Method Development for Next-Generation Sequencing Data in Population Genetics

A PhD thesis by Jonas Meisner
Submitted December 2020

Supervisors:
Main supervisor: Professor Anders Albrechtsen

Affiliations:
Section for Computational and RNA Biology, Department of Biology, Faculty of Science,
University of Copenhagen, Denmark

This thesis has been submitted to the PhD School of Science, University of Copenhagen,
Denmark
Preface

This thesis has been submitted with the purpose of obtaining a PhD degree at the Faculty of Science, University of Copenhagen. The PhD was started in January 2018 and ended in December 2020 and was supervised by professor Anders Albrechtsen. It was carried out in the Section for Computational and RNA Biology, Department of Biology at the University of Copenhagen, Denmark. For change of academic environment, I spent two months in the lab of Matteo Fumagalli at Imperial College London, United Kingdom, and I had a two weeks visit in the lab of Karsten Kristiansen at BGI-Shenzhen, China.
Abstract

This thesis covers work in aspects of population genetics, statistical genetics and machine learning, and it consists of several new statistical methods as well as a novel machine learning framework for usage in population genetics.

The first paper presents two new methods for low-depth next-generation sequencing data, where the first method infers population structure using principal component analysis and the other method estimates admixture proportions. The first method accommodates the uncertainty in the genotypes by working directly on genotype likelihoods in an iterative approach for estimating individual allele frequencies. The method is shown to be more accurate than existing methods for inferring population structure. The individual allele frequencies can be used for estimating admixture proportions in a matrix factorization approach that is much faster than existing methods for estimating admixture proportions.

The second paper presents a new method for inferring population structure using principal component analysis in the presence of rampant non-random missingness. The method directly models the missingness in an expectation-maximization algorithm to impute missing data. We demonstrated that the method is more accurate than competing methods for inferring population structure, since most other methods fail due to not accounting for missingness. The method is further shown to scale to very large genetic datasets in terms of computational runtime.

The third paper describes a new method that tests for and quantifies deviations from Hardy-Weinberg Equilibrium in structured populations using genotype or low-depth next-generation sequencing data. It naturally accounts for population structure by incorporating individual allele frequencies in a likelihood framework that works directly on genotype likelihoods. The method is shown to be more accurate at detecting and quantifying deviations from Hardy-Weinberg Equilibrium in structured populations than existing methods.

The fourth paper introduces a new method for inferring local haplotype structure by estimating latent encodings and clusterings of haplotypes in phased haplotype data using neural networks. It is based on a variational autoencoder model that can be used to infer population structure as well as to estimate admixture proportions in a novel likelihood framework. We demonstrate that this method is able to capture global fine-scale population structure by utilizing haplotype information, which is not performed in standard approaches.
Dansk Resumé

Denne afhandling beskriver arbejde i emner indenfor populationsgenetik, statistisk genetik og machine learning, og den indeholder flere nye statistiske metoder samt en ny machine learning metode til brug i populationsgenetik.

Det første manuskript præsenterer to nye metoder til lavdybde næste generations sekventeringsdata, hvor den første metode udleder populationsstruktur ved brug af principal component analysis og den anden metode estimerer admixture proportions. Den første metode tager højde for usikkerheden i genotyperne ved direkte at arbejde med genotype likelihoods i en iterativ tilgang for at estimere individuelle allelfrekvenser. Det vises at metoden er mere præcis end eksisterende metoder til at udlede populationsstruktur. De individuelle allelfrekvenser kan bruges til at estimere admixture proportions i en matrix faktorisering tilgang som er meget hurtigere end eksisterende metoder til at estimere admixture proportions.

Det andet manuskript præsenterer en ny metode til at udlede populationsstruktur ved brug af principal component analysis i tilstedeværelse af meget manglende data. Metoden modellerer direkte det manglende data i en expectation-maximization algoritme ved imputation. Det vises at metoden er mere præcis end andre konkurrerende metoder til at udlede populationsstruktur, hvor de fleste andre metoder fejler på grund af at de ikke tager højde for manglende data. Derudover vises det at metoden skalerer til meget store genetiske datasæt på baggrund af beregningsmæssig køretid.

Det tredje manuskript beksriver en ny metode, der tester og kvantificerer afvigelser fra Hardy-Weinberg-ligevægt i strukturerede populationer ved brug af genotyper eller lavdybde næste generations sekventeringsdata. Den tager naturligt højde for populationsstruktur ved at inkorporere individuelle allelfrekvenser i et likelihood framework, der bruges direkte på genotype likelihoods. Det vises at metoden er mere præcis til at påvise og kvantificere afvigelser fra Hardy-Weinberg-ligevægt i strukturerede populationer end eksisterende metoder.

Det fjerde manuskript introducerer en ny metode til at udlede lokal haplotypestruktur ved at estimere haplotypers latente indkodninger og klynger i phased haplotype data ved brug af neurale netværk. Den bruger en variational autoencoder model, som kan bruges til at udlede populationsstruktur samt estimere admixture proportions i et nyt likelihood framework. Det demontreres at metoden kan fange global finskala populationsstruktur ved hjælp af haplotype information, som ikke kan gøres i standard metoder.
List of publications

Publications included in the thesis


Jonas Meisner, Siyang Liu, Mingxi Huang and Anders Albrechtsen. Large-scale Inference of Population Structure in Presence of Missingness using PCA. (Second round of review in *Bioinformatics*)


Other publications since start of PhD (* = joint first author)


Other publications in preparation


Rasmus Heller, Casia Nursyifa, Genis Garcia-Erill, Jordi Salmona, Lounes Chikhi, Jonas Meisner, Thorfinn Sand Korneliussen and Anders Albrechtsen. A reference-free approach to analyze non-model RADseq data using standard Next Generation Sequencing toolkits. (Second round review in *Molecular Ecology Resources*)

I will start by thanking my supervisor Anders Albrechtsen. I think it is fair to say that we have had a lot of great experiences together during my PhD, and I can only look back at it all and smile. I know that I have always been driving you a bit crazy with my choice of programming language and machine learning ideas but I feel like I have made you see and appreciate some of their features. I really admire your open door policy that makes it feel like you always have time for a quick or a long chat no matter the topic. I’m really happy for you with all your recent growth and success, and I really appreciate all the opportunities you have created for me during this time.

I would of course like to thank the entire PopGen group with all its former and current members for creating so many great memories. Especially the PIs, Anders, Hans, Ida and Rasmus that do so much to foster academic collaborations and to create a social environment even with COVID-19 on the doorstep. Let’s see if I can remember everybody that I have crossed path with during this time: Anna, Anne, Aviaja, Casia, Casper-Emil, Christian, Cindy, Emil, Frederik, Genis, Kristian, Liam, Malthe, Mengyuan, Patricia, Rute, Ryan, Samuelle, Xiaodong, Yorgos and Zilong.

A heartfelt shout-out to my people in the ‘Island’ office. You will always have a special place in my heart and I feel that we have created something rare in a research environment. It has been amazing to have a workspace where you feel that you can work hard if you have to and have fun when you want to.

Also a special thanks to Emil for being a great PhD companion these years, especially on our travels together as well as in our in-depth discussions of the Tolkien universe or football. The photo of you posing with a sword wearing sunglasses at an Italian medieval festival will forever be a treasure in my photo library.

I also want to thank all the people at the Bioinformatics Centre for providing me with so many great and mostly drunk memories through the years.

My family also deserves a huge thank you for always supporting me and making me feel special and proud of my work, even though they have no clue of what I’m doing.

Finally, I want to thank mi amor, Alba, for always supporting me and especially through the strange times that COVID-19 brought upon us. I can’t wait to see where our future is headed.
Contents

Preface i
Abstract iii
Dansk Resumé v
List of publications vii
Acknowledgements ix
1 Introduction 1
  1.1 Background ................................................. 2
    1.1.1 Genetic variation ..................................... 2
    1.1.2 Whole-genome sequencing and SNP arrays ............. 3
    1.1.3 Inference of population structure ..................... 5
    1.1.4 Variational autoencoders ............................. 9
  1.2 Contributions and perspectives .............................. 12
    1.2.1 Paper I .................................................. 12
    1.2.2 Paper II ............................................... 14
    1.2.3 Paper III ............................................... 15
    1.2.4 Paper IV ............................................... 16
2 Paper I 25
3 Paper II 53
4 Paper III 85
5 Paper IV 105
Introduction

The main content of this thesis is four papers of which I am the first author. Two of the papers have been published, one has been submitted to a journal with the aim of publication and is currently undergoing the second round of review, while the fourth paper is available on bioRxiv. The first two papers describe methods for inferring population structure using principal component analysis. The first method accommodates the uncertainty in the genotypes from low-depth whole-genome sequencing data, while the second method iteratively models missingness of various sources. The third paper introduces a test statistic to address the problem of testing for Hardy-Weinberg equilibrium in structured populations. Lastly, the fourth paper provides a new framework for modelling local haplotype structure using neural networks from where it also is able to infer fine-scale population structure by utilizing haplotype information.

This chapter is intended to have the reader gain intuition of the concepts and methods in population genetics that my included papers touch upon. The introduction is therefore not meant to be a thorough review of these topics but rather a broad overview of the concepts as well as the terminology used throughout the thesis. In the first part of this chapter, I will introduce central concepts of genetic variation, genetic data, approaches for inferring population structure and variational autoencoders that have all been essential parts of my research. In the second part of this chapter, I will introduce my included papers with their main points and contributions as well as some examples of usage.
1.1 Background

1.1.1 Genetic variation

Humans are diploid organisms as we carry two homologous copies of our 22 autosomal chromosomes and a pair of sex chromosomes that may differ. The chromosomes are long linear genomic sequences that together comprise the human genome with the addition of a circular chromosome in the mitochondria of our cells. Even though most individuals share a large majority of the human genomic sequence, the copies of the chromosomes will have different variations that may summarize years of evolutionary history of a population or species. The simplest variation in the genome is a single-nucleotide polymorphism (SNP) which is when a mutation occurs at a site or locus in the genome. This process gives rise to two versions of the locus, which are denoted alleles. A common assumption in population genetics as well as in this thesis is that SNPs throughout the genome are diallelic, thus, two versions at each locus. At a locus, the allele occurring at the highest frequency in a population is commonly called the major allele, and the other allele, the minor allele. The minor allele frequency is therefore used to indicate the rarity of a genetic variant in a population. An individual is homozygous at a locus if the two alleles are the same, whether it be the major or minor allele, and is heterozygous if the two alleles are different such that both the major and the minor allele are present. This information is summarized by the genotype which represents the count of minor alleles at a site, taking a value of either 0, 1 or 2. By seeing the genotype as a sampling of two independent alleles in a population, the genotype of an individual at a site can be described through a Binomial model with the minor allele frequency as the parameter.

The Hardy-Weinberg equilibrium (HWE) [Hardy et al., 1908; Weinberg, 1908] describes the relation between allele frequency and genotype frequencies in a population. The genotype frequencies will follow the expected frequencies of the Binomial model in one generation of random mating and will remain constant in a population under some assumptions. These assumptions include a very large randomly mating population, the absence of natural selection, mutations and migration. The HWE is therefore a very useful null model in population genetics. Genetic variation will vary in distinct populations as allele frequencies may differ due to different population histories. The Wright-Fischer model [Wright, 1931] deals with the effect of genetic drift on the allele frequency by sampling a finite number of alleles in a randomly mating population with non-overlapping generations of constant population size. It additionally assumes that there are no other evolutionary forces acting in the process such as selection, mutation and migration. The model describes the expected allele frequency to be constant over generations, however, its variance depends on the population size and it increases with a decrease in population size. Allele frequency distributions of distinct populations can therefore evolve very differently due to stochastic fluctuations in different population sizes even without other evolutionary forces acting. Therefore, genetic drift will have a strong effect in smaller populations that would, for example,
be relevant in demographic events such as a bottleneck or a founder effect. Natural selection is another evolutionary process that changes the allele distribution in populations over generations by for example increasing the frequency of alleles that are advantageous, e.g. due to adaption in extreme environments [Yi et al., 2010], or decreasing the frequency of deleterious alleles.

Populations therefore become genetically differentiated over time due to the evolutionary forces of genetic drift, selection and mutations. Population structure is an umbrella term to describe the systematic difference in allele frequencies between populations or sub-populations due to non-random mating, which is often geographically determined. Population structure can also arise from admixture, when distinct populations have mixed such that the genomes of mixed individuals will have fractional membership from multiple ancestral sources. Chromosomal segments or haplotype blocks in the genome can originate from different ancestral sources since segments are inherited in chunks due to a finite number of recombination events. This will further create linkage disequilibrium (LD), which is the non-random correlation between alleles, and it is especially strong in alleles from the same haplotype block. Thus, population structure caused by recent admixture will lead to LD beyond the local haplotype blocks in the chromosomes. Therefore, systematic differences in allele frequencies due to population structure must be accounted for in many analyses such as relatedness [Moltke and Albrehtsen, 2014], inbreeding, HWE [Meisner and Albrehtsen, 2019] and in large-scale association studies as it could mimic the association signals and thus lead to false positives [Marchini et al., 2004].

1.1.2 Whole-genome sequencing and SNP arrays

The development of whole-genome sequencing, or next-generation sequencing, technologies have transformed genetic studies as it allows researchers to generate and study full genomes of organisms at a much faster and cheaper scale. It works by generating large amounts of sequencing reads that are short fragments of DNA, which need to either be mapped to a reference genome or used in de novo assembly [Metzker, 2010]. For the mapping process, the sequencing reads will have associated base qualities and mapping qualities that describe the confidence of each called base in the read by the sequencing machine and the confidence of the mapping by a mapping algorithm, respectively. The information of these qualities can be used in downstream analyses to account for potential errors, issues or other uncertainties that might be caused by the sequencing or mapping process. This is usually accounted for in the process of genotype calling, where the genotypes of an individual are inferred, and in the process of SNP calling, where sites with multiple alleles are detected [Nielsen et al., 2011].

There is an ever-increasing demand for larger sample sizes in modern genetic studies that has lead to the choice of relying on medium and low-depth sequencing (< 20X and < 5X per individual, respectively) as a cost-effective strategy. This is the case in large-scale sequencing projects such as the 1000 Genomes Project [1000 Genomes Project Consortium et al., 2015]. It has also been
shown that larger sample sizes sequenced at low-depth provide greater accuracy in the estimation of many parameters in population genetics in comparison to limited sample sizes sequenced at higher sequencing depths [Fumagalli, 2013]. A sequencing depth of 5X means that on average a position in genome will only be covered by 5 sequencing reads and will therefore provide less evidence and more uncertainty in the genotype calling and SNP calling of the individuals. The uncertainty arises due to the difficulty of distinguishing a SNP from a sequencing error and the problem of not having both alleles sampled in the sequencing process, which would lead to heterozygous genotypes being called as homozygous [Nielsen et al., 2011]. We have visualized a low-depth sequencing scenario in Figure 1.1.

Several methods and statistical frameworks have been developed to account for the uncertainty in the genotypes in medium and low-depth whole-genome sequencing by utilizing genotype likelihoods. Genotype likelihoods accommodate the uncertainty in the genotypes by modelling the probability of the observed sequencing data while conditioning on the possible genotypes for a given individual. The genotype likelihoods can simply be calculated with an assumption of reads being independent of each other while accounting for base qualities to retain information from the sequencing process [Korneliussen et al., 2014; Nielsen et al., 2011]. Paper I and III both introduce methods that deal with the uncertainty in the genotypes for low-depth whole-genome sequencing data on the basis of genotype likelihoods.

A cheaper alternative to high depth whole-genome sequencing is using genotypes from SNP arrays, where a given set of pre-ascertained SNPs in the genome is genotyped. However, this leads to ascertainment bias as estimates from SNP arrays can be different from estimates of whole-genome sequencing data due to the non-uniform ascertainment [Lachance and Tishkoff, 2013]. Genotyping

---

**Figure 1.1:** The blue blocks represent sequencing reads of an individual mapped to a reference genome, while red and yellow blocks represent genetic variants in the reads with high and low base qualities, respectively. The position with the red mutation blocks has more evidence of being a true heterozygous than the position with yellow mutation block, which could be a sequencing error.
1.1. Background

errors frequently occur due to various factors such as allelic dropout [Pompanon et al., 2005]. However, SNP arrays are popular in genome-wide association studies due to the demand of very large samples sizes and the possibility of haplotype imputation to inexpensively increase statistical power by increasing the number of SNPs being available for testing [Li et al., 2009].

1.1.3 Inference of population structure

The most common approaches for inferring population structure are based on either principal component analysis (PCA) or clustering algorithms. In both cases, it is usually assumed that SNPs or variable sites are independent thus ignoring linkage-disequilibrium (LD) or LD is reduced by pruning the number of SNPs. PCA projects individuals onto inferred axes of genetic variation that capture population structure, while the clustering algorithms assign individuals with fractional membership to a predefined number of ancestral populations. However as shown in Engelhardt and Stephens [Engelhardt and Stephens, 2010], the two approaches are connected when viewed as matrix factorizations with different sets of constraints. For example, each of the $K - 1$ top principal components of PCA will model a linear combination of the admixture proportions of $K$ ancestral populations. We have illustrated the connection in Figure 1.2.

![Figure 1.2: Connection between inferred population structure using PCA and admixture proportions. Analyses has been performed in the CEU, CHB and YRI populations of the 1000 Genomes Project [1000 Genomes Project Consortium et al., 2015]. In the first row, the first plot simply represents an intercept of 0 due to standardization of the genotype matrix while the subsequent plots display the first and second principal components, respectively. The second row displays the estimated admixture proportions using $K = 3$. Individuals are colored by their population label with blue, red and yellow for CEU, CHB and YRI, respectively.](image)

PCA [Pearson, 1901] is a dimensionality reduction technique to summarize the variables of the data with a set of orthogonal principal components that are linear combinations of the original variables, which maximize the variance in the data. The principal components can, in our case, be seen as axes of genetic variation as they will capture population structure or cryptic structure when individuals are projected onto them. One of its first applications in genetics was by Menozzi et al. (1978) [Menozzi et al., 1978] for gene frequencies in Europeans. However, PCA
was popularized in population genetics for modern genetic datasets with the EIGENSOFT software [Price et al., 2006] and the theoretical work in Patterson et al. (2006) [Patterson et al., 2006].

We define a genotype matrix $\mathbf{G}$ with $N$ samples and $M$ diallelic SNPs or variable sites. An entry in the matrix can take values of either 0, 1 or 2, specifying the number of minor alleles observed for a given individual at a given site. The covariance between individual $i$ and $j$ is then computed as follow under the assumption of a Binomial model

$$\text{cov}(i, j) = \frac{1}{M} \sum_{m=1}^{M} \frac{(g_{im} - 2f_{m})(g_{jm} - 2f_{m})}{2f_{m}(1 - f_{m})},$$  

(1.1)

where $g_{im}$ is the genotype of individual $i$ at site $m$ and $f_{m}$ is the minor allele frequency at site $m$.

The full $N \times N$ covariance matrix $\mathbf{C}$ is more efficiently estimated using matrix multiplications, $\mathbf{C} = \frac{1}{M} \mathbf{X} \mathbf{X}^T$ with $\mathbf{X}$ being the standardized genotype matrix. To obtain the eigenvectors $\mathbf{U}$, which are the individuals projected onto the inferred axes of genetic variation, eigendecomposition is performed on the covariance matrix, $\mathbf{C} = \mathbf{U} \Sigma \mathbf{U}^T$. $\Sigma$ is here the diagonal matrix of eigenvalues, which describes how much of the variance each of the inferred axes of genetic variation explain in the dataset. Another approach for obtaining the eigenvectors is directly through singular value decomposition (SVD) on the standardized genotype matrix such that covariance matrix does not have to be estimated

$$\mathbf{X} = \mathbf{U} \Sigma \mathbf{V}^T. \tag{1.2}$$

Here the $\mathbf{U}$ matrix is the same eigenvectors obtained through eigendecomposition of the covariance matrix, $\mathbf{S}$ is the diagonal matrix of the singular values and the $\mathbf{V}$ matrix is the eigenvectors of the covariance matrix of the SNPs instead of individuals. The $\mathbf{V}$ matrix explains how much each SNP contributes to the axes of genetic variation, while the singular values are related to the eigenvalues with a simple transformation. Due to its lower dimension, it has traditionally been much faster to perform PCA on the covariance matrix. However, with advances in large-scale eigenvector problems and random matrix theory [Lehoucq et al., 1998, Halko et al., 2011], it is possible to only infer the top first $K$ eigenvectors based on a low-rank approximation of $\mathbf{X}$ commonly named truncated SVD. This has allowed for the development of methods that scale to very large genetic datasets of biobank magnitude due to its much lower computational complexity [Galinsky et al., 2016, Abraham et al., 2017, Meisner et al., 2020]. An example of a PCA plot is visualized in Figure 1.3.

The clustering algorithm introduced in STRUCTURE [Pritchard et al., 2000] clusters individuals based on a probabilistic model where each individual allele is assigned to a cluster and the posterior probabilities are approximated using a Bayesian Markov chain Monte Carlo (MCMC) framework. A key assumption in this algorithm is that the number of populations is assumed
1.1. Background

Figure 1.3: PCA performed on the CEU, CHB and YRI populations of the 1000 Genomes Project [1000 Genomes Project Consortium et al., 2015]. Left plot displays the inferred population structure using PCA, and the right plot displays a heatmap of the corresponding covariance matrix. The brighter the color in the heatmap, the higher the covariance between two individuals.

to be known, where it models the observed genotypes by two factor matrices $Q$ and $F$, the individual admixture proportions and ancestral allele frequencies, respectively, in a full Bayesian setting. Due to scalability issues of MCMC, maximum likelihood methods based on the same likelihood model have been proposed [Tang et al., 2005, Alexander et al., 2009]:

$$L(Q, F; G) = \prod_{i=1}^{N} \prod_{m=1}^{M} p(g_{im} | Q_i, F_m). \ (1.3)$$

Here $g_{im}$ is the genotype of individual $i$ at site $m$, $Q_i$ is the admixture proportions of individual $i$ and $F_m$ is the ancestral allele frequencies at site $m$. The ADMIXTURE software [Alexander et al., 2009] has especially gained a lot of popularity due to its speed. The maximum likelihood model has been extended to genotype likelihoods in NGSadmix [Skotte et al., 2013] where the unobserved genotype is treated as a latent variable. An example of an admixture plot with estimated admixture proportions is visualized in Figure 1.4.

Both paper I and paper II introduce methods for inferring population structure using PCA while accounting for missingness with different approaches and data input. Additionally, a faster method for estimating individual admixture proportions is introduced in Paper I based on the PCA. Paper I performs PCA by estimating the covariance matrix for low-depth sequencing data while accounting for uncertainty using genotype likelihoods with inner iterations of truncated SVD. Paper II introduces another method for datasets with rampant missingness using called
genotypes or a single-allele sampling approach, where eigenvectors are iteratively inferred using truncated SVD. Lastly, paper IV introduces a novel approach to infer population structure from phased haplotype data using neural networks trained in windows to model LD heuristically along the genome.

![Figure 1.4: Estimated admixture proportions in the CEU, CHB and YRI populations of the 1000 Genomes Project [1000 Genomes Project Consortium et al., 2015], where each bar on the horizontal axis represents an individual.](image)

**Individual allele frequencies**

The concept of individual allele frequencies arises from the model in STRUCTURE, and the term is used in [Thornton et al., 2012], where they can be seen as the linear combination of admixture proportions and ancestral allele frequencies, $\Pi = QF^T$. In this way, the sampling distribution of the genotype given population structure can be described in an individual-specific manner such that $g_{im} \sim \text{Binomial}(2, \pi_{im})$, for individual $i$ at site $m$. This has been extended to PCA instead of the STRUCTURE model [Hao et al., 2016] using the top inferred principal components that capture population structure in a low-rank approximation of the genotype matrix. However as with clustering algorithms, this approach will also need to have predefined the number of components to retain, which is not a trivial task but tests have been derived to select an optimal number [Patterson et al., 2006, Shriner, 2011].

Paper I introduces a novel approach to use the low-rank approximation on the basis of population structure to model missingness for genotype likelihoods in low-depth sequencing data. This is further extended to genotype data or similar in the method introduced in paper II, where missingness can be modelled in an efficient accelerated approach that broadens its application to large-scale genetic datasets. The concept of individual allele frequencies is useful for many population genetic analyses in the presence of population structure. These include estimation of relatedness coefficients [Moltke and Albrechtsen, 2014, Moltke and Albrechtsen, 2014], per-
1.1. Background

individual inbreeding coefficients \cite{Moltke2015}, per-site inbreeding coefficients (HWE), PCA based selection scans and association analysis \cite{Song2015}. In paper III, the individual allele frequencies are used to account for population structure in detecting deviations from HWE in structured populations.

**Haplotype-based analysis**

The methods and approaches discussed until this point have focused on genotype data under the assumption that SNPs are independent such that LD is ignored. ChromoPainter \cite{Lawson2012} is instead an example of a method that utilizes the information in LD to infer population structure and haplotype sharing. It uses phased haplotype data, which means that haplotypes have been estimated from genotype data in a process called 'phasing', such that parental chromosomes of individuals have been inferred \cite{Scheet2006, Browning2007}. ChromoPainter then 'paints' the haplotypes as mosaics of each other using a hidden markov model \cite{Li2003} that directly models LD to achieve state-of-the-art results for inferring fine-scale population structure.

**1.1.4 Variational autoencoders**

The aforementioned PCA and clustering algorithms are all unsupervised learning methods, where the goal usually is to uncover patterns or structure in the data. A related unsupervised method for dimensionality reduction is the autoencoder that learns an efficient lower dimensional encoding of the data using neural networks \cite{Rumelhart1985, Baldi2012}. Neural networks, as multilayer feed-forward networks, are universal function approximators \cite{Hornik1990} and are therefore capable of learning a mapping from high-dimensional space of the input data to a lower dimensional latent space and the reverse mapping as well. The network mapping from the data to the latent encoding is commonly named the encoder and the network mapping the latent encoding back to the input data space is commonly named the decoder. This reconstruction of the data from its encoding is not lossless as was also not the case for the low-rank approximation in truncated SVD introduced earlier for estimating individual allele frequencies. A lot of work is being done to regularize the mappings of the autoencoder such that they capture meaningful and interpretable structure \cite{Bank2020}.

A variational autoencoder (VAE) is the merge of standard autoencoders with variational inference in order to regularize the mappings learnt from the neural networks \cite{Kingma2013, Rezende2014}. The latent space is forced to follow a probability distribution and the neural networks will encode and decode for parameters of the stochastic processes. The goal is to learn the unknown data generating distribution by introducing a latent variable to define a generative model that models the joint probability distribution of the data and the latent variable. Variational inference is a technique that is usually used to approximate the posterior
distribution of an unknown latent variable, and it is often used as an alternative to MCMC that is based on Markov chain sampling [Blei et al., 2017]. In the following section, I will use the mathematical notation used in the machine learning literature and the original paper of the variational autoencoder [Kingma and Welling, 2013]. The marginal log-likelihood of the data, \( x \), can be written as follows:

\[
\log p_\theta(x) = \log \int_z p_\theta(x, z)dz = \log \int_z p_\theta(x | z)p(z)dz,
\]

where the probability distribution of the data is parameterized by parameters \( \theta \). Variational inference introduces a function to describe a family of distributions over the latent variable, \( z \), that approximates its true posterior \( p(z | x) \approx q_\phi(z | x) \), and the function is parameterized by variational parameters, \( \phi \). In the VAE framework, stochastic variational inference is used to optimize both sets of parameters, \( (\theta, \phi) \), by maximizing the evidence lower bound (ELBO) of the marginal log-likelihood using neural networks. Thus, the optimization is based on amortized inference, where the parameters are estimated through mappings parameterized by the neural networks [Shu et al., 2018]. This means that the number of parameters depends on the network size and not on the sample size. The two neural networks are trained in an autoencoder architecture. For VAEs, the encoder and the decoder are not learning direct mappings to and from the latent encoding but rather probability distributions from which the latent encoding and the reconstruction can be sampled, respectively. An overall depiction of the VAE architecture is visualized in Figure [1.5]. The ELBO, which is the loss function, is defined as follows for the latent variable model in the VAE framework:

\[
\log p_\theta(x) \geq \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z | x)} \left[ \log p_\theta(x | z) - \log \frac{q_\phi(z | x)}{p(z)} \right].
\]

The first term in the expectation corresponds to the log-likelihood and finding the latent encoding, \( z \), which best reconstructs the input data, and it is therefore commonly called the reconstruction term. This is the direct connection to standard autoencoders. The second term regularizes the latent encoding by constraining the flexibility of the approximate posterior distribution by forcing it to be close to a chosen prior distribution. The ELBO is approximated by Monte Carlo samples of the latent encoding from the approximate posterior distribution \( q_\phi(z | x) \).

Paper IV introduces a new method for inferring haplotype and population structure using a variational autoencoder framework. With the assumption of a Gaussian mixture prior, haplotypes are jointly clustered and encoded in windows along the genome.
1.1. Background

Figure 1.5: Neural network architecture of the variational autoencoder with one latent variable. The input data $\mathbf{x}$ is mapped to a probability distribution through the encoder from which the latent encoding, $\mathbf{z}$, is sampled. The latent encoding is then mapped to a probability distribution from which the reconstruction of the input data, $\hat{\mathbf{x}}$, is sampled.
1.2 Contributions and perspectives

1.2.1 Paper I

The main contribution of the first paper is the development of two new methods in the \texttt{PCAngsd} framework [Meisner and Albrechtsen, 2018] for performing PCA and estimating individual admixture proportions in low-depth next-generation sequencing data. The study was motivated by an assumption in an existing tool for performing PCA, namely \texttt{ngsTools} [Fumagalli et al., 2014], that caused problems for datasets with low and variable sequencing data. In \texttt{ngsTools} and \texttt{PCAngsd}, it is assumed that genotypes are not observed and thus treated as a latent variable due to working on low-depth next-generation sequencing data. Therefore, both approaches adjust the standard approach of estimating the sample covariance matrix of the individuals by weighing the latent genotypes by their posterior probability on the basis of genotype likelihoods. However, \texttt{ngsTools} uses a prior genotype probability based on the estimated population allele frequency such that individuals are assumed to be from a homogeneous population. This is violated when dealing with structured populations or most modern genetic datasets. This assumption is commonly used by PCA software when dealing with missing data [Price et al., 2006, Abraham et al., 2017].

The problem with this choice of prior genotype probability is revealed when having samples of variable sequencing depth, where individuals will be modelled by their amount of missing information instead of population structure. \texttt{PCAngsd} tries to solve this problem by introducing individual allele frequencies to model the missing information, which are updated in an iterative algorithm such that individuals are conditionally independent given population structure. The standard approach for estimating the sample covariance (Equation 1.1.3) has been extended to the following for individual $i$ and $j$ in \texttt{PCAngsd}:

$$\text{cov}(i, j) = \frac{1}{M} \sum_{m=1}^{M} \sum_{g_i=0}^{2} \sum_{g_j=0}^{2} (g_i - 2f_m)(g_j - 2f_m)p(G_i = g_i, G_j = g_j | X_i, X_j, \pi_{im}, \pi_{jm}) \frac{2f_m(1 - f_m)}{2f_m(1 - f_m)}.$$ \hspace{1cm} (1.6)

Here $M$ is the number of variable sites, $G_i$ is the unobserved genotype of individual $i$, $X_i$ is the observed sequencing data of individual $i$, $f_m$ is the allele frequency at site $m$ and $\pi_{im}$ is the individual allele frequency of individual $i$ at site $m$. We show through simulations, examples from the 1000 Genomes Project [1000 Genomes Project Consortium et al., 2015] and a waterbuck dataset that we are capable of inferring the correct population structure from PCA using low and variable depth next-generation sequencing data by estimating individual allele frequencies in an iterative approach, which is of particular interest in large-scale studies of population and ancient genetics.

As described previously, there is a strong connection between admixture proportions and PCA since they both indirectly model the individual allele frequencies. We further show that we can...
utilize the individual allele frequencies estimated from the above PCA approach to estimate per-individual admixture proportions and ancestral allele frequencies using non-negative matrix factorization (NMF) and an assumption of $K$ ancestral populations. NMF is another dimensionality reduction method similarly to PCA, where two non-negative factor matrices are optimized to approximate another non-negative matrix. In our case, this boils down to the individual allele frequencies being approximated as follows, $\Pi \approx QF^T$, where $Q$ is the admixture proportions and $F$ is the ancestral allele frequencies. We derive an algorithm that shows comparable results to NGSadmix in a much faster procedure which is a very appealing choice for large datasets.

**Examples of usage of PCAngsd**

The two methods in PCAngsd are already widely used and have been applied in many studies to infer population structure in low-depth sequencing data. PCAngsd has, for example, been used in a study on killer whales with samples sequenced at 5X and 2X, in which the authors explore the relationship between defined ecotypes in the killer whales of Antarctic and non-Antarctic lineages [Foote et al., 2019]. Another example of usage is in a study of honey bees in the Americas, where PCAngsd is used to infer population structure in data with a mean sequencing depth of 5.1X (Figure 1.6). The authors investigate the spread of introduced African honey bee species including hybridization with European species that had previously invaded the Americas [Callee et al., 2020].

![Figure 1.6: Inferred population structure in honey bees of the Americas using PCAngsd. The PCA plot captures different species of honey bees including hybrids between African and European species that have been used for crop pollination. Figure S1 in original study Callee et al., 2020.](image)
The individual allele frequencies that are iteratively estimated using PCA are very informative of population structure. We have therefore incorporated them in a test statistic for HWE to account for population structure as an extension to the PCAngsd framework that is presented in paper III.

1.2.2 Paper II

The main contribution of the second paper is a method called EMU for inferring population structure using PCA in large-scale genetic datasets with missingness. EMU iteratively estimates individual allele frequencies based on PCA in a similar fashion to PCAngsd, however with different data applications. There are two main motivations for this project. This first is the recent use of ultra low-depth sequencing data from non-invasive prenatal tests (NIPT) in population genetics, where hundreds of thousands [Liu et al., 2018] and now millions of women are sequenced as part of the prenatal testing for chromosomal abnormalities of the fetus. The second motivation is the issues of merging different types of genetic data, for example, SNP arrays with exome sequencing data. The merging gives rise to patterns of non-random missingness that current methods can not deal with. Missing data in genetic datasets, commonly named missingness, is usually ignored or removed by most methods for inferring population structure. In the advent of popular fast SVD-based approaches, methods ignoring missingness performs mean imputation where the mean genotype value is used because the full genotype matrix is needed in the computations [Galinsky et al., 2016, Abraham et al., 2017]. However, this will lead the inferred population structure to correlate with the amount of missingness for the individual samples. Instead, EMU models the missingness explicitly in an iterative approach of estimating individual allele frequencies. It utilizes an expectation-maximization algorithm for PCA [Kiers, 1997] that keeps observed data as fixed values while modelling the missing data. This allows us to use an EM acceleration scheme in the iterative approach to speed up convergence in the estimation of individual allele frequencies. The EM algorithm is equivalent [Josse and Husson, 2012] to finding the individual allele frequencies that minimize the following expression for genotype data:

$$\min_\Pi \left\| W \odot (G - 2\Pi) \right\|^2. \quad (1.7)$$

Here $W$ is a weight matrix excluding entries of missing data, $\Pi$ is the individual allele frequencies and $\odot$ represents element-wise matrix multiplication. We show that a lot of commonly used methods for performing PCA in genetic data infer inaccurate population structure in the presence of missingness, whereas EMU is fully capable of capturing accurate population structure by modelling the missingness and being much faster than most other methods. We also implemented a memory-efficient variant of our method that scales well with the evergrowing number of large-scale genetic datasets.

The idea of EMU started with the emerge of very large genetic datasets such as the Chinese
1.2. Contributions and perspectives

Millionome Project\footnote{http://cmdb.bgi.com/}. The samples are based on DNA from non-invasive prenatal tests (NIPT), which in its phase 1 consisted of 140K samples and have since grown larger than millions. The samples have a reported sequencing depth of 0.06X-0.1X, where the concept of genotype likelihoods does not apply anymore due to the low amount of information for each site in the genome. In the original study [Liu et al., 2018], population structure is inferred using PCA by estimating the covariance matrix where sites are only used in pairwise estimation if both individuals have information to avoid dealing with missingness. The runtime for estimating the covariance matrix is reported as days without accounting for the following computationally expensive eigendecomposition to extract the eigenvectors. We show that using \texttt{EMU} we can infer population structure in less than two hours while also modelling the missing information for all individuals. We have now successfully run \texttt{EMU} on over one million Chinese ultra low-depth genomes.

1.2.3 Paper III

Paper III contributes with a new test statistic for Hardy-Weinberg equilibrium (HWE) in structured populations using low-depth sequencing data or genotype data. The HWE describes the relationship between allele frequencies and genotype frequencies between non-overlapping generations under the assumptions of a random mating population and in the absence of other evolutionary forces acting on the allele distributions [Hardy et al., 1908, Weinberg, 1908]. Due to the simplicity of the model, it has become a very common tool for quality control of variable sites by detecting deviations from HWE in a population as a null model [Wigginton et al., 2005, Waples, 2015]. Deviations from HWE may arise from effects leading to non-random mating, e.g. population structure, as well as genotyping or sequencing errors. Modern genetic datasets usually include samples from diverse ancestries and some degree of population structure is almost inevitable which will lead to deviations from HWE and obscuring potential errors or findings. Hence, we introduce a new likelihood ratio test to test for deviations from HWE while accounting for population structure, and it is implemented into the \texttt{PCAngsd} framework [Meisner and Albrechtsen, 2019]. The test is based on the introduction of a per-site inbreeding coefficient, $F$, which is based on Wright’s coefficient of inbreeding [Wright, 1949], but altered to quantify deviation from HWE due to either increased homozygosity or heterozygosity. The genotype probabilities with the extension of the per-site inbreeding coefficient can be described as follows, while also conditioning on individual allele frequencies to naturally account for population structure:
\[ p(G = g \mid \pi_{im}, F_m) = \begin{cases} 
(1 - \pi_{im})^2 + \pi_{im}(1 - \pi_{im})F_m, & g = 0, \\
2\pi_{im}(1 - \pi_{im})(1 - F_m), & g = 1, \\
\pi_{im}^2 + \pi_{im}(1 - \pi_{im})F_m, & g = 2. 
\end{cases} \] (1.8)

Where \(\pi_{im}\) is the individual allele frequency of individual \(i\) at site \(m\), \(F_m\) is the per-site inbreeding coefficient at site \(m\) and \(G\) is the genotype. We extended the likelihood model in Vieria et al. (2013) [Vieira et al., 2013] to account for population structure using individual allele frequencies. The model is based on genotype likelihoods to accommodate the uncertainty in the genotypes for low-depth next-generation sequencing data. The per-sites inbreeding coefficients are estimated using an EM algorithm and a likelihood ratio test can be defined to test for deviations from HWE. We show through simulations and applications to examples of the 1000 Genomes Project [1000 Genomes Project Consortium et al., 2015] that we are more accurate at detecting and quantifying deviations from HWE in structured populations in comparison to competing methods for both genotype and low-depth next-generation sequencing data.

**Examples of usage of structured HWE test**

Our test statistic has been applied in a study of songbirds with samples of varying sequencing depth from 4X-50X investigating differentiating populations in western North America. PCAngsd has been used to infer population structure and detect deviations from HWE in order to remove false variants because of paralogous loci in the genome, as paralogous loci are quantified with excess of heterozygosity. PCAngsd has also been applied in a study on African leopards. Here, it has also been used to filter out problematic regions in the genome that may be due to mapping problems or sequencing errors as a quality control step which are common in studies of non-model organisms [Pečnerová et al., 2020]. See Figure 1.7.

### 1.2.4 Paper IV

The main contribution of the last paper is a novel framework, HaploNet, which performs haplotype clustering and learns haplotype encodings using neural networks for phased haplotype data. The popularity of machine learning and deep learning has exploded over the past decade with successes in computer vision, natural language processing and even biology with the recent results of AlphaFold2. Unfortunately in the field of population genetics we usually do not have the luxury of big labelled datasets as in other fields that can be used to train big models through supervised learning. However, most of the recent studies of utilizing machine learning in population genetics are still methods based on supervised learning. Evolutionary simulators [Kelleher and Lohse, 2020] [Haller and Messer, 2019] are used to simulate training

[https://doi.org/10.1038/d41586-020-03348-4](https://doi.org/10.1038/d41586-020-03348-4)
Figure 1.7: Example of quantification of deviations from HWE in problematic regions in 3 different scaffolds of a leopard assembly for quality control using PCAngsd. Authors have removed the regions with excess of heterozygosity that are marked with boundaries of red lines. Figure S5E in original study [Pečnerová et al., 2020].

data from demographic models and the real data are assumed to fit the trained problem scenario [Sheehan and Song, 2016, Schrider and Kern, 2018, Flagel et al., 2019, Gower et al., 2020, Chan et al., 2018]. Instead HaploNet is a completely data-driven unsupervised learning method that models the latent structure of the data with an autoencoder approach. Using a variational autoencoder architecture with a Gaussian mixture prior, HaploNet learns latent clusterings and encodings of phased haplotype data in windows of a fixed number of SNPs. It takes phased haplotype data, where the haplotypes of each of the two chromosomes have been inferred, such that LD can implicitly be modelled and utilized in inferring haplotype and population structure.

We introduce an additional latent variable \( y \) to induce more flexibility and interpretability in the latent structure such that the marginal approximate posterior of \( z \) will be a mixture of Gaussians, while \( y \) will capture the clustering of the haplotypes. The following ELBO is maximized in HaploNet:

\[
\log p_\theta(\mathbf{x}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}, y | \mathbf{x})} \left[ \log p_\theta(\mathbf{x} | \mathbf{z}) - \log \frac{q_\phi(\mathbf{z} | \mathbf{x}, y)}{p_\theta(\mathbf{z} | y)} - \frac{q_\phi(y | \mathbf{x})}{p(y)} \right]. \tag{1.9}
\]
Where the hierarchical dependency of $z$ on $y$ is visible in the penalizing term for regularizing the latent encoding. The haplotype cluster, $y$, is sampled from a Categorical distribution conditioning on the data, $x$, and the latent encoding, $z$, is sampled from a Gaussian distribution conditioning on both the data and the sampled haplotype cluster, $y$. In this way, HaploNet learns a latent encoding of the haplotypes as well as their clustering in a given genomic window. We demonstrate that we are able to infer population structure using a PCA approach and an admixture model approach, which both capture fine-scale structure that are not possible in standard approaches using unphased genotype data, by utilizing haplotype information. HaploNet performs similarly to the state-of-the-art ChromoPainter software [Lawson et al., 2012], which explicitly models LD in a probabilistic framework, in the PCA approach but in a much faster and more scalable procedure. By having an interpretable latent variable, we are able to use HaploNet to model parameters of interest which we show by estimating admixture proportions in a completely unsupervised manner. HaploNet is a novel framework that models the haplotypes directly using neural networks, and with the current pace of developments in deep learning, the importance of integrating these frameworks with population genetics can only be understated with the number of large-scale genetic datasets being generated.
Bibliography


Paper I

Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data

By
Jonas Meisner$^1$ and Anders Albrechtsen$^1$

$^1$The Bioinformatics Centre, Department of Biology, University of Copenhagen, DK-2200, Denmark

Publication details
Published in *Genetics*, vol. 210, no. 2, 719-731, 2018.
Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data

Jonas Meisner1 and Anders Albrechtsen
The Bioinformatics Centre, Department of Biology, University of Copenhagen, DK-2200, Denmark
ORCID IDs: 0000-0002-9540-6673 (J.M.); 0000-0001-7306-031X (A.A.)

ABSTRACT
We here present two methods for inferring population structure and admixture proportions in low-depth next-generation sequencing (NGS) data. Inference of population structure is essential in both population genetics and association studies, and is often performed using principal component analysis (PCA) or clustering-based approaches. NGS methods provide large amounts of genetic data but are associated with statistical uncertainty, especially for low-depth sequencing data. Models can account for this uncertainty by working directly on genotype likelihoods of the unobserved genotypes. We propose a method for inferring population structure through PCA in an iterative heuristic approach of estimating individual allele frequencies, where we demonstrate improved accuracy in samples with low and variable sequencing depth for both simulated and real datasets. We also use the estimated individual allele frequencies in a fast non-negative matrix factorization method to estimate admixture proportions. Both methods have been implemented in the PCAngsd framework available at http://www.popgen.dk/software/.

KEYWORDS
Population structure; PCA; admixture; ancestry; next-generation sequencing; genotype likelihoods; low depth

POPULATION genetic studies often consist of individuals of diverse ancestries, and inference of population structure therefore plays an important role in population genetics and association studies. Population stratification can act as a confounding factor in association studies as it can lead to spurious associations (Marchini et al. 2004). Principal component analysis (PCA) has been used in genetics for a long time, such as in Menozzi et al. (1978) where synthetic maps were produced in an exploratory analysis of genetic variation. PCA is now a common tool in population genetic studies, where its dimension reduction properties can be used to visualize population structure by summarizing the genetic variation through principal components (Novembre and Stephens 2008), correct for population stratification in association studies, and investigate demographic history (Patterson et al. 2006; Price et al. 2006; Fumagalli et al. 2013) as well as perform genome selection scans (Hao et al. 2015; Galinsky et al. 2016; Luu et al. 2017). PCA is an appealing approach to infer population structure as the aim is not to classify the individuals into discrete populations, but instead to describe continuous axes of genetic variation such that heterogeneous populations and admixed individuals can be better represented (Patterson et al. 2006). Another successful approach in modeling complex population structure is to estimate admixture proportions based on clustering-based methods (Pritchard et al. 2000; Tang et al. 2005; Alexander et al. 2009; Skotte et al. 2013), such as the popular software ADMIXTURE, which have also been used for correction of population stratification in association studies (Price et al. 2010).

Next-generation sequencing (NGS) methods (Metzker 2010) produce a large amount of DNA sequencing data at low cost and are commonly used in population genetic studies (Nielsen et al. 2012). But NGS methods are associated with high error rates usually caused by several factors such as sampling, alignment, and sequencing errors. Many NGS studies are based on medium (\(<15\times\)) and low (\(<5\times\)) depth data due to the demand for large sample sizes as seen in large-scale sequencing studies, e.g., 1000 Genomes Project Consortium (2010, 2012). However, the use of medium- and, especially, low-depth sequencing data introduces challenges rooted in the statistical uncertainty induced when calling genotypes and variants in these scenarios (Nielsen et al. 2012). The statistical uncertainty increases for low-depth...
samples due to the increased difficulty of distinguishing a variable site from a sequencing error with the information provided. Problems can arise due to chromosomes being sampled with replacement in the sequencing process, and both alleles may not have been sampled for a heterozygous individual in low-depth scenarios. Homozygous genotypes may also be wrongly inferred as heterozygous due to sequencing errors. Thus, genotype calling will associate individuals with a statistical uncertainty that should be taken into account (Nielsen et al. 2011, 2012).

To overcome these problems related to NGS data and genotype calling, probabilistic methods have been developed to take use of genotype likelihoods in combination with external information for various population genetic parameters (Kim et al. 2011; Nielsen et al. 2012; Fumagalli et al. 2013; Skotte et al. 2013; Vieira et al. 2013; Korneliusen et al. 2014; Kousathanas et al. 2017), such that posterior genotype probabilities can be used to model the related uncertainty. Genotype likelihoods can be estimated to incorporate errors of the sequencing process such as the base quality uncertainty. Genotype likelihoods can be estimated to incorporate the related uncertainties of the individuals on axes of genetic variation in the dataset. The top principal components capture most of the population structure as they represent the projection of the individuals on axes of genetic variation in the dataset (Patterson et al. 2006; Engelhardt and Stephens 2010).

External information can be incorporated to define posterior genotype probabilities using Bayes’ theorem in combination with genotype likelihoods (Nielsen et al. 2011). The population allele frequency is often used as information in the estimation of prior genotype probability \( P(G_i | p_i) \), for an individual \( i \) in site \( s \) (Kim et al. 2011; Nielsen et al. 2012; Fumagalli et al. 2013; Vieira et al. 2013). Assuming the population is in Hardy-Weinberg equilibrium (HWE) for a site \( s \), the prior genotype probability is then given as \( P(G_i = 0 | p_i) = (1 - p_i)^2 \), \( P(G_i = 1 | p_i) = 2p_i(1 - p_i) \) and \( P(G_i = 2 | p_i) = p_i^2 \) for the three different possible genotypes. As defined in Kim et al. (2011), using the estimated population allele frequency \( \bar{p}_i \), the posterior genotype probability is computed as follows for individual \( i \) in site \( s \):

\[
P(G_i = g | X_i, \bar{p}_i) = \frac{P(X_i | G_i = g)p(G_i = g | \bar{p}_i)}{\sum_{g \in S} P(X_i | G_i = g')p(G_i = g' | \bar{p}_i)}
\]  

(1)

**PCA**

The standard way of performing PCA in population genetics and using it to infer population structure is based on the method defined in Patterson et al. (2006). For a genotype matrix \( G \) of \( n \) individuals and \( m \) variable sites, the \( n \times n \) covariance matrix \( C \), also known as the genetic relationship matrix (GRM), is computed as follows for two individuals \( i \) and \( j \):

\[
c_{ij} = \frac{1}{m} \sum_{s=1}^{m} \frac{g_{is} - 2\bar{p}_i}{2\bar{p}_i (1 - \bar{p}_i)} \frac{g_{js} - 2\bar{p}_j}{2\bar{p}_j (1 - \bar{p}_j)}
\]

(2)

Here, \( g_{is} \) is the observed genotype for individual \( i \) in site \( s \), to distinguish it from \( G \) defined above for unobserved genotypes, and \( \bar{p} \) is the estimated population allele frequency. The principal components are then inferred by performing an eigendecom-position of the covariance matrix, such that \( C = V \Sigma V^T \) with \( V \) being the matrix of eigenvectors and \( \Sigma \) the diagonal matrix of the corresponding eigenvalues. Principal components and eigenvectors will be used interchangeably throughout this study. The top principal components capture most of the population structure as they represent the projection of the individuals on axes of genetic variation in the dataset (Patterson et al. 2006; Engelhardt and Stephens 2010).

This method has been extended to NGS data in Fumagalli et al. (2013), as well as in Skotte et al. (2012), using the
probabilistic framework described in Equation 1, by summing over the genotypes of each individual weighted by the joint posterior genotype probabilities under the assumption of HWE in the whole sample. The method has been implemented in the ngsTools framework (Fumagalli et al. 2014). The covariance matrix is estimated as follows for NGS data using only known variable sites for two individuals i and j:

\[
q_{ij} = \sum_{g=0}^{1} \sum_{g'=0}^{1} (g - 2p_j) (g' - 2p_i) P(G_{ij} = g, g' | X_i, X_j, \hat{p}_s) \frac{2p_j (1-p_j)}{2p_i (1-p_i)}
\]

(3)

ngsTools splits up the joint posterior probability, \( P(G_{ij} | X_i, X_j, \hat{p}_s) \), into \( P(G_{ij} | X_i, \hat{p}_s) P(G_{ij} | X_j, \hat{p}_s) \) for \( i \neq j \) by assuming conditional independence between individuals given the estimated population allele frequencies. The non-diagonal entries in the covariance matrix are now directly estimated from the posterior expectations of the genotype instead of the observed genotypes as described in Equation 2. The original method weights each site by its probability of being a variable site such that SNP calling is not needed prior to the covariance matrix estimation. This is not taken into account in this study as we are using called variable sites to infer population structure. The population allele frequencies are estimated from the genotype likelihoods using an expectation maximization (EM) algorithm (Kim et al. 2011) as described in the supplemental material.

The problem with this approach is that the assumption of conditional independence between individuals given the population allele frequency is only valid when there is no population structure. Here, we propose a novel approach of estimating the covariance matrix using iteratively estimated individual allele frequencies to update the prior information of the posterior genotype probability. Thereby, we condition on the individual allele frequencies as in the clustering-based approaches such as Pritchard et al. (2000), Tang et al. (2005), Alexander et al. (2009), Skotte et al. (2013).

**Individual allele frequencies**

A model for estimating individual allele frequencies based on population structure was introduced in STRUCTURE (Pritchard et al. 2000), as later described in Equation 13. Hao et al. (2015) proposed a different model for estimating individual allele frequencies \( \mathbf{H} \) by using the information in the principal components instead of having an assumption of K ancestral populations. The model is defined as the matrix product,

\[ \mathbf{H} = \mathbf{S} \mathbf{A} \]

(4)

where \( \mathbf{S} \) represents the population structure such that \( \mathbf{A} \) represents the mapping of the population structure \( \mathbf{S} \) to the allele frequencies. Hao et al. (2015) estimated the individual allele frequencies through a singular value decomposition (SVD) method, where genotypes are reconstructed using only the top D principal components such that they will be modeled by population structure. A similar approach has been proposed by Conomos et al. (2016), where the inferred principal components are used to estimate individual allele frequencies in a simple linear regression model. However, due to working on NGS data and not knowing the genotypes, we are extending the method of Hao et al. (2015) to NGS data by using posterior expectations of the genotypes, referred to as genotype dosages, instead of genotypes. Thus, we will be using

\[
\mathbf{E}(G_{is} | \chi_{is}, \hat{p}_s) = \sum_{g=0}^{2} g P(G_{is} = g | \chi_{is}, \hat{p}_s),
\]

(5)

for individual \( i \) in site \( s \).

The individual allele frequencies are then estimated by performing a SVD on the centered genotype dosages, and reconstructing them using only the top D principal components. \( \mathbf{S} \) is then added to the reconstruction and scaled by 1/2 based on a binomial distribution assumption of \( G_{is} \) for \( i = 1, \ldots, n \) and \( s = 1, \ldots, m \), to produce the individual allele frequencies. Since SVD is a method that takes real-valued input, we will have to truncate the estimated individual allele frequencies in order to constrain them in the range \([0, 1]\). However, Hao et al. (2015) showed that the resulting estimates were still very accurate for common variants considering this limitation.

For ease of notation, let \( \mathbf{E} \) be the \( n \times m \) matrix of genotype dosages, \( e_{is} = \mathbb{E}(G_{is} | \chi_{is}, \hat{p}_s) \), for \( i = 1, \ldots, n \) and \( s = 1, \ldots, m \). The following steps for estimating the individual allele frequencies are adopted from the SVD method (Hao et al. 2015) to work on NGS data:

1. Form matrix \( \mathbf{S} = [\mathbf{W}_1, \ldots, \mathbf{W}_D] \) and all representing column vectors, such that Equation 4 can be approximated as \( \mathbf{H} = \mathbf{S} \mathbf{A} \).
2. If the estimated allele frequency estimates in a range based on a small value \( \gamma \) (0.1 x 10^-4), such that \( \hat{p}_s \in [\gamma, 1-\gamma] \) for \( i = 1, \ldots, n \) and \( s = 1, \ldots, m \).

**Algorithm 1: SVD method for estimating individual allele frequencies**

1. The centered genotype dosages are constructed as \( \mathbf{E}^{(C)} = \mathbf{E} - \mathbf{2}\mathbf{p} \) for \( i = 1, \ldots, n \).
2. Perform SVD on the centered genotype dosages, \( \mathbf{E}^{(C)} = \mathbf{W} \mathbf{A} \mathbf{U}^T \), where \( \mathbf{W} \) will represent population structure similarly to \( \mathbf{V} \).
3. Define \( \mathbf{E}^{(C)}_{is} \) to be the prediction of the centered genotype dosages using only the top D principal components, \( \mathbf{E}^{(C)}_{is} = \mathbf{W}_i \mathbf{A}_i \mathbf{U}_i^T \).
4. Estimate \( \mathbf{H} \) by adding \( \mathbf{2p} \) to \( \mathbf{E}^{(C)}_{is} \) row-wise and scaling by 1/2 based on \( \hat{p}_s = 1/2G_{is} \).

We now incorporate the individual allele frequencies into the estimation of posterior genotype probabilities. The estimated individual allele frequencies are used as updated prior information instead of the population allele frequencies, and will be able to model missing data with the inferred population structure of the individuals. Thus, the posterior genotype probabilities are estimated as follows for individual \( i \) in site \( s \):
Each individual is now seen as a single population with allele frequency \( \hat{\pi}_u \), where as the prior genotype probability are estimated assuming HWE, such that \( P(G = 0|\hat{\pi}_u) = (1-\hat{\pi}_u)^2 \), \( P(G = 1|\hat{\pi}_u) = 2(1-\hat{\pi}_u)\hat{\pi}_u \) and \( P(G = 2|\hat{\pi}_u) = \hat{\pi}_u^2 \). An updated definition of the posterior expectations of the genotypes is then given as:

\[
E[G|X_u, \hat{\pi}_u] = \sum_{g=0}^{2} g P(G = g|X_u, \hat{\pi}_u).
\] (7)

This procedure of updating the prior information can be iterated to estimate new individual allele frequencies on the basis of updated population structure. Therefore, we propose the following algorithm for an iterative procedure of estimating the individual allele frequencies.

**Convergence of our iterative method is defined as when the root-mean-square deviation (RMSD) of the inferred population structure in the SVD \( W \) is smaller than a value \( \mu(1.0 \times 10^{-5}) \) between two successive iterations. The RMSD of iteration \( t \) for \( D \) principal components is given as,

\[
RMSD = \sqrt{\frac{1}{2D} \sum_{d=1}^{D} \sum_{i=1}^{n} \left( W^2_{id} - \bar{W}^2_{id} \right)^2}. \] (8)

**Covariance matrix**

We now use the final set of individual allele frequencies to estimate an updated covariance matrix in a similar model as in Equation 3, but incorporating the individual allele frequencies into the joint posterior probability. The entries of the covariance matrix \( C \) are now defined as follows for individuals \( i \) and \( j \):

\[
c_{ij} = \frac{1}{m} \sum_{u=1}^{m} \sum_{d=1}^{D} \left( \left( \hat{\pi}_u^2 - 2\hat{\pi}_u \right) P(G_i = g_i|X_u, \hat{\pi}_u) \right) \left( \left( \hat{\pi}_j^2 - 2\hat{\pi}_j \right) P(G_j = g_j|X_u, \hat{\pi}_j) \right) \frac{1}{2\hat{p}_i(1-\hat{p}_j)}.
\] (9)

For \( i \neq j \), the joint posterior probability can be computed as \( P(G_i|X_u, \hat{\pi}_u)P(G_j|X_u, \hat{\pi}_j) \), since, in contrast to the assumption made in the model of Fumagalli et al. (2013) using population allele frequencies, the individuals are conditionally independent given the individual allele frequencies. The above equation can be expressed in terms of the genotype dosages for ease of notation and computation for \( i \neq j \):

\[
c_{ij} = \frac{1}{m} \sum_{u=1}^{m} \left( E[G_i|X_u, \hat{\pi}_u] - 2\hat{p}_i \right) \left( E[G_j|X_u, \hat{\pi}_j] - 2\hat{p}_j \right) \frac{1}{2\hat{p}_i(1-\hat{p}_j)}. \] (10)

However, for \( i = j \) (diagonal of the covariance matrix), the joint posterior probability is simplified to \( P(G_i|X_u, \hat{\pi}_u) \), such that the estimation of the diagonal covariance entries is given as:

\[
c_{ii} = \frac{1}{m} \sum_{u=1}^{m} \sum_{g=0}^{2} \left( \hat{\pi}_u^2 - 2\hat{\pi}_u \right) P(G_i = g_i|X_u, \hat{\pi}_u) \frac{1}{2\hat{p}_i(1-\hat{p}_i)}. \] (11)

An eigendecomposition of the updated estimated covariance matrix is then performed to obtain the principal components as described earlier, \( C = \Sigma \Sigma^T \). Note that \( V \) and \( W \) from algorithm 1 are not the same even though they both represent population structure through axes of genetic variation in the dataset. This is due to a different scaling, and the joint posterior probability of Equation 11 is not taken into account in \( W \) for \( i = j \).

**Number of principal components**

It can be hard to determine the optimal number of principal components that represent population structure. In our method, we are using Velicer’s minimum average partial (MAP) test as proposed by Shriner (2011) to automatically detect the number of top principal components \( D \) used for estimating the individual allele frequencies. Shriner showed that the test based on a Tracy-Widom distribution (Patterson et al. 2006) systematically overestimates the number of significant principal components, and performs even worse for datasets including admixed individuals. However, in order to be able to perform the MAP test and detect the optimal \( D \), an initial covariance matrix is estimated based on the model in Equation 3.

The MAP test is performed on the estimated initial covariance matrix \( C \) for NGS data as an approximation of the Pearson correlation matrix used by Shriner. Using the notion of Shriner, \( C^i_j \) is defined as the matrix of partial correlations after having partialled out the first \( d \) principal components. Velicer (1976) proposed the summary statistic \( l_d = \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} \left( C^i_j \right)^2 \), where \( C^i_j \) represents the entry in \( C^i_j \) for individuals \( i \) and \( j \). Thus, the test statistic \( l_d \) represents the average squared correlation after partialling out the top \( d \) principal components. The number of top principal components that represent population structure is then chosen as \( d = \arg\min_d l_d \) for \( d = 0, \ldots, m-1 \). We have used the same implementation of the MAP test as Shriner.
The MAP test, and the preceding estimation of the initial covariance matrix, can be avoided by having prior knowledge of an optimal $D$ for the dataset being analyzed and manually selecting $D$.

**Genotype calling**

As previously shown in Nielsen et al. (2012) and Fumagalli et al. (2013), genotypes can be called from posterior genotype probabilities to achieve higher accuracy in low-depth NGS scenarios. We can adapt this concept to our posterior genotype probabilities based on individual allele frequencies, such that genotypes can be called at a higher accuracy in structured populations from low-depth NGS data. The genotype for individual $i$ in site $s$ is called as follows:

$$
\hat{g}_{is} = \arg\max_{g \in \{0,1,2\}} P(G_{is} = g | X_{is}, \pi_{is}).
$$

(12)

**Admixture proportions**

Based on the likelihood model defined in STRUCTURE (Pritchard et al. 2000), individual allele frequencies $\mathbf{H}$ can be estimated using admixture proportions $\mathbf{Q}$ and population-specific allele frequencies $\mathbf{F}$ (Alexander et al. 2009), such that:

$$
\pi_{is} = \sum_{k=1}^{K} q_{ik} f_{ik};
$$

(13)

for an individual $i$ in a variable site $s$. This is based on an assumption of $K$ ancestral populations where $\sum_{k=1}^{K} q_{ia} = 1$ and $0 \leq q, f \leq 1 \forall a, f \in \{Q, F\}$. Here $\mathbf{Q}$ and $\mathbf{F}$ must be inferred in order to estimate the individual allele frequencies, whereas $K$ is assumed to be known. One probabilistic approach for inferring population structure through admixture proportions for low-depth NGS data has been implemented in the NGSadmix software (Skotte et al. 2013). Here both parameters, $\mathbf{Q}$ and $\mathbf{F}$, are jointly estimated in an EM algorithm using genotype likelihoods.

In our case, we have already estimated the individual allele frequencies based on our iterative procedure using PCA described above. $K$ can be chosen as the number of principal components $D + 1$, since it would explain the number of distinct ancestral population from which the individual allele frequencies have been estimated. There is, however, not always a direct interpretation between principal components and admixture proportions (Alexander et al. 2009; Engelhardt and Stephens 2010). Therefore, we propose an approach based on NMF to infer $\mathbf{Q}$ and $\mathbf{F}$ using only our estimated individual allele frequencies as information for low depth NGS data. NMF has previously been applied directly on genotype data to infer population structure and admixture proportions by Frichot et al. (2014), where their method showed comparable accuracy and faster runtime in comparison to ADMIXTURE.

NMF is a dimension reduction and factor analysis method for finding a low-rank approximation of a matrix, which is similar to PCA, but NMF is constrained to find non-negative low dimensional matrices. For an non-negative matrix $\mathbf{H} \in \mathbb{R}^{n \times m}$, the goal of NMF is to find an approximation of $\mathbf{H}$ based on two non-negative factor matrices $\mathbf{Q} \in \mathbb{R}^{n \times K}$ and $\mathbf{F} \in \mathbb{R}^{m \times K}$, such that:

$$
\mathbf{H} \approx \mathbf{QF}^T.
$$

(14)

$\mathbf{Q}$ will consist of columns of non-negative basis vectors such that linear combinations of these approximates $\mathbf{H}$ through $\mathbf{F}$. Thus, based on the non-negative nature of our parameters, we can apply the ideas of NMF to infer admixture proportions $\mathbf{Q}$ and population-specific allele frequencies $\mathbf{F}$ from our individual allele frequencies. We use a combination of recent research in NMF to minimize the following least squares problem with a sparseness constraint on $\mathbf{Q}$:

$$
\min_{\mathbf{Q,F}} \|\mathbf{H} - \mathbf{QF}^T\|_F^2 + \alpha \sum_{i=1}^{m} \sum_{k=1}^{K} |q_{ik}|.
$$

(15)

for $\mathbf{Q} \geq 0$, $\mathbf{F} \geq 0$, and $\alpha \geq 0$. Here $\|\cdot\|_F$ is the Frobenius norm of a matrix and $\alpha$ is the regularization parameter controlling the sparseness enforced as also introduced in Frichot et al. (2014).

Lee and Seung (1999, 2001) proposed a multiplicative update (MU) algorithm to solve the standard NMF problem without the sparseness constraint included above. Their update rules can be seen as conservative steps in a gradient descent optimization problem for updating $\mathbf{F}$ and $\mathbf{Q}$, which ensure that the non-negative constraint holds for each update. Hoyer (2002) extended the MU to incorporate the sparseness constraint described in Equation 15 for $\mathbf{Q}$. For $\alpha > 0$, the regularization parameter is used to reduce noise, especially induced by the uncertainty of low-depth NGS data, in the estimated admixture proportions by enforcing sparseness in the solution. An iteration of using the MU rules is then described as follows:

$$
\tilde{\mathbf{F}}^{(t+1)} = \tilde{\mathbf{F}}^{(t)} \odot \frac{\mathbf{Q}^{T(i)} \mathbf{H}}{\mathbf{Q}^{T(i)} \mathbf{Q}^{(i)}/\mathbf{F}^{(t)})};
$$

(16)

$$
\tilde{\mathbf{Q}}^{(t+1)} = \tilde{\mathbf{Q}}^{(t)} \odot \frac{\mathbf{F}^{(t+1)}}{\mathbf{Q}^{(i)} \mathbf{Q}^{(i)} \mathbf{F}^{(t+1)}/\mathbf{F}^{(t+1)}} + \alpha
$$

(17)

where $\odot$ represents element-wise multiplication, and the division operator is element-wise as well.

However, MU has been shown to have a slow convergence rate, especially for dense matrices, and our approach is therefore to accelerate MU by combining two different techniques. We propose an algorithm of combining the acceleration scheme described by Gillis and Glineur (2012) with the asymmetric stochastic gradient descent algorithm (ASG-MU) of Serizel et al. (2016) for updating $\mathbf{F}$ and $\mathbf{Q}$ in a fast approach. The acceleration scheme of Gillis and Glineur (2012) updates each acceleration scheme of Gillis and Glineur (2012) updates each
II into a set of $B$ mini-batches, which are then updated sequentially in a permuted order to improve the convergence rate and performance of MU (Serizel et al. 2016). After each update, we truncate the entries of both $F$ and $Q$ to be in range $[0, 1]$ and normalize the rows of $Q$ to sum to one. The concept of combining an acceleration scheme with a stochastic gradient descent approach for MU has also been explored in Kasai (2017).

The algorithm is iterated until the admixture proportions has converged. Convergence is defined as when the RMSD of estimated admixture proportions of two successive iterations are smaller than a value $\phi (1.0 \times 10^{-4})$. The RMSD of iteration $t + 1$ is given as,

$$\text{RMSD} = \sqrt{\frac{1}{NK} \sum_{i=1}^{n} \sum_{k=1}^{K} (q_{ik}^{(t+1)} - q_{ik}^{(t)})^2}.$$  \hspace{1cm} (18)

The $\phi$ parameter enforcing sparseness in the estimated solution of $Q$ is arbitrarily specified. However the use of the likelihood measure in the NGSdix (Skotte et al. 2013) model can be used to determine the $\phi$ parameter fitting the dataset.

The likelihood measure is defined as:

$$L(\hat{Q}, \tilde{F}) = \prod_{i=1}^{n} \prod_{u=1}^{m} \sum_{g=d}^{2} P(X_{iu} = g | G_{iu} = g, \pi_{iu}),$$  \hspace{1cm} (19)

where $\pi_{iu} = \sum_{k=1}^{K} q_{uk} f_{ik}$. Based on the fast estimation of admixture proportions using our NMF algorithm, an appropriate $\phi$ can easily be found by scanning a specified interval in an automated fashion based on the likelihood measure. This can be performed without sacrificing significant runtime compared to NGSdix due to already having estimated the individual allele frequencies for a particular $K$.

**Implementation**

Both presented methods have been implemented in a Python framework named PCAngsd. The framework is freely available at [http://www.popgen.dk/software/](http://www.popgen.dk/software/).

The memory requirements of PCAngsd is $O(mn)$ as the entire matrix of genotype likelihoods needs to be stored in memory for both methods. The most computationally expensive step is the estimation of individual allele frequencies and covariance matrix $[O(m^2n)]$. However, a fast SVD method for only computing the top $D$ eigenvectors, implemented in the Scipy library (Jones et al. 2014) using ARPACK (Lehoucq et al. 1998) as an eigensolver, has been used to speed up the iterative estimations of the individual allele frequencies. PCAngsd is also multithreaded to take advantage of several cores, and the backbone of the framework is based on Numpy data structures (van der Walt et al. 2011) using the Numba library (Lam et al. 2015) to speed up bottlenecks with just-in-time (JIT) compilation.

**Simple simulation of genotypes and sequencing data**

To test the capabilities of our two presented methods, we simulated low-depth NGS data and generated genotype likelihoods. Allele frequencies of the reference panel of the Human Genome Diversity Project (HGDP) (Cann et al. 2002) were used to generate a total of 380 individuals from three distinct populations (French, Han Chinese, Yoruba) including admixed individuals in ~0.4 million SNPs across all autosomes. As the allele frequencies are known for each population, the genotypes of each individual can be sampled from a binomial distribution for each di-allelic SNP, using the population-specific allele frequency or an admixed allele frequency as parameter. No linkage disequilibrium (LD) was simulated. The genotypes are therefore known and are used in the evaluation of our methods in our low-depth scenarios.

The number of reads in each SNP were sampled from a Poisson distribution with a mean parameter resembling the average sequencing depth of the individual, and the genotype was used to sample the number of derived alleles from a binomial distribution using the sampled depth as parameter. The average sequencing depth of each individual was sampled uniformly random from a range of $[0.5, 5]$. Sequencing errors were incorporated by sampling each read with a probability $e = 0.01$ of being an error. The genotype likelihoods were then finally generated from the probability mass function of a binomial distribution using the sampled parameters and $e$. This approach of genotype likelihood simulation has previously been used in Kim et al. (2011), Skotte et al. (2013), and Vieira et al. (2013).

A complex admixture scenario was constructed to test the capabilities of our methods; 100 individuals were sampled directly from each of the population-specific allele frequencies (non-admixed), while 50 individuals were sampled to have equal ancestry from each of the three distinct populations (three-way admixture). Finally, 30 individuals were sampled from a gradient of ancestry between all pairs of the ancestral populations (two-way admixture).

**100 Genomes low-depth sequencing data**

We also analyzed human low-coverage NGS data of 193 individuals from the 100 Genomes Project Consortium et al. (2010, 2012). The individuals were from four different populations consisting of 41 from CEU (Utah residents with Northern and Western European ancestry), 40 from CHB (Han Chinese in Beijing), 48 from YRI (Yoruba in Ibadan), and 64 individuals from MXL (Mexican ancestry in Los Angeles), representing an admixed population of European and Native American ancestry. The individuals from the low-coverage datasets have a varying sequencing depth from $1.5 \times$ to $12.5 \times$ after site filtering. An advantage of using the low-coverage data of the 100 Genomes Project data are that reliable genotypes are available that can be used for validation purposes.

SNP calling and estimation of genotype likelihoods of the 100 Genomes dataset was performed in ANGSD (Korneliussen et al. 2014) using simple read quality filters. A significance threshold of $1.0 \times 10^{-6}$ was used for SNP calling alongside a MAF threshold of 0.05 to remove rare variants. A total number of 8 million variable sites across all autosomes was used in the analyses. The full ANGSD
command used to generate the genotype likelihoods is provided in the supplemental material.

**Waterbuck low-depth sequencing data**

Lastly, an animal dataset (nonmodel organism) as also included in our study. A reduced low-depth NGS dataset of the waterbuck (*Kobus ellipsiprymnus*) originating from C. Pedersen et al. (University of Copenhagen, unpublished data) was analyzed. The dataset consists of 73 samples that were sampled at five different sites in Africa with a varying sequencing depth from 2.2× to 4.7× aligned to 88,935 scaffolds. The dataset was reduced to only include sampling sites with >10 samples such that the inferred axes of genetic variation will reflect true population structure. As performed for the 1000 Genomes dataset, genotype likelihoods were estimated in ANGSD with the same SNP and MAF filters. A total number of 9.4 million SNPs across the autosomes of the waterbuck was analyzed in this study.

**Data availability**

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables. The waterbuck dataset analyzed in our study is publicly available in the European Nucleotide Archive (ENA) repository (PRJEB28089). Supplemental material available at Figshare: https://doi.org/10.25386/genetics.6953243.

**Results**

For the simulated and 1000 Genomes datasets, results estimated in PCAngsd on low-depth NGS data were evaluated against the results estimated from genotype data, as well as naïvely called genotypes from genotype likelihoods. The model in Equation 2 was used to perform PCA, while ADMIXTURE was used to estimate admixture proportions on the “true” genotype datasets. The performance of PCAngsd was also compared to existing genotype likelihood methods, with the ngsTools model (Equation 3) for performing PCA, and NGSAdmix (Equation 19) for estimating admixture proportions. In all the following cases of admixture plots estimated by PCAngsd, we used $B = 5$, and $\alpha$ was chosen as the one maximizing the likelihood measure described above (Equation 19), also shown in Supplemental Material, Figure S5.

RMSD was used to evaluate the performances of both NGS methods for estimating admixture proportions in terms of accuracy:

$$\text{RMSD} = \sqrt{\frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} \left( q_{ik}^{(\text{geno})} - q_{ik}^{(\text{NGS})} \right)^2},$$

where $q_{ik}^{(\text{geno})}$ and $q_{ik}^{(\text{NGS})}$ represent the estimated admixture proportion for individual $i$ in ancestral population $k$ from known genotypes and NGS data, respectively. The accuracy of the inferred PCA plots of both NGS methods was also compared to the PCA plots of known genotypes for the simulated and 1000 Genomes datasets using RMSD. However, a Procrustes analysis (Wang et al. 2010; Fumagalli et al. 2013) had to be performed prior to the comparison as the direction of the principal components can differ based on the eigendecomposition of the covariance matrices.

All tests in this study were performed server-side using 32 threads (Intel Xeon CPU ES-2690) for both PCAngsd and NGSAdmix.

**Simulation**

The results of performing PCA on the simulated dataset based on frequencies from three human populations are displayed in Figure 1, where we simulated unadmixed, two-way admixed and three-way admixed individuals. The MAP test reported two significant principal components, which was also expected for individuals simulated from three distinct populations. The inferred principal components clearly show the importance of taking individual allele frequencies into account in the probabilistic framework. Here, PCAngsd was able to infer the population structure of individuals from distinct populations and admixed individuals nicely, as also verified by a Procrustes analysis obtaining a RMSD of 0.00121, when compared to the PCA inferred from the true genotypes. There is clear bias in the results of the ngsTools model, where the patterns represent sequencing depth rather than population structure, as seen in Figure S1. The individuals are acting as a gradient toward the origin due to their

Figure 1 PCA plots of the top two principal components in the simulated dataset consisting of 380 individuals and 0.4 million variable sites. The left-hand plot shows the PCA performed on the known genotypes using Equation 2. The middle plot shows the PCA performed by PCAngsd, and the right-hand plot displays the PCA performed by the ngsTools model (Equation 3).
varying sequencing depth. The biased performance of ngsTools was also reflected in the corresponding Procrustes analysis, with a RMSD of 0.0174.

To ensure that the individual allele frequencies estimated using PCAngsd are representative estimates, we compared them to the allele frequencies of the HGDP reference panel from which the genotypes of each individual has been sampled. Sampling errors were therefore not taken into account in the comparison. The estimates obtained from NGSadmix were also compared. The estimates of PCAngsd obtain a RMSD value of 0.0330, and the estimates of NGSadmix a value of 0.0327 based on low-depth NGS data. The results of PCAngsd are displayed in Figure S9.

The estimated admixture proportions of the simulated dataset are displayed in Figure 2. PCAngsd estimated the admixture proportions well with a RMSD of 0.00476 compared to the ADMIXTURE estimates of the known genotypes, but was, however, outperformed by NGSadmix with a RMSD of 0.00184. For the 380 individuals and 0.4 million SNPs using K = 3, PCAngsd had an average runtime of only 2.9 min while NGSadmix had an average runtime of 7.9 min (Table 1).

### 1000 Genomes

We also applied the methods of PCAngsd to the CEU (European ancestry), CHB (Chinese ancestry), YRI (Nigerian ancestry), and MXL (Mexican ancestry) populations of the low-coverage 1000 Genomes dataset. The MAP test indicated evidence of three significant principal components, meaning that the Native American ancestry explains enough genetic variance in the dataset to represent an axis of its own. The results of the PCA are displayed in Figure 3. As was also seen for the simulated dataset, PCAngsd is able to cluster all individuals almost perfectly, while the ngsTools model is only able to capture some of the same population structure patterns with some of the populations looking admixed. Its results are still biased by the variable sequencing depth, as also seen in Figure S2. The RMSD values of the Procrustes analyses verify the observations, where PCAngsd has a RMSD of 0.00182 compared to ngsTools with a RMSD of 0.0075.

The admixture plots are displayed in Figure 4. was is not able to outperform NGSadmix in terms of accuracy; however, it is still able to estimate a very similar result. PCAngsd has some issues with noise in its estimation, but is, however, able to reduce it with the use of the sparseness parameter, \( \alpha = 1500 \). The likelihood measure in Equation 19 was used to easily find an optimal \( \alpha \), as seen in Figure S10. PCAngsd estimates the admixture proportions with a RMSD of 0.0108 compared to NGSadmix with a RMSD of 0.007148. The average runtime for 193 individuals and 8 million SNPs using \( K = 4 \) was 27.3 min for PCAngsd, and 7.1 hr for NGSadmix, making PCAngsd \( >15 \times \) faster than NGSadmix while both performing PCA and estimating admixture proportions.

### Waterbuck

Lastly, we analyzed the low-depth whole genome sequencing waterbuck dataset consisting of 73 individuals from five localities. The MAP test reported four significant principal components explaining the genetic variation in the dataset, which also fits with having five distinct waterbuck sampling sites. The PCA plots are visualized in Figure 5, where the top four principal components for each method are plotted. Once again, PCAngsd is able to cluster the populations much better than the ngsTools model; however, the effect is not as apparent as for the other datasets. Interestingly, populations can switch positions between the two methods, as seen with Samole on the second principal component, and Samburu and Matetsi on the third principal component.

As a few clusters are not so well defined, they will affect the admixture plots seen in Figure 6, where the increased level of noise is hard to remove without also affecting the true ancestry signals. Still, PCAngsd is capturing the same ancestry signals as NGSadmix with the use of the sparseness parameter. It is worth noting that an admixed individual of Ugalla and QENP was captured in both PCA and admixture estimation of PCAngsd, as also verified by the NGSadmix method. The runtime for the waterbuck dataset consisting of 73 samples and 9.4 million SNPs using \( K = 5 \) was an average of 14.5 min for PCAngsd, while NGSadmix had an average runtime of 3.2 hr, thus making PCAngsd \( >13 \times \) faster.

### Naively called genotypes

We used inferred population structure from naively called genotypes of the simulated and 1000 Genomes datasets, and the results are visualized in Figures S7 and S8. Genotypes were called by choosing the genotypes with the highest genotype likelihoods. No filters were applied in the genotype calling, since Skotte et al. (2013) showed that naively called genotypes had higher accuracy of inferred admixture proportion when no filters were used. The Procrustes analyses report RMSD values of 0.0123 and 0.00310 for performing PCA on the simulated and the 1000 Genomes dataset, respectively (cf. RMSD values of 0.00121 and 0.00182 using PCAngsd). Here, the naively called genotypes performed slightly better.
than ngsTools in both cases, but the results were still biased by sequencing depth. ADMIXTURE estimates admixture proportions from the called genotypes, with RMSD values of 0.00995 and 0.00865 for the two datasets, respectively, thus performing slightly better than PCAngsd for the 1000 Genomes dataset.

Discussion

We have presented two methods for inferring population structure and admixture proportions in low-depth NGS data, and both methods have been implemented in a framework named PCAngsd. We developed a method to iteratively estimate individual allele frequencies based on PCA using genotype likelihoods in a heuristic approach. We connected principal components to admixture proportions such that we are able to infer and estimate both in a very fast approach, making it feasible to analyze large datasets.

Based on the results when inferring population structure using PCA, it is clear that the increased uncertainty of low-depth sequencing data biases the clustering of populations using the ngsTools model, which also takes genotype uncertainty into account. Contrary to PCAngsd, population structure is not taken into account when using the posterior genotype probabilities to estimate the covariance matrix. The ngsTools model uses population allele frequencies as prior information for all individuals, such that individuals are assumed to be sampled from a homogeneous population. This assumption is, of course, violated when individuals are sampled from structured populations with diverse ancestries. Missing data are therefore modeled by population allele frequencies that resemble an average across the entire sample, which is similar to setting standardized genotypes to 0 in the estimation of the covariance matrix for genotype data. As an effect of this, the low-depth individuals are modeled by sequencing depth instead of population structure. These
results may lead to misinterpretations of population structure or admixture only due to low and variable sequencing depth. But the bias is not seen for individuals with equal sequencing depth, as shown in Figure S4 for the ngsTools model. Here, all individuals have been simulated with an average sequencing depth of 2.5×, such that individuals will inherit approximately the same amount of missing data. However, PCAngsd is able to overcome the observed bias of low and variable sequencing depth by using individual allele frequencies as prior information, which leads to more accurate results in all datasets of the study, as missing data are modeled accounting for inferred population structure. The assumption of conditional independence between individuals in the estimation of the covariance matrix (Equation 10) also holds for structured populations by conditioning on individual allele frequencies.

The number of significant eigenvectors used in the estimation of individual allele frequencies is determined by the MAP test. The MAP test is performed on the covariance matrix estimated from the ngsTools model. Thus, in cases of complex population structure, and low and variable sequencing depth, it is possible that the MAP test will not find a suitable number of significant eigenvectors to represent the genetic variation of the dataset. It could, therefore, be more relevant to use prior information regarding the number of eigenvectors needed for the dataset instead. However, for each of the cases analyzed in this study, the MAP test inferred the expected number of significant eigenvectors to describe the population structure.

PCAngsd is able to approximate the results of NGSadmix to a high degree when estimating admixture proportions using solely the estimated individual allele frequencies. However, although PCAngsd is not able to outperform NGSadmix in terms of accuracy, it is able to capture the exact same ancestry patterns as the clustering-based methods in a much faster approach, as shown by the runtimes of each method. Another advantage of PCAngsd is that the estimated individual allele frequencies need to be computed only once for a specific $K$, thus multiple different random seeds can be tested in the same run for an even greater speed advantage over NGSadmix, as the iterative estimation of individual allele frequencies is the most computational expensive step in PCAngsd. A proper $\alpha$ value, controlling the sparseness enforced in the estimated admixture proportions, can also be found through an automated scan implemented in our framework based on the likelihood measure of NGSadmix. PCAngsd is therefore an appealing alternative for estimating admixture proportions for low-depth NGS data as convergence and runtime can be a problem for a large number of parameters in NGSadmix.

PCAngsd was only seen to converge to a single solution for all our practical tests, where we used five batches for all analyses ($B = 5$).

Both methods of the PCAngsd framework rely on a representative set of individual allele frequencies, which we model using the inferred principal components of the SVD on the genotype dosages. The number of individuals representing each population or subpopulation is essential for inferring principal components that describe true population structure, as each individual will contribute to the construction of these axes of genetic variation. This particular effect can be seen in the PCA results of the waterbuck dataset where the populations are described only by a low number of individuals, such that some of the clusters are not as well defined as for the other datasets. The admixture proportions estimated from the waterbuck dataset are therefore affected as well, which can be seen by the additional noise in the admixture plots.

The PCAngsd framework may be able to push the lower boundaries of sequencing depth required to perform population genetic analyses on NGS data in large-scale genetic studies. This is also demonstrated by downsampling the...
1000 Genomes dataset in Figures S5 and S6, which display the robustness of PCAngsd in fairly low sequencing depth. However when down-sampling to only 1% of the reads, the PCA and admixture results become very noisy. PCAngsd also demonstrates an effective approach for dealing with merged datasets of various sequencing depths, as missing data will be modeled by population structure. Further, the estimated individual allele frequencies open up the development and

Figure 5 PCA plots of the top four principal components for the waterbuck dataset with 73 individuals and 9.4 million variable sites. The first row displays the plots of the first and second principal components for PCAngsd and the ngsTools model, respectively, while the second row displays the plots of the third and fourth principal components.

Figure 6 Admixture plots for $K = 5$ of the waterbuck dataset where each bar represents a single individual and the different colors reflect each of the $K$ components. The first plot is the admixture proportions estimated in PCAngsd with parameter $\alpha = 5000$, and the second plot shows the admixture proportions estimated in NGSadmix.
extension of population genetic models based on a similar probabilistic framework, such that population structure can be taken into account in heterogeneous populations.

Acknowledgments

This project was funded by the Lundbeck foundation (R215-2015-4174).

Literature Cited


Communicating editor: J. Novembre
Inferring Population Structure and Admixture Proportions in Low Depth NGS Data

Jonas Meisner & Anders Albrechtsen

Supplementary Material

Genotype likelihoods

Genotype likelihoods are the probability of the observed sequencing data given the unobserved genotypes. They can be computed from next-generation sequencing (NGS) data using the uncertainty of each base from the raw quality scores of sequencing machines. The base quality score $\text{baseQ}$ is usually in Phred scale such that the probability of an error in the observed base call is given by $\epsilon = 10^{-\text{baseQ}/10}$. The probability of observing a base $b$ of read $r$ in a site $s$ can be seen as the likelihood of the given allele. For having $L$ reads covering $s$ and assuming independence between the reads (and the error probabilities), the genotype likelihood can be computed by the product of the allelic likelihoods for the site $[1, 2]$. The genotype likelihood for individual $i$ in site $s$ can be defined as follows for a multi-allelic case derived from the approach in [3]:

$$P(X_{is} | G = A_1A_2) \propto \prod_{r=1}^{L} \left( \frac{P(b_r^{(i)} | A_1)}{2} + \frac{P(b_r^{(i)} | A_2)}{2} \right).$$

(1)

Here $X_{is}$ is the sequencing data, $P(b | A)$ = $1 - \epsilon$, for $b = A$, and $P(b | A) = \frac{\epsilon}{3}$, for $b \neq A$, with $\epsilon$ being the probability of error in the observed base call. This is for an arbitrary genotype $A_1A_2$.

Population allele frequencies

The population allele frequencies $p$ can be estimated from NGS data using an Expectation Maximization (EM) algorithm to compute the maximum likelihood estimator for each site. The likelihood function of $p$ in a site $s$ is defined in Kim et al. (2011) [4] as follows by assuming independence between all $n$ individuals:

$$\mathcal{L}(p_s) = P(X_s | p_s) \propto \prod_{i=1}^{n} P(X_{is} | p_s).$$

(2)

Here $X_s$ is the observed sequencing data in site $s$. Since the genotype is not observed for NGS data, a latent variable $G$ is introduced by taking the sum over the possible genotypes. Thus for individual $i$ in site $s$, $P(X_{is} | p_s)$ can now be defined as:
\[ P(X_{is} | p_s) = \sum_{g=0}^{2} P(X_{is} | G = g)P(G = g | p_s), \quad (3) \]

where \( P(X_{is} | G_{is} = g) \) is the genotype likelihood and \( P(G_{is} = g | p_s) \) is the genotype probability. By assuming Hardy-Weinberg equilibrium (HWE) in the whole sample, the genotype probabilities are estimated as \( P(G_{is} = 0 | p_s) = (1 - p_s)^2 \), \( P(G_{is} = 1 | p_s) = 2p_s(1 - p_s) \) and \( P(G_{is} = 2 | p_s) = p_s^2 \). The maximum likelihood estimator of \( p_s \) is then defined as follows:

\[ \hat{p}_s^{(ML)} = \arg\max_{p_s} \prod_{i=1}^{n} P(X_{is} | p_s). \quad (4) \]

The maximum likelihood solution is found by estimating the mean posterior expectations of the latent variable \( G \) iteratively for all individuals. The posterior genotype probability for individual \( i \) in site \( s \) is given as:

\[ P(G_{is} = g | X_{is}, \hat{p}_s) = \frac{P(X_{is} | G_{is} = g)P(G_{is} = g | p_s)}{\sum_{g' = 0}^{2} P(X_{is} | G_{is} = g')P(G_{is} = g' | p_s)}. \quad (5) \]

And the posterior expectation of the genotype is then given as:

\[ \mathbb{E}[G_{is} | X_{is}, \hat{p}_s] = \sum_{g=0}^{2} gP(G_{is} = g | X_{is}, \hat{p}_s). \quad (6) \]

Now the update step for iteration \( t + 1 \) in the EM algorithm can be defined as the mean of the posterior expectations of the genotype. The population allele frequency for each site is then obtained by scaling with 2 based on an assumption of \( G \) being Binomial distributed (\( \mathbb{E}[G] = 2p \)):

\[ \hat{p}_s^{(t+1)} = \frac{\sum_{i=1}^{m} \mathbb{E}[G | X_{is}, \hat{p}_s^{(t)}]}{2m}. \quad (7) \]
PCA - Sequencing depth

Simulated dataset

Figure S1: PCA plots of the simulated dataset as in Figure 1 but with individuals colored by their individual sampled sequencing depth. The upper PCA plot is of PCAngsd and the bottom is of the ngsTools model.
1000 Genomes dataset

**Figure S2:** PCA plots of the 1000 Genomes dataset as in Figure 3 but with individuals colored by their individual sequencing depth. The upper PCA plot is of PCAngsd and the bottom is of the ngsTools model. The sequencing depths are estimated in ANGSD [5].
Waterbuck dataset

Figure S3: PCA plots of the waterbuck dataset as in Figure 6 but with individuals colored by their individual sequencing depth. The PCA plots of the left column are of PCAngsd and the plots of the right column are of the ngsTools model. The sequencing depths are estimated in ANGSD [5].
Simulated dataset - Equal sequencing depth (2.5X)

Figure S4: PCA plots of a simulated dataset consisting of 380 individuals and 0.4 million variable sites. Each individual has been simulated with a sequencing depth of 2.5X in the same way as described in the Material and Methods section. The upper left plot shows the PCA performed on the known genotypes using equation 2, the upper right plot shows the PCA performed using PCAngsd and the bottom plot is the PCA performed using ngsTools model. Procrustes analyses showed a RMSD value of 0.000508 for PCAngsd and 0.000721 for the ngsTools model when compared to the inferred principal components of the genotypes.
**Downsampling 1000 Genomes dataset**

The 1000 Genomes dataset has been downsampled at different rates $S$ using ANGSD [5] to test the robustness of PCAngsd. $S$ represents the fraction of sequencing reads kept in the estimation of genotype likelihoods, and here $S = 0.25, 0.10, 0.05, 0.01$ have been tested. The same filters have been used to call SNPs and generate genotype likelihoods for all four downsampling rates in ANGSD, which will have an effect when the downsampling rate becomes small as seen below for the number of variable sites evaluated. The filters applied can be seen in the command-line example for ANGSD below. The four different downsampling rates yield the following sequencing depths:

- $S = 0.25$ : $0.42 - 3.23X$ (7.5 million sites)
- $S = 0.10$ : $0.18 - 1.38X$ (6.5 million sites)
- $S = 0.05$ : $0.11 - 0.79X$ (3.4 million sites)
- $S = 0.01$ : $0.23 - 3.30X$ (2412 sites)

**PCA plots**

<table>
<thead>
<tr>
<th>$S$</th>
<th>Sequencing Depth</th>
<th>Number of Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.42 - 3.23X</td>
<td>7.5 million</td>
</tr>
<tr>
<td>0.10</td>
<td>0.18 - 1.38X</td>
<td>6.5 million</td>
</tr>
<tr>
<td>0.05</td>
<td>0.11 - 0.79X</td>
<td>3.4 million</td>
</tr>
<tr>
<td>0.01</td>
<td>0.23 - 3.30X</td>
<td>2412 sites</td>
</tr>
</tbody>
</table>

**Figure S5:** PCA plots of the downsampled 1000 Genomes datasets performed using PCAngsd. The upper left plot is performed with a downsampling rate of $S = 0.25$, the upper right plot is performed with a downsampling rate of 0.10, the bottom left plot is performed with a downsampling rate of 0.05 and the bottom right is performed with a downsampling rate of 0.01. Procrustes analyses reported RMSD values of 0.00292, 0.00356, 0.00399 and 0.0827 for the four plots, respectively, in comparison to the PCA performed on the reliable genotypes using equation 2.
Admixture proportions

Figure S6: Admixture plots for $K = 4$ of the downsampled versions of the 1000 Genomes dataset performed using PCAngsd. The upper plot shows the admixture plot with a downsampling rate of $S = 0.25$ and $\alpha = 815$, the second plot shows the admixture plot with a downsampling rate of 0.10 and $\alpha = 665$, the third plot is the admixture plot with a downsampling rate of 0.05 and $\alpha = 555$ and the bottom plot is the admixture plot with a downsampling rate of 0.01 and $\alpha = 15$. The RMSD values are 0.0173, 0.0331, 0.0530 and 0.347 for the four plots, respectively, in comparison to the admixture proportions estimated using ADMIXTURE [6] from the reliable genotypes.
Naive genotype calling

Genotypes have naively been called from genotype likelihoods with a uniform prior such that the calling is based on the highest genotype likelihood. Skotte et al. [7] showed that genotypes called from the genotype likelihoods performed better than genotypes called from posterior genotype probabilities using the population allele frequencies as prior for inferring population structure in low depth scenarios for samples of diverse ancestry. Here we show the population structure inferences from the called genotypes in the simulated and 1000 Genomes datasets including a downsampling scenario for $S = 0.05$ as seen above.

![Figure S7: PCA plots performed on the called genotypes from genotype likelihoods with a uniform prior. The upper left plot is performed on the simulated dataset, the upper right plot is performed on the 1000 Genomes dataset and the bottom plot is performed on the downsampling 1000 Genomes dataset using $S = 0.05$. Procrustes analyses showed RMSD values of 0.0123, 0.00310 and 0.0296, respectively, in comparison to the PCA performed on the known genotypes of both datasets (compared to RMSD values of 0.00121, 0.00182 and 0.00399 using PCAnsgd).]
Figure S8: Admixture plots estimated using ADMIXTURE [6] on the called genotypes from genotype likelihoods with a uniform prior. The top plot is estimated from the simulated dataset, the middle plot is estimated from the 1000 Genomes dataset and the bottom plot is performed on the downsampled 1000 Genomes dataset using $S = 0.05$. The RMSD values are 0.00995, 0.00865 and 0.0994 for the three plots, respectively, in comparison to the admixture proportions estimated from the genotypes of both datasets (compared to RMSD values of 0.00476, 0.0108 and 0.0530 using PCAngsd).
Individual allele frequencies

Figure S9: Smooth scatterplot of the frequencies from which the simulated genotypes have been sampled from against the individual allele frequencies estimated using PCAngsd. Outliers are highlighted with black dots.
NMF $\alpha$ parameter

**Figure S10**: Combined plots of the Frobenius error and likelihood measure obtained using different $\alpha$ values in the estimation of admixture proportions for the real datasets. The left figure shows the plot for the 1000 Genomes dataset with an optimal $\alpha = 1500$ in terms of maximizing the likelihood measure. The right figure shows the same for the waterbuck dataset with an optimal $\alpha = 5000$. $B = 5$ was used in both cases.

**Command-line examples**

**ANGSD**

Call SNPs and estimate genotype likelihoods for the low coverage 1000 Genomes dataset using ANGSD [5].

```
./angsd -bam 1000g.bamlist -GL 1 -out 1000g_GL -doGlf 2 -doMajorMinor 1 -doMaf 2 -minMaf 0.05 -SNP_pval 1e-6 -minQ 20 -minMapQ 30 -skipTriallelic 1 -minInd 50 -rf chrFile -doDepth 1 -doCounts 1 -P 20
```

**PCAngsd**

Perform PCA and estimate admixture proportions for the genotype likelihoods of the low coverage 1000 Genomes dataset using PCAngsd.

```
python pcangsd.py -beagle 1000g_GL.beagle.gz -o 1000g_pcangsd -threads 20 -admix -admix_alpha 1500
```

**References**


Large-scale Inference of Population Structure in Presence of Missingness using PCA

By
Jonas Meisner\textsuperscript{1}, Siyang Liu\textsuperscript{2}, Mingxi Huang\textsuperscript{2} and Anders Albrechtsen\textsuperscript{1}

\textsuperscript{1}The Bioinformatics Centre, Department of Biology, University of Copenhagen, DK-2200, Denmark
\textsuperscript{2}BGI-Shenzhen, Shenzhen, 518083, China

Publication details
In second round of review in Bioinformatics
Genetic and population analysis

Large-scale Inference of Population Structure in Presence of Missingness using PCA

Jonas Meisner¹,*, Siyang Liu², Mingxi Huang² and Anders Albrechtsen¹

¹Department of Biology, University of Copenhagen, Copenhagen, DK-2200, Denmark and
²BGI-Shenzhen, Shenzhen, 518083, China

*To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Principal component analysis (PCA) is a commonly used tool in genetics to capture and visualize population structure. Due to technological advances in sequencing, such as the widely used non-invasive prenatal test, massive datasets of ultra-low coverage sequencing are being generated. These datasets are characterized by having a large amount of missing genotype information.

Results: We present EMU, a method for inferring population structure in the presence of rampant non-random missingness. We show through simulations that several commonly used PCA methods can not handle missing data arisen from various sources, which leads to biased results as individuals are projected into the PC space based on their amount of missingness. In terms of accuracy, EMU outperforms an existing method that also accommodates missingness while being competitively fast. We further tested EMU on around 100K individuals of the Phase 1 dataset of the Chinese Millionome Project, that were shallowly sequenced to around 0.08x. From this data we are able to capture the population structure of the Han Chinese and to reproduce previous analysis in a matter of CPU hours instead of CPU years. EMU’s capability to accurately infer population structure in the presence of missingness will be of increasing importance with the rising number of large-scale genetic datasets.

Availability: EMU is written in Python and is freely available at https://github.com/rosemeis/emu.

Contact: jonas.meisner@bio.ku.dk

1 Introduction

The advent of whole-genome sequencing technologies has brought the opportunity of generating large amount of genomic data at low cost (Metzker (2010)). Large-scale sequencing studies are therefore becoming more prevalent (Consortium et al. (2015a,b); Francioli et al. (2014); Gudbjartsson et al. (2015); Liu et al. (2018)) as they help researchers understand genetic variation in populations on a much broader scale than previously possible using genotyping arrays. A cost-effective strategy with the ever-increasing demand for larger sample sizes seems to advocate for the use of medium or low coverage sequencing (Nielsen et al. (2011); Li (2011)). Larger sample sizes sequenced at lower depths will generally lead to better population-scale estimates of genetic variation compared to sequencing at higher depths at the cost of limited sample sizes (Fumagalli (2013)). With this appealing trade-off, a genomic study (Liu et al. (2018)) was recently conducted on ultra-low coverage sequencing data of 141K Chinese pregnant women as part of the Chinese Millionome Project. The participants in the study underwent a non-invasive prenatal test (NIPT) which is common for testing fetal chromosomal abnormalities by shallowly sequencing the cell-free DNA from the maternal plasma. The study provided insight into the genetic structure and history of the Chinese population as well as performing genome-wide association studies (GWAS) with principal components as covariates. The study had an average depth of < 0.1X, which allowed for the much larger sample size compared to other sequencing projects.

These large-scale sequencing studies will usually consist of individuals from diverse ancestries and may include cryptic structure not accounted for. Population structure plays a major role in population genetics for understanding population demography (Patterson et al. (2006)), as well as in association studies where it acts as a confounding factor and must be accounted for (Marchini et al. (2004); Price et al. (2006)). One approach for
inferring population structure is based on the use of principal component analysis (PCA). PCA has the appealing feature of projecting individuals onto inferred axes of genetic variation that capture population structure in a continuous fashion. The standard way to infer population structure using PCA has been to construct a genetic relationship matrix (GRM) and perform eigen decomposition on this matrix to infer the axes of genetic variation (Patterson et al. (2006)). However as the sample size in large scale studies is constantly increasing, faster and more scalable methods based on various low-rank approximations (Lehoucq et al. (1998); Halko et al. (2011)) have been developed to only infer the top axes of genetic variation almost directly from the genotype matrix (Meisner and Albrechtsen (2018); Galinsky et al. (2016); Abraham et al. (2017)). The problem for the majority of these methods is that they cannot handle missing data in an appropriate manner.

Common approaches for dealing with missingness when inferring population structure are either to thin the dataset by removing sites or individuals with missingness rates above a certain threshold or to simply ignore the presence of missingness by using the mean genotype value (Patterson et al. (2006); Galinsky et al. (2016); Abraham et al. (2017)), also called mean imputation. The problem with the first approach is that one would lose information that could potentially be crucial in downstream analyses, and especially when merging datasets of different sources. The problem with second approach is that the inferred axes will correlate with the amount of missingness and thus no longer only represent population structure. This is due to missingness being modelled as the average across the entire dataset.

We therefore propose a new method that is specially designed to deal with large-scale genetic datasets with very high levels of missingness using a novel accelerated approach. EMU (EM-PCA for Ultra-low Coverage Sequencing Data) is an accelerated expectation-maximization (EM) algorithm for PCA to model the missingness in an iterative fashion. The concept of an iterative PCA for imputing missing values is not novel and has been formally described (Kiers (1997); Josse and Husson (2012)). A similar method (Meisner and Albrechtsen (2018)) has also been developed for low coverage sequencing data on the basis of genotype likelihoods which however is not optimal for ultra-low coverage sequencing data, where individuals almost never have more than one read at a locus. We apply and show that our method displays high accuracy, robustness and scalability on both simulated and real data with very high missingness, where we apply it to around 100K non-invasive prenatal test data (Liu et al. (2018)). In relation to missingness patterns, we also demonstrate that EMU is robust to having different SNP ascertainment schemes in a dataset as would be a result of merging different data types. Additionally, we compare our method in terms of accuracy, computational speed and memory usage against other popular choices of methods for inferring population structure on the basis of PCA.

2 Methods

We will now describe our method EMU for inferring population structure in ultra-low coverage sequencing data. As we assume to be working on datasets with extensive amount of missingness for ultra-low coverage sequencing data (≤ 0.1X), we use a single-read sampling approach to best describe our data in a similar fashion to (Liu et al. (2018)). This means that we expect on average ≤ 0.1 sequencing reads to map to a given position in the covered genome on average.

We therefore define a data matrix $D$ with its entries representing the output of a single-read sampling approach for observed sequencing reads in $n$ individuals and $m$ variable sites. Thus for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, $d_{ij}$ can take values from $\{0, 1, -9\}$ where 0 and 1 are the sampling of the major and minor allele respectively, while $-9$ refers to missing data for the given individual $i$ in the given site $j$. We thereby assume that all sites are diallelic and that both major and minor alleles are known and we ignore sequencing errors. The sampling of an allele can therefore be seen as a Bernoulli process. EMU is also capable of working with diploid genotype data, where $d_{ij}$ can take values from $\{0, 0.5, 1, -9\}$ such that heterozygous genotype information is kept, but here we describe it for pseudo-haploid data (single-read sampling).

2.1 Population allele frequencies

The allele frequency across all samples in the dataset $\hat{f}_j$ is estimated as follows for a single site $j$ by counting the observed number of minor alleles.

We denote this as the population allele frequency:

$$\hat{f}_j = \frac{\sum_{i=1}^{n} I[d_{ij} = 1]}{\sum_{i=1}^{n} I[d_{ij} = 0] + \sum_{i=1}^{n} I[d_{ij} = 1]},$$

where $I[X]$ being the indicator function. I.e. the allele frequencies are calculated such that individuals with missing information are not counted for the given site. For diploid genotype data, we also count the heterozygous individuals ($d_{ij} = 0.5$) in both numerator and denominator.

2.2 Individual allele frequencies

Pritchard et al. (Pritchard et al. (2000)) introduced the concept of individual allele frequencies in STRUCTURE for genotype data. Under the assumption of a fixed number of ancestral populations, admixture proportions $Q$ and ancestral allele frequencies $P$ are inferred that when multiplied represent individual allele frequencies, $\Pi = QF^T$. More recently Hao et al. (Hao et al. (2015)) constructed a similar approach for genotype data, instead based on PCA, where the top principal components are used to reconstruct the genotype matrix, and the individual allele frequencies can be derived from a low-rank approximation. Several methods have applied this idea (Meisner and Albrechtsen (2018); Conomos et al. (2016); Meisner and Albrechtsen (2019)). In this study, we present an iterative variant for pseudo-haploid and genotype data that accounts for missingness.

For individual $i$ at site $j$ the individual allele frequency $\pi_{ij}$ can be seen as the underlying parameter in the binomial sampling process of genotype $g_{ij}$, conditioned on the population structure.

2.3 Iterative PCA

EMU is based on the iterative PCA algorithm of Kiers (1997) (EM-PCA) which deals with finding a low-rank approximation of $D$ iteratively, where missing values are imputed by reconstruction from the estimated low-rank approximation of the previous iteration using singular value decomposition (SVD). We use a low value $K$ as the rank of this approximation and this value should ideally reflect the number of ancestral populations from which the individuals are derived. When choosing $K$ as the rank we indirectly model both the admixture proportions for each individual and the $K + 1$ ancestral allele frequency for each site in order to predict the individual genotypes (Engelhardt and Stephens (2010); Meisner and Albrechtsen (2018)). This iterative procedure corresponds to an EM algorithm and it is equivalent to finding the matrix of individual allele frequencies that minimize the expression (Josse and Husson (2012))

$$\min_{\Pi} \|\mathbf{C} \odot (\mathbf{D} - \Pi)\|_F^2.$$  

(2)

Here $\mathbf{C}$ is a weight matrix with entries such that $c_{ij} = 0$ if $d_{ij} = -9$ and $c_{ij} = 1$ otherwise for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, while $\odot$ represents element-wise multiplication. Thus, only entries with information are evaluated. However, note that $\Pi$ will be estimated from the full dataset. The iterative procedure and its updating scheme to estimate the individual allele frequencies are described in Algorithm 1, where
missing values are initialized as the population allele frequencies $\hat{f}$ and $K$ dimensions are used in the low-rank approximation.

We define convergence as when the root-mean-square deviation (RMSD) of $U_{[1:K]}^{(t)}$ between two successive iterations of the iterative procedure is less than some small value $\epsilon = 5 \times 10^{-7}$.

$$\text{RMSD} \left( U_{[1:K]}^{(t)} , U_{[1:K]}^{(t+1)} \right) = \sqrt{ \frac{1}{mK} \sum_{j=1}^{m} \sum_{k=1}^{K} (u_{jk}^{(t)} - u_{jk}^{(t+1)})^2 }.$$  \hspace{1cm} (3)

**Algorithm 1: EM-PCA in EMU**

**Input:** $D$, $C$, $\hat{f}$

1. Initialize centered $E$ matrix, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$:
   - Set $e_{ij} = d_{ij} - f_{j}$, if $c_{ij} = 1$.
   - Set $e_{ij} = 0$, if $c_{ij} = 0$.
2. SVD on centered matrix: $E = W S U^T$.
3. Individual allele frequencies: $\hat{f} = W_{[1:K]} S_{[1:K]} U_{[1:K]}^T + \hat{f}$, where $\hat{f}$ is added row-wise, and $\hat{f}$ is mapped to domain $[0, 1]$.
4. Update centered $E$ matrix, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$:
   - Set $e_{ij} = d_{ij} - f_{j}$, if $c_{ij} = 1$.
   - Set $e_{ij} = e_{ij} - f_{j}$, if $c_{ij} = 0$.
5. Repeat step 2, 3 and 4 until convergence.

**Output:** $\hat{f}$

After obtaining the final set of individual allele frequencies that minimizes Equation 2 from our EM-PCA algorithm, we can infer the population structure from a standardized matrix such that the variable sites are weighted by their population allele frequencies, $\hat{f}$ (Patterson et al. (2006)). Thus, define the standardized matrix $X$ with entries:

$$x_{ij} = \frac{-e_{ij}}{\sqrt{f_{j}(1-f_{j})}},$$  \hspace{1cm} (4)

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ with $e_{ij}$ being defined as step 4 in the final iteration of Algorithm 1. Performing SVD on the standardized matrix will infer and map individuals onto axes of genetic variation that represents population structure (principal components):

$$X = V S P^T.$$  \hspace{1cm} (5)

Here $V$ is the principal components capturing population structure, which are identical to the principal components obtained from performing full eigendecomposition on the genetic relationship matrix constructed using the same standardized matrix, $X$ (Patterson et al. (2006); Engelhardt and Stephens (2010)).

### 2.5 Implementation

EMU has been implemented in Python using Numpy (Van Der Walt et al. (2011)) data structures and Cython (Behnel et al. (2011)) for parallelization and to speed up computational bottlenecks. We are using a truncated SVD implementation of the scikit-learn library (Pedregosa et al. (2011)) that uses a randomized PCA procedure (Halko et al. (2011)) to only compute the $K$ largest eigenvectors of a given matrix. We have also extended our method to work on diploid genotypes such that the information of heterozygous sites is retained.

The code is freely available at https://github.com/Rosemies/emu and works using both Python 2.7 and 3.7.

The computational complexity for one iteration in our EM-PCA algorithm will be $O(nrm)$ for low-rank approximations using the truncated SVD procedure, and the number of iterations will depend on the amount of missingness and number of samples and variable sites in the dataset. In total, our algorithm will have a memory requirement of $O(nrm)$ bytes. However in modern large-scale datasets, the constant to this bound will be important for actual applications. A more detailed description of the memory usage is described in the supplementary material. The algorithm is linear in both the number of samples and the number of sites for both computational speed and memory usage.

We have also implemented an alternative variant of the iterative update in our algorithm that is much more memory efficient by slightly sacrificing computational speed. This variant uses $\sim 20x$ less memory than the previously described procedure by using a lazy evaluation approach for $E$ in custom matrix multiplications, based on the same randomized SVD procedure, such that only the low-rank factor matrices of $E$ are considered. A description of this variant can also be found in the supplementary material and we will regard to this variant as EMU-mem in the main results.

### 2.6 Simulation of single-read data matrix

We have simulated genotype data to test the capabilities of our method. The simulations are based on allele frequencies from three Chinese populations (Han, Uygur and Dai) generated from genotyped individuals (Lazaridis et al. (2014)) of the Human Genome Diversity Project (Cann et al. (2002)). These populations were selected to make the scenario somewhat similar to the real data used in this study with assumed low genetic differentiation between populations. We simulate individuals under a Binomial model, such that a sampled genotype $g_{ij}$ can take values of 0, 1, and 2, representing the minor allele count for individual $i$ at site $j$. We then transform the simulated genotype matrix into a single-read sampling matrix (pseudo-haploid) using the following scheme for individual $i$ at site $j$. A homozygous genotype ($g_{ij} \in \{0, 2\}$) is directly converted to either 0 or 1, respectively, while a heterozygous genotype ($g_{ij} = 1$) is simply converted to either 0 or 1 with equal probability in the single-read sampling matrix. To test and showcase our method in various cases of extreme missingness, we have simulated three different scenarios of missingness as well as five additional scenarios, related to emulating different SNP ascertainment schemes. In all simulated scenarios, the number of variable sites is $\sim 350K$ after filtering out rare variants based on a minor allele frequency threshold (3%). To simplify the different scenarios for the reader, we have provided a graphical overview of the simulation procedures in the supplementary material (Figure S1).
2.6.1 Missingness rate scenarios
The three scenarios for different degrees of missingness have been simulated with a total of 900 individuals (250 from each distinct population and 150 being admixed) and a total of 9000 individuals (2500 from each distinct population and 1500 being admixed) to evaluate the effect of sample size.

In Scenario 1, we have simulated individuals with a randomly assigned missingness rate from 5% to 50%. This means that an individual i with a missingness rate of 50% for example would have a probability of 0.5 to keep the sampled allele at a given site \( j \) \((d_{ij} \in \{0, 1\})\) and a probability of 0.5 to remove the sampled information \((d_{ij} = -9)\).

For Scenario 2, we have made the missingness rate gradient more extreme such that the simulated individuals are randomly assigned a missingness rate between 90% and 99%. We have also used this scenario to perform computational tests regarding speed, memory and accuracy for much larger sample sizes later on.

And last in Scenario 3, we have simulated half of the individuals from the three distinct population with a missingness rate sampled from \(\mathcal{N}(0.95, 0.001)\) (≈ 95%). The other half and all admixed individuals have been simulated with a missingness rate sampled from \(\mathcal{N}(0.5, 0.01)\) (≈ 50%).

2.6.2 SNP ascertainment scheme scenarios
For the five scenarios emulating different SNP ascertainment schemes, we have simulated 750 individuals with 250 from each population such that admixed individuals have not been included.

In Scenario 4, we have simulated a scenario that tries to emulate the merging of a SNP array dataset with a whole-genome sequencing (WGS) dataset. Therefore, we have simulated half of the individuals in each of the three different populations to only retain information of very common variants (MAF ≥ 0.25) and set other variants as missing, while the other halves of the three populations have information from all simulated variants.

For Scenario 5, we have simulated half of the individuals in one of the three different populations (Han) to only retain information in approximately a third of the variants (≈130K), while the other half of the population retain information in the other two-thirds of the variants (≈233K). However, there is a small overlap of 13K between the two subsets. Thus, we create a scenario where one population mostly has to rely on the other two populations to estimate within-population correlations between its two halves. The other two populations (Uygur and Dai) are simulated with full information. The case is similar for Scenario 6, but here two populations (Han and Uygur) have been simulated such that half of their individuals has information in approximately a third of the variants while the other half has information in the other two-thirds of the variants with a slight overlap. The Dai population is simulated with full information.

Scenario 7 and 8 are almost identical to Scenario 5 and 6, however now there is no overlap between the halves of a population. This means that the correlation within a population between its two halves must solely rely on the correlation with the other populations.

The last four scenarios (5, 6, 7 and 8) tries to emulate cases where either different datasets of different SNP arrays have been merged, which may be almost or entirely non-overlapping, or when dealing with ultra-low coverage sequencing data.

2.7 Phase 1 of the Chinese Millionome Project
The Chinese Millionome Project (CMDP) aims at analyzing millions of Chinese sequencing genomes to understand the genetic diversity of the Chinese population and to promote the precision medicine initiatives in China. In the Phase 1 study (Liu et al. 2018), Liu et al. analysed the low coverage genomes of about 141K female participants that were recruited via the NIPT test during pregnancy. The individuals were sequenced at ultra-low depth with 5-10 million using either 38bp or 89bp single end reads, corresponding to an average sequencing depth of 0.08X. We restricted our analysis to sites that are known to be common (MAF ≥ 0.05) both in our project and in the East Asian populations of the 1000 Genomes Project and that had a sequencing depth ≥ 0.1X to remove outlier sites with potential mapping issues and that additionally provide very little information. We additionally kept individuals sequenced with 38bp reads and removed individuals with sequencing error rate greater than 0.00325 as done in the original study (Liu et al. 2018). This resulted in pseudo-haplotype genotype matrix of 97K individuals and 440K sites.

3 Results
For the simulated datasets, we test and compare EMU against other commonly used methods for inferring population structure using PCA. These include PLINK (version 2.0) (Chang et al. 2015), smartpca and FastPCA (Galinsky et al. (2016)) from the EIGENSOFT package (version 7.2.1) (Patterson et al. (2006)), FlashPCA (version 2.0) (Abraham et al. (2017)). PLINK and smartpca estimate the GRM and perform full eigendecomposition on this, while FastPCA and FlashPCA use low-rank approximation approaches on the standardized data matrix. From these methods, PLINK is the only other method besides EMU that accounts for missingness in the dataset, while the rest perform mean imputation. We also tested multidimensional scaling (MDS) from IBS distances and SNPRelate (Zheng et al. (2012)) for all simulated scenarios, however we did not include them in the main results. Their results are shown in Figure S15 and S16.

To assess the performance of our method in the various simulated scenarios, we perform Procrustes analyses (Dryden and Mardia (1998)). We infer population structure from the full genotype datasets, from which the single-read sampling matrices have been generated from, to serve as ground-truth in comparisons with all the tested methods. A Procrustes analysis will then find scaling and rotation components to best represent the inferred principal components of a tested method in comparison with the ground-truth and the RMSD is reported.

3.1 Simulation
In the following results, we have tested eight different scenarios of missingness. The first three scenarios are related to individuals having different missingness rates, while the last five scenarios are related to the merging of different datasets and using different SNP ascertainment schemes.

3.1.1 Missingness rate scenarios
There are two cases for each of the three scenario (A and B) and these distinguish the sample size of the simulated dataset. Case A has 900 individuals with 250 individuals sampled from each distinct population and 150 admixed individuals, while case B has 9000 individuals with 2500 individuals sampled from each distinct population and 1500 admixed. We only display the results of case A in the main results while the results of case B are displayed in the supplementary material.

The results of Scenario 1, where the missingness rate is between 5% – 50% for all individuals, are shown in Figure S2 and S3. It can be seen that the three methods, which are not accounting for missingness in the dataset (smartpca, FastPCA and FlashPCA), are struggling to separate the individuals from the three populations into distinct clusters in the presence of missingness and they produce almost identical results where the individuals with more missingness is closer to the origin (Figure S4). In contrast, both EMU and PLINK are able to infer the population structure accurately as also verified in the Procrustes analyses shown in Table 1.
Large-scale PCA in Presence of Missingness

For Scenario 2, the interval of the missingness rate was increased to 90 – 99% in order to simulate a more extreme scenario as also seen in the Chinese Millionome Project. The results are displayed in Figure 1 and S5 for 900 and 9000 individuals, respectively, and the 900 individuals colored by their missingness rate in Figure S6. Again, EMU and PLINK are once again able to infer the simulated population structure but EMU is able to recover a slightly more accurate PCA plot as shown in Table 1. It can also be seen that the estimates of PLINK for the individuals of the three populations are slightly more noisy than the estimates of EMU, and that these individuals generally have a higher missingness rate. Here FastPCA is not able to infer any meaningful structure in the dataset, while smartpca and FlashPCA finds some overall pattern but it is heavily biased by the missingness in the dataset.

In Scenario 3, the individuals sampled from the three population were simulated with two different settings of missingness and the results are displayed in Figure S7, S8 and S9. EMU and PLINK are capturing the population structure but PLINK still has more noisy estimates for the individuals of the three distinct populations. Due to the smaller variance in the missingness rate intervals, smartpca, FastPCA and FlashPCA are capturing the population structure accurately of the individuals simulated under one missingness setting. However, the individuals simulated under the other missingness setting now clusters together, which in particular illustrates the problem of not accounting for missingness as these clusters may be interpreted as separate populations.

We also tried to increase the number of principal components to $K = 4$ which corresponds to assuming 5 ancestral populations. This has only lowered the performance slightly (Table 1).

<table>
<thead>
<tr>
<th>Method</th>
<th>1A</th>
<th>1B</th>
<th>2A</th>
<th>2B</th>
<th>3A</th>
<th>3B</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMU</td>
<td>0.000644</td>
<td>0.000190</td>
<td>0.00390</td>
<td>0.00101</td>
<td>0.00197</td>
<td>0.000556</td>
</tr>
<tr>
<td>PLINK</td>
<td>0.000650</td>
<td>0.000192</td>
<td>0.00567</td>
<td>0.00110</td>
<td>0.00262</td>
<td>0.000622</td>
</tr>
<tr>
<td>smartpca</td>
<td>0.00897</td>
<td>0.00266</td>
<td>0.0299</td>
<td>0.00681</td>
<td>0.0292</td>
<td>0.00904</td>
</tr>
<tr>
<td>FastPCA</td>
<td>0.00897</td>
<td>0.00266</td>
<td>0.0420</td>
<td>0.00684</td>
<td>0.0292</td>
<td>0.00904</td>
</tr>
<tr>
<td>FlashPCA</td>
<td>0.00897</td>
<td>0.00266</td>
<td>0.0420</td>
<td>0.00684</td>
<td>0.0292</td>
<td>0.00904</td>
</tr>
<tr>
<td>EMU-mem</td>
<td>0.000879</td>
<td>0.000272</td>
<td>0.03407</td>
<td>0.00102</td>
<td>0.00207</td>
<td>0.000595</td>
</tr>
<tr>
<td>EMU (K=4)</td>
<td>0.000877</td>
<td>0.000271</td>
<td>0.03389</td>
<td>0.00101</td>
<td>0.00220</td>
<td>0.000601</td>
</tr>
</tbody>
</table>

**3.1.2 SNP Ascertainment scheme scenarios**

For the SNP ascertainment scenarios, we simulate 750 individuals from the three populations with 250 from each, thus excluding admixed individuals. We note that PLINK could not be run for Scenario 7 and 8 as it relies on pairwise estimates across all samples and can not run if there is no overlap in variants between a single pair of individuals. This would be a potential problem in its usage for ultra-low coverage sequencing data.

For Scenario 4, we are emulating the merging of whole-genome sequencing data and SNP array data by having half of the individuals in each of the three populations only have information in variants with a MAF $\geq 0.25$. The results are visualized in Figure S10. Once again the methods...
performed mean imputation are biased by the different missingness rate, which is now created by having different SNP ascertainment schemes instead of being random, while EMU and PLINK are able to infer correct population structure. As the simulated missingness is non-random, we also see that a method like MDS is failing at capturing population structure (Figure S15).

In Scenario 5, half of the Han population is simulated such that it only had a small overlap with the other half (13K), while the two other populations are simulated with full information. The results are shown in Figure S11. All methods performing mean imputation interpret the two halves of the Han population as two separate populations, while EMU and PLINK still infer correct population structure.

Scenario 6 are similar to Scenario 5, but now two of the populations are simulated with two different halves (Han and Uygur). The results are visualized in Figure S12. Even with almost no overlap between the halves within two of the populations, EMU and PLINK are able to infer correct population structure. Of course, it is seen that mean imputation methods are biased once again and they show the four halves as different clusters.

The last two SNP ascertainment scheme scenarios are almost simulated identically as 5 and 6 but with no overlap in the subsets of a population. We see almost the exact same results as for the previous two scenarios except that PLINK could not be run, and EMU is now failing to converge in the last scenario as there is no solution for the iterative procedure. The results are visualized in Figure S13 and S14 for Scenario 7 and 8, respectively. A lack of overlap can in many scenarios lead to non-identifiability issues which will cause EMU to not converge.

3.2 Computational tests

We have tested runtimes and memory usage of EMU for different simulated sample sizes and compared it to the other tested methods. The different datasets were simulated under the same settings as Scenario 2 with individual missingness rates between 90%−99%. The different sample sizes simulated were 900, 9000, 18000, 36000, 54000 and 90000. smart pca could not be run for sample sizes > 36000. All datasets have ~350K variable sites after filtering out rare variants with a threshold of 5%.

All analyses were performed server-side using 64 threads (2.10 GHz; Intel Xeon Gold 6152), and the results of the computational tests are summarized in Figure 2. It is clear that the methods based on low-rank approximation (EMU, FastPCA and FlashPCA) are faster than the methods that construct the GRM followed by eigendecomposition. However, the implementation in PLINK does keep up with the low-rank approximation approaches to a certain extent but its approach would become unfeasible for large sample sizes. To demonstrate this, we also performed curve-fitting of the runtimes for EMU and PLINK for comparison with more extreme sample sizes. The curves are shown in Figure S11. Here we can derive from extrapolation that it would take PLINK ~98 days to infer population structure for 1 million individuals, while it would only take ~4.1 hours for EMU. This is due to PLINK having to construct the GRM (O(n^2)) and additionally having to perform eigendecomposition on it.

It is noteworthy that the number of iterations performed in EMU decrease when the sample size is increased (see Figure S10). This is due to having more individuals contributing to the axes of genetic variation, such that the eigenspectrum of the top principal components representing population structure is increased and random fluctuations are proportionally decreased. Thus, it becomes easier for EMU to impute the missing values in a more well-defined PC space.

3.3 Analysis on the 96,800 ultra-low-pass genomes

We apply EMU and PLINK to analyse 96,880 NIPT ultra-low pass genomes (Liu et al. (2018)). The participants came from 31 provinces throughout mainland China. After performing individual and site filtering based on read length, sequencing error rate, minor allele frequency (≤ 0.05) and sequencing depth (≤ 0.1X), we obtain a genotype matrix consisting of 96,880 individuals and 440,183 sites. Again, both EMU and PLINK are run server-side with 64 threads under the same configurations as in the simulations. We use 3 eigenvectors to estimate individual allele
frequencies in EMU. The inferred population structure of EMU and PLINK are visualized in Figure 3 and S17, respectively. Runtime information of both methods is displayed in Table 2, where EMU is shown to be ~6.4x faster than PLINK. However due to the weak population structure in the Chinese individuals, EMU needs to perform 72 iterations that also naturally affects its runtime. As seen in the simulations, the results are similar but EMU seems to cluster the individuals better in comparison to PLINK.

In the results, PC1 captures the cline of genetic variation from North to South China, while the PC2 captures the cline from West to East which is less apparent. PC2 is mainly driven by individuals with no province information. Some of these individual reported their ethnicity which were all Uygur - the main ethnic group in most western province (Xinjiang).

### Table 2. Runtime and memory usage for dataset of 96,880 individuals and 440,813 sites.

<table>
<thead>
<tr>
<th>Method</th>
<th>Memory (GB)</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMU</td>
<td>198.8</td>
<td>72</td>
</tr>
<tr>
<td>PLINK</td>
<td>87.7</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### 4 Discussion

We have implemented a novel method, EMU, for inferring population structure in large-scale genetic studies that accounts for random and non-random missingness. It estimates individual allele frequencies in an iterative manner based on low-rank approximations to impute missing information in the dataset through an EM-PCA approach. We have shown that EMU outperforms other commonly used approaches in population genetics and it is very competitive in terms of runtime for large-scale datasets. EMU will also work for smaller datasets, however if there is a very large amount of missing data then EMU will need either more individuals or more sites to achieve a similar performance as having full information.

Through simulation studies, we demonstrated that the majority of existing methods does not account for missingness in a meaningful way, which introduces major biases in the results of these methods. This bias is even observed for low missingness rates. PLINK also corrects for missingness and was shown to infer accurate population structure. However its approach is not feasible for very large-scale genetic studies as it performs full eigendecomposition on the constructed GRM. By extrapolation from runtimes of computational tests, PLINK would need to run for ~98 days (compared to ~4.1 hours for EMU) to infer population structure for 1 million individuals under the tested configurations of this
study. We further note that EMU outperformed PLINK in every scenario regarding inference of principal components.

As previously mentioned, the results of smartpca, FlashPCA and FastPCA were almost seen to be identical for all the tested scenarios, however FastPCA deviated from the other two in Scenario 2A, 2B and 3A where it performed slightly worse. This is due to FastPCA, also based on random matrix theory, not performing any power iterations including normalization for stabilizing results as is done in the randomized PCA procedure (Halko et al. (2011)) (EMU) as well as methods based on ARPACK (Lehoucq et al. (1998)) (FlashPCA). This shows the importance of having stabilization steps in small noisy datasets, as the results of FlashPCA are similar to smartpca that performs full eigendecomposition on the GRM with mean imputation, and thus appear more robust.

We also applied EMU to the dataset of Phase 1 study of the Chinese Millionome Project, where we show that EMU is able to reconstruct the previous findings in a matter of CPU hours compared to CPU hours of the original study. We are able to identify population structure across mainland China with two clines from North to South and from West to East, respectively. In this analysis LD pruning was not needed, however since EMU takes PLINK files as input, the user can perform LD pruning using PLINK. This can be done using the non-missing part of the data. However, if there is too much missingness then pruning can be performed by thinning sites such that only a fraction of the sites are retained.

One limitation of the iterative nature of EMU is the fact that the individual allele frequencies need to be updated, and thus stored in memory. This can become a problem for large-scale genetic studies with sample sizes greater than 100K to run on standard server solutions. We therefore also implemented a more memory-efficient variant of EMU, which only keep the low-rank factor matrices in memory and compute individual allele frequencies when needed at the cost of computational speed and only a slight loss of accuracy. We believe that our method may be of use in the field of applied statistics as we successfully combine EMU-PCA with a SQUAREM acceleration scheme on top of truncated SVD methods to infer population structure in the presence of large missingness.

5 Funding

The study was supported by the Lundbeck foundation (R215-2015-4174) (A.A & J.M.). S.L. & M.H. was supported by National Natural Science Foundation of China (31900487).

6 Author contributions

J.M. & A.A developed the method. J.M. wrote the algorithm, wrote the manuscript and performed simulation analyses. S.L. & M.H. performed the analysis using the NIPT data. All authors reviewed and contributed to the final manuscript.

7 Data availability

The data underlying this article were provided by the Chinese National GenBank under licence by permission. Data requests should be directed to the corresponding author of the original study (Liu et al. (2018)).

References


Large-scale Inference of Population Structure in Presence of Missingness using PCA

Jonas Meisner, Siyang Liu, Mingxi Huang and Anders Albrechtsen

Supplementary Material

EMU-mem - Memory-efficient implementation

Here we describe a more memory-efficient implementation of EMU. Instead of computing and storing the individual allele frequencies $\hat{\Pi}$ after performing SVD (step 3, algorithm 1), then only the decomposition matrices $W_{[1:K]}S_{[1:K]}U_{[1:K]}^T$ are retained at each iteration. By only using these matrices alongside the original data matrix $D$ and population allele frequencies $\hat{f}$, we will never have to construct the memory heavy matrix $E$ to keep track of the individual allele frequencies. We will instead estimate the values that are needed on the fly. This is possible through the use of the low-rank approximation variant of Halko et al. [1], such that the decomposition of the truncated SVD is based on lazy evaluation of $E$ in its matrix multiplication steps. We are performing no normalization in the power iterations [1] of the inner steps of the low-rank approximation in EMU-mem in order to speed up computations but it can lead to a slight loss in accuracy. Additionally, $D$ is also stored in a 2-bit integer format as only 2-bits are needed for an individual in a given site. Thus each byte will contain the genotypes of 4 individuals.

Hence only $W_{[1:K]}$, $S_{[1:K]}$ and $U_{[1:K]}^T$ will be updated in each iteration.

Memory usage of EMU variants

The memory usage in EMU is mainly governed by:

- $D$ $n \times m$ signed char matrix,
- $E$ $n \times m$ float matrix,

such that roughly $5nm$ bytes are used in memory.

While for EMU-mem, the memory usage is mainly governed by:

- $D$ $b \times m$ unsigned char matrix, where $b = \frac{n}{4}$.

This is a 95% reduction in memory usage compared to EMU, by using roughly $\frac{nm}{4}$ bytes.
Accelerated EM algorithm

Here we describe the acceleration scheme in the expectation-maximization algorithm for performing PCA. We use the SqS3 acceleration scheme [2] and have the following definitions:

\( \tilde{W}^{(t)} \) is the \( n \times K \) factor matrix in iteration \( t \), \( \tilde{W} = WS \)

\( U^{(t)} \) is the \( m \times K \) factor matrix in iteration \( t \)

\( F(\tilde{W}^{(t)}) \) is an EM update of \( \tilde{W}^{(t)} \) based on Algorithm 1

\( F(U^{(t)}) \) is an EM update of \( U^{(t)} \) based on Algorithm 1

Two normal EM steps are taken for the two factor matrices as inner updates in the acceleration step. As in Algorithm 1 in the main paper, we map the individual allele frequencies to domain at each inner update in the acceleration step. The accelerated update is defined as follows for \( \tilde{W} \):

\[
\tilde{W}^{(t+1)} = \tilde{W}^{(t)} + 2\alpha_{\tilde{W},t} R_{\tilde{W},t} + \alpha_{\tilde{W},t}^2 V_{\tilde{W},t},
\]

(S1)

where \( R_{\tilde{W},t} = F(\tilde{W}^{(t)}) - \tilde{W}^{(t)} \) and \( V_{\tilde{W},t} = F(F(\tilde{W}^{(t)})) - 2F(\tilde{W}^{(t)}) + \tilde{W}^{(t)} \). And the accelerated update is likewise defined as follows for \( U \):

\[
U^{(t+1)} = U^{(t)} - 2\alpha_{U,t} R_{U,t} - \alpha_{U,t}^2 V_{U,t},
\]

(S2)

where \( R_{U,t} = F(U^{(t)}) - U^{(t)} \) and \( V_{U,t} = F(F(U^{(t)})) - 2F(U^{(t)}) + U^{(t)} \).

The steplengths in iteration \( t \), which are defined by \( \alpha_{\tilde{W},t} \) and \( \alpha_{U,t} \), are computed in the following way for \( \tilde{W} \) and \( U \), respectively:

\[
\alpha_{\tilde{W},t} = -\frac{\|R_{\tilde{W},t}\|_F}{\|V_{\tilde{W},t}\|_F}
\]

(S3)

\[
\alpha_{U,t} = -\frac{\|R_{U,t}\|_F}{\|V_{U,t}\|_F}
\]

(S4)

\( \| . \| \) represents the Frobenius norm.
Supplementary Figures

Figure S1: Graphical overview of the simulated scenarios. Rows represent SNPs and columns represent individuals and each dot is colored by missingness such that red means information while sand-white means missing value.
Figure S2: PCA plots of tested methods for Scenario 1A displaying the top two axes of genetic variation for 900 individuals. Black dots represent two-way admixed individuals. The top left plot shows the PCA performed on the full dataset, such that it acts as ground-truth. Individuals were simulated with low to moderate missingness rates between 5 – 50%.
Figure S3: PCA plots of tested methods for Scenario 1B displaying the top two axes of genetic variation, similarly to Figure S2, however performed on 9000 individuals. Black dots represent two-way admixed individuals. The top left plot shows the PCA performed on the full dataset, such that it acts as ground-truth. Individuals were simulated with low to moderate missingness rates between 5 – 50%.
Figure S4: The same PCA plots as in Figure S2, however the individuals have been coloured by their sampled missingness rate. Here in Scenario 1A, the individuals were assigned rates between 5 – 50%.
Figure S5: PCA plots of tested methods for Scenario 2B displaying the top two axes of genetic variation, similarly to Figure 1, however performed on 9000 individuals. Black dots represent two-way admixed individuals. The top left plot shows the PCA performed on the full dataset, such that it acts as ground-truth. Individuals were simulated with extreme missingness rates between 90 – 99%.
Figure S6: The same PCA plots as in Figure 1, however the individuals have been coloured by their sampled missingness rate. Here in Scenario 2A, the individuals were assigned rates between 90 – 99%.
Figure S7: PCA plots of tested methods for Scenario 3A displaying the top two axes of genetic variation for 900 individuals. Black dots represent two-way admixed individuals. The top left plot shows the PCA performed on the full dataset, such that it acts as ground-truth. Individuals were simulated with missingness rates sampled as either \( \approx 50\% \) or \( \approx 95\% \).
Figure S8: PCA plots of tested methods for Scenario 3B displaying the top two axes of genetic variation, similarly to Figure S3, however performed on 9000 individuals. Black dots represent two-way admixed individuals. The top left plot shows the PCA performed on the full dataset, such that it acts as ground-truth. Individuals were simulated with missingness rates sampled as either $\approx 50\%$ or $\approx 95\%$. 
Figure S9: The same PCA plots as in Figure S3, however the individuals have been coloured by their sampled missingness rate. Here in Scenario 3A, the individuals were assigned rates around either 50% and 95%.
Figure S10: PCA plots of tested methods for Scenario 4 displaying the top two axes of genetic variation. The top left plot shows the PCA performed on the full dataset, such that it acts as ground-truth. Half of the individuals from each population were simulated using only common variants (MAF $\geq 0.25$), while the other halves were simulated using the entire set of variants to emulate the merging of datasets from SNP arrays with whole-genome sequencing.
Figure S11: PCA plots of tested methods for Scenario 5 displaying the top two axes of genetic variation. The top left plot shows the PCA performed on the full dataset, such that it acts as ground-truth. In this scenario, one half of the simulated Han individuals only has information in \( \sim 130K \) variants, while the other half has information in \( \sim 233K \) variants with \( \sim 13K \) variants overlapping between the halves. The other two populations have full information.
Figure S12: PCA plots of tested methods for Scenario 6 displaying the top two axes of genetic variation. The top left plot shows the PCA performed on the full dataset, such that it acts as ground-truth. The simulation procedure is almost identical to Scenario 5 but it is performed for two populations (Han and Uygur) instead of one. The Dai population is simulated with full information.
Figure S13: PCA plots of tested methods for Scenario 7 displaying the top two axes of genetic variation. The top left plot shows the PCA performed on the full dataset, such that it acts as ground-truth. The simulation scenario is almost identical to Scenario 5, however there is no overlap in variants between the two subsets of the Han population. Thus, one half of the simulated Han individuals only has information in \( \sim 117K \) variants while the other half has information in \( \sim 233K \) variants. PLINK was not able to run in this scenario.
Figure S14: PCA plots of tested methods for Scenario 8 displaying the top two axes of genetic variation. The top left plot shows the PCA performed on the full dataset, such that it acts as ground-truth. The simulation procedure is almost identical to Scenario 7 but it is performed for two populations (Han and Uygur) instead of one. The Dai population is simulated with full information. PLINK was not able to run in this scenario.
Figure S15: MDS using PLINK for all simulated scenarios. Black dots represent two-way admixed individuals. MDS could not be run for the last two scenarios due to the same reason as for PLINK.
Figure S16: PCA plots using the R package SNPRelate for all simulated scenarios. SNPRelate suffers from the bias of mean imputation as seen in the other methods. Black dots represent two-way admixed individuals.
Figure S17: Number of iterations until convergence in EMU and EMU-mem as a function of the sample size of the dataset.
A linear model has been fitted for EMU and an exponential model for PLINK. Assuming that the number of iterations of EMU will not decrease further, then the runtime can be approximated by the following function:

\[ f(n) = 0.000246n + 0.786 \]

While the runtime for PLINK is approximated by:

\[ g(n) = 10^{-10.2} \times n^{2.558} \]

Here \( n \) is the number of individuals, while the number of sites is assumed constant (350K). The points represent the actual runtimes, while the lines follow the fitted runtimes. The first runtime point was not included in the curve fitting.

Figure S18: Curve fitting of runtimes for EMU and PLINK. Data points and fitted lines are in log_{10} scaling.
Figure S19: Inferred population structure using EMU and PLINK on the Phase 1 dataset of 96,880 low-pass genomes. Individuals have been colored by their reported province of sampling where individuals colored grey have no information.
References


Paper III

Testing for Hardy-Weinberg equilibrium in structured populations using genotype or low-depth next generation sequencing data

By
Jonas Meisner\(^1\) and Anders Albrechtsen\(^1\)

\(^1\)Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Denmark

Publication details
Published in *Molecular Ecology Resources*, vol. 19, no. 5, 1144-1152, 2019.
INTRODUCTION

Genotype frequencies in a population are normally described using the principle of Hardy–Weinberg equilibrium (HWE) (Hardy et al., 1908; Weinberg, 1908). Under the assumption of HWE, genotype frequencies can be defined as functions of allele frequencies which are conveniently captured as the binomial distribution for diallelic sites. HWE states that genotype and allele frequencies will remain constant in nonoverlapping generations in the absence of other evolutionary forces given an assumption of random mating. Testing for HWE in a population has therefore become a very common tool for detecting possible sequencing or genotyping errors, population stratification as well as other effects leading to nonrandom mating, acting as a quality control step in genetic analyses (Waples, 2014; Wigginton, Cutler, & Abecasis, 2005). Extensions to HWE have been defined to incorporate an inbreeding coefficient in the statistical models to quantify deviations from HWE as an excess or deficiency in observed heterozygotes. However, population structure will also lead to violations of the expected Hardy–Weinberg (HW) proportions by increasing the observed homozygosity due to the Wahlund effect, or increasing the observed heterozygosity due to recent admixture.

Recent methods have been developed to account for population structure using individual allele frequencies estimated from principal component analysis (PCA) (Hao, Song, & Storey, 2015; Korneliussen, Albrechtsen, & Nielsen, 2014; Meisner & Albrechtsen, 2018). PCA is a commonly used tool in population genetics for inferring population structure, as it has an advantage of describing individuals
along axes of genetic variation instead of having to assign them in clusters (Patterson, Price, & Reich, 2006). The individual allele frequencies represent probabilities of the distribution from which the genotypes of each individual are sampled given their inferred population structure (Hao et al., 2015). These methods have been shown to be very effective in large data set with diverse ancestries, where population structure can be naturally taken into account using principal components. We have recently demonstrated the effectiveness of individual allele frequencies in next-generation sequencing (NGS) data as well, where we accurately infer population structure using an iterative algorithm in low-depth sequencing data (Meisner & Albrechtsen, 2018).

Additionally, test statistics for testing for HWE in the presence of population structure have been constructed by incorporating axes of genetic variation from PCA for genotype data (Hao & Storey, 2017; Sha & Zhang, 2011). The methods utilize similar ideas of adjusting genotype frequencies based on fitted logistic models using the inferred principal components as covariates to account for the population stratification.

Next-generation sequencing data have become more prevalent in genetic studies as the cost of whole-genome sequencing has decreased (Chiang et al., 2017; Franciolli et al., 2014; Liu et al., 2018; UK10K Consortium, 2015). This has also led to an increased number of large-scale sequencing studies of samples with diverse ancestries (1000 Genomes Project Consortium et al., 2010, 2015). However, sequencing depth is usually lowered to meet the demand of the large- scale populations, but recent methods have used approaches to estimate individual allele frequencies from the inferred population structure using PCA instead (Hao et al., 2015; Meisner & Albrechtsen, 2018). In this study, we use individual allele frequencies estimated from PCAngsd (Meisner & Albrechtsen, 2018), of which our method will act as a natural extension. The number of eigenvectors used to model the individual allele frequencies are automatically detected in PCAngsd (Meisner & Albrechtsen, 2018), although if the data consist of K homogeneous populations then a reasonable choice to explain population structure would be $D = K - 1$ (Patterson et al., 2006).

Genotypes are not observed in sequencing data and we will instead work directly on genotype likelihoods to avoid introducing uncertainty through genotype calling. The genotype likelihood, $p(X_i | G = g)$, can be described as the probability of observing the sequencing data $X_i$ given the genotype $g$ for individual $i$ at site $s$. One model that retains information of the sequencing process when estimating genotype likelihoods is described in section 1 of the Supporting Information. We are therefore proposing a method for estimating per-site inbreeding coefficients and computing likelihood ratios to test for HWE.

## 2 | MATERIALS AND METHODS

We assume that individuals are diploid and variable sites are diallelic with genotypes coded as the number of minor alleles, $g = \{0, 1, 2\}$, and that the major and minor alleles are known for a dataset of $n$ individuals and $m$ sites. Based on individual allele frequencies, the genotypes of the individuals are assumed to be sampled as follows given their inferred population structure, under the assumption of HWE:

$$g_i \sim \text{Binomial}(2, x_i)$$

with $x_i$ being the individual allele frequency for individual $i$ at site $s$. The concept of individual allele frequencies were introduced in STRUCTURE (Pritchard, Stephens, & Donnelly, 2000) based on admixture proportions, population-specific allele frequencies and an assumption of $K$ ancestral populations, but recent methods have used approaches to estimate individual allele frequencies from the inferred population structure using PCA instead (Hao et al., 2015; Meisner & Albrechtsen, 2018).

In this study, we propose a method to test for HWE in structured populations on the basis of genotype likelihoods. The method incorporates individual allele frequencies to account for population structure, such that we are able to test for effects leading to nonrandom mating other than population structure. Our method is implemented into the PCAngsd framework (Meisner & Albrechtsen, 2018) for ease of use with both low-depth NGS and genotype data. We demonstrate its usefulness in both simulated and real data set.

### 2.1 | Expectation-maximization algorithm

We now describe our method, which is an extension to an expectation-maximization (EM) algorithm derived by Vieira et al. (2013) using genotype likelihoods. The EM algorithm is based on Wright’s coefficient of inbreeding (Wright, 1949) for site $s$ defined as:

$$f_s = 1 - \frac{H_s}{H_e}$$

where $H_s$ is the observed frequency of heterozygotes and $H_e$ is the expected frequency of heterozygotes. We extend the model by substituting individual allele frequencies for population allele frequencies in the likelihood function of the model to take population structure into account. In this way, we are able to estimate per-site inbreeding coefficients in structured and admixed populations. The likelihood of the inbreeding coefficient at site $s$ is defined as follows by assuming independence between individuals conditional on the population structure captured by individual allele frequencies:
Here $F_i$ is the per-site inbreeding coefficient, $n_i$ is the individual allele frequency and $X_i$ is the observed sequencing data for individual $i$ at site $s$. The genotype probability, $P(G = g | n_i, F_i)$, is computed from HW proportions with the inbreeding coefficient incorporated to model deviations from HWE. Thus, for $g = 0, 1, 2$:

$$P(G = g | n_i, F_i) = \begin{cases} 
(1 - n_i)^2 + n_i(1 - n_i)F_i & g = 0, \\
2n_i(1 - n_i)(1 - F_i) & g = 1, \\
n_i^2 + n_i(1 - n_i)F_i & g = 2.
\end{cases}$$

(4)

In this likelihood framework, $F_i$ is normally restricted to $[0, 1]$, although we allow it to be in the interval of $[-1, 1]$, where a negative estimate indicates an excess of heterozygosity and a positive estimate indicates an excess of homozygosity at site $s$. While a positive inbreeding coefficient does not change the allele frequency, a negative inbreeding coefficient will increase the sample minor allele frequency as the fraction of heterozygous individuals increases. At certain allele frequencies a negative inbreeding coefficient can lead to negative probabilities for the homozygous genotypes in Equation and thus make the distribution invalid. The minimum value, which the inbreeding coefficient can take for keeping genotype probabilities $\geq 0$, is analytically expressed as $F_i \geq \max (\frac{n_i}{1 - n_i} - 1, \frac{n_i - 1}{n_i})$. For example, $F_i = -1$ can only have an allele frequency of 0.5 because all individuals will be heterozygous, and when having to adjust for a negative genotype probability then the sample minor allele frequency will be altered due to a shift in genotype probabilities. To prevent this, we propose a heuristic distribution by truncating negative frequencies to 0 and rescaling the distribution to sum to one. The results of the truncation are visualized in Figure S1. $F_i$ does not have the same biological interpretation for negative values, but instead only act as some measure of deviation from HW proportions in the direction of excess of heterozygosity. It is therefore noteworthy that $F_i$ will not behave the same in the negative domain as for the positive. We will still refer to $F_i$ as per-site inbreeding coefficient for convenience.

The likelihood is maximized using the proposed EM algorithm. The EM algorithm is fully described in of the Supporting Information. Using the maximum likelihood estimate, we construct a likelihood ratio test (LRT) statistic, $D_i$, to test for deviations from HWE at each site. The null model is defined as $F^{\star}_i = 0$ and the alternative model is defined as the maximum likelihood estimate, $F_i = \hat{F}_i$.

$$D_i = -2 \ln \left( \frac{L(F_i = \hat{F}_i)}{L(F_i = 0)} \right).$$

(5)

$D_i$ will be $\chi^2$ distributed with 1 degree of freedom.

### 2.2 | Simulation of genotypes and low-depth sequencing data

We simulated genotypes and low-depth sequencing data to test the performance of the method. Using allele frequencies from three populations (French, Han Chinese, Yoruba) generated from genotyped individuals (Lazaridis et al., 2014) of the Human Genome Diversity Project (HGDP) (Cann et al., 2002), we simulated genotypes of 330 individuals using a binomial model. One hundred individuals were simulated from each of the three populations, while 30 individuals where simulated with different degrees of admixture between pairs of the three populations based on varying admixture proportions to represent admixed individuals. The data consist of 340,000 variable sites and linkage disequilibrium was not simulated between sites. Low-depth sequencing data are represented by genotype likelihoods, which are simulated from the sampled genotypes based on a previously used approach (Kim et al., 2011; Meinsner & Albrechtsen, 2018; Skotte et al., 2013). For one individual, the number of reads at each site is sampled from a Poisson distribution with mean $d_i$. Here $d_i$ represents the average sequencing depth of individual $i$, which is sampled from a normal distribution with mean 5 and variance 1, $d_i \sim \mathcal{N}(5, 1)$. Thus, the average sequencing depth of the simulated sequencing data is $\sim 5 \times$. The simulated genotypes are then used to sample the number of reads with the minor allele based on a binomial model also using $d_i$ as parameter. Finally, the genotype likelihoods of the three genotypes are obtained from the probability mass function of the binomial distribution. Sequencing errors are incorporated using 0.01 as the probability of a sampled read being an error.

The data generated from the described procedure is regarded as Scenario 1, where all sites are sampled in HWE. However, we also generate a second scenario, regarded as Scenario 2, where half of the sites deviate from HWE. This is done by changing the observed heterozygous genotypes to either of the homoygous genotypes with a probability of 0.5, such that $F = 0.5$ for a quarter of the sites. Likewise in the opposite direction, we change the homoygous genotypes to heterozygous with a probability of 0.5 in order to simulate $F = -0.5$ for a different quarter of the sites. However, the latter case will cause changes in allele frequencies due to imbalance between the numbers of homoygous genotypes, such that $F = -0.5$ does not hold when using sample allele frequencies.

Lastly, we generate Scenario 3 based on the same principles as in Scenario 2 but having half of the sites deviate from HWE in a more continuous fashion. A quarter of the sites deviate from HWE with $F \in [0.1, 1]$ and a quarter of the sites deviate from HWE with $F \in [-1, -0.1]$, where the latter case of negative inbreeding coefficients will cause a change in the allele frequencies again.

### 2.3 | 1000 Genomes Project data

We also test PCANOSS on genotype and sequencing data of the phase 3 release from the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2010, 2015). The dataset consists of 366 unrelated individuals from four populations with 56 individuals from ASW (Americans of African Ancestry), 99 from CEU (Utah residents with Northern and Western European ancestry), 103 from CHB (Han Chinese in Beijing) and 108 from YRI (Yoruba in
Ibadan). The genotype data are based on variant calls that consist of 7.4 million variable sites after data filtering, and the sequencing data are based on low-coverage NGS data of the same individuals in 7.9 million variable sites with 6.9 million overlapping between the data set. The variable sites in the genotype data are highly curated such that information from low-coverage sequencing, exome sequencing and other genotyping platforms (single nucleotide polymorphism [SNP] chips) have been used to produce a high-quality phased variant callset (1000 Genomes Project Consortium et al., 2015). Data filtering and generation of genotype likelihoods from the low-coverage dataset have been performed in ANGSD (Korneliussen et al., 2014) version 0.929. The genotype likelihoods are therefore covering all variable sites in the low-coverage sequencing data, called in ANGSD, and not only the curated variant calls as in the genotype data. The average sequencing depth in the low-coverage dataset is estimated to be 6.1× (varying from 1.7 to 13.6×) based on the used filters. The filtering options used for both data set are described in the Supporting Information.

3 | RESULTS

We have implemented our method into the PCAngsd framework as we will also refer to it in the following results. Through the framework, the method will work on both genotype likelihoods in Beagle format and standard genotype data in binary PLINK format. PCAngsd converts the genotype matrix into the genotype likelihood format on-the-go for ease of use and introduces the possibility of incorporating a genotype error model (Albrechtsen et al., 2009).

Other methods for testing for HWE in structured populations are hwes (Sha & Zhang, 2011) as well as a recently proposed method shWE (Hao & Storey, 2017). Both use individual allele frequencies, but only work on genotype data. However, PCAngsd has the advantage of estimating per-site inbreeding coefficients in addition to a test statistic as well as working on both low-depth sequencing data and genotype data. We compare PCAngsd version 0.98 to both shWE version 1.9 and the commonly used implementation in PLINK (Purcell et al., 2007) version 1.9 which does not accommodate population structure. We have used r version 3.4.4 to run the analyses of shWE and for plotting results. All the following results performed on sequencing data relate to genotype likelihoods.

3.1 | Simulations

As proof of concept, we have applied PCAngsd to simulated genotype and low-depth sequencing data of 330 individuals from three different populations (French, Han Chinese, Yoruba) in 340,000 variable sites. The inferred population structure using PCAngsd is visualized in Figure 1 and we have used the top two eigenvectors to model the individual allele frequencies (D = 2). As described in the Materials and Methods section, three scenarios have been simulated: Scenario 1, where none of the sites is sampled to be out of HW proportions (F = 0); Scenario 2, where half of the sites are sampled with F = 0 while the other half is equally sampled with either F = 0.5 or F = -0.5 being out of HW proportions; and Scenario 3, where half of the sites are deviating from HWE with either F ∈ [0.1, 1] or F ∈ [-1, −0.1].

In Scenario 1, we are able to estimate per-site inbreeding coefficients that follow a normal distribution around the expected value (F = 0) with the spread representing sampling variance, displayed in Figure 2. As seen by the quantile-quantile (QQ) plots in Figure 3, our test statistic is also behaving as expected under
the null for both genotype and low-depth sequencing data. For genotype data, PCAngsd performs very similarly to shwe while the test statistics of Plink are inflated and biased due to not taking population structure into account. shwe is used with three logistic factors, as one factor represents the intercept. shwe and Plink are also applied to naively called genotypes of the sequencing data in Scenario 1 to demonstrate the difficulties in analysing low-depth sequencing data. Genotypes are called by choosing the most likely genotype based on the generated genotype likelihoods as also performed in the commonly used gAtk software (McKenna et al., 2010), and the results are also displayed in Figure 3. Here it is clearly seen that the two methods have inflated test statistics as the statistical uncertainty in the genotypes is not taken into account, whereas PCAngsd is able to account for this uncertainty by working directly on the genotype likelihoods. Additionally, naively called genotypes at different sequencing depths (10, 15, 20×) have also been tested for shwe in order to see when its performance becomes similar to PCAngsd (Figure S4), which appear to be around ~15×. We have also tested the method hwes (Sha & Zhang, 2011) on the simulated genotype data set and its results are shown in Figure S5 but not included in the main results as it is performing slightly worse in comparison to shwe.

When applied to the simulated data in Scenario 2, the effect of negative inbreeding coefficients is seen in the estimates of PCAngsd. The estimates for the sites sampled with $F = -0.5$ are slightly biased, as expected, due to the sample frequencies being affected by the negative per-site inbreeding coefficients, as displayed in Figure 4. However, again the estimates for the sites sampled with $F = 0$ and $F = 0.5$ follow normal distributions around the expected values. Table 1 further shows that PCAngsd performs well in terms of power and false positive rate (FPR) in comparison to shwe and Plink, and it works well for detecting sites that deviate from HWE with negative per-site inbreeding coefficients. PCAngsd slightly loses power when using low-depth sequencing data but it is still able to keep the expected FPR.

The results of Scenario 3 are shown in Figure S3 and Table 1. Even in this more complex scenario, PCAngsd is able to estimate very reasonable inbreeding coefficients that follow our expectations. Again PCAngsd is shown to have more power than shwe while maintaining a low FPR. Similarly to Scenario 2, PCAngsd slightly loses some power for the low-depth sequencing data.

Additionally, we tested the effect of choosing different numbers of eigenvectors $D$ for estimating the individual allele frequencies in PCAngsd (Figure S6). As expected by choosing $D = 3$, we do not observe any difference in comparison to $D = 2$, as all population structure has already been captured by the top two eigenvectors. However, when choosing $D = 1$, we see the effect of not capturing all population structure, which leads to a skewed distribution with more positive inbreeding coefficients.

### 3.2 1000 Genomes Project

We also applied PCAngsd to genotype and low-coverage sequencing data of 366 individuals from four populations in the 1000 Genomes Project (ASW, CEU, CHB, YRI). The two data set consist of 7.4 million and 7.9 million variable sites, respectively, where 6.9 million of the sites are overlapping. The population structure inferred using PCAngsd is displayed in Figure 5 and again the top two eigenvectors have been used to model the individual allele frequencies ($D = 2$). The computational runtimes of PCAngsd were

![FIGURE 2](https://example.com/figure2.png)

**Figure 2** Histograms of the estimated per-site inbreeding coefficients for the simulated scenario with sites sampled from Hardy–Weinberg proportions. The left plot displays the estimates from simulated genotype data and the right plot shows the estimates from simulated NGS data with a sequencing depth of ~5×. Sites with a $p$-value lower than 0.05 are coloured blue [Colour figure can be viewed at wileyonlinelibrary.com]
6 min for the genotype data which is >10× faster than shwe, and 20 min for the low-coverage sequencing data. Both runtimes include reading of data and estimation of individual allele frequencies performed server-side using 64 threads (2.10 GHz; Intel Xeon Gold 6152).

In both data set, we are able to estimate per-site inbreeding coefficients that follow a normal distribution around 0 (Figure 6), as expected, although the negative estimates are seen to be slightly skewed with a heavy tail for the low-coverage sequencing data when using all sites. As seen for the simulations, our test statistic performs

![QQ plots of the test statistics in -log10 scale for Scenario 1, where all sites are sampled from Hardy-Weinberg proportions. The left plot is the test statistics of using simulated genotype data while the right plot is the test statistics of using simulated low-depth sequencing data (~5×), with PCAngsd based directly on genotype likelihoods, and shwe and PLINK based on naively called genotypes. The red line visualizes similarity between expected and observed p-values while the black curves describe the 90% concentration bands of the expected p-values](Colour figure can be viewed at wileyonlinelibrary.com)

![Histograms of the estimated per-site inbreeding coefficients for the simulated scenario with half of the sites sampled with \( F = 0 \), a quarter sampled with \( F = 0.5 \) and quarter sampled with \( F = -0.5 \). The left plot displays the estimates from simulated genotype data and the right plot shows the estimates from simulated NGS data with a sequencing depth of ~5×. Sites with a p-value lower than 0.05 are coloured blue, while sites with a p-value lower than 1.0 \( \times 10^{-6} \) are coloured red](Colour figure can be viewed at wileyonlinelibrary.com)
very similarly to shwe using three logistic factors for the genotype data and both methods are able to reduce the number of sites that deviate from HWE by a factor of ~8 ($\alpha = 10^{-6}$) in comparison to Plink, which is expected to be inflated due to population structure (Figure S2 and Table S1). This effect is considerably smaller when analysing the full low-coverage sequencing data using PCAngsd, where we would also expect to see more technical and random effects causing deviations from HWE in comparison to the curated set of sites of the genotype data. However, we show that the performance of PCAngsd, when only analysing overlapping sites with the curated callset for the low-coverage sequencing data, is very similar to that using the genotype data. Most of the significant sites have therefore seemingly been filtered out in the variant calling for the available genotype data of the 1000 Genomes Project phase 3 release, thus reinforcing the capabilities of PCAngsd in low-depth sequencing data. We tested two other methods based on genotype likelihoods on the low-coverage sequencing data as well (Korneliussen et al., 2014; Vieira et al., 2013). However, neither method takes population structure into account and therefore perform similarly to Plink (Figure S7).

### DISCUSSION

We have proposed a method to test for HWE in structured populations and integrated it into the PCAngsd framework. This is made possible by incorporating individual allele frequencies, which are modelled from population structure, into a likelihood framework that estimate deviations from HWE. The method is able to work on both low-depth sequencing data, in the form of genotype likelihoods in Beagle format, and genotype data in binary PLINK format.

<table>
<thead>
<tr>
<th>Scenario 2</th>
<th>Method</th>
<th>Data</th>
<th>Power</th>
<th>FPR</th>
<th>FPR ($\alpha = 0.05$)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCAngsd</td>
<td>Genotypes</td>
<td>0.990</td>
<td>0.000</td>
<td>0.0511</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>PCAngsd</td>
<td>NGS</td>
<td>0.941</td>
<td>0.000</td>
<td>0.0510</td>
<td>0.971</td>
</tr>
<tr>
<td></td>
<td>shWE</td>
<td>Genotypes</td>
<td>0.988</td>
<td>0.000</td>
<td>0.0485</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>Plink</td>
<td>Genotypes</td>
<td>0.990</td>
<td>0.0589</td>
<td>0.575</td>
<td>0.965</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 3</th>
<th>Method</th>
<th>Data</th>
<th>Power</th>
<th>FPR</th>
<th>FPR ($\alpha = 0.05$)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCAngsd</td>
<td>Genotypes</td>
<td>0.787</td>
<td>0.000</td>
<td>0.0503</td>
<td>0.892</td>
</tr>
<tr>
<td></td>
<td>PCAngsd</td>
<td>NGS</td>
<td>0.707</td>
<td>0.000</td>
<td>0.0487</td>
<td>0.863</td>
</tr>
<tr>
<td></td>
<td>shWE</td>
<td>Genotypes</td>
<td>0.671</td>
<td>0.000</td>
<td>0.0479</td>
<td>0.836</td>
</tr>
<tr>
<td></td>
<td>Plink</td>
<td>Genotypes</td>
<td>0.800</td>
<td>0.0584</td>
<td>0.364</td>
<td>0.871</td>
</tr>
</tbody>
</table>

**TABLE 1: Performance of methods on classification of sites out of Hardy–Weinberg (HW) proportions for simulated data with half of the sites sampled with $F = 0$, which are used to measure FPR, and the other half sampled out of HW proportions according to the described scenario. The half sampled out of HWE are used to measure power. A site is classified as out of HWE for a p-value < $1.0 \times 10^{-6}$, except for the fifth column where a less stringent threshold is evaluated.**
We have applied PCAngsd to both simulated and real data to test for HWE in structured populations. When we demonstrate that PCAngsd performs well using both low-depth sequencing and genotype data in the presence of population structure. The simulation studies show that PCAngsd performs very similarly to an existing method for genotype data, shwe, and both methods are able to detect deviations from HWE caused by factors other than population structure. However, PCAngsd also performs well on simulated low-depth sequencing data (~5×) where we are able to keep the statistical power high while keeping the false-positive rate low. The bias of calling genotypes for low-depth sequencing data has also been demonstrated using shwe and Plink, but PCAngsd is able to overcome this bias by working directly on genotype likelihoods.

Our results for the 1000 Genomes Project data set are not as clean as seen in the study of Hao and Storey (Hao & Storey, 2017), as we analyse all variable sites across the whole genome without filtering for sites overlapping the genotyping chip from the phase 3 release. We further show that many of the sites that deviate from HWE in the full low-coverage sequencing data are not present in the phase 3 variant callset of the 1000 Genomes Project, thus verifying the usefulness of PCAngsd, as it is able to detect the deviations from HWE due to technical errors when using low-coverage sequencing data in structured populations.

Inferred population structure can be obscured by relatedness as both will manifest as genetic similarity (Conomos, Reiner, Weir, & Thornton, 2016), and therefore we recommend excluding any known close relatives before applying PCAngsd as it would also violate the model assumptions. However in large-scale studies, we would expect that only population structure is representing the top eigenvectors and using only these could eliminate the possible confounding effect of smaller cryptic relatedness (Astle & Balding, 2009). As a limitation of PCA, it can be sensitive to batch effects that reflect other confounding factors as well, which will be a problem if they are captured contrary to population structure by the top eigenvectors. Even with these potential limitations, PCAngsd seems robust in its inference of population structure as seen in the results of the low-coverage 1000 Genomes Project dataset as well as another application of PCAngsd in a recent songbird study (Friis et al., 2018). Here PCAngsd was used to remove paralogous loci with an excess of heterozygotes based on our HWE test. Paralogous loci, and other regions where mapping is problematic, may also explain the heavy tail seen for the full 1000 Genomes Project low-coverage sequencing dataset as they would contribute to negative inbreeding coefficients.

Therefore, in addition to testing for deviations from HWE, PCAngsd estimates per-site inbreeding coefficients that quantify the deviation from HWE and are useful for understanding why sites are out of HWE. We have proposed a heuristic extension to a likelihood framework such that we are able to estimate negative per-site inbreeding coefficients. In this way, we can model deviations from HWE in both directions that may provide deeper insight into sites of evolutionary interest regardless of observing excess of homozygosity or excess of heterozygosity in structured populations.

ACKNOWLEDGEMENTS

This project was funded by the Lundbeck foundation (R215-2015-4174).

AUTHOR CONTRIBUTIONS

J.M. and A.A. developed the method. J.M. implemented the method and performed the analyses, while A.A. supervised the findings of this work. Both authors have discussed the results and contributed to the final manuscript.

DATA ACCESSIBILITY

The method is integrated in the PCAngsd framework which is freely available at https://github.com/Roseméis/pcangsd. The datasets...
used from the phase 3 release of the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2010, 2015) are available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/ and ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ for the low coverage sequencing data and the variant callset, respectively.

A list of the used individuals is included as a Supporting Information. The dataset (Lazaridis et al., 2014) used to generate population-specific allele frequencies for our simulation studies is available at https://reich.hms.harvard.edu/datasets. The generated allele frequencies from the three populations are available as a Supporting Information including the script used for simulations.

REFERENCES


SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Meisner J, Albrechtsen A. Testing for Hardy-Weinberg equilibrium in structured populations using genotype or low-depth next generation sequencing data. Mol Ecol Resour. 2019;19:1144–1152. https://doi.org/10.1111/1755-0998.13019
1 Genotype likelihoods

Genotype likelihood is defined as the probability of observing the sequencing data $X_{is}$ given an unobserved genotype $G$. Here the sequencing data describes the $L$ sequencing reads covering site $s$ for individual $i$ such that the genotype likelihood is summarized as $P(X_{is} | G = g)$ for $g$. Here we describe the approach of estimating genotype likelihoods as performed in the GATK framework [1], where error probabilities of the sequencing are incorporated with an assumption of independence. In the diploid case, the model can be described as follows for an arbitrary genotype $g = A_1A_2$: \[
P(X_{is} | G = A_1A_2) = \prod_{r=1}^{L} \left( \frac{P(b_r | A_1)}{2} + \frac{P(b_r | A_2)}{2} \right),
\] where $b_r$ represents the observed base in the $r$-th sequencing read. $P(b | A)$ will therefore be $1 - \epsilon$ for $b = A$ and $\frac{1}{2}$ for $b \neq A$, with $\epsilon$ being the error probability.

2 EM algorithm

The following EM algorithm is an extension of the model described in Vieira et al. (2013) [3]. It is based on Wright’s coefficient of inbreeding, $F_s$: 
\[
F_s = 1 - \frac{H_O}{H_E}.
\]

With $H_O$ being the observed heterozygote genotype counts, and $H_E$ being the expected heterozygote genotype count for site $s$. Note that we allow this particular inbreeding coefficient to be negative and we are therefore able to model an excess of heterozygous individuals at a given site. The inbreeding coefficient is estimated based on the following likelihood model: 
\[
\mathcal{L}(F_s) = P(X_s | \pi_s, F_s) \propto \prod_{i=1}^{n} P(X_{is} | \pi_{is}, F_s),
\] where $X_{is}$ is the observed sequencing data and $\pi_{is}$ is the individual allele frequency for individual $i$ at site $s$ with an assumption of independence between individuals conditional on the inferred population structure. The genotypes are not observed in NGS data, and we therefore introduce
the genotypes as a latent variable using the law of total probability. The likelihood for individual $i$ at site $s$ is thus given as:

$$P(X_{is} | \pi_{is}, F_s) = \sum_{g=0}^{2} P(X_{is} | G = g)P(G = g | \pi_{is}, F_s).$$  \hspace{1cm} (4)

Here $P(X_{is} | G = g)$ is the genotype likelihood and $P(G = g | \pi_{is}, F_s)$ is the genotype probability for a genotype $g = 0, 1, 2$. The genotype probabilities, conditioned on the individual allele frequency, $\pi_{is}$, and the per-site inbreeding coefficient, $F_s$, are estimated based on an assumption of extended HWE, where the inbreeding coefficient has been incorporated to model deviations:

$$P(G = g | \pi_{is}, F_s) = \begin{cases} (1 - \pi_{is})^2 + \pi_{is}(1 - \pi_{is})F_s, & g = 0, \\ 2\pi_{is}(1 - \pi_{is})(1 - F_s), & g = 1, \\ \pi_{is}^2 + \pi_{is}(1 - \pi_{is})F_s, & g = 2. \end{cases}$$ \hspace{1cm} (5)

As described in the Materials and methods section, the genotype probabilities need to be rescaled in the case of negative per-site inbreeding coefficients. Using the information of the genotype likelihoods and genotype probabilities, posterior genotype probabilities can be estimated using Bayes’ theorem in order to model the uncertainty of sequencing data with external information of population structure and inbreeding.

$$P(G = g | X_{is}, \pi_{is}, F_s) = \frac{P(X_{is} | G = g)P(G = g | \pi_{is}, F_s)}{\sum_{g'=0}^{2} P(X_{is} | G = g')P(G = g' | \pi_{is}, F_s)}. \hspace{1cm} (6)$$

The sum over individuals of the posterior probabilities of being heterozygous, $\sum_{i=1}^{n} P(G = 1 | X_{is}, \pi_{is}, F_s)$, will be used as an estimate of the observed number of heterozygous individuals at site $s$ in NGS data. $H_E$ is approximated by taking the sum over individuals of the expected frequencies of being heterozygous based on the individual allele frequencies, under the assumption of HWE. Using equation 1, the per-site inbreeding coefficient is updated as follows for iteration $t$:

$$F_s^{(t+1)} = 1 - \frac{H_O}{H_E} = 1 - \frac{\sum_{i=1}^{n} P(G = 1 | X_{is}, \pi_{is}, F_s^{(t)})}{\sum_{i=1}^{n} 2\pi_{is}(1 - \pi_{is})}$$ \hspace{1cm} (7)

### 3 Likelihood ratio test

A likelihood ratio test is simply defined using $F_s = 0$ as the null model and the maximum likelihood estimate of our EM algorithm, $F_s = \hat{F}_s$, as the alternative model at site $s$:

$$\mathcal{L}(F_s = 0) = \prod_{i=1}^{n} \sum_{g=0}^{2} P(X_{is} | G = g)P(G = g | \pi_{is})$$ \hspace{1cm} (8)

$$\mathcal{L}(F_s = \hat{F}_s) = \prod_{i=1}^{n} \sum_{g=0}^{2} P(X_{is} | G = g)P(G = g | \pi_{is}, \hat{F}_s)$$ \hspace{1cm} (9)

$$D_s = 2\ln \left( \frac{\mathcal{L}(F_s = \hat{F}_s)}{\mathcal{L}(F_s = 0)} \right), \quad D_s \sim \chi_1^2.$$ \hspace{1cm} (10)
4 Software

4.1 PCAngsd

Command to estimate per-site inbreeding coefficients and LRT statistics in PCAngsd [4] for genotype likelihoods in Beagle format:

```python
python pcangsd.py -beagle gl.beagle.gz -o gl.pcangsd -inbreedSites -threads 20
```

Command to estimate per-site inbreeding coefficients and LRT statistics in PCAngsd for genotype data in binary PLINK format:

```python
python pcangsd.py -plink geno -o geno.pcangsd -inbreedSites -threads 20
```

4.2 ANGSD

Command used for calling SNPs and generating genotype likelihoods, including filtering, in ANGSD [5] as performed for the low coverage 1000 Genomes Project sequencing data:

```bash
angsd -bam 1000G.bamlist -rf chrlist -minQ 20 -minMapQ 30 -skipTriallelic 1 -doMajorMinor 1 -doGlf 2 -doMaf 2 -minMaf 0.05 -SNP_pval 1e-6 -GL 2 -doDepth 1 -doCounts 1 -minInd 300 -out 1000G.angsd -P 20
```

Thus, we use a minor allele frequency (MAF) filter of 0.05 and only include diallelic sites that are covered with at least 1 read for 300 of the 366 individuals. SNPs are called using a p-value threshold of $1.0 \times 10^{-6}$.

Commands used to estimate per-site inbreeding coefficients (Figure S7):

```bash
angsd -bam 1000G.bamlist -rf chrlist -minQ 20 -minMapQ 30 -skipTriallelic 1 -doMajorMinor 1 -doMaf 2 -minMaf 0.05 -SNP_pval 1e-6 -GL 2 -minInd 300 -HWE_pval_F 1 -out 1000G.angsd.F1 -P 20
angsd -bam 1000G.bamlist -rf chrlist -minQ 20 -minMapQ 30 -skipTriallelic 1 -doMajorMinor 1 -doMaf 2 -minMaf 0.05 -SNP_pval 1e-6 -GL 2 -minInd 300 -doHWE 1 -out 1000G.angsd.F2 -P 20
```

4.3 PLINK

PLINK [6] was used with a MAF filter of 0.05 and a maximum missing rate of 0.01 for diallelic SNPs only.

4.4 Software versions

Versions of used software in analyses:

- PCAngsd (0.98)
- sHWE (lfa R package) (1.9)
- PLINK (1.9)
- ANGSD (0.929)
5 Figures & tables

Figure S1: Distributions of the genotype probabilities (Equation 4) are displayed for $F = -0.5$ and $F = -1$ as well as the effect of rescaling. The blue, red and green curves describe $P(G = 0 \mid \pi, F)$, $P(G = 1 \mid \pi, F)$ and $P(G = 2 \mid \pi, F)$, respectively. The top row features the unscaled distributions for the two negative inbreeding coefficients, and the bottom row visualizes the two corresponding rescaled distributions.
Figure S2: QQ plots of the test statistics in -log_{10} scale of the 1000 Genomes Project datasets. The top plots shows the results of PCAngsd, where the left plot is based on the genotype data, and the middle and right plots show the estimates from low coverage sequencing data. Here the middle plot is filtered for overlapping sites with the curated set of the genotype data while the right plot includes all sites. The bottom row displays the QQ plots of the test statistics from sHWE and PLINK, respectively. Due to precision in the outputted \( p \)-values of sHWE, all \( p \)-values < \( 1.0 \times 10^{-16} \) are truncated to 16 in -log_{10}-scale for convenience in visualization. The red line visualizes similarity between expected and observed \( p \)-values while the black curves describe the 90\% concentration bands of the expected \( p \)-values.

### 1000 Genomes Project

<table>
<thead>
<tr>
<th>Threshold</th>
<th>PCAngsd - Geno</th>
<th>PCAngsd - NGS</th>
<th>sHWE</th>
<th>PLINK</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha = 0.05 )</td>
<td>0.0704</td>
<td>0.0628 (0.137)</td>
<td>0.0702</td>
<td>0.381</td>
</tr>
<tr>
<td>( \alpha = 1.0 \times 10^{-6} )</td>
<td>0.00775</td>
<td>0.00391 (0.0231)</td>
<td>0.00705</td>
<td>0.0612</td>
</tr>
</tbody>
</table>

Table S1: Proportions of total sites that are classified to deviate from HWE for each of the methods at two different thresholds. The reported values for using PCAngsd on the low coverage sequencing data are filtered for variable sites overlapping with the genotype data, while the values in the parentheses are when using all sites.
Figure S3: Histograms of the estimated per-site inbreeding coefficients for the Scenario 3 with half of the sites sampled with $F = 0$, a quarter of the sites uniformly sampled with positive inbreeding coefficients $F \in [0,1]$ and a quarter of the sites uniformly sampled with negative inbreeding coefficients $F \in [-1,-0.1]$. The top plots display the estimates from simulated genotype data and the bottom plots show the estimates from simulated NGS data with a sequencing depth of $\sim 5X$. The plots in the left column are estimates from sites sampled as $F = 0$, and the plots in the right column are estimates from sites sampled as $F \neq 0$. Sites with a $p$-value lower than 0.05 are colored blue, while sites with a $p$-value lower than $1.0 \times 10^{-6}$ are colored red.
Figure S4: QQ plots of the test statistics using sHWE [7] in $-\log_{10}$ scale using naively called genotypes from simulated sequencing data at 10, 15 and 20X, respectively, where all sites are sampled from HW proportions. Due to precision in the outputted $p$-values of sHWE, all $p$-values $< 1.0 \times 10^{-16}$ are truncated to 16 in $-\log_{10}$-scale for convenience in visualization. The red line visualizes similarity between expected and observed $p$-values while the black curves describe the 90% concentration bands of the expected $p$-values.

Figure S5: QQ plots of the test statistics using HWES [8] with $K = 3$ in $-\log_{10}$ scale using the known and naively called genotypes from simulated sequencing data, where all sites are sampled from HW proportions (Scenario 1). All $p$-values $< 1.0 \times 10^{-16}$ are truncated to 16 in $-\log_{10}$-scale for convenience in visualization. The red line visualizes similarity between expected and observed $p$-values while the black curves describe the 90% concentration bands of the expected $p$-values.
Figure S6: Histograms of the estimated per-site inbreeding coefficients for the Scenario 1 using different numbers of eigenvectors (1, 2 or 3) in the estimation of individual allele frequencies in PCAngsd. The results using $D = 2$ are the same from the main results. The top plots display the estimates from simulated genotype data and the bottom plots show the estimates from genotype likelihoods of simulated NGS data with a sequencing depth of $\sim 5X$. 

Figure S6: Histograms of the estimated per-site inbreeding coefficients for the Scenario 1 using different numbers of eigenvectors (1, 2 or 3) in the estimation of individual allele frequencies in PCAngsd. The results using $D = 2$ are the same from the main results. The top plots display the estimates from simulated genotype data and the bottom plots show the estimates from genotype likelihoods of simulated NGS data with a sequencing depth of $\sim 5X$. 

102 Chapter 4. Paper III
Figure S7: Results of ANGSD [5] methods on the low coverage sequencing data of the 1000 Genomes Project that are filtered for overlapping sites with the corresponding genotype data. The top plots show the performance of the ANGSD implementation of the ngsF [3] method and the bottom plots show an implementation, where negative inbreeding coefficients are allowed in ANGSD. The left plots show the histograms of the estimated per-site inbreeding coefficients, while the right plots display the QQ plots of the test statistics in -log_{10} scale. In the histograms, sites with a p-value lower than 0.05 are colored blue, while sites with a p-value lower than 1.0 x 10^{-6} are colored red. The red line visualizes similarity between expected and observed p-values while the black curves describe the 90% concentration bands of the expected p-values.
References


Paper IV

Haplotype and Population Structure Inference using Neural Networks in Whole-Genome Sequencing Data

By
Jonas Meisner¹ and Anders Albrechtsen¹

¹Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Denmark

Publication details
On bioRxiv.
Haplotype and Population Structure Inference using Neural Networks in Whole-Genome Sequencing Data

Jonas Meisner & Anders Albrechtsen
Department of Biology
University of Copenhagen

December 28, 2020

Abstract
Accurate inference of population structure is important in many studies of population genetics. In this paper we present, HaploNet, a novel method for performing dimensionality reduction and clustering in genetic data. The method is based on local clustering of phased haplotypes using neural networks from whole-genome sequencing or genotype data. By utilizing a Gaussian mixture prior in a variational autoencoder framework, we are able to learn a low-dimensional latent space in which we cluster haplotypes along the genome in a highly scalable manner. We demonstrate that we can use encodings of the latent space to infer global population structure using principal component analysis with haplotype information. Additionally, we derive an expectation-maximization algorithm for estimating ancestry proportions based on the haplotype clustering and the neural networks in a likelihood framework. Using different examples of sequencing data, we demonstrate that our approach is better at distinguishing closely related populations than standard principal component analysis and admixture analysis. We show that HaploNet performs similarly to ChromoPainter for principal component analysis while being much faster and allowing for unsupervised clustering.

1 Introduction
Understanding population structure is a cornerstone in population and evolutionary genetics as it provides insights into demographic events and processes that have affected a population. The most common approaches for inferring population structure from genetic data are using principal component analysis (PCA) [33] and using a clustering algorithm such as STRUCTURE [34], or derivations thereof. PCA infers continuous axes of genetic variation that summarize the genetic relationship between samples and the clustering algorithm assigns samples to a predefined number of ancestral sources while allowing for fractional membership. The inferred axes of PCA are very useful to account for population or cryptic structure in association studies or even to simply visualize the genetic data. A limitation of PCA and clustering algorithms are that they usually assume all single-nucleotide polymorphisms (SNPs) to be independent, and they do therefore not benefit from the information of correlated sites or they may be biased thereof in their global estimates [44, 33]. ChromoPainter [26] employs the Li and Stephens hidden markov model [27] for haplotype sampling in order to model and utilize correlations between SNPs, by letting samples be a mosaic of each
other’s haplotypes. In this way, ChromoPainter has become state-of-the-art for inferring fine-scale population structure.

Gaussian mixture models and $k$-Means are commonly used methods for performing unsupervised clustering. However, these methods suffer from the curse of dimensionality where distances will become almost equidistant between samples in high-dimensional space [50]. A popular approach to overcome the curse of dimensionality is to perform dimensionality reduction, e.g. using PCA, and then perform clustering in the low-dimensional space that still captures most of the variation in the full dataset [13]. Recently, deep autoencoders methods have been very successful for large-scale datasets as they perform dimensionality reduction and clustering either sequentially or jointly to benefit from induced non-linearity and scalability of deep learning architectures using neural networks [48, 47]. Deep autoencoders have also been introduced in generative models, e.g. variational autoencoders, where the unknown data generating distribution is learnt by introducing latent random variables, such that new samples can be generated from this distribution [25, 36].

Most studies in population genetics utilizing neural networks for parameter inference have mainly focused on supervised learning through simulations from demographic models [40, 38, 14, 15, 8]. Here, an overall demography is assumed, based on previous literature, and a lot of different datasets are simulated using small variations in model parameters, e.g. selection coefficient or recombination rate, with evolutionary simulators (e.g. msprime [21]). The studies are usually converting a simulated haplotype matrix into a downscaled fixed sized image with rows and/or columns sorted based on some distance measure. The network is then trained on the simulated datasets to learn the specified model parameters with feature boundaries in convolutional layers, and in the end, the model is tested on a real dataset. However, recently more studies have instead focused on deep generative models for data-driven inference or simulation using state-of-the-art unsupervised learning approaches [31, 49, 4, 46].

We here present HaploNet, a method for inferring haplotype and population structure using neural networks in an unsupervised approach for phased haplotypes of whole-genome sequencing (WGS) data. We utilize a variational autoencoder (VAE) framework to learn mappings to and from a low-dimensional latent space in which we will perform indirect clustering of haplotypes with a Gaussian mixture prior. Therefore, we do not have to rely on simulated training data from demographic models with a lot of user-specified parameters, but we are able to construct a fully data-driven inference framework. We demonstrate that we can use haplotype encodings from our latent space to infer fine-scale global population structure with PCA, similarly to the state-of-the-art software ChromoPainter while being much faster, as we model correlated SNPs in windows using non-linear layers in neural networks. However unlike ChromoPainter, we rely on clustering in our inferred latent space instead of pairwise distances between haplotypes. This allows us to compute likelihoods of the observed haplotypes in each window given a latent state, where each state represents an inferred haplotype cluster. From the computed likelihoods, we build a clustering model similar to NGSadmix [42] which we fit using an accelerated expectation-maximization (EM) algorithm to estimate ancestry proportions as well as frequencies of our neural network inferred haplotype clusters. We show that HaploNet has much higher resolution than the widely used ADMIXTURE software [2]. Additionally, we are also capable of simulating entire new chromosomes based on the inferred population structure and haplotype clusters with the generative properties of a VAE framework. This is done by sampling haplotypes from the most probable sequence of haplotype clusters in a population and the sampled haplotypes will still reflect properties of real samples, such as LD patterns. HaploNet is freely available at https://github.com/rosemeis/HaploNet.
2 Material and Methods

The method is based on phased haplotype data from diallelic markers. We define $X$ as a $2N \times M$ haplotype matrix for a single chromosome, where $N$ is the number of individuals and $M$ is the number of SNPs along the chromosome. The entries of the matrix are encoded as either 0 or 1, referring to the major and minor allele, respectively. For unphased data, we can generalize this formulation to a standard genotype matrix where heterozygous genotypes are encoded as 0.5.

For each chromosome, we divide the sites into windows of a fixed length of $L$ SNPs, which we assume is much smaller than $M$. The windows are non-overlapping and we will further assume that the parameters estimated in a window are independent from parameters estimated in adjacent windows. The length of the genomic windows can also be defined by a recombination map but we have kept it fixed in this study for the matter of generalizability and ease of application. For each defined window along a chromosome, we will independently train neural networks in a variational autoencoder framework to learn haplotype clusters and encodings using a Gaussian mixture prior. From these trained networks, we will be able to calculate distances between pairs of haplotypes using their encodings, and we will additionally be able to calculate a likelihood for an observed haplotype for each haplotype cluster similarly to calculating genotype likelihoods for WGS data.

2.1 Variational autoencoder

An autoencoder is a state-of-the-art approach for performing dimensionality reduction by learning a mapping of the space of the input data, $X$, to a low-dimensional space $Z$ and a mapping back to the input data [37, 3]. More formally, we can describe the two mapping functions as $g : X \mapsto Z$ and $f : Z \mapsto X$, which are commonly called the encoder and the decoder, respectively. Both the encoder and the decoder are parameterized by (deep) neural networks to learn the mapping, as multilayer feed-forward neural networks are universal function approximators [17].

A probabilistic variant of this neural network architecture is used in a variational autoencoder (VAE) to learn the unknown generating process of the input data by introducing latent variables and modelling the joint probability through variational inference. As an optimization method, variational inference is often used to approximate the posterior distribution of a latent variable $z$, $p(z|x)$, by fitting a function that describes a chosen family of distributions, $q_\phi(z|x)$. Thus, variational inference turns it into an optimization problem, where the objective is to maximize the evidence lower bound (ELBO) of the marginal log-likelihood of the data using an iterative procedure in contrast to Monte Carlo Markov Chain methods that approximate it by sampling from the posterior distribution [6]. The function approximating the posterior distribution is parameterized with variational parameters $\phi$. Kingma and Welling introduced the Stochastic Gradient Variational Bayes (SGVB) estimator [25] of the ELBO for approximate posterior inference in a VAE framework (as well as [36]), where a set of parameters, $(\theta, \phi)$, are optimized with amortized inference using mappings parameterized by neural networks. Here, the marginal log-likelihood, $p_\theta(x)$, is parameterized with parameters $\theta$. This amortization means that the number of parameters does not depend on sample size as in traditional variational inference, but it only depends on the network size [41]. The VAE can be seen as an autoencoder with its latent space being regularized by a chosen prior distribution to make the inferred latent space more interpretable and prevent overfitting. A standard Gaussian prior is the most common choice, however, it is often too simple and a lot of effort has been made to make the approximate posterior richer with normalizing flows and additional stochastic layers [35, 24, 43].
In our proposed method, we will construct a generative model and use a variational autoencoder framework to learn low-dimensional encodings of haplotypes in windows along the genome. However, we also introduce an additional categorical variable $y$ to represent haplotype clusters such that we enforce a Gaussian mixture prior to increase the flexibility of our approximate posterior distribution. In this way, we are able to jointly perform dimensionality reduction and clustering of the haplotypes in a highly scalable approach. In the following model descriptions, we will follow the mathematical notation used in the machine learning literature and Kingma and Welling (2013) [25], where $p_\theta$ and $q_\phi$ are probability functions that define the decoder and encoder part respectively. $\theta$ and $\phi$ are the parameters (biases and weights) in the neural networks. We have a simple visualization of the directed graphical model with our probabilistic encoder and decoder networks in Figure 1. We define the following latent variable model in a genomic window with data $x \in \{0, 1\}^L$. Gaussian latent variable $z \in \mathbb{R}^D$ and categorical latent variable $y \in \{0, 1\}^C$ (one-hot encoded) as follows:

$$p_\theta(x, z, y) = p_\theta(x | z)p_\theta(z | y)p(y), \quad (1)$$

with generative processes defined as:

$$p(y) = \text{Cat}(y; C^{-1}1), \quad (2)$$
$$p_\theta(z | y) = \mathcal{N}(z; \mu_\theta(y), \sigma_\theta^2(y)1), \quad (3)$$
$$p_\theta(x | z) = \text{Ber}(x; \pi_\theta(z)). \quad (4)$$

Here $z$ is a $D$-dimensional vector representing the latent haplotype encoding and $C$ is the number of haplotype clusters, while $\mu_\theta : \{0, 1\}^C \rightarrow \mathbb{R}^D$, $\sigma_\theta : \{0, 1\}^C \rightarrow \mathbb{R}^D$ and $\pi_\theta : \mathbb{R}^D \rightarrow [0, 1]^L$ are mapping functions parameterized by neural networks with network parameters $\theta$. In this case, $\text{Ber}(x; \pi_\theta(z))$ is a vectorized notation of Bernoulli distributions and each of the $L$ sites will have an independent probability mass function. We assume that the covariance matrix of the multivariate Gaussian distribution is a diagonal matrix which will promote disentangled factors. We have the following inference (encoder) model:

$$q_\phi(z, y | x) = q_\phi(z | x, y)q_\phi(y | x), \quad (5)$$
$$q_\phi(y | x) = \text{Cat}(y; \pi_\phi(x)), \quad (6)$$
$$q_\phi(z | x, y) = \mathcal{N}(z; \mu_\phi(x, y), \sigma_\phi^2(x, y)1), \quad (7)$$

where $\mu_\phi : \{0, 1\}^{L+C} \rightarrow \mathbb{R}^D$, $\sigma_\phi^2 : \{0, 1\}^{L+C} \rightarrow \mathbb{R}^D$ and $\pi_\phi : \{0, 1\}^L \rightarrow [0, 1]^C$ again are mapping functions parameterized by neural networks with network parameters $\phi$. Therefore the marginal posterior distribution and marginal approximate posterior distribution of $z$ will both be a mixture of Gaussians. Thus, $q_\phi(z, y | x)$ and $p_\theta(x | z)$ will constitute the probabilistic encoder and decoder, respectively, in comparison to the deterministic encoder and decoder of the standard autoencoder.

From the marginal log-likelihood of the data, we derive the following ELBO of our variational autoencoder model for haplotype $i$,

$$\log p_\theta(x_i) \geq L(\phi, \theta; x_i) = E_{q_\phi(z, y | x_i)} \left[ \log p_\theta(x_i | z) - \log \frac{q_\phi(z | x_i, y)}{p_\theta(z | y)} - \log \frac{q_\phi(y | x_i)}{p(y)} \right], \quad (8)$$
Figure 1: Directed graphical model of HaploNet, where the solid edges captures the generative model described in Equation 1, $p(\mathbf{x}, \mathbf{z}, y)$, with decoder $p(\mathbf{x} | \mathbf{z})$. The solid edges are parameterized by neural networks with parameters $\theta$. The dashed edges captures the inference model (encoder) of the variational approximation to the posterior, $q(\mathbf{z}, y | \mathbf{x})$. The dashed edges are parameterized by neural networks with parameters $\phi$. Color of circular nodes indicate whether the variable is latent or observed.

where the marginal log-likelihood of the full data in a window is given by:

$$\log p(\mathbf{x}_1^N) = \sum_{i=1}^{2N} \log p(\mathbf{x}_i).$$

The derivation of this ELBO is described in the supplementary material. We immediately see that the first term describes the reconstruction error of mapping from the latent space back to the input space as in an autoencoder framework. The next two terms act as regularization on the learnt latent spaces, where the second term encourages the variational Gaussian distribution to be close to the approximated prior distribution, while the last term encourages anti-clustering behaviour to prevent all haplotypes to cluster in one component. This is a modification of the unsupervised loss of the M2 model [23] as described by Rui Shu\(^1\) where information of the haplotype cluster is also propagated through $p(\mathbf{z} | y)$. However, we further approximate the categorical latent variable with samples from a Gumbel-Softmax distribution [19, 28] instead of the categorical distribution. The Gumbel-Softmax distribution is a continuous approximation to the categorical distribution that can be easily reparameterized for differentiable sampling and gradient estimations. In this way, we can avoid an expensive computational step of having to marginalize over the categorical latent variable in the SGVB estimator of the ELBO as is done in the original model. A lot of different interpretations and implementations of the Gaussian Mixture Variational Autoencoder (GMVAE) have been proposed [12, 11, 7, 20], and a similar architecture to ours has been implemented\(^2\).

\(^1\)http://ruishu.io/2016/12/25/gmvae/
\(^2\)https://github.com/jariasf/GMVAE
2.1.1 Reparameterization tricks

End-to-end learning is a desired trait in deep learning models as it enables joint training of model parameters through backpropagation. However with the network architectures of variational autoencoders, the sampling steps of the latent variables are not letting information flow end-to-end as they are non-differentiable. The reparameterization trick is a simple approach to express a random variable as deterministic using the introduction of an auxiliary variable as defined in the SGVB estimator [25]. In this way, the stochastic process of sampling is removed from the computation graph and we can therefore efficiently estimate gradients and perform end-to-end learning of our network parameters, \((\phi, \theta)\), through Monte Carlo samples of the latent variables to approximate the ELBO in Equation 8.

A \(D\)-dimensional sample from a multivariate Gaussian distribution, \(z\), can be expressed using the introduction of auxiliary noise variable \(\epsilon\), with \(\epsilon_d \sim \mathcal{N}(0, 1)\), for \(d = 1, \ldots, D\):

\[
z_d = \mu_d + \sigma_d \epsilon_d.
\]

(10)

Similarly, the reparameterization trick for the Gumbel-Softmax distribution is an approach for providing differentiable samples that approximate samples from a categorical distribution. It is an extension of the Gumbel-Max trick [29] for sampling from a categorical distribution, where the softmax function is used as a continuous differentiable approximation to the arg max function. The Gumbel-Softmax distribution was derived simultaneously in two independent studies [19, 28]. A \(C\)-dimensional one-hot encoded categorical sample \(y\) can be approximated as sample \(\hat{y}\) from the Gumbel-Softmax distribution, and it is defined as follows for its entry \(c\):

\[
\hat{y}_c = \frac{\exp((\log(\pi_c) + g_c)/\tau)}{\sum_{c' = 1}^C \exp((\log(\pi_{c'}) + g_{c'}/\tau))}.
\]

(11)

Here \(g_c, \ldots, g_C\) are i.i.d. samples from Gumbel(0,1) (auxiliary variables) and \(\pi_c\) is the probability of haplotype cluster \(c\). The temperature parameter, \(\tau\), is governing the smoothness of the distribution such that when \(\tau\) approaches 0, samples become identical to samples from the categorical distribution. We have provided examples of the Gumbel-Softmax distribution in the supplementary material to provide the reader with a better intuition of it. Note that \(\hat{y} \in [0, 1]^C\) unlike \(y\), and the functions \(\mu_\theta, \sigma_\theta, \mu_\phi, \sigma_\phi\) parameterized by the neural networks are therefore redefined as mapping to or from a continuous space in the practical implementation.

2.1.2 Simulation from haplotype clusters

We can use the trained neural networks to generate the most probable haplotypes of the inferred haplotype clusters or to simulate entirely new data from the haplotype clusters using the functions parameterizing the \(p_\theta(z \mid y)\) and \(p_\phi(x \mid z)\) distributions. The distribution means of \(p_\theta(z \mid y)\) are learnt for each haplotype cluster and we use them to estimate the mean haplotype representing each of the \(C\) clusters. The mean latent encodings, using only information of the haplotype clusters, \(\mathbb{E}[p_\theta(z \mid y)] = \mu_\theta(y)\), can be decoded to generate the most probable haplotypes as follows:

\[
\hat{x}_c = \arg \max_x p_\phi(x \mid \mu_\phi(y = c)),
\]

(12)

where \(\hat{x}_c \in \{0, 1\}^L\) represents the mean haplotype of the \(c\)-th haplotype cluster for \(c = 1, \ldots, C\) (one-hot encoded). The information learnt by our mappings parameterized using neural networks is 

\[111\]
therefore very interpretable. We can also sample from the distributions using the reparameterization trick to generate random haplotype sequences from a haplotype cluster. For example, given the most probable sequence of haplotype clusters a population, we can easily simulate entire chromosomes that will resemble the real data of the population.

2.2 Global population structure inferred using PCA

The inferred local encodings can be used to estimate global population structure by summarizing the information across windows using PCA. There is no obvious way to combine the two haplotypes of a single individual in our setting. However, we propose an approach that computes the covariance between individuals using the mean latent encodings of their haplotypes, \( E[q_\theta(z \mid x, y)] = \mu_\theta(x, \pi_\theta(x)) \), where we condition directly on the mean haplotype cluster probabilities, \( E[q_\theta(y \mid x)] = \pi_\theta(x) \). For two individuals, we can average the covariance across the four different pairwise combinations of their haplotypes in each window, and then average across all windows for a global estimate. The sample covariance between individual \( i \) and \( j \) is computed as follows:

\[
\text{cov}(i, j) = \frac{1}{4KW} \sum_{w=1}^{W} \sum_{a=1}^{2} \sum_{b=1}^{2} \left( \mu_{i,a}^{(w)} - \hat{\mu}^{(w)} \right)^T \left( \mu_{j,b}^{(w)} - \hat{\mu}^{(w)} \right),
\]

where \( \mu_{i,a}^{(w)} = p_{\phi}^{(w)}(x_{i,a}^{(w)}, \pi_\theta(x_{i,a}^{(w)})) \in \mathbb{R}^D \) represent the mean latent encoding for the \( a \)-th haplotype in the \( w \)-th window for individual \( i \), and \( \hat{\mu}^{(w)} \in \mathbb{R}^D \) represent the average across individuals for the \( w \)-th window. We can then perform eigendecomposition of the covariance matrix to extract the top eigenvectors that capture population structure as in a standard PCA approach.

2.3 Ancestry proportions and haplotype cluster frequencies

Another often used approach for inferring population structure are by estimating admixture proportions. We propose a model for estimating ancestry proportions and haplotype cluster frequencies assuming \( K \) ancestral components based on the model introduced in NGSAdmix [42], as an extension to the ADMIXTURE model [2], where instead of latent states of unobserved genotypes, we have latent states of haplotype clusters. We can exploit and utilize the generative nature of our GMVAE model to the \( K \) states of haplotype clusters. We can exploit and utilize the generative nature of our GMVAE model to the \( K \)-way admixture proportions and haplotype cluster frequencies \( Q \) and window-based ancestral haplotype frequencies \( F \):

\[
\mathcal{L}(Q, F; X) \propto \prod_{w=1}^{W} \prod_{i=1}^{N} \prod_{a=1}^{2} \prod_{k=1}^{K} \sum_{c=1}^{C} p(x_{i,a}^{(w)} \mid y = c) f_{wkc} q_{ik}
\]

with \( k \) describing the ancestral state, for \( Q \in [0,1]^{N \times K} \) with constraint \( \sum_{k=1}^{K} q_{ik} = 1 \) and \( F \in [0,1]^{W \times K \times C} \) with constraint \( \sum_{c=1}^{C} f_{wkc} = 1 \). The NN likelihoods are calculated as follows using the probability mass function of the Bernoulli distribution and the properties, \( E[p_\theta(z \mid y)] = \mu_\theta(y) \) and \( E[p_\theta(x \mid \mu_\theta(y))] = \pi_\theta(\mu_\theta(y)) \):
\[
p(x_{i,a}^{(w)} | y = c) \propto \prod_{l=1}^{L} \pi_{\theta}(\mu_{\theta}(y = k))^{x_{i,a,l}^{(w)}}(1 - \pi_{\theta}(\mu_{\theta}(y = k)))^{1-x_{i,a,l}^{(w)}},
\]
for \( c = 1, \ldots, C \) (one-hot encoded) and \( x_{i,a}^{(w)} \in \{0,1\}^{L} \) is the data of the \( a \)-th haplotype of individual \( i \) in window \( w \). Maximum likelihood estimates of \( F \) and \( Q \) are obtained using an expectation-maximization (EM) algorithm. The full description of the EM algorithm is detailed in the supplementary material. We use the S3 scheme of the SQUAREM methods [45] for accelerating our EM implementation, such that one large step is taken in parameter space based on the linear combination of two normal steps.

2.4 Implementation

We have implemented \texttt{HaploNet} as a \texttt{Python} program using the \texttt{PyTorch} library (v.1.7) [32], and it is freely available on https://github.com/rosemeis/HaploNet. We have used the \texttt{NumPy} [16] and \texttt{scikit-allel} [30] libraries for preprocessing the data from Variant Call Format (VCF) into data structures to be used in \texttt{HaploNet}. The EM algorithm and other postprocessing scripts for estimating covariance matrices have been written in \texttt{Cython} [5] for speed and parallelism.

Our proposed overall network architecture in \texttt{HaploNet}, including descriptions of major substructures, are displayed in Figure 2, and we have used fully-connected layers throughout the network. For all inner layers in our neural networks, we are using exponential linear unit (ELU) [10] activations to induce non-linearity into the networks, followed by batch normalization [18], while all outer layers are modelled with linear activations. This means that we are estimating the logits of the probabilities instead the probabilities directly in \( \pi_{\theta}(z) \) and \( \pi_{\phi}(x) \) for computational stability, as well as for \( \sigma_{\theta}^{2}(y) \) and \( \sigma_{\phi}^{2}(x,y) \) that represent \( \log \sigma^{2} \) in inner computations.

We are training our networks with the Adam optimizer [22] with a learning rate of \( 1.0 \times 10^{-3} \), \( \beta_1 = 0.9 \) and \( \beta_2 = 0.999 \) with a batch-size of 128. We have used a fixed temperature in the sampling from the Gumbel-Softmax distribution of \( \tau = 0.1 \) to approximate and encourage a categorical sampling, and we use one Monte Carlo sample of the latent variables to approximate the expectation in equation 8. In practice we have an additional weight term on the categorical loss term in the ELBO, \( \beta_{\text{cat}} \), such that we are able to adjust the anti-clustering effect of the uniform prior \( p(y) \), by weighing the impact of this regularization term in the ELBO (see Equation S11). In all analyses, we have used \( \beta_{\text{cat}} = 0.1 \). We decided to use all data in the training to account for and model all possible haplotypes present in the dataset as we also have no way of validating our training in this unsupervised setup. We are therefore training our models using a fixed number of 200 epochs.

2.5 1000 Genomes Project

We applied \texttt{HaploNet} to the phase 3 data of the 1000 Genomes Project (TGP) [1]. The entire dataset consists of 2504 unrelated phased individuals from 26 different populations that are assigned to 5 super populations, which consist of African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS), and we inferred local and global population structure for each super population.

We chose three smaller subsets of the data with populations of lower genetic differentiation to compare the inferred global population structure against existing methods. For the first subset, we use four European populations (CEU, GBR, IBS, TSI) that span Northern and Southern Europe,
Figure 2: The neural network architecture of HaploNet split into three major substructures. Here the solid lines represent the estimation of distribution parameters, while the dashed lines represent sampling of latent variables. 

a) Displays the neural network parameterizing the distribution $q_\phi(y \mid x)$, for sampling the haplotype cluster, b) the network parameterizing the regularizing distribution of the sampled encoding, $p_\theta(z \mid y)$, and c) shows the network parameterizing the distribution $q_\phi(z \mid x, y)$, for sampling the haplotype encoding, as well as the network decoding the sampled encoding to reconstruct our input. Note that the colors of the network blocks are coherent across substructures such that the sampled $y$ in a) is used in both b) and c).
and in the second subset we use the three Chinese populations (CDX, CHB, CHS), while the third subset consists of five of the African populations (GWD, ESN, MSL, YRI, LWK), excluding ACB and ASW that have European admixture.

For all subsets of the data, we have filtered the variable sites based on minor allele frequency at a standard threshold of 0.05. In HaploNet, we used a window size of 1024 SNPs, a latent dimension of 64 for the Gaussian latent variables representing the haplotype encodings and a latent dimension of 16 for the categorical latent variables representing the haplotype clusters. We also tried to use different window-sizes (512, 2048, 5120) for the European super population (EUR) to investigate the effect and robustness of our method, as well as only using SNPs that overlap with the available high density genotype chip data to also explore HaploNet’s capabilities for genotype chip datasets. A more detailed description of the network architecture used in each scenario can be found in the supplementary material.

2.5.1 Computational comparisons

All models of HaploNet presented in this study have been trained on a machine with a NVIDIA GeForce RTX 2060 Super GPU (8GB VRAM), using CUDA v.11.0 and cuDNN v.8.0, and an Intel Core i5-9600 CPU (3.1 GHz). All other analyses have been performed on a cluster with an Intel Xeon E5-2699 CPU (2.2 GHz). The runtimes listed are therefore related to the system on which they have been run, unless otherwise specified.

In the smaller subsets of the 1000 Genomes Project dataset, we have used ChromoPainter (v.4.1.0) [26] to estimate the shared genome chunks between individuals in an unsupervised manner such that no population information is given and all individuals can be used as donors of each other. We are using their linked model that utilizes pre-estimated recombination rates from genetic maps of the human chromosomes to model the correlation between SNPs using default parameters. We used their own R library for performing PCA on the estimated chunk-counts matrix, and we have used PLINK (v.1.9) [9] for performing standard PCA on unphased genotypes. We heuristically evaluate the population structure inferred in the different methods by clustering accuracy in principal component space using Gaussian mixture models (GMMs) of the mclust R library [39] based on population labels. We have only used the principal components that capture population structure. We further compare the performance of our EM algorithm for estimating ancestry proportions on our NN likelihoods to the widely used ADMIXTURE (v.1.3) [2] software across all scenarios. ADMIXTURE uses unphased genotypes.

3 Results

3.1 Subsets of the 1000 Genomes Project

We applied HaploNet to three subsets of the phased whole-genome sequencing data of the 1000 Genomes Project in order to identify fine-scale population structure using haplotype information, and we compare our method to standard PCA, PCA from ChromoPainter and admixture proportions estimated from ADMIXTURE. We have summarized the runtimes for HaploNet and ChromoPainter in each of the three scenarios in Table 1.

We first analyzed a subset of the European populations (CEU, GBR, IBS, TSI) excluding FIN to capture fine-scale structure patterns. After filtering, the dataset consists of 404 individuals and 6.0 million SNPs. We have displayed the inferred population structure in Figure 3 and S2. The
benefit of using haplotype information is immediately clear, as HaploNet and ChromoPainter are able to separate the Southern European populations on the second principal component, which is not possible in standard PCA. In this scenario, ChromoPainter had a runtime of 2.7 days for the longest chromosome (chromosome 2), while it took 1.0 hours of training in HaploNet on the GPU setup. When looking at the estimated ancestry proportions, it is also clear that HaploNet has estimated more distinct components between Northern and Southern Europe in comparison to ADMIXTURE, assuming $K = 2$ (Figure S2).

We also analyzed the Chinese populations (CHB, CHS and CDX) that consist of 301 individuals and 5.5 million common SNPs. The results are visualized in Figure 4 and S3, where we see that HaploNet is very good at capturing fine-scale structure at both PCA and ancestry proportions level. We again capture similar information in comparison to ChromoPainter, where both methods reduce the within population variance, except for a few outliers, however, standard PCA also captures similar overall population structure. The runtime was 1.3 days for ChromoPainter, while it was 0.7 hours for training HaploNet on chromosome 2. However in terms of estimating ancestry proportions, we are able to separate the sources of ancestry in the two Han Chinese populations (CHB and CHS) to a degree which is not possible in ADMIXTURE for $K = 3$. In the PCAs, a CHS individual clusters in the center of the CHB population, which is also perfectly captured by HaploNet while not being captured by ADMIXTURE due to noise. We did observe convergence issues for ADMIXTURE in this scenario for $K = 3$.

In the last of the smaller subsets, we analyzed the African populations excluding ACB and ASW that have European admixture, such that the dataset consists of 504 individuals and 8.5 million SNPs. The PCA plots are displayed in Figure S4 and the estimated ancestry proportions in Figure 5. It appears, as populations are more genetically differentiated, that the three PCA methods are performing similarly and there are no immediate advantages of using HaploNet and ChromoPainter for PCA. However, ChromoPainter seems to have less variance in its populations leading to tighter clustering but also less separation between the two Nigerian populations, ESN and YRI. On chromosome 2, it took 6.1 days to run ChromoPainter on this African subset of the 1000 Genomes Project data, while it only took 1.8 hours to train HaploNet. For ancestry proportions, we see a striking difference between HaploNet and ADMIXTURE, assuming $K = 5$, where ADMIXTURE

![Figure 3](image-url)
is not able to utilize the fifth ancestry component for capturing structure at all. HaploNet is able to split up the ESN and YRI, while overall also estimating more distinct components of the populations. We did observe convergence issues for ADMIXTURE in this scenario for $K = 5$.

We evaluated the inferred global population structure by the three approaches using a Gaussian mixture model (mclust) to cluster the populations and the results are described in Table S1. We see that the two methods based on haplotype information are much better at distinguishing the European populations (here assuming CEU and GBR to be a single population), which would also be expected based on the PCA plots. For the other two scenarios, we do not observe any major differences between the methods.

**Figure 4:** Estimated ancestry proportions of the three Chinese populations in the 1000 Genomes Project based on HaploNet and ADMIXTURE, respectively, for $K = 3$. 
Figure 5: Estimated ancestry proportions of the five African populations in the 1000 Genomes Project without European admixture based on HaploNet and ADMIXTURE, respectively, for $K = 5$.

### 3.2 Super populations

We also applied HaploNet to all super populations in the 1000 Genomes Project and inferred global population structure and estimated ancestry proportions. The results are visualized for AFR, AMR, EAS, EUR and SAS in Figure S5, S6, S7, S8 and S9, respectively. We observe a general pattern when performing PCA in the different super populations, where HaploNet is inferring similar global population structure to standard PCA, though with tighter population clustering. However for EAS and EUR, HaploNet is also capturing structure on the third principal component, which is not captured using standard PCA. When estimating ancestry proportions, we see a general pattern of HaploNet capturing much more fine-scale structure by improving the signal-to-noise ratio in comparison to ADMIXTURE, as was also seen for the smaller subsets. For the EAS super population, the admixed Japanese (JPT) individual is also perfectly captured by HaploNet.

We filtered out variable sites not overlapping with the high density genotype chip of the Omni
Table 1: Runtimes for chromosome 2 in the three smaller subsets of the 1000 Genomes Project using HaploNet and ChromoPainter. HaploNet has been trained on a GPU machine as well as on the same cluster that ChromoPainter has been tested on using CPU with 64 threads.

<table>
<thead>
<tr>
<th></th>
<th>HaploNet (GPU)</th>
<th>HaploNet (CPU)</th>
<th>ChromoPainter</th>
</tr>
</thead>
<tbody>
<tr>
<td>African</td>
<td>1.8 hours</td>
<td>4.0 hours</td>
<td>146.0 hours</td>
</tr>
<tr>
<td>Chinese</td>
<td>0.7 hours</td>
<td>1.5 hours</td>
<td>30.3 hours</td>
</tr>
<tr>
<td>European</td>
<td>1.0 hours</td>
<td>2.4 hours</td>
<td>64.0 hours</td>
</tr>
</tbody>
</table>

platform [1] in European super population (EUR) to test how well HaploNet would perform in this scenario. After filtering, the dataset contained of 1.3 million SNPs. The results are visualized in Figure S12, where HaploNet has similar performance in comparison to when it is used on the full dataset, however, it has a slight loss of resolution for estimating ancestry proportions.

3.3 Changing window size

To test the robustness of HaploNet in terms of model hyperparameters, we applied it to the European super population with varying window sizes and evaluated its performance of inferring global population structure based on both PCA and ancestry proportions. The results are displayed in Figure S10 and S11. For the PCA plots that are inferred using haplotype encodings, we observe that HaploNet is not able to capture the structure on the third principal component for a small window size of 512, however, it is capable of capturing the structure for all larger window sizes. Small differences between the larger window sizes could be due to the stochastic nature of our model framework. The ancestry proportions are estimated based on our generated neural network likelihoods, where we see an opposite pattern with smaller window sizes showing fine-scale resolution and the usage of larger window sizes results in coarser ancestry proportions estimations.

3.4 Reconstruction and simulation

With the generative properties of the HaploNet framework, we are able to visualize and evaluate what our neural networks are learning. In Figure 6 and S13, we show two arbitrary chosen segments on chromosome 2 trained on the European super population (EUR). In both segments we see that HaploNet is fully capable of reconstructing the haplotype matrix from the haplotype encodings with accuracies of 99.7% and 99.2% for predicting the correct alleles of the haplotypes in the two segments, respectively. We also display the most probable haplotypes of the inferred haplotype clusters, as described in section 2.1.2, to gain insight into the haplotype clustering process in HaploNet. It has used 7 and 6 haplotype clusters to model the haplotype encodings in the two different segments, respectively.

We are able to generate new haplotypes by sampling from the learnt generative probability distributions. For each of the five European populations, we used the haplotype encoding of the most probable haplotype cluster in the population of each window to sample the full genome of 50 new haplotypes. As shown in Figure S14, the generated haplotypes demonstrate the same structure patterns as seen in the real data, Figure S8.
4 Discussion

We have presented our new framework, HaploNet, which performs dimensionality reduction and clustering on haplotypes using neural networks. We explored its capability to infer population structure based on local parameters, which are estimated in windows along the genome, as well as its generative properties for reconstruction and simulation. We show the benefits of merging machine learning with traditional statistical frameworks, as we develop a novel method for estimating ancestry proportions from neural networks in a likelihood model. We tested HaploNet on data from the 1000 Genomes Project and compared its results to commonly used software for inferring population structure using PCA or ancestry proportions. We demonstrate that HaploNet is capable of utilizing haplotype information for inferring fine-scale population structure similarly to ChromoPainter while being much faster. Besides performing PCA, we also demonstrate that we are capable of estimating ancestry proportions to a degree which is not possible for standard software, such as ADMIXTURE, where we utilize neural network architecture to incorporate haplotype information. This architecture further allows us to explore and understand the processes for encoding and clustering the haplotypes through the learnt mappings.

The number of clusters, $C$, is usually a non-trivial hyperparameter to set in a Gaussian mixture model or in a genetic clustering setting. Multiple runs of varying $C$ are usually performed and evaluated based on some criteria. In our study, we saw that HaploNet seemed capable of inferring the number of haplotype clusters to use by setting $C$ to a high fixed number. It would then only use a subset of the haplotype clusters to model the haplotype encodings, which has also been observed in a different application of the GMVAE model [7]. This is especially beneficial for our approach, as it would not be feasible to verify or validate $C$ along the genome of varying window sizes, though future work into this behaviour would be of high interest for the GMVAE model.

A limitation of our model, as for all deep learning models, is that only relatively small genetic datasets are usually available to researchers, which introduces problems with training convergence and overfitting. We have therefore kept the number of parameters in our model low, while using
an autoencoder architecture that naturally regularizes its reconstruction performance. Another advantage of having a small model configuration is observed with the low training times on both GPU and CPU setups that broadens the application opportunities of HaploNet. However, the difference between GPU and CPU will be larger when running chromosomes in parallel. For all analyses in this study, we have been using the entire data for training our neural networks to model all available haplotypes, whereas with larger datasets, one would usually partition the data in three; one for training, one for tuning hyperparameters of the model and one for ensuring that the trained model generalizes to new data. We have also limited ourselves to using fixed window lengths across the chromosomes for the matter of simplicity and ease of use, where we instead could have included external information from genetic maps to define windows of variable size. We show that HaploNet is somewhat robust to changes in window size when inferring global population structure, either using PCA or ancestry proportions, however we find a trade-off in resolution between the two measures as a function of the window size that is subject for future research. We further show that we are still able to capture much of the fine-scale structure when only evaluating variable sites available on a common genotype chip that allows for broader applications of our method.

Our model also serves as a proof-of-concept and an exploration for how non-linear neural networks and specialized architectures can be utilized to learn haplotype clusters and encodings across a full genome in a very scalable procedure. We hypothesize that as the number of large-scale genetic datasets are growing, we will see the increasing importance of deep learning in population genetics, as deeper models can be trained and more bespoke architectures can be developed. As also shown in our study, we can even use learnt encodings or estimated parameters together with standard statistical frameworks to further improve our understanding of genetic variation. Future developments of our framework would be to use the haplotype clusters and encodings in sequential models, to e.g. infer local ancestry in a semi-supervised approach, as well as investigate its potential properties for imputation based on reconstruction.

5 Data availability statement

The 1000 Genomes Project phase 3 data used in this study is publically available at https://www.internationalgenome.org/category/phase-3/.

6 Acknowledgements

The study was supported by the Lundbeck foundation (R215-2015-4174).

References


Supplementary Material

Derivation of the evidence lower bound (ELBO)

We here derive the ELBO from the marginal log-likelihood of a haplotype, $\mathbf{x}$, using Jensen’s inequality and the following definitions, $p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{y}) = p_\theta(\mathbf{x} | \mathbf{z})p_\theta(\mathbf{z} | \mathbf{y})p(\mathbf{y})$, $q_\phi(\mathbf{z}, \mathbf{y} | \mathbf{x}) = q_\phi(\mathbf{z} | \mathbf{x}, \mathbf{y})q_\phi(\mathbf{y} | \mathbf{x})$.

We follow the mathematical used in the machine learning literature and Kingma and Welling [25].

$$\log p_\theta(\mathbf{x}) = \log \sum_y \int_q p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{y}) \, d\mathbf{z}$$

(S1)

$$= \log \sum_y \int_q p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{y}) \frac{q_\phi(\mathbf{z}, \mathbf{y} | \mathbf{x})}{q_\phi(\mathbf{z} | \mathbf{x})} \, d\mathbf{z}$$

(S2)

$$= \log \mathbb{E}_{q_\phi(\mathbf{z}, \mathbf{y} | \mathbf{x})} \left[ \frac{p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{y})}{q_\phi(\mathbf{z} | \mathbf{x})} \right]$$

(S3)

$$\geq \mathbb{E}_{q_\phi(\mathbf{z}, \mathbf{y} | \mathbf{x})} \left[ \log \frac{p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{y})}{q_\phi(\mathbf{z} | \mathbf{x})} \right]$$

(S4)

$$= \mathbb{E}_{q_\phi(\mathbf{z}, \mathbf{y} | \mathbf{x})} \left[ \log p_\theta(\mathbf{x} | \mathbf{z}) - \log \frac{q_\phi(\mathbf{z} | \mathbf{x}, \mathbf{y})}{p_\theta(\mathbf{z} | \mathbf{y})} - \log \frac{q_\phi(\mathbf{y} | \mathbf{x})}{p(\mathbf{y})} \right].$$

(S5)

We can rewrite the last term in the expectation using $p(\mathbf{y} = c) = C^{-1}$, for $c = 1, \ldots, C$:

$$\mathbb{E}_{q_\phi(\mathbf{z}, \mathbf{y} | \mathbf{x})} \left[ \log \frac{q_\phi(\mathbf{y} | \mathbf{x})}{p(\mathbf{y})} \right] = \mathbb{E}_{q_\phi(\mathbf{y} | \mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}, \mathbf{y} | \mathbf{x})} \left[ \log \frac{q_\phi(\mathbf{y} | \mathbf{x})}{p(\mathbf{y})} \right]$$

(S7)

$$= \mathbb{E}_{q_\phi(\mathbf{y} | \mathbf{x})} \left[ \log \frac{q_\phi(\mathbf{y} | \mathbf{x})}{p(\mathbf{y})} \right]$$

(S8)

$$= \mathbb{E}_{q_\phi(\mathbf{y} | \mathbf{x})} \left[ \log q_\phi(\mathbf{y} | \mathbf{x}) \right] - \mathbb{E}_{q_\phi(\mathbf{y} | \mathbf{x})} \left[ \log p(\mathbf{y}) \right]$$

(S9)

$$= -H(q_\phi(y | x)) + \log C,$$

(S10)

where $H(X)$ is the entropy of a given distribution. In that way, both the second and third term will act as regularization on the reconstruction term through the latent space distributed by $q_\phi(\mathbf{z}, \mathbf{y} | \mathbf{x})$.

ELBO with $\beta_{cat}$ weight term

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}, \mathbf{y} | \mathbf{x})} \left[ \log p_\theta(\mathbf{x} | \mathbf{z}) - \log \frac{q_\phi(\mathbf{z} | \mathbf{x}, \mathbf{y})}{p_\theta(\mathbf{z} | \mathbf{y})} \right] - \beta_{cat} (-H(q_\phi(\mathbf{y} | \mathbf{x})) + \log C)$$

(S11)
Derivation of EM algorithm

Here we derive the EM algorithm for estimating ancestry proportions $Q \in [0,1]^{N \times K}$ and window-based haplotype frequencies $F \in [0,1]^{W \times K \times C}$, and we will use the following notation:

- $W$ number of windows
- $N$ number of individuals
- $C$ number of haplotype states
- $K$ number of ancestral states
- $x_{i,a}^{(w)}$ observed data of the $a$-th haplotype of individual $i$ in window $w$
- $\psi^{(n)}$ auxiliary variable in iteration $n$
- $q_{ik}^{(n)}$ ancestry proportion of state $k$ for individual $i$ in iteration $n$
- $f_{wkc}^{(n)}$ ancestral frequency of haplotype state $c$ in ancestral state $k$ in window $w$ in iteration $n$

We define the likelihood function:

$$L(Q,F;X) \propto W \prod_{w=1}^{W} \prod_{i=1}^{N} \prod_{a=1}^{2} p(x_{i,a}^{(w)} | \psi)$$  \hspace{1cm} (S12)

$$p(x_{i,a}^{(w)} | \psi) = \sum_{k=1}^{K} \sum_{c=1}^{C} p(x_{i,a}^{(w)} | \psi, k, y = c) p(y = c | \psi, c) p(k | \psi) \hspace{1cm} (S13)$$

$$= \sum_{k=1}^{K} \sum_{c=1}^{C} p(x_{i,a}^{(w)} | y = c) f_{wkc} q_{ik}. \hspace{1cm} (S14)$$

The EM update for $q_{ik}$ is defined as follows for iteration $n+1$:

$$q_{ik}^{(n+1)} = \frac{\sum_{w=1}^{W} \sum_{a=1}^{2} p(k | x_{i,a}^{(w)}, \psi^{(n)})}{\sum_{k=1}^{K} \sum_{w=1}^{W} \sum_{a=1}^{2} p(k | x_{i,a}^{(w)}, \psi^{(n)})} \hspace{1cm} (S15)$$

$$= \frac{\sum_{w=1}^{W} \sum_{a=1}^{2} p(k | x_{i,a}^{(w)}, \psi^{(n)})}{2W} \hspace{1cm} (S16)$$

and the update for $f_{wkc}$:

$$f_{wkc}^{(n+1)} = \frac{\sum_{i=1}^{N} \sum_{a=1}^{2} p(k,y = c | x_{i,a}^{(w)}, \psi^{(n)})}{\sum_{c'=1}^{C} \sum_{i=1}^{N} \sum_{a=1}^{2} p(k,c = k' | x_{i,a}^{(w)}, \psi^{(n)})} \hspace{1cm} (S17)$$

where
\[
p(k | x_{i,a}^{(w)} , \psi^{(n)}) = \sum_{c=1}^{C} p(k, y = c | x_{i,a}^{(w)} , \psi^{(n)}), \tag{S18}
\]

and
\[
p(k, y = c | x_{i,a}^{(w)} , \psi^{(n)}) = \frac{p(x_{i,a}^{(w)} | k, y = c, \psi^{(n)}) p(y = c, x_{i,a}^{(w)} | \psi^{(n)})}{\sum_{k'=1}^{K} \sum_{c'=1}^{C} p(x_{i,a}^{(w)} | k', y = c', \psi^{(n)}) p(y = c', x_{i,a}^{(w)} | \psi^{(n)})} \tag{S19}
\]
\[
= \frac{p(x_{i,a}^{(w)} | y = c) p(y = c | k, \psi^{(n)}) p(k | \psi^{(n)})}{\sum_{k'=1}^{K} \sum_{c'=1}^{C} p(x_{i,a}^{(w)} | y = c') p(y = c' | k', \psi^{(n)}) p(k' | \psi^{(n)})} \tag{S20}
\]
\[
= \frac{p(x_{i,a}^{(w)} | y = c) f_{w k c}^{(n)} q_{ik}^{(n)}}{\sum_{k'=1}^{K} \sum_{c'=1}^{C} p(x_{i,a}^{(w)} | y = c') f_{w k' c'}^{(n)} q_{ik'}}. \tag{S21}
\]
Network architectures

In all models, the architectures for the networks of GMVAE were defined as follows:

- $\pi_\phi(x): L-H-ELU-BN-C$
- $\mu_\phi(x, y), \sigma_2^2(x, y): (L + C)-H-ELU-BN-D$
- $\mu_\theta(y), \sigma_2^2(y): C-D$
- $\pi_\theta(z): D-H-ELU-BN-L$

with a hidden layer of 256 units ($H = 256$), ELU being an ELU activation layer, and BN a batch normalization layer. For a window size of $L = 1024$, the number of model parameters is $8.5 \times 10^5$.

AFR - including smaller subset (GWD, ESN, MSL, YRI, LWK)

- Window size: $L = 1024$
- Dimensions: $D = 64, C = 16$
- Batch size: 128
- Epochs: 200

AMR

- Window size: $L = 1024$
- Dimensions: $D = 64, C = 16$
- Batch size: 128
- Epochs: 200

EAS - including smaller subset (CDX, CHB, CHS)

- Window size: $L = 1024$
- Dimensions: $D = 64, C = 16$
- Batch size: 128
- Epochs: 200

EUR - including smaller subset (CEU, GBR, IBS, TSI)$^3$

- Window size: $L = \{512, 1024, 2048, 5120\}$
- Dimensions: $D = 64, C = 16$
- Batch size: 128
- Epochs: 200

SAS

- Window size: $L = 1024$
- Dimensions: $D = 64, C = 16$
- Batch size: 128
- Epochs: 200

$^3 L = 256, H = 128, D = 32$ were used for only evaluating sites of the high density genotype chip.
Supplementary figures and tables
Gumbel-Softmax sampling

Figure S1: Two examples of approximate sampling from different categorical distributions using the Gumbel-Softmax distribution. First column displays the two categorical distributions with the two examples distinguishable by color, and the next three columns display random samplings from the distributions. The random sampling is performed using the reparameterization trick as described in Equation 11 with $\tau = 10, 1, 0.1$.

**a)** A discrete uniform distribution, **b)** a categorical distribution with decreasing probabilities. Here it can be seen that as $\tau$ approaches 0, the sample from the Gumbel-Softmax distribution will converge to a categorical sample, while as $\tau$ approaches $\infty$, it will converge towards a uniform sample.
Figure S2: Estimated ancestry proportions of the European populations excluding FIN in the 1000 Genomes Project based on HaploNet and ADMIXTURE, respectively, for $K = 2$.

Figure S3: Inferred population structure based on PCA in the Chinese populations (CDX, CHB, CHS) of the 1000 Genomes Project. The PCA plots show the top two inferred principal components using standard PCA, HaploNet and ChromoPainter, respectively.
Figure S4: Global population structure of the African populations without European admixture in the 1000 Genomes Project. Top left figure displays the pairwise plots of the top 3 inferred principal components using standard PCA, the top right figure displays the pairwise plots of the principal components inferred by HaploNet and the bottom figure shows the pairwise plots of the principal components inferred by ChromoPainter.

<table>
<thead>
<tr>
<th>Population</th>
<th>PCA</th>
<th>HaploNet</th>
<th>ChromoPainter</th>
</tr>
</thead>
<tbody>
<tr>
<td>African (4 clusters)</td>
<td>0.784 (3)</td>
<td>0.784 (3)</td>
<td>0.784 (3)</td>
</tr>
<tr>
<td>Chinese (3 clusters)</td>
<td>0.837 (1)</td>
<td>0.841 (1)</td>
<td>0.837 (1)</td>
</tr>
<tr>
<td>European (3 clusters)</td>
<td>0.592 (1)</td>
<td>1.000 (2)</td>
<td>1.000 (2)</td>
</tr>
</tbody>
</table>

Table S1: Clustering accuracy in inferred PC-space based on population labels using GMMs in mclust for standard PCA, HaploNet and ChromoPainter. The number of PCs used for the clustering for each method is specified in parentheses, where only PCs capturing global structure are used.
Super populations

Figure S5: Global population structure inferred for the African (AFR) super population in the 1000 Genomes Project. Top left figure displays the pairwise plots of the top four inferred principal components from HaploNet, while the top right figure displays the results of standard PCA. The two bottom figures show the estimated ancestry proportions of HaploNet and ADMIXTURE for $K = 5$, respectively.
Figure S6: Global population structure inferred for the American (AMR) super population in the 1000 Genomes Project. Top left figure displays the pairwise plots of the top four inferred principal components from HaploNet, while the top right figure displays the results of standard PCA. The two bottom figures shows the estimated ancestry proportions of HaploNet and ADMIXTURE for $K = 3$, respectively.
Figure S7: Global population structure inferred for the East Asian (EAS) super population in the 1000 Genomes Project. Top left figure displays the pairwise plots of the top three inferred principal components from HaploNet, while the top right figure displays the results of standard PCA. The two bottom figures shows the estimated ancestry proportions of HaploNet and ADMIXTURE for $K = 4$, respectively.
Figure S8: Global population structure inferred for the European (EUR) super population in the 1000 Genomes Project. Top left figure displays the pairwise plots of the top three inferred principal components from HaploNet, while the top right figure displays the results of standard PCA. The two bottom figures shows the estimated ancestry proportions of HaploNet and ADMIXTURE for $K = 3$, respectively.
Figure S9: Global population structure inferred for the South Asian (SAS) super population in the 1000 Genomes Project. Top left figure displays the pairwise plots of the top four inferred principal components from HaploNet, while the top right figure displays the results of standard PCA. The two bottom figures shows the estimated ancestry proportions of HaploNet and ADMIXTURE for $K = 3$, respectively.
Window size in HaploNet

Figure S10: Inferred global population structure using PCA by HaploNet with different window sizes (512, 1024, 2048, 5120) in the European super population (EUR).
Figure S11: Estimated ancestry proportions by HaploNet using different window sizes (512, 1024, 2048, 5120) in the European super population (EUR) for $C = 3$. 
Figure S12: Inferred population structure in the European super population (EUR) using only sites of the high density genotype chip of the Omni platform in the 1000 Genomes Project. First row shows pairwise plots of the three top PCs inferred using HaploNet, while the second row shows the estimated ancestry proportions by HaploNet for $K = 3$. 
Reconstruction and simulation

**Figure S13:** The 200th window in chromosome 2 for the European super population (EUR). Left plot shows the true haplotype matrix with haplotypes ordered by their inferred haplotype cluster in this window by HaploNet. Middle plot shows the reconstruction of the haplotype matrix by HaploNet, while the last plot shows the reconstructions from the mean haplotype encoding of each haplotype cluster.

**Figure S14:** Inferred population structure of genomes generated from sequences of haplotype clusters. 50 haplotypes have been generated from each population in the European super population (EUR).