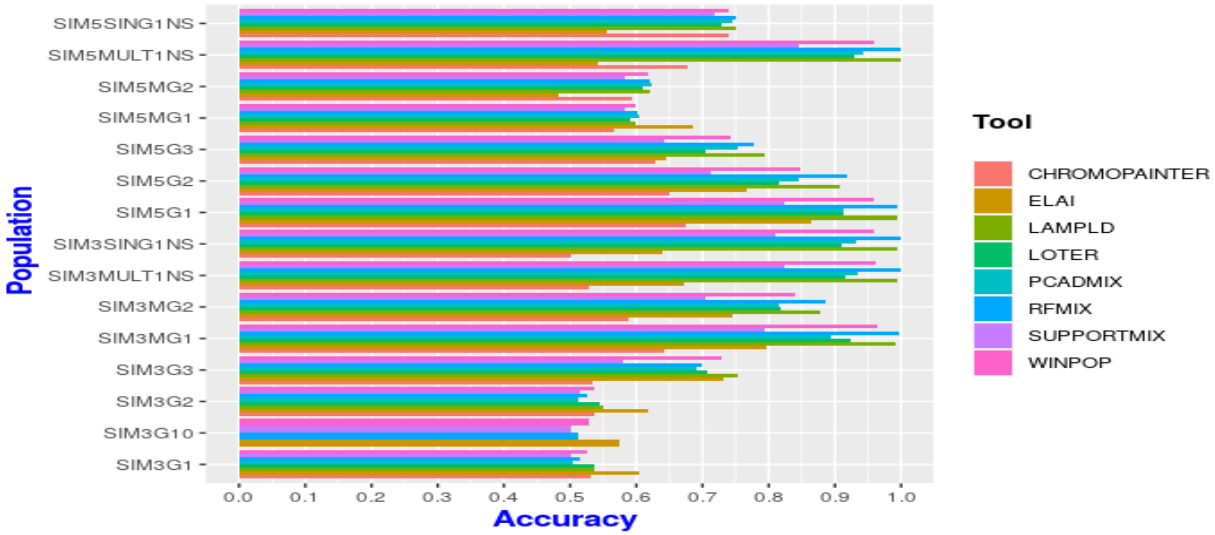


(e) True positive rate when identifying KHS

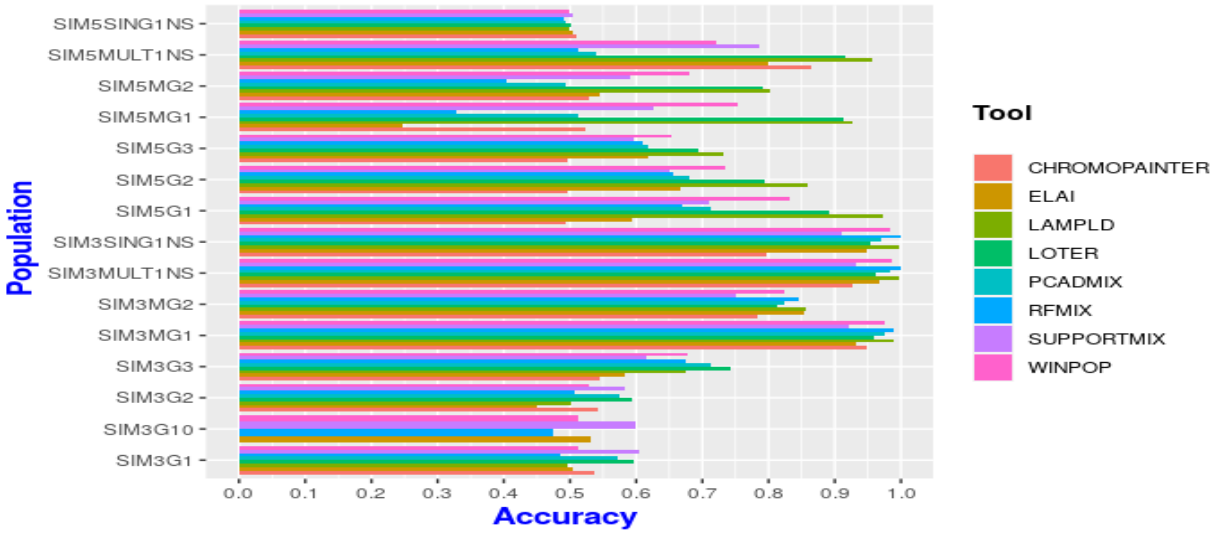
Figure 4.2: Performance of each model in each admixed simulation as measured by the TPR when identifying: (a) CEU, (b) CHB, (c) YRI, (d) GIH and (e) KHS copies.



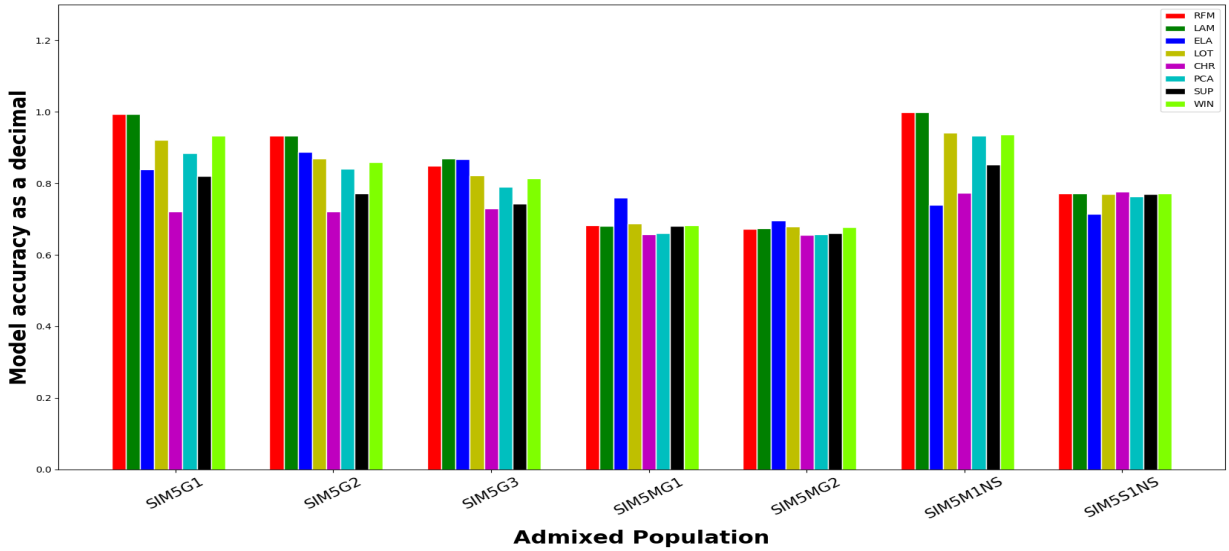
(a) Model accuracy in estimating CEU



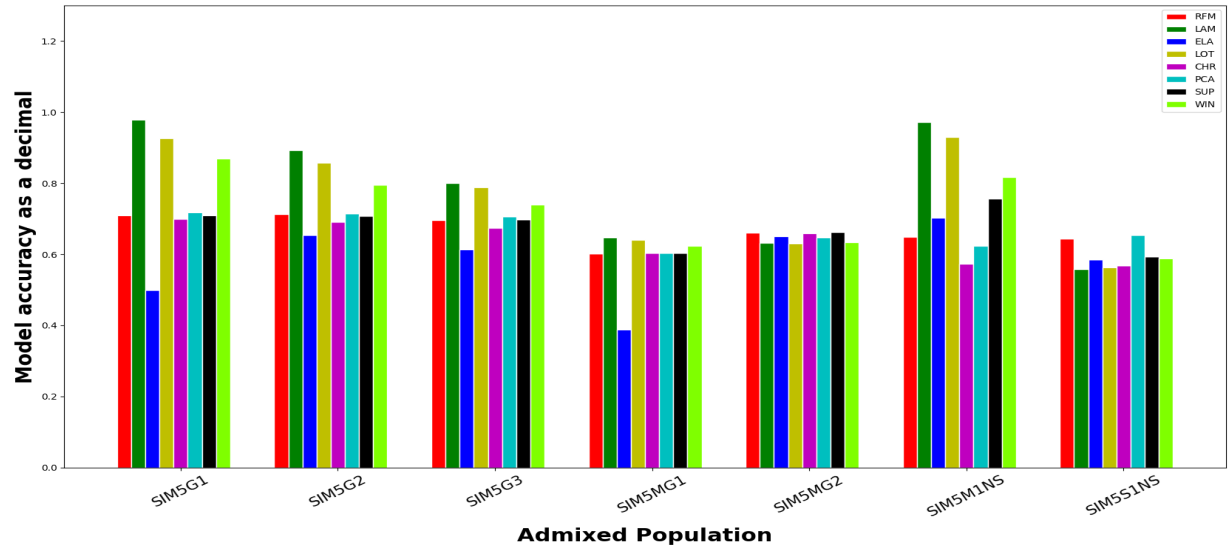
(b) Model accuracy in estimating CHB



(c) Model accuracy in estimating YRI

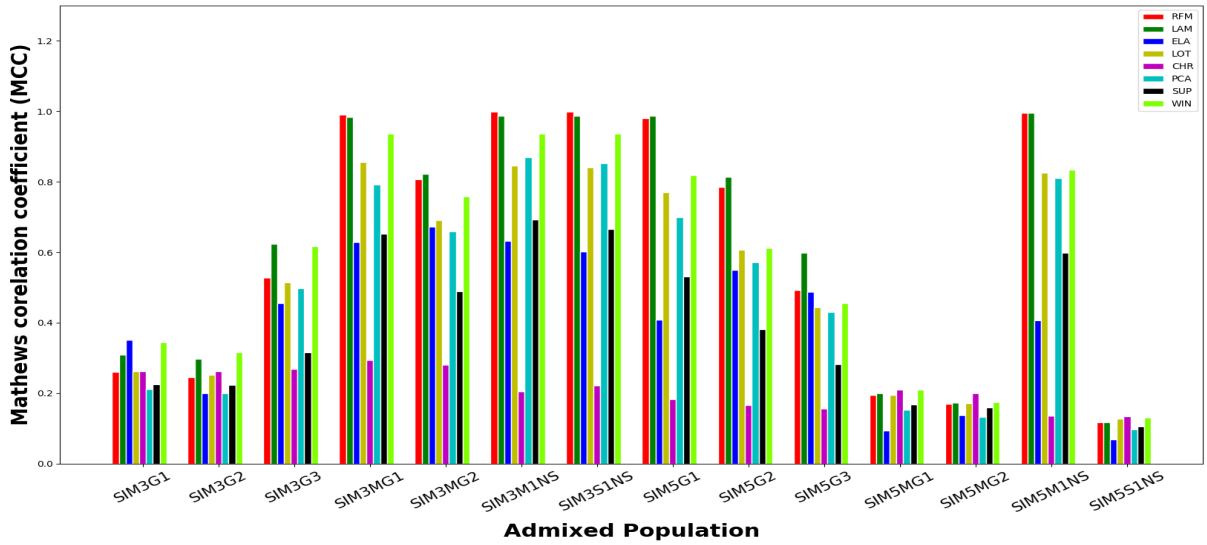


(d) Model accuracy in estimating GIH

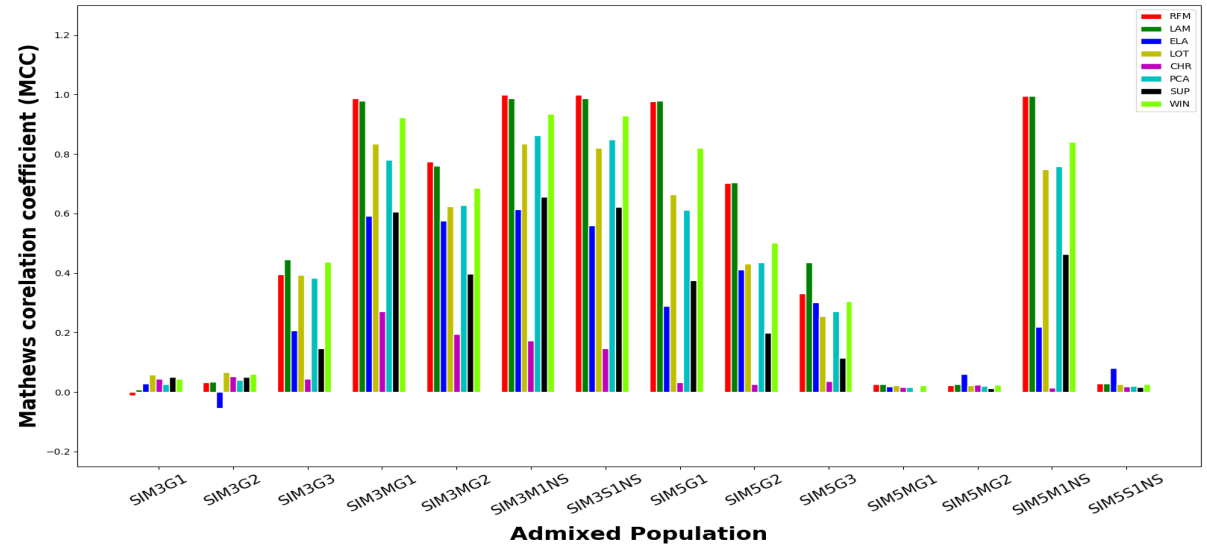


(e) Model accuracy in estimating KHS

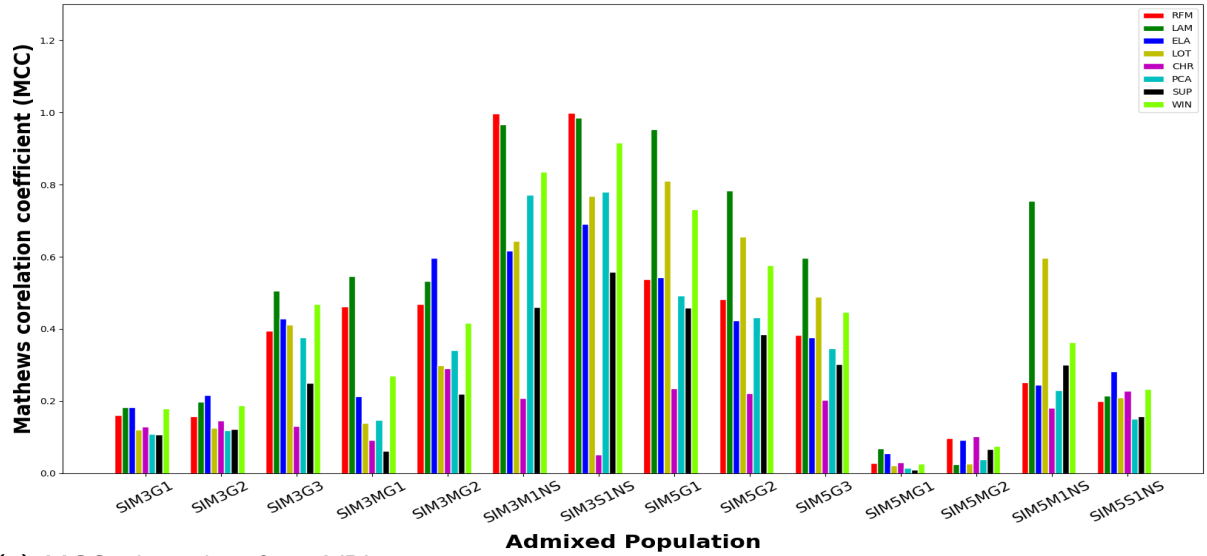
Figure 4.3: Performance of each model in each admixed simulation given the ACC when identifying: (a) CEU, (b) CHB (c) YRI, (d) GIH and (e) KHS copies.



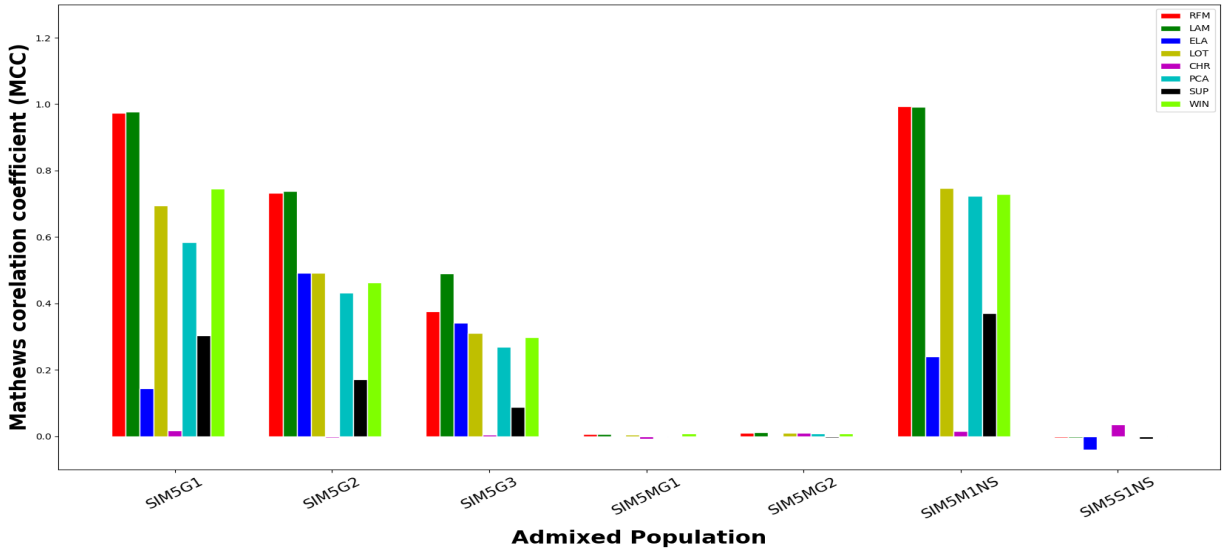
(a) MCC when identifying CEU copies



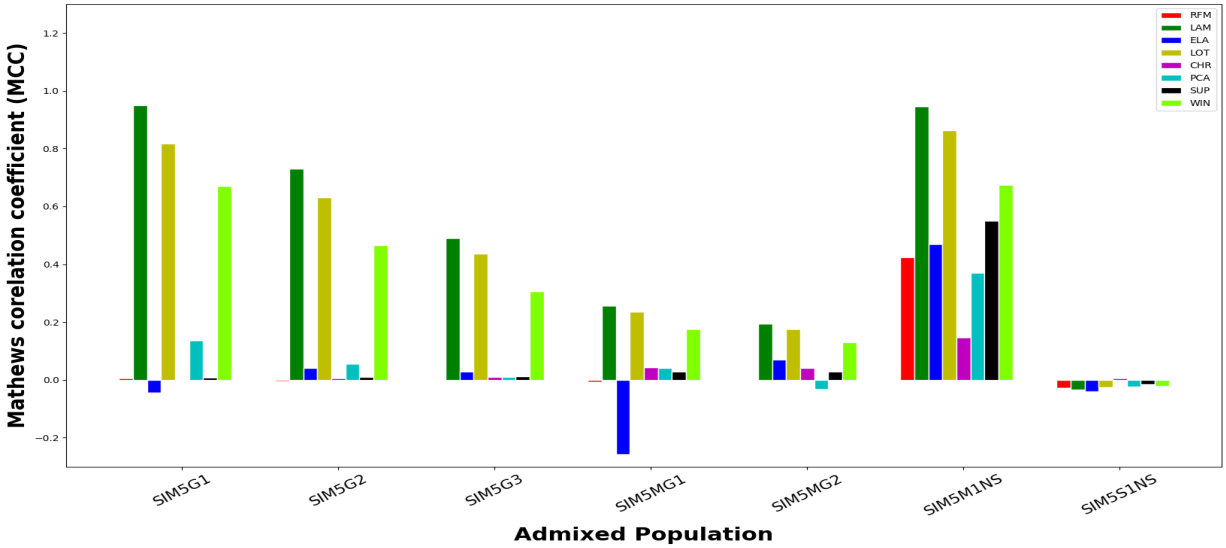
(b) MCC when identifying CHB copies



(c) MCC when identifying YRI copies



(d) MCC when identifying GIH copies



(e) MCC when identifying KHS copies

Figure 4.4: Performance of each model in each admixed simulation based on the MCC metric when identifying: (a) CEU, (b) CHB, (c) YRI, (d) GIH and (e) KHS copies.

Tables 4.10 and **4.11** represent the top five models when local ancestry is deconvoluted in three- and five-way admixtures. **Tables 4.12**, **4.13**, **4.14** and **4.15** represent the model performance as determined by the overall Mathews correlation coefficient (OMCC), overall accuracy (OACC), the Kappa (κ) and the F_1-micro average score for three- and five-way admixtures. The first leading model is contained in the dark-grey coloured cell, while the second is in light-grey coloured cells. They are constructed according to the PyCM compare method of comparing confusion matrices of different models (**Tables 4.12**, **4.13**, **4.14** and **4.15**). "Admix" stands for admixture model, "admix popn" for admixed population, "S-W" and "M-

W" for single-wave and multiple-wave, "H" for healthy; where ✓ shows a population consists of disease-unaffected individuals only and ✗ consists of disease-affected and-unaffected population individuals.

Table 4.10: The five leading models in simulated three-way admixtures.

Admix Model	Admix Popn	G	H	Rank
S-W	SIM3G1	15	✓	WIN ELAI LAM LOT CHR
	SIM3G2	100	✓	WIN LAM CHR LOT RFM
	SIM3G3	600	✓	LAM WIN LOT RFM PCA
	SIM3S1NS	15	✗	RFM LAM WIN PCA LOT
M-W	SIM3MG1	15	✓	RFM LAM WIN LOT PCA
	SIM3MG2	100	✓	LAM RFM WIN LOT ELA
	SIM3M1NS	15	✗	RFM LAM WIN PCA LOT

Table 4.11: The five leading models in simulated five-way admixtures.

Admix Model	Admix Popn	G	H	Rank
S-W	SIM5G1	15	✓	LAM LOT WIN RFM PCA
	SIM5G2	100	✓	LAM LOT RFM WIN PCA
	SIM5G3	600	✓	LAM LOT WIN RFM ELA
	SIM5S1NS	15	✗	CHR WIN LOT ELA RFM
M-W	SIM5MG1	15	✓	LAM LOT WIN ELAI CHR
	SIM5MG2	100	✓	LAM LOT WIN CHR ELA
	SIM5M1NS	15	✗	LAM LOT WIN RFM PCA

Table 4.12: The overall Mathews correlation coefficient (OMCC) measuring performance of each of the eight state-of-the-art models in three- and five-way admixture simulations.

Admix Population	Overall Mathews correlation coefficient (OMCC)							
	WIN	SUP	RFM	ELA	LAM	PCA	LOT	CHR
Three-way admixtures								
SIM3G1	0.1933	0.1334	0.1336	0.1839	0.1648	0.1180	0.1533	0.1478
SIM3G2	0.1900	0.1357	0.1433	0.0983	0.1733	0.1209	0.1527	0.1550
SIM3G3	0.5129	0.2352	0.4416	0.3361	0.5240	0.4252	0.4447	0.1508
SIM3MG1	0.9186	0.6055	0.9827	0.5930	0.9753	0.7763	0.8280	0.2769
SIM3MG2	0.6744	0.4070	0.7367	0.6154	0.7493	0.5996	0.6066	0.2463
SIM3M1NS	0.9301	0.6587	0.9991	0.6226	0.9853	0.8620	0.8281	0.1902
SIM3S1NS	0.9301	0.6335	0.9987	0.5908	0.9866	0.8433	0.8241	0.1708
Five-way admixtures								
SIM5G1	0.7459	0.3710	0.6947	0.2422	0.9655	0.5061	0.7674	0.1184
SIM5G2	0.5279	0.2586	0.5505	0.3528	0.7585	0.4017	0.5811	0.1042
SIM5G3	0.3698	0.1790	0.3316	0.2704	0.5306	0.2857	0.3979	0.0963
SIM5MG1	0.1099	0.0504	0.0576	-0.0710	0.1413	0.0478	0.1307	0.0622
SIM5MG2	0.0914	0.0560	0.0666	0.0687	0.1096	0.0417	0.1017	0.0779
SIM5M1NS	0.6964	0.4893	0.6765	0.3225	0.9568	0.5491	0.7953	0.1052
SIM5S1NS	0.0896	0.0619	0.0819	0.0767	0.0752	0.0633	0.0776	0.09666
SIM5G1 tested with unequal ancestral population sizes								
SIM5SKEW	0.7116	0.3757	0.6947	0.2555	0.8155	0.4874	0.6745	0.1223

Table 4.13: The overall accuracy (OACC) attained by each of the eight state-of-the-art models in three- and five-way admixture simulations.

Admix Population	Overall Accuracy (OACC)							
	WIN	SUP	RFM	ELA	LAM	PCA	LOT	CHR
Three-way admixtures								
SIM3G1	0.4851	0.4665	0.4283	0.4696	0.4599	0.4434	0.4764	0.4529
SIM3G2	0.4853	0.4604	0.4352	0.3854	0.4651	0.4449	0.4755	0.4571
SIM3G3	0.6869	0.5118	0.6273	0.5310	0.6857	0.6357	0.6571	0.4557
SIM3MG1	0.9585	0.7931	0.9912	0.7879	0.9875	0.8864	0.9118	0.6280
SIM3MG2	0.8090	0.6499	0.8439	0.7564	0.8518	0.7674	0.7719	0.5538
SIM3M1NS	0.9653	0.827	0.9995	0.7897	0.9927	0.9317	0.9138	0.5938
SIM3S1NS	0.9649	0.8104	0.9994	0.7634	0.9932	0.9206	0.9098	0.57312
Five-way								
SIM5G1	0.8003	0.4792	0.7011	0.4326	0.9731	0.5811	0.8187	0.2964
SIM5G2	0.6293	0.3944	0.6061	0.4890	0.8123	0.5090	0.6681	0.2891
SIM5G3	0.5001	0.3285	0.4602	0.4276	0.6326	0.4204	0.5139	0.2817
SIM5MG1	0.3155	0.1974	0.1646	0.1636	0.3739	0.1865	0.3607	0.2027
SIM5MG2	0.2915	0.2165	0.2080	0.2225	0.3281	0.2032	0.3180	0.2331
SIM5M1NS	0.75735	0.6106	0.6476	0.4483	0.9698	0.5611	0.8572	0.3947
SIM5S1NS	0.3060	0.2785	0.2965	0.2709	0.2989	0.2787	0.2961	0.3132
SIM5G1 tested with unequal ancestral population sizes								
SIM5SKEW	0.7669	0.4829	0.7011	0.4398	0.8385	0.5648	0.7362	0.3006

Table 4.14: The Cohen's Kappa metric measuring performance of each of the eight state-of-the-art models in three- and five-way admixture simulations.

Admix Population	Cohen's Kappa (κ)							
	WIN	SUP	RFM	ELA	LAM	PCA	LOT	CHR
Three-way admixtures								
SIM3G1	0.1904	0.1328	0.1290	0.1785	0.1614	0.1168	0.1529	0.1453
SIM3G2	0.1877	0.1349	0.1386	0.0894	0.1699	0.1198	0.1524	0.1525
SIM3G3	0.5046	0.231	0.4269	0.3037	0.5110	0.4198	0.4427	0.1477
SIM3MG1	0.9183	0.6033	0.9826	0.5922	0.9753	0.7762	0.8278	0.2768
SIM3MG2	0.6740	0.4066	0.7367	0.5978	0.7491	0.5986	0.6053	0.2463
SIM3M1NS	0.9300	0.6578	0.9991	0.6017	0.9852	0.8619	0.8277	0.1901
SIM3S1NS	0.9310	0.6330	0.9987	0.5683	0.9865	0.8433	0.8239	0.1707
Five-way								
SIM5G1	0.7425	0.3434	0.6156	0.2237	0.9652	0.4659	0.7667	0.1109
SIM5G2	0.5254	0.2441	0.4894	0.3352	0.7576	0.3742	0.5774	0.1000
SIM5G3	0.366	0.1699	0.3055	0.2609	0.5290	0.2682	0.3907	0.0935
SIM5MG1	0.1075	0.0436	0.0469	-0.0617	0.1411	0.0411	0.1303	0.0544
SIM5MG2	0.0907	0.0510	0.0587	0.0635	0.1091	0.0381	0.1013	0.0718
SIM5M1NS	0.6723	0.4733	0.5755	0.293	0.9556	0.4714	0.792	0.1051
SIM5S1NS	0.0883	0.0616	0.0757	0.0748	0.0743	0.0591	0.0772	0.0959
SIM5G1 tested with unequal ancestral population sizes								
SIM5SKEW	0.7000	0.3431	0.6155	0.2370	0.7916	0.4460	0.6623	0.1132

Table 4.15: The F_1-micro metric measuring performance of each of the eight state-of-the-art models in three- and five-way admixture simulations.

Admix Population	F_1-micro							
	WIN	SUP	RFM	ELAI	LAM	PCA	LOT	CHR
Three-way admixtures								
SIM3G1	0.4466	0.4115	0.4056	0.4442	0.4316	0.4025	0.4260	0.4182
SIM3G2	0.4483	0.4150	0.4120	0.3783	0.4389	0.4062	0.4272	0.4248
SIM3G3	0.6516	0.4788	0.5970	0.5211	0.6586	0.5969	0.6183	0.4174
SIM3MG1	0.7342	0.5590	0.8006	0.5922	0.8393	0.6460	0.6603	0.4542
SIM3MG2	0.7224	0.5615	0.7649	0.7286	0.7838	0.6679	0.6588	0.5072
SIM3M1NS	0.9214	0.6976	0.9989	0.7319	0.9843	0.8791	0.8200	0.4613
SIM3S1NS	0.9495	0.7302	0.9991	0.7419	0.9905	0.8778	0.8657	0.4109
Five-way admixtures								
SIM5G1	0.8058	0.4196	0.7185	0.3495	0.9758	0.5554	0.7955	0.2396
SIM5G2	0.6127	0.3426	0.5923	0.4789	0.7997	0.4661	0.6387	0.2407
SIM5G3	0.4803	0.2947	0.4188	0.4104	0.6113	0.3753	0.4916	0.2480
SIM5MG1	0.2471	0.1610	0.1561	0.1291	0.2734	0.1675	0.2622	0.1791
SIM5MG2	0.2602	0.1944	0.1915	0.2217	0.2589	0.1891	0.2559	0.2182
SIM5M1NS	0.7179	0.5333	0.7276	0.3925	0.9390	0.5986	0.7863	0.2749
SIM5S1NS	0.2500	0.2365	0.2238	0.2360	0.2442	0.2160	0.2500	0.2614
SIM5G1 tested with unequal ancestral population sizes								
SIM5SKEW	0.7761	0.4166	0.7186	0.3598	0.8727	0.5323	0.7211	0.2345

Table 4.16: The overall mean absolute deviations ± SD of RFMIX, LAMP-LD, ELAI, LOTER, CHROMOPAINTER, PCADMIX, SUPPORTMIX and WINPOP in three- and five-way admixture simulations. Dark-grey coloured cells show the leading model for each admixed population and light-grey coloured cells show the second best model.

Admixed Population	**Mean Absolute Deviation**							
	RFM	LAM	ELA	LOT	CHR	PCA	SUP	WIN
Three-way admixtures								
SIM3G1	$0.33\pm6.3\times10^{-5}$	$0.31\pm5.7\times10^{-5}$	$0.289\pm4.9\times10^{-5}$	$0.31\pm6.6\times10^{-5}$	$0.331\pm6.4\times10^{-5}$	$0.346\ \pm7.1\times10^{-5}$	$0.333\pm7.1\times10^{-5}$	$0.304\pm5.6\times10^{-5}$
SIM3G2	$0.34\pm6.4\times10^{-5}$	$0.31\pm5.7\times10^{-5}$	$0.334\pm5.6\times10^{-5}$	$0.320\pm6.6\times10^{-5}$	$0.33\pm6.4\times10^{-5}$	$0.346\pm7.1\times10^{-5}$	$0.334\pm6.9\times10^{-5}$	$0.306\pm5.8\times10^{-5}$
SIM3G3	$0.227\pm6.7\times10^{-5}$	$0.18\pm5.7\times10^{-5}$	$0.251\pm6.0\times10^{-5}$	$0.21\pm6.6\times10^{-5}$	$0.33\pm6.4\times10^{-5}$	$0.226\pm7.1\times10^{-5}$	$0.299\pm6.6\times10^{-5}$	$0.183\pm5.7\times10^{-5}$
SIM3MG1	$0.0054\pm6.8\times10^{-5}$	$0.007\pm6.8\times10^{-5}$	$0.134\pm5.0\times10^{-5}$	$0.055\pm6.9\times10^{-5}$	$0.243\pm6.8\times10^{-5}$	$0.074\pm7.6\times10^{-5}$	$0.13\pm7.5\times10^{-5}$	$0.025\pm6.8\times10^{-5}$
SIM3MG2	$0.097\pm6.8\times10^{-5}$	$0.088\pm6.4\times10^{-5}$	$0.146\pm6.0\times10^{-5}$	$0.141\pm6.8\times10^{-5}$	$0.283\pm6.4\times10^{-5}$	$0.148\pm7.4\times10^{-5}$	$0.217\pm7.1\times10^{-5}$	$0.114\pm6.4\times10^{-5}$
SIM3M1NS	$0.0003\pm7.3\times10^{-5}$	$0.0047\pm7.3\times10^{-5}$	$0.138\pm5.7\times10^{-5}$	$0.053\pm7.3\times10^{-5}$	$0.265\pm7.2\times10^{-5}$	$0.044\pm7.7\times10^{-5}$	$0.109\pm7.7\times10^{-5}$	$0.022\ \pm7.2\times10^{-5}$
SIM3S1NS	$0.0004\pm7.1\times10^{-5}$	$0.004\pm7.1\times10^{-5}$	$0.154\pm5.4\times10^{-5}$	$0.057\pm7.1\times10^{-5}$	$0.275\pm7.0\times10^{-5}$	$0.052\pm7.6\times10^{-5}$	$0.12\ \pm7.6\times10^{-5}$	$0.022\pm7.0\times10^{-5}$
Five-way admixtures								
SIM5G1	$0.119\pm4.6\times10^{-5}$	$0.01\pm4.1\times10^{-5}$	$0.164\pm3.6\times10^{-5}$	$0.065\pm4.2\times10^{-5}$	$0.27\pm4.4\times10^{-5}$	$0.16\pm4.9\times10^{-5}$	$0.194\pm4.8\times10^{-5}$	$0.075\pm4.2\times10^{-5}$
SIM5G2	$0.154\pm4.7\times10^{-5}$	$0.066\pm4.0\times10^{-5}$	$0.184\pm4.5\times10^{-5}$	$0.12\ \pm4.2\times10^{-5}$	$0.27\pm4.3\times10^{-5}$	$0.188\pm4.9\times10^{-5}$	$0.228\ \pm4.6\times10^{-5}$	$0.136\pm4.1\times10^{-5}$
SIM5G3	$0.206\pm4.6\times10^{-5}$	$0.131\pm3.9\times10^{-5}$	$0.201\pm4.2\times10^{-5}$	$0.18\pm4.3\times10^{-5}$	$0.27\ \pm4.3\times10^{-5}$	$0.221\pm4.7\times10^{-5}$	$0.253\pm4.4\times10^{-5}$	$0.183\pm4.0\times10^{-5}$
SIM5MG1	$0.318\pm4.3\times10^{-5}$	$0.232\pm4.2\times10^{-5}$	$0.259\pm3.6\times10^{-5}$	$0.237\pm4.3\times10^{-5}$	$0.3\pm4.2\times10^{-5}$	$0.311\pm4.6\times10^{-5}$	$0.305\pm4.4\times10^{-5}$	$0.251\pm4.1\times10^{-5}$
SIM5MG2	$0.3\pm4.2\times10^{-5}$	$0.247\pm4.2\times10^{-5}$	$0.288\pm4.2\times10^{-5}$	$0.251\pm4.2\times10^{-5}$	$0.287\ \pm4.2\times10^{-5}$	$0.304\pm4.9\times10^{-5}$	$0.296\pm4.5\times10^{-5}$	$0.26\pm4.1\times10^{-5}$
SIM5M1NS	$0.141\pm4.6\times10^{-5}$	$0.012\pm4.6\times10^{-5}$	$0.191\pm3.8\times10^{-5}$	$0.053\pm4.5\times10^{-5}$	$0.231\pm4.5\times10^{-5}$	$0.171\pm4.7\times10^{-5}$	$0.141\pm4.7\times10^{-5}$	$0.09\pm4.2\times10^{-5}$
SIM5S1NS	$0.268\pm4.2\times10^{-5}$	$0.255\pm4.4\times10^{-5}$	$0.254\pm3.7\times10^{-5}$	$0.256\pm4.4\times10^{-5}$	$0.25\pm4.4\times10^{-5}$	$0.277\pm4.4\times10^{-5}$	$0.27\pm4.8\times10^{-5}$	$0.253\ \pm4.4\times10^{-5}$

4.3.1.1 Healthy, three-way admixtures formed by a single admixture event

In the context of local ancestry, predictions should be identical to the true ancestry (or ground truth). For confusion matric metrics, high metric values show that model predictions are a good representation of the ground truth. Based on the ability of a model to correctly estimate segments with CEU origins (TPR), WINPOP exceptionally performed in all healthy three-way single-wave admixtures. It was followed by SUPPORTMIX in SIM3G1 and SIM3G2, and by LAMP-LD in SIM3G3 (**Figure 4.2a**). Although the difference is small, according to the accuracy (ACC) and the Mathews correlation coefficient (MCC), ELAI performed better than WINPOP when classifying segments with CEU origins of SIM3G1 population (**Figures 4.3a** and **4.4a**), see for instance, the MCC values: 0.3503 (ELAI) versus 0.3440 (WINPOP) and ACC values: 67.60% (ELAI) versus 67.52% (WINPOP). **Table 4.17** represents the TPR and FPR evaluated when model predictions for each ancestry were compared to the ground truth in healthy single-wave admixtures. Although the FPR of WINPOP is not the lowest, its TPR is the highest given CEU estimates of SIM3G1, SIM3G2 and SIM3G3 populations. Thus, we note that all the considered measures (TPR, ACC, MCC and ROC) agree that WINPOP is best in identifying CEU copies of SIM3G2 and SIM3G3 (**Figures 4.2a**, **4.3a**, **4.4a** and **Table 4.17**). Although LAMP-LD is second best according to ACC and MCC, according to TPR it was surpassed by SUPPORTMIX in predicting the CEU segments of SIM3G2 (**Figures 4.2a**, **4.3a**, **4.4a** and **Table 4.17**). Considering the CEU segments of SIM3G1 and SIM3G2, our findings suggest that models should be improved so as to increase the TPR and reduce the FPR. Nevertheless, the overall model performance is good (**Table 4.17**).

Regardless of having the highest TPR when identifying CHB copies for SIM3G1 and SIM3G2 (**Figure 4.2b, Table 4.17**), ELAI is almost random in SIM3G1 and performs below average in SIM3G2, that is, FPR is higher than TPR (**Table 4.17**). When considering the ROC space (TPR and the FPR), LOTER and CHROMOPAINTER performed better than other models in CHB of SIM3G1 and SIM3G2 (**Figure 4.2b** and **Table 4.17**) while, based on ACC, WINPOP and SUPPORTMIX are better (**Figure 4.3b**). However, according to MCC, LOTER and SUPPORTMIX are best in CHB copies of SIM3G1 and ELAI and SUPPORTMIX in CHB copies of SIM3G2. The two leading models in predicting segments with CHB origins in SIM3G3 are LAMP-LD and WINPOP (**Figures 4.3b** and **4.4b**). In contrast, TPR and FPR suggests LOTER and LAMP-LD (**Figure 4.2b** and **Table 4.17**). RFMIX performed poorly, it predicted other ancestries as CHB more often than it could correctly estimate CHB in SIM3G1, while LAMP-LD was almost random (**Table 4.17**).

Based on the TPR, in SIM3G1, RFMIX leads in predicting segments with YRI origins while,

in SIM3G2 and SIM3G3, ELAI leads in predicting segments with YRI origins (**Figure 4.2c**). When considering the trade-off between the gains (TPR) and loses (FPR) then LAMP-LD and WINPOP outperformed ELAI when predicting YRI segments in the ancient three-way admixture (**Figures 4.2c** and **4.4c** and **Table 4.17**). However, based on the ACC, in SIM3G1, SUPPORTMIX and WINPOP lead in predicting such YRI segments, yet in SIM3G2, LOTER and SUPPORTMIX lead and finally, LOTER and WINPOP lead in SIM3G3 (**Figure 4.3c**). In any case (regardless of the metric and ancestry), WINPOP is one of the two topmost models in estimating the number of copies from a particular population in SIM3G3 (**Table 4.17**).

Generally, in healthy three-way single-wave admixtures formed within 100 generations, all model predictions slightly agree [243] (or have no relationship [241]) with the ground truth (**Tables 4.12** and **4.14**). This concurs with the overall accuracy (OACC) and the F_1-micro average where all values are below 0.50 (**Tables 4.13** and **4.15**). **Table 4.16** shows how each model performed based on the mean absolute error. Now, given a recent three-way admixture: SIM3G1, the two leading models are WINPOP and ELAI (**Tables 4.12, 4.14, 4.15** and **4.16**) while, in an admixture formed 100 generations ago: SIM3G2, WINPOP and LAMP-LD lead (**Tables 4.12**, **4.14** **4.15** and **4.16**). Contrastingly, according to the OACC metric, the two leading models in SIM3G1 and SIM3G2 are WINPOP and LOTER (**Table 4.13**). In SIM3G3, LAMP-LD (**Table 4.16**) and WINPOP lead with predictions that are moderately related to the ground truth, $0.41 \leq$ OMCC < 0.6 [241] (**Table 4.12**).

4.3.1.2 Healthy five-way single-wave admixtures

LAMP-LD and RFMIX lead in correctly predicting segments with CEU (**Figures 4.2a**, **4.3a** and **4.4a**), CHB (**Figures 4.2b**, **4.3b** and **4.4b**) and GIH (**Figures 4.2d**, **4.3d** and **4.4d**) origins in SIM5G1 and SIM5G2. This is consistent with the TPR and FPR since LAMP-LD and RFMIX have the highest TPR and lowest FPR as compared to other models when classifying segments with CEU, CHB and GIH origins of SIM5G1 and SIM5G2 (**Table 4.17**). The two also lead in CHB segments of SIM5G3 (**Figures 4.3b** and **4.4b**). Since LOTER is more stable with regards to gains versus loses (TPR and FPR), it outperforms RFMIX such that the two top models in predicting segments with YRI and KHS origins in SIM5G3 are LAMP-LD and LOTER (**Figures 4.3c** and **4.4c**, **Figures 4.3e** and **4.4e**). It is important to note that considering TPR and FPR in predicting segments with CHB origins in SIM5G3, ELAI performs better than RFMIX (**Figure 4.2b** and **Table 4.17**) while, in segments with GIH origins, PCADMIX is better than RFMIX (**Figure 4.2d**). Although LOTER has the lowest FPR given segments with YRI origins in healthy five-way single-wave admixtures (SIM5G1,

SIM5G2 and SIM5G3), its TPR is less than that of RFMIX (**Table 4.17**). We also note that except for LAMP-LD, LOTER and WINPOP, all models performed poorly when identifying KHS copies in all the healthy five-way single-wave admixtures (**Table 4.17**).

Unlike in three-way admixtures where different metrics yield different model ranks, in healthy five-way single-wave admixtures, all considered metrics (OACC, OMCC, F_1-micro, κ and MAD) yield the same conclusions. Since it performed poorly in estimating the KHS ancestry of all healthy five-way single-wave admixtures, RFMIX was surpassed by LOTER (**Table 4.17**). The two leading models in estimating the local ancestry of SIM5G1, SIM5G2 and SIM5G3 are LAMP-LD and LOTER (**Tables 4.12**, **4.13**, **4.14**, **4.15** and **4.16**). For both models, predictions very strongly relate to the ground truth (MCC≥ 0.75: **Table 4.12**) with LAMP-LD predictions agreeing almost perfectly with the ground truth and LOTER predictions substantially agreeing with the ground truth (**Table 4.14**).

4.3.1.3 Healthy three-way multiple-wave admixture simulations

RFMIX and LAMP-LD performed better than other models in estimating CEU (**Figures 4.2a**, **4.3a** and **4.4a**, and **Table 4.18**), CHB copies of SIM3MG1 and SIM3MG2 (**Figures 4.2b**, **4.3b** and **4.4b**, and **Table 4.18**), and YRI copies of SIM3MG1 (**Figures 4.3c** and **4.4c**, and **Table 4.18**). Unlike in three-way single-wave admixtures, RFMIX and LAMP-LD are the two best models overall in SIM3MG1 and SIM3MG2 (**Tables 4.12**, **4.13**, **4.14**, **4.15** and **4.16**). Their predictions agree almost perfectly with the ground truth (**Table 4.14**). However, as compared to LAMP-LD, we noticed that the performance of RFMIX decreases as the length of segments shortens.

4.3.1.4 Healthy five-way multiple-wave admixture simulations

When classifying CHB (**Figures 4.3b**, **4.4b** and **Table 4.18**) and GIH (**Figures 4.2d**, **4.4d** and **Table 4.18**), in five-way multi-point admixtures (SIM5MG1 and SIM5MG2), all classifiers/models are random. Nevertheless, SUPPORTMIX and CHROMOPAINTER performed better than others in correctly estimating segments with CEU origins of SIM5MG1 and SIM5MG2 (**Figure 4.2a** and **Table 4.18**), while LAMP-LD and LOTER performed better in KHS (**Figure 4.2e** and **Table 4.18**), and ELAI and RFMIX performed better in estimating segments with YRI origins of SIM5MG1 and SIM5MG2 (**Figure 4.2c** and **Table 4.18**). ELAI did not perform well in KHS of SIM5MG1; its FPR almost doubles its TPR (**Table 4.18**) and its MCC is negative (**Figure 4.4e**).

Overall, in five-way admixtures (SIM5MG1 and SIM5MG2), LAMP-LD and LOTER performed better than other models (**Tables 4.12**, **4.13**, **4.14** and **4.15**). We observed significant changes in metric values in three- and five-way multiple-wave admixtures. In five-way, estimates slightly agree with the ground truth (**Table 4.12**, **4.13**, **4.14**, **4.15** and **4.16**).

4.3.1.5 Disease-affected and -unaffected admixture simulations with SNPs under selection

In three-way single- and multiple-wave admixtures, RFMIX and LAMP-LD performed better than other models in inferring local ancestry of disease-affected and -unaffected individuals with some SNPs under selection, SIM3M1NS and SIM3S1NS (**Figures 4.2**, **4.3**, **4.4** and **Table 4.19**). Overall, given three-way disease-affected and -unaffected individuals (SIM3M1NS and SIM3S1NS), estimates of the two leading models (LAMP-LD and LOTER) agree almost perfectly with the ground truth (**Tables 4.12**, **4.13**, **4.14**, **4.15** and **4.16**).

Overall, given five-way disease-affected and -unaffected individuals that resulted from a single-wave admixture (SIM5S1NS), nearly all models randomly performed in classifying segments with CHB and GIH origins. We noticed poor performance when classifying segments with KHS origins (**Table 4.19**). This is also supported by **Figure 4.5** which represents the deviations in local ancestry of cases and controls when identifying the five ancestral populations given disease-affected and -unaffected five-way admixture formed by a single-wave: SIM5S1NS and two-wave (multiple-wave) admixture: SIM5M1NS. In comparison to other tested models, in SIM5S1NS, CHROMOPAINTER performed better although almost random in GIH with TPR below 0.2 (**Figures 4.3d** and **4.4d**) and random in KHS and YRI copies (**Figures 4.2e**, **4.4e** and **Table 4.19**,**Figures 4.3c**, **4.4c, 4.5k** and **Table 4.19**). On the other hand, ELAI leads in estimating CHB copies of SIM5S1NS (**Figures 4.2b, 4.4b, 4.5m** and **Table 4.19**) yet, when estimating the segments with CEU origins, PCADMIX (**Figure 4.5l**), LOTER (**Figure 4.5j**) and LAMP-LD (**Figure 4.5n**) performed better.

In SIM5S1NS (the single-wave five-way disease-affected and -unaffected) and SIM5M1NS (five-way disease-affected and -unaffected individuals with some SNPs under selection formed by two admixture waves), all models did not perform well. LAMP-LD and RFMIX performed better in identifying segments of CEU (**Figures 4.2a**, **4.3a**, **4.4a** and **Table 4.19**), CHB (**Figures 4.2b**, **4.3b**, **4.4b** and **Table 4.19**), GIH (**Figures 4.2d**, **4.3d**, **4.4d** and **Table 4.19**) and YRI origins (**Figures 4.2c**, **4.3c**, **4.4c** and **Table 4.19**). Whereas LAMP-LD and LOTER performed better in identifying segments of KHS origins (**Figures 4.2e**, **4.3e**, **4.4e** and **Table 4.19**).

Overall, models performed better in SIM5M1NS than in SIM5S1NS, for example, LAMP-LD (**Figures 4.5g** and **4.5n**) and ELAI (**Figures 4.5f** and **4.5m**). In SIM5M1NS, the two leading models are LAMP-LD and LOTER (**Tables 4.12**, **4.13**, **4.14** and **4.15**). While in SIM5S1NS, the estimates of the two leading models: CHROMOPAINTER and WINPOP slightly agree with the ground truth (**Tables 4.12**, **4.13**, **4.14**, **4.15** and **4.16**).

4.3.2 Model performance given incorrect admixture dates: a recent three-way single-wave admixture

This section presents the results obtained when a population formed 15 generations ago is analysed with 10 generations. **Table 4.20** shows the performance of each of the four models that depend on admixture generations (WINPOP, SUPPORTMIX, RFMIX and ELAI) when deconvoluting ancestry in recently formed healthy three-way admixed individuals (retaining the name SIM3G1) given accurate (15) and inaccurate (10) admixture generations (SIM3G10) based on confusion matrix class metrics. In short, "SIM3G10" represents the dataset in which the admixture is analysed with 10 generations. We observed the following, when identifying the ancestral populations:

a. CEU - based on MCC, the rank of models changed with ELAI moving from the first to the second position, while based on TPR, the rank of models remained the same. Based on F_1-micro, ELAI moved from the second to the fourth position.

b. CHB - according to ACC, TPR, F_1-micro and MCC, the rank remained the same.

c. YRI - based on ACC, there is no change in rank, while for TPR, F_1-micro and MCC, ELAI moved from the first to the third position.

This therefore shows that during local ancestry inference, ELAI is sensitive to the accuracy of admixture generations, particularly in CEU and YRI, while, RFMIX, WINPOP and SUPPORTMIX are less sensitive to admixture date underestimation (**Table 4.20**). As shown in **Table 4.21**, overall metric values of ELAI dropped by 3-5%, while those of SUPPORTMIX, RFMIX and WINPOP dropped by $< 1\%$. Nevertheless, globally the model rank remains the same, except when conclusions are based on the overall accuracy: OACC (**Table 4.21**). Hence, concurring with class metrics, we conclude that ELAI is sensitive to admixture generations underestimation when deconvoluting local ancestry.

Table 4.20: Model performance based on ACC, TPR, F_1-micro and MCC (row-wise) when identifying CEU, CHB and YRI copies (column-wise) in a recent three-way admixture, using 15 (SIM3G1) and 10 (SIM3G10) generations. The model in the row containing the darker grey shade represents the leading model and light grey shade represents the second leading model in that population when identifying copies from the given ancestry.

Metric	Model	CEU		CHB		YRI	
		SIM3G1	SIM3G10	SIM3G1	SIM3G10	SIM3G1	SIM3G10
ACC	WIN	0.6752	0.6749	0.5516	0.5507	0.7433	0.7426
	SUP	0.6132	0.6137	0.5536	0.5547	0.7661	0.7609
	RFM	0.6351	0.6308	0.5256	0.5214	0.6960	0.6829
	ELA	0.6760	0.6375	0.5272	0.5232	0.7359	0.7173
TPR	WIN	0.6057	0.6041	0.3627	0.3612	0.4373	0.4402
	SUP	0.6010	0.6010	0.3734	0.3664	0.2806	0.2879
	RFM	0.4965	0.4761	0.3299	0.3202	0.4910	0.5112
	ELA	0.4914	0.4478	0.4488	0.4341	0.4565	0.4231
F_1	WIN	0.6333	0.6323	0.3956	0.3941	0.3107	0.3115
	SUP	0.5900	0.5903	0.4037	0.3998	0.2409	0.2416
	RFMIX	0.5575	0.5442	0.3601	0.3513	0.2994	0.2990
	ELAI	0.5842	0.5336	0.4345	0.4243	0.3138	0.2836
MCC	WIN	0.3440	0.3434	0.0443	0.0422	0.1803	0.1813
	SUP	0.2244	0.2252	0.0512	0.0510	0.1074	0.1059
	RFMIX	0.2606	0.2518	-0.0120	-0.0225	0.1611	0.1599
	ELAI	0.3504	0.2673	0.0294	0.0178	0.1832	0.1426

Table 4.21: Global performance (OACC, OMCC, Kappa and F_1-micro score) of WINPOP, SUPPORTMIX, RFMIX and ELAI in recent three-way admixtures given 15 and 10 generations.

Admix Population	Model	Confusion matrix metric			
		OACC	OMCC	Kappa	F_1
SIM3G1	WIN	0.4851	0.1933	0.1904	0.4466
	SUP	0.4665	0.1334	0.1328	0.4115
	RFM	0.4283	0.1336	0.1290	0.4057
	ELA	0.4696	0.1839	0.1785	0.4442
SIM3G10	WIN	0.4841	0.1924	0.1895	0.4461
	SUP	0.4647	0.1334	0.1326	0.4106
	RFM	0.4176	0.1258	0.1206	0.3982
	ELA	0.4390	0.1392	0.1349	0.4138

4.3.3 Model performance given skewed reference population sizes

As mentioned in **Section 2.4**, the size of reference ancestral panels may affect the accuracy of local ancestry estimates. Therefore, this section compares the results of estimating the local ancestry of a recent five-way admixture that mimics the South Africans of mixed ancestry using equal (100 individuals, as in **Section 4.3.1.1**) and unequal (unbalanced) ancestral sample sizes. "SIM5G1" represents the dataset in which we analyse using equal ancestral population sizes and "SIM5SKEW" represents the dataset in which we analyse using unequal ancestral population sizes. We increased CEU and YRI samples to 140 and 160, respectively, retained the CHB sample size (100) and reduced GIH and KHS sample sizes to 80 and 20, respectively.

Although changes are small in general, the accuracy of estimates changed in almost all of the considered models (**Figures 4.6a**, **4.6b**, **4.7a**, **4.7b**, **4.8a** and **4.8d**). A huge effect is visualised when ancestry is deconvoluted using the LAMP-LD model; specifically, the KHS predictions changed from having a strong to a moderate relationship with the ground truth (**Figures 4.7e** and **4.8e**) and the YRI predictions changed from an almost perfect (very strong) to just a strong relationship with the ground truth (**Figures 4.7c** and **4.8c**). The same applies to LOTER's KHS predictions that changed from having a strong to a moderate relationship with the ground truth (**Figures 4.7e** and **4.8e**).

Table 4.17: The true and false positive rate of estimating each ancestry given healthy single-wave admixtures. Cell colouring: red represents poor model performance (the FPR is higher than the TPR), orange represents a random classifier (the model cannot discriminate between positives and negatives), green represents an almost random classifier (which can either improve or deteriorate) and white represents a better classifier (the TPR $>$ FPR).

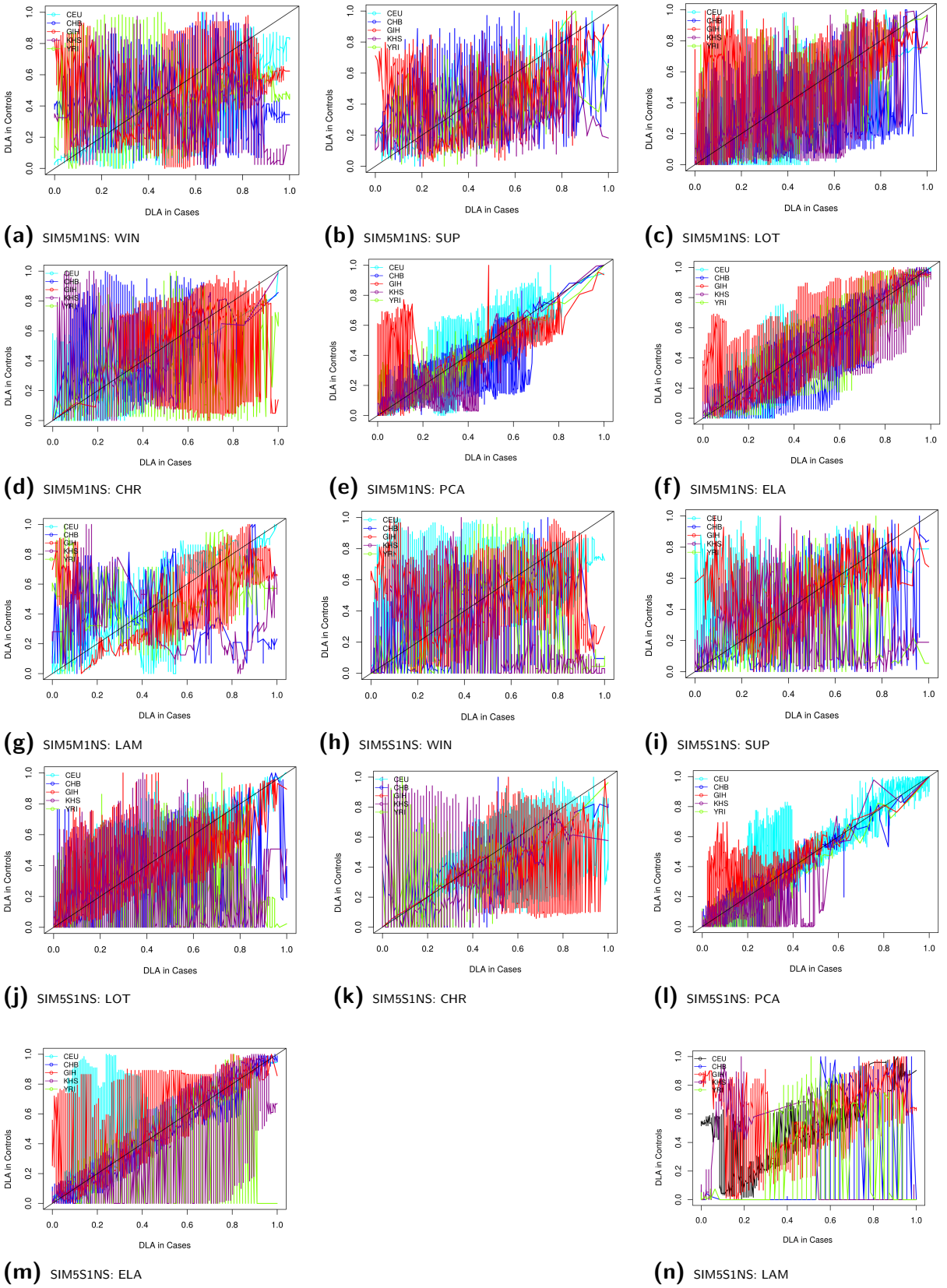
Admix	Anc	Rate	Model							
			WIN	SUP	RFM	ELA	LAM	PCA	LOT	CHR
Three-way admixtures										
SIM3G1	CEU	TPR	0.6057	0.6010	0.4965	0.4914	0.5419	0.5390	0.5687	0.5261
		FPR	0.2648	0.3761	0.2453	0.1648	0.2410	0.3290	0.3081	0.2705
	CHB	TPR	0.3627	0.3734	0.3299	0.4488	0.3639	0.3695	0.4306	0.389
		FPR	0.3200	0.3238	0.3414	0.4195	0.3558	0.3442	0.3718	0.3455
	YRI	TPR	0.4373	0.2806	0.4910	0.4565	0.4664	0.3351	0.2933	0.3917
		FPR	0.2100	0.1599	0.2728	0.2215	0.2292	0.2017	0.1577	0.2261
SIM3G2	CEU	TPR	0.5972	0.5727	0.4936	0.2693	0.5363	0.5332	0.5631	0.5268
		FPR	0.2836	0.3489	0.2567	0.1132	0.2472	0.3349	0.3137	0.2712
	CHB	TPR	0.3760	0.3787	0.3529	0.4489	0.3787	0.3779	0.4351	0.3941
		FPR	0.3176	0.3291	0.3222	0.5040	0.3457	0.3385	0.3691	0.3421
	YRI	TPR	0.4309	0.3222	0.4825	0.5918	0.4808	0.3441	0.2975	0.4086
		FPR	0.1995	0.1797	0.2710	0.2930	0.2263	0.1999	0.1566	0.2230
SIM3G3	CEU	TPR	0.7499	0.5716	0.6464	0.3741	0.7018	0.6874	0.6982	0.5275
		FPR	0.1384	0.2606	0.1345	0.0187	0.0945	0.1939	0.1890	0.2656
	CHB	TPR	0.5901	0.4383	0.5638	0.6118	0.6156	0.5748	0.6271	0.3890
		FPR	0.1682	0.2957	0.1827	0.4025	0.1816	0.2019	0.2356	0.3452
	YRI	TPR	0.7694	0.5278	0.7685	0.8701	0.8613	0.6400	0.5984	0.4014
		FPR	0.1596	0.1995	0.2200	0.2515	0.1785	0.1618	0.1178	0.2270
Five-way admixtures										
SIM5G1	CEU	TPR	0.8887	0.7777	0.9875	0.2060	0.9887	0.8079	0.8535	0.3861
		FPR	0.0460	0.1706	0.0045	0.0013	0.0026	0.0774	0.0576	0.1917
	CHB	TPR	0.8716	0.5561	0.9787	0.1455	0.9852	0.7004	0.7772	0.1682
		FPR	0.0203	0.0975	0.0022	0.0066	0.0023	0.0448	0.0451	0.1285
	GIH	TPR	0.7795	0.4037	0.9767	0.1048	0.9785	0.7006	0.7271	0.2010
		FPR	0.0381	0.1037	0.0036	0.0272	0.0030	0.0823	0.0432	0.1832
	KHS	TPR	0.6595	0.0037	0.0002	0.4215	0.9371	0.0427	0.8393	0.0251
		FPR	0.0443	0.0026	0.0000	0.4698	0.0036	0.0052	0.0385	0.0243
	YRI	TPR	0.8760	0.7911	0.9977	0.8868	0.9934	0.8867	0.8374	0.6176
		FPR	0.1106	0.2815	0.3985	0.2801	0.0242	0.3330	0.0412	0.3569
SIM5G2	CEU	TPR	0.7334	0.6353	0.8024	0.4327	0.8244	0.6964	0.7420	0.3587
		FPR	0.0959	0.2021	0.0350	0.0181	0.0293	0.1025	0.1022	0.1863
	CHB	TPR	0.6725	0.4365	0.7939	0.6644	0.8354	0.5949	0.6577	0.1819
		FPR	0.0734	0.1584	0.0355	0.1149	0.0422	0.0790	0.1016	0.1470
	GIH	TPR	0.5520	0.3297	0.7298	0.3902	0.7721	0.5756	0.5627	0.1752
		FPR	0.0852	0.1487	0.0298	0.0223	0.0386	0.1121	0.0753	0.1797
	KHS	TPR	0.5141	0.0242	0.0059	0.2009	0.7610	0.0291	0.6465	0.0606
		FPR	0.0924	0.0210	0.0067	0.1654	0.0553	0.0129	0.060	0.0580
	YRI	TPR	0.7019	0.6220	0.9522	0.8231	0.8100	0.8008	0.6983	0.5651
		FPR	0.1227	0.2190	0.4218	0.3530	0.0740	0.3230	0.0699	0.3241
SIM5G3	CEU	TPR	0.6137	0.5509	0.5547	0.3459	0.6307	0.5619	0.6260	0.3332
		FPR	0.1332	0.2279	0.0827	0.0143	0.0609	0.1222	0.1478	0.1745
	CHB	TPR	0.5097	0.3596	0.4967	0.6241	0.6223	0.4683	0.5286	0.2167
		FPR	0.1208	0.1960	0.0994	0.1841	0.0924	0.1239	0.1695	0.1681
	GIH	TPR	0.4256	0.2694	0.4301	0.2255	0.5717	0.4457	0.419	0.1775
		FPR	0.1174	0.1726	0.0767	0.0183	0.0782	0.1485	0.1057	0.1730
	KHS	TPR	0.3888	0.0496	0.0421	0.2826	0.5798	0.0241	0.4471	0.1047
		FPR	0.1192	0.0443	0.0430	0.2549	0.1123	0.0208	0.0746	0.0984
	YRI	TPR	0.5713	0.4795	0.8253	0.6793	0.7244	0.6967	0.5503	0.5030
		FPR	0.1358	0.1823	0.3982	0.2712	0.1212	0.3158	0.0967	0.2880

Table 4.18: The true and false positive rate of estimating each ancestry given healthy multiple-wave admixtures. Cell colouring represent the following: red–poor model performance (the FPR is higher than the TPR), orange–random classifier (the model cannot discriminate between positives and negatives), green–the classifier is almost random (classifier might either improve or deteriorate) and white–the classifier is good (the TPR is greater than the FPR).

Admix	Anc	Rate	Model							
			WIN	SUP	RFM	ELA	LAM	PCA	LOT	CHR
Three-way admixtures										
SIM3MG1	CEU	TPR	0.9772	0.8435	0.9976	0.7944	0.9935	0.8977	0.9293	0.6375
		FPR	0.0401	0.1908	0.0073	0.1671	0.0095	0.1059	0.0741	0.3435
	CHB	TPR	0.9498	0.7531	0.9958	0.7863	0.9898	0.8866	0.9053	0.6256
		FPR	0.02694	0.1508	0.0087	0.1946	0.01203	0.1067	0.0709	0.3546
	YRI	TPR	0.3180	0.1680	0.2841	0.4713	0.4478	0.1537	0.2201	0.1737
		FPR	0.0078	0.0349	0.0007	0.0302	0.0017	0.0065	0.0152	0.0205
SIM3MG2	CEU	TPR	0.8908	0.7463	0.8874	0.6483	0.8928	0.8404	0.8488	0.6001
		FPR	0.1297	0.2563	0.0818	0.0248	0.0716	0.1790	0.1562	0.3198
	CHB	TPR	0.8136	0.6353	0.8728	0.8708	0.8761	0.7990	0.8117	0.5498
		FPR	0.1301	0.2404	0.0983	0.2937	0.1139	0.1700	0.1846	0.3555
	YRI	TPR	0.4469	0.3039	0.5442	0.7494	0.5819	0.3340	0.2891	0.3758
		FPR	0.0527	0.0841	0.0630	0.0725	0.0505	0.0440	0.0437	0.0804
Five-way admixtures										
SIM5MG1	CEU	TPR	0.4221	0.4993	0.3858	0.1122	0.3910	0.3930	0.4065	0.4375
		FPR	0.2094	0.3124	0.1930	0.0563	0.1930	0.2355	0.2089	0.2214
	CHB	TPR	0.1725	0.1991	0.1729	0.0655	0.1747	0.182	0.1870	0.2059
		FPR	0.1500	0.1964	0.1462	0.0539	0.1485	0.1638	0.1636	0.1885
	GIH	TPR	0.2013	0.1929	0.1967	0.0582	0.2019	0.2221	0.1883	0.2212
		FPR	0.1934	0.1932	0.1908	0.0576	0.1948	0.2252	0.1835	0.2291
	KHS	TPR	0.3601	0.0089	0.0009	0.2781	0.5256	0.0377	0.4867	0.0416
		FPR	0.2015	0.0044	0.0014	0.5388	0.2729	0.0233	0.2576	0.0261
	YRI	TPR	0.2468	0.305	0.6093	0.7539	0.2140	0.402	0.1035	0.4635
		FPR	0.1243	0.2447	0.4147	0.3795	0.0349	0.3048	0.0415	0.2731
SIM5MG2	CEU	TPR	0.4168	0.4783	0.3964	0.2773	0.3917	0.3917	0.4063	0.4508
		FPR	0.2317	0.2985	0.2192	0.1510	0.2119	0.2494	0.2258	0.2351
	CHB	TPR	0.1726	0.2159	0.1694	0.3615	0.1721	0.1802	0.1825	0.2096
		FPR	0.1473	0.2017	0.1451	0.2790	0.1434	0.1573	0.1572	0.1805
	GIH	TPR	0.2039	0.2140	0.2134	0.1545	0.2142	0.2386	0.2012	0.2418
		FPR	0.1951	0.2182	0.2030	0.1565	0.2022	0.2303	0.1916	0.2322
	KHS	TPR	0.3313	0.0131	0.0024	0.1252	0.4937	0.0210	0.4556	0.0419
		FPR	0.2123	0.0074	0.0023	0.0817	0.2964	0.0324	0.2807	0.0268
	YRI	TPR	0.2032	0.3191	0.5316	0.4065	0.0462	0.3520	0.0556	0.4056
		FPR	0.1181	0.2212	0.3695	0.2651	0.03078	0.2917	0.0374	0.2508

Table 4.19: The true and false positive rate of estimating each ancestry given disease-affected and -unaffected admixed individuals with some SNPs under selections. Cell colouring represent the following: red–poor model performance (the FPR is higher than the TPR), orange–random classifier (the model cannot discriminate between positives and negatives), green–the classifier is almost random (it may either improve or deteriorate) and white–the classifier is good (the TPR is greater than FPR).

Admix	Anc	Rate	Model							
			WIN	SUP	RFM	ELA	LAM	PCA	LOT	CHR
Three-way admixtures										
SIM3M1NS	CEU	TPR	0.9732	0.8678	0.9998	0.7229	0.9932	0.9465	0.9303	0.6717
		FPR	0.0369	0.1720	0.0006	0.0770	0.0064	0.0766	0.0811	0.4650
	CHB	TPR	0.9528	0.7662	0.9993	0.8980	0.9918	0.9128	0.8895	0.4836
		FPR	0.0210	0.1156	0.0003	0.2642	0.0055	0.0504	0.0582	0.310
	YRI	TPR	0.9612	0.7382	0.9975	0.7757	0.9976	0.8411	0.8692	0.3040
		FPR	0.0071	0.0334	0.0000	0.0152	0.0013	0.0066	0.0182	0.0292
SIM3S1NS	CEU	TPR	0.9739	0.8607	0.9997	0.7118	0.9938	0.9440	0.9304	0.6881
		FPR	0.0371	0.1910	0.0011	0.0938	0.0058	0.0923	0.0859	0.4653
	CHB	TPR	0.9484	0.7340	0.9989	0.8703	0.9910	0.8973	0.8802	0.4298
		FPR	0.02109	0.1166	0.0003	0.2757	0.0047	0.0490	0.0588	0.2827
	YRI	TPR	0.9611	0.6961	0.9984	0.7048	0.9988	0.7930	0.8527	0.1207
		FPR	0.0072	0.0406	0.0000	0.0164	0.0016	0.0119	0.0190	0.0629
Five-way admixtures										
SIM5M1NS	CEU	TPR	0.8995	0.7653	0.9995	0.2915	0.9961	0.9089	0.8725	0.3177
		FPR	0.0464	0.1253	0.0018	0.0225	0.0012	0.0619	0.0413	0.1829
	CHB	TPR	0.8756	0.5966	0.9977	0.5738	0.9975	0.8259	0.8164	0.1231
		FPR	0.0197	0.0807	0.0008	0.2521	0.0007	0.0334	0.0346	0.1088
	GIH	TPR	0.7875	0.472	0.9963	0.5489	0.9929	0.8167	0.7849	0.1486
		FPR	0.0407	0.0903	0.0013	0.2315	0.0011	0.05027	0.0344	0.1326
	KHS	TPR	0.6554	0.5811	0.3068	0.4615	0.9447	0.2655	0.8787	0.5578
		FPR	0.0162	0.0644	0.0000	0.0499	0.0003	0.0093	0.0191	0.4105
	YRI	TPR	0.8707	0.6410	0.9991	0.4670	0.9989	0.8917	0.8451	0.2675
		FPR	0.1512	0.1132	0.3634	0.0863	0.0286	0.3251	0.0419	0.0491
SIM5S1NS	CEU	TPR	0.2467	0.2677	0.2184	0.0864	0.2173	0.2313	0.2501	0.2335
		FPR	0.1288	0.1639	0.1171	0.0473	0.1162	0.1418	0.1334	0.1163
	CHB	TPR	0.0982	0.1200	0.0922	0.3250	0.0922	0.0961	0.1106	0.0945
		FPR	0.0731	0.1014	0.0664	0.2085	0.0661	0.0775	0.0842	0.0765
	GIH	TPR	0.1084	0.1000	0.1017	0.1368	0.1014	0.1162	0.1083	0.1354
		FPR	0.1070	0.1074	0.1055	0.1799	0.1049	0.1184	0.1098	0.1048
	KHS	TPR	0.2546	0.2540	0.1220	0.2335	0.3023	0.1052	0.3065	0.3488
		FPR	0.2752	0.2681	0.1418	0.2721	0.3368	0.1216	0.3323	0.3426
	YRI	TPR	0.5833	0.4655	0.7161	0.5023	0.5333	0.6496	0.484	0.5044
		FPR	0.3267	0.2966	0.4929	0.2139	0.3015	0.4809	0.2643	0.2633



(a) SIM5M1NS: WIN **(b)** SIM5M1NS: SUP **(c)** SIM5M1NS: LOT

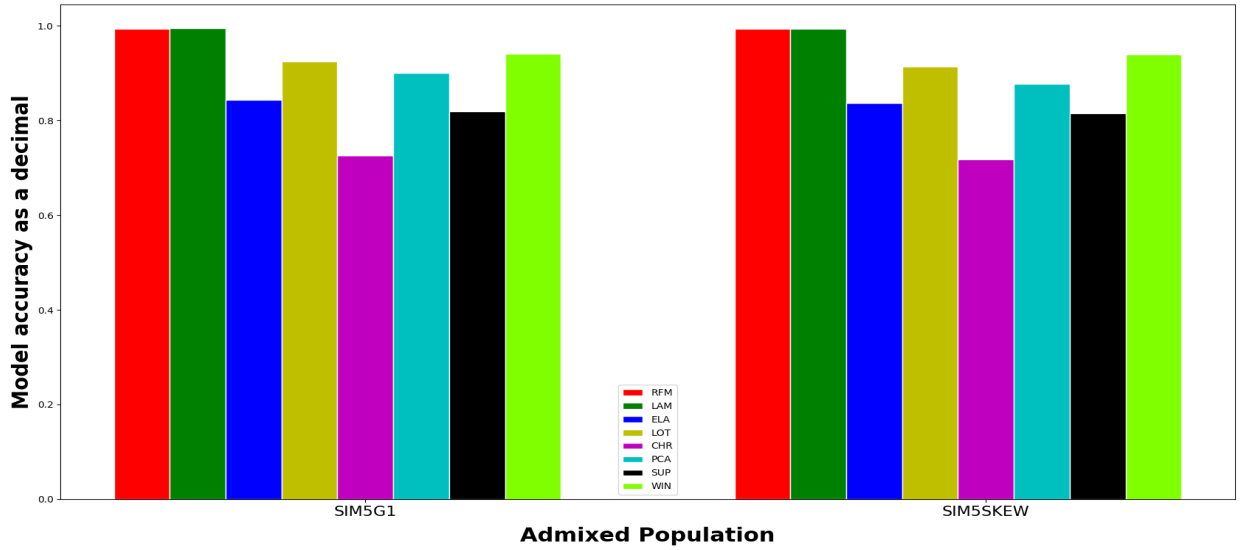
(d) SIM5M1NS: CHR **(e)** SIM5M1NS: PCA **(f)** SIM5M1NS: ELA

(g) SIM5M1NS: LAM **(h)** SIM5S1NS: WIN **(i)** SIM5S1NS: SUP

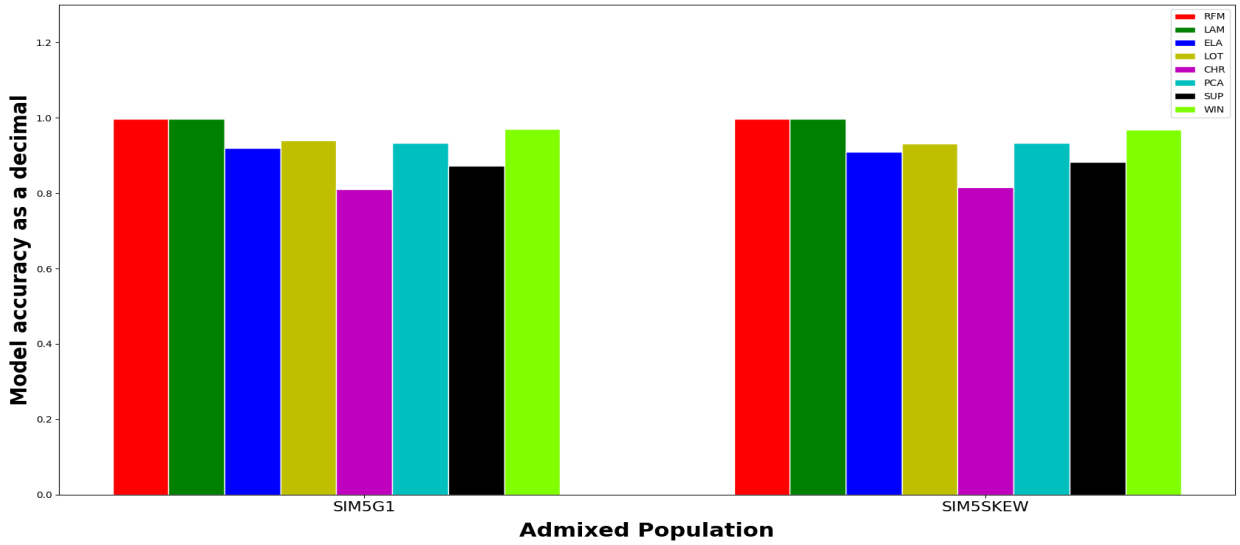
(j) SIM5S1NS: LOT **(k)** SIM5S1NS: CHR **(l)** SIM5S1NS: PCA

(m) SIM5S1NS: ELA **(n)** SIM5S1NS: LAM

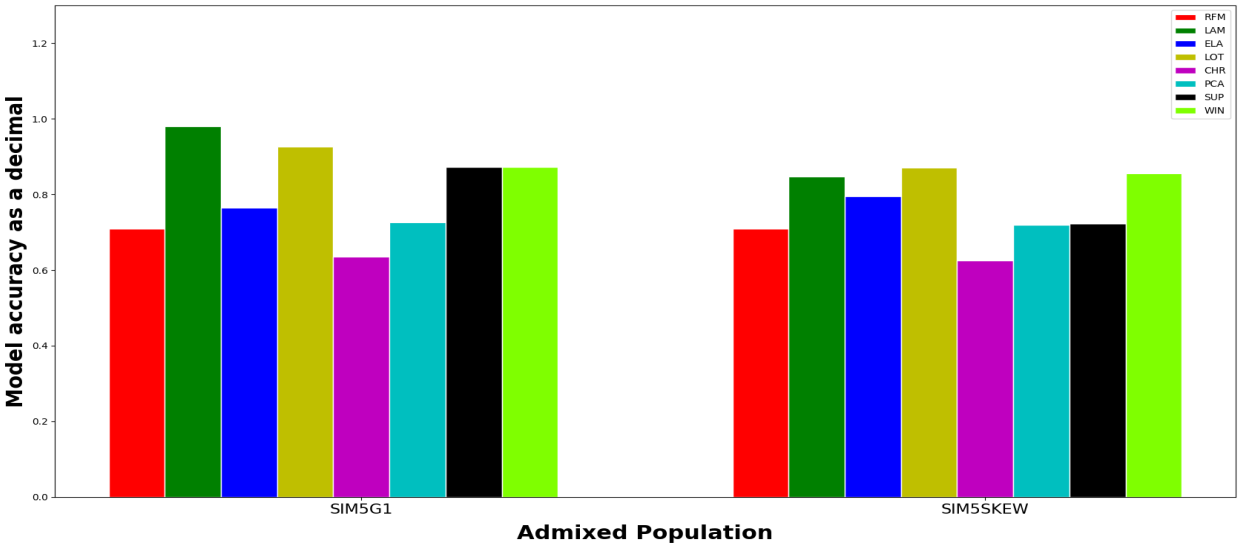
Figure 4.5: Deviations in local ancestry between unhealthy and healthy individuals based on WINPOP, SUPPORTMIX, LOTER, CHROMOPAINTER, PCADMIX, ELAI and LAMP-LD in SIM5M1NS and SIM5S1NS, respectively (row-wise) when identifying CEU, CHB, GIH, KHS and YRI.



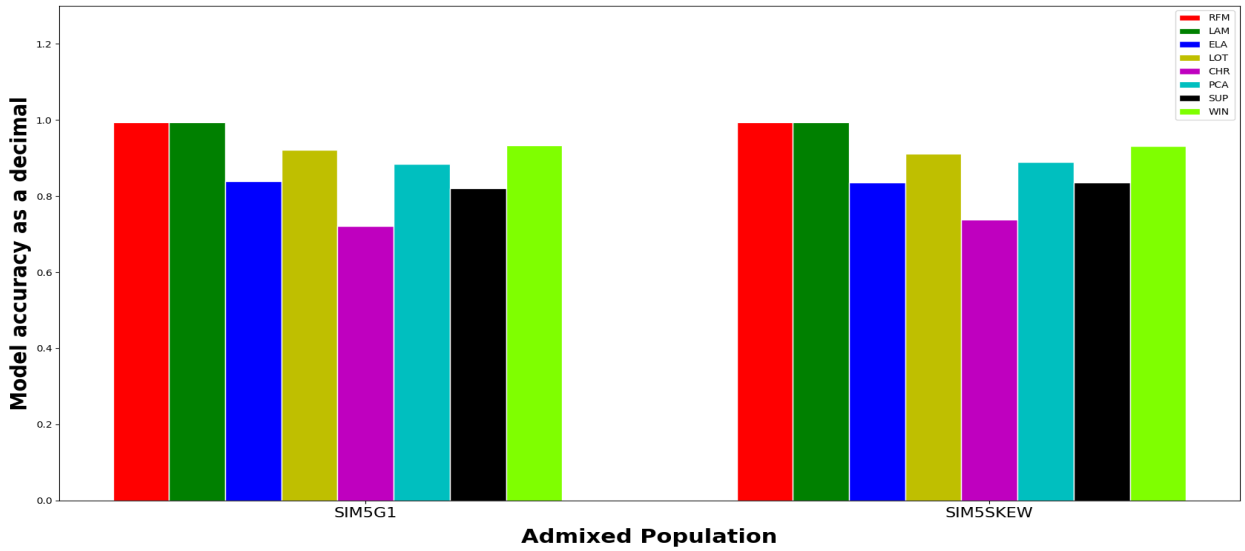
(a) Accuracy of models in estimating CEU



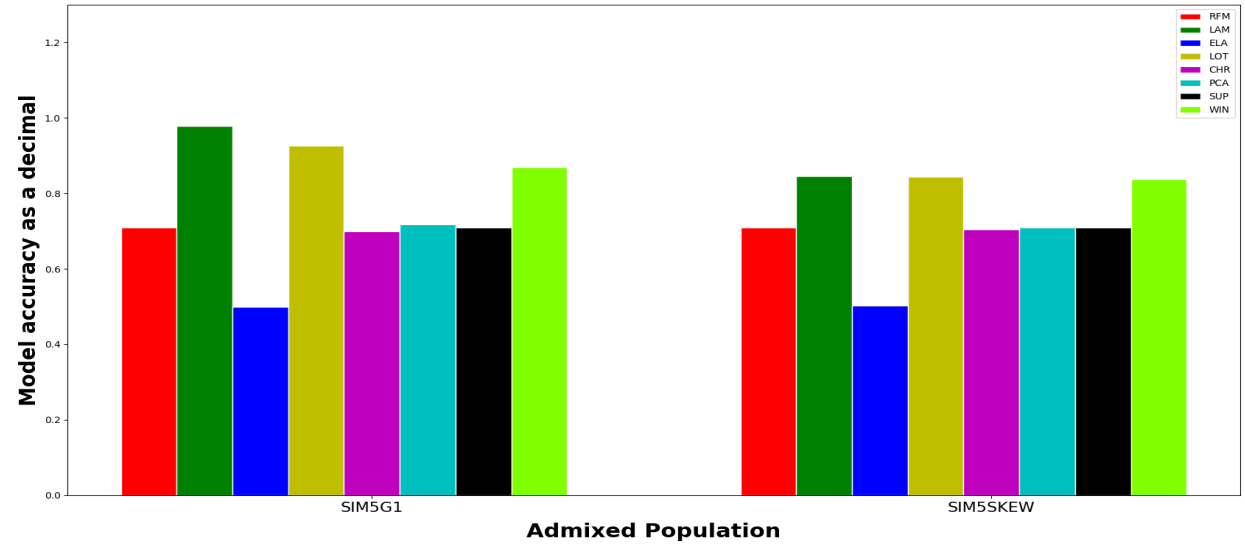
(b) Accuracy of models in estimating CHB



(c) Accuracy of models in estimating YRI

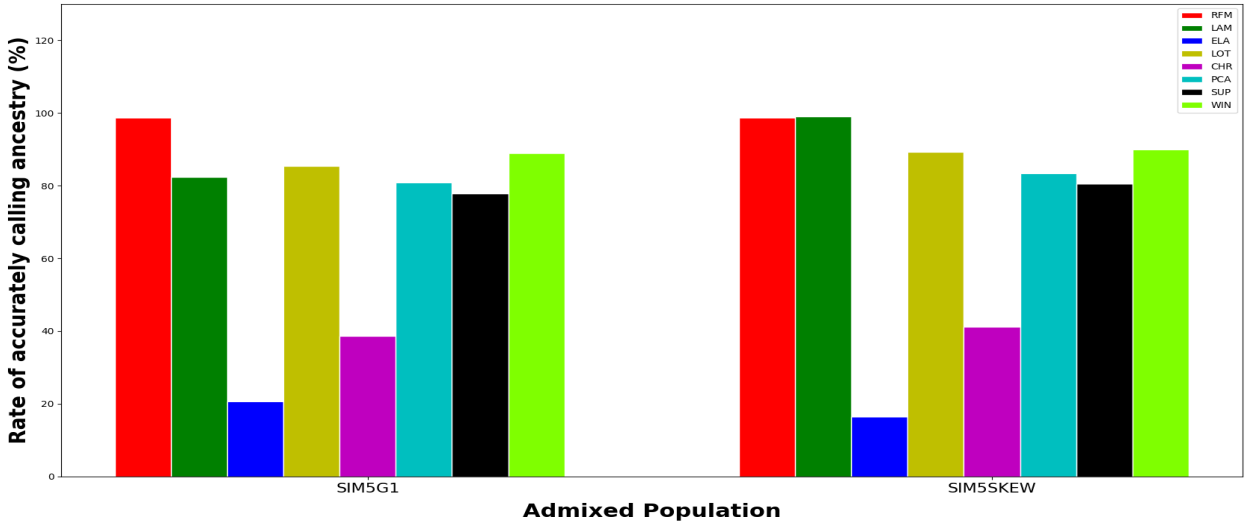


(d) Accuracy of models in estimating CHB

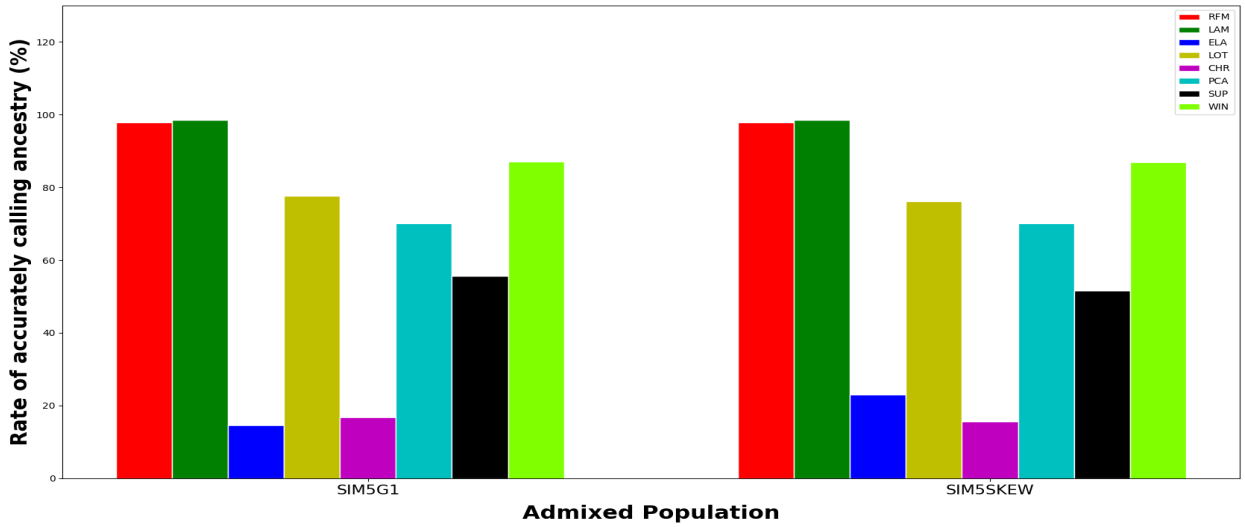


(e) Accuracy of models in estimating KHS

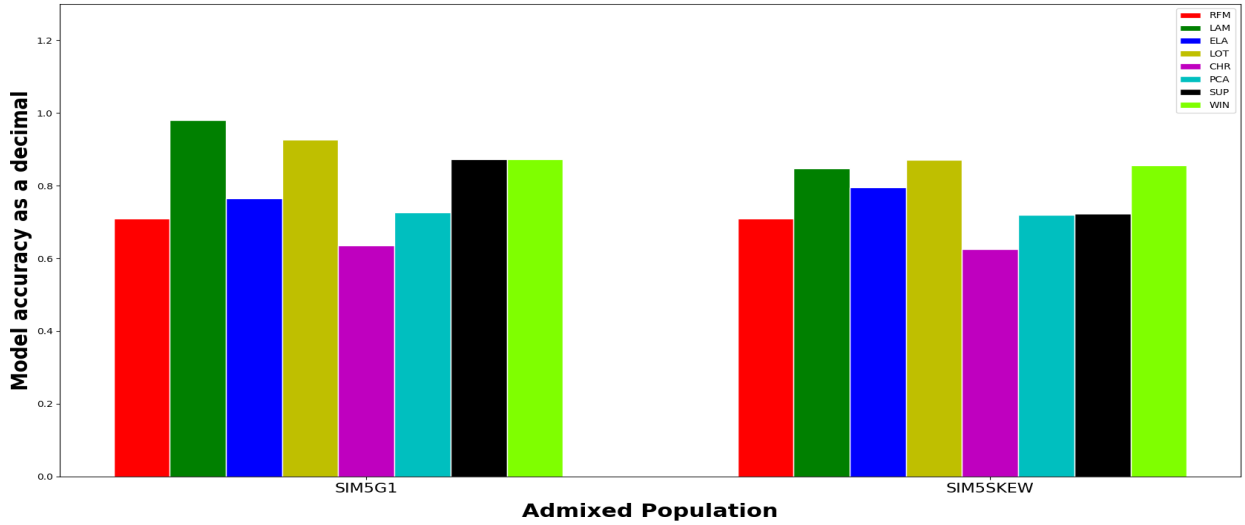
Figure 4.6: Performance of models as measured by accuracy when identifying: (a) CEU, (b) CHB, (c) YRI, (d) GIH and (e) KHS copies for SIM5G1 based on balanced and unbalanced reference ancestral population sizes.



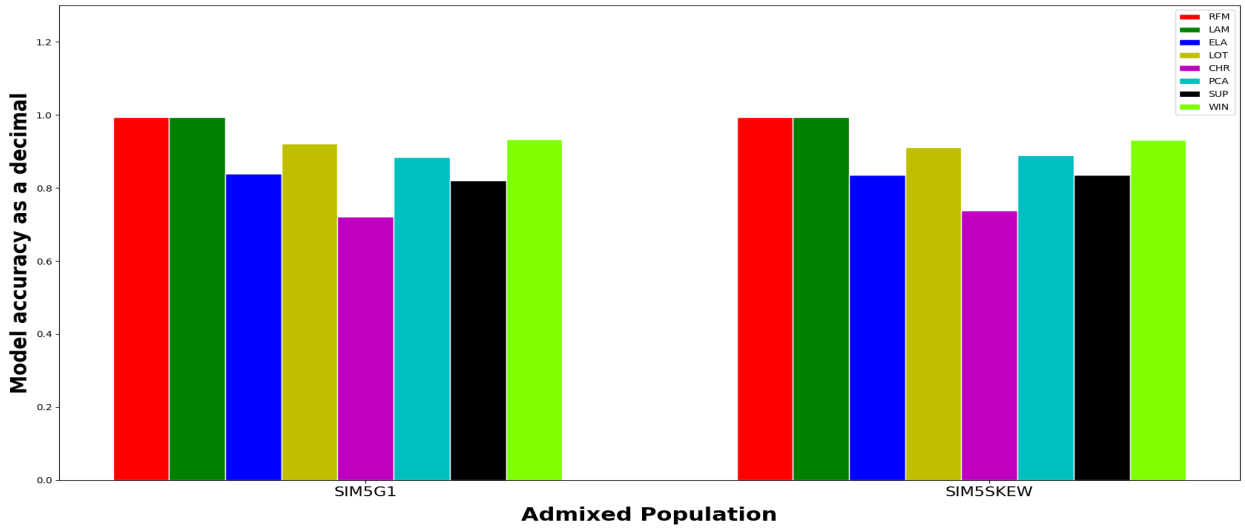
(a) The TPR in classifying CEU segments



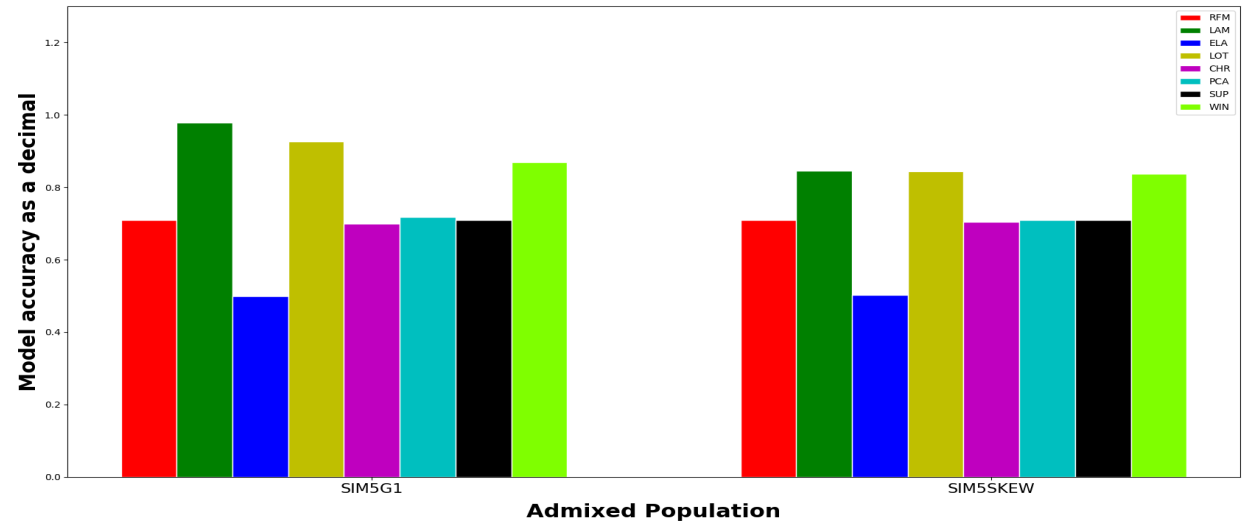
(b) The TPR in classifying CHB segments



(c) The TPR in classifying YRI segments

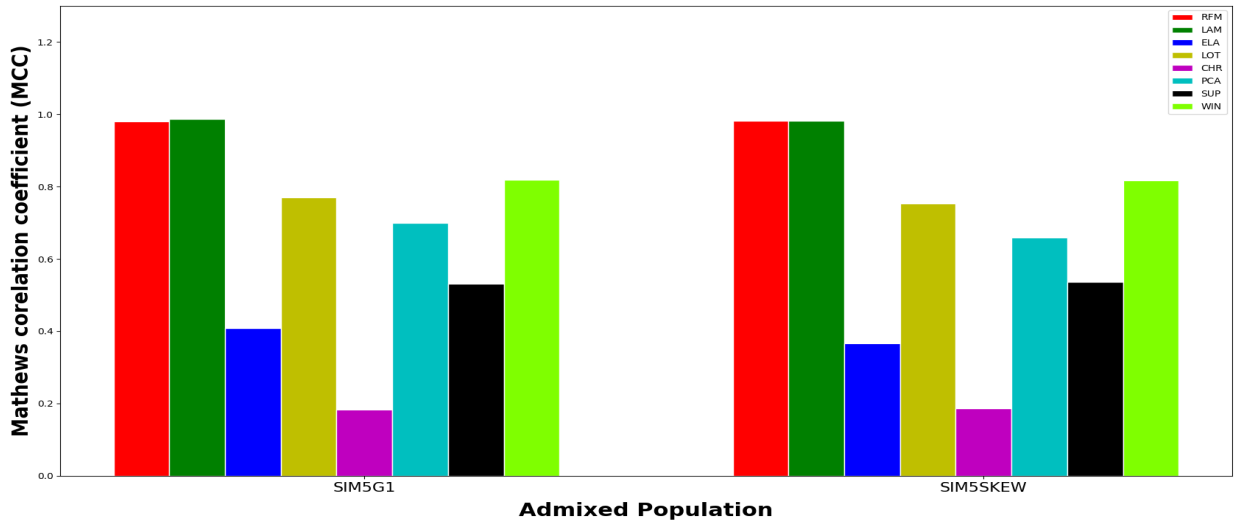


(d) The TPR in classifying GIH segments

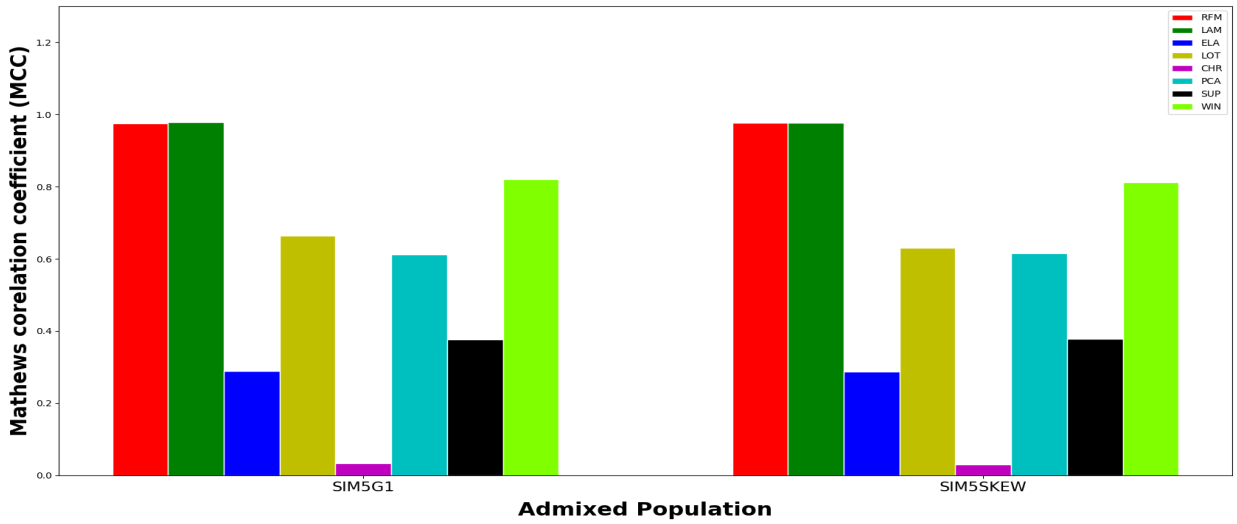


(e) The TPR in classifying KHS segments

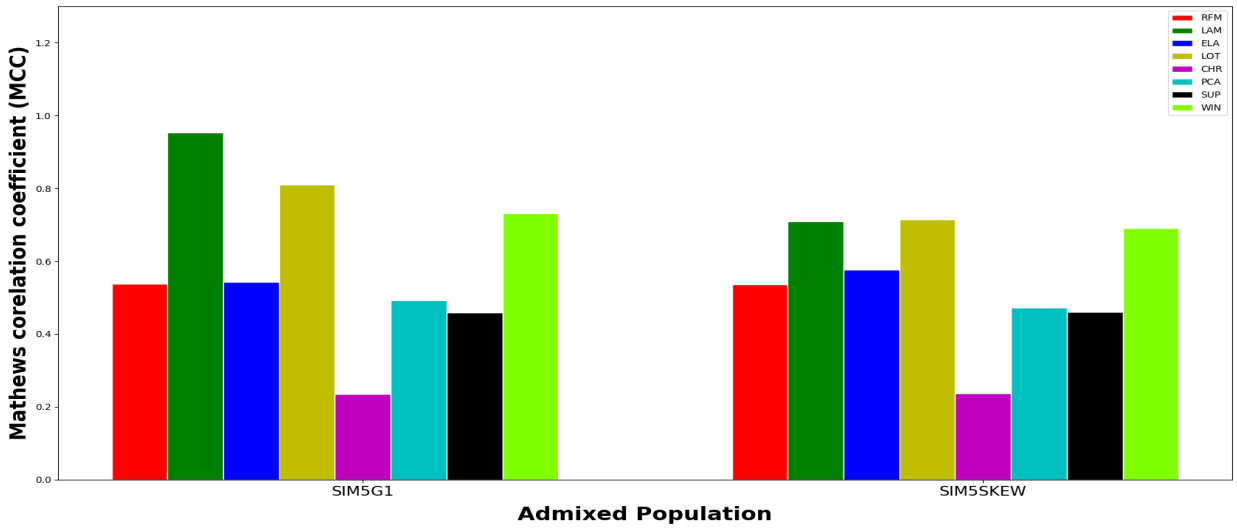
Figure 4.7: Performance of models based on the true positive rate (TPR) metric in identifying: (a) CEU, (b) CHB, (c) YRI, (d) GIH and (e) KHS copies using balanced and unbalanced reference ancestral population sizes.



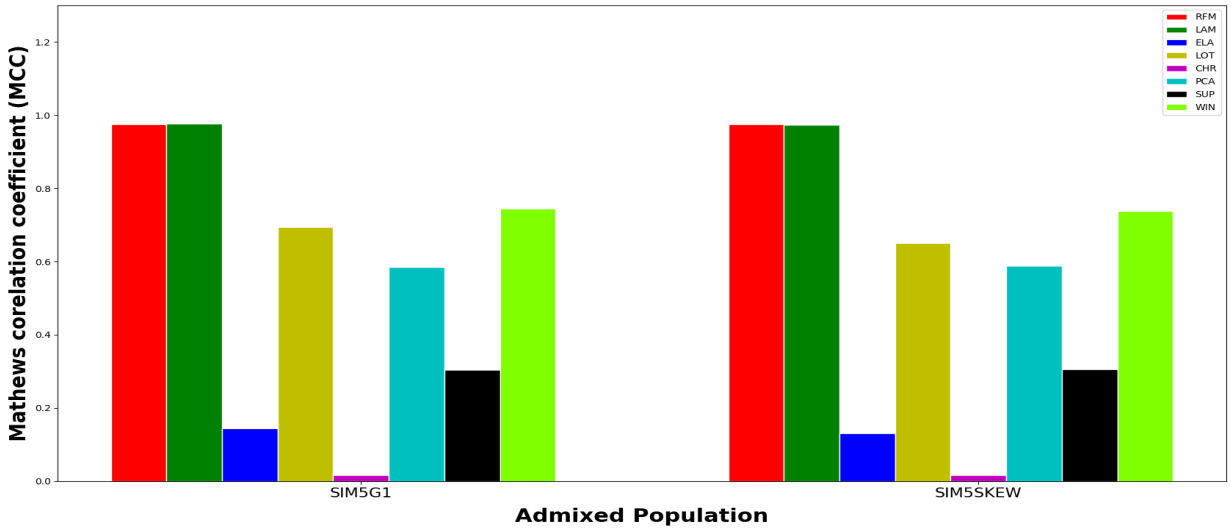
(a) MCC when identifying CEU



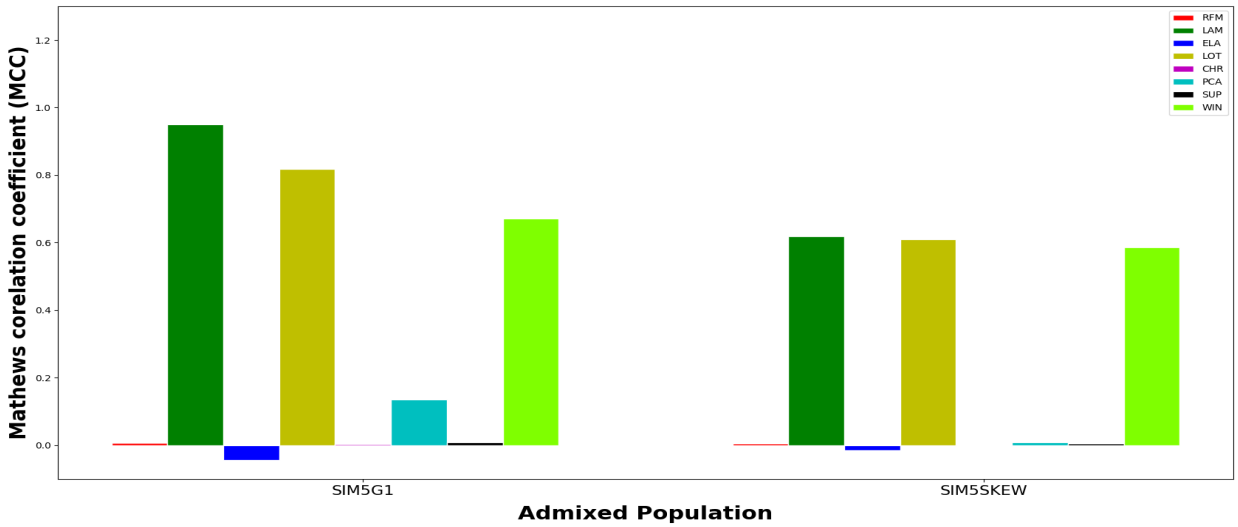
(b) MCC when identifying CHB



(c) MCC when identifying YRI



(d) MCC when identifying GIH



(e) MCC when identifying KHS

Figure 4.8: Performance of models as measured by MCC when identifying: (a) CEU, (b) CHB, (c) YRI, (d) GIH and (e) KHS copies for a recent five-way admixture SIM5G1 using equal and unequal reference ancestral population sizes (SIM5SKEW).

Although it was challenging to detect KHS copies for SIM5G1 using equal population sizes, it worsened when unbalanced ancestral population sizes were used (**Figures 4.8e** and **4.7e**). For example, PCADMIX failed to detect the few segments it detected with equal ancestral population sizes (**Figure 4.7e** and **Table 4.22**). Despite improving the TPR of LOTER in identifying YRI copies, unbalanced reference population sizes also increased the number of false positives for the same ancestry (**Table 4.22**).

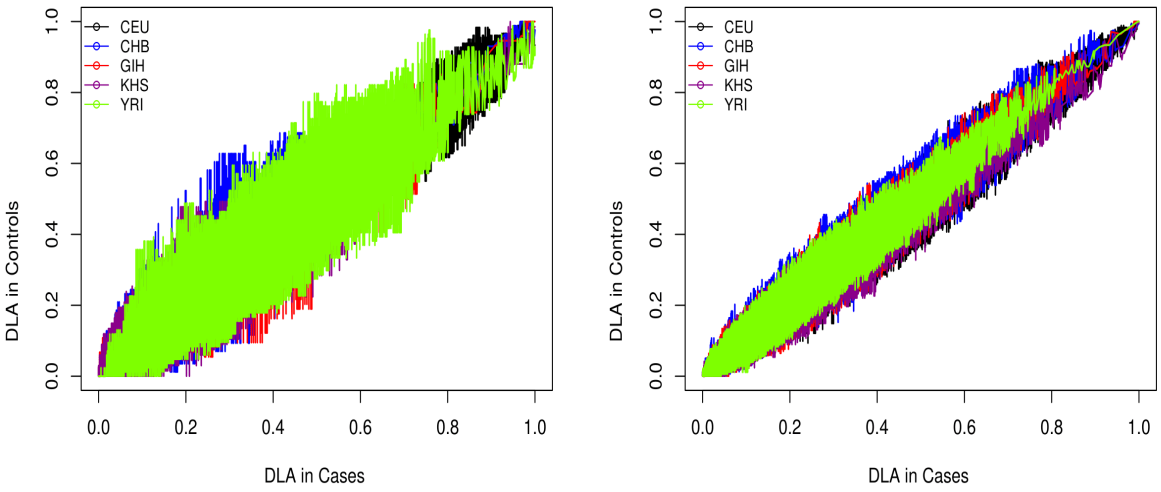
Table 4.22: The true and false positive rate when identifying CEU, CHB, GIH, KHS and YRI given SIM5G1 with equal and unequal population sizes. Red cell colour represents poor model performance (the FPR is higher than the TPR), orange represents a random classifier (the model cannot discriminate between positives and negatives), green represents an almost random classifier (which can either improve or deteriorate) and white represents a better classifier (the TPR is greater than the FPR).

Admix	Anc	Rate	Model							
			WIN	SUP	RFM	ELA	LAM	PCA	LOT	CHR
SIM5G1	CEU	TPR	0.8887	0.7777	0.9875	0.2060	0.9887	0.8079	0.8535	0.3861
		FPR	0.0460	0.1706	0.0045	0.0013	0.0026	0.0774	0.0576	0.1917
	CHB	TPR	0.8716	0.5561	0.9787	0.1455	0.9852	0.7004	0.7772	0.1682
		FPR	0.0203	0.0975	0.0022	0.0066	0.0023	0.0448	0.0451	0.1285
	GIH	TPR	0.7795	0.4037	0.9767	0.1048	0.9785	0.7006	0.7271	0.2010
		FPR	0.0381	0.1037	0.0036	0.0272	0.0030	0.0823	0.0432	0.1832
	KHS	TPR	0.6595	0.0037	0.0002	0.4215	0.9371	0.0427	0.8393	0.0251
		FPR	0.0443	0.0026	0.0000	0.4698	0.0036	0.0052	0.0385	0.0243
	YRI	TPR	0.8760	0.7911	0.9977	0.8868	0.9934	0.8867	0.8374	0.6176
		FPR	0.1106	0.2815	0.3985	0.2801	0.0242	0.3330	0.0412	0.3569
SIM5SKEW	CEU	TPR	0.8987	0.8059	0.9875	0.1641	0.9896	0.8338	0.8921	0.4114
		FPR	0.0501	0.1828	0.0043	0.0004	0.0048	0.1114	0.0809	0.2080
	CHB	TPR	0.8699	0.5162	0.9792	0.2293	0.9846	0.7016	0.7614	0.1557
		FPR	0.0222	0.0819	0.0022	0.0245	0.0027	0.0442	0.0523	0.1208
	GIH	TPR	0.7630	0.3454	0.9764	0.099	0.9704	0.6763	0.6658	0.1770
		FPR	0.0367	0.0745	0.0034	0.0287	0.0027	0.0719	0.0431	0.1600
	KHS	TPR	0.5095	0.0013	0.0002	0.4670	0.4764	0.0023	0.4966	0.0124
		FPR	0.027	0.0010	0.0000	0.4838	0.0027	0.0014	0.0131	0.0122
	YRI	TPR	0.9164	0.8332	0.9976	0.8709	0.9945	0.8649	0.9124	0.6459
		FPR	0.1671	0.3180	0.3989	0.2335	0.2074	0.3343	0.1444	0.3817

Based on the global (overall) performance of models, the two leading models given unbalanced reference population sizes are LAMP-LD and WINPOP (**Tables 4.12**, **4.13**, **4.14** and **4.15**). This shows that WINPOP performed better than LOTER in this case. We observe that the unbalanced population sizes did not only resulted in reduced estimates accuracy in LOTER and LAMP-LD but also that of WINPOP, and PCADMIX. However, it seem not to have an effect on SUPPORTMIX, RFMIX, ELAI and CHROMOPAINTER (**Tables 4.12**, **4.13**, **4.14** and **4.15**).

4.3.4 Application of existing models in the mixed ancestry of South African

From **Section 4.3.1**, we realised that when deconvoluting local ancestry in five-way admixtures, LAMP-LD and LOTER perform better than WINPOP, RFMIX, ELAI, SUPPORTMIX, PCADMIX and CHROMOPAINTER. As a result, we show the plots of the deviations in controls against cases of these two models when applied to the real tuberculosis data of the South African coloured population: SAC (**Figure 4.9**). Although LAMP-LD performed better in simulated five-way admixtures, it was outperformed by LOTER in the real Tuberculosis data of the SAC (**Figures 4.9a** and **4.9b**). This could be due to the unbalanced reference population panels we used as it was shown in **Section 4.3.3** that LAMP-LD is greatly affected by the skewness of reference ancestral panels. The reference ancestral population sizes were: 165 CEU, 203 YRI, 137 CHB, 101 GIH and 24 KHS given 733 admixed individuals.



(a) Deviations in case-control: LAMP-LD **(b)** Deviations in case-control: LOTER

Figure 4.9: Deviations in local ancestry between disease-affected and -unaffected individuals when LAMP-LD and LOTER (two best models) are applied to the real Tuberculosis data of the South Africans of mixed ancestry to estimate CEU, CHB, GIH, KHS and YRI ancestries.

4.4 Discussion

Regardless of the disease status (healthy/controls or case-control individuals), all considered models performed better in multiple-wave than single-wave three-way admixtures (**Tables 4.12**

and **4.14**). This is possibly because the multiple waves we simulated are from different ancestral populations and do not consist of more than one pulse from a single source population which is challenging [245, 246]. Conversely, consistent with previous studies, models performed better in healthy single-wave than in healthy multiple-wave five-way admixtures [245] while, in five-way case-control individuals, models performed better in multiple-wave admixtures (**Tables 4.12** and **4.14**).

Similarly to Cottin et al. [247] and Medina et al. [245], accuracy deteriorates with a decrease in the length of ancestry segments. In five-way (single- or multiple-wave) admixtures, the shorter the segments length, the less accurate models become. However, this was not the case in three-way admixtures where estimates are better in the shorter segments than they are in longer segments (SIM3G3 vs SIM3G2 and SIM3G1). More so, the accuracy of existing models deteriorates with increase in the number of contributing populations [245, 247]. Nonetheless, when deconvoluting local ancestry with SUPPORTMIX, RFMIX, LAMP-LD, WINPOP, PCADMIX and LOTER, our results differ given the healthy single-wave admixtures formed after 15 and 100 generations. Since all models performed poorly in healthy three-way single-wave admixtures: SIM3G1 and SIM3G2, the increase in the number of ancestral populations did not result in a decrease in the accuracy of estimates in five-way admixtures: SIM5G1 and SIM5G2.

Given all four considered models, the difference between metric values when correct and incorrect admixture dates are used to deconvolve local ancestry in a three-way admixture formed 15 generations ago is small. This could be due to the recency of the admixture we considered since segments are too long to be too different. Nonetheless, our results are consistent with previous studies which have shown that (1) SUPPORTMIX is less sensitive to admixture date underestimates within 20-folds [125] and (2) Underestimating admixture generations may not yield to huge changes in ELAI estimates than it does for overestimating [97]. Given that RFMIX deteriorates with increase in admixture generations and that WINPOP has been performed better than SUPPORTMIX in most of our admixed populations, in three-way single-wave admixed populations whose generations since admixture is less known, the most appropriate model among those that rely on admixture generations is WINPOP (**Table 4.20**).

Our **Section 4.3.3** findings concur with previous studies which reported that RFMIX is not affected by unequal reference population sizes [129]. Finally, we note to which the improvements in the LOTER model have increased the accuracy of estimates in complex multi-way admixed populations, specifically in cases when unbalanced populations sizes are used as reference ancestral panels.

4.4.1 Comparing a model to itself in different admixtures

In accordance with the "no free lunch theorem", we compare a model to itself in different admixture scenarios. **Tables 4.23** and **4.24** tabulate how each considered model performed when estimating local ancestry in different admixture scenarios, based on their independence/dependence on admixture generations, respectively. These tables are based on the PyCM compare method of comparing confusion matrices given different models (**Tables 4.12**, **4.13**, **4.14** and **4.15**). We note that the order of performance of LAMP-LD and LOTER is the same in both three- and five-way admixtures (**Figure 4.10**), where both (1) Performed best in three-way disease-affected and unaffected, followed by healthy multiple-wave and lastly, single-wave admixtures, and (2) In five-way admixtures, they performed best in diseased multiple-wave followed by healthy single-wave and lastly, healthy multiple-wave admixtures. On the other hand, the performance of WINPOP in three-way admixtures concurs with both LAMP-LD and LOTER (**Figure 4.10**). However, unlike LAMP-LD and LOTER, WINPOP performed better in SIM5G1 than it did in SIM5M1NS.

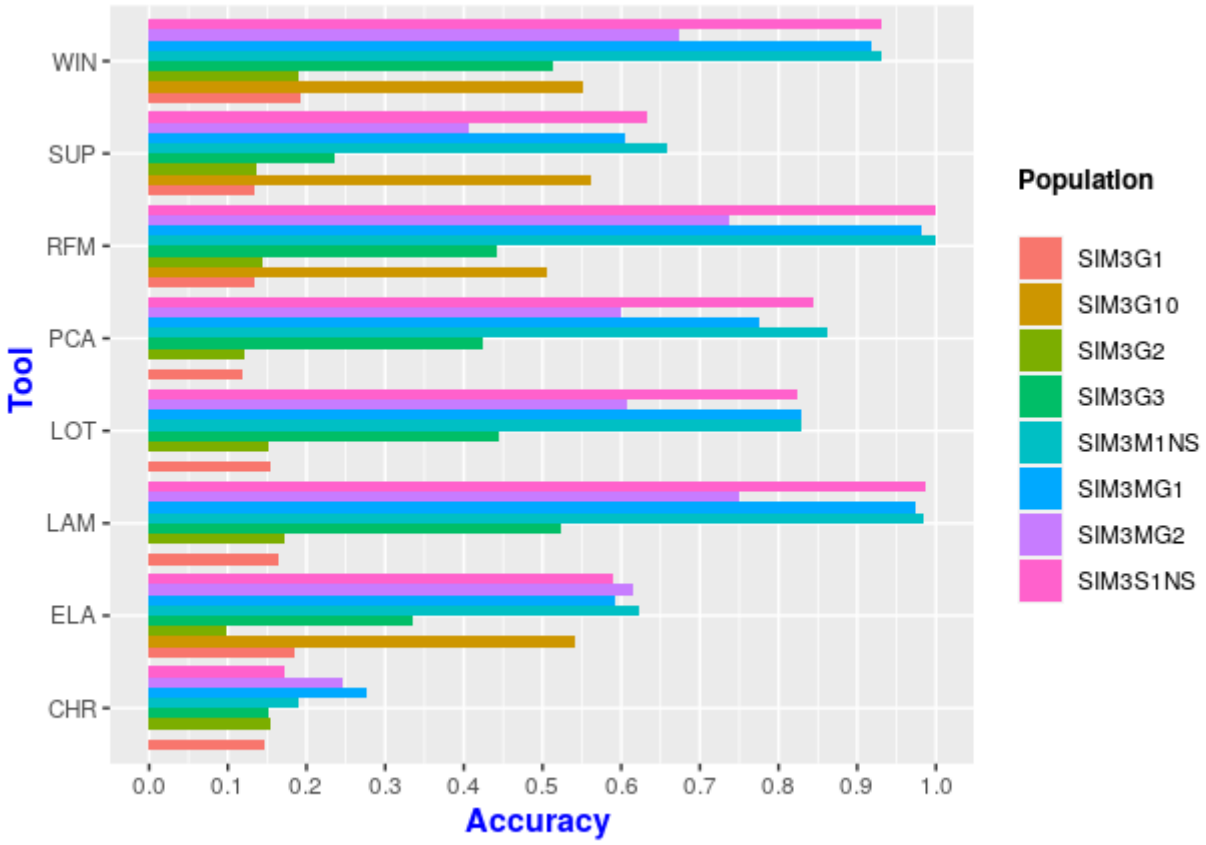


Figure 4.10: Overall accuracy in estimating three-way admixed population segments

4.5 Summary

Based on the qualitative assessment and recent application studies (**Chapter 2**), within the same framework, we assessed eight cutting edge models: LAMP-LD, LOTER, RFMIX, WINPOP, ELAI, PCADMIX, SUPPORTMIX and CHROMOPAINTER using confusion matrix metrics and the mean absolute error on different admixture scenarios. In agreement with the conclusions made in **Chapter 2**, deviations in local ancestry still exist, but we report some improvements has been noticed through the introduction of the LOTER model. Overall, models are affected by: the skewness of reference ancestral populations, the complexity of the admixture dynamics (that is, the number of populations involved) and the statistical or biological parameters they require including the admixture generations. Based on our simulated three-way admixtures, models performed better in the ancient in comparison to recent admixtures, we therefore conclude that it is inadequate to evaluate models based on the number ancestral populations and the skewness of reference ancestral panels on only. Thus, it is necessary to assess different models on a given dataset to select the most appropriate model. However, the performance of LOTER in the mixed ancestry population of South Africa shows that the multi-way local ancestry problem can be efficiently tackled and optimally solved. In **Chapter 3**, we design a framework that paves the way for local ancestry inference model assessment.

Table 4.23: Rank of each model performance in different admixtures: models that do not require admixture generations to deconvolute local ancestry.

Tool	Three-way	Five-way
LAMP-LD	SIM3S1NS	SIM5M1NS
	SIM3M1NS	SIM5G1
	SIM3MG1	SIM5G2
	SIM3MG2	SIM5G3
	SIM3G3	SIM5MG1
	SIM3G1	SIM5MG2
	SIM3G2	SIM5S1NS
LOTER	SIM3S1NS	SIM5M1NS
	SIM3M1NS	SIM5G1
	SIM3MG1	SIM5G2
	SIM3MG2	SIM5G3
	SIM3G3	SIM5MG1
	SIM3G1	SIM5MG2
	SIM3G2	SIM5S1NS
PCADMIX	SIM3M1NS	SIM5M1NS
	SIM3S1NS	SIM5G1
	SIM3MG1	SIM5G2
	SIM3MG2	SIM5G3
	SIM3G3	SIM5MG1
	SIM3G1	SIM5MG2
	SIM3G2	SIM5S1NS
CHROMOPAINTER	SIM3M1NS	SIM5M1NS
	SIM3MG1	SIM5G1
	SIM3MG2	SIM5G3
	SIM3G1	SIM5MG2
	SIM3G3	SIM5S1NS
	SIM3G2	SIM5MG1
	SIM3S1NS	SIM5MG2

Table 4.24: Rank of each model performance in different admixtures: models that require admixture generations to deconvolve ancestry.

Tool	Three-way	Five-way
WINPOP	SIM3S1NS	SIM5M1NS
	SIM3M1NS	SIM5G1
	SIM3MG1	SIM5G2
	SIM3MG2	SIM5G3
	SIM3G3	SIM5S1NS
	SIM3G1	SIM5MG2
	SIM3G2	SIM5MG1
RFMIX	SIM3M1NS	SIM5M1NS
	SIM3S1NS	SIM5G1
	SIM3MG1	SIM5G2
	SIM3MG2	SIM5G3
	SIM3G3	SIM5S1NS
	SIM3G1	SIM5MG2
	SIM3G2	SIM5MG1
ELAI	SIM3M1NS	SIM5G2
	SIM3MG2	SIM5G3
	SIM3S1NS	SIM5G1
	SIM3MG1	SIM5M1NS
	SIM3G3	SIM5MG1
	SIM3G1	SIM5MG2
	SIM3G2	SIM5S1NS
SUPPORTMIX	SIM3M1NS	SIM5M1NS
	SIM3S1NS	SIM5G1
	SIM3MG1	SIM5G2
	SIM3MG2	SIM5G3
	SIM3G3	SIM5MG2
	SIM3G2	SIM5MG1
	SIM3G1	SIM5S1NS

Chapter 5

General discussion, conclusion and recommendations

5.1 General discussion

5.1.1 Introduction

Admixture, recombination, mutation, genetic drift and selection are major drivers of human evolution (**Section 1.2.2**). The admixture process introduces alleles into a specific population, often high ancestry proportions from contributing populations provide room for selection to quicken the adaptation process [248]. In short, the admixture process facilitates natural selection by introducing the patterns of variation in a population upon which selection can act. Similarly to selection, admixture may lead to increased linkage disequilibrium and affect allele frequencies [248, 249, 250]. However, despite the deviations that exist in ancestry at particular regions, existing local ancestry inference models do not account for selection in complex multi-way admixture processes [4]. Therefore, accounting for genetic variation patterns including those due to natural selection may improve the accuracy of estimates.

To date, ancestry inference models depend on genotype/haplotype information of the previously isolated populations to infer the ancestry of every chromosomal segment. As such, the relationships that may exist between these populations may provide insights into the ancestry inference. Since mSPECTRUM [165] estimates were shown to be comparable to HAPMIX, modifying it to account for post-admixture selection and an unknown number of contributing populations (as in HDPStructure) may improve the local ancestry estimates in multi-way

admixtures. As aforementioned in **Chapter 2**, mSPECTRUM assumes all contributing populations have common origins and hence uses infinite state space hidden Markov models. It places a hierarchical Dirichlet process prior on the transition probabilities distribution. Given these modifications, we provide details on how the problem can be solved and described that in the next section.

5.1.2 Population relationships, post-admixture selection and admixture analysis

Consider a pool of founder haplotypes with an unknown number of individuals. We assume $K \geq 2$ groups of individuals diverge from the pool in different numbers at different times to distinct geographical regions. Denote each region for each group of individuals by R_k for $k \in \{1,\ldots,K\}$. We refer to each of the K groups as a population. After some time, a subset of each of the $K-1$ populations migrate to one region $R'_k \in \{R_1,\ldots,R_K\}$. At region R'_k, individuals from each population k breed among themselves in a process called isolated growth. As such, thereafter, each of population k's individual is formed from the set of founder haplotypes through a series of recombination and mutation. Later, the K different population groups interbreed for G generations resulting in a completely new population. This new population is the admixed population while each of the K populations that interbred at R'_k are ancestral (or source). Admixed population individuals may adapt to their environment. Therefore, each admixed haplotype is created by a series of recombination, mutation and probably natural selection. Thus, an admixed haplotype consists of segments that originate from the K previously isolated (ancestral/source) populations. As such, recombination governs the way founder haplotypes and ancestries switch along the admixed individual genome, while mutation and selection determine the allele that is observed at every genomic position in the admixed individual genome. **Figure 5.1** shows the haplotype inheritance model of admixed individuals that were formed when three previously isolated populations (with common origins) interbred G generations ago. The first column represents the pool of founder haplotypes, the second represents three groups of individuals (P_1, P_2 and P_3) that diverge from the pool of founders to the three different regions (R_1, R_2 and R_3). The third column represents a subset of the populations P_2 and P_3 from R_2 and R_3 that migrate to R_1 and the natives, P_1. In column four, P_2, P_3 and a subset of P_1 later interbreed for G generations to form the admixed individuals. **Table 5.1** represents the symbols and notation of the modified mSPECTRUM.

Based on the previous paragraph, the following are the assumptions made when accounting for selection and state persistence in mSPECTRUM:

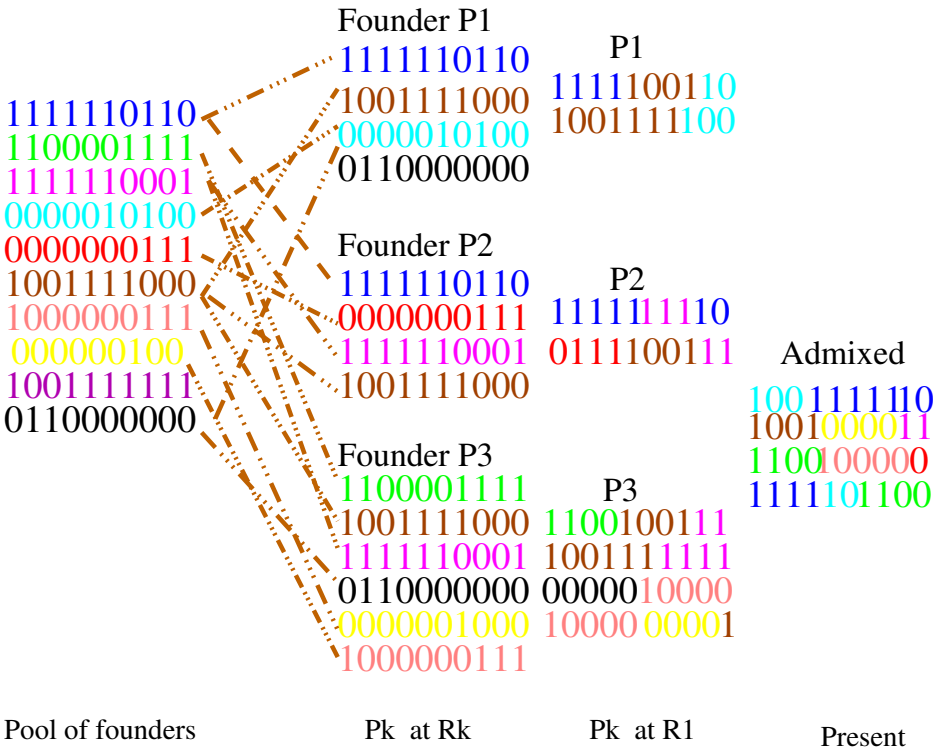


Figure 5.1: Illustrating the haplotype inheritance model in admixed individuals given ancestral populations have common origins. Pk represents P_k and R1 represents R_1, for $k \in \{1,2,3\}$.

1. There exist a pool of an unknown number of founder haplotypes.
2. $K \geq 2$ previously isolated populations formed as a result of migration from a pool of founder haplotypes interbreed G generations ago to form an admixed population whose individuals are a mosaic of segments originating from different ancestral populations.
3. Ancestral or admixed population individuals are haploids.
4. The form of variation in each population is single nucleotide polymorphisms (SNPs) and they are bi-allelic.
5. Mutation is the main source of variation and it does not result in the reversion of base pairs.
6. At every SNP t, an admixed allele $(y_{\imath t})$ is generated by two hidden variables: $z_{\imath t} \in \{1,2,\ldots\}$ and $x_{\imath t} \in \{1,2,\ldots\}$, representing variables that select for the founder haplotype and the ancestry of admixed individual $\imath$ at SNP t.
7. Continuous gene flow (CGF) is a rare event.

Table 5.1: Table of notation and symbols of the modified mSPECTRUM

Symbol	Description
ι	individual index,
ℓ	founder haplotype index, $\ell \in \{1,2,\ldots\}$,
T	number of SNP positions under study,
$F_{\ell t}$	allele of founder haplotype ℓ at SNP t, $F_{\ell t} \in \{0,1\}$, 0 represents the minor allele,
G	time since admixture (in generations),
K	number of previously isolated (ancestral) and unadmixed populations,
k	source population index, $k \in \{1,2,\ldots,K\}$,
n_k	number of individuals in ancestral population k,
N	number of individuals in the admixed population,
$\mathfrak{p}_{tk1}$	frequency of allele 1 in source population k at SNP t,
$y_{\iota t}$	allele of individual ι at SNP t,
$\mathbf{y}_\iota$	haplotype of individual ι, $\mathbf{y}_\iota = (y_{\iota 1},\ldots,y_{\iota T})$,
$x_{\iota t}$	locus-specific ancestry for individual ι at SNP t,
$x_{1:T}$	$x_{1:T} = (x_1,x_2,\ldots,x_T)$,
x_{-t}	all x_r except for x when $r = t$; $x_{-t} = (x_1,x_2,\ldots,x_{t-1},x_{t+1},x_{t+2},\ldots,x_T)$,
η_ℓ	rate at which mutation is inherited from the founder haplotype ℓ to individual ι,
r_t	recombination rate between consecutive SNPs: $t-1$ and t; it is the same for all ι,
r	rate at which breakpoints occur,
d_t	physical distance between consecutive SNPs: $t-1$ and t,
$s_{\ell'\ell}$	probability of switching from copying from founder ℓ' at SNP $t-1$ to founder ℓ at t,
ν_ℓ^k	"initial and background" probability of founder ℓ in population k,
τ_k	parameter that scales for recombination in population k, $\tau_k \in (0,\infty)$,
$\mathfrak{q}_{\iota k}$	ancestry proportion contributed by population k to admixed individual ι,
$\mathfrak{e}_t$	determines the probability of post-admixture selection at SNP t,
$I(\cdot)$	an indicator function.

8. In every population, crossovers between chromosomes occur as a Poisson process at a rate r per genetic distance [78, 80, 92, 117, 165, 163]. Thus, at least one recombination occurs with probability $1-\exp(-d_t r)$ and no recombination occurs with probability $\exp(-d_t r)$, where d_t is the physical distance and r is such that

    a. In ancestral population k, splits occur a rate, $r = r_t \tau_k$ [78, 201].

    b. In admixed individuals, breakpoints occur between consecutive SNPs: $t-1$ and t at a rate $r = G r_t$ per unit distance.

Since each ancestral population haplotype is a mosaic of unknown founder haplotypes, founder haplotypes are discrete state variables that we assume to be a Markov chain. Therefore, the relationship between population k haplotypes and founder haplotypes can be modelled by hidden Markov models (HMMs), such that each admixed haplotype ι is a mosaic of ancestral

blocks originating from founder haplotypes. Thus, we model the inheritance process using HMMs, where admixed alleles are observations generated by two Markov chain processes: the founder haplotype and ancestry sequence; we note that, the two chains are not independent of each other. Conventionally, we seek to determine for each admixed individual haplotype $1 \leq \iota \leq N$, a sequence of ancestries $(x_{\iota t})_{1\leq t\leq T}$ copied from the founder haplotype at t $(z_{\iota t})$ that generated the sequence of observed admixed alleles $(y_{\iota t})_{1\leq t\leq T}$. That is, given an admixed haplotype y_ι, we aim to recover the sequence of ancestries x_ι and founder haplotypes z_ι from which these ancestries were copied

$$(z_{\iota 1}, x_{\iota 1}), (z_{\iota 2}, x_{\iota 2}), \ldots, (z_{\iota T}, x_{\iota T}) \big| y_{\iota 1}, y_{\iota 2}, \ldots, y_{\iota T}. \tag{5.1.1}$$

As a matter of fact, most real world problems involve datasets with state persistence [195, 196]. State persistence has been shown to be severe in Bayesian nonparametrics (BNPs) due to the simplicity of Bayesian models and the flexible nature of hierarchical Dirichlet process hidden Markov models (HDP-HMMs) which mostly cause unrealistic switching of states [195, 196]. Furthermore, most existing inference algorithms cannot adjust well to these unrealistic fast dynamics [195]. Although Beal et al. [192] pioneered the modelling of state persistence in infinite hidden Markov models (iHMM) using a self-transition parameter, it has been reported that this does not fully integrate with inference in BNPs [195]. Actually, existing BNP local ancestry inference models such as mSPECTRUM and HDPStructure do not fully account for state persistence. To fully account for state persistence, previous studies recommended using the sticky HDP-HMMs [195] or the HDP hidden semi-Markov models (HDP-HSMMs) [251, 252]. Therefore, we recommend an mSPECTRUM modification to leverage the use of sticky HDP-HMMs; details of sticky HDP-HMMs are provided in the next section.

5.1.3 Sticky hierarchical Dirichlet process hidden Markov models

Here, we propose how the Sticky hierarchical Dirichlet process hidden Markov models (sticky HDP-HMMs) can be extended from the original HDP-HMMs, described in **Section 2.3.2.2** to account for state persistence through the κ parameters. **Figure 5.2** illustrates sticky HDP-HMMs: GEM is the Griffiths Engen McCloskey distribution or stick breaking process with a parameter α_1; $s_{\ell'}$ is the probability of transition from state ℓ' to any other state. It is a Dirichlet process with a concentration parameter $\alpha_2 + \kappa$ and a base measure $\frac{\alpha_2\beta+\kappa\delta_\ell}{\alpha_2+\kappa}$, such that δ_ℓ is the point mass at ℓ [191]; and H is the prior distribution of emission parameters θ_ℓ that parameterises the emission model distribution F [253]. Placing an additional weight κ on each transition state of Relation (2.3.18) yields (5.1.2) [195]. As for the Chinese restaurant

$$
\begin{aligned}
\beta &\sim \text{GEM}(\alpha_1), \\
s_{\ell'} &\sim \text{DP}\left(\alpha_2 + \kappa, \frac{\alpha_2\beta + \kappa\delta_\ell}{\alpha_2 + \kappa}\right), \\
z_t &\sim s_{z_{t-1}}, \\
\theta_\ell &\sim H, \quad \ell = 1, 2, \ldots, \\
y_t &\sim F(y_t|\theta_{z_t}).
\end{aligned}
\tag{5.1.2}
$$

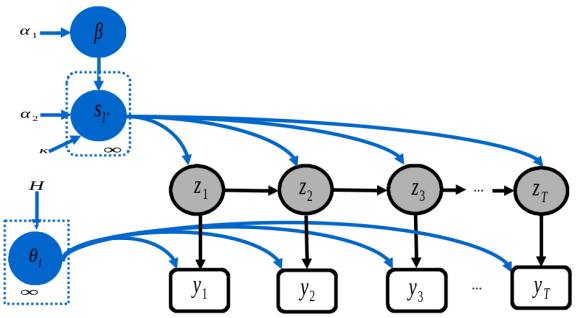


Figure 5.2: The graphical model of the sticky HDP-HMMs: arrows represent the dependence relationship, blue circles and arrows represent the prior of the iHMMs while black arrows and nodes represent the iHMMs. Blue circle nodes represent random variable parameters and grey circle nodes represent unobserved variables that generate observed variables (rectangular nodes). Boundless nodes represent hyper-parameters.

franchise representation problem, the change introduces some loyal customers in each of the restaurants in the franchise [195, 254].

5.1.4 Sticky HDP-HMMs and the modified mSPECTRUM model

As previously mentioned in **Section 5.1.2**, observations or admixed alleles $(y_{\iota t})_{1\leq\iota\leq N, 1\leq t\leq T}$ are generated by two hidden indicator variables which select for the: ancestral population and founder haplotype from which the ancestral population is copied at SNP t, that is, $(x_{\iota t})_{1\leq\iota\leq N, 1\leq t\leq T}$ and $(z_{\iota t})_{1\leq\iota\leq N, 1\leq t\leq T}$. Assuming an unknown number of founder haplotypes and the interbreeding of $K \geq 2$ ancestral populations, we model the haplotype inheritance model in admixed individuals using sticky HDP-HMMs. The transition distribution is a matrix such that, for each ancestral population k, each row is a Dirichlet process. The row defines the probability of transitioning from copying founder haplotype ℓ' to any other founder, ℓ. Since we assume that admixed individuals are independent, we omit the index ι and write, z_t to represent $z_{\iota t}$ and x_t to represent $x_{\iota t}$. **Figure 5.3** is a graphical representation of the mSPECTRUM modified model that accounts for natural selection in local ancestry deconvolution. The natural selection parameter is not shown in the representation since it is contained in the observed information y_t. Similarly to the original sticky HDP-HMM, H is the prior distribution of emission parameters $\theta_{\ell k}$ that parameterises the emission model distribution F, $\alpha_1 \sim \text{Gamma}(a_1, b_1)$ [191, 195, 201], $\alpha_2 + \kappa \sim \text{Gamma}(a_2, b_2)$ [195, 254], $\frac{\kappa}{\alpha_2+\kappa} \sim \text{Beta}(a_3, b_3)$ [195, 254], $\sigma \sim \text{Gamma}(a_4, b_4)$ and $\gamma \sim \text{Gamma}(a_5, b_5)$, where a_1, a_2, a_3, a_4, a_5, b_1, b_2, b_3, b_4 and b_5 are arbitrary constants.

$$
\begin{aligned}
\beta &\sim \text{GEM}(\alpha_1), \\
s_{\ell'} &\sim \text{DP}\left(\alpha_2+\kappa, \frac{\alpha_2\beta+\kappa\delta_\ell}{\alpha_2+\kappa}\right), \quad \ell = 1,2,\ldots \\
z_t &\sim s_{z_{t-1}}, \qquad 1 < t \leq T \\
\varepsilon &\sim \text{GEM}(\sigma), \\
q_{\ell k} &\sim \text{DP}(\gamma,\varepsilon), \qquad \ell = 1,2,\ldots, \quad k = 1,2,\ldots \\
x_t &\sim q_{z_t x_{t-1}}, \qquad 1 < t \leq T \\
\theta_{\ell k} &\sim H, \qquad \ell = 1,2,\ldots, \quad k = 1,2,\ldots \\
y_t &\sim F(y_t|\theta_{z_t x_t}), \quad 1 \leq t \leq T.
\end{aligned}
$$

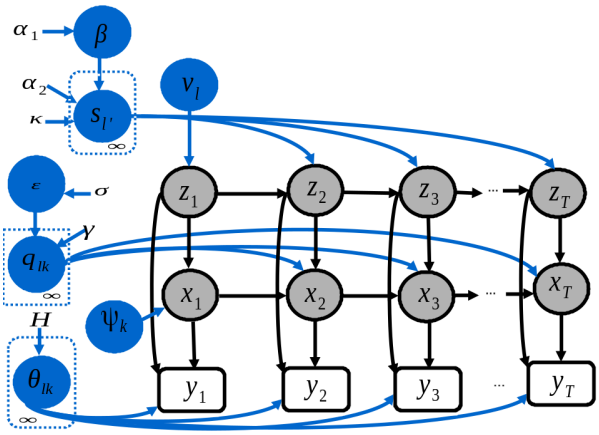


Figure 5.3: A graphical representation of an extension of the mSPECTRUM model to account for post-admixture selection and population relationships based on the modern humans origins in local ancestry. Arrows represent dependence relationship. Blue circles and arrows represent the prior of the iHMM and black represent the iHMM. Blue circle nodes represent random variable parameters and grey circle nodes represent unobserved variables that generate observed variables (rectangular nodes), and boundless nodes represent hyper-parameters.

5.1.5 Probability of observing an admixed allele

Since admixed individuals descend from founders their haplotypes reflect the occurrence or non-occurrence of mutation. As highlighted in **Section 5.1.2**, the modified mSPECTRUM assumes mutation is the chief source of variation [46] and yields no reversion to the original allele. Thus, if the admixed allele is different from the founder allele, then chances are high that a mutation recently occurred in the individual at that SNP (*denovo* mutation). Although variation is introduced by mutation, recombination and gene flow or admixture, mutation has shown to be a major mechanism which selection acts on [46]. We describe the variation introduced by recombination and gene flow or admixture as standing genetic variation[1]. Therefore, observed variables account for a case (1) when mutation is the major mechanism for selection and (2) when selection acts on standing genetic variation. Hence, the probability of the occurrence of selection is higher on a SNP that underwent a recent mutation (not inherited from a parent) than it is the one that did not (we refer to as a non-mutant SNP), such that

$$
P(y_t|F_{\ell t},\eta_\ell,\mathfrak{e}_t) = \left(\frac{1}{1+\exp(-\mathfrak{e}_t)}\right)\left(\eta_\ell I(y_t \neq F_{\ell t}) + (1-\eta_\ell) I(y_t = F_{\ell t})\right), \tag{5.1.3}
$$

where y_t is the allele of an admixed haplotype at SNP t, $F_{\ell t}$ is the allele of founder ℓ at t. Since SNPs are biallelic, the probability of observing an admixed allele given it's a copy from

[1]Standing genetic variation is the presence of alternative alleles at a locus in a population [111, 255].

founder ℓ in ancestry population k is given by

$$\begin{aligned} P(y_t|z_t=\ell,x_t=k) &\propto P(y_t)P(z_t=\ell,x_t=k|y_t) \\ &= P(y_t)P(z_t=\ell|y_t)P(x_t=k|y_t) \\ &= \cancel{P(y_t)}\left(\frac{P(z_t=\ell)P(y_t|z_t=\ell)}{\cancel{P(y_t)}}\right)P(x_t=k) \\ &= P(z_t=\ell)P(y_t|z_t=\ell)P(x_t=k), \end{aligned} \tag{5.1.4}$$

where $P(y_t|z_t)$ is equivalent to Equation (5.1.3), $P(x_t=k)$ is the proportion of ancestry contributed by population k and $P(z_t=\ell)$ is the probability of founder ℓ at SNP t. Since SNPs are biallelic, each admixed allele $y_t \in \{0,1\}$, hence, an admixed allele at every SNP can be drawn from a Bernoulli distribution with parameter $\theta_{\ell k}$, that is,

$$y_t|z_t,x_t \sim \text{Ber}(\theta_{\ell kt}) \tag{5.1.5}$$

where $\theta_{\ell kt} = P(y_t|z_t,x_t)$ as in Equation (5.1.4).

Although mutation is the main mechanism for selection, a subset of mutant and/or non-mutant SNPs (standing variation) are selection targets. In order to account for this selection, we let $0 \leq \frac{1}{1+\exp(-\mathfrak{e}_t)} \leq 1$ such that,

$$\mathfrak{e}_t \sim \begin{cases} \text{Uniform}(0,10) & y_t \neq F_{\ell t} \\ \text{Uniform}(-10,0) & y_t = F_{\ell t} \end{cases}.$$

Figure 5.4 represents a random draw of the selection variable $(\mathfrak{e}_t)$ and their genomic interpretation regarding the admixed SNPs genotype data. Let $P(\mathfrak{e}_t) = \frac{1}{1+\exp(-\mathfrak{e}_t)}$, then $0.0 \leq P(\mathfrak{e}_t) \leq 0.25$ shows SNPs that are non-mutant and not selection target regions, $0.25 \leq P(\mathfrak{e}_t) \leq 0.5$ shows SNPs that are non-mutant and selection target regions, $0.5 < P(\mathfrak{e}_t) \leq 0.75$, shows mutant SNPs that are not selection target regions, and lastly $0.75 < P(\mathfrak{e}_t) \leq 1$, shows SNPs that are mutant and also selection target regions.

5.1.6 Sampling techniques and Dirichlet process properties

Markov chain Monte Carlo (MCMC) algorithms, specifically Gibbs and Metropolis Hastings samplers have been popularly used to determine the posterior distribution of model parameters given observed data in sticky HDP-HMMs. In such cases, the most popular MCMC algorithms have been: collapsed and blocked Gibbs samplers [195, 196, 254]. These sampling techniques

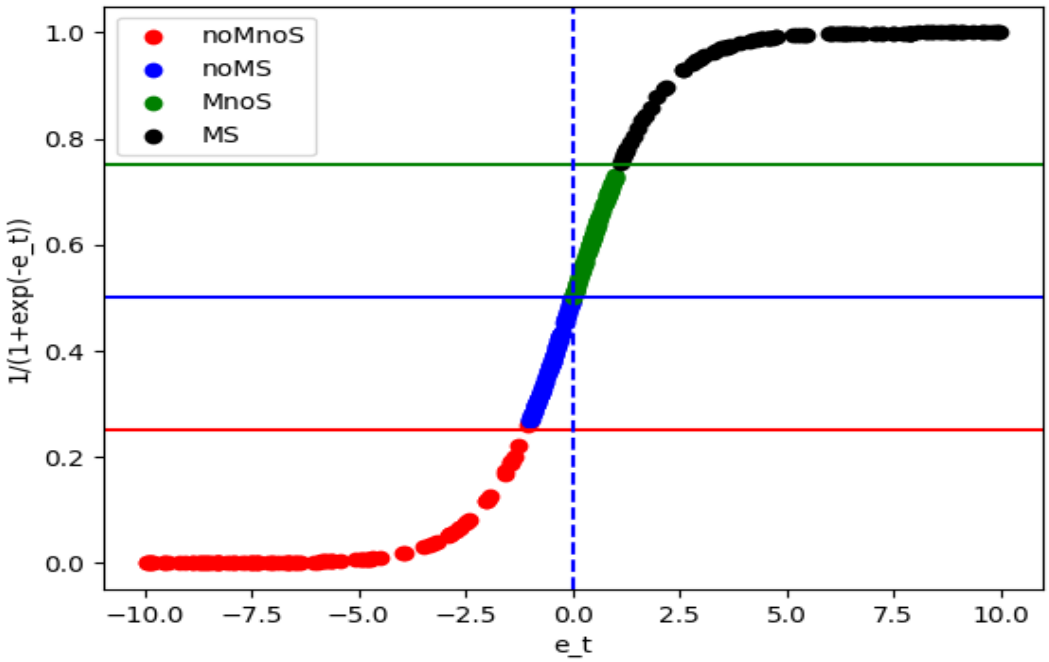


Figure 5.4: A graphical representation of a variable that determines selection in local ancestry inference. MS represents mutant SNPs that are detected as selection targets, MnoS represents mutant SNPs that are not detected as selected SNPs, noMS these are non-mutant SNPs that are detected as selected SNPs and noMnoS are non-mutant SNPs that are not detected as selected SNPs.

use the DP properties including the discreteness, clustering and its posterior probability. In this regard, the collapsed Gibbs sampler is based on the Chinese restaurant franchise (CRF). On the contrary, the blocked Gibbs sampler uses the indicator variables to directly associate observations to model parameters, hence avoiding some cumbersome book-keeping of the CRF [195, 254]. Although sampling the indicator variables instead of table and dish assignments in CRF has shown to be sufficient, we note that this idea still adds a set of auxiliary variables [195, 254]. We discuss the CRF in the admixture modelling context in the next section and later demonstrate how the indicator variables relate with observation parameters.

5.1.6.1 The Chinese Restaurant Franchise in the admixture modelling context

Given K populations with common origins, population individuals share founder haplotypes (**Figure 5.1**). Therefore, for the haplotype inheritance problem, all the ancestral populations form a "franchise". Each population k is a "restaurant", each admixed population individual is the "customer" and the founder haplotype is the "served dish". Let the population, founder

haplotype and individual index be k, ℓ and ι, respectively. Our state value is the founder haplotype which as mentioned before is shared across ancestral populations. However, the distribution of founder haplotypes differ among ancestral populations. It is highly probable to copy from the same founder haplotype in the same population at two consecutive loci, often called state persistence. This means, if we start copying from a certain founder at SNP t, chances are high that the same founder maybe copied at SNP $t+1$. This is similar to the case when some customers become loyal to a specific restaurant due to a "specialty" dish which happens to be a restaurant's namesake. Ideally, this places some weight on same state transitions, such that, in population k a haplotype founder is drawn according to $\frac{(\alpha_2\beta+\kappa\delta_k)}{(\alpha_2+\kappa)}$, showing that every population reconsiders the shared base measure β to account for state persistence [195, 254]. Here, the index of δ is k because the founder haplotype is population k's namesake, that is, $\ell = k$.

It has been noted that variation can be between- or within-population. Given a population, individuals can be further assigned to groups and we shall call these groups "clusters". Just as the first occupant of a table considers the dish for their table in a CRF framework, initially, the founder haplotype (corresponding to the "considered dish") is drawn according to β, we will call this "considered founder". In population k, the founder choice for cluster c can be overridden by a Bernoulli distributed variable w_{kc} with parameter ρ. Thus, w_{kc} is the override variable [195, 254]. Now, the founder haplotype after override (corresponding to the "served dish") is ℓ_{kc}. We call this "final founder". We give the relationship between these three variables in Relation (5.1.6).

$$\bar{\ell}_{kc}|\beta \sim \beta, \qquad w_{kc}|\alpha_2,\kappa \sim \text{Ber}(\rho), \qquad \ell_{kc}|\bar{\ell}_{kc},w_{kc} = \begin{cases} \bar{\ell}_{kc}, & \text{if } w_{kc}=0 \\ k, & \text{if } w_{kc}=1 \end{cases}, \quad (5.1.6)$$

where $\rho = \frac{\kappa}{\alpha_2+\kappa}$ is the prior probability that the founder haplotype was overridden ($w_{kc} = 1$) [195, 254]. Without considering the ancestral population, the first individual to belong to a cluster selects a founder as in original CRF denoted by $\bar{\ell}_{kc}$. We refer to it as the considered founder. Due to some update information on the population, the considered haplotype may be different from final haplotype. In the original HDP-HMM, $w_{kc} = 0$ for all clusters and the considered haplotype is always the final. Similarly to the original HDP-HMM in Relation (2.3.16), we marginalise over stick breaking measures to get the predictive distributions over clusters

where a founder haplotype is shared and the considered founder haplotypes are as follows

$$\begin{aligned} p(c_{k\iota}|c_{k1},c_{k2},\ldots,c_{k\iota-1},\alpha_2) &\propto \sum_{t=1}^{C_k} n_{k\ell} I(c_{k\iota},c) + (\alpha_2+\kappa) I(c_{k\iota},\tilde{c}), \\ \bar{\ell}_{kc}|\underline{\bar{\ell}}_1,\underline{\bar{\ell}}_2,\ldots,\underline{\bar{\ell}}_{k-1},\bar{\ell}_{k1},\ldots,\bar{\ell}_{k\iota-1},\alpha_1 &\propto \alpha_1 I(\bar{\ell}_{kc},\tilde{\ell}) + \sum_{\ell=1}^{\bar{L}} \bar{m}_{\cdot\ell} I(\bar{\ell}_{kc},\ell). \end{aligned} \tag{5.1.7}$$

where C_k is the number of instantiated clusters, $\tilde{c}$ is the unseen (unrepresented) cluster, $\bar{\ell}_{kc}$ is the considered population k haplotype, $\underline{\bar{\ell}}_k = \{\ell_{k1},\ell_{k2},\ldots,\ell_{kC_k}\}$, $n_{k\ell}$ is the number of individuals in population k whose founder is ℓ in cluster c, $\bar{L}$ is the number of considered haplotypes and $\bar{m}_{\cdot\ell} = \sum_k \bar{m}_{k\ell}$ is the total number of clusters that considered founder ℓ in all ancestral populations such that $\bar{m}_{k\ell}$ is the number of clusters that considered founder ℓ in population k. Since we have relations that link w_{kc}, ℓ_{kc} and $\bar{\ell}_{kc}$, the next section provide details on sampling $\bar{m}_{k\ell}$ using these auxiliary variables.

5.1.6.2 Sampling $\bar{m}_{k\ell}$

In this section, let $\mathbf{m}$ represent $m_{k\ell}$ and $\mathbf{w}$ represent w_{kc}, and $\omega = \beta,\alpha_2,\kappa,\gamma,\varepsilon$. Following Fox et al. [195], the joint distribution of $\mathbf{m}$, $\mathbf{w}$ and $\bar{\mathbf{m}}$ is

$$p(\mathbf{m},\mathbf{w},\bar{\mathbf{m}}|z_{1:T},\omega) = p(\bar{\mathbf{m}}|\mathbf{m},\mathbf{w},z_{1:T},\omega)p(\mathbf{w}|\mathbf{m},z_{1:T},\omega)p(\mathbf{m}|z_{1:T},\omega). \tag{5.1.8}$$

Since $m_{k\ell}$ is the number of clusters that inherited founder ℓ in population k, given the founder assignment. Using the Polya urn representation of a DP, the number of clusters that inherited founder haplotype ℓ is distributed according to

$$p(m_{k\ell}=m|n_{kc},\omega) = \frac{\Gamma(\alpha_2\beta_\ell+\kappa\delta(\ell,k))}{\Gamma(\alpha_2\beta_\ell+\kappa\delta(\ell,k)+n_{k\ell})} s(n_{k\ell},m)(\alpha_2\beta_\ell+\kappa\delta(\ell,k))^m. \tag{5.1.9}$$

Given a large number of individuals in population k whose origins are founder haplotype ℓ in cluster c ($n_{k\ell}$), it is computationally expensive to compute the Stirling numbers. Therefore, it is recommended to sample $m_{k\ell}$ by simulating cluster assignments as in Equation (5.1.10). Thus, sample $m_{k\ell}$ by sampling cluster assignments for each population individual ι, that is,

$$\begin{aligned} p(c_{k\iota}=c|\ell_{kc}=\ell,\mathbf{c}^{-k\iota},\ell^{-kc},y_{1:T},\omega) &\propto p(c_{k\iota}|c_{k1},\ldots,c_{k(\iota-1)},c_{k(\iota+1)},\ldots,c_{kC_k},\alpha_2,\kappa)p(\ell_{kc}=\ell|\omega) \\ &\propto \begin{cases} \tilde{n}_{kc}^{-k\iota} & c\in\{1,\ldots,C_k\} \\ \alpha_2\beta_\ell+\kappa\delta(\ell,k) & c=\tilde{c} \end{cases}, \end{aligned} \tag{5.1.10}$$

where $\tilde{n}_{kc}^{-k\iota}$ is the count of cluster c individuals in population k excluding individual ι, $\mathbf{c}^{-k\iota}$ are cluster assignments for all individuals up to and including the $(\iota-1)^{\text{th}}$ and ℓ^{-kc} are founder haplotype assignments for all clusters in population k excluding that for cluster c. As highlighted before, C_k represents the number of currently instantiated clusters in population k. Equation (5.1.10) shows that given founder haplotype assignments, in a population, individuals are assigned to clusters according to a DP with a concentration parameter $\alpha_2\beta_\ell + \kappa\delta(\ell,k)$. That is,

$$c_{k\iota} = c|\ell_{kc_{k\iota}} = \ell, \mathbf{t}^{-k\iota}, \ell^{-kc}, y_{1:T}, \boldsymbol{\omega} \;\sim\; s^*_\ell, \qquad s^*_\ell|\alpha_2,\kappa,\beta_\ell \;\sim\; \text{GEM}(\alpha_2\beta_\ell + \kappa\delta(\ell,k))$$

As seen in Relation (5.1.6), in population k, the founder choice for a cluster can be overridden by $w_{kc} \sim \text{Ber}(\rho)$ resulting in a specialty founder whose index is k. Now, if a cluster inherits a founder ℓ which is not specialty, the number of clusters inheriting this founder is $m_{k\ell}$ and therefore $w_{kc} = 0$; otherwise we assume the considered founder index $\bar{\ell}_{kc}$ is known for all clusters inheriting from the specialty founder, hence simplifying the inference of w_{kc}.

$$\begin{aligned}
p(w_{kc}|\ell_{kc}=k,\beta,\rho) &= \sum_{\bar{\ell}_{kc=1}}^{\bar{L}} p(\bar{\ell}_{kc}, w_{kc}|\ell_{kc}=k,\beta) + p(\bar{\ell}_{kc}=\tilde{\ell}, w_{kc}|\ell_{kc}=k,\beta) \\
&\propto p(w_{kc}|\rho) \sum_{\bar{\ell}_{kc=1}}^{\bar{L}} p(\bar{\ell}_{kc}=\tilde{\ell}|\beta)\left[p(\ell_{kc}=k|\bar{\ell}_{kc},w_{kc}) + p(\ell_{kc}=k|\bar{\ell}_{kc}=\tilde{\ell},w_{kc})\right] \\
&\propto \begin{cases} \beta_k(1-\rho), & w_{kc}=0; \\ \rho, & w_{kc}=1, \end{cases}
\end{aligned} \tag{5.1.11}$$

where, $\rho = \frac{\kappa}{\alpha_2+\kappa}$ and $1-\rho$ is the prior probability that $w_{kc}=1$ and $w_{kc}=0$ so that the considered founder $(\bar{\ell}_{kc}=k)$ occurs with probability β_k. Equation (5.1.11) shows that after knowing that the final founder is specialty, chances are high that the considered founder $(\bar{\ell}_{kc})$ that was suggested by the prior was disregarded. Although $w_{kc}=0$ with probability $1-\rho$, it is possible to still inherit a specialty founder (i.e, the considered founder is specialty, $\bar{\ell}_{kc}=k$) [195, 254]. We assume that the total number of founder overrides in population k, $w_{k\bullet} = \sum_c w_{kc}$ is distributed according to Binomial(n,ρ). Given the number of clusters in each population that inherited a particular founder $(m_{k\ell})$ and the override variable for each instantiated cluster (w_{kc}), we compute the number of clusters that considered founder ℓ in

population k ($\bar{m}_{k\ell}$) as follows,

$$\bar{m}_{k\ell} = \begin{cases} m_{k\ell}, & \ell \neq k \text{ (if final founder is not specialty);} \\ m_{kk} - w_{k\bullet}, & \ell = k \text{ (if final haplotype is specialty),} \end{cases} \tag{5.1.12}$$

where, m_{kk} is the number of clusters that inherited the specialty founder in population k and $w_{k\bullet}$ is the total number of overrides in population k. Using Equation (5.1.12), the count of distinct considered founders is equal to that of distinct final founders (that is, $\bar{L} = L$) [195, 254].

5.1.6.3 Sampling β, ε, s and $\mathfrak{q}$

In order to sample β, ε, s and $\mathfrak{q}$, we use the derivation of a DP as an infinite limit of a finite mixture model. Assuming L and K are counts of components of finite mixture models, the prior distributions of the finite mixtures are

$$\begin{aligned} \beta|\alpha_1 &\sim \text{Dir}\left(\frac{\alpha_1}{L}, \ldots, \frac{\alpha_1}{L}\right) \\ \mathbf{s}_\ell|\alpha_2, \kappa, \beta &\sim \text{Dir}\left(\alpha_2\beta_1, \ldots, \alpha_2\beta_\ell + \kappa, \ldots, \alpha_2\beta_L\right) \\ \varepsilon|\sigma &\sim \text{Dir}\left(\frac{\sigma}{K}, \ldots, \frac{\sigma}{K}\right) \\ \mathfrak{q}_k|\gamma, \varepsilon &\sim \text{Dir}\left(\gamma\varepsilon_1, \ldots, \gamma\varepsilon_K\right), \end{aligned} \tag{5.1.13}$$

and the posterior distributions are then given by

$$\begin{aligned} \beta|\bar{m}, \alpha_1 &\sim \text{Dir}\left(\frac{\alpha_1}{L} + \bar{m}_{\bullet 1}, \ldots, \frac{\alpha_1}{L} + \bar{m}_{\bullet L}\right) \\ \varepsilon|\bar{m'}, \sigma &\sim \text{Dir}\left(\frac{\sigma}{K} + \bar{m'}_{\bullet 1}, \ldots, \frac{\sigma}{K} + \bar{m'}_{\bullet K}\right) \\ \mathbf{s}_\ell|z_{1:T}, \alpha_2, \kappa, \beta &\sim \text{Dir}(\alpha_2\beta_1 + n_{\ell 1}, \ldots, \alpha_2\beta_\ell + \kappa + n_{\ell\ell}, \ldots, \alpha_2\beta_L + n_{\ell L}) \\ \mathfrak{q}_k|x_{1:T}, z_t, \gamma, \varepsilon &\sim \text{Dir}(\gamma\varepsilon_1 + n^*_{k1}, \ldots, \gamma\varepsilon_K + n^*_{kK}), \end{aligned} \tag{5.1.14}$$

where, $n_{\ell\ell'}$ and $n^*_{kk'}$ is the number of transitions that have occurred when copying from founder haplotype ℓ to founder haplotype ℓ' in the state sequence $z_{1:T}$ and number of times the founder ℓ of population k was followed by founder ℓ' of population k' in the state sequence $x_{1:T}$, $\bar{m}_{k\ell}$ and $\bar{m'}_{k\ell}$ corresponds to the number of clusters in population k that considered founder ℓ and founder ℓ' clusters that were inherited by population k.

5.1.6.4 Sampling $(z_{1:T}, x_{1:T})$

This section highlights how we sample (z_t, x_t) jointly. We derive the conditional distribution of (z_t, x_t) using forward procedure following the ideas of Fox et al. [195].

$$p(z_t, x_t | z_{t-1}, x_{t-1}, y_{1:T}, \mathbf{s}, \mathbf{q}, \theta) \propto p(z_t | s_{z_{t-1}}) p(x_t | \mathbf{q}_{z_t x_{t-1}}) g(y_t | \theta_{z_t x_t}) m_{t+1,t}(z_t, x_t). \quad (5.1.15)$$

Since both z_t and x_t have a Markovian structure, then the backward message from (z_t, x_t) to (z_{t-1}, x_{t-1}) denoted by $m_{t,t-1}(z_{t-1}, x_{t-1})$ depends on z_{t-1} and x_{t-1}. That is,

$$m_{t,t-1}(z_{t-1}, x_{t-1}) \propto \begin{cases} \sum_{z_t} \sum_{x_t} p(z_t | s_{z_{t-1}}) p(x_t | \mathbf{q}_{z_t x_{t-1}}) g(y_t | \theta_{z_t x_t}) m_{t+1,t}(z_t, x_t), & t \leq T \\ 1, & t = T+1. \end{cases} \quad (5.1.16)$$

Using **Figure 5.3**, Equation (5.1.15) cn be rewritten as follows:

$$p(z_t = \ell, x_t = k | z_{t-1} = \ell', x_{t-1} = k', y_{1:T}, \mathbf{s}, \mathbf{q}, \theta) \propto s_{z_{t-1}}(\ell) \mathbf{q}_{\ell k}(k, \ell) \text{Ber}(y_t; \theta_{\ell k}) m_{t+1,t}(k, \ell) \quad (5.1.17)$$

where,

$$\begin{aligned} m_{t+1,t}(k, \ell) &= \sum_{u=1}^{L} \sum_{v=1}^{K} s_v(u) \mathbf{q}_{uv}(k, \ell) \text{Ber}(y_{t+1}; \theta_{uv}) m_{t+2,t+1}(u, v) \\ m_{T+1,T}(k, \ell) &= 1, \quad k = 1, \ldots, K, \quad \ell = 1, \ldots, L. \end{aligned}$$

As previously highlighted, we place Gamma priors on $\alpha_1, \alpha_2 + \kappa, \sigma$ and γ, and Beta prior on ρ [191, 193, 195, 196, 254].

5.1.7 Summary of differences between the proposed mSPECTRUM extension and other local ancestry models

We summarise the differences between the modified mSPECTRUM and existing local ancestry inference models in **Table 5.2**. We believe the proposed mSPECTRUM modification will improve the accuracy of estimates since they address some modelling challenges that exist in local ancestry inference.

Table 5.2: Comparing mSPECTRUM extension to local ancestry models (the BNPs and non-BNP models).

Model assumptions	mSPECTRUM extension	BNP models	non-BNP models
Models post-admixture selection in multi-way admixtures	✓	✗	✗
Ancestral populations have common origins	✓	✓for mSPECTRUM, ✗ for HDPStructure	✗
Number of hypothetical founders and/or populations	Both unknown	unknown founders in mSPECTRUM	number of populations is known
Independent contributing population	✗	✓for HDPStructure, ✗ for mSPECTRUM	independent
Fully model state persistence	✓	✗	✗
Model for state persistence	Fox et al. [195]	Beam sampling	✗
State dependent model	✓	✓for HDPStructure, ✗ for mSPECTRUM	✗
Haplotype-based model	✓	✓	Some are and others are not.

5.2 Overall thesis conclusion

Admixed individuals such as African-Americans, South African Coloureds and complex admixed individuals from south-America and Africa have been among the most marginalised stratum of the society globally. However, migration and integration may continue, and rates of admixture are expected to increase in the near future. Numerous studies have applied admixture association analyses to African-Americans and Hispanic cohorts, demonstrating added value beyond standard association testing. Admixture association is critically reliant on accurate locus ancestry inference (LAI), which requires well-specified founding population reference samples. Power can be optimised by combining admixture mapping and association testing, but this approach is rarely adopted because of the multi-stage process required and the challenge in application to complex multi-way admixed samples. To better understand admixed individuals, we have investigated local ancestry models for multi-way admixtures. In this thesis,

1. We observed that it is important to choose an appropriate multi-way local ancestry model for each admixed population dataset because model accuracy differs with datasets.

2. We noted that LOTER [56] brought a major improvement in local ancestry inference as it only requires ancestral and admixed haplotype information to deconvolve ancestry [4]. Based on the quantitative review (**Chapter 4**), we confirm that LOTER performs well in ancient admixtures and is comparable to the best model in five-way admixtures.

3. Unlike LAMP-LD (an LD-based model which outperformed six other state-of-the-art models in five-way simulated datasets), LOTER is less affected by the skewness in the reference ancestral population sizes (**Chapter 4**).

4. Different from other models, LOTER has been designed for non-model species [56] and requires longer running time. Therefore, if run time is a concern for users, then, LOTER may not be the most appropriate model choice.

5. WINPOP is robust to admixture generations, but, requires information on the genome-wide average ancestry proportions which might not always be accurate if known. In such a case, SUPPORTMIX might be a better choice since it estimates the global ancestry proportions. Also, if ancestral population information is less known and a pool of reference ancestral panels (which may perhaps contain the proxies of the ancestral information) is available, it is best to deconvolve ancestry with SUPPORTMIX [4]. If the information about the admixture generations is unknown, then PCADMIX could be a better option as it doesn't require admixture generations to infer ancestry. However, we noted that PCADMIX is not ideal when the SNPs under study are less than double the number of individuals for a given chromosome.

6. Consistent with existing studies where RFMIX performed better than LAMP-LD 75% of the time in three-way admixtures (**Table B.1**) [56, 128, 129], RFMIX is comparable to LAMP-LD. This could therefore explain why RFMIX is more popular than LAMP-LD. This could be due to the fact that most existing studies are benchmarked for three-way admixed populations and also compared to LAMP-LD, RFMIX has a shorter running time (see, **Figure 5.5**). Finally, the output format of RFMIX may require no/little programming background to be manipulated for further applications compared to the LAMP-LD formats.

7. Although its performance tends to decrease as the number of source populations increases, the performance of ELAI in three-way admixed populations is comparable to the best models [97, 102]. More so, it is useful in estimating recombination [4].

8. We facilitated the local ancestry inference process and application studies by designing a Python framework (FRANC), which eases the evaluation of different models on a given admixture dataset and enables user to easily select their tool of choice or run multiple tools in a single framework.

In the next section, we provide some recommendations based on these findings, as well as highlight some future trends in local ancestry inference.

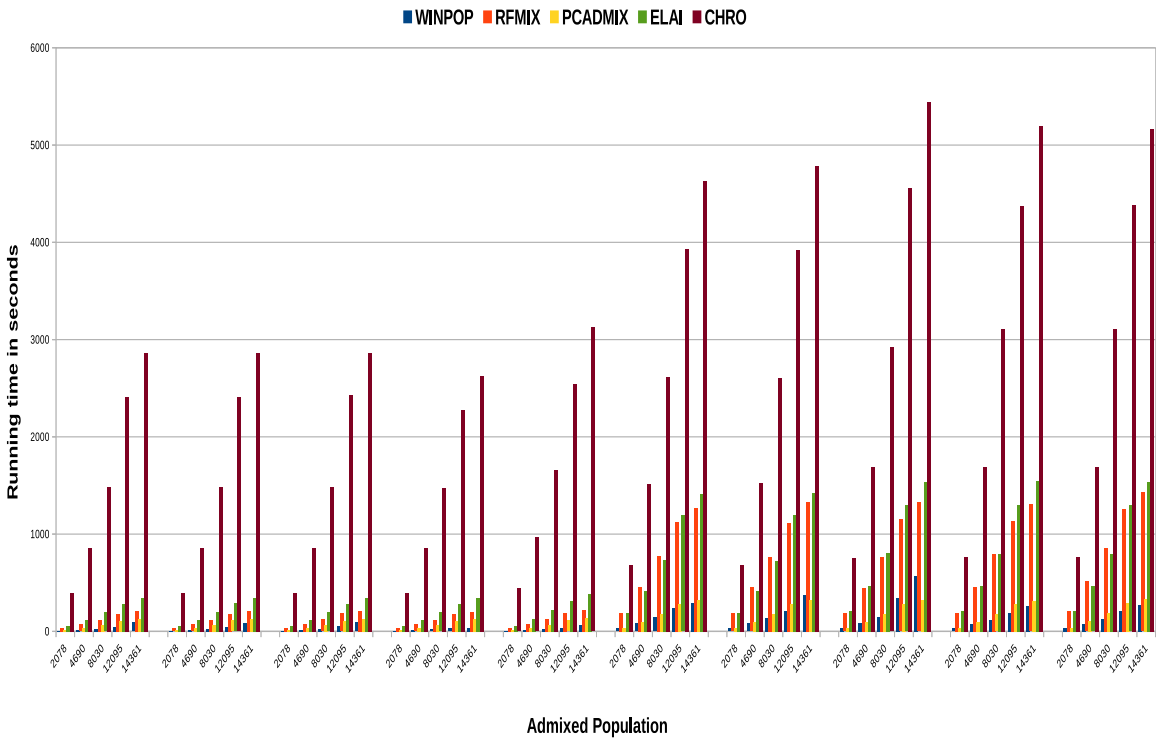


Figure 5.5: A time graph representation of the first five fastest models, given SIM3G1 (diseased unaffected three-way admixed individuals formed 12 generations ago), SIM3G2 (diseased unaffected three-way formed 100 generations ago), SIM3G3 (diseased unaffected three-way formed 600 generations ago), SIM3MG1 (diseased affected and unaffected three-way), SIM3MG2, SIM5G1 (diseased unaffected five-way formed 12 generations ago), SIM5G2, SIM5G3, SIM5MG1 and SIM5MG2, respectively each for 2 078, 4 690, 12 095 and 14 361 SNPs in that order.

5.3 Recommendations

1. We recommend that future local ancestry inference models account for post-admixture selection and population relationships based on natural human evolution and in this thesis we presented a theoretical framework extending the mSPECTRUM model.

2. For existing implemented local ancestry inference, we recommend WINPOP and/or SUPPORTMIX if the users are taking into account the running time of the model, in case of the 4^{th} point of **Section 5.2**. Although the accuracy of estimates may be compromised, in five-way admixtures, we observed that WINPOP estimates were comparable to LAMP-LD and LOTER (the best models in five-way admixtures).

3. In order to make informed decisions, we recommend that users assess the performance of the different state-of-the-art models on a given admixed population dataset before

any estimate application.

4. Given an admixture formed by the interbreeding of three or more ancestral populations, if the reference ancestral panels are

   (a) Equal in size, we recommend LAMP-LD and

   (b) Skewed, we recommend LOTER.

5. Where necessary, instead of relying on default parameters, we recommend users to provide biologically relevant population parameters as these may differ among admixed populations.

6. Model evaluation studies should assess both LD- and non-LD-based models within the same framework (at least two from each category). FRANC developed in **Chapter 3** provides a tool to enable this.

7. Since the local ancestry problem is not completely solved in multi-way admixtures, we recommend developers/modellers to improve/modify existing models rather than developing new models unless absolutely necessary. For example, we suggest modifying LAMP-LD to improve the accuracy in skewed reference ancestral panels and adjusting LOTER to improve on the running time. Further, it may be practical to model more realistic biological assumptions, including accounting for post-admixture selection and the origins of modern humans as proposed in **Section 5.1**.

5.4 Future work

In future, we aim to incorporate other existing tools within the FRANC framework and extend it to take advantage of high performance computing and workflow framework (nextflow). Additionally, modelling more realistic biological assumptions including natural selection, founder bottlenecks and population relationships may provide more insights into local ancestry estimates accuracy. Thus, it may be interesting to examine the modification of mSPECTRUM including the incorporation of state persistence even when founders switch between consecutive SNPs. Finally, there is a need to investigate Bayesian nonparametric models given different inference schemes, including the particle Gibbs sampling.

References

[1] John Murray, Heinz Peter Nasheuer, Cathal Seoighe, Grace P McCormack, D Michael Williams, and David AT Harper. The contribution of william king to the early development of palaeoanthropology. *Irish Journal of Earth Sciences*, 33:1–16, 2015.

[2] Gaston K. Mazandu, Ephifania Geza, Milaine Seuneu, and Emile R. Chimusa. Orienting Future Trends in Local Ancestry Deconvolution Models to Optimally Decipher Admixed Individual Genome Variations.

[3] World Health Organization. WHO Global Health Estimates. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates, 2020. [Online; accessed 18-November-2021].

[4] Ephifania Geza, Jacquiline Mugo, Nicola J Mulder, Ambroise Wonkam, Emile R Chimusa, and Gaston K Mazandu. A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. *Briefings in bioinformatics*, 20(5):1709–1724, 2018. doi: 10.1093/bib/bby044.

[5] Ephifania Geza, Nicola J Mulder, Emile R Chimusa, and Gaston K Mazandu. FRANC: A unified framework for multi-way local ancestry deconvolution with high density SNP data. *Briefings in bioinformatics*, 21(5):1837–1845.

[6] Joseph K Pickrell, Nick Patterson, Chiara Barbieri, Falko Berthold, Linda Gerlach, Tom Güldemann, Blesswell Kure, Sununguko Wata Mpoloka, Hirosi Nakagawa, Christfried Naumann, et al. The Genetic Prehistory of Southern Africa. *Nature communications*, 3:1143, 2012.

[7] Ying Zhou, Hongxiang Qiu, and Shuhua Xu. Modeling continuous admixture using admixture-induced linkage disequilibrium. *Scientific reports*, 7:43054, 2017.

[8] Bruce S Weir and C Clark Cockerham. Estimating f-statistics for the analysis of population structure. *evolution*, pages 1358–1370, 1984.

[9] Wenfei Jin. *Admixture dynamics, natural selection and diseases in admixed populations*. Springer, 2015.

[10] Kwang Hyun Ko. Hominin interbreeding and the evolution of human variation. *Journal of Biological Research-Thessaloniki*, 23(1):17, 2016.

[11] Michael F Hammer, August E Woerner, Fernando L Mendez, Joseph C Watkins, and Jeffrey D Wall. Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences*, 108(37):15123–15128, 2011.

[12] Joel T Dudley and Konrad J Karczewski. *Exploring personal genomics*. Oxford University Press, 2013.

[13] John H Relethford and Rosalind M Harding. Population genetics of modern human evolution. *eLS*, 2001.

[14] Marta Melé, Asif Javed, Marc Pybus, Pierre Zalloua, Marc Haber, David Comas, Mihai G Netea, Oleg Balanovsky, Elena Balanovska, Li Jin, et al. Recombination gives a new insight in the effective population size and the history of the old world human populations. *Molecular biology and evolution*, 29(1):25–30, 2011.

[15] Michael Dannemann and Fernando Racimo. Something old, something borrowed: admixture and adaptation in human evolution. *Current opinion in genetics & development*, 53:1–8, 2018.

[16] Olga Dolgova and Oscar Lao. Evolutionary and medical consequences of archaic introgression into modern human genomes. *Genes*, 9(7):358, 2018.

[17] Fernando A Villanea and Joshua G Schraiber. Multiple episodes of interbreeding between neanderthal and modern humans. *Nature ecology & evolution*, 3(1):39, 2019.

[18] Emile R Chimusa, Joel Defo, Prisca K Thami, Denis Awany, Delesa D Mulisa, Imane Allali, Hassan Ghazal, Ahmed Moussa, and Gaston K Mazandu. Dating admixture events is unsolved problem in multi-way admixed populations. *Briefings in bioinformatics*, 21 (2):751, 2018.

[19] Virginie Orgogozo, Baptiste Morizot, and Arnaud Martin. The differential view of genotype–phenotype relationships. *Frontiers in genetics*, 6:179, 2015.

[20] Kay Young McChesney. Teaching diversity: The science you need to know to explain why race is not biological. *SAGE Open*, 5(4):1–13, 2015.

[21] Alejandro Burga and Ben Lehner. Beyond genotype to phenotype: why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience. *The FEBS journal*, 279(20):3765–3775, 2012.

[22] Aravinda Chakravarti. Perspectives on human variation through the lens of diversity and race. *Cold Spring Harbor perspectives in biology*, 7(9):a023358, 2015.

[23] Ning Yu, Zhi wei Zhao, Y X Fu, Nyamkhishig Sambuughin, Michelle Ramsay, T Jenkins, Elina Leskinen, László Patthy, Lynn B. Jorde, Takashi Kuromori, and Wei Li. Global patterns of human dna sequence variation in a 10-kb region on chromosome 1. *Molecular Biology and Evolution*, 18(2):214–222, 2001.

[24] Amelie Baud, Megan K Mulligan, Francesco Paolo Casale, Jesse F Ingels, Casey J Bohl, Jacques Callebert, Jean-Marie Launay, Jon Krohn, Andres Legarra, Robert W Williams, et al. Genetic variation in the social environment contributes to health and disease. *PLoS genetics*, 13(1):e1006498, 2017.

[25] Patrick J Stover. Human nutrition and genetic variation. *Food and nutrition bulletin*, 28(1_suppl1):S101–S115, 2007.

[26] Jing Li, Luyong Zhang, Hang Zhou, Mark Stoneking, and Kun Tang. Global patterns of genetic diversity and signals of natural selection for human adme genes. *Human molecular genetics*, 20(3):528–540, 2010.

[27] Wayne A Schröder, Andreas K Klein, Stefan Winter, Mathias Schwab, Michael Bonin, Andreas Zell, and Ulrich M Zanger. Genomics of adme gene expression: Mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver. *The pharmacogenomics journal*, 13(1):12–20, 2013.

[28] Ulrich M Zanger. Pharmacogenetics–Challenges and Opportunities Ahead. *Frontiers in pharmacology*, 1:112, 2010.

[29] Jeffrey C Long. Human genetic variation: The mechanisms and results of microevolution. *Ann Arbor*, 1001:48109, 2004.

[30] Yusuke Nakamura. DNA variations in human and medical genetics: 25 years of my experience. *Journal of human genetics*, 54(1):1, 2009.

[31] David E Reich, Stephen F Schaffner, Mark J Daly, Gil McVean, James C Mullikin, John M Higgins, Daniel J Richter, Eric S Lander, and David Altshuler. Human genome

sequence variation and the influence of gene history, mutation and recombination. *Nature genetics*, 32(1):135, 2002.

[32] Anthony JF Griffiths, Susan R Wessler, Carroll Sean B, and Doebley John. *An Introduction to genetic analysis*. W.H. Freeman, New York, 11th ed edition, 2015.

[33] Kyung-Ah Sohn, Eric P Xing, et al. A hierarchical dirichlet process mixture model for haplotype reconstruction from multi-population data. *The Annals of Applied Statistics*, 3(2):791–821, 2009.

[34] Francis S Collins and Monique K Mansoura. The human genome project. revealing the shared inheritance of all humankind. *Cancer*, 91(1 Suppl):221–225, 2001.

[35] Lynn B Jorde and Stephen P Wooding. Genetic variation, classification and 'race'. *Nature genetics*, 36:S28–S33, 2004.

[36] Samuel Levy, Granger Sutton, Pauline C Ng, Lars Feuk, Aaron L Halpern, Brian P Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F Kirkness, Gennady Denisov, et al. The diploid genome sequence of an individual human. *PLoS Biol*, 5(10):e254, 2007.

[37] Noah A Rosenberg. A population-genetic perspective on the similarities and differences among worldwide human populations. *Human biology*, 83(6):659, 2011.

[38] Mehrdad Bakhtiari, Sharona Shleizer-Burko, Melissa Gymrek, Vikas Bansal, and Vineet Bafna. Targeted genotyping of variable number tandem repeats with adVNTR. *Genome research*, 28(11):1709–1719, 2018.

[39] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[40] Wong Limsoon. *The practical bioinformatician*. World scientific, 2004.

[41] Francis S Collins, Lisa D Brooks, and Aravinda Chakravarti. A DNA polymorphism discovery resource for research on human genetic variation. *Genome research*, 8(12): 1229–1231, 1998.

[42] Michael Lynch. Mutation and human exceptionalism: our future genetic load. *Genetics*, 202(3):869–875, 2016.

[43] Ziyue Gao, Minyoung J Wyman, Guy Sella, and Molly Przeworski. Interpreting the dependence of mutation rates on age and time. *PLoS biology*, 14(1):e1002355, 2016.

[44] Graham Coop and Molly Przeworski. An evolutionary view of human recombination. *Nature Reviews Genetics*, 8(1):23, 2007.

[45] Na Li and Matthew Stephens. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots using Single-Nucleotide Polymorphism data. *Genetics*, 165(4): 2213–2233, 2003.

[46] Nicholas Hamilton Barton. Genetic linkage and natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552):2559–2569, 2010.

[47] Robert G Latta. Natural selection, variation, adaptation, and evolution: a primer of interrelated concepts. *International journal of plant sciences*, 171(9):930–944, 2010.

[48] Leopoldo Sánchez and John A Woolliams. Impact of nonrandom mating on genetic variance and gene flow in populations with mass selection. *Genetics*, 166(1):527–535, 2004.

[49] Philip W Hedrick, Elaina M Tuttle, and Rusty A Gonser. Negative-assortative mating in the white-throated sparrow. *Journal of Heredity*, 109(3):223–231, 2017.

[50] PWE Kearns, IPM Tomlinson, P O'Donald, and CJ Veltman. Non-random mating in the two-spot ladybird (adalia bipunctata): I. a reassessment of the evidence. *Heredity*, 65(2):229–240, 1990.

[51] Gilean AT McVean. A genealogical interpretation of linkage disequilibrium. *Genetics*, 162(2):987–991, 2002.

[52] Joanna Masel. Genetic drift. *Current Biology*, 21(20):R837–R838, 2011.

[53] Sarah M Brown, Katherine A Harrisson, Rohan H Clarke, Andrew F Bennett, and Paul Sunnucks. Limited population structure, genetic drift and bottlenecks characterise an endangered bird species in a dynamic, fire-prone ecosystem. *PloS One*, 8(4), 2013.

[54] Bennett M Berger. How long is a generation? *The British Journal of Sociology*, 11(1): 10–23, 1960.

[55] David E Reich and Eric S Lander. On the allelic spectrum of human disease. *TRENDS in Genetics*, 17(9):502–510, 2001.

[56] Thomas Dias-Alves, Julien Mairal, and Michael GB Blum. LOTER: A software package to infer local ancestry for a wide range of species. *Molecular Biology and Evolution*, 35 (9):2318–2326, 2018.

[57] Jun J Yang, Cheng Cheng, Meenakshi Devidas, Xueyuan Cao, Yiping Fan, Dario Campana, Wenjian Yang, Geoff Neale, Nancy J Cox, Paul Scheet, et al. Ancestry and Pharmacogenomics of Relapse in Acute Lymphoblastic Leukemia. *Nature Genetics*, 43 (3):237–241, 2011.

[58] Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

[59] Cindy George, Tandi E Matsha, Rajiv T Erasmus, and Andre P Kengne. Haematological profile of chronic kidney disease in a mixed-ancestry south african population: a cross-sectional study. *BMJ open*, 8(11):e025694, 2018.

[60] Katarzyna Bryc, Adam Auton, Matthew R Nelson, Jorge R Oksenberg, Stephen L Hauser, Scott Williams, Alain Froment, Jean-Marie Bodo, Charles Wambebe, Sarah A Tishkoff, et al. Genome-wide Patterns of Population Structure and Admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences*, 107 (2):786–791, 2010.

[61] Sarah A Tishkoff, Floyd A Reed, Françoise R Friedlaender, Christopher Ehret, Alessia Ranciaro, Alain Froment, Jibril B Hirbo, Agnes A Awomoyi, Jean-Marie Bodo, Ogobara Doumbo, et al. The Genetic Structure and History of Africans and African Americans. *Science*, 324(5930):1035–1044, 2009.

[62] Shuhua Xu, Wei Huang, Ji Qian, and Li Jin. Analysis of Genomic Admixture in Uyghur and its Implication in Mapping Strategy. *The American Journal of Human Genetics*, 82 (4):883–894, 2008.

[63] Tesfaye B Mersha. Mapping asthma-associated variants in admixed populations. *Frontiers in genetics*, 6:292, 2015.

[64] Emile R Chimusa, Michelle Daya, Marlo Möller, Raj Ramesar, Brenna M Henn, Paul D Van Helden, Nicola J Mulder, and Eileen G Hoal. Determining Ancestry Proportions in Complex Admixture Scenarios in South Africa Using a Novel Proxy Ancestry Selection Method. *PloS One*, 8(9):e73971, 2013.

[65] Erika de Wit, Wayne Delport, Chimusa E Rugamika, Ayton Meintjes, Marlo Möller, Paul D van Helden, Cathal Seoighe, and Eileen G Hoal. Genome-wide Analysis of the Structure of the South African Coloured Population in the Western Cape. *Human genetics*, 128(2):145–153, 2010.

[66] Emile R Chimusa, Noah Zaitlen, Michelle Daya, Marlo Möller, Paul D van Helden, Nicola J Mulder, Alkes L Price, and Eileen G Hoal. Genome-wide Association Study of Ancestry-specific TB Risk in the South African Coloured Population. *Human molecular genetics*, 23(3):796–809, 2014.

[67] Michelle Daya, Lize Van der Merwe, Christopher R Gignoux, Paul D Van Helden, Marlo Möller, and Eileen G Hoal. Using multi-way admixture mapping to elucidate tb susceptibility in the south african coloured population. *BMC genomics*, 15(1):1021, 2014.

[68] JPA Angseesing and D Sutton. Investigating the causes of changes in allele frequency during evolution. *Journal of Biological Education*, 9(6):251–255, 1975.

[69] Masatoshi Nei. F-statistics and analysis of gene diversity in subdivided populations. *Annals of human genetics*, 41(2):225–233, 1977.

[70] Bernard Wood. *Wiley-Blackwell encyclopedia of human evolution*. John Wiley & Sons, 2011.

[71] Liang Ma, Ya-Jie Ji, and De-Xing Zhang. Statistical measures of genetic differentiation of populations: Rationales, history and current states. *Current Zoology*, 61(5):886–897, 2015.

[72] Eva-Maria Willing, Christine Dreyer, and Cock Van Oosterhout. Estimates of genetic differentiation measured by fst do not necessarily require large sample sizes when using many snp markers. *PLoS One*, 7(8), 2012.

[73] Kent E Holsinger and Bruce S Weir. Genetics in geographically structured populations: defining, estimating and interpreting f st. *Nature Reviews Genetics*, 10(9):639–650, 2009.

[74] Sunhee Kim and Chang-Yong Lee. A consistent approach to the genotype encoding problem in a genome-wide association study of continuous phenotypes. *PloS one*, 15 (7):e0236139, 2020.

[75] Richard C Lewontin. The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, 49(1):49, 1964.

[76] Ji-Qian Fang. *Handbook of Medical Statistics*. World Scientific, 2018.

[77] Brigitte Mangin, Aurélie Siberchicot, S Nicolas, Agnes Doligez, Patrice This, and Christine Cierco-Ayrolles. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*, 108(3):285, 2012.

[78] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.

[79] Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

[80] Yael Baran, Bogdan Pasaniuc, Sriram Sankararaman, Dara G Torgerson, Christopher Gignoux, Celeste Eng, William Rodriguez-Cintron, Rocio Chapela, Jean G Ford, Pedro C Avila, et al. Fast and Accurate Inference of Local Ancestry in Latino Populations. *Bioinformatics*, 28(10):1359–1367, 2012.

[81] 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, 2013.

[82] Badri Padhukasahasram. Inferring ancestry from population genomic data and its applications. *Frontiers in genetics*, 5:204–204, 2014.

[83] Zachariah Gompert and C Alex Buerkle. Analyses of genetic ancestry enable key insights for molecular ecology. *Molecular Ecology*, 22(21):5278–5294, 2013.

[84] Jonathan K Pritchard and Peter Donnelly. Case–control studies of association in structured or admixed populations. *Theoretical population biology*, 60(3):227–237, 2001.

[85] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

[86] Kai Yuan, Ying Zhou, Xumin Ni, Yuchen Wang, Chang Liu, and Shuhua Xu. Models, methods and tools for ancestry inference and admixture analysis. *Quantitative Biology*, 5(3):236–250, 2017.

[87] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[88] Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology*, 28(4):289–301, 2005.

[89] Eric Frichot, François Mathieu, Théo Trouillon, Guillaume Bouchard, and Olivier François. Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196(4):973–983, 2014.

[90] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[91] Apichart Intarapanich, Philip J Shaw, Anunchai Assawamakin, Pongsakorn Wangkumhang, Chumpol Ngamphiw, Kridsadakorn Chaichoompu, Jittima Piriyapongsa, and Sissades Tongsima. Iterative pruning pca improves resolution of highly structured populations. *BMC bioinformatics*, 10(1):382, 2009.

[92] Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genet*, 5(6):e1000519, 2009.

[93] Nick Patterson, Neil Hattangadi, Barton Lane, Kirk E Lohmueller, David A Hafler, Jorge R Oksenberg, Stephen L Hauser, Michael W Smith, Stephen J O'Brien, David Altshuler, et al. Methods for High-density Admixture Mapping of Diseases Genes. *The American Journal of Human Genetics*, 74(5):979–1000, 2004.

[94] Clive J Hoggart, Mark D Shriver, Rick A Kittles, David G Clayton, and Paul M McKeigue. Design and Analysis of Admixture Mapping Studies. *The American Journal of Human Genetics*, 74(5):965–978, 2004.

[95] Andreas Sundquist, Eugene Fratkin, Chuong B Do, and Serafim Batzoglou. Effect of Genetic Divergence in Identifying Ancestral Origin Using HAPAA. *Genome research*, 18 (4):676–682, 2008.

[96] Sriram Sankararaman, Srinath Sridhar, Gad Kimmel, and Eran Halperin. Estimating Local Ancestry in Admixed Populations. *The American Journal of Human Genetics*, 82 (2):290–303, 2008.

[97] Yongtao Guan. Detecting Structure of Haplotypes and Local Ancestry. *Genetics*, 196 (3):625–642, 2014.

[98] Adam Auton, Katarzyna Bryc, Adam R Boyko, Kirk E Lohmueller, John Novembre, Andy Reynolds, Amit Indap, Mark H Wright, Jeremiah D Degenhardt, Ryan N Gutenkunst, et al. Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Research*, 19(5):795, 2009.

[99] Michael Dodson and Robert Williamson. Indigenous peoples and the morality of the human genome diversity project. *Journal of Medical Ethics*, 25(2):204–208, 1999.

[100] Ganguly Nirmal K, Rahmat Bano, and SD Seth. Human genome project: Pharmacogenomics and drug development. *Indian Journal of Experimental Biology*, 39:955–961, 2001.

[101] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299, 2005.

[102] Xiangqun Zheng-Bradley and Paul Flicek. Applications of the 1000 genomes project resources. *Briefings in functional genomics*, 16(3):163–170, 2016.

[103] Michael F Seldin, Bogdan Pasaniuc, and Alkes L Price. New Approaches to Disease Mapping in Admixed Populations. *Nature Reviews Genetics*, 12(8):523–528, 2011.

[104] Mengjie Chen, Can Yang, Cong Li, Lin Hou, Xiaowei Chen, and Hongyu Zhao. Admixture mapping analysis in the context of gwas with gaw18 data. In *BMC proceedings*, volume 8, page S3. BioMed Central, 2014.

[105] Zuben E Sauna and Chava Kimchi-Sarfaty. Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, 12(10):683, 2011.

[106] Ryan Hunt, Zuben E Sauna, Suresh V Ambudkar, Michael M Gottesman, and Chava Kimchi-Sarfaty. Silent (synonymous) snps: should we care about them? *Single nucleotide polymorphisms*, pages 23–39, 2009.

[107] Tuuli Lappalainen, Elina Salmela, Peter M Andersen, Karin Dahlman-Wright, Pertti Sistonen, Marja-Liisa Savontaus, Stefan Schreiber, Päivi Lahermo, and Juha Kere. Genomic landscape of positive natural selection in northern european populations. *European Journal of Human Genetics*, 18(4):471, 2010.

[108] Murray Cadzow, James Boocock, Hoang T Nguyen, Phillip Wilcox, Tony R Merriman, and Michael A Black. A bioinformatics workflow for detecting signatures of selection in genomic data. *Frontiers in genetics*, 5:293–293, 2014. doi: 10.3389/fgene.2014.00293.

[109] Kevin N Laland, John Odling-Smee, and Sean Myles. How culture shaped the human genome: bringing genetics and the human sciences together. *Nature Reviews Genetics*, 11:137, 2010.

[110] Joseph J Vitti, Sharon R Grossman, and Pardis C Sabeti. Detecting natural selection in genomic data. *Annual review of genetics*, 47:97–120, 2013.

[111] Rowan DH Barrett and Dolph Schluter. Adaptation from standing genetic variation. *Trends in ecology & evolution*, 23(1):38–44, 2008.

[112] Wenfei Jin, Shuhua Xu, Haifeng Wang, Yongguo Yu, Yiping Shen, Bailin Wu, and Li Jin. Genome-wide detection of natural selection in african americans pre-and post-admixture. *Genome research*, 22(3):519–527, 2012.

[113] Lian Deng, Andrés Ruiz-Linares, Shuhua Xu, and Sijia Wang. Ancestry variation and footprints of natural selection along the genome in latin american populations. *Scientific reports*, 6:21766–21766, 2016.

[114] Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–595, 1989.

[115] Justin C Fay and Chung-I Wu. Hitchhiking under positive darwinian selection. *Genetics*, 155(3):1405–1413, 2000.

[116] Rodrigo Secolin, Alex Mas-Sandoval, Lara R Arauna, Fábio R Torres, Tânia K de Araujo, Marilza L Santos, Cristiane S Rocha, Benilton S Carvalho, Fernando Cendes, Iscia Lopes-Cendes, et al. Distribution of local ancestry and evidence of adaptation in admixed populations. *Scientific reports*, 9(1):1–12, 2019.

[117] Heming Wang, Tamar Sofer, Xiang Zhang, Robert C Elston, Susan Redline, and Xiaofeng Zhu. Local ancestry inference in large pedigrees. *Scientific Reports*, 10(1):1–8, 2020.

[118] Hua Tang, Shweta Choudhry, Rui Mei, Martin Morgan, William Rodriguez-Cintron, Esteban González Burchard, and Neil J Risch. Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans. *The American Journal of Human Genetics*, 81(3):626–633, 2007.

[119] Bogdan Pasaniuc, Sriram Sankararaman, Dara G Torgerson, Christopher Gignoux, Noah Zaitlen, Celeste Eng, William Rodriguez-Cintron, Rocio Chapela, Jean G Ford, Pedro C Avila, et al. Analysis of latino populations from gala and mec studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics*, 29(11):1407–1415, 2013.

[120] Gaurav Bhatia, Arti Tandon, Nick Patterson, Melinda C Aldrich, Christine B Ambrosone, Christopher Amos, Elisa V Bandera, Sonja I Berndt, Leslie Bernstein, William J Blot, et al. Genome-wide scan of 29,141 african americans finds no evidence of directional selection since admixture. *The American Journal of Human Genetics*, 95(4):437–444, 2014.

[121] Zhaoming Wang, Kevin B Jacobs, Meredith Yeager, Amy Hutchinson, Joshua Sampson, Nilanjan Chatterjee, Demetrius Albanes, Sonja I Berndt, Charles C Chung, W Ryan

Diver, et al. Improved Imputation of Common and Uncommon SNPs with a New Reference Set. *Nature genetics*, 44(1):6–7, 2012.

[122] Giulio Genovese, Robert E Handsaker, Heng Li, Eimear E Kenny, and Steven A McCarroll. Mapping the human reference genome's missing sequence by three-way admixture in latino genomes. *The American Journal of Human Genetics*, 93(3):411–421, 2013.

[123] Helena Martins, Kevin Caye, Keurcien Luu, Michael GB Blum, and Olivier Francois. Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *Molecular ecology*, 25(20):5029–5042, 2016.

[124] Nick Patterson, Desiree C Petersen, Richard E van der Ross, Herawati Sudoyo, Richard H Glashoff, Sangkot Marzuki, David Reich, and Vanessa M Hayes. Genetic Structure of a Unique Admixed Population: Implications for Medical Research. *Human Molecular Genetics*, 19(3):411–419, 2010.

[125] Larsson Omberg, Jacqueline Salit, Neil Hackett, Jennifer Fuller, Rebecca Matthew, Lotfi Chouchane, Juan L Rodriguez-Flores, Carlos Bustamante, Ronald G Crystal, and Jason G Mezey. Inferring Genome-wide Patterns of Admixture in Qataris Using Fifty-five Ancestral Populations. *BMC genetics*, 13(1):49, 2012.

[126] Marco Galaverni, Romolo Caniglia, Luca Pagani, Elena Fabbri, Alessio Boattini, and Ettore Randi. Disentangling timing of admixture, patterns of introgression, and phenotypic indicators in a hybridizing wolf population. *Molecular biology and evolution*, 34 (9):2324–2339, 2017.

[127] Po-Ru Loh, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K Pickrell, David Reich, and Bonnie Berger. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, pages genetics–112, 2013.

[128] James Xue, Todd Lencz, Ariel Darvasi, Itsik Pe'er, and Shai Carmi. The time and place of european admixture in ashkenazi jewish history. *PLoS genetics*, 13(4):e1006644, 2017.

[129] Brian K Maples, Simon Gravel, Eimear E Kenny, and Carlos D Bustamante. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-ancestry Inference. *The American Journal of Human Genetics*, 93(2):278–288, 2013.

[130] Hugues Aschard, Alexander Gusev, Robert Brown, and Bogdan Pasaniuc. Leveraging local ancestry to detect gene-gene interactions in genome-wide data. *BMC genetics*, 16 (1):124, 2015.

[131] Yik Y Teo. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Current opinion in lipidology*, 19(2):133–143, 2008.

[132] Daniel Shriner, Adebowale Adeyemo, and Charles N Rotimi. Joint ancestry and association testing in admixed individuals. *PLoS computational biology*, 7(12):e1002325, 2011.

[133] Pere P Simarro, Jean Jannin, and Pierre Cattand. Eliminating human african trypanosomiasis: where do we stand and what comes next? *PLoS medicine*, 5(2):e55, 2008.

[134] Gengxin Li and Hongjiang Zhu. Genetic studies: The linear mixed models in genome-wide association studies. *Open Bioinformatics Journal*, 7(1):27–33, 2013.

[135] Xuexia Wang, Xiaofeng Zhu, Huaizhen Qin, Richard S Cooper, Warren J Ewens, Chun Li, and Mingyao Li. Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics*, 27(5):670–677, 2011.

[136] Daniel Shriner. Overview of admixture mapping. *Current Protocols in Human Genetics*, pages 1–23, 2013.

[137] Michael W Smith and Stephen J O'Brien. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nature Reviews Genetics*, 6(8):623–632, 2005.

[138] Piotr Szulc, Malgorzata Bogdan, Florian Frommlet, and Hua Tang. Joint genotype- and ancestry-based genome-wide association studies in admixed populations. *Genetic epidemiology*, 41(6):555–566, 2017.

[139] H Gui, AM Levin, S Xiao, M Yang, J Yang, S Hochstadt, K Whitehouse, A Carlton, D Rynkowski, D Lanfear, et al. Joint test of allelic dosage and local ancestry identifies ints3 as new susceptibility gene for asthma among african american individuals. In *A33. Asthma Mechanisms*, pages A1331–A1331. American Thoracic Society, 2018.

[140] Marisa Oliveira, Worachart Lert-itthiporn, Bruno Cavadas, Verónica Fernandes, Ampaiwan Chuansumrit, Orlando Anunciação, Isabelle Casademont, Fanny Koeth, Marina Penova, Kanchana Tangnararatchakit, et al. Joint ancestry and association test indicate two distinct pathogenic pathways involved in classical dengue fever and dengue shock syndrome. *PLoS neglected tropical diseases*, 12(2):e0006202, 2018.

[141] Bogdan Pasaniuc, Noah Zaitlen, Guillaume Lettre, Gary K Chen, Arti Tandon, WH Kao, Ingo Ruczinski, Myriam Fornage, David S Siscovick, Xiaofeng Zhu, et al. 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.

[142] Jorge Duconge and Gualberto Ruaño. The emerging role of admixture in the pharmacogenetics of puerto rican hispanics. *Journal of pharmacogenomics & pharmacoproteomics*, 1(101), 2010.

[143] Laura H Goetz, Liliana Uribe-Bruce, Danjuma Quarless, Ondrej Libiger, and Nicholas J Schork. Admixture and clinical phenotypic variation. *Human heredity*, 77(1-4):73–86, 2014.

[144] Tomris Cesuroglu, Elena Syurina, Frans Feron, and Anja Krumeich. Other side of the coin for personalised medicine and healthcare: content analysis of 'personalised' practices in the literature. *BMJ open*, 6(7):e010243, 2016.

[145] Wolfgang Sadée and Zunyan Dai. Pharmacogenetics/genomics and personalized medicine. *Human molecular genetics*, 14(suppl 2):R207–R214, 2005.

[146] Matthias Schwab and Elke Schaeffeler. Pharmacogenomics: a key component of personalized therapy. *Genome medicine*, 4(11):1, 2012.

[147] Giulio Genovese, Robert E Handsaker, Heng Li, Nicolas Altemose, Amelia M Lindgren, Kimberly Chambert, Bogdan Pasaniuc, Alkes L Price, David Reich, Cynthia C Morton, et al. Using population admixture to help complete maps of the human genome. *Nature genetics*, 45(4):406, 2013.

[148] L Luca Cavalli-Sforza and Marcus W Feldman. The Application of Molecular Genetic Approaches to the Study of Human Evolution. *Nature genetics*, 33:266–275, 2003.

[149] Daniel Wegmann, Darren E Kessner, Krishna R Veeramah, Rasika A Mathias, Dan L Nicolae, Lisa R Yanek, Yan V Sun, Dara G Torgerson, Nicholas Rafaels, Thomas Mosley, et al. Recombination Rates in Admixed Individuals Identified by Ancestry-based Inference. *Nature genetics*, 43(9):847–853, 2011.

[150] David Reich, Nick Patterson, Philip L De Jager, Gavin J McDonald, Alicja Waliszewska, Arti Tandon, Robin R Lincoln, Cari DeLoa, Scott A Fruhan, Philippe Cabre, et al. A Whole-genome Admixture Scan Finds a Candidate Locus for Multiple Sclerosis Susceptibility. *Nature genetics*, 37(10):1113–1118, 2005.

[151] Xiaofeng Zhu, Richard S Cooper, and Robert C Elston. Linkage Analysis of a Complex Disease Through Use of Admixed Populations. *The American Journal of Human Genetics*, 74(6):1136–1153, 2004.

[152] Claire Churchhouse and Jonathan Marchini. Multiway Admixture Deconvolution Using Phased or Unphased Ancestral Panels. *Genetic epidemiology*, 37(1):1–12, 2013.

[153] Eric Y Durand, Chuong B Do, Joanna L Mountain, and J Michael Macpherson. Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution. *bioRxiv*, page 010512, 2014.

[154] Brenna M Henn, Laura R Botigué, Simon Gravel, Wei Wang, Abra Brisbin, Jake K Byrnes, Karima Fadhlaoui-Zid, Pierre A Zalloua, Andres Moreno-Estrada, Jaume Bertranpetit, et al. Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. *PLoS Genet*, 8(1):e1002397, 2012.

[155] Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of Population Structure Using Dense Haplotype Data. *PLoS Genet*, 8(1):e1002453–e1002453, 2012.

[156] Bogdan Pașaniuc, Sriram Sankararaman, Gad Kimmel, and Eran Halperin. Inference of Locus-specific Ancestry in Closely Related Populations. *Bioinformatics*, 25(12):i213–i221, 2009.

[157] Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Reconstructing Genetic Ancestry Blocks in Admixed Individuals. *The American Journal of Human Genetics*, 79(1):1–12, 2006.

[158] Jesse M Rodriguez, Sivan Bercovici, Megan Elmore, and Serafim Batzoglou. Ancestry Inference in Complex Admixtures via Variable-length Markov Chain Linkage Models. *Journal of Computational Biology*, 20(3):199–211, 2013.

[159] James J Yang, Jia Li, Anne Buu, and L Keoki Williams. Efficient Inference of Local Ancestry. *Bioinformatics*, 29(21):2750–2756, 2013.

[160] Salma Khalaf Al-Kaabi and Andrew Atherton. Impact of noncommunicable diseases in the State of Qatar. *ClinicoEconomics and outcomes research: CEOR*, 7:377–385, 2015.

[161] Hebe N Gouda, Fiona Charlson, Katherine Sorsdahl, Sanam Ahmadzada, Alize J Ferrari, Holly Erskine, Janni Leung, Damian Santamauro, Crick Lund, Leopold Ndemnge Aminde, et al. Burden of non-communicable diseases in sub-saharan africa, 1990–2017:

results from the global burden of disease study 2017. *The Lancet Global Health*, 7(10): e1375–e1387, 2019.

[162] Denis Pierron, Margit Heiske, Harilanto Razafindrazaka, Veronica Pereda-loth, Jazmin Sanchez, Omar Alva, Amal Arachiche, Anne Boland, Robert Olaso, Jean-Francois Deleuze, et al. Strong selection during the last millennium for african ancestry in the admixed population of madagascar. *Nature communications*, 9(1):932, 2018.

[163] Sriram Sankararaman, Gad Kimmel, Eran Halperin, and Michael I Jordan. On the inference of ancestries in admixed populations. *Genome research*, 18(4):668–675, 2008.

[164] Bogdan Paşaniuc, Justin Kennedy, and Ion Măndoiu. Imputation-based local ancestry inference in admixed populations. In *International Symposium on Bioinformatics Research and Applications*, pages 221–233. Springer, 2009.

[165] Kyung-Ah Sohn, Zoubin Ghahramani, and Eric P Xing. Robust estimation of local genetic ancestry in admixed populations using a nonparametric bayesian approach. *Genetics*, 191(4):1295–1308, 2012.

[166] Abra Brisbin, Katarzyna Bryc, Jake Byrnes, Fouad Zakharia, Larsson Omberg, Jeremiah Degenhardt, Andrew Reynolds, Harry Ostrer, Jason G Mezey, and Carlos D Bustamante. PCAdmix: Principal Components-based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Human biology*, 84(4):343, 2012.

[167] Youna Hu, Cristen Willer, Xiaowei Zhan, Hyun Min Kang, and Gonçalo R Abecasis. Accurate local-ancestry inference in exome-sequenced admixed individuals via off-target sequence reads. *The American Journal of Human Genetics*, 93(5):891–899, 2013.

[168] Po-Ru Loh, Pier Francesco Palamara, and Alkes L Price. Fast and accurate long-range phasing in a uk biobank cohort. *Nature genetics*, 48(7):811, 2016.

[169] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A New Multipoint Method for Genome-wide Association Studies by Imputation of Genotypes. *Nature genetics*, 39(7):906–913, 2007.

[170] Aude Saint Pierre and Emmanuelle Génin. How important are rare variants in common disease? *Briefings in functional genomics*, 13(5):353–361, 2014.

[171] Michael Salter-Townshend and Simon Myers. Fine-scale inference of ancestry segments without prior knowledge of admixing groups. *Genetics*, pages genetics–302139, 2019.

[172] Gad Kimmel and Ron Shamir. A block-free hidden markov model for genotypes and its application to disease association. *Journal of Computational Biology*, 12(10):1243–1260, 2005.

[173] Pasi Rastas, Mikko Koivisto, Heikki Mannila, and Esko Ukkonen. Phasing genotypes using a hidden markov model. *Bioinformatics Algorithms: Techniques and Applications*, pages 353–372, 2008.

[174] Howard Wolinsky. Genetic genealogy goes global: Although useful in investigating ancestry, the application of genetics to traditional genealogy could be abused. *EMBO reports*, 7(11):1072–1074, 2006.

[175] Juan-Camilo Chacón-Duque, Kaustubh Adhikari, Macarena Fuentes-Guajardo, Javier Mendoza-Revilla, Victor Acuña-Alonzo, Rodrigo Barquera, Mirsha Quinto-Sánchez, Jorge Gómez-Valdés, Paola Martínez Everardo, Hugo Villamil-Ramírez, et al. Latin americans show wide-spread converso ancestry and imprint of local native ancestry on physical appearance. *Nature communications*, 9(1):5388–5388, 2018.

[176] Martin Mächler and Peter Bühlmann. Variable length markov chains: methodology, computing, and software. *Journal of Computational and Graphical Statistics*, 13(2): 435–455, 2004.

[177] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006.

[178] Garrett Hellenthal, George BJ Busby, Gavin Band, James F Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, 2014.

[179] Calvin Chi, Xiaorong Shao, Brooke Rhead, Edlin Gonzales, Jessica B Smith, Anny H Xiang, Jennifer Graves, Amy Waldman, Timothy Lotze, Teri Schreiner, et al. Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry. *PLoS genetics*, 15(1):e1007808, 2019.

[180] Emily T Norris, Lu Wang, Andrew B Conley, Lavanya Rishishwar, Leonardo Mariño-Ramírez, Augusto Valderrama-Aguirre, and I King Jordan. Genetic ancestry, admixture and health determinants in latin america. *BMC genomics*, 19(8):861, 2018.

[181] Farid Rajabli, Briseida E Feliciano, Katrina Celis, Kara L Hamilton-Nelson, Patrice L Whitehead, Larry D Adams, Parker L Bussies, Clara P Manrique, Alejandra Rodriguez,

Vanessa Rodriguez, et al. Ancestral origin of apoe $\varepsilon 4$ alzheimer disease risk in puerto rican and african american populations. *PLoS genetics*, 14(12):e1007791, 2018.

[182] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[183] Mariette Awad and Rahul Khanna. *Hidden Markov Model*. Apress, Berkeley, CA, 2015. ISBN 978-1-4302-5990-9. doi: 10.1007/978-1-4302-5990-9_5. URL https://doi.org/10.1007/978-1-4302-5990-9_5.

[184] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[185] Marius Bartcus, Faicel Chamroukhi, and Hervé Glotin. Hierarchical dirichlet process hidden markov model for unsupervised bioacoustic analysis. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2015.

[186] Nilesh Tripuraneni, Shixiang Shane Gu, Hong Ge, and Zoubin Ghahramani. Particle gibbs for infinite hidden markov models. In *Advances in Neural Information Processing Systems*, pages 2395–2403, 2015.

[187] Ioan Stanculescu, Christopher KI Williams, and Yvonne Freer. Autoregressive hidden markov models for the early detection of neonatal sepsis. *IEEE journal of biomedical and health informatics*, 18(5):1560–1570, 2013.

[188] Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.

[189] Nuria Oliver, Ashutosh Garg, and Eric Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, 96(2):163–180, 2004.

[190] Lorenzo Rimella and Nick Whiteley. Exploiting locality in high-dimensional factorial hidden markov models. *J. Mach. Learn. Res.*, 23:4–1, 2022.

[191] Yee Whye Teh and Michael I Jordan. Hierarchical bayesian nonparametric models with applications. *Bayesian Nonparametrics*, 28(158):42, 2009.

[192] Matthew J Beal, Zoubin Ghahramani, and Carl E Rasmussen. The infinite hidden markov model. In *Advances in neural information processing systems*, pages 577–584, 2002.

[193] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. doi: 10.1198/016214506000000302.

[194] Subhashis Ghosal. The dirichlet process, related priors and posterior asymptotics. *Bayesian nonparametrics*, 28:35, 2010.

[195] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. A sticky hdp-hmm with application to speaker diarization. *The Annals of Applied Statistics*, pages 1020–1056, 2011.

[196] Edoardo M Airoldi, David Blei, Elena A Erosheva, and Stephen E Fienberg. *Handbook of mixed membership models and their applications*. CRC press, 2014.

[197] Zoubin Ghahramani, Michael I Jordan, and Padhraic Smyth. Factorial hidden markov models. *Machine learning*, 29(2-3):245–273, 1997.

[198] Tingting Liu and Jan Lemeire. Effective and efficient identification of persistent-state hidden (semi-) markov models. In *STAIRS 2014*, pages 171–180. IOS Press, 2014.

[199] F Abegaz, K Chaichoompu, E Genin, DW Fardo, I König, JM Mahachie John, and Kristel Van Steen. Principals about principal components in statistical genetics. *Briefings in Bioinformatics*, 20(6):2200–2216, 2018.

[200] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genet*, 2(12):e190, 2006.

[201] Lloyd T Elliott, Maria De Iorio, Stefano Favaro, Kaustubh Adhikari, Yee Whye Teh, et al. Modeling population structure under hierarchical dirichlet processes. *Bayesian Analysis*, 14(2):313–339, 2019.

[202] Jianzhong Ma and Christopher I Amos. Theoretical formulation of principal components analysis to detect and correct for population stratification. *PloS one*, 5(9):e12510, 2010.

[203] Paul HC Eilers and Renée X De Menezes. Quantile smoothing of array cgh data. *Bioinformatics*, 21(7):1146–1153, 2005.

[204] Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support vector machines and kernels for computational biology. *PLoS Comput Biol*, 4(10):e1000173, 2008.

[205] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[206] Charles Sutton and Andrew McCallum. An Introduction to Conditional Random Fields. *Machine Learning*, 4(4):267–373, 2011.

[207] Yizhen Zhong, Minoli A Perera, and Eric R Gamazon. On using local ancestry to characterize the genetic architecture of human traits: Genetic regulation of gene expression in multiethnic or admixed populations. *The American Journal of Human Genetics*, 2019.

[208] Jacquiline W Mugo, Ephifania Geza, Joel Defo, Samar SM Elsheikh, Gaston K Mazandu, Nicola J Mulder, and Emile R Chimusa. A multi-scenario genome-wide medical population genetics simulation framework. *Bioinformatics*, 33(19):2995–3002, 2017.

[209] Evans Kiptoo Cheruiyot, Rawlynce Cheruiyot Bett, Joshua Oluoch Amimo, Zhang Yi, Raphael Mrode, and Fidalis Denis Mujibi. Signatures of selection in admixed dairy cattle in tanzania. *Frontiers in genetics*, 9:607, 2018.

[210] Qianqian Zhang, Mario PL Calus, Mirte Bosse, Goutam Sahana, Mogens Sandø Lund, and Bernt Guldbrandtsen. Human-mediated introgression of haplotypes in a modern dairy cattle breed. *Genetics*, 209(4):1305–1317, 2018.

[211] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. PLINK: A Tool Set for Whole-genome Association and Population-based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[212] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, et al. Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443, 2016.

[213] Daniel Hui, Zhou Fang, Jerome Lin, Qing Duan, Yun Li, Ming Hu, and Wei Chen. Lait: a local ancestry inference toolkit. *BMC genetics*, 18(1):83, 2017.

[214] Batool Mutar Mahdi. Introductory chapter: Concept of human leukocyte antigen (hla). In Batool Mutar Mahdi, editor, *Human Leukocyte Antigen (HLA)*, chapter 1. IntechOpen, Rijeka, 2019. doi: 10.5772/intechopen.83727.

[215] Diogo Meyer, Vitor RC Aguiar, Bárbara D Bitarello, Débora YC Brandt, and Kelly Nunes. A genomic perspective on hla evolution. *Immunogenetics*, 70(1):5–27, 2018.

[216] Benjamin Lê Cook and Willard G Manning. Thinking beyond the mean: a practical guide for using quantile regression methods for health services research. *Shanghai archives of psychiatry*, 25(1):55, 2013.

[217] Alireza Baratloo, Mostafa Hosseini, Ahmed Negida, and Gehad El Ashal. Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. *Emergency*, 3(2):48–49, 2015.

[218] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

[219] Tom Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine learning*, 31(1):1–38, 2004.

[220] Yeliz Karaca and Carlo Cattani. *Computational Methods for Data Analysis*. Walter de Gruyter GmbH & Co KG, 2018.

[221] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 2018.

[222] Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 28(4):289–301, 2005.

[223] Alkes L Price, Nick Patterson, Fuli Yu, David R Cox, Alicja Waliszewska, Gavin J McDonald, Arti Tandon, Christine Schirmer, Julie Neubauer, Gabriel Bedoya, et al. A Genomewide Admixture Map for Latino Populations. *The American Journal of Human Genetics*, 80(6):1024–1036, 2007.

[224] Katarzyna Bryc, Eric Y Durand, J Michael Macpherson, David Reich, and Joanna L Mountain. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*, 96(1):37–53, 2015.

[225] Andries T Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, 27(2):e1608, 2018.

[226] François Pompanon, Aurélie Bonin, Eva Bellemain, and Pierre Taberlet. Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, 6(11):847, 2005.

[227] Mathieu Bourgey, Mathieu Lariviere, Chantal Richer, and Daniel Sinnett. Alg: automated genotype calling of luminex assays. *PLoS One*, 6(5):e19368, 2011.

[228] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564, 2010.

[229] Morgan Mayer-Jochimsen, Shannon Fast, and Nathan L Tintle. Assessing the impact of differential genotyping errors on rare variant tests of association. *PloS one*, 8(3): e56626, 2013.

[230] Bo Peng, Christopher I Amos, and Marek Kimmel. Forward-time simulations of human populations with complex diseases. *PLoS genetics*, 3(3):e47, 2007.

[231] Guillaume Laval and Laurent Excoffier. Simcoal 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, 20(15):2485–2487, 2004.

[232] Rui Zhang, Chang Liu, Kai Yuan, Xumin Ni, Yuwen Pan, and Shuhua Xu. Admixsim 2: a forward-time simulator for modeling complex population admixture. *BMC bioinformatics*, 22(1):1–15, 2021.

[233] Robert Brown and Bogdan Pasaniuc. Enhanced Methods for Local Ancestry Assignment in Sequenced Admixed Individuals. *PLoS Comput Biol*, 10(4):e1, 2014.

[234] Zachariah Gompert. A continuous correlated beta process model for genetic ancestry in admixed populations. *PloS one*, 11(3):e0151047, 2016.

[235] Negar Khayatzadeh, Gabor Mészáros, YT Utsunomiya, Jose Fernando Garcia, Urs Schnyder, Birgit Gredler, Ino Curik, and Johann Soelkner. Locus-specific ancestry to detect recent response to selection in admixed swiss fleckvieh cattle. *Animal genetics*, 47(6):637–646, 2016.

[236] Romuald Laso-Jadart, Christine Harmant, Hélène Quach, Nora Zidane, Chris Tyler-Smith, Qasim Mehdi, Qasim Ayub, Lluis Quintana-Murci, and Etienne Patin. The genetic legacy of the indian ocean slave trade: recent admixture and post-admixture selection in the makranis of pakistan. *The American Journal of Human Genetics*, 101 (6):977–984, 2017.

[237] Burak Yelmen, Mayukh Mondal, Davide Marnetto, Ajai K Pathak, Francesco Montinaro, Irene Gallego Romero, Toomas Kivisild, Mait Metspalu, and Luca Pagani. Ancestry-Specific Analyses Reveal Differential Demographic Histories and Opposite Selective Pressures in Modern South Asian Populations. *Molecular Biology and Evolution*, 36 (8):1628–1642, 04 2019.

[238] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2): 442–451, 1975.

[239] Jan Gorodkin. Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, 28(5-6):367–374, 2004.

[240] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[241] Sepand Haghighi, Masoomeh Jasemi, Shaahin Hessabi, and Alireza Zolanvari. Pycm: Multiclass confusion matrix library in python. *J. Open Source Software*, 3(25):729, 2018.

[242] Yasen Jiao and Pufeng Du. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology*, 4(4):320–330, 2016.

[243] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363, 2005.

[244] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[245] Paloma Medina, Bryan Thornlow, Rasmus Nielsen, and Russell Corbett-Detig. Estimating the timing of multiple admixture pulses during local ancestry inference. *Genetics*, 210(3):1089–1107, 2018.

[246] Xumin Ni, Kai Yuan, Chang Liu, Qidi Feng, Lei Tian, Zhiming Ma, and Shuhua Xu. Multiwaver 2.0: modeling discrete and continuous gene flow to reconstruct complex population admixtures. *European Journal of Human Genetics*, page 1, 2018.

[247] Aurélien Cottin, Benjamin Penaud, Jean-Christophe Glaszmann, Nabila Yahiaoui, and Mathieu Gautier. Simulation-based evaluation of three methods for local ancestry deconvolution of non-model crop species genomes. *G3: Genes, Genomes, Genetics*, 10(2): 569–579, 2020.

[248] Hassan Aliloo, Raphael Mrode, AM Okeyo, and John P Gibson. Ancestral haplotype mapping for gwas and detection of signatures of selection in admixed dairy cattle of kenya. *Frontiers in Genetics*, 11:544, 2020.

[249] Cesar Fortes-Lima, Antoine Gessain, Andres Ruiz-Linares, Maria-Cátira Bortolini, Florence Migot-Nabias, Gil Bellis, J Víctor Moreno-Mayar, Berta Nelly Restrepo, Winston Rojas, Efren Avendaño-Tamayo, et al. Genome-wide ancestry and demographic history of african-descendant maroon communities from french guiana and suriname. *The American Journal of Human Genetics*, 101(5):725–736, 2017.

[250] Iman Hamid, Katharine L Korunes, Sandra Beleza, and Amy Goldberg. Rapid adaptation to malaria facilitated by admixture in the human population of cabo verde. *Elife*, 10: e63177, 2021.

[251] Matthew J Johnson and Alan S Willsky. Bayesian nonparametric hidden semi-markov models. *Journal of Machine Learning Research*, 14(Feb):673–701, 2013.

[252] Colin Reimer Dawson, Chaofan Huang, and Clayton T Morrison. An infinite hidden markov model with similarity-biased transitions. In *International Conference on Machine Learning*, pages 942–950, 2017.

[253] Ioannis Sgouralis and Steve Pressé. An introduction to infinite hmms for single-molecule data analysis. *Biophysical journal*, 112(10):2021–2029, 2017.

[254] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. An HDP-HMM for systems with state persistence. In *Proceedings of the 25th international conference on Machine learning*, pages 312–319. ACM, 2008.

[255] John J Welch and Chris D Jiggins. Standing and flowing: the complex origins of adaptive variation. *Molecular ecology*, 23(16):3935–3937, 2014.

[256] Lawrence M Leemis and Jacquelyn T McQueston. Univariate distribution relationships. *The American Statistician*, 62(1):45–53, 2008.

[257] Yousry Abdelkader and Zainab Al-Marzouq. Probability distribution relationships. *Statistica*, 70(1):41–51, 2010.

[258] Ivo D Dinov, Kyle Siegrist, Dennis K Pearl, Alexandr Kalinin, and Nicolas Christou. Probability distributome: a web computational infrastructure for exploring the properties, interrelations, and applications of probability distributions. *Computational statistics*, 31(2):559–577, 2016.

[259] Lawrence M Leemis. Relationships among common univariate distributions. *The American Statistician*, 40(2):143–146, 1986.

[260] Robert V Hogg and Allen T Craig. *Introduction to mathematical statistics. (7th Edition)*. Pearson, 2012.

[261] Catherine Forbes, Merran Evans, Nicholas Hastings, and Brian Peacock. *Statistical distributions*. John Wiley & Sons, 2011.

[262] Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. *Dirichlet and related distributions: Theory, methods and applications*. John Wiley & Sons, 2011.

[263] Yee Whye Teh. Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer, 2011.

[264] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[265] David Blackwell, James B MacQueen, et al. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[266] Zoubin Ghahramani. Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110553, 2013.

[267] Viet-An Nguyen, Jordan Boyd-Graber, and Stephen F Altschul. Dirichlet mixtures, the dirichlet process, and the structure of protein space. *Journal of computational biology*, 20(1):1–18, 2013.

[268] Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

[269] Quinn T Ostrom, Kathleen M Egan, L Burt Nabors, Travis Gerke, Reid C Thompson, Jeffrey J Olson, Renato LaRocca, Sajeel Chowdhary, Jeanette E Eckel-Passow, Georgina Armstrong, et al. Glioma risk associated with extent of estimated european genetic ancestry in african americans and hispanics. *International journal of cancer*, 00:00–00, 2019.

[270] Lesedi M Williams, Zhihua Qi, Ken Batai, Stanley Hooker, Nancy J Hall, Roberto F Machado, Alice Chen, Sally Campbell-Lee, Yongtao Guan, Rick Kittles, et al. A locus on chromosome 5 shows african ancestry–limited association with alloimmunization in sickle cell disease. *Blood advances*, 2(24):3637–3647, 2018.

[271] Aditi Shendre, Howard Wiener, Marguerite R Irvin, Degui Zhi, Nita A Limdi, Edgar T Overton, Christina L Wassel, Jasmin Divers, Jerome I Rotter, Wendy S Post, et al. Admixture mapping of subclinical atherosclerosis and subsequent clinical events among african americans in 2 large cohort studies. *Circulation: cardiovascular genetics*, 10(2): e001569, 2017.

[272] Maud Duranton, François Allal, Christelle Fraïsse, Nicolas Bierne, François Bonhomme, and Pierre-Alexandre Gagnaire. The origin and remolding of genomic islands of differentiation in the european sea bass. *Nature communications*, 9(1):2518, 2018.

[273] Denis Pierron, Margit Heiske, Harilanto Razafindrazaka, Ignace Rakoto, Nelly Rabetokotany, Bodo Ravololomanga, Lucien M-A Rakotozafy, Mireille Mialy Rakotomalala, Michel Razafiarivony, Bako Rasoarifetra, et al. Genomic landscape of human diversity across madagascar. *Proceedings of the National Academy of Sciences*, 114(32):E6498–E6506, 2017.

Appendix A

Some important mathematical concepts

A.1 Probability distributions common in genetic ancestry

Although several probability distributions exist, they tend to relate to each other [256, 257]. Relationships among distributions simplify modelling in biological, physical, medical and environmental applications. More so, they ease computational simulations which are helpful in understanding important process characteristics [258, 259]. Distributions can related by: (1) special case, for example, a Gamma distribution whose shape and scale parameters are $\alpha = 1$ and $\beta > 0$, respectively is exponential with mean β; (2) transformation, for example, the Gamma distribution can be transformed into a Chi-square and vice versa; (3) Bayesian, for example, the binomial distribution with parameters (n, p) where p is Beta distributed, and; (4) limiting (asymptotic), for example, the Beta with equal parameters (shape and scale: $\alpha = \beta$) approximates a normal distribution as $\beta \to \infty$ [258]. This section gives a brief introduction on some of the most widely used probability distributions in genetic inference, including the Gamma, Dirichlet, Dirichlet process, multinomial and multivariate normal distribution.

A.1.1 The Gamma distribution

The Gamma distribution models waiting times in econometrics and Bayesian statistics where it is a conjugate prior to the Poisson and exponential distributions. It has two parameters, $\alpha, \beta > 0$ representing the shape and scale parameter of the Gamma density, respectively. A

Gamma distributed random variable X, has a probability distribution function given by

$$f(x) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} I_{x>0},$$

where

$$I_{\{x>0\}} = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

and $\Gamma(.)$ is the Gamma function [260]. The Gamma distribution is either denoted by $\text{Gamma}(\alpha,\beta)$ or $G(\alpha,\beta)$ or $\Gamma(\alpha,\beta)$. Given $(X_\iota)_{1\leq\iota\leq n} \sim \text{Gamma}(\alpha,\beta)$ then, $\sum_{\iota=1}^{n} X_\iota \sim$ $\text{Gamma}(\sum_{\iota=1}^{n}\alpha_\iota,\beta)$. More so, given two independent Gamma distributed random variables, say $X_j \sim \text{Gamma}(\alpha_j,\beta)$ for $j = 1,2$ then, a random variable Z which is Beta distributed can be formed from Gamma variables [260], such that

$$Z = \frac{X_1}{X_1 + X_2} \sim \text{Beta}(\alpha_1 + \alpha_2, \beta).$$

A.1.2 The Dirichlet distribution

Similarly to the Gamma distribution, the Dirichlet distribution belongs to the exponential family of distributions. However, it is most suitable for modelling the random behaviour of percentages or proportions [261] It is most popular in genetics statistics, nonparametric inferences, reliability theory and stochastic processes [262]. For example, it has been successful in modelling the distribution of ancestry proportions contributed by each ancestral population in admixed individuals [78, 87, 165]. Dirichlet distribution belongs to the continuous multivariate probability distributions family. It is parametrised by a vector of positives [262]. A vector of random variables $\mathbf{X} = (X_k)_{1\leq k\leq K}$ is Dirichlet distributed with a vector of parameters $\alpha = (\alpha_1,\alpha_2,\ldots,\alpha_K)$, denoted by $(X_1,\ldots,X_K) \sim \text{Dir}(\alpha_1,\ldots,\alpha_K)$ if its probability density function is given by

$$f(x_1,x_2,\ldots,x_K) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} x_\iota^{\alpha_\iota - 1} I_{\{x_\iota>0\}} \tag{A.1.1}$$

where $\alpha_0 = \sum_{k=1}^{K}\alpha_k$, $\sum_{k=1}^{K-1} x_k < 1$ and $x_K = 1 - \sum_{\iota=1}^{K} x_\iota$. Now, if $K = 2$, Equation (A.1.1) reduces to the probability density function of a Beta distribution. Thus, the Beta distribution is a particular instance of a Dirichlet distribution. The Dirichlet distribution is a conjugate

prior to the categorical and multinomial distributions. Assuming $\mathbf{X} = (X_1, \ldots, X_K)$ follows a multinomial distribution with parameters n and $\mathbf{p} = (p_1, \ldots, p_K)$, the probability mass function is given by

$$f(X_1 = x_1, \ldots, X_K = x_K) = \frac{n!}{x_1! x_2! \ldots x_K!} \prod_{i=1}^{K} p_i^{x_i}, \tag{A.1.2}$$

where $p_\iota \geq 0$ is the probability parameter such that $\sum_{\iota=1}^{K} p_\iota = 1$, $(x_\iota)_{1 \leq \iota \leq K} \in \{0, 1, \ldots, n\}$ and $x_K = n - \sum_{\iota=1}^{K-1} x_\iota$.

A.1.3 The Dirichlet process

A Dirichlet process (DP) is a probability distribution over discrete probability distributions [263]. It generalises the Dirichlet distribution (**Section A.1.2**) [263]. Unlike the Dirichlet distribution, the DP allow the number of parameters to grow with data [194]. It consists of two parameters: the concentration, $\alpha_1 > 0$ and the base measure distribution, H which is the expectation of the DP. **Definition A.1.4** is the formal definition of DP.

A.1.4 Definition. A random measure G_0 on $(\Theta, \mathscr{B})$ is generated by a Dirichlet process, $DP(\alpha_1 H)$, if for each finite measurable partition, $\{T_1, T_2, \ldots, T_K\}$ on Θ, the joint vector of probabilities is Dirichlet distributed, i.e,

$$(G_0(T_1), G_0(T_2), \ldots, G_0(T_K)) \sim \text{Dir}\left(\alpha H(T_1), \alpha H(T_2), \ldots, \alpha H(T_K)\right), \tag{A.1.3}$$

then

$$G_0 \sim DP(\alpha_1, H) \tag{A.1.4}$$

where $\alpha_1 > 0$ and H are the concentration parameter and base measure of the DP, respectively [191, 193, 194, 195, 263].

There are different perspectives of a DP, these include: the explicit representation according to the stick breaking construction [264] and the implicit representation through the Polya urn scheme or the Chinese restaurant process [265]. We provide more details on theses in the next sections.

A.1.4.1 Stick breaking construction (SBC)

According to Sethuraman [264], given two independent sequences (β_ℓ) and (θ_ℓ) for $\ell \geq 1$, a random draw G_0 in Equation (A.1.4) can be represented as

$$G_0 = \sum_{\ell=1}^{\infty} \beta_\ell \delta_{\theta_\ell}, \qquad \ell = 1, 2, \ldots \tag{A.1.5}$$

where δ_{θ_ℓ} is a point mass function at θ, for θ distinct valued atoms, where $\theta_\ell \overset{\text{i.i.d}}{\sim} H$ and β_ℓ are non-negative weights, so that $\sum_{\ell=1}^{\infty} \beta_\ell = 1$ [165, 264] with probability 1. β_ℓ's are sampled according to the stick breaking construction (SBC) [194, 196, 266] as follows

$$\beta'_\ell = \text{Beta}(1, \alpha_1), \qquad \beta_\ell = \beta'_\ell \prod_{m=1}^{\ell-1} (1 - \beta'_m). \tag{A.1.6}$$

The distribution of $\beta = (\beta_1, \beta_2, \ldots)$ in Equation (A.1.6) is also called the Griffiths Engen McCloskey and is often represented as

$$\beta \sim \text{GEM}(\alpha_1) \tag{A.1.7}$$

Metaphorically, Equation (A.1.6) or (A.1.7) represents the process of breaking a stick of unit length into pieces each of weight β_l, which is a random proportion of the stick that remains after considering all the $\ell - 1$ weights [195]. The first piece is of weight $\beta_1 = \beta'_1 \sim \text{Beta}(1, \alpha_1)$ and the remaining stick is $(1 - \beta_1)$. Breaking the stick for the second time yields a piece of weight $\beta_2 = \text{Beta}(1, \alpha_2)(1 - \beta'_1)$. The process continues until the length of the remaining piece is zero [196]. Equation (A.1.5) shows that a DP can be expressed as an infinite mixture of point masses, δ_{θ_l} and with probability 1, it is discrete [196, 195].

Unlike the explicit representation which models a random draw G_0 itself, implicit representations model probability measures drawn from G_0 [267, 254]. Consider

$$\theta'_u | \theta'_1, \theta'_2, \ldots, \theta'_{u-1}, G_0 \sim G_0, \tag{A.1.8}$$

Since draws from a DP are discrete, it is always possible for a sample of draws from G_0 to have repeated values [196].

Let z_u be the indicator random variable for the distinct values $(\theta_l)_{l=1}^{K}$ such that $\theta'_u = \theta_{z_u}$.

Therefore, given that N observed, the distribution of the $N+1^{\text{th}}$ observation is as follows

$$p(z_{N+1}|z_1,\ldots,z_N,\alpha_1) = \frac{1}{N+\alpha_1}\left(\sum_{l=1}^{L} n_\ell I(z,l) + \alpha_1 I(z,L+1)\right). \tag{A.1.9}$$

where $I(z,\ell)$ is Kronecker delta, n_ℓ and L counts the number of $z_u = l$ and distinct values that have been seen so far, and and $L+1$ is a previously unseen value.

Given a Polya urn model, θ'_u corresponds to the ball and θ_ℓ is its colour. Initially, the urn is empty. Eventually, choose a colour from H, that is, $\theta_1 \sim H$, paint a ball in that colour and drop the ball in the urn. Subsequently, either reach the urn, draw a ball and replace it by two balls of the same colour with a probability proportional to the quantity of balls of that colour in the urn (n_l) or choose a new colour from H, paint the ball in this colour and drop it in the urn [191, 192, 254, 263].

Apart from describing a Polya urn scheme, Equation (A.1.9) also metaphorically describes the Chinese restaurant process [193, 254]. Details on this metaphor are given in the next section [33, 185, 263].

A.1.4.2 The Chinese restaurant process (CRP)

This is an analogy of the way customers sit in a restaurant assuming an infinite number of tables and dishes [194, 196]. Since each table serves a unique dish θ_ℓ which can be served at multiple tables, the dish characterises cluster (class) membership [196]. The u^{th} customer denoted by θ'_u enters the restaurant and sits at table indicated by z_u with a probability determined by either the number of customers occupying that table (n_ℓ) or the concentration parameter α_1 if the customer is the first occupant of the table. Thus, increasing the number of occupied tables by 1 [191]. Considering L as the number unique dishes served to the N customers already in the restaurant, we note Equation (A.1.9) describes the CRP. Thus, the CRP is a Polya urn with customers corresponding to balls and colours to tables which automatically is the dish the customer is served.

Mainly, the CRP asserts that, given

$$z_u \sim \beta, \qquad\qquad \beta \sim \text{GEM}(\alpha_1)$$

integrating out β yields the predictive distribution on z in Equation (A.1.9). Now, given N draws (draw as previously described), the question is, what is the distribution over L the

number of distinct values of z_u? Antoniak [268] showed that, given Equation (A.1.7), then

$$p(L|N,\alpha_1) = \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1+N)} s(N,L)\alpha_1^L \tag{A.1.10}$$

where $s(N,L)$ defines unassigned Stirling numbers of the first kind [186, 195].

Putting everything together, the generative model of a DP mixture model is as follows

$$\begin{aligned} G_0 &= \sum_{\ell=1}^{\infty} \beta_l \delta_\ell, & \beta &\sim \text{GEM}(\alpha_1), & \theta_l &\sim H, \\ z_u &\sim \beta, & y_u &\sim f(\theta_{z_u}), \end{aligned} \tag{A.1.11}$$

where z_u is the indicator variable (just as defined before).

A.1.5 Multivariate normal distribution

Assume phased ancestral and admixed individuals. Let the number of haplotypes in each ancestry be $2n_k$ for haplotypes in every population k given by $H^{(k)} = (h_1^{(k)}, \ldots, h_{2n}^{(k)})$, where haplotype ι of population k is $h_\iota^{(k)} = (h_{\iota 1}^{(k)}, h_{\iota 2}^{(k)}, \ldots, h_{\iota T}^{(k)})$ such that $h_{\iota t}^{(k)} \in \{0,1\}$, $\iota \in \{1,\ldots,2n_k\}$. MULTIMIX [152] and PCADMIX [166] are the two PCA-based method. They both split the genome into windows, each with at least m SNPs so that altogether there are a total of $W = T/m$ windows. MULTIMIX first infer the *fitted ancestry* as an estimate of the local ancestry. Therefore, the fitted ancestry may be the same or differ from the actual local ancestry at a site. Given the fitted ancestry in the w^{th} window, X_w, the observed admixed haplotypes in window w, Y_w, are normally distributed. That is,

$$Y_w|X_w \sim N(\mu_{X_w w}, \Sigma_{X_w w} + \lambda \mathbb{I}_m) \tag{A.1.12}$$

where μ_{kw} and Σ_{kw} for $k = X_w$ estimates the mean (for the first $K-1$ principal components: PCs, in the case of PCADMIX), and variance-covariance matrix of SNPs (scores) that were copied from population k within window $w \in \{1,\ldots,W\}$. The parameter $\lambda > 0$, guarantees the invertibility of the variance-covariance matrix, while $\mathbb{I}_m$ is a unit matrix. In the PCADMIX model, Equation (A.1.12), models correlations between observations. Unlike in MULTIMIX, the observed haplotype, Y_w is replaced by a vector of ancestry scores for admixed haplotype ι in the w^{th} window denoted by $Y_{\iota w}$ on the first $(K-1)$ PCs. The fitted ancestry, X_w, is replaced by the ancestry of haplotype ι in the w^{th} window $X_{\iota w}$. Contrary to MULTIMIX, PCADMIX does not depend on $\lambda \mathbb{I}_m$, yet, requires $(\Sigma_{X_w w}) = \max\{\varsigma, 0.0001\}$ where ς is an

empirical covariance so that

$$\text{Cov } (X_{\iota w}^{u}, X_{\iota w}^{v}) < \sqrt{\text{Var } (X_{\iota w}^{u})\ \text{Var } (X_{\iota w}^{v})}$$

where u and v are two PCs, and $X_{\iota w}^{a}$ defines a vector of ancestry scores for haplotype ι in window w for principal component a.

A.2 Inference and finite space Markov models

As aforementioned, HMM and their extensions all have three model parameters, i.e, the transition, initial and emission probability models. HMM and extensions aim at recovering the hidden state sequence that generated the observed sequence. This is done by answering three questions: to what extent are model parameters able to match the observed sequence (evaluation), what is the most likely hidden state sequence that generated the observed sequence (decoding) and what are the optimal HMM parameters that describe the given observed sequence (learning)? In the next sections, we discuss the solutions to these three questions.

A.2.1 To what extent can HMM parameters match the observed sequence?

Knowing the observed sequence and the HMM parameters, the probability of a state k at t is

$$\begin{aligned} P(X_t = k|Y_{1:T}) =& \frac{P(\overbrace{X_t = k, Y_{1:t}}^{A}, \overbrace{Y_{t+1:T}}^{B})}{P(Y_{1:T})} \\ \propto& P(\overbrace{X_t = k, Y_{1:t}}^{A}, \overbrace{Y_{t+1:T}}^{B}) \\ =& P(\overbrace{Y_{t+1:T}|X_t = k, Y_{1:t}}^{B|A}) P(\overbrace{X_t = k, Y_{1:t}}^{A}) \end{aligned} \tag{A.2.1}$$

Considering the joint probability in Equation (A.2.1) we have

$$\begin{aligned}
P(X_t = k, Y_{1:t}) =& \sum_{k'} P(X_t = k, X_{t-1} = k', Y_{1:t-1}, Y_t) \\
=& \sum_{k'} P(X_t = k | X_{t-1} = k') P(Y_t | X_t = k, Y_{1:t-1}) P(X_{t-1} = k', Y_{1:t-1}) \\
=& \sum_{k'} \overbrace{P(X_t = k | X_{t-1} = k')}^{\text{transition}} \overbrace{P(Y_t | X_t = k)}^{\text{emission}} P(X_{t-1} = k', Y_{1:t-1}) \\
=& e_k(Y_t) \sum_{k'} s_{k'k} P(X_{t-1} = k', Y_{1:t-1})
\end{aligned} \tag{A.2.2}$$

Setting $\alpha_t(k) = P(X_t = k, Y_{1:t})$ and substituting into Equation (A.2.2) yields

$$\alpha_t(k) = e_k(Y_t) \sum_{k'} s_{k'k} \alpha_{t-1}(k') \tag{A.2.3}$$

Also, setting $\beta_t(k) = P(Y_{t+1:T} | X_t = k)$, and, $\gamma_t(k) = P(X_t = k | Y_{1:T})$ and substituting for $\beta_t(k)$, $\alpha_t(t)$ and $\gamma_t(k)$ in Equation (A.2.1) yields

$$\gamma_t(k) = \frac{\alpha_t(k)\beta_t(k)}{\sum_k \alpha_t(k)\beta_t(k)}. \tag{A.2.4}$$

Therefore, in addition to its usefulness in the computation of the most likely state at t, the forward-backward probabilities compute the marginal likelihood, given in the denominator of Equation (A.2.4). $\beta_t(k)$ is called the backward probability and similarly to the recursive formula in Equation (A.2.3) it has the recursive expression given by

$$\begin{aligned}
\beta_t(k) =& \sum_{k'} P(Y_{t+1:T}, X_{t+1} = k' | X_t = k) \\
=& \sum_{k'} P(Y_{t+1} | X_{t+1} = k') P(Y_{t+2:T} | X_{t+1} = k') P(X_{t+1} | X_t = k) \\
=& \sum_{k'} e_{k'}(Y_{t+1}) s_{kk'} \beta_{t+1}(k')
\end{aligned} \tag{A.2.5}$$

A.2.2 What is the most likely unobserved sequence

The Viterbi algorithm is one of the most common algorithm for determining the most probable hidden state sequence. Denote the most probable hidden state sequence by X^* is

$$X^* = \text{argmax}_{X_{1:T}} P(X_{1:T}|Y_{1:T}). \tag{A.2.6}$$

Since, $P(X_{1:T}|Y_{1:T}) \propto P(X_{1:T}, Y_{1:T})$, then $\text{argmax}_{X_{1:T}} P(X_{1:T}|Y_{1:T}) = \text{argmax}_{X_{1:T}} P(X_{1:T}, Y_{1:T})$. Now, setting $\mu_t(X_t) = \max_{X_{1:t-1}} [P(X_{1:t}, Y_{1:t})]$ and using the HMM probabilistic graph we have

$$\begin{aligned}
\mu_t(X_t) &= \max_{X_{1:t-1}} [P(X_t|X_{t-1})P(Y_t|X_t)P(X_{t-1}, Y_{1:t-1})] \\
&= \max_{X_{t-1}} [P(X_t|X_{t-1})P(Y_t|X_t)] \max_{X_{t-2}} P(X_{t-1}, Y_{1:t-1}) \\
&= \max_{X_{t-1}} \left[e_{X_t}(Y_t) s_{X_{t-1}X_t}\right] \mu_{t-1}(X_{t-1})
\end{aligned} \tag{A.2.7}$$

Therefore,

$$\mu_t(X_t) = \begin{cases} P(X_1)P(Y_1|X_1) & \text{if } t = 1 \\ \max_{X_{t-1}} \left[e_{X_t}(Y_t) s_{X_{t-1}X_t}\right] \mu_{t-1}(X_{t-1}) & \text{if } t \geq 2 \end{cases}$$

Appendix B

Existing models, application and evaluation studies

Table B.1: Evaluation studies on local ancestry inference models.

Study description	Study population (samples and markers)	Rank as per findings	Reference
Introducing a method: WINPOP	Simulated data on two- and three-way admixtures. Panel samples from WTCCC control groups, HGDP and HapMap	WINPOP, LAMP, HAPAA, SABER	[156]
Introducing a method: LAMP-LD	Simulated two- and three-way admixtures and real data (Mexicans and Purto Ricans) on one chromosome	LAMP-LD, WINPOP, GEDI-ADMIX and HAPMIX	[80]
Introducing a method: SUPPORTMIX	Qatar population of 156 individuals on 71 982 SNPs	SUPPORTMIX, WINPOP	[125]
Introducing a method: PCADMIX	Simulated two- and three-way admixture and NYU Latino data	PCADMIX, LAMP, HAPMIX (in three-way admixtures)	[166]
Introducing a method: mSPECTRUM	Simulated (two- and three-way admixtures) and real (22 chromosomes of HGDP) dataset	mSPECTRUM, LAMP, HAPMIX	[165]
Introducing a method: ALLOY	Six simulated populations (two-, three- and four-way admixtures), each with 100 chromosome 1 HapMap samples	ALLOY, WINPOP	[158]
Introducing a method: RFMIX	Simulated two- and three-way admixtures on 519 937 SNPs from chromosome 1 and 11	RFMIX, LAMP-LD, SUPPORTMIX	[129]
Introducing a method: EILA	Six two-way simulated admixed populations each with 30 samples from chromosome 1 on $\approx$ 70 000 SNPs	EILA, LAMP, HAPMIX	[159]
Association studies	489 related Puerto and Mexican ethnic groups and 3 204 unrelated Latino individuals from MEC on 127 935 SNPs	WINPOP, LAMP-LD, ALLOY and PCADMIX	[119]
Association studies	749 simulated five-way admixture and 733 real South African coloureds on 229 076 SNPs	LAMP-LD, WINPOP	[66]
Admixture mapping	GAW18 Mexican American cohort on 37 438 chromosome 3 SNPs	LAMP-LD, MULTIMIX, LAMP	[104]
Introducing a method: ELAI	Simulated two- and three-way admixtures, 58 HapMap3 Mexican and 66 samples from the 1000 GP on the whole genome	ELAI, LAMP-LD, HAPMIX	[97]
Determining selection signals	two dbGaP datasets: 815 Mexicans children from 261 families on 352 754 Viva SNPs, and 1 117 cases and 1 112 controls on 479 757 Lipid SNPs	ELAI, LAMP-LD, RFMIX and MULTIMIX	[102]
Determining place and admixture time	2 540 Ashkenazi Jews based on the following reference panels, 217 west Europeans, 112 east Europeans, 162 south Europeans, 52 Iberians, 146 Levant, 77 South-ME and 70 Druze	RFMIX, LAMP-LD	[128]
Introducing a method: LOTER	Simulated two- and three-way human admixtures (20 samples) and 36 real Populus individuals on 500 000 SNPs	LOTER, RFMIX, LAMP-LD	[56]
Detecting selection signals	700 Malagasy individuals on 1.9 million SNPs	ELAI, PCADMIX and RFMIX	[162]

Table B.2: A partial list of estimate application studies conducted between the years 2015 and 2019.

Study description	**Study population (samples and markers)**	**Tool/(s) used**	**Reference paper**
Detecting selection signals	249 individuals from 13 Latinos on 678 micro-satellite markers	STRUCTURE V2	[113]
Detecting ancestry-enriched SNPs in Latinos	347 Latinos (Columbians, Mexican, Peru and Puerto Ricans) 1 000 GP phase 3 and 1 102 HGDP reference panels including Africans, Europeans and Native Americans on 435 782 SNPs.	RFMIX	[180]
Investigating association between genetic ancestry and multiple sclerosis prevalence	African Americans: 1 081 cases and 2 611 controls, Asian Americans: 64 cases and 4 851 controls and Hispanics: 326 cases and 3 451 controls	RFMIX	[179]
Association between ancestry and Alzheimer diseases	Puerto Ricans (220 cases and 169 controls) and African Americans (1 766 cases and 3 730 controls)	RFMIX	[181]
Effects of ancestry on physical appearance	6 589 Latinos (Brazilians, Mexicans, Peru, Chileans and Columbians), 2 058 reference panels (Native Americans, Europeans, east Asians, south/east Mediterraneans and Africans) on 546 780 SNPs.	RFMIX	[175]
Identifying signatures of selection	324 individuals (cattle) on 111 836 SNPs	ELAI	[209]
Association of ancestry in glioma	832 cases and 675 controls (African Americans and Hispanics)	RFMIX	[269]
GWAS on sickle cell diseases	African Americans including 144 cases and 123 controls on 1 471 970 SNPs	ELAI	[270]
Admixture mapping of Subclinical Atherosclerosis	African Americans: 1 554 MESA samples on 611 449 SNPs and 3 000 ARIC samples genotyped at 579 847 SNPs	LAMP-LD	[271]
Understanding population demographics and history	4 trios each from a mixture of: eastern and western Mediterranean, and Mediterranean and Atlantic sea bass on 2 628 725 SNPs	CHROMOPAINTER	[272]
Understanding origins and demographic history	2 130 Noir Maroon (eastern Brazilian and Colombian) individuals on 1 782 673 SNPs	RFMIX	[249]
Understanding population history and demographics	700 Malagasy individuals on 2 268 323 SNPs	PCADMIX	[273]

Table B.3: Existing ancestry deconvolution tools.

SOFTWARE	Multi way	Account LD	LD model	Biological/statistical parameters	Reference populations	Admixed populations	Reference
STRUCTURE v2*	✓	✓	HMM	Markers and ancestry proportions	Unphased	Unphased	[78]
ANCESTRYMAP*	✗	✓	HMM	Physical, recombination map and ancestry proportions	Unphased	Unphased	[93]
ADMIXMAP*	✓	✓	HMM	Physical map, ancestry proportions	Unphased	Unphased	[94]
SABER	✓	✓	MHMM	Physical map/recombination distance	Phased/ Unphased	Phased/ Unphased	[157]
LAMP	✓	✗	✗	Admixture generations, LD threshold, physical map, window length, ancestry proportions and recombination rate	Unphased	Unphased	[96]
HAPAA	✓	✓	HHMM	Admixture generations and genetic divergence	Phased	Phased	[95]
SWITCH	✓	✓	MHMM	Recombination rate	Phased	Phased	[163]
GEDI-ADMIX	✓	✓	Fixed size FHMM	Admixed and ancestral SNPs	Phased	Phased	[164]
WINPOP	✓	✗	✗	Admixture generations, LD threshold, offset, ancestry proportions and recombination	Unphased	Unphased	[156]
HAPMIX	✗	✓	HHMM	Genetic map, mutation rate, admixture generations, ancestral and admixed SNPs	Phased	Phased/ Unphased	[92]
CHROMOPAINTER	✓	✓	Coancestry matrix	Genetic map	Phased	Phased	[155]
LAMP-LD	✓	✓	HHMM	Window, hidden state size and physical map	Phased	Unphased	[80]
SUPPORTMIX*	✓	✓	HMM	Admixture generations, genetic map and window size	Phased	Phased	[125]
PCADMIX*	✓	✓	Window blocks of SNPs	Genetic map and window size	Phased	Phased	[166]
mSPECTRUM	✓	✓	HMM-HDP	Mutation and recombination rate	Phased	Phased	[165]
MULTIMIX	✓	✓	MVN	Genetic map, misfit probabilities, window size and legend	Phased/ Unphased	Phased/ Unphased	[152]
ALLOY	✓	✓	Non-inhomogeneous VLMC	Admixture generations, markers, ancestry proportions and genetic map	Phased	Phased	[158]
EILA	✓	✗	✗	Physical map	Unphased	Unphased	[159]
SEQMIX*	✓	✓	✓	Genetic map	Unphased	Unphased	[167]
RFMIX	✓	✗	✗	Admixture generations, window size, genetic map and number of threads	Phased	Phased	[129]
ELAI	✓	✓	Two layer HMM	Admixture generations, upper and lower cluster	Phased/ Unphased	Phased/ Unphased	[97]
LOTER	✓	✗	✗	-	Phased	Phased	[56]

Appendix C

AdmixSim 2 simulator and parameters

Table C.1: Ancestry proportion and population sizes for a hybrid-isolation model given three-way admixture

*1-6	CEU,CHB,YRI,CEUCHBYRI
100	0.25,0.65,0.1,0
200	0,0,0,1
250	0,0,0,1
500	0,0,0,1
500	0,0,0,1
500	0,0,0,1

Table C.2: Ancestry proportion and population sizes for a gradual admixture model given three-way admixture

*1-6	CEU,CHB,YRI,CEUCHBYRI
100	0.25,0.65,0.1,0
150	0.1,0.35,0.05,0.5
200	0.1,0.35,0.05,0.5
250	0.1,0.35,0.05,0.5
500	0.1,0.35,0.05,0.5
500	0.1,0.35,0.05,0.5

Table C.3: Ancestry proportion and population sizes for a cgf given three-way admixture

*1-6	CEU,CHB,YRI,CEUCHBYRI
100	0.25,0.65,0.1,0
150	0.2,0.3,0,0.5
200	0.2,0.3,0,0.5
250	0.2,0.3,0,0.5
500	0.2,0.3,0,0.5
500	0.2,0.3,0,0.5