Methodology and Infrastructure for Statistical Computing in Genomics

Applications for Ancient DNA

KRISTIINA AUSMEES
Abstract

This thesis concerns the development and evaluation of computational methods for analysis of genetic data. A particular focus is on ancient DNA recovered from archaeological finds, the analysis of which has contributed to novel insights into human evolutionary and demographic history, while also introducing new challenges and the demand for specialized methods.

A main topic is that of imputation, or the inference of missing genotypes based on observed sequence data. We present results from a systematic evaluation of a common imputation pipeline on empirical ancient samples, and show that imputed data can constitute a realistic option for population-genetic analyses. We also develop a tool for genotype imputation that is based on the full probabilistic Li and Stephens model for haplotype frequencies and show that it can yield improved accuracy on particularly challenging data.

Another central subject in genomics and population genetics is that of data characterization methods that allow for visualization and exploratory analysis of complex information. We discuss challenges associated with performing dimensionality reduction of genetic data, demonstrating how the use of principal component analysis is sensitive to incomplete information and performing an evaluation of methods to handle unobserved genotypes. We also discuss the use of deep learning models as an alternative to traditional methods of data characterization in genomics and propose a framework based on convolutional autoencoders that we exemplify on the applications of dimensionality reduction and genetic clustering.

In genomics, as in other fields of research, increasing sizes of data sets are placing larger demands on efficient data management and compute infrastructures. The final part of this thesis addresses the use of cloud resources for facilitating data analysis in scientific applications. We present two different cloud-based solutions, and exemplify them on applications from genomics.

Keywords: statistical computing, genotype imputation, ancient DNA, deep learning, dimensionality reduction, genetic clustering, distributed computing

Kristiina Ausmees, Department of Information Technology, Division of Scientific Computing, Box 337, Uppsala University, SE-751 05 Uppsala, Sweden.

© Kristiina Ausmees 2022

ISSN 1651-6214
ISBN 978-91-513-1457-0
URN urn:nbn:se:uu:diva-470703 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-470703)
Dedicated to my family
List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Kristiina Ausmees, Federico Sanchez-Quinto, Mattias Jakobsson and Carl Nettelblad. An empirical evaluation of genotype imputation of ancient DNA. G3. Accepted.


Reprints were made with permission from the publishers.
List of additional papers

Although not explicitly discussed in the comprehensive summary, the following papers are related to the contents of this thesis.

1 Camille Clouard, Kristiina Ausmees and Carl Nettelblad. A Joint Use of Pooling And Imputation For Genotyping SNPs. *Submitted*.

# Contents

1 Introduction .......................................................................................................................... 11  
1.1 DNA Sequencing ........................................................................................................... 11  
1.2 Ancient DNA .................................................................................................................. 14  
1.3 Overview ......................................................................................................................... 14  

2 Imputation of Genotype Data .............................................................................................. 15  
2.1 Approximate Coalescent Models ....................................................................................... 15  
2.1.1 The Li and Stephens Model ....................................................................................... 16  
2.1.2 The Browning and Browning Model ......................................................................... 18  
2.2 Imputation of Ancient DNA ............................................................................................ 20  
2.2.1 Evaluation of a Prevalent Pipeline .......................................................................... 22  
2.2.2 Towards a Specialized Method .................................................................................. 23  

3 Characterization of Genotype Data .................................................................................... 26  
3.1 Principal Component Analysis ......................................................................................... 26  
3.1.1 Methods to Handle Missing Data in PCA ................................................................ 27  
3.2 The Structure Model ...................................................................................................... 29  
3.3 Deep Learning ................................................................................................................. 30  
3.3.1 A Convolutional Autoencoder for Characterization of Genetic Variation .............. 31  
3.3.2 Notes on Applying Deep Learning to Genotype Data ............................................. 33  

4 Scalable Data Analysis via Cloud Computing ....................................................................... 36  

5 Conclusion and Outlook ..................................................................................................... 40  

6 Summary of Papers ............................................................................................................ 42  
Sammanfattning på Svenska ........................................................................................................ 45  
Acknowledgments ..................................................................................................................... 47  
References .................................................................................................................................. 48
1. Introduction

This thesis concerns the development of computational methods in the field of genomics. The main topic is statistical methods for analysis of genetic sequencing data, with a focus on addressing the challenges posed by ancient DNA (aDNA) recovered from archaeological finds.

Various topics that are encompassed by the greater subject of scientific computing are addressed, including the development of theory as well as achieving its efficient implementation in terms of software and execution. This chapter provides a brief overview of the biological background relevant to understanding the data of interest, followed by an outline of the remainder of the thesis.

1.1 DNA Sequencing

The advent of next-generation sequencing (NGS) technologies has removed many of the barriers previously associated with genome sequencing [1], leading to a rapid increase in the amount of available data and accelerating various areas of genomic research [2]. The set of technologies referred to as NGS generally involves the breaking of the genetic sequence into smaller fragments, determining their base pairs, and subsequently reconstructing the sequence [3]. The methodology has led to great improvements in throughput and cost compared to earlier approaches, and allowed for the automation of large parts the sequencing process.

The analysis process of NGS data can be divided into three stages, visualized in Figure 1.1. In the primary stage, raw data from the sequencing machine is analyzed to determine the bases of the generated fragments, producing reads, or short sequences of nucleotides. The second stage consists of reconstructing the reads to get a cohesive representation of the sample’s DNA sequence. This can be done either by means of de-novo assembly or by alignment and variant calling [4, 5].

In this thesis, the main focus is on the latter method, in which the reads are aligned, or mapped, to a reference genome based on similarity. Once the alignment is done, the process of variant calling proceeds to determine differences between the sample’s genome and that of the reference. Such variations can be changes in a single base, denoted as single nucleotide polymorphisms (SNPs) or larger structural variations like inversions and deletions. Alignment and variant calling pipelines typically contain several data correction and filtering steps and may also make use of additional information such as variants that are previously known to exist.
Figure 1.1. Stages of analysis of NGS data. Sequencing machines fragment the DNA molecule, after which primary analysis determines the sequence of base pairs, producing reads. In the secondary analysis step, the reads are reconstructed by means of de-novo assembly (not pictured) or alignment to a reference and subsequent variant calling.

Tertiary analysis, in turn, comprises a wide variety of methods for interpretation of the data generated in the previous steps. Highly application-specific, this stage is of a more exploratory nature and can consist of various data aggregation, improvement and inference steps, often in combination with additional data such as patterns of genetic variation or phenotypic information.

The tertiary analysis is often based on statistical models and the ability to draw conclusions is largely dependent on the quality of the data studied. All three analysis stages typically contain quality-assessment steps and produce some measure of certainty of the results. In the alignment step, the degree of similarity between a read and the reference is an indication of the quality of the mapping. Some degree of inconsistency is typically allowed, as it can be due to base errors in the read or actual genetic differences between sample and reference. As reads carrying the allele present in the reference will have a higher chance of obtaining a high-quality mapping, a preference is introduced towards this allele. Denoted reference bias, this can also have an effect on downstream analysis of the data. Another factor that can impact the quality of aligned data is that a read with several similar matches within the reference genome is more uncertain. Such issues are more prevalent for short reads.

The number of reads that overlap a given nucleotide is denoted as the read depth, or coverage, at that locus. The average read depth over the sequence is referred to as the coverage of the sample. This impacts the subsequent variant calling step, as a higher number of overlapping reads facilitates determining the presence of a variant at a position.
In the variant calling step in Figure 1.1, the determined sequence of the target sample is displayed at the top. The leftmost red rectangle exemplifies a case in which a read is considered incorrect, e.g. due to errors in base calling or alignment, and disregarded. The red rectangle in the middle exemplifies a location with no read coverage at which the target sequence cannot be determined, and the rightmost one a case where a SNP is called. Variant calling is thus facilitated by a larger number of reads overlapping a position, and those called at loci with low coverage have a higher degree of uncertainty.

In the schematic of variant calling in Figure 1.1, the target comprises a single genetic sequence, which is the case when analyzing haploid organisms that have a single set of chromosomes. Humans are diploid organisms, as we inherit a copy of every chromosome from each parent. In this case, the reads that overlap a genomic position in the alignment can come from either of the two copies, which needs to be taken into account in the variant calling process.

The unordered pair of alleles that is determined at each genomic location for a diploid sample is denoted a genotype. For SNPs that have two possible alleles, the genotype can be coded as 0, 1, or 2, representing the count of a given allele. If the number of overlapping reads is low or genotypes cannot be confidently determined for other reasons, genotype likelihoods, defined as the probability of the read data given a genotype, might be preferable to use in downstream analysis rather than called genotypes.

In determining genotypes, information about the sequence of alleles in each individual chromosome copy, or haplotype, is lost. While genotypes can be informative about various aspects of genetic variation, as discussed in Chapter 3, haplotypes as fundamental units of inheritance have an essential role in the description and interpretation of genomes [6]. Estimation of haplotypes from genotype data, also called phasing, can be performed using statistical models, and such approaches form a basis for the methods of genotype imputation discussed in Chapter 2.

Such methods make use of the fact that certain combinations of variants are shared between individuals, forming patterns of haplotypes in populations. This haplotype structure, or non-random sharing of combinations of variants along the genome, is referred to as linkage disequilibrium (LD), and is characterized by a deviation of observed haplotype frequencies from those expected given the frequencies of the individual alleles they comprise [7]. LD can be informative of various genomic and historical processes and is influenced by different factors including population structure, migration, selection and genetic linkage. The latter is the tendency of sites that are closely located on a chromosome to be inherited together. During the cell division in which the genetic material that is passed on to the next generation is created, fragments from the paternal and maternal copies of each chromosome can cross over, or recombine, to create new combinations of alleles in chromosomes. The rate of recombination between two sites thus also affects their level of linkage.
1.2 Ancient DNA

The analysis of genetic information from aDNA has been a critical component in numerous discoveries regarding the history of modern humans. An important addition to paleontological and archaeological evidence, it has provided complementary information to existing theories, perhaps most notably the support leading to wide-spread acceptance of the out-of-Africa model for the origin of modern humans [8]. It has also led to novel insights into questions previously not easily explored, such as whether observed cultural variation has been driven by migration and admixture or the spread of ideas [9].

Several intrinsic challenges exist when analyzing aDNA. Contamination from other organisms requires rigorous laboratory practices as well as methods to distinguish endogenous DNA [10, 11]. Further, post-mortem damage causes fragmentation of the DNA molecule [12] as well as chemical changes to its nucleotides [13]. This can pose problems in the the alignment process described in Section 1.1, as shorter reads are more difficult to uniquely map to a reference [14] and damaged nucleotides can lead to higher mismatch rates. Such traits can also cause a particular sensitivity to reference bias [15].

While the number of genetically investigated ancient genomes grows every year, and had reached more than 1000 as of 2018, limitations in sample quality and prohibitive costs have led to most of the data having relatively low coverage of 1x or less [16]. The accuracy of variant calls can be greatly affected for such data, and determining diploid genotypes may not be feasible [17, 18]. These characteristics can limit the applicability of standard tools, and have led to the need for custom methods for variant calling and downstream analysis.

1.3 Overview

The remainder of the thesis is organized as follows. Chapter 2 concerns the subject of imputation, or the prediction of unobserved genotypes. We review two statistical frameworks based on haplotype phasing and outline Papers I and II in which different aspects of imputing aDNA are addressed.

The topic of Chapter 3 is the characterization of patterns in genetic variation based on genotype data. We first consider the use of principal component analysis for dimensionality reduction and the problems posed by sparse sample data, and introduce Paper III in which different strategies to address this issue are evaluated. We further discuss the use of data-driven models based on artificial neural networks as an alternative to standard methods. This is also addressed in Paper IV, where we propose a deep learning framework for genotype data and demonstrate its utility on various data characterization tasks.

Finally, Chapter 4 addresses practical issues related to computational infrastructure in light of the increasingly large amounts of genetic data that are becoming available. We introduce Papers V and VI, in which the use of cloud computing to enable scalable analysis of large data sets is explored.
2. Imputation of Genotype Data

Imputation in the genetic context refers to the prediction of unobserved variants and has become a standard tool in the analysis of sequencing data. In genome-wide association studies (GWAS), imputation has frequently been used to boost the power of inference [19] and allow for analyses that require high-resolution data, such as the identification of causal variants for diseases via fine-mapping [20]. Imputation can also facilitate population genetic analyses that require dense genotypes, and has been incorporated into several large-scale sequencing projects to increase the information content [21, 22, 23].

Many common methods for imputation target SNP data and are based on statistical models that use known variation from a set of reference individuals to infer genotypes at untyped sites. In Section 2.1, two such frameworks that form the basis for commonly used imputation tools are described: the Li and Stephens model and the Browning and Browning model.

As a means of boosting information content, imputation has the potential to increase the scientific returns of aDNA studies, which are often limited to low coverage data. However, the characteristics of ancient samples can cause lower imputation performance than for typical present-day data, and the applicability of existing imputation methods on aDNA is still not fully explored. In Section 2.2, we outline some of the issues surrounding imputation of aDNA, and present Papers I and II, in which the performance of different imputation methods on aDNA are evaluated.

2.1 Approximate Coalescent Models

Most widely used methods for genotype imputation are based on statistical models that infer haplotype phase of a sample conditional on its observed genotypes, and use the phase information to predict unobserved variants [24]. The basic idea that underlies many such phasing methods is that since genetic material is passed on via inheritance, new haplotypes are derived from old ones through the processes of mutation and recombination [25]. As these events tend to be rare, haplotypes in a population can be expected to be similar to each other.

This idea originates in coalescence theory [26], which describes how alleles sampled from a population may originate from a common ancestor. The methods most commonly used for phasing are based on approximations of the coalescent [27], which yield similar expected patterns of haplotypes, but are
computationally less expensive [28]. The approximate coalescent model gives rise to a Markovian structure along the genetic sequence that can be expressed in terms of a Hidden Markov Model (HMM) [29].

HMMs are a class of models that describe stochastic processes that generate sequences. They represent observed data being generated by an underlying Markov process that is unobserved, or hidden. Figure 2.1 shows the graphical representation of a HMM of length $M$ with hidden states $X_j$ and observations $Y_j$. Transition probabilities $P(X_j | X_{j-1})$ describe the probability of transitioning to a state $X_j$ at position $j$ in the sequence, given that the state in the previous position is $X_{j-1}$. Observation probabilities $P(Y_j | X_j)$ define the probability of observing value $Y_j$ given that the hidden state is $X_j$. The joint probability of the HMM model is given by Equation 2.1.

$$P(Y, X) = P(X_1) \prod_{j=2}^{M} P(X_j | X_{j-1}) \prod_{j=1}^{M} P(Y_j | X_j)$$ (2.1)

Figure 2.1. Graphical representation of a HMM, with hidden states $X_j$ and observed values $Y_j$. Arrows indicate conditional dependence relations.

Let $\mathcal{H}$ denote a set of $H$ reference haplotypes typed at $M$ loci, and $S$ a target sample with partially observed genotypes at the same set of positions. In HMM-based models for phasing and imputation, the observed values are the genotypes of $S$, and the hidden states represent the corresponding unobserved haplotypes. Sites in $\mathcal{H}$ at which $S$ has observed genotypes are used to inform the phasing, and imputation is performed at loci in $\mathcal{H}$ for which $S$ has unobserved genotypes. Based on this general framework, a variety of methods have been developed. These mainly differ in how the reference haplotypes $\mathcal{H}$ are used to model the hidden states, as well as in the type of inference that is performed. Two of the most prominent frameworks are the Li and Stephens model and the Browning and Browning model, which are outlined in Sections 2.1.1 and 2.1.2.

2.1.1 The Li and Stephens Model
The Li and Stephens framework [30] provides a means of modeling haplotype frequencies in a population, and forms the basis for many phasing and
imputation tools, including PHASE [31, 32], MACH [33], IMPUTE2 [34], Beagle versions 4.1 [35] and up, and GLIMPSE [36]. In this model, \( \mathcal{H} \) is a set of template haplotypes, and sample haplotypes are represented as mosaics of these templates. Below, we describe the diploid version of the framework, in which hidden states represent the pair of unobserved template haplotypes used for each of the sample’s two chromosomes.

For a set of reference haplotypes \( \mathcal{H} \) of size \( H \), typed at \( M \) loci and a sample individual \( S \), the following HMM over sites \( j \in \{1 \ldots M\} \) is defined:

- The observed states \( Y_j \) are the genotypes of the sample \( S \).
- The hidden states \( X_j = (x^1_j, x^2_j) \) are pairs that indicate the phase of the genotype of \( S \), where \( x^k \in \mathcal{H} \) denotes the template haplotype used for the \( k \)th chromosome. The state space of the HMM is thus \( (\mathcal{H} \times \mathcal{H}) \).
- The observation probabilities \( P(Y_j|X_j) \) are given in Table 2.1.
- The transition probabilities \( P(X_j|X_{j-1}) \) given in Equation 2.2.

| \( Y_j \) | \( G(X_j) \) |
|---|---|---|
| 0 | \( (1 - \epsilon_j)^2 \) | \( 1 - \epsilon_j \epsilon_j \) | \( \epsilon_j^2 \) |
| 1 | \( 2 \epsilon_j(1 - \epsilon_j) \) | \( (1 - \epsilon_j)^2 + \epsilon_j^2 \) | \( 2 \epsilon_j(1 - \epsilon_j) \) |
| 2 | \( \epsilon_j^2 \) | \( (1 - \epsilon_j) \epsilon_j \) | \( (1 - \epsilon_j)^2 \) |

Table 2.1. The observation probabilities \( P(Y_j|X_j) \) for a position \( j \) of the HMM based on the Li and Stephens framework, where \( G(X_j) \) denotes the genotype implied by the hidden state \( X_j \) and \( Y_j \) is the observed genotype.

The observation probabilities depend on the similarity between the genotype implied by the reference haplotypes used as a template for hidden state \( X_j \), denoted \( G(X_j) \), and the observed genotype \( Y_j \). Changes in alleles are modeled as occurring independently along a chromosome, with an error parameter \( \epsilon_j \) that represents the probability of a mutation or genotyping error causing the observed allele to change to its alternative. The probability of an observed genotype at location \( Y_j \) is thus obtained by summing over the possible allele changes that can lead the genotype \( G(X_j) \) to change to the genotype \( Y_j \).

\[
P(X_j|X_{j-1}) = \begin{cases} 
\frac{\theta_j^2}{H^2} & \text{if 2 switch} \\
(1 - \theta_j)\theta_j + \frac{\theta_j^2}{H^2} & \text{if 1 switch} \\
(1 - \theta_j)^2 + \frac{2(\theta_j - \theta_j^2)}{H} + \frac{\theta_j^2}{H^2} & \text{if 0 switch}
\end{cases} 
\]

(2.2)
The transition probabilities depend on how many switches occur between the states $X_{j-1}$ and $X_j$. A switch is defined as a change in reference haplotype along a chromosome, i.e. if $x_{j-1}^k \neq x_j^k$. The switches can be seen as representing historical recombination events occurring between loci $j-1$ and $j$, and are modeled using the parameter $\theta_j = 1 - e^{-\frac{4N_e d_j}{H}}$. This parameter captures the fact that the transition rate depends on the genetic distance $d_j$ between loci $j-1$ and $j$ and the effective population size $N_e$. The presence of $H$ in the formulas reflects that the more reference haplotypes we consider, the lower the transition rate becomes, and the longer we can expect the stretches of sequence that are shared with the sample to be.

2.1.2 The Browning and Browning Model

The Browning and Browning framework is based on a model of local haplotype clusters [37] that are defined from the reference haplotypes depending on their similarity at nearby sites. Using an example from [38], Figure 2.2 shows the localized haplotype cluster model for the set of haplotypes of length four in Table 2.2. The model is illustrated as a directed graph in which the leftmost node corresponds to the source, and there is a level for each of the four loci in the haplotypes. Nodes with solid circles represent allele 1 and those with dotted circles allele 2. Each node represents a local haplotype cluster. Every path from the source to a node in the final level represents a reference haplotype. Every edge $(n_i, n_j)$ of the graph is labeled with the relative frequency of haplotypes traversing the node $n_i$ that also traverse the edge $(n_i, n_j)$.

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1111</td>
<td>21</td>
</tr>
<tr>
<td>1112</td>
<td>79</td>
</tr>
<tr>
<td>1122</td>
<td>95</td>
</tr>
<tr>
<td>1221</td>
<td>116</td>
</tr>
<tr>
<td>2111</td>
<td>25</td>
</tr>
<tr>
<td>2112</td>
<td>112</td>
</tr>
<tr>
<td>2122</td>
<td>152</td>
</tr>
</tbody>
</table>

Table 2.2. Example set of haplotypes used to define a localized haplotype cluster. The haplotypes are of length four, and the numbers 1 and 2 denote different alleles. The count describes the number of each haplotype present in the set of individuals of interest.

For a set $H$ of $H$ reference haplotypes typed at $M$ loci, the corresponding localized haplotype cluster model $D$ is used to define a haploid HMM over a single chromosome of a sample $S$ in which
Figure 2.2. Graphical representation of the localized haplotype cluster model for the haplotypes in Table 2.2. Every path from the source node $I$ to a node in the final level $(m = 4)$ of the graph represents a haplotype in the table. Nodes with solid lines represent allele 1 and nodes with dotted lines represent allele 2. Labels on the edges represent the proportion of haplotypes leaving a node that traverse the given edge.

- the observed states $Y_j$ are the allele of the sample $S$ at sites $j \in \{1 \ldots M\}$
- the hidden states $X_j$ are haplotype clusters

Let $T(X_j)$ denote the allele that haplotype cluster $X_j$ represents, and $F(X_i, X_j)$ the relative frequency on the edge $(X_i, X_j)$ in $D$.

A haplotype cluster emits its associated allele with probability 1, and has observation probability 0 for all other alleles (Equation 2.3).

$$P(Y_j | X_j) = \begin{cases} 1 & \text{if } Y_j = T(X_j) \\ 0 & \text{otherwise} \end{cases} \quad (2.3)$$

The transition probabilities of the HMM are based on the edges in $D$ and their labels of relative haplotype frequencies (Equation 2.4), ensuring that the state space at each position $j$ and the possible transitions between states follow the structure of the localized haplotype cluster $D$.

$$P(X_j | X_{j-1}) = \begin{cases} F(X_{j-1}, X_j) & \text{if } (X_{j-1}, X_j) \text{ is an edge in } D \\ 0 & \text{otherwise} \end{cases} \quad (2.4)$$

The described haploid model can subsequently be turned into a diploid HMM that models the observed diploid genotypes of $S$ by considering ordered pairs of haplotype clusters. For each level $m$ of the graph $D$, let $L_m$ denote the set of haplotype clusters at level $m$. The state space of the diploid HMM is then $\bigcup_{m=1}^{M} (L_m \times L_m)$. For a state $(X_a, X_b)$, the unordered pair of alleles emitted by
and \( X_b \) in \( D \) represents a diploid genotype that is emitted with probability 1 by the state. Transition probabilities are obtained by the product of the haploid probabilities (Equation 2.5).

\[
P((X_a, X_b) | (X_c, X_d)) = P(X_a | X_c) P(X_b | X_d)
\]  

(2.5)

The Browning and Browning framework differs from that of Li and Stephens in several ways. The localized haplotype cluster model results in a more par-simonious method in the sense that it uses a smaller set of possible haplotype configurations to represent the target sample’s haplotypes. This has the effect of reducing the size of the model’s state space, as well as the possible number of transitions between them, lowering the computational burden. However, it also results in the fact that only haplotype configurations that are present in the reference can be expressed, whereas all possible haplotypes can be obtained for the target sample in the Li and Stephens model.

Further, recombination and mutation are not explicitly modeled as parameters and do therefore not need to be estimated or fit to the data. As observed genotypes that differ from the alleles associated with a haplotype cluster have observation probability 0, the effects of genotyping error are also not modeled.

The Browning and Browning model is implemented in the software Beagle, which is available in several versions. The algorithm described here corresponds to Beagle 4.0 [39], which is also the version used in Paper I. In this version, phasing and imputation proceeds by an iterative process in which a localized haplotype cluster is built based on phased input data, and used to build a HMM from which phased haplotypes are subsequently sampled. For a set of target samples, the phased input data can consist of both phased reference haplotypes as well as current estimates of haplotypes for the target samples.

The forward-backward algorithm is used to sample a set of phased haplotypes for each target sample conditional on the observed genotypes, which are subsequently used as input for the next iteration. If both alleles of \( Y_j \) are missing, the observation probability \( P(Y_j | X_j) = 1 \). If one allele is observed, then \( P(Y_j | X_j) = 1 \) if the observed allele is in the unordered pair of alleles emitted by \( X_j \), and \( P(Y_j | X_j) = 0 \) otherwise. In the final iteration, the Viterbi algorithm is used to calculate the most likely sequence of hidden states, resulting in a set of phased haplotypes for each sample.

### 2.2 Imputation of Ancient DNA

There are several challenges associated with imputation of aDNA. As imputation tends to perform best on high-quality and densely genotyped data [28], one issue is the relatively high prevalence of errors and missing data in ancient samples. It has been argued that in such cases it is preferable to use models that incorporate the uncertainty in the data by considering genotype
likelihoods rather than discrete, called genotypes [40, 41], an option that is supported by some tools.

Another challenge is the fact that only genetic variants present in the reference can be recreated using imputation. The panels of high-quality phased reference haplotypes that are available today consist of present-day samples, and may consequently not be representative of the genetic variation of ancient individuals. Some imputation models allow the set of reference haplotypes to be constructed from study individuals as well as the phased reference panel, which could alleviate this issue by the incorporation of ancient genetic variation in the model.

Previous studies have performed imputation on aDNA and successfully used the results in various population-genetic analyses [42, 43, 44, 45, 46]. Section 2.2.1 summarizes Paper I in which a study designed according to the imputation methodologies previously used in the aDNA community was performed. The goal was to provide a systematic evaluation that could inform study design and aid the interpretation of results from genetic analyses based on imputed data.

Existing imputation tools have generally been designed with present-day data in mind, with algorithm development largely driven by advances in genotyping technology that have lead to increasing quality and quantity of sequenced individuals [47]. While this development has resulted in highly effective and specialized tools that can be applied to samples of tens or hundreds of thousands of individuals, it is not necessarily reflective of the requirements of aDNA studies. In this field, smaller samples with lower coverage and quality are more typical, and algorithm simplifications made for the sake of optimizing speed may have larger trade-offs in accuracy than for present-day data.

The question of adapting imputation methods to ancient DNA is addressed in Paper II, where we develop an imputation tool based on the full probabilistic Li and Stephens model. In Section 2.2.2, we outline the model and its implementation, show that it achieves comparatively high accuracy on empirical and simulated low coverage ancient data, and discuss potential extensions for further customization to aDNA.

Similarly to previous studies of aDNA, the methodology in Papers I and II for assessment of imputation performance was based on using ancient individuals for which high coverage data is available. The same set of 5 individuals, with coverages between 19x and 57x, was used in both papers, using the original dense read data to call a set of high-quality genotypes considered as gold standard. Low coverage data for these individuals was subsequently generated by subsampling reads and inferring genotype likelihoods from the sparser data. These likelihoods were used as input to imputation, and the resulting genotypes were finally compared to the gold standard ones to give an indication of imputation accuracy.

The five high coverage samples used for performance evaluation were sf12 [48], Loschbour and LBK [49], ne1 [42] and ans17 [50]. The first two, sf12
and Loschbour, stem from hunter-gatherer individuals from around 8000 years ago. The sf12 sample is from a site on the Baltic Sea island of Stora Karlsö in Sweden, and the other sample is named after the Loschbour rock shelter in Luxembourg where it was found. The remaining samples are derived from individuals from early farming cultures in Europe. The LBK sample is an individual from around 7000 years ago found in Stuttgart, Germany, and ne1 is a roughly contemporary individual from Polgár Ferenci hát, Hungary. The ans17 individual is the most recent sample, found at the Ansarve site on the island of Gotland in Sweden, and dated to roughly 5000 years ago.

2.2.1 Evaluation of a Prevalent Pipeline

In the study performed in Paper I, imputation was performed using the software Beagle 4.0, which implements the Browning and Browning model outlined in Section 2.1.2, using genotype likelihoods as input. The phased reference panel was the 1000 Genomes [21] phase 3 data set, using chromosomes 1-22 filtered to keep biallelic SNPs with a Minor Allele Count (MAC) of at least 5.

As the five high coverage individuals used for performance evaluation were found in European locations, it is reasonable to consider the European subset of the reference panel as genetically most representative. Computational time can be saved by using a smaller reference, and some studies of imputation of present-day samples have indicated that population-specific panels result in the highest performance [51, 52]. We compared use of the entire reference to only considering the European subset, and found that the larger reference did improve accuracy, suggesting that considering larger and more diverse references is motivated when imputing aDNA.

As previously discussed, the fact that only panels of modern individuals are available for use as a reference may be problematic for the imputation of ancient samples. While the selected imputation method allows for the inclusion of other ancient samples in the inference, the benefits of considering more closely related samples may be diminished by the fact that the majority are of low coverage and quality.

This was investigated by comparison of performance of imputation when performed independently on each evaluation individual to performing joint imputation of a sample consisting of 61 ancient individuals with coverages between 0.1x and 16x, in addition to the evaluation individuals. The inclusion of the additional ancient samples resulted in increased accuracy, indicating that information in low to moderate coverage samples can be leveraged to improve imputation performance.

An in-depth analysis of imputed data showed that overall concordance of over 99% with the gold standard genotypes was obtained for data with 1x coverage, and that error in heterozygote genotypes reached similar levels for
variants with population minor allele frequencies of at least 10%. Our results confirm and extend those from previous studies, and motivate the use of imputed data for downstream population-genetic analyses.

The study also suggested areas of improvement for imputation of aDNA. For rare variation, the effects of different reference sets was more pronounced, and the tendency to reference bias was higher. A significant decrease in imputation accuracy was also observed from around coverage 0.75x and below. The two hunter-gatherer samples, which are more distant from the genetic profile of modern populations than the individuals from early agricultural cultures, exhibited lower performance overall.

2.2.2 Towards a Specialized Method

In Paper II, we present a tool for genotype imputation based on the Li and Stephens model described in Section 2.1.1. Unlike many other implementations based on this underlying model, we did not make modifications or simplifications to reduce the state space, but considered the full probabilistic model of all haplotype configurations and performed phasing and imputation of missing data jointly. The motivation was to assess if this approach could prove advantageous for particularly challenging data in terms of low coverage and sequencing errors, and to lay the foundation for further development towards a method that is customized for aDNA.

The implemented method, denoted prophaser, was designed to incorporate uncertainty in the observed data by integrating over the possible genotypes at each site. This was done by modifying the HMM described in Section 2.1 to have as observations vectors of genotype likelihoods rather than called genotypes. Only biallelic SNPs were considered, yielding observations on the form \( Y = [L(0), L(1), L(2)] \), where \( L(g) \) denotes the likelihood of genotype \( g \). This resulted in a HMM with the joint probability given in Equation 2.6.

\[
P(Y, X) = P(X_1) \prod_{j=2}^{M} P(X_j|X_{j-1}) \prod_{j=1}^{M} \sum_{g=0}^{2} P(g|X_j)L(g) \quad (2.6)
\]

Posterior probabilities \( P(X_j|Y_1, \ldots, Y_M) \) for each state were calculated using the forward-backward algorithm for the entire sequence \( j = 1, \ldots, M \), after which the most likely genotype at each site was obtained by marginalizing over the possible states.

As the state space of the model is of size \( H^2 \) and the number of computations of the forward-backward algorithm scales quadratically with number of states, optimization is required to increase computational efficiency. A technique described in [33] that makes use of regular patterns in the transition matrices was implemented, reducing the computational requirements from \( O(H^4) \) to \( O(H^2) \). An extension of the scheme to increase memory efficiency described in the same paper was also implemented, which allowed the
code to be run in an efficient manner on a conventional graphics processing unit (GPU).

The proposed method was compared to two imputation methodologies based on the Beagle software, which is available in several versions that vary in underlying algorithm and implementation. The first pipeline corresponded to the one in Paper I, where Beagle 4.0 was used with genotype likelihoods as input to perform imputation in a single step. The second pipeline consisted of two steps and was recommended in [53], based on an evaluation of several different imputation methodologies on a low coverage ancient sample. In this pipeline, Beagle 4.1 was first used to determine sites where genotypes could be confidently called based on the likelihoods, which were then used as input to Beagle 5.2 for performing imputation at unobserved sites.

Imputation accuracy was evaluated for the three pipelines on two data sets, one consisting of the five high coverage empirical ancient samples described in Section 2.2 and the second consisting of 10 present-day samples for which simulated low coverage data was generated. For all methods, imputation was performed on chromosome 20 using the 1000 Genomes [21] phase 3 data set, filtered to only include biallelic SNPs.

![Figure 2.3](image)

**Figure 2.3.** Top: concordance at heterozygote sites for genotypes imputed from empirical ancient samples downsampled to different coverage levels. Bottom: the fraction of SNPs retained for evaluation after application a post-imputation filter based on posterior genotype probability (GP). For both plots, results are averaged over the five individuals, and standard deviation is indicated by shaded regions. The legend indicates the GP threshold used in the post-imputation filter for the different methodologies.

Imputation tools typically output a measure of confidence of the results which can be used to filter out genotypes with high uncertainty. In assessment
of imputation performance, it is thus relevant to consider both accuracy as well as the amount of sites retained after the post-imputation filter. In Paper II, we measured accuracy by the genotype concordance, defined as the fraction of sites where the imputed genotype was the same as the gold standard one, and performed post-imputation filtering by requiring the posterior genotype probability (GP) of the most probable genotype to reach a certain threshold. As heterozygote sites tend to be more difficult to impute correctly and are less subject to chance agreement to the reference majority allele, we focused on evaluating performance for sites in which the gold standard genotype was heterozygous.

Figure 2.3 shows results on the empirical data set for different coverage levels, with post-imputation GP thresholds adjusted for the different imputation methodologies, prioritizing similar numbers of evaluated sites at lower coverages in particular. We found that prophaser outperformed the other methodologies overall, with advantages most pronounced at coverages below 0.5x, where more variants could be retained, while maintaining a higher concordance.

Analysis over the allele frequency spectrum further indicated that rare variants were captured to a higher extent by prophaser, a difference that was discernible even for higher coverages where the methodologies tended to have more similar overall performance. Similar observations were made on the simulated data set, and execution of prophaser on GPUs yielded comparable runtimes to the other pipelines. The study suggested that the proposed approach is a promising alternative for imputation of particularly challenging data and provides a basis for further customization for aDNA, which we discuss more in Chapter 5.
3. Characterization of Genotype Data

Characterization methods that identify and condense relevant features from complex data are an important tool for making sense of the increasingly large and diverse genetic data sets that are becoming available. A central component of exploratory data analysis, characterization methods can highlight potential inquiries of interest or otherwise inform the choice of methodology, such as checking the validity of modeling assumptions for a particular data set in population genetic analyses [54]. Descriptive methods are also used in quantitative and medical genomics, e.g. to incorporate ancestry information into predictive models of traits and individual disease risk [55, 56], where it can improve accuracy and aid in reducing health disparities due to genetic differences [57].

A very widely used method for summarizing high-dimensional data is PCA, in which a linear transformation is performed to obtain a lower-dimensional representation that allows for visualization and identification of relevant patterns in the data. In section 3.1, we outline the application of PCA for genotype data, discuss how the presence of missing values can pose a problem, and present Paper III in which methods to address this issue are evaluated.

A central concern in the analysis of genotype data is the description of similarity between samples that allows for portrayal of genetic relationships and clustering of genetically linked individuals in a population. While PCA can be used to describe such population structure along continuous axes of variation in a data-driven manner, another class of methods that are standard in the field are instead based on discrete population admixture models [58]. This approach defines cluster membership in terms of a number of genetically distinct source populations, and was first introduced in the Structure method, which is outlined in Section 3.2.

Finally, Section 3.3 is dedicated to the use of deep learning (DL) methods in genomics and population genetics. We discuss how DL represents a different paradigm to many standard methods used in these fields, and summarize Paper IV in which we present a DL-based framework for characterization of genotype data and demonstrate its utility for performing dimensionality reduction and genetic clustering.

3.1 Principal Component Analysis

PCA [59] is a statistical method for analyzing multivariate data in which a transformation is defined from a set of possibly correlated variables to a set
of linearly uncorrelated variables, or principal components. These make up
new dimensions for representing the data and are defined so that they maxi-
mize its variance. The new dimensions are orthogonal, resulting in mutually
uncorrelated variance explained by each dimension. The idea is that this will
emphasize relevant patterns in the data, and that considering only the few prin-
cipal components that represent the most variance will be enough to accurately
represent it.

One of the earliest uses of PCA in population genetics was performed in
[60], and it has since been used for various applications, including the detec-
tion of population structure [61] and assessing genetic similarities between
ancient and present-day individuals [49].

In applying PCA to SNP data, each individual is usually represented by
its genotype at a given set of loci, and the resulting principal components are
viewed as axes of genetic variation due to the ancestry of the samples. Samples
with similar genetic affinities are thus expected to be projected to positions
with close proximity in PCA space.

### 3.1.1 Methods to Handle Missing Data in PCA

When performing PCA, the presence of missing variables influences the defi-
nition of the model as well as the projected coordinates of samples. As miss-
ing information is common in empirical sequencing data, this can introduce
bias in the interpretation of PCA results. Such effects risk becoming partic-
ularly pronounced when dealing with low-coverage data, as is often the case
for aDNA.

This section concerns the particular scenario in which a PCA model is de-
defined on one set of samples, and a new sample that has missing values for
a subset of the variables used is to be projected onto that model. This is a
problem that occurs in the field of aDNA, where it is common to visualize the
genetic affinities of ancient individuals in the context of modern genetic varia-
tion by projecting them onto a PCA model defined using a panel of present-day
individuals [17].

A method to address this issue that has been used in the field of aDNA is to
define separate PCA models for each ancient sample in which only those sites
in the modern panel that are present in the ancient sample are used, and subse-
quently merging the individual projections using the Procrustes transformation
[62, 63].

When applying this technique to evaluate imputation performance by means
of PCA in Paper I, significant artefacts due to missing genotypes were ob-
erved (not published). Further investigation showed that this was largely due
to the influence of different sets of SNPs overlapping the modern reference,
with effects particularly noticeable due to the fact that different data points for
the same individual were compared.
Other techniques for handling missing data have been introduced in the PCA literature. In [64], several strategies are defined, and two regression-based methods are shown to be statistically superior on simulated industrial data sets. However, as PCA is highly data-dependent, such results are not directly transferable to the genomics scenario. Firstly, genotype data differs from many common applications due to the fact that there are generally many more variables than samples. Further, LD and the existence of variants that give certain SNPs a large impact in the transformation may result in covariance structures that are atypical, and in some cases particularly sensitive to missing data.

In order to evaluate methods in the particular context of genotype data, we performed a study on empirical SNP data, presented in Paper III. Five methods were selected and tested on four panels of samples based on public data sets of present-day individuals that are commonly used in practical applications. We considered both array and sequencing data, with and without the removal of linked variants, in order to obtain panels with different correlation structures.

We selected for evaluation the methods of trimmed scores (TRI), projection to the model plane (PMP), trimmed score regression (TSR), known data regression (KDR), as well as the strategy of performing individual projections and merging using Procrustes transformation (PRCR). TRI and PMP correspond to the default and lsqproject options for dealing with missing genotypes offered by the widely used software SMARTPCA from EIGENSOFT 7.2.1 [61, 55]. As previously mentioned, the PRCR method has been employed within the aDNA community. TSR and KDR are regression-based methods that were found to give best performance on the industrial data set considered in [64].

For all panels considered, TRI resulted in significantly higher errors than the other methods, whose performance tended to be on a more similar level. This is in line with previous results from the literature, and shows that the more involved approaches might be worth considering over this very simple method, which consists of filling in missing values with their unconditional mean (0 for centered data).

The evaluation yielded varying results, with no method consistently outperforming the others. However, the regression-based methods displayed high relative performance overall, with TSR in particular showing consistently low relative errors on different data sets. For the panel that was used to define a PCA model in Paper I, the KDR method was shown to result in lowest error for most levels of missing data, motivating its use in the paper. An impact of the filtering of correlated sites by LD pruning was also visible on the relative performance of the considered methods.
3.2 The Structure Model

The Structure model presented in [65] defines a discrete population admixture model that forms the basis for many widely used software implementations, including STRUCTURE, ADMIXTURE [66] and FRAPPE [67]. In this model, individuals are given a proportional assignment to $K$ clusters, characterized by a set of allele frequencies at each location considered.

For a data set consisting of $N$ individuals typed at $L$ loci, with two allele copies at each loci, a Structure model with $K$ populations contains:

- observed genotypes $X$, where $x_{ilc}$ denotes the variant observed for individual $i$ at locus $l$, copy $c$
- unknown origin populations of the genotypes $Z$, where $z_{ilc}$ denotes the population of origin for allele copy $c$ at locus $l$ for individual $i$
- unknown admixture proportions for each individual $Q$, where $q_{ik}$ denotes the proportion of the genome that originates from population $k$ for individual $i$
- unknown allele frequencies $F$, where $f_{klj}$ denotes the frequency of variant $j$ at at locus $l$ in population $k$

The probability of allele observations for an allele copy $c$ at locus $l$ is given by a categorical distribution with probabilities corresponding to the allele frequencies $(f_{kl1}, f_{kl2}, \ldots, f_{klJ_l})$ of the allele copy’s population $k$, where $J_l$ denotes the number of variants at locus $l$:

$$P(x_{ilc} = j | F, z_{ilc} = k) = f_{klj} \quad (3.1)$$

The origin populations of allele copies are similarly given by a categorical distribution with probabilities corresponding to the admixture proportions $(q_{i1}, q_{i2}, \ldots, q_{iK})$:

$$P(z_{ilc} = k | F, Q) = q_{ik} \quad (3.2)$$

The probabilities in Equations 3.1 and 3.2 are given independently for $i \in 1, \ldots, N$, $l \in 1, \ldots, L$, $c \in 1, 2$. The allele frequencies are modeled by a Dirichlet distribution, $f_{kl} \sim Dirichlet(\lambda_1, \lambda_2, \ldots, \lambda_{J_l})$ with a default of $\lambda_1 = \lambda_2 = \cdots = \lambda_{J_l} = 1.0$, yielding a uniform prior. For the admixture proportions, a Dirichlet prior $q_i \sim Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_K)$ is used. They choose to use a symmetric prior with $\alpha_1 = \alpha_2 = \cdots = \alpha_K$, and attempt to find an appropriate value by setting a uniform hyperprior $\alpha \sim Uniform(0, 10)$. This parameter controls the nature of the clustering, with $\alpha < 1$ modeling each individual originating in few populations, and $\alpha > 1$ that individuals tend to have admixture components from many populations in similar amounts.

In the software STRUCTURE, a Bayesian approach using MCMC is implemented to sample from the posterior in order to obtain estimates of the ad-
mixture proportions as well as the allele frequencies of the populations. The widely used ADMIXTURE software is based on the same probabilistic model, but has a faster implementation that instead focuses on maximizing the likelihood.

In the described model, the allele frequencies characterize the populations and the allele copies in the observed genotypes are modeled as independent draws from these frequencies. This corresponds to the assumption that Hardy-Weinberg and linkage equilibrium between loci hold within the populations, and that accounting for the presence of disequilibrium by means of the populations groupings is a meaningful characterization of population structure.

The authors point out that inference depends heavily on these assumptions, and that assessing their validity is important in the interpretation of the results. They also discuss the difficulties involved in inferring appropriate values for $K$, which is also brought up in e.g. [68]. That care must be taken in interpretation of the population clusters, and that they do not necessarily have a clear biological or demographic significance, is also addressed elsewhere, e.g. in [69].

### 3.3 Deep Learning

As larger amounts of genetic sequencing data are becoming available, the use of artificial neural networks (ANNs) and deep learning in particular is increasing in the fields of genomics and population genetics [70]. Such methods are based on nonlinear, hierarchical models that learn abstract features in a data-driven manner [71] and represent a different paradigm to many methods traditionally used in these fields that are based on statistical inference and the incorporation of assumptions or knowledge about the biological processes underlying the data [72].

A central population genetic research question that DL has been applied to is the inference of evolutionary and genomic parameters from observed patterns of genetic variation. In [73], a DL model was used for joint estimation of population size and selection parameters, a problem that has been considered particularly challenging for traditional models as the effects of demography and selection on the genome can be difficult to untangle [74].

Their approach relied on the use of summary statistics that represent some aspect of genetic variation. While having the benefit of reducing the number of variables to consider, summary statistics are not able capture all information present in the raw genomic data and must be carefully chosen to be informative for the parameters of interest. The use of summary statistics has traditionally been widely adopted in the field of population genetics, e.g. to avoid prohibitive costs of complex likelihood-based methods or to employ simulation-based methods such as Approximate Bayesian Computation (ABC) which can suffer from decreased performance as the number of inputs grows [75].
It has been argued that DL can avoid this ‘curse of dimensionality’ in certain cases [76], and DL methods that do not rely on summary statistics but operate directly on sequencing data have successfully been used for population genetic inference in e.g. [77] and [78]. Possible benefits for the analysis of genomic data of such ‘end-to-end’ DL models that incorporate the entire data-processing pipeline and avoid the use of handcrafted features are further discussed in e.g. [70].

DL has also been applied beyond predictive modeling for more general characterization of genomic data sets by means of generative models that estimate the distribution of the data and allow sampling from the learned distribution. In [79], generative models based on shallow and deep ANNs were used to create artificial genomes that retained key characteristics of empirical data, providing an alternative that, in contrast to many model-based simulation methods, does not require the prior estimation of demographic parameters for the population of interest. They also demonstrated how such models were possible to use for dimensionality reduction of genotype data. In [80], the utility of another generative DL model for nuanced visualization of genetic variation and capturing of population structure was demonstrated.

3.3.1 A Convolutional Autoencoder for Characterization of Genetic Variation

Paper IV introduces a DL framework denoted Genotype Convolutional Autoencoder (GCAE) designed to learn patterns of genetic variation from genotype data. This section provides an outline of the model architecture and training strategies employed, focusing on the design choices made to adapt it to the application of genotype data.

Figure 3.1 shows the architecture of the proposed framework. The structure is that of an autoencoder, where an encoder transforms the input data to a lower-dimensional representation from which a decoder subsequently aims to recreate the original input [81]. The two components are trained together by minimizing the discrepancy between the original input and its reconstruction, with the idea of finding a latent representation, or encoding, that captures key characteristics of the data. The model alternates convolutional and pooling layers in the encoder, contains a series of fully-connected layers in the middle, after which convolutional and upsampling layers are used in the decoder to reconstruct the input.

The use of convolution is a differentiating factor between our approach and those of [80] and [79] discussed in the previous chapter. Convolutional networks have been successfully applied to data ranging from time-series and text to images and video, and have become the dominant approach in the field of computer vision [82]. The use of convolutional layers for this application was motivated by the fact that they are sensitive to the order of the input variables
Figure 3.1. Architecture of the GCAE model from paper IV. The binary mask representing missing data is displayed in blue, and marker-specific variables in green and red. Black dashed lines indicate residual connections.

[71] and therefore take into account the sequential nature of genetic data, with variants located near each other often being correlated. Convolution was applied by means of a sliding dot-product with a weight matrix, or kernel, along the spatial dimension of the input, in a one-dimensional version of the operation typically applied to image data. As convolutional kernels have a finite spatial extent with shared weights operating over the entire input, they tend to be useful when its reasonable to expect that similar local patterns will appear in multiple locations in the data, e.g. the presence of edges in images [71]. While patterns due to LD may be captured by such constructs, a key difference between typical image and DNA sequence data is that genotype information is also characterized by a dependence of absolute position. In order to allow the model to learn global positional information that may be lost with convolution, two sets of what we denote marker-specific variables were also incorporated in the architecture. The two sets are displayed in red and green in Figure 3.1 and each have one variable per genotype in the input sequence, the values of which are updated by the model training procedure. Aside from the genotype sequence and a set of marker-specific variables, the model input also includes a dimension representing missing data, displayed in blue in Figure 3.1. An explicit representation of missing data was opted for in order to aid the model in learning features that are not sensitive to missing genotype values, as this is a common occurrence in empirical genomic data sets. The training procedure also made use of data augmentation by randomly setting inputs to missing or erroneous values in order to improve the robustness of the model to missing and incorrectly genotyped sites. The loss function was also adapted for application to sequence data. For each loci \( m \in 1, \ldots, M \) in the sequence data, model input was represented by a scalar value consisting of a scaling of the alternate allele count (0, 1 or 2...
for diploid genotypes), and the loss function was expressed in terms of assigning one of 3 discrete values. This was done by evaluating the cross-entropy between the target \((y)\) and predicted \((\hat{y})\) genotype probabilities for a sample, shown in the first term in Equation 3.3, where \(y^i_m\) denotes the probability of genotype \(i\) at location \(m\).

\[
L(y, \hat{y}) = -\frac{1}{M} \sum_{m} \sum_{i} y^i_m \log(\hat{y}^i_m) + \alpha \sum_{j} e^2_j \tag{3.3}
\]

Target genotype probabilities were represented using one-hot encoding, and predictions for a location \(m\) were obtained by applying the sigmoid function to the raw model output in order to obtain a value \(o_m\) between 0 and 1, and then calculating the expected genotype frequencies \(f(g)\) under Hardy-Weinberg equilibrium for \(g \in \{0, 1, 2\}\) (Equation 3.4).

\[
\begin{align*}
f(0) &= (1 - o_m)^2 \\
f(1) &= 2(1 - o_m)o_m \\
f(2) &= o_m^2
\end{align*}
\tag{3.4}
\]

The loss function also comprised an L2 penalty on the values of the latent representation \((e)\) for regularization of the model. This is shown in the second term in Equation 3.3, where \(d\) is the number of dimensions of the latent representation and \(\alpha\) is a parameter controlling the contribution of the penalty term to the loss.

In Paper IV, the utility of the proposed framework for different data characterization tasks was evaluated. For the application of dimensionality reduction, it was shown to result in a lower-dimensional representation that was more nuanced and visually informative than that of PCA, while retaining information about global patterns in genetic variation to a higher extent than t-SNE and UMAP. Comparison to the variational autoencoder approach considered in [80] suggested that the two DL methods resulted in qualitatively similar latent representations. Figure 3.2 shows the dimensionality reduction obtained using GCAE on a set of individuals from 166 populations representing worldwide genetic variation.

Assessment of the output genotypes further showed that spatial patterns as measured by the decay of LD along the chromosome were preserved to an extent, and application to the problem of genetic clustering yielded results that were comparable to those of the model-based approach implemented in the software ADMIXTURE.

### 3.3.2 Notes on Applying Deep Learning to Genotype Data
Two central difficulties associated with developing DL methods for genotype data are large data volumes and high dimensionality [83]. These pose com-
Figure 3.2. Dimensionality reduction results of GCAE applied to genotypes of individuals from 166 worldwide populations. Coloring is by population, grouped into eight superpopulations, according to the legend in Paper IV.

Computational challenges that require efficient algorithms and large-scale computational resources for model training. Such considerations also served as an initial motivation for considering convolutional layers as a means to reduce the number of trainable weights in our study. Data with high dimensionality, in which the number of variables is high and typically larger than the number of samples, poses several challenges to statistical data analysis in general, often leading to issues such as estimation instability and overfitting [84, 85].

Data augmentation by means of removing information and introducing noise turned out to be an approach which improved the ability of the network to generalize. In our implementation this was done in a uniformly random manner, but for some data sets it might be beneficial to incorporate information about known patterns, e.g. reference bias, the post-mortem damage that is known to affect aDNA, or batch effects from combining sequencing data from multiple experiments. Other regularization techniques used included dropout, addition of noise and constriction of the magnitude of the encoded values.

Further exploration of model regularization could achieve improved latent structure and ability to scale to higher number of latent dimensions, at which performance was found to drop off for both DL-based methods evaluated in Paper IV. However, we also note that the k-Nearest Neighbors classifier itself is subject to decrease in performance for higher dimensions [86], and a different metric might therefore give a more informative assessment of performance.
An aspect of genotype data that posed difficulties was the presence of rare variants, with low frequency alleles in particular becoming fixed in the reconstructed data. For the model that was used for analysis of output genotypes, a re-weighing of the loss by means of the effective number of samples, as defined in [87], was performed to improve the reconstruction of low-frequency genotypes. For a sample with target genotype $y$ at a given position in the genome, the contribution to the loss function was multiplied by a factor $\frac{1-\beta}{1-\beta^{n_y}}$, where $n_y$ denotes the number of samples with target genotype $y$ at the position, and $\beta$ is a hyperparameter.

This has the effect of increasing the relative contribution to the loss of rare genotypes, with $\beta$ controlling the degree of the weighting. Higher values of $\beta$ were found to reduce the issue of low-frequency variants becoming fixed in the output genotypes, but a trade-off for performance in the middle of the allele frequency spectrum was observed. A possible future extension could be to have different values of $\beta$ for different SNPs depending on the allele frequency.

The format of model output and the loss function used was also motivated by an observed tendency to underestimate the presence of rare genotypes. In our model, the problem of genotype reconstruction is represented as a categorization problem in which a discrete count is predicted. We consider the genotype as being drawn from a distribution which we aim to reconstruct. A biological interpretation is that each individual is then seen as a sample from a population, where the genotype distribution can be characterized by an expected allele frequency. On a more general level, this approach is related to others that have been used to enforce structure on the predicted probability distributions of neural networks. In e.g. [88], a similar binomial construction was used to impose a unimodal distribution in order to preserve ordering between labels for an ordinal classification problem. They also considered a loss function based on the Wasserstein distance that, unlike cross-entropy, considers inter-label relationships under a one-hot encoding of targets [89]. Loss functions with this characteristic would also be interesting to explore as a means to obtain more realistic posterior probabilities for this application.

Inspection of the output genotypes showed that the ability of GCAE to retain sequential information in the form of LD between sites was highest for pairs of loci that were short distances from each other, with overestimated correlation for sites that were further apart. As the hierarchical structure and sparsity of convolutional neural networks cause units in deeper layers to be influenced by larger segments of the input sequence than units in more shallow layers [71], experimentation with deeper networks may improve the ability to capture LD over larger genomic distances. The explicit inclusion of spatial information in the form of genetic distance or position, as in e.g. [90] or [77], is also an interesting approach to pursue in the future.
4. Scalable Data Analysis via Cloud Computing

The rise of high-throughput technologies has caused an explosion in the size of genomic data sets, and led to increasingly data-driven research in the life-sciences [91]. Facilitating many types of statistical analysis, the availability of large data sets has led to significant progress in areas such as investigating the association between diseases and human genetic variation as well as in the fields of personalized medicine and drug development [92, 93]. Population-genetic analyses of large cohorts of individuals can also give insights about our evolutionary past [9], as well as aid in the reconstruction of genetic information from ancient samples, as exemplified in Paper I.

With increasingly data-intensive applications, larger demands are also placed on computational infrastructure. As a consequence, the use of cloud computing has seen an increased use in genomics research. Under this model, computing infrastructure or software is made available in an on-demand fashion and accessed via networks, essentially allowing users to rent resources rather than to purchase, own, and maintain them. A wide variety of cloud-based resources are now offered by various actors, including commercial cloud providers such as Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure, and IBM Cloud as well as academic initiatives like Open Science Data Cloud and the Swedish National Infrastructure for Computing (SNIC) Science Cloud.

Potential advantages of cloud computing for scientific users are facilitated reproducibility and collaboration, as well as the global availability that allows access to large-scale resources regardless of the in-house facilities of one’s institution [94]. For organizations, the cloud model can provide advantages compared to up-front acquisition of local resources. The often significant initial investment can be reduced. Further, as cost-effectiveness depends largely on the level of resource utilization that can be maintained, the pay-per-use pricing model may have benefits in scenarios where future resource usage is difficult to predict [95]. The inherent flexibility of cloud computing can also facilitate handling of short and intense peaks in resource demand at which on-premise resources reach full utilization by enabling cloud bursting, a hybrid solution in which resources from public cloud providers can be used to temporarily boost local capacity.

Among the challenges posed by increasing sizes of genomic data sets is the storage needed to hold raw data. This, together with the associated difficulties of transfer and processing may hinder access and sharing of data between...
institutions [96]. The migration of storage and processing to the cloud is a possible means to mitigate this, and many consortia now make data sets available via public repositories hosted on cloud resources, e.g. the ENCODE [97] and the 1000 Genomes [21] projects.

This issue is addressed in Paper V, where we introduce the BAM Search Infrastructure (BAMSI); a scalable, distributed platform for data filtering and streaming using resources from multiple cloud providers. Largely intended for preprocessing, our tool can be used to perform an initial reduction in size of a data set, allowing a smaller subset to be downloaded or staged into other environments for further analysis. We exemplified the use of BAMSI by performing an analysis of structural variation on the entire 1000 Genomes phase 3 low-coverage set of aligned reads. The roughly 60TB data set was scanned for alignments indicative of possible inversion events, and a genome-wide overview of the results was presented.

We also demonstrated the scalability of such a multi-cloud setup by analyzing the aggregated filtering throughput for deployments making use of different mirrors of the data. One deployment was set up with filtering workers on SSC, accessing the data from local shared storage at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). Another set of workers was set up on Amazon EC2, accessing the data from the publicly available Amazon S3 bucket.

Figure 4.1 shows the throughput for combinations of these configurations. When only using SSC resources, saturation of the link to UPPMAX was reached at around 170 MB/s at 12 workers. As expected, saturation of the link to S3 was not reached, and using Amazon resources resulted in roughly linear scaling. The third setup, in which resources on EC2 were added to those on SSC from the point of saturation, again shows continued scaling as a function of number of workers. This illustrates how the use of multiple data sources along with distributed processing located near the data allows bandwidth bottlenecks to be minimized. In scenarios such as this, where a local community cloud is available, this allows for increasing filtering throughput while maintaining cost-effectiveness.

While cloud-based technologies abstract away much of the underlying compute resources, the traditional Infrastructure as a Service (IaaS) model nonetheless requires the allocation and maintenance of clusters of virtual machines. As pricing depends on provisioned capacity, maximizing cost-effectiveness requires achieving and maintaining high cluster utilization rates. The setup and maintenance of clusters that can scale elastically according to current usage is not trivial and requires technical knowledge, resulting in management overhead or even the rendering of cloud resources inaccessible for some users [98].

An alternative paradigm to IaaS that has been gaining traction is that of serverless computing. In this model, the provisioning of resources is hidden from the user, and the running of code or containers, rather than vir-
Figure 4.1. Filtering throughput of BAMSI as a function of numbers of workers for different setups of data and compute. The solid line indicates runs in which all workers were deployed on SSC. The dotted line indicates runs in which the number of SSC workers was fixed to 12, with additional workers deployed on Amazon EC2. The dashed line indicates runs in which workers were deployed on Amazon EC2 only. The maximum throughput over three runs is plotted for each setup of workers.

Virtual infrastructure, is provided on demand. The serverless paradigm offers consumption-based pricing and inherent elasticity of resources and promises a simpler means to obtain ‘pay for what you use’ rather than ‘pay for what you provision’.

Most commercial cloud vendors now offer serverless options, but as the increased simplicity comes at a higher price, the comparative cost-effectiveness would depend on cluster utilization levels that could be maintained for dedicated cloud resources. Other potential issues of serverless technologies include vendor lock-in resulting from the adaption of applications to the interface of a particular service, as well as inefficiencies resulting from handling of storage and state persistence, discussed further below.

In Paper VI, we present a workflow management system named Serverless Workflow Enablement and Execution Platform (SWEEP). Built on the serverless paradigm, SWEEP allows for cloud-based computation without the need of provisioning and maintaining clusters. SWEEP workflows are defined in terms of functions or containers as executable units that can be run on different serverless engines, and are thus not inherently tied to a single cloud provider. The use of SWEEP was demonstrated and evaluated on two workflows from different scientific domains; a canonical sequence alignment and variant calling pipeline, based on the Genome Analysis Toolkit (GATK) [99] best practices documentation and a workflow from the field of geosciences in which fine-resolution satellite imagery was used to analyze surface reflectance in lakes.
The use of distributed cloud environments for scientific applications has several implications for workflow design that can affect performance in terms of scalability and correctness. The execution framework needs to be able to detect and handle failures in individual tasks in order to achieve resiliency. By implementing smaller executable units, retries due to random failures can have less effect on overall runtime, but can also incur higher communication overhead. This is also influenced by the amount of inter-task communication and level of I/O performed, with network performance affecting both. Large, long-running tasks can on the other hand reduce the potential concurrency of the system and lead to unbalanced computational loads. This can potentially render it more difficult to maintain high cluster utilization rates, particularly in cases with unpredictable runtimes.

Workflows that have the potential to scale out or are subject to intensive bursts of resource utilization might benefit from the extremely broad parallelism of the serverless model. Under such circumstances, the use of serverless resources can provide a cost-effective alternative with lower response times for the researcher.

The serverless execution paradigm can be a means to achieve elasticity of resources, but it is associated with other challenges, including overhead due to the setup of the computational runtime for each executable unit, and the maintenance of data persistence. As serverless execution is performed in a stateless manner in which the invoker is unaware of the underlying execution environment, remote data storage must be used for persisting information between individual tasks. Efficiency and scalability thus requires both low-latency and high-throughput remote storage, as well as high network bandwidth to avoid large data transfer overheads.

It has been argued in e.g. [98] that as network bandwidths are approaching those of storage, disk locality is becoming less important in determining performance. For many practical applications, the nature of the workflow and the tasks that it comprises, including the level of I/O, will determine the optimal approach.
The work presented in this thesis concerns computational methods in the field of genomics. The studies assessing statistical methods have focused on evaluation in the context of sparse, uncertain, or otherwise challenging data and aim to contribute to the further development of models and tailoring of methodology for such scenarios.

The empirical study of imputation confirmed that application to ancient data can give satisfactory results, but also pointed out accuracy for low coverage data and rare variants as potential areas of improvement. Our later results suggested that the proposed implementation of the full Li and Stephens probabilistic framework is a promising alternative method in such cases, and future work in this area will involve exploring the incorporation of ancient-specific information into the model.

This could be done by tuning error and recombination probabilities, either by fitting them to the data as in e.g. [33] or by letting them be informed by external information. A model of the specific damage profile of aDNA could be used to adjust the error parameter, and the recombination parameter could be weighted according to temporal information or some measure of expected level of divergence between an individual and the template haplotypes used.

The latter could also be done in combination with construction of the haplotype model based on other study samples in addition to the phased reference panel. A simple iterative scheme in which current phase estimates for remaining study samples can be included in the set of template haplotypes used for a sample’s HMM is currently implemented, but might be improved by such a weighting, or by using a sampling scheme like the one used in [33].

By developing an implementation that is optimized to run in a highly parallel manner we hope to harness the increased availability of GPUs that has resulted from recent trends in the field of machine learning for making more comprehensive, and therefore costly, imputation methods more practical to employ in cases where maximizing accuracy is the main priority.

The study of methods to handle missing data in PCA confirmed the data-sensitive nature of the problem, and suggested that method evaluation may be necessary to establish an appropriate approach on a case-to-case basis. Future work in this area will include more rigorous testing on different data sets, including an evaluation of the effects of preprocessing and filtering the data. Performing a characterization of the covariance structure of genomic data sets and how it can influence effects of missing data might also aid in developing more general guidelines and best practices.
With the presented framework for data characterization based on convolutional autoencoders, we aim to contribute to the development of DL methods for applications in genomics. As more implementations become available, both challenges associated with this type of data discussed in Section 3.3.2 can be addressed. The establishment of appropriate architecture designs, training strategies and areas of hyperparameter space, together with the possibility to generate artificial genomes with realistic characteristics, could aid in handling the high-dimensional nature of the data. As prediction is relatively fast once a model has been trained, the public availability of models pre-trained to characterize genetic variation could further aid in addressing the computational challenges associated with sizes of genomic data sets. This would also allow for re-use of the coefficients of trained models for similar applications, or to speed up the development of new models by using them as a starting point for additional optimization.

The correctness of dimensionality reduction and genetic clustering results is not necessarily precisely defined, and performance comparison between methods is therefore not straightforward. However, the intrinsic differences between a highly data-driven approach like DL to alternative methods, e.g. those based on neighborhood graphs or the Structure model of discrete population admixture, can lead it to reveal different aspects of the data. We argue that such perspectives can be useful, particularly when other methods are difficult to interpret or will not function as intended due to modeling assumptions not being met.

A commonly brought up limitation of DL is the fact that the information learned by a model tends to be represented in a way that is complicated to decipher, making it difficult to assess how features of the input impact the results [72]. Advances have been made in achieving interpretability of DL methods in genomics [100], and it will be interesting to investigate adaptions of these for increasing the ability to draw conclusions about underlying biological processes for the application considered here.

Future work will also involve extension of the proposed framework for other applications. We plan to investigate modifications of network architectures to incorporate phenotypic information and explore questions related to quantitative genetics. Another area of interest is that of imputation, where DL is relevant to explore as an alternative to traditional methods based on population genetic models both as a different, more-data driven approach as well as a means to tackle the challenges associated with scaling to increasingly large reference sets.

Further, we have argued that as demands on storage and computational resources are rising, cloud infrastructures can facilitate scalable data analysis in genomics. Two frameworks that exemplify the use of such resources for scientific workflows were introduced with the aim to contribute to the utilization of a variety of computational infrastructures for applications in genomics and other data-intensive sciences.
6. Summary of Papers

Paper I
This paper describes a study in which we evaluated an imputation pipeline that has previously been used in the aDNA community. Our results suggested practical guidelines for designing imputation studies, including the fact that a large reference panel of phased haplotypes can be beneficial, and that the information in low- and moderate-coverage ancient data can be leveraged to improve imputation accuracy further, particularly for rare variants. An analysis of error patterns for different genetic variants and allele frequencies was also performed, which may be of use in post-processing and interpretation of results from downstream analysis.

Contributions
The author of this thesis performed the data processing relating to downsampling and genotype likelihood generation of sample data and executed the imputation experiments. The study was designed by the second author, in collaboration with the other authors. The author of this thesis produced the paper with feedback from the remaining authors.

Paper II
In this paper, we implement a method for imputation that is based on the full probabilistic Li and Stephens model for haplotype frequencies, and compare its performance to alternative pipelines based on the Beagle software suite. We consider empirical ancient data downsampld to lower coverages as well as artificially thinned out genotypes of modern individuals, and show that the proposed method yields higher genotype concordance at low coverages and improved ability to capture rare variation. Our implementation is optimized for running in a massively parallel setting, e.g. on GPUs, on which we show that it achieved reasonable runtimes on the experiments performed.

Contributions
The author of this thesis drafted the paper and the main structure of the software, and contributed to ideas and study design. The second author implemented software optimizations, in addition to writing the related sections of the paper and overall proofreading.
Paper III
Principal component analysis (PCA) is a widely used data characterization method in the field of genomics. The results of the method are affected by missing values in the data, which is a common feature of genetic data, and ancient DNA in particular. In this paper, we consider the problem of projecting samples with incomplete information onto an existing PCA model in the context of analysis of genetic data, and evaluate different methods from the literature. We found that the widely used method of imputing missing values with their unconditional mean resulted in highest errors overall, while regression-based methods gave consistently high relative performance on the different genomic data sets used for evaluation.

Contributions
The author of this thesis designed the study, implemented the methods, ran experiments, and wrote the paper.

Paper IV
The use of deep learning methods is relatively new in the field of genomics and population genetics, but it has shown to be a promising alternative to traditional methods for various applications. In this paper, we present a deep learning model based on a convolutional autoencoder architecture and demonstrate its utility for performing dimensionality reduction and genetic clustering of genotype data. A comparison of dimensionality reduction results highlighted qualitative differences to alternative methods including PCA, t-SNE and UMAP, and an analysis of output from the model showed that it was able to learn a representation that preserved sequential properties of the data in the form of decay of linkage disequilibrium with distance along the chromosome. Our results demonstrated that the use of convolution is feasible for genotype data, and we suggest customized network architectures and training strategies, providing a foundation for further development of this type of neural network models for genotype data.

Contributions
The author of this thesis took part in discussing study design and developing ideas for the GCAE model, in addition to having the main responsibility for developing the code. Data preparation, running of experiments and writing of the paper were also primarily done by the author of this thesis.
Paper V

We address some of the challenges posed by large data sets in genomics by presenting a tool for scalable, distributed filtering of genomic data using cloud technologies. The BAM Search Infrastructure (BAMSI) is a Software as-a-Service (SaaS) solution that can leverage compute resources from multiple cloud providers to perform initial filtering of large, public data sets, and thus make re-analysis available without the need to have access to a large-scale computing facility. We exemplify the use of BAMSI on the 1000 Genomes data set by performing a simple analysis of structural variation on the entire set of low-coverage reads that totals 60TB in size, and demonstrate how the use of multiple mirrors of the data can improve the horizontal scalability of the system.

Contributions

The author of this thesis developed the software and performed the experiments in close collaboration with the second author. Ideas and manuscript were developed in cooperation with all authors.

Paper VI

In this paper, we evaluate the utility of the serverless execution model for running scientific data analysis pipelines in the cloud. We present a workflow management system named Serverless Workflow Enablement and Execution Platform (SWEEP) that allows workflows comprising functions and containers to be executed on one or several cloud providers, without the need to provision and maintain virtual infrastructure. We exemplify the use of the proposed system on two applications; a variant calling pipeline from the field of genomics and a geoscientific workflow based on satellite remote sensing, and evaluate performance and scaling.

Contributions

The author of this thesis participated in study design and discussions regarding the design of the workflows, and implemented the software framework in close collaboration with the first author. The writing of the paper was done jointly between all authors.
De senaste årtiondenas teknikutveckling har gjort det möjligt att läsa av av kompletta genetiska sekvenser hos individer i större detalj och mer kostnadseffektivt än tidigare. Detta har lett till en stor ökning av mängden genetiska data som finns tillgänglig och möjliggjort stora framsteg i genetikforskningen, samtidigt som allt större krav ställs på metoder och infrastruktur på att utföra beräkningar på effektiva sätt.

Även inom området arkeogenetik, där DNA från arkeologiska fynd studeras, har stora tekniska framsteg gjorts som lett till att genetisk information har kunnat utvinnas från kvarlevor av förhistoriska och historiska människor samt andra hominider. Sådana data har blivit ett viktigt komplement till arkeologiska och paleontologiska metoder i studerandet av människans evolutionära och demografiska historia. Undersökandet av nya frågeställningar har också möjliggjorts, såsom huruvida kulturella skiften huvudsakligen varit resultatet av fysisk migration av människor eller spredandet av idéer.

Det finns dock flera utmaningar med att använda äldre material i genetiska studier. Proverna tenderar att innehålla mindre mängder DNA som är i sämre skick än vad som är typiskt för modernt material, vilket kräver specialiserade metoder både laboratorietekniskt samt för efterföljande dataanalys.

Fokus för detta avhandlingsarbete har varit utvecklandet av beräkningsmetoder för att analysera genetiska data, framför allt statistiska modeller och hur de kan anpassas för scenarier med särskilda utmaningar, till exempel de som uppstår i undersökningar av uråldrigt DNA.

Ett centralt ämne är metoder för imputation, eller ifyllandet av saknad information. Imputation är vanligt att utföra i studier av moderna individers DNA, och det finns många standardmetoder som har visat sig fungera bra och ge korrekta resultat. De vanligaste metoderna är baserade på information om vilka mönster av haplotyper, eller sammanhängande fragment i de kromosomer vi ärver från våra föräldrar, som är vanliga inom dagens befolkningar.

Imputation är lovande för studier av uråldrigt DNA eftersom det kan möjliggöra utvinnandet av mer information från det ofta begränsade biologiska materialet som är tillgängligt. Ett antal arkeogenomiska studier har framgångsrikt använt sig av imputation, men det har inte undersömts till fullo om metoder som är designade för moderna data verkligen är tilllämpliga. Det finns flera omständigheter som kan försvara imputation av uråldrigt DNA. Dels tenderar arkeogenomiska data vara både osäkra och glesare, och vidare är den genetiska variationen som finns inom dagens befolkning troligtvis mindre representativ för individer som levde en lång tid tillbaka i tiden.
I denna avhandling ingår en empirisk undersökning av en imputationsmetodologi som tidigare har tillämpats inom arkeogenomiska studier. Våra resultat bekräftade att imputerade data kan utgöra ett realistiskt alternativ för populationsgenetiska analyser och indikerade hur tillvägagångssätt kan anpassas för denna typ av data. Vi utvecklade även en mjukvara för imputation som är baserad på en mer fullständig probabilistisk modell för haplotypfrekvenser än vad många andra metoder tillämpar, och visade att det kan ge förbättrade resultat på särskilt utmanande data.

En annan central frågeställning inom genomik och populationsgenetik är karaktärisering av data på ett sätt som möjliggör visualisering och utforskande dataanalys. Sådana metoder är ett viktigt verktyg för att begripliggöra stora och komplexa datamängder, och kan till exempel användas för att kartlägga genetisk variation och upptäcka mönster och struktur inom populationer.

En vida använd metod är principalkomponentanalys (PCA), som utgör en linjär projektion för att representera data på ett sätt som maximerar dess variationer. Vi diskuterar utmaningar förknippade med att använda PCA på data med ofullständiga genotyper och presenterar en utvärdering av tillvägagångssätt för att hantera detta på olika empiriska datamängder. Våra resultat indikerade att att kvaliteten på de resultaten man fick från metoderna i hög grad var beroende av egenskaper i de data som användes, men att regressionsbaserade metoder konsekvent gav relativt läga fel i skattningsarna.

Vi har vidare undersökt användningen av datadrivna metoder baserade på artificiella neuronnät som ett alternativ till standarmodeller för olika applikationer inom genetisk kartläggning. I avhandlingen presenteras ett sådant ramverk baserat på djupinläsning, som vi visar gav en mer nyanserad dimensionsreducering än PCA samtidigt som den bibehöll global information till en högre grad än andra icke-linjära metoder som t-SNE och UMAP. Vi diskuterar även andra tillämpningar av ramverket, och visar att det ger upphov till en representation där sekventiell information bevaras, i meningen att det finns en stark korrelation mellan närliggande positioner i genomet.

Inom genomik, liksom inom andra forskningsområden, ställer ökande datamängder högre krav på effektiv datahantering och beräkningsinfrastruktur. Den sista delen av avhandlingen behandlar användandet av molnresurser för att underlätta dataanalys i vetenskapliga tillämpningar. I den första studien exemplifieras hur sådana distribuerade resurser kan användas för effektiv filtrering och analys av genetiska data och därmed tillgängliggöra stora informationsmängder för användare som inte har tillgång till högprestandaresurser för lagring och beräkning. Slutligen diskuteras användandet av datorresurser i en serverfri molnmodell, i vilken användaren inte är anvarig från att sätta upp och underhålla virtuell infrastruktur. Vi presenterar ett system som möjliggör användandet av sådana beräkningsresurser från flera olika kommersiella moln och visar att data- och beräkningsintensiva arbetsflöden såsom variantbestämmning av genetiska data kan utföras inom ett sådant ramverk.

46
Acknowledgments

I would like to thank my main supervisor Carl Nettelblad for his guidance, inspiring ideas and discussions. I’ve worked on quite a broad range of topics during these years, but I have yet to come across one that you don’t have extensive knowledge and insightful new ideas about. I would also like to thank my second supervisor Mattias Jakobsson for introducing me to the exciting field of population genetics and ancient DNA.

A big thanks also to all my friends and colleagues at TDB throughout the years for providing such a great working environment. Thanks to Slobodan, Simon, Fredrik and Robin in particular for being fantastic office mates. And to Sonja and Ylva for being good friends that I’m glad I got to share the PhD experience with.

I am also grateful to the Centre for Interdisciplinary Mathematics, Uppsala University for partially funding my position, as well as to Ångpanneföreningen’s forskningsstiftelse, Anna Maria Lundins stipendiefond and Helge Ax:son Johnsons stiftelse for their generous travel grants.

Finally, the biggest thank you goes to Ben for always being incredibly supportive and for keeping me alive, sane and happy. And to our daughter Esther for reminding me about what is important in life.
References


[33] Y. Li, C. J. Willer, J. Ding, et al. MaCH: using sequence and genotype data to


[83] L. Koumakis. Deep learning models in genomics; are we there yet?


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2129

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)