Chapter 1

Introduction

1.1 Chromopainter and ancient DNA

In this introduction I will outline the following: i) What are 'haplotype-based' methods and what advantages and disadvantages do they offer over 'unlinked' methods, ii) a summary of different methods used to analyse ancient DNA and iii) the need to merge datasets genotyped on different arrays and options for imputation.

1.1.1 Gains to be made with haplotype information

1.1.1.1 History

Haplotype-based methods are statistical approaches in genetic analysis which explicitly model Linkage Disequilibrium (LD), or correlation in frequency, between neighbouring genetic markers along a haplotype [1]. This is in contrast to 'unlinked' methods, which assume a model of Linkage Equilibrium between SNPs. In this scenario, a 'haplotype' is a contiguous sequence of alleles which are located on the same chromosome. In this thesis, I will concentrate on haplotype-based methods in the context of identifying shared haplotypes between individuals in order to understand the population structure and history of a population(s).

Linkage Disequilibrium (LD), which is the key concept underpinning haplotype-based approaches, is the non-independence of alleles carried at different positions in the genome. It has been studied since the earliest days of genetics [2, 3] and has since been a fundamental aspect of virtually all areas of genetics. The primary advantage of accounting for LD in a

[1] Note that other methods, for example `octopus` are referred to as 'haplotype-based' genotype callers, but they represent a distinct group of methods to e.g. ChromoPainter.

model is that information about the frequency of an allele in a population also provides 'free' information about the frequency of neighbouring alleles.

Some of the earliest uses of LD information for the study of population history came in the form of microsatellite markers, whose linked tandem repeats can be thought of as analogous to linked alleles on a haplotype. Microsatellites were, and still are, commonly applied to study the population structure of wild animal systems; for instance, Amos et al (1993) used microsatellites markers to examine the population structure of whales [4]. Later, microsatellites at the CD4 locus were leveraged to show the preferred model of Human population history was a recent African origin [5]. This was deduced as Sub-Saharan Africans had substantially more variability in their frequencies of haplotypes and a higher diversity of STRP alleles associated with the Alu deletion than non-Africans, strongly suggesting Africa was the common origin of these haplotypes. This study outlined the insights into population history that can be obtained from the analysis of a very small number of linked markers.

The next major advance was the development of methods to use LD information between SNP markers rather than within microsatellites, as SNPs are substantially more numerous across the human genome. Studies in the early 2000s utilised the then-new Hap-Map results [6] to show LD varies markedly across worldwide populations, and that such variation can be used to make inferences about human colonisation history [7]. Conrad et al calculated the proportion of unique haplotypes that were shared between two geographic regions, and by showing that the number of distinct haplotypes per region declines from Africa, provided additional evidence to support the previously proposed recent African origin of humanity [8]. It was also shown that isolated Native American populations had approximately 3 times fewer haplotypes per genomic region, implicating recent endogamy plays a large role in shaping patterns of haplotype variation.

The 2000s saw a rapid increase in the number of SNP markers and individuals which had been sequenced. Accounting for LD and recombination within a model is necessarily computationally complex, as the number of combinations of alleles and their possible evolutionary histories balloons as the number of loci considered increases (does it scale quadratically?). Therefore, the new era of sequencing demanded new and more efficient methods to cope with such data. The development of the Li and Stephens copying model (LSM) [9] was instrumental in the development of such methods [10] and provided an elegant solution to the increased complexity modelling recombination between linked loci. As such, it is now a critical model in virtually all areas of genomic methodology, such as gene conversion parameters, admixed populations, human colonization history, local ancestry in admixed populations and imputation. The LSM was, and still is, the foundation for methods of

the haplotype phasing methods needed for haplotype-based methods [11, 12]. Briefly, the LSM provides a way to generate the next haplotype in a sample, conditional upon a set of previously sampled haplotypes. Importantly, the LSM allows recombination rates to vary over small genetic distances, and thus allows for the estimation of recombination rates over fine-scales from genomic data.

The first paper to formalise a haplotype-based approach for the study of population history was that of Hellenthal et al 2008 [13]. The ancestry model of Hellenthal et al is based upon the LSM, using a Hidden Markov Model to reconstruct each target haploid as a mosaic of *donor* haplotypes. The conditional probability that a given haploid 'copies' from a particular reference haplotype is obtained by observing whether the alleles at the same position match between haplotypes. One drawback of the LSM was that the ordering of the haplotypes sampled influenced the likelihood underlying the recombination rate estimate; for example, if a subset of haplotypes in the sample happen to show higher levels of diversity, then the computed likelihood will generally be higher if these individuals are generated early rather than late in the ordering. By assuming the individuals within each population share the same demographic history, Hellenthal et al were able to circumvent this issue and provide more reliable (?) results. Similar to the results of Conrad et al (2006), Hellenthal et al's analysis of the structure of global haplotype sharing provided strong evidence of a recent African origin of modern humans.

In the same year, Jakobsson et al (2008) analysed a much larger number of SNPs (n=525,910) [14]. It was demonstrated that haplotype clusters show an elevated ability to determine local structure than unlinked SNPs alone; 51.87% of haplotype clusters were found in at most two regions, in contrast with 4.66% of SNP alleles. This seems naturally intuitive, as haplotype clusters are formed from different combination of SNP alleles, which are necessarily more unique than single SNPs (know what I want to say, but not sure if this is the right way of saying it).

Building on the copying model proposed by Hellenthal et al (2008), Lawson et al (2015) [15] created ChromoPainter, again based the LSM. ChromoPainter is a more general model than that of Hellenthal 2008; whereas the Hellenthal 2008 model was explicitly formulated to determine the ordering oh human colonisation, ChromoPainter generates a co-ancestry matrix, which gives information on the level of recent shared ancestry between each donor and recipient individual. ChromoPainter also allowed for variable recombination rates between neighbouring SNPs. ChromoPainter was shown to have an enhanced ability to separate closely related populations when plotted on a PCA compared to unlinked methods. It was originally developed in tandem with its own clustering method fineSTRUCTURE, and

has since been extended into methods to detect and date admixture [16], and infer ancestry proportions [17].

The 'next-generation' of chromosome painting methods had to confront the same issue that Li and Stephens; how to adapt methodology to larger and larger sample sizes; ChromoPainter was designed with datasets of <10,000 people in mind, whereas biobank-scale datasets typically contain 500,000+ individuals. As such, ChromoPainter does not scale well to large datasets, especially when there are a large number of donor haplotypes.

One approach is to use the Burrows-Wheeler transform (PBWT) [18, 19], which is a method to efficiently find matching haplotypes in large datasets. The insight to apply the PBWT to genetic data has been one of the most crucial insights into computation biology, as it allows for substantial increases in efficiency across a wide range of applications such as sequence alignment [20], phasing [21] and data compression [22]. PBWT has been applied to Chromosome Painting on Biobank-scale datasets in several recent papers [23, 24]. Similarly, methods to detect IBD in Biobank-scale cohorts have leveraged the PBWT [25, 26]. However, PBWT-based approaches are still relatively immature; for example, they do not allow for the use of a reference panel and all haplotypes must be compared to all other haplotypes in an 'all-v-all' manner (further explanation given in Appendix section B.1). Despite their current limitations, it seems that the future of Chromosome Painting will at least in part be based on the PBWT.

1.1.1.2 Advantages of accounting for haplotypes

ChromoPainter can be run in either 'linked' or 'unlinked' mode. In the linked mode, described in detail in later sections, LD between neighbouring SNPs is accounted for. 'Unlinked' mode assumes a model of linkage equilibrium between markers and has been shown to be statistically identical to the likelihood model underlying the commonly used ADMIXTURE algorithm [15].

A typical case study, and one which I will return to in later chapters, was a study which attempted to identify population structure among individuals from the British Isles. This study, hereafter referred to as POBI, genotyped 2039 people from England, Wales and Scotland [27]. In summary, it was possible to detect structure down to the level of Devon and Cornwall (two neighbouring counties) using ChromoPainter. On the other hand, little structure was apparent when using unlinked methods (PCA). This outlines the benefits of incorporating linkage information when attempting to identify fine-scale structure between closely related populations.

Gattepaille and Jakobson (2012) [28] provided the mathematical foundations for the

advantage of using linked markers over unlinked ones. They describe a metric, GIA (gain of informativeness for assignment), a term borrowed from information theory to describe the additional amount of information gained when using haplotype data instead of individuals alleles separately. They showed that whilst combining two markers is not necessarily advantageous for ancestry inference, GIA is often positive for markers in LD with one another, demonstrating the advantage of haplotypes. Under a variety of simulated scenarios, incorrect assignment of individuals into populations was reduced between 26% and 97% when using haplotype data. They showed that using empirical data of individuals from France and Germany, accounting for haplotypes could reduce the rate of mis-assignment by 73%.

One less considered advantage of using haplotype information is that it may mitigate ascertainment bias. Ascertainment bias occurs when a subset of SNPs are chosen for analysis. SNPs are typically chosen because they display variation. If this variation is determined in one population, say British, then there is no guarantee that the variation will also be seen in another population, say Han Chinese. Therefore, including these SNPs can often provide misleading estimates of genetic diversity and commonly estimated parameters such as F_{st} [29]. Conrad et al (2006) showed that, owing to the lack of African individuals used in the SNP discovery process, populations from the Middle East, Europe and South Asia showed the highest levels of heterozygosity. These findings were in stark disagreement with the currently accepted model of human history and studies which demonstrated Africans have the highest levels of genetic diversity [30–32]. However, when instead of SNP heterozygosity, haplotype heterozygosity is used as a metric for diversity, African populations consistently have the highest values. The reason for this is, although the ascertainment for a particular SNP may depend strongly upon the ascertainment scheme, the same underlying haplotypes are likely to be observed, regardless of which SNPs are used to tag them. Thus, ascertainment is less likely to ascertain

In a similar manner, another advantage of using haplotype-based methods is that rare alleles are not required. Rare alleles are highly informative about recent, fine-scale population structure, as they are shared by the fewest number of individuals (max n=2) within a dataset. Methods which leverage this information have been used to model the population history of large datasets [33–35]. However, rare alleles are harder to genotype, as they are more difficult to distinguish from sequencing errors, and they are often not included on standard genotyping arrays. Because of this, allele-frequency filters are often applied in population genetic studies to reduce the risk of incorporating incorrectly genotyped SNPs. Further, more SNPs need to be sequenced in order to find rare variants in a wide range of populations. Using haplotype information negates the needs for using rare variants; if individuals share long haplotypes in common, then by default will also share rare variants that occur on those

haplotypes.

However, the usage of haplotype-based methods is not without drawbacks. They are typically slower by an of magnitude, as the computational complexity is [what?]. Secondly, the nature of haplotype-based methods means they require the data to be phased. Phasing is a statistical procedure [2] that requires substantial computation resources. The inconvenience of introducing an additional time and resource intensive step to the analysis means that many studies opt to not perform ChromoPainter analyses.

Lastly, 'switch-errors' may often occur, when the incorrect ordering of alleles on a haplotype is inferred. Whilst Lawson and Falush (2012) showed that sporadic, randomly distributed switch-errors are unlikely to significantly affect the overall ChromoPainter analysis, systemic errors, where haplotypes from particular individuals are made to look more like each other than they do those of other members of the sample, may be more problematic [36].

1.2 Methods used to analyse ancient DNA

Here, I will outline some of the most widely used methods to analyse ancient DNA.

1.2.1 Unlinked methods

The first studies into ancient DNA mostly used statistical methods which compare allele-sharing or allele-frequencies between populations or individuals. These methods, in particular f-statistics and their extensions [37–40] and Principle Component Analysis [41], can address a wide-range of questions pertaining to population structure, admixture, shared drift and population graphs.

A key reason why methods based on allele-sharing and allele-frequency differences were, and still are widely used in ancient DNA is that they can easily be modified to use data in pseudo-haploid format. Pseudo-haploid genotypes are generated by sampling a read at random to represent a single allele at a given SNP. This is often necessary, because ancient samples routinely do not enough reads covering a SNP to confidently call heterozygous genotypes. Pseudo-haploid calls are therefore used widely, including currently (e.g. [42]), in most studies of ancient humans.

Whilst pseudo-haploid genotype calls circumvent the problem of calling heterozygous genotypes at low coverage positions, they necessarily holds less information relative to true diploid genotypes and are thus less powerful at e.g. identifying population structure or

[2] Phasing can also be performed using other methods, such as sequencing family trios. However, this is rarely used in population genetic studies and so I will not discuss it here

genetic similarity. Further, the use of pseudo-haploid calls may result in an elevated level of reference bias [43–45]. Reference bias occurs because the reference fasta file which is used to align reads only contains a single allele at each position. Therefore, reads which contain a non-reference allele (i.e. an allele not represented in the reference fasta) contain more mismatches with the reference than reads which contain the reference allele, and accordingly are given a lower mapping quality score. Accordingly, when selecting a read at random, reads with the reference allele are more likely to be selected as the pseudo-haploid call, generating a bias towards the reference allele. Attempts are being made to represent non-linear reference genomes as graphs in order to mitigate the effect of reference bias [45, 46].

For many of the early ancient DNA studies, such as that of Green et al 2010 [37] and Lazaridis et al 2014 [47], powerful methods for detecting population substructure and admixture were not required, as the questions asked primarily considered broad questions about human history, such as the nature of human-archaic interactions and whether there was significant genetic differences between the first farmers and the preceding hunter-gatherers. These populations, particularly humans and Neanderthals, are highly diverged and hence do not require powerful methods. For example, in the case of Lazaridis et al (2014), simply plotting Loschbour and Stuttgart on a PCA of modern individual showed they had substantially different ancestries.

Perhaps the most widely used method amenable to pseudo-haploid data is the family of F-statistics [3], which were first outlined in a 2009 study into the population history of India [49]. These methods use the principle of shared drift in order to estimate genetic similarity (f_2), branch-length and admixture (f_3) and tests of treeness (f_4). Since 2009, F-statistics have been extended into multiple, more advanced, frameworks which are able to answer more complex questions about population history through the generation of population admixture graphs. In particular, qpAdm has been shown to be a flexible and coverage-robust method of estimating individual and population level admixture fractions [40]. An attractive feature of F-statistics is that they explicitly test models of population history and can provide readily interpretable results with associated jacknifed confidence intervals. A related method is the so-called ABBA-BABA test, developed by Green et al (2010) [37] in order to determine whether, and to what extent, admixture between humans and the newly sequenced Neanderthal genome had occurred. This simple test counts the number of times across the genome a 4 population phylogenetic tree shows a particular configuration at a given locus.

One possible issue of f-statistics is that of drifted populations; f_3 tends to pick out drifted

[3] Although related, they should not to be confused with Sewall Wright's F-statistics [48].

populations.

In contrast to the F-statistics, which explicitly tests models of population relationships, Principle Component Analysis (PCA) is a 'model-free' method typically used to obtain a visual summary of the genetic ancestry of the sample being analysed. PCA is commonly used as it is typically very fast and is easily interpretable. Several methods have been developed which adapt the standard PCA approach (e.g. eigenstrat [41]) to low coverage ancient DNA [50–52]. I note that PCA may also be performed on matrices obtained from linked analysis, such as a matrix of pairwise IBD sharing or ChromoPainter coancestry matrix.

1.2.2 ChromoPainter ancient DNA

1.2.2.1 History

In recent years, the 'low hanging fruit' of broad-scale questions have mostly been answered and studies into more fine-scale populations structures have become more prevalent. Accordingly, methods which can detect more subtle population structure have been required. However, the incorporation of ChromoPainter analysis into studies of ancient DNA was slow, in part because of the difficult of phasing low-coverage genomes and concerns over introducing bias towards present-day populations during imputation.

ChromoPainter can be used to answer a variety of questions relating to the genetic variation and population history of groups of samples. It can provide an overview of genetic ancestry through Principle Component Analysis of the co-ancestry matrix. Differential haplotype donation to different worldwide populations, as shown in Fig 1.1, can reveal geographic correlates of genetic variation.

The first use of ChromoPainter on ancient DNA was in the seminal paper of Lazaridis et al (2014) [47]. Through the generation of two high-coverage ancient genomes, they were the first to pos that most present-day Europeans can be modelled as a mixture of three ancestral populations. For the ChromoPainter analysis, they did not impute missing genotypes in the ancient samples, as the possible bias effects had yet to be studied; only positions with non-missing genotypes were retained (as the samples were of high coverage, this was not an issue, as 495,357 SNPs were kept). The the ability of fineSTRUCTURE to meaningfully cluster ancient individuals was confirmed by recapitulating previous results that identified different present-day European populations as being more closely related to Early Farmers and hunter-gatherers than others.

In-between 2014 and the present-day, there have been approximately 31 studies which

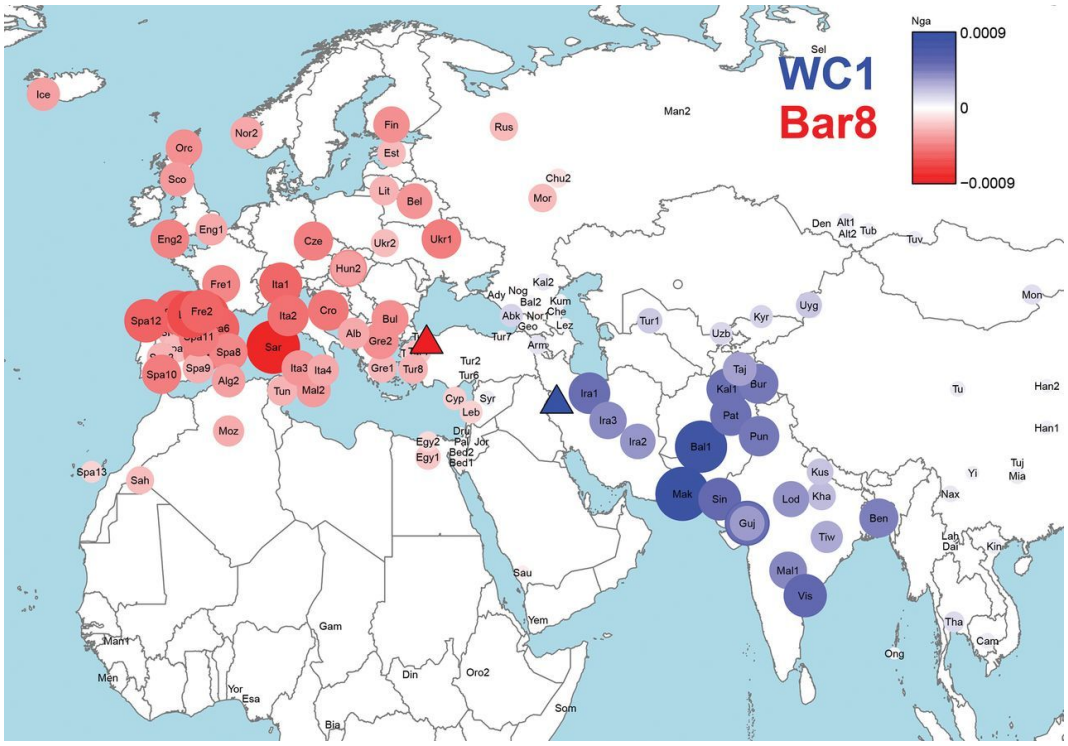

Figure 1.1: Map of differential haplotype sharing with present-day populations between WC1 (Iranian Farmer) and Bar8 (Anatolian Farmer) from Broushaki et al (2016) [1]. Bar8 copies relatively more from red populations and WC1 from blue populations.

have used ChromoPainter on ancient samples (based on Web of Science search results).

As of writing (September 2021), the study of Margaryan et al (2020) is the biggest so far to use ChromoPainter, with over 400 samples used [53]. This study concluded that detecting structure within the dataset using 'traditional' methods was not possible and so opted to use haplotype-based analyses on all samples above 0.1x mean depth. Another recent large study into the genomic history of the Roman Empire and surrounding regions leveraged ChromoPainter [54].

More recently, ChromoPainter has been used to study aspects of archaic hominin ancestry in present-day humans [55,56]. Whilst ChromoPainter is not specifically designed to accurately estimate local ancestry, it is possible to infer identify potentially introgressed Denisovan regions of DNA by determining whether a haplotype which is more similar to the Denisovan genome than to a panel of sub-Saharan Africans. ChromoPainter has also been extended to studying the ancient DNA of non-human organisms, such as plants, bacteria [57].

1.2.2.2 Benchmarking ChromoPainter and imputation

Many studies which have used ChromoPainter on ancient samples have performed tests and benchmarks to various degrees of detail.

An early study to explicitly investigate the reliability of ChromoPainter on ancient DNA was Martiniano et al (2017) [44]. This study explored various aspects of ChromoPainter analysis on ancient samples. Testing whether including imputed genotypes introduced bias towards particular present-day populations was key, as if it were the case, it would potentially invalidate all results obtained from using ChromoPainter on ancient samples. Potential bias was estimated by plotting normal quantile-quantile plots of the copyvectors obtained from imputed and non-imputed markers. Whilst the differences in amount of copying differed by up to 14%, most percentage differences were substantially lower and there was no evidence of structured bias towards or against particular geographic regions, with the authors concluding "There is no strong evidence for systematic changes being caused by genotype imputation.".

The same study also investigated the impact of filtering genotypes based on genotype probabilities by creating two datasets, one containing hard filtered genotypes and one not, and performing fineSTRUCTURE clustering. They inferred 7 more clusters when using filtered genotypes. Whilst this could perhaps be an indication of improved performance, it is hard to draw solid conclusions from these data. The overall number of fineSTRUCTURE clusters can not be seen as a direct measurement of performance; for example, the additional clusters inferred may simply be a result of the stochastic nature of MCMC sampling, and given only a single replicate of each test was performed, it is not possible to rule this out. Performing the same analysis on simulated data, where the population labels of individuals are known in advance, would be a more controlled test.

Since the study of Martiniano et al, many papers which incorporated ChromoPainter analysis into studies of ancient DNA have included their own set of benchmarks. Antonio et al (2019) [54] analysed 127 ancient genomes of a mean coverage of 1x and tested imputation accuracy on a single individual (NE1) downsampled to different levels of coverage. However, this analysis was only performed on a single sample and the effect of imputation on the ChromoPainter process was not evaluated. Margaryan et al (2020) performed a downsampling test on two high coverage genomes down to 1x mean coverage and concluded that, whilst there was some suggestion that the 1x downsample tended to a more mixed ancestry profile, there was no evidence that incorrect ancestries have been inferred or that major changes in ancestries have occurred.

Imputation is a necessary pre-processing step for ChromoPainter analysis on low-medium

coverage ancient DNA samples for two primary reasons. Firstly, ChromoPainter does not allow for missing genotypes and so imputation is required to estimate missing genotypes. Secondly, whilst they are covered by reads, non-missing positions may still be low in coverage and thus require to be re-estimated, particularly when the true genotype is heterozygous. Therefore, it is important to determine to what extent it is possible to accurately impute genotypes at different levels of mean coverage.

The accuracy of imputation has been tested in various studies. There is difficulty in comparing the estimated accuracies between studies, however, due to differences in factors such as samples analyses, software used to call genotypes and impute samples, the regions analysed and filters applied. However, all investigations have reported a 'high' concordance between imputed and non-imputed genotypes.

check this sentence

The most systematic and thorough evaluation of imputation in ancient genomes was performed by Hui et al (2020) [58]. This study noted that it is possible to impute using a one or two step approach and, through the use of downsampled genomes, showed that the two-step approach provides more accurate imputed genotypes. This study also showed that whilst most genotype likelihood callers (e.g. GATK, atlas) performed similarly well, atlas was preferred because of it's ability to model post-mortem damage (PMD) in ancient samples. Accordingly, I will use atlas to call genotype likelihoods in the rest of my thesis.

It should be noted that the study only considered a single ancient genome (NE1) and it is therefore unclear how generalisable these results are when applied to samples with ancestries more or less prevalent in a reference panel. In particular, the results may not be applicable to ancient samples from Africa, which would likely harbour more diversity, much of which would be unlikely to be present in any reference panels. However, this study provided important benchmarks for many critical steps in the analysis of low coverage samples which had previously been missing from the literature, such as selection of a reference panel, the feasibility of local imputation and the application of pre and post imputation filters.

1.3 Issues and solution to low coverage data

Coverage and lack of, is an issue which has plagued the field of ancient DNA since its inception. Compared to DNA obtained from present-day samples, ancient DNA samples typically have a much lower proportion of endogenous DNA. This is because DNA degrades over time from environmental factors. Therefore, when the DNA fragments are sequenced, relatively few of them will align to the human reference. The coverage of a genome is therefore the mean number of reads mapped to each position in the genome.

The primary issue with low-coverage data is the increased uncertainty when calling diploid genotypes, particularly when the true genotype is heterozygous. Several methodological adaptations have been applied to existing methods in order to adapt them to low coverage ancient DNA. These approaches primarily attempt to circumvent making diploid genotype calls; for example, the previously mentioned strategy of pseudo-haploid genotype calling.

Alternatively, methods may avoid making diploid calls by working on genotype likelihoods. Genotype likelihoods represent a posterior estimate of the confidence of the 3 different genotypes at a bi-allelic locus, and thus allow the method to appropriately propogate that certainty throughout the analysis. A wide array of complex statistical approaches have been developed in order to accurately estimate the posterior genotype likelihoods. These approaches integrate factors such as sequencing-machine reported base-quality scores and estimates of read-mapping / sequencing errors [59]. Common methods to estimate likelihoods include the GATK model [60], SAMtools [61], SOAPsnp [62] and SYK model [63]. Genotype likelihoods can either be estimated prior to the analysis from aligned reads (BAM files), using software such as ANGSD [64], ATLAS [65] or GATK [60]. Other softwares will take BAM files directly as input and estimate genotype likelihoods during the analysis process (e.g. STITCH [66]).

Once genotype likelihoods have been estimated, population level parameters such as inbreeding coefficients and F_{st} can be estimated directly [64] with greater accuracy than direct genotype calls. Similarly, modifications of the ADMIXTURE [67] algorithm and PCA have been developed in order to analyse low coverage samples more effectively [68,69]. Recent advances have allowed the identification of 1st and 2nd-degree relatives from as low as 0.02x coverage samples [70, 71].

Several methods exist which jointly estimating ancient DNA specific confounding factors, such as contamination and post-mortem damage, alongside the demographic parameter of interest [72]. Schraiber (2018) [73] developed a novel maximum-likelihood approach which leverages information from different low-coverage samples from within the same population to infer population-level parameters, such as genetic continuity between ancient and modern populations.

Viera et al (2016) developed a method (ngsF-HMM) to infer matching identical-by-descent (IBD) segments from low-coverage data [74]. This program is mostly designed for demographic inference in the context of conservation genetics - for example, estimating the relation of inbreeding to fitness decline. To account for the uncertainty, all 3 genotype likelihoods are integrated over in order to estimate whether or not a genomic region is IBD

given the likelihoods. This method showed that there is a substantial gain in power when likelihoods are used compared to genotype calls. Whilst similar to ChromoPainter in terms of modelling SNPs as linked markers, ngsF-HMM differs in that it estimates pairwise IBD segments rather than comparing each haplotype to all other haplotypes. Whilst the use of pairwise IBD segments in clustering analysis has been shown to be more powerful than using unlinked methods, ChromoPainter provides more power than IBD sharing [36].

1.4 Combining data from multiple chips

A related issue stems from the current practice of developing a large number of genotyping arrays. Different cohorts are genotyped on different arrays and sets of SNPs, as different SNPs have different characteristics. For example, some SNPs are known to be associated with particular phenotypes, some SNPs are known to be more variable (and therefore more informative at identifying structure) in certain populations. Whilst this generation of custom genotyping arrays has meant a wider variety of questions and populations can be studied using genoyping arrays, it also makes combining data from across different arrays potentially troublesome, as they often have a small overlap in the SNPs upon which they have been genotyped.

For example, in my thesis, I have worked with at least 3 genotyping arrays; 'Human Origins', 'Hell Bus' and the UK Biobank. Often I have wanted to compare populations on different arrays, such as the African populations on the Human Origins array and UK Biobank individuals on the UK Biobank array. After merging the datasets, the overlap was small, only 70,000 SNPs. This is around an order of magnitude fewer SNPs than a typical ChromoPainter analysis.

Having a smaller number of SNPs may reduce power in two ways. Firstly, there is simply fewer informative data points to use when comparing the SNP patterns between two populations and therefore fewer possible data points which can be used to identify populations. Secondly, ChromoPainter derives parts of its power from the LD between neighbouring SNPs. LD between two neighbouring SNPs is correlated with their physical distance. Fewer overall SNPs means each neighbouring pair of SNPs are physically further away from one another and thus have less LD information.

One solution to the issue of a small number of SNP would be to impute the remaining SNPs. In this context, imputation refers to estimating missing genotypes using of a model usually based upon the LSM and a large reference panel. Imputation is widely used in e.g. GWAS to generate sequence-level data.

However, it is possible that imputation may cause a bias in the data. If missing genotypes are imputed incorrectly more often from one population than another, this will result in an increased, but spurious genetic similarity between the target and reference population. This may be a particular issue when analysing populations which are not well represented in imputation reference panels, such as non-Europeans. The nature and magnitude of this bias, however, is yet to be fully understood, particularly in the context of ChromoPainter.

Therefore, one question to ask is the following; is it more desirable to impute the missing positions or to use use a smaller number of overlapping SNPs. This is something which I will investigate in chapter 3 with a case study investigating African ancestry in the UK Biobank dataset.