Demographic and population separation history inference based on whole genome sequences

by

Shiya Song

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2016

Doctoral Committee:

Assistant Professor Jeffrey Kidd, Chair
Associate Professor Jun Li
Assistant Professor Hyun Min Kang
Professor Kerby A. Shedden
Professor Patricia Wittkopp
Associate Professor Sebastian K. Zöllner

This dissertation is dedicated to my family,
whose love and support have carried me through.

ACKNOWLEDGEMENTS

I would like to express my greatest gratitude to Dr. Jeffrey Kidd for his guidance in my research and study during my Ph.D. Dr. Jeffrey Kidd is a solid mentor and role model to me. He encouraged me to pursue the topics that I'm interested in and offered me tremendous help when I needed it. He has broad interests and maintains several collaborations, which broaden my horizons. Through collaborative projects, I have learned how to work cooperatively and take an active part in collaboration. I'm also grateful to have an interdisciplinary committee from different departments. I would like to thank Dr. Sebastian Zöllner and Dr. Jun Li for discussion on the PSMC models and ABC analysis, Dr. Hyun Min Kang for expert knowledge on NGS analysis, Dr. Patricia Wittkopp for expert knowledge on evolutionary biology, and Dr. Kerby Shedden for guidance on the statistics part of my thesis.

Apart from my committee members, I would like to thank all my collaborators. Without them contributing data and analysis, I couldn't complete the story. In particular, I'd like to thank Kimberly McManus, Joanna Kelley for $\partial a \partial i$ and population structure analysis and Krishna Veeramah for guidance on G-PhoCS analysis. I'd like to thank Laura Botigué and Krishna Veeramah for their contribution and guidance on the ancient dog project.

I would like to thank all the past and current members of Kidd lab for giving me great companion and support all the time. Some of them provide me with spiritual support and encourage me whenever I face trouble. I'd like to thank Julia Wildschutte, Amanda Pendleton and Feichen Shen for being my listeners and sharing experiences. Some of them provide enourmous help for my projects. In particular, I'd like to thank Ela Sliwerskas and Sarah Emery for making fosmid pools libraries, Matthew Oetjens for mitochondria analysis on dogs, Angela Taravella for the population structure analysis on dogs, and Amanda Pendleton for the selection analysis on dogs.

I would also like to thank all my friends here at University of Michigan, especially my peers. I'm very grateful to have Shweta Ramdas, Sai Chen, Xuefang Zhao, Raymond Cavalcante, Ridvan Eksi, Byran Moyers, Craig Biwer and Lauren Jepsen as my classmates. We took classes and studied together, and inspired each other. This greatly helped to relieve pressure from Ph.D study. I'd like to thank all of my friends who make my life colorful and meaningful. Because of their presence and numerous memories we shared, I feel like having Ann Arbor as my second hometown.

Last but not least, I'm very grateful to my parents who support my pursuit of Ph.D abroad both materially and spiritually. Without them, I couldn't follow my dreams of being a researcher. And also to my beloved husband, Haipeng, who has always been taking care of me and been there for me when needed. Thank you so much for giving me love and faith to carry on.

TABLE OF CONTENTS

DEDICATION . . . . **ii**

ACKNOWLEDGEMENTS . . . . **iii**

LIST OF FIGURES . . . . **viii**

LIST OF TABLES . . . . **x**

ABSTRACT . . . . **xi**

CHAPTER

I. Introduction 1

1.1 Introduction 1
1.2 Background 3
1.2.1 High-throughput resequencing technologies 3
1.2.2 Pairwise sequentially Markovian coalescent (PSMC) model and other derivatives 6
1.2.3 Bayesian Coalescence-based demographic inference 9
1.2.4 Approximate Bayesian Computation in population genetics . . . 11
1.2.5 Ancient DNA studies 13
1.2.6 Overview of human, gorilla and canine evolutionary history . . 14
1.3 Dissertation overview 17

II. Resolving haplotypes using fosmid pool sequencing 30

2.1 Introduction 30
2.2 Methods 32
2.2.1 Fosmid pool sequencing 32
2.2.2 Haplotype reconstruction 32
2.3 Results 35
2.3.1 Haplotype reconstruction 35
2.3.2 Comparison with statistical phasing 37
2.4 Discussion 38

III. Human separation history inference using physically phased genomes 52

3.1 Introduction 52
3.2 Methods 54
3.2.1 MSMC analysis 54
3.2.2 PSMC on pseudo-diploid genome 54
3.2.3 ABC analysis 55
3.3 Results 57
3.3.1 Demography inference using PSMC' 57
3.3.2 Split time inferred using MSMC and the effect of phasing error 58
3.3.3 An ABC method to infer population split time using PSMC on pseudo-diploid genomes 59
3.4 Discussion 61

IV. Demographic history inference on gorilla species 85

4.1 Introduction 85
4.2 Methods 87
4.2.1 Whole genome sequences of gorilla 87
4.2.2 Genotype calling 87
4.2.3 Population structure 88
4.2.4 Demographic analysis using G-PhoCS 88
4.2.5 Demographic Inference of Western Lowland Gorilla 90
4.3 Results 91
4.3.1 Gorilla Population Structure 91
4.3.2 Demographic inference on gorilla 92
4.3.3 Western Gorilla Demographic Inference 94
4.4 Discussion 94

V. Evolutionary and demographic history of canine species 116

5.1 Introduction 116
5.2 Methods 118
5.2.1 Sequencing of two ancient dog genomes 118
5.2.2 Genotype calling for ancient samples and contemporary samples 121
5.2.3 Mitochondrial Analysis 125
5.2.4 Population Structure 125
5.2.5 f-statistics analysis 126
5.2.6 G-PhoCS and HXH divergence estimation 129
5.2.7 Analysis of loci associated with the domestication process . . . 133
5.3 Results 134
5.3.1 Sequencing of two Neolithic canid genomes from Germany . . 134
5.3.2 Mitochondrial similarity amongst German Paleolithic and Neolithic dogs 135
5.3.3 Genetic clustering of the European Neolithic dogs 136

5.3.4 Signatures of admixture within canids 137
5.3.5 Canid demographic history 139
5.3.6 Functional variants associated with the domestication process . 144
5.4 Discussion 145

VI. Conclusions and Future Work 175

6.1 Conclusions 175
6.2 Future directions 178

LIST OF FIGURES

Figure

2.1 Haplotype assembly results 40
2.2 Illustration of ReFHap's phasing result and a switch error 41
2.3 Illustration of the metrics used to quantify phasing errors 41
3.1 PSMC' inferred population history 65
3.2 MSMC inferred split times 66
3.3 PSMC on pseudo-diploid genomes 67
3.4 Split times and migration rate inferred using PSMC and ABC 68
3.5 Relative cross coalescence rate inferred using MSMC 69
3.6 Relative cross coalescence rate inferred using MSMC on simulated data 70
3.7 PSMC on pseudo-diploid genomes 71
3.8 Simulation results on inferring split time and migration using the combined approach of PSMC and ABC 72
3.9 Simulation results on different approaches to deal with unphased SNPs 73
3.10 Posterior distribution of split time and migration rate inferred using ABC 74
3.11 TMRCA distribution inferred using PSMC 75
3.12 Chi-square distance between TMRCA distributions using different haplotypes . . 76
3.13 Recent European ancestry inferred by RFMix in Massai individual NA21302 . . 77
3.14 Relative cross coalescence rate inferred using MSMC with and without masking out European ancestry from Massai individual 77
3.15 Comparison of relative cross coalescence rate inferred using MSMC on fosmid haplotypes constructed by trio phasing data or by Prism 78
4.1 Genome wide heterozygosity values 98
4.2 PCA of three gorilla species 99
4.3 PCA of western lowland gorilla 99
4.4 Admixture results for the three gorilla subspecies 100
4.5 Admixture results for 14 western lowland gorillas 101
4.6 Four-population phylogeny as inferred by TreeMix 102
4.7 Phylogeny and explored migration bands for G-PhoCS analysis 102
4.8 Total migration rate inferred by G-PhoCS, given by $m_{AB} \times \tau_{AB}$ 103
4.9 MCMC traces for three demographic parameters inferred using G-PhoCS 104
4.10 Demographic parameters inferred by G-PhoCS 105
4.11 Inferred best-fit demographic model of western lowland gorillas 106
4.12 Comparison of $\partial a \partial i$ and PSMC Models 106
5.1 Archaeological remains 148
5.2 MapDamage analysis showing the deamination changes typical of ancient DNA . 149

5.3 Performance of ancient genotype caller 150
5.4 Phylogeny of ancient and contemporary canids 151
5.5 PCA and population structure between ancient and contemporary canids 152
5.6 Genetic affinity of ancient samples 153
5.7 ADMIXTURE analysis based on SNP array data 154
5.8 CTC-Wolf admixture as inferred by SpaceMix 155
5.9 Inferred admixture proportion from f_4-ratio tests 156
5.10 Demographic model regarding ancient and contemporary dogs and wolves 157
5.11 NJ tree based on 16,434 “neutral” loci of selected samples used in G-PhoCS . . . 158
5.12 Divergence time estimates in G-PhoCS analysis when including and excluding ancient samples 159
5.13 Population size estimates in G-PhoCS analysis when including and excluding ancient samples 160
5.14 Total migration rate estimates in G-PhoCS analysis when including and excluding ancient samples 161
5.15 Comparison of divergence time estimates with/without NGD in G-PhoCS analysis 162
5.16 Comparison of divergence time estimates with/without migration setting in G-PhoCS analysis 162
5.17 Illustration of the tree structure used for estimating HXH/Europe divergence time 163
5.18 Results for numerical analysis 163
5.19 Haplotype and copy-number variation at the amylase 2B (AMY2B) locus 164
6.1 Structural Variants discovered using fosmid pool sequencing data 181

LIST OF TABLES

Table

2.1 Summary of population geographic information and presence in HapMap or 1000 Genomes Project . . . 42
2.2 Phasing statistics from fosmid pool sequencing . . . 43
2.3 Comparison between statistical phasing . . . 44
2.4 Summary of clone statistics of fosmid pool sequencing . . . 45
2.5 Summary of variant calling for whole genome sequencing . . . 45
2.6 Comparison of fosmid-resolved haplotype for NA19240 with Sanger sequenced fosmid clones . . . 46
3.1 Percentage of unphased SNPs . . . 79
3.2 Split time estimation from previous studies . . . 80
3.3 Posterior estimates of split time and migration rate using IM model . . . 81
4.1 Gorilla population history estimates using G-PhoCS . . . 107
4.2 Western lowland gorillas demographic history estimates from $\partial a \partial i$ . . . 108
4.3 Comparison of published estimates of gorilla population split times . . . 109
4.4 Comparison of D-statistics . . . 110
5.1 f_4-ratio test results . . . 165
5.2 D-statistics supporting the migration band setting used in G-PhoCS . . . 166
5.3 Comparison of divergence time estimates from previous studies using G-PhoCS . 166
5.4 The range of parameters sampled for estimating HXH/European divergence time 167
5.5 Jackknife estimates of P(110)/P(101) . . . 167

ABSTRACT

Patterns of DNA sequence variation among present day individuals contain rich information about past population history. The recent availability of whole genome sequences provides both challenges and opportunities for developing computational methods to infer detailed models of population history. The goal of this thesis is to extend current methodology and apply available techniques to answer questions about population history in human, gorilla and canine species.

Recent methodologies based on the sequentially Markovian coalescent model permit the inference of population history using single or several whole genome sequences. However, these approaches fail to generate parametric estimates for split times and split times estimation is confounded by subsequent migration. Additionally, the effect of switch errors resulted from statistical phasing on split time estimation is largely unknown. We reconstructed phased haplotypes of nine individuals from diverse populations using fosmid pool sequencing and assessed the accuracy of statistically phased haplotypes in the 1000 Genomes Phase III project. We analyzed population size and separation history using the Pairwise Sequentially Markovian Coalescent model (PSMC) and Multiple Sequentially Markovian Coalescent model (MSMC) and found that applying MSMC on statistically phased haplotypes results in a more recent split time estimation compared with physically phased haplotypes due to switch errors, especially for separation involving the San and Mbuti populations. We further extended PSMC with Approximate Bayesian Computation to infer split time and migration rates under a standard isolation with migration model. We dated several key events in human separation history using these methods.

Gorillas are nonhuman primate, listed as threatened on the IUCN Red List of Threatened

Species. We analyzed whole genome sequencing data of thirteen gorilla individuals and applied GPhoCS, a Bayesian, coalescent-based approach to infer ancestral population sizes, divergence times and migration rates amongst the three gorilla subspecies, shedding light on the evolutionary forces that have uniquely influenced patterns of gorilla genetic variation.

The origins and dynamics of dog domestication has been a controversial and intriguing problem. Ancient DNA can improve our understanding of older events in human/dog prehistory. We analyzed two ancient dog genomes from the Neolithic and over 100 contemporary canine genomes. While both dogs show signatures of admixture, they predominantly share ancestry with modern European dogs, contradicting a late Neolithic population replacement suggested previously by mitochondrial studies. We developed a numerical method to date the divergence time between ancient dogs and European dogs. By calibrating the mutation rate using our oldest dog, we narrowed the timing of dog domestication to a window of 20-40 kyrs ago.

CHAPTER I

Introduction

1.1 Introduction

Patterns of DNA sequence variation that exist among present day individuals contain rich information about past population history, such as the change of effective population size through time, population structure, isolation and migration (Schraiber and Akey, 2015). Learning population history is crucial to undesrtand the origin and evolution of species, historical events happened around populations, and how present-day genomic variations are shaped. However, inferring population history from DNA sequence variation is challenging. On one hand, correct handling of data and the use of appropriate data is crucial to answer specific questions. On the other hand, such inference largely relies on comprehensive and statistical methods to account for the interplay of forces including mutation, recombination, admixture and selection that shape present day genetic variation.

The first wave of studies extensively used uniparentally inherited systems, such as mitochondrial DNA (mtDNA) (Walker et al., 1987) and the non-recombining portion of the Y chromosome (NRY) (Underhill et al., 2000), providing the matrilineal and patrilineal perspective of the past. Although mtDNA and NRY have now been characterized to high resolution and sampled from tens of thousands of individuals (Behar et al., 2012; Karmin et al., 2015), single-locus based analysis provides only limited view because of the inherent high stochasticity of the evolutionary process.

The precision of evolutionary-inference methods increases rapidly with the number of genealogical trees studied, and very slowly with the number of sampled individuals (Rosenberg and Nordborg, 2002). Subsequent studies focused on short tandem repeat (STR) loci (Rosenberg et al., 2002; Tishkoff et al., 2009) and single nucleotide polymorphisms (SNPs) (reviewed in Novembre and Ramachandran (2011)) that are distributed across the genome. Although hundreds of thousands of STR loci or SNPs provide higher resolution for population structure, uncertainty of the STR mutation rates and ascertainment bias in the way SNPs were discovered limit their power to infer population history (Veeramah and Hammer, 2014).

Since the advent of high-throughput next generation resequencing technologies, whole genome sequences are becoming more available. Contiguous DNA sequence data reduces ascertainment biases, increases power to infer demographic processes, and also raises statistical and computational challenges (Schraiber and Akey, 2015). There are several existing challenges regarding whole genome sequence based demographic inference. Firstly, quality control is of great importance since next generation sequencing is more error prone compared with Sanger sequencing and the use of short reads (100-150 bp) restricts the use of methods that require haplotype-phased data. Secondly, full-likelihood methods that aim to compute or approximate the probability of observing the genetic variation given the assumed coalescent model and demographic parameters remain computationally demanding, especially for large data sets (Beerli and Felsenstein, 2001). However, methods that compress the data into summary statistics such as the allele frequency spectrum (Gutenkunst et al., 2009; Excoffier et al., 2013) or rely on clever simplification and approximation of the likelihood function (Li and Stephens, 2003; McVean and Cardin, 2005) are being developed and applied to large scale genomics data. Moreover, two recent methodologies permit the inference of population history using whole genome sequences. Li and Durbin (2011) developed the Pairwise Sequentially Markovian Coalescent model (PSMC) to reconstruct the distribution of the time since the most recent common ancestor (TMRCA) between the two alleles of an individual and infer population size changes over time. Gronau et al. (2011) extracts information from gene trees relating haplotypes sampled from multiple populations to estimate population sizes, divergence times and migration rates. Both methods use one or several whole genome sequences of each population.

This dissertation focuses on demographic and population separation history inference based on whole genome sequences, an effort to extend current methodology and apply available techniques to answer questions about population history in human, gorilla and canine species. In this chapter, I will introduce the sequencing technologies used in this dissertation and review several state of the art methodologies for population history inference. Moreover, I will introduce existing knowledge of human, gorilla, canine evolution based on preivous studies. Finally, I will give an overview of the following chapters.

1.2 Background

1.2.1 High-throughput resequencing technologies

The first finished-grade human genome sequence was released in 2004 (Consortium et al., 2004), leading to a new era of genomics research. The limitations of automated Sanger sequencing called for a need for new and improved technologies for sequencing large numbers of genomes. Newer methods are referred to as next-generation sequencing (NGS), capable of sequencing millions of DNA molecules in parallel (reviewed in Metzker (2010)).

The main characteristics of NGS sequencing is the production of millions of short reads (100-150 bp) and its application to resequencing studies. A typical analysis involves several steps: the sequence reads are first mapped to a reference genome, then single-nucleotide polymorphisms (SNPs) are discovered or genotyped at given alleles. The choice of sequencing strategy, mapping/SNP calling algorithm and quality checking is crucial to downstream analysis, especially for population genetics based analysis. For example, accurate calling of heterozygous sites requires high-coverage data to account for the effects of sequencing errors and stochasticity in sampling each allele.

Read mapping algorithms like BWA can efficiently align short sequencing reads against a large reference sequence allowing mismatches and gaps (Li and Durbin, 2009). For species for which only the genome assembly of close relatives exist, tools like Stampy can deal with reads that include sequence variation, achieving both speed and sensitivity (Lunter and Goodson, 2011). The most commonly used software for making hard genotype calls by computing genotype likelihhods, namely the probability of the observed data given every possible diploid genotype, is the Genome Analysis Toolkit (GATK) (McKenna et al., 2010; DePristo et al., 2011). Since reads are aligned independently, initial mapping tends to have misalignment of reads spanning indels. The per-base quality scores are also inaccurate and vary with sequence technology, machine cycle and sequence context. GATK takes the initial SAM/BAM files from the aligner and assembles methods to perform local realignment, base quality recalibration and multi-sample SNP calling. A standard workflow for variant discovery and genotyping requires the following steps: 1) mark duplicates: identify duplicated reads since the same DNA molecules can be sequenced several times 2) local realignment around indels: identify intervals that need to be realigned and then perform the actual realignment of reads 3) base quality score recalibration: analyze patterns of covariation in the sequence and then apply an empirically accurate error model to the bases 4) discover or genotype variants using UnifiedGenotyper or HaplotypeCaller.

After genotypes have been called, there are several options to do variant filtration to ensure that only the most accurate data are used in downstream analysis (Auwera et al., 2013). GATK incorporates Variant Qaulity Score Recalibration (VQSR), which uses machine-learning methods to assign a well calibrated probability to each variant call in a raw call set. We can use this variant quality score to filter the raw call set to produce a subset of calls with the desired level of quality. In detail, VQSR trains a recalibration model using user-defined statistics (eg. Coverage, Qual-

ByDepth, MappingQualityRankSumTest) and known variants sites (eg. 1000 Genomes Project, HapMap, Omni) to estimate the relationship between variant call annotations and the probability that a variant call is a true genetic variant rather than sequencing error. The program then applies this model to all the variants and annotates each variant with a quality score called VQSLOD. It's recommended to retrieve 99.9% of true variants from the truth-training sets to achieve both high sensitivity and specificity. VQSR should be applied to SNPs and indels separately. An alternative approach is to apply hard filtering, if there are not enough variants for the statistical methods like VQSR to be applicable or there are no truth-site resources available. Hard filtering can use any annotations like Depth (DP), QualByDepth (QD), Fisherstrand (FS), RMSMappingQuality (MQ), HaplotypeScore and so on. Many population genetics statistics or algorithms, like nucleotide diversity, also require knowledge of the callable sites, namely the sites that are reliably called. Thus, it's critical to apply the same filtering criteria to all sites.

SNP arrays are a type of DNA microarray designed to genotype thousands of SNPs across individuals. SNP arrays are widely used in GWAS studies as well as population genetics analysis, but ascertainment bias needs to be considered since the SNP discovery panel is often small and focused in specific populations (Clark et al., 2005).

Standard next generation sequencing technology is ”phase-insensitive”, ignoring the diploid nature of most organisms. Although there are statistical phasing techniques, modeling individual haplotypes as a mosaic of known haplotypes (Li and Stephens, 2003), experimental phasing approaches including physically separating the chromosomes during cell division Dear and Cook (1989), and recovering haplotypes from long DNA fragments, by using fosmid clones (Burgtorf et al., 2003; Kitzman et al., 2011) or dilution haplotyping are also developed (Kaper et al., 2013; Kuleshov et al., 2014).

1.2.2 Pairwise sequentially Markovian coalescent (PSMC) model and other derivatives

Coalescent theory is the most popular model that relates mutation, recombination and population structure with genetic variation (Wakeley and Wakeley, 2009). It's a natural extension of classical population genetics theory and was first proved mathematically by Kingman (Kingman, 1982). The basic idea underlying the coalescent is that, in the absence of selection, sampled lineages can be viewed as randomly choosing parents as we go backward in time (Rosenberg and Nordborg, 2002). The times back to common ancestors among members of the sample are modeled by a Poisson process. Eventually, all lineages coalesce into a single lineage, namely the most recent common ancestor (MRCA) of the sample. The number of mutations that occur over a branch of a given length is also poisson distributed with rate proportional to the mutation parameter and the branch length.

Recombination can also be added into this framework. Recombination allows linked sites to have different genealogical trees. As we proceed from one end of the sequence to the other, the genealogy changes as recombination happens. As the recombination approaches infinity, the genealogies of unlinked loci are independent given the historical demography. When modeling continuous DNA sequences instead of multiple unlinked loci, recombination events result in a correlation structure that is complex and non-Markovian since the distribution of the next genealogy depends not just on the current genealogy but also all previous ones. The state space of the ancestral recombination graph (ARG) is huge and likelihood-based inference is restricted because there exists no numerical expression and the construction of efficient Monte Carlo methods for estimating the likelihood is technically challenging. McVean and Cardin (2005) proposed a simplification of the coalescent process in which coalescence between lineages with no overlapping ancestral material is banned. The resulting process has a simple Markovian structure when generating genealogies sequentially along a sequence, also called 'Sequential Markovian Coalescent' (SMC).

Li and Durbin (2011) developed the first algorithm that applies SMC on whole genome sequences to do population history inference. The idea is that a diploid genome sequence contains hundreds of thousands of indepedent loci, each with its own TMRCA between two alleles. The distribution of TMRCA is informative about population history as the rate of coalescent events in any given epoch is an inverse function of population size at that time. They proposed the pairwise sequenctially Markovian coalescent (PSMC) model, which is a specialization to the case of two chromosomes of the SMC model. PSMC builds a hidden Markov model to infer local TMRCA based on the local density of heterozygotes, whereas hidden states are discretized time to the most recent common ancestors (TMRCA), and transition represents ancestral recombination events. The discrete-time Markov chain is derived from the continous-time Markov chain by integrating probability densities in time intervals. Time intervals are chosen to be approximately evenly distributed in the log space. Given a maximum TMRCA T_{max} and a number of atomic time intervals n, the boundaries of these intervals are $t_i = 0.1 \exp[i/nlog(1 + 10T_{max})] - 0.1, i = 0, ..., n$. On autosome, $T_{max} = 15, n = 64$. A hidden state k in the HMM means a coalescence between the two haplotypes at this point in the sequence lies in the time interval $[t_k, t_{k+1})$. When using expectation-maximisation (EM), parameters at the last iteration (by default the twentieth) were taken as the final results. The stationary distribution of the hidden states is σ_k, namely the probability that a segment coalesced at time interval $[t_k, t_{k+1})$. Given this distribution, we can calculate the effective population size for each time interval, which gives us the PSMC curve (step wise function of time vs effective population size). However, for both small and large t, the expected number of segments in an interval is very small, leading to overfitting due to insufficient data. In such case, blocks of adjacent atomic intervals were combined to have the same population-size parameter. The free parameters of this model include the scaled mutation rate, the recombination rate and piecewise constant ancestral population sizes λ_t.

PSMC is relatively straightforward to apply, with whole genome sequences of at least 20X coverage to guarantee that the majority of heterozygous sites are not sequencing errors and most homozygous sites are called reliably. The criteria for callable regions are 1) read depth is within half and twice of the average read depth 2) the root mean squared mapping quality of reads covering the locus is above 25 3) no indels within 10bp. The consensus sequences were further divided into 100-bp non-overlapping bins, decoded by '.' (missing) if $\geqslant 90$ bases were filtered or uncalled, '1' if >10bp were called and there was at least one heterozygote, or '0' otherwise, as input of the PMSC. The estimated TMRCA is in units of $2N_0$ time and λ_t is scaled to N_0 as well. In order to convert to real value, mutation rate and generation time needs to be provided.

Since PSMC only models two sequences, the coalescent event joining the sequences at the most recent common ancestor almost always occurred more than 20,000 years ago, setting a lower bound for population history inference. Several studies developed methods to model more than two sequences, while overcoming the problem of extensive computation. Sheehan et al. (2013) approximate the conditional sampling process for adding an $(n+1)$th sequence to the distribution of genealogies connecting n sequences. Schiffels and Durbin (2014) extends PSMC to the mulitple sequential Markovian coalescent (MSMC) that simplifies the relationship at a given location between multiple samples by looking at 1) the time to the most recent common ancestor of any two sequences, that happens the first 2) the total length of all singleton branches in the tree. They derive approximate transition and emission rates using the sequentially Markovian coalescent (SMC') framework (McVean and Cardin, 2005; Marjoram and Wall, 2006). If sequences are sampled from different subpopulations, an additional free parameters for coalescence rates within and across population boundaries are added to infer how subpopulations separated over time. When having two haplotypes, MSMC is called PSMC', slightly different from PSMC because it uses the SMC' model, which accounts for recombination events between segments with the same time to coalescence. Wilton et al. (2015) showed that SMC' is a highly accurate approximation to the ancestral recombination graph. MSMC can be applied to eight haplotypes at most, giving us picture of population history as recently as 70 generations ago. MSMC can also model population separation, using the ratio between the cross-population and within-population coalescence rates- 'relative cross coalescence rate'. The relative cross coalescence rate should be close to 1 when the two populations are well mixed and 0 after they have fully separated, providing an intuitive rather than parametric way to model population separations. Similarly, PSMC can also model population split by looking at pseudo-diploid genomes in which two haplotypes obtained from different populations. The inferred TMRCA distribution is informative about the timing of population splits since the time after which nearly no coalescence events occur is a good estimate for the population split time. When modeling population splits, both PSMC and MSMC require phased haplotypes, raising the question of how phasing errors affect such analysis. Besides, both methods fail to provide quantified estimates of split times and migration rates, as split times estimates can be confounded by subsequent migration.

1.2.3 Bayesian Coalescence-based demographic inference

The mathematical aspects of the coalescent model make it possible for likelihood based inference.

$$L = \sum_{G} P(D|G,\mu)P(G,\alpha) \tag{1.1}$$

where α is the collection of parameters (such as populations sizes and migration rates), G is the genealogy, and D is the observed data (typically DNA sequences). However, this is not feasible because summing over all possible genealogies is computationally expensive and difficult. Advanced computational techniques, such as importance sampling (Stephens and Donnelly, 2000) and Markov chain monte carlo (Nielsen and Wakeley, 2001; Beerli and Felsenstein, 2001) have been applied to this problem.

Gronau et al. (2011) adapted the MCMCcoal program by Burgess and Yang (2008); Rannala and Yang (2003) and added two new features to the G-PhoCS program: 1) modeling of gene flow between populations 2) handling of unphased diploid genotypes in the input. G-PhoCS assume a known population phylogeny tree T, in which each population p (current and ancestral) is associated with a parameter θ_p for effective population size and each ancestral population is associated with a divergence time τ_p. G-PhoCS takes a set of mulitple sequence alignments as input. G-PhoCS makes the following assumptions to enable factorization for efficient calculation: 1) independent priors are assumed for model parameters θ_p and τ_p, usually Gamma distributions defined by the user 2) genealogies at distinct loci are assumed to be independent given the model parameters (thus requiring loci to be with sufficient interlocus recombination) 3) each locus-specific genealogy G_i induces conditional independence of the corresponding alignment, X_i and the model parameters. Hence the contribution of a locus i to the complete data density, $P(G_i, X_i|T, \theta_p, \tau_p)$, can be expressed as a product of a genealogy prior, $P(G_i|T, \theta_p, \tau_p)$, and a locus data likelihood, $P(X_i|G_i)$.

Migration is introduced in G-PhoCS through migration bands, a directed pair of populations $S \rightarrow T$, where S is the source population and $T$is the target population. The life span of the migration band is the time interval during which both S and T exist. Each migration band is associated with a rate parameter $m_{ST} = M_{ST}/\mu$, where M_{ST} is defined as the proportion of individuals in population T that arose by migration from population S per generation.

When applying G-PhoCS, users need to specify the phylogeny, migration scenarios, fine-tune parameters for the update steps of the MCMC, priors for parameters, initial value for each parameter and the number of iterations. The success of MCMC requires that the Markov chain has converged and it is mixing adequately well to explore the parameter space. The convergence of the Markov chain needs to be checked. The mean value of the retained samples are used as point estimates and 95% higest posterior density intervals are used as credible intervals. Compared with PSMC/MSMC methods, G-PhoCS takes much longer time to achieve convergence and the running time increase dramatically with the number of haplotpyes and migration scenarios.

G-PhoCS assumes the input alignments represent a set of putative ”neutral loci” in which recombination occurred at negligible rates, and between which recombination occurred at sufficiently high rates. Gronau et al. (2011) have identified contiguous intervals of 1000 bp and then select a subset of these intervals that ensured a minimum inter-locus distance of 50,000 bp. Indels, simple repeats, recent transposable elements, segmental duplications, exons of protein-coding genes and conserved noncoding elements are avoided. Positions corresponding to hypermutable CpG dinucleotides are also masked. G-PhoCS only infers ratios between model parameters. To obtain absolute values, it's necessary to calibrate the estimates using either a known mutation rate or an estimated divergence time for outgroup species.

1.2.4 Approximate Bayesian Computation in population genetics

Likelihood-based inference methods are quite powerful, making it possible to explore complex and parameter-rich demographic models. However, for many problems in population genetics inference, the likelihood function is intractable. In these cases, an attractive alternative method is approximate Bayesian computation (ABC). ABC approaches bypass exact likelihood calculation by using summary statistics and simulations. This is more applicable to population genetics studies since simulation based on the coalescent is relative easy and fast (Csilléry et al., 2010).

The most basic ABC algorithm is based on rejection sampling (Pritchard et al., 1999). We sample parameters from a prior distribution and generate datasets from the model using candidate parameters. We then compare the simulated data to the observed data by computing a distance between them. If this distance is small enough, meaning that the simulated data is close enough to the observed data, we keep the sampled parameters, otherwise we discard it. The obtained parameter values constitute the resulting posterior distribution and a credibility interval for the parameter of interest.

ABC based on rejection sampling is of low efficiency since the parameter space increase exponentially with the number of parameters. Other statistical techniques have been applied to ABC algorithm. Markov chain Monte Carlo (MCMC) sampling has been used in an ABC framework, called ABC-MCMC, which explores the parameter space iteratively using the distance between the simulated and the observed summary statistics to update the current parameter values (Wegmann et al., 2009). However, MCMC chains are easily in danger of getting stuck causing a high rejection rate. Besides, MCMC chains cannot be parallelized.

Sequential Monte Carlo sampling differs from the MCMC sampling by its use of a particle filter. The algorithm works with large pools of candidates, called particles. At each stage of the algorithm, the particles are perturbed and then filtered, bringing the pool closer and closer to a sample drawn from the desired posterior. Toni et al. (2009) derived the ABC sequential Monte Carlo sampling (ABC SMC) algorithm that applied a sequential importance sampling algorithm. ABC-SMC starts with a number of sampled parameter values θ (called particles) from the prior distribution $\pi(\theta)$. For each iteration t, the normal ABC rejection sampling are applied to this pool until the number of accepted particles reaches N. The weight for each particle in the accepted pool is updated

$$w_t^{(i)} = \begin{cases} 1, & \text{if } t = 0 \\ \frac{\pi(\theta_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}, \theta_t^{(i)})}, & \text{if } t > 0 \end{cases} \tag{1.2}$$

For next iteration, sample parameter θ from the previous population θ_{t-1} with weights w_{t-1} and perturb the particle using a perturbation kernel. As iteration goes up, the tolerances ϵ_t is chosen to be smaller, $\epsilon_1 > ... > \epsilon_T \geq 0$, thus the distributions gradually evolve towards the target posterior. The perturbation kernel K_t is usually a random walk (uniform or Gaussian). When $T = 1$, the ABC SMC algorithm corresponds to the ABC rejection algorithm.

ABC-SMC can be easily parallelized. The choice of the tolerance ϵ can affect the variance of the posterior distribution. Sometimes ϵ is chosen to keep a certain percentage of the sampled parameters. As the models become more complex with more parameters and summary statistics, finding accepted parameters becomes more difficult. Wegmann et al. (2009) proposed to transform large number of summary statistics using partial least square (PLS) regression to pull out the most important latent structure of the summary statistics.

ABC algorithms are widely used in population genetics analysis. Numerous studies have applied approximate Bayesian Computation to study demographic history of human, chimpanzee and gibbon (Prado-Martinez et al., 2013; Veeramah et al., 2015; 2011). Frantz et al. (2015) applied ABC framework to do model testing on different domestication models of pigs.

1.2.5 Ancient DNA studies

Ancient DNA (aDNA) gives us unique information about the past that is normally unavailable when using data from contemporary individuals - ancient DNA offers the ability to analyze the genetic patterns that existed at a particular time and geographical location (Pickrell and Reich, 2014). aDNA enables direct inference about the relationship of historical population to populations living today. For example, direct comparison of human and Neanderthal DNA confirmed Neanderthal admixture in modern humans and provided estimates of the time period of admixture as well as the proportion of Neanderthal ancestry existed in modern humans (Green et al., 2010; Prüfer et al., 2014). Apart from providing insights into demographic history, aDNA studies are crucial in understanding the effect of selection. For example, the absence of lactase persistence allele in ancient Europeans supports the hypothesis of strong positive selection at the lactase locus in modern Europeans (Burger et al., 2007). aDNA also helps determine mutation rate. For example, the 35,000 year old wolf genome allowed recalibration of the lupine mutation rate (Skoglund et al., 2015).

Obtaining reliable sequence data from ancient DNA specimens is challenging because of DNA degradation after the death of the organism, resulting in low levels of DNA, degradation of DNA into small fragments and post-mortem damage of DNA (Veeramah and Hammer, 2014). Advances in second-generation sequencing lead to a bulk of studies in analyzing ancient DNA genomes (Stoneking and Krause, 2011; Knapp and Hofreiter, 2010). However, due to low endogenous DNA content, some ancient specimens are only sequenced in very low coverage, unable to produce reliable calls for every nucleotide sites. One way of overcoming this uncertainty is to focus the analysis on sites that are confidently known to be biallelic in modern samples, which allows descriptive assessments of the relationship between the ancient specimen and contemporary samples through principal component analysis, structure type analysis and f statistics analysis (Green et al., 2010). Even if the DNA content from ancient remains is sufficient to generate high coverage sequencing data, aDNA sequences tend to be substantially enriched for C>T substitutions at the 5' end of reads, and G>A substitutions at the 3' end. One simple solution is to remove such sites either bioinformatically or through chemical treatment with uracil DNA glyosylase. Algorithms are also developed to quantify the effect of post-mortem damage, thus mitigate its effect for downstream analysis (Jónsson et al., 2013).

1.2.6 Overview of human, gorilla and canine evolutionary history

There are two competing models to explain how members of the genus *Homo* underwent the transition from archaic humans to anatomically modern humans: multiregional evolution (MRE) model and the recent African origin (RAO) model. The mtNDA study reported a phylogenetic tree that relates all human mtDNA to a common African ancestor that lived around 200,000 years ago, supporting the recent African origin model (Walker et al., 1987). Results from the non-recombining region of Y chromosome (Thomson et al., 2000) and from autosomal loci (Hawks and Wolpoff, 2001) were also consistent with an African origin of anatomically modern humans. Furthermore, analysis of genome-wide data from globally distributed modern-day individuals showed a pattern of decreased genetic diversity and increased linkage disequilibrium (LD) as a function of the distance from eastern or southern Africa (Ramachandran et al., 2005), suggesting that anatomically modern humans experienced a serires of founder events as they expanded out of Africa.

The timing of dispersal from Africa is still under great controversy. Single locus analysis based on non-recombining region of Y chromosome suggested that non-African samples shared a common ancestor around 40,000 to 60,000 years ago (Thomson et al., 2000). However, a recent mtDNA study that applied ancient DNA to calibrate mtDNA mutation rate suggested that the timing of the most recent common ancestor of African and non-African mitochondrial DNA, around 78,300 years ago (62,400-94,900 years ago) and the timing of the MRCA for all modern humans at 157,000 years ago (134,000-188,000 years ago) (Fu et al., 2013). Studies that applied whole genome sequence data and a range of statistical modelling approaches such as by evaluating the allele frequency spectrum, the distribution of tracts of identity by state or coalescent patterns at short interspersed loci reached different conclusions. Most studies assumed a mutation rate of 2.5×10^{-8} bp per generation. Gronau et al. (2011) applied applied Bayesian coalescence based model (G-PhoCS) on 37,574 1kb 'neutral loci' and inferred that west African diverged around 47,000 years ago. Harris and Nielsen (2013) using IBS sharing inferred that the divergence between African and non-African happened around 55,0000 years ago. Gravel (2012) using allele frequency spectrum inferred that the divergence happened around 51,000 years ago. Li and Durbin (2011) applied pairwise sequentially Markovian coalescent (PSMC) model and paired X chromosomes from African and non-African males and suggested that the two groups remained as one population until 60-80 kyrs ago with substantial genetic exchange up until 20-40 kyrs ago. However, recent studies on mutation rate suggested a rate of $1.25 \times 10^{-8)}$ bp per generation (Kong et al., 2012). Using this mutation rate, Schiffels and Durbin (2014) applied multiple sequentially Markovian coalescent (MSMC) model on statistically phased genomes (two or four haplotypes per population) and suggested that African and non-African populations exhibited a slow, gradual separation beginning earlier than 200,000 years ago and lasting until about 40,000 years ago, while the median point of such divergence was around 60,000 – 80,000 years ago. Part of the controversy lies in the different mutation rate assumed to recalibrate divergence time. Despite mutation rate difference, subsequent migration and ancestral population structure also affect divergence time estimates. Most studies indicate that the separation of human populations was a gradual event with substantial genetic exchange after the initial split. However, some models such as PSMC or MSMC failed to provide quantitative measure of split times and migration rates.

Gorillas are human's closet living relatives other than chimpanzees and are as well important for the study of human evolution. Hobolth et al. (2007) inferred that the divergence time between human and chimpanzee is around 5.5 to 7 Myr ago, and the divergence time between human, chimpanzee and gorilla is 8.5 to 12 Myr ago considering possible variation of mutation rates. There has been considerable effort to estimate split times and population sizes for western and eastern gorillas (Ackermann and Bishop, 2010; Prado-Martinez et al., 2013; Scally et al., 2012; Thalmann et al., 2011; 2007). These studies make use of disparate data sets and modeling assumptions, particularly in terms of the treatment of gene flow subsequent to initial population separations. Early studies based on eight microsatellites Thalmann et al. (2011) suggested that the separation of Cross River and western lowland gorilla populations occurred 17.8 kyrs ago, followed by a comparatively high level of gene flow. On the other hand, Prado-Martinez et al. (2013) estimated this population divergence time at 114 kyrs ago based on a modified PSMC approach (Note: the above mentioned values have been adjusted to match the mutation rate used in this study where appropriate). The random phasing procedure applied in the modified PSMC approach may not be appropriate for such recent population split times. Moreover, estimates of the separation of eastern gorillas from the western lowland/Cross River gorillas range from about 100 kyrs to 450 kyrs, with varying degrees, lengths and directions of gene flow (Ackermann and Bishop, 2010; Becquet and Przeworski, 2007; Mailund et al., 2012; Prado-Martinez et al., 2013; Scally et al., 2012; Thalmann et al., 2007).

The origins and dynamics of dog (*Canis familiaris*) domestication have long been an interesting and controversial questions for both geneticists and archaeologists. There are a great number of whole genome sequences available on various village dogs and breeds and over 5000 samples with

SNP array data. This faciliate the study of dog evolutionary history. Although it is widely accepted that dogs were domesticated from a gray wolf (*Canis lupus*) (Vilà et al., 1997), the location and timing of such domestication and the specific wolf population that dogs are domesticated from are still largely debated (Perri, 2016; Horard-Herbin et al., 2014). Savolainen et al. (2002) used markers on mtDNA to suggest that dogs originated from Southeast Asia 15,000 years ago. VonHoldt et al. (2010) showed that dog breeds shared a higher proportion of multi-locus haplotypes unique to grey wolves from the Middle East, suggesting that they are a dominant source of genetic diversity. Shannon et al. (2015) proposed Central Asia as a domestication origin based on LD-decay curves: LD is lowest in Afghanistan and Central Asia at short inter-SNP distances. Wang et al. (2016) included dogs from southern East Asia and suggested they have higher genetic diversity compared to other populations and are the basal group relating to gray wolves, indicating southern East Asian origin of domestic dogs 33,000 years ago. The analysis of complete mitochondrial genomes of ancient canids suggest that dogs are domesticated in Europe since all modern dogs are phylogenetically most closely related to either ancient or modern canids of Europe (Thalmann et al., 2013). Recently, Frantz et al. (2016) analyzed mtNDA loci from several ancient Neolithic dogs and suggested a prehisotric turnover of mtDNA lineages in the European continuent sometime between the late Neolithic and today. They also proposed a dual origin for domestication given that the inferred split time between Asian and European dogs occurred later than the oldest fossil remains found in Europe.

1.3 Dissertation overview

With the advance of sequencing technologies and computational methods to infer population history, a great number of analysis can be performed to answer questions about the past history of not only humans, but also our close relatives and companions. The goal of this dissertation is to extend current methodology and apply available techniques to answer questions about population history in human, gorilla and canine species.

In chapter II, we reconstruct highly accurate phased haplotypes using fosmid pool sequencing and assess the accuracy of statistically phased haplotypes in the 1000 Genomes Phase III project. We also assess the accuracy of haplotypes phased by ShapeIT (Delaneau et al., 2008) using an existing reference panel. This helps us quantify the extent of phasing errors using existing statistical phasing algorithms. In chapter III, we analyzed population size and separation history using the pairwise sequenctially Markovian coalescent (PSMC) and multiple sequenctially Markovian coalescent (MSMC) model on our reconstructed haplotypes. We aimed to see the effect of phasing errors on such analysis and compare the results with other methods. We further extended PSMC with Approximate Bayesian Computation to model population splits by fitting the inferred TMRCA distribution obtained from PSMC on pseudo-diploid genomes to a standard Isolation-with-Migration model. We improve the inference of population separation history using PSMC and MSMC, clarify the strengths and limitations of existing methods and data, and contribute additional knowledge about several key events in human history. The fosmid pool sequencing experiments were performed by Elzbieta Sliwerskas and Sarah Emery.

In chapter IV, we applied G-PhoCS, a Bayesian coalescent-based approach to infer ancestral population sizes, divergence times and migration rates among the three gorilla subspecies. We evaluated different migration settings, adjusted model parameters to help the MCMC algorithm converge to a stationary distribution, summarized and recalibrated each statistic to provide reasonable estimates. We also used a diffusion approximation approach (using site frequency spectrum) to infer temporal changes in western lowland gorilla effective population size. The SFS analyses were performed by Kimberly McManus and Joanna Kelley.

New technologies offer the ability to directly sample genetic material from ancient samples, a key methodology for testing hypothesis about historical population processes. This is especially useful in studying evolutionary history of canine species and the dynamics of domestication. All the sampled Old World wolves form a sister monophyletic clade and so does all the dogs. It is difficult to find the putative domestication location by looking at genetic data of contemporary canines alone since it is possible that dogs are domesticated from a wolf population that later went extinct. Thus it's time to joint efforts from archaeologists and geneticists and extend population genetics analysis to ancient remains. In chapter V, we analyzed two ancient dog genomes from the Neolithic and over 100 contemporary canine genomes to understand the evolutionary history of dogs in Europe since the primary wolf divergence. A previous study based on mitochondrial DNA and a late Neolithic Irish genome suggested a Late Neolithic population replacement and a dual origin of domestic dogs (Frantz et al., 2016). We aimed to examine whether and to what extent a large-scale replacement occurred during the Neolithic by performing comprehensive population genetics analysis on three ancient dog genomes and other contemporary canine genomes. We also applied G-PhoCS to infer a complete demographic model for dogs and wolves and further developed a numerical method to date the divergence time between ancient dogs and modern European dogs. The mitochondria analysis was performed by Matthew Oetjens, population structure analysis (PCA,ADMIXTURE) were performed by Aangela M. Taravella, population structure analysis (NGSadmix, SpaceMix,ADMIXTUREGRAPH) were performed by Krishna R. Veeramah, domestication loci analysis were performed by Amanda L. Pendleton. DNA isolation and screening were performed by Amelie Scheu, Kevin Daly, Shyamie Gopalan, Martina Unterländer, Joachim Burger. I performed the variants calling of ancient and contemporary dog genomes, f statistics based analysis, G-PhoCS analysis. Krishna R. Veeramah and I developed the numerical method to infer divergence time between ancient samples and European populations.

Bibliography

Ackermann, R. R. and Bishop, J. M. Morphological and molecular evidence reveals recent hybridization between gorilla taxa. *Evolution*, 64(1):271–290, 2010.

Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J. et al. From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, pages 11–10, 2013.

Becquet, C. and Przeworski, M. A new approach to estimate parameters of speciation models with application to apes. *Genome research*, 17(10):1505–1519, 2007.

Beerli, P. and Felsenstein, J. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, 98(8):4563–4568, 2001.

Behar, D. M., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N. M., Kivisild, T., Torroni, A. and Villems, R. A “copernican” reassessment of the human mitochondrial dna tree from its root. *The American Journal of Human Genetics*, 90(4):675–684, 2012.

Burger, J., Kirchner, M., Bramanti, B., Haak, W. and Thomas, M. G. Absence of the lactase-persistence-associated allele in early neolithic europeans. *Proceedings of the National Academy of Sciences*, 104(10):3736–3741, 2007.

Burgess, R. and Yang, Z. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Molecular biology and evolution*, 25(9):1979–1994, 2008.

Burgtorf, C., Kepper, P., Hoehe, M., Schmitt, C., Reinhardt, R., Lehrach, H. and Sauer, S. Clone-based systematic haplotyping (csh): a procedure for physical haplotyping of whole genomes. *Genome research*, 13(12):2717–2724, 2003.

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. and Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome research*, 15(11):1496–1502, 2005.

Consortium, I. H. G. S. et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

Csilléry, K., Blum, M. G., Gaggiotti, O. E. and François, O. Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

Dear, P. and Cook, P. Happy mapping: a propoaal for linkage mapping the human genome. *Nucleic acids research*, 17(17):6795–6807, 1989.

Delaneau, O., Coulonges, C. and Zagury, J.-F. Shape-it: new rapid and accurate algorithm for haplotype inference. *BMC bioinformatics*, 9(1):1, 2008.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M. et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498, 2011.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. and Foll, M. Robust demographic inference from genomic and snp data. *PLoS Genet*, 9(10):e1003905, 2013.

Frantz, L. A., Mullin, V. E., Pionnier-Capitan, M., Lebrasseur, O., Ollivier, M., Perri, A., Linderholm, A., Mattiangeli, V., Teasdale, M. D., Dimopoulos, E. A. et al. Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*, 352(6290):1228–1231, 2016.

Frantz, L. A., Schraiber, J. G., Madsen, O., Megens, H.-J., Cagan, A., Bosse, M., Paudel, Y., Crooijmans, R. P., Larson, G. and Groenen, M. A. Evidence of long-term gene flow and selection during domestication from analyses of eurasian wild and domestic pig genomes. *Nature genetics*, 47(10):1141–1148, 2015.

Fu, Q., Mittnik, A., Johnson, P. L., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J. et al. A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology*, 23(7):553–559, 2013.

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

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y. et al. A draft sequence of the neandertal genome. *science*, 328(5979):710–722, 2010.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. and Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics*, 43(10):1031–1034, 2011.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. and Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet*, 5(10):e1000695, 2009.

Harris, K. and Nielsen, R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet*, 9(6):e1003521, 2013.

Hawks, J. D. and Wolpoff, M. H. The four faces of eve: hypothesis compatibility and human origins. *Quaternary International*, 75(1):41–50, 2001.

Hobolth, A., Christensen, O. F., Mailund, T. and Schierup, M. H. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS Genet*, 3(2):e7, 2007.

Horard-Herbin, M.-P., Tresset, A. and Vigne, J.-D. Domestication and uses of the dog in western europe from the paleolithic to the iron age. *Animal Frontiers*, 4(3):23–31, 2014.

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. and Orlando, L. mapdamage2. 0: fast approximate bayesian estimates of ancient dna damage parameters. *Bioinformatics*, page btt193, 2013.

Kaper, F., Swamy, S., Klotzle, B., Munchel, S., Cottrell, J., Bibikova, M., Chuang, H.-Y., Kruglyak, S., Ronaghi, M., Eberle, M. A. et al. Whole-genome haplotyping by dilution, amplification, and sequencing. *Proceedings of the National Academy of Sciences*, 110(14):5552–5557, 2013.

Karmin, M., Saag, L., Vicente, M., Sayres, M. A. W., Järve, M., Talas, U. G., Rootsi, S., Ilumäe, A.-M., Mägi, R., Mitt, M. et al. A recent bottleneck of y chromosome diversity coincides with a global change in culture. *Genome research*, 25(4):459–466, 2015.

Kingman, J. F. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982.

Kitzman, J. O., MacKenzie, A. P., Adey, A., Hiatt, J. B., Patwardhan, R. P., Sudmant, P. H., Ng, S. B., Alkan, C., Qiu, R., Eichler, E. E. et al. Haplotype-resolved genome sequencing of a gujarati indian individual. *Nature biotechnology*, 29(1):59–63, 2011.

Knapp, M. and Hofreiter, M. Next generation sequencing of ancient dna: requirements, strategies and perspectives. *Genes*, 1(2):227–243, 2010.

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A. et al. Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature*, 488(7412):471–475, 2012.

Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M. and Snyder, M. Whole-genome haplotyping using long reads and statistical methods. *Nature biotechnology*, 32(3):261, 2014.

Li, H. and Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

Li, H. and Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

Li, N. and Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

Lunter, G. and Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome research*, 21(6):936–939, 2011.

Mailund, T., Halager, A. E., Westergaard, M., Dutheil, J. Y., Munch, K., Andersen, L. N., Lunter, G., Prüfer, K., Scally, A., Hobolth, A. et al. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genet*, 8(12):e1003125, 2012.

Marjoram, P. and Wall, J. D. Fast" coalescent" simulation. *BMC genetics*, 7(1):1, 2006.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

McVean, G. A. and Cardin, N. J. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1459):1387–1393, 2005.

Metzker, M. L. Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1):31–46, 2010.

Nielsen, R. and Wakeley, J. Distinguishing migration from isolation: a markov chain monte carlo approach. *Genetics*, 158(2):885–896, 2001.

Novembre, J. and Ramachandran, S. Perspectives on human population structure at the cusp of the sequencing era. *Annual review of genomics and human genetics*, 12:245–274, 2011.

Perri, A. A wolf in dog's clothing: Initial dog domestication and pleistocene wolf variation. *Journal of Archaeological Science*, 68:1–4, 2016.

Pickrell, J. K. and Reich, D. Toward a new history and geography of human genes informed by ancient dna. *Trends in Genetics*, 30(9):377–389, 2014.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G. et al. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, 2013.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C. et al. The complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505(7481):43–49, 2014.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W. and Cavalli-Sforza, L. L. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942–15947, 2005.

Rannala, B. and Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4):1645–1656, 2003.

Rosenberg, N. A. and Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5):380–390, 2002.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. and Feldman, M. W. Genetic structure of human populations. *science*, 298(5602):2381–2385, 2002.

Savolainen, P., Zhang, Y.-p., Luo, J., Lundeberg, J. and Leitner, T. Genetic evidence for an east asian origin of domestic dogs. *Science*, 298(5598):1610–1613, 2002.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T. et al. Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388):169–175, 2012.

Schiffels, S. and Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919–925, 2014.

Schraiber, J. G. and Akey, J. M. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 2015.

Shannon, L. M., Boyko, R. H., Castelhano, M., Corey, E., Hayward, J. J., McLean, C., White, M. E., Said, M. A., Anita, B. A., Bondjengo, N. I. et al. Genetic structure in village dogs reveals a central asian domestication origin. *Proceedings of the National Academy of Sciences*, 112(44):13639–13644, 2015.

Sheehan, S., Harris, K. and Song, Y. S. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. *Genetics*, 194(3):647–662, 2013.

Skoglund, P., Ersmark, E., Palkopoulou, E. and Dalén, L. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology*, 25(11):1515–1519, 2015.

Stephens, M. and Donnelly, P. Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635, 2000.

Stoneking, M. and Krause, J. Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics*, 12(9):603–614, 2011.

Thalmann, O., Fischer, A., Lankester, F., Pääbo, S. and Vigilant, L. The complex evolutionary history of gorillas: insights from genomic data. *Molecular biology and evolution*, 24(1):146–158, 2007.

Thalmann, O., Shapiro, B., Cui, P., Schuenemann, V. J., Sawyer, S. K., Greenfield, D., Germonpré, M., Sablin, M., López-Giráldez, F., Domingo-Roura, X. et al. Complete mitochondrial genomes of ancient canids suggest a european origin of domestic dogs. *Science*, 342(6160):871–874, 2013.

Thalmann, O., Wegmann, D., Spitzner, M., Arandjelovic, M., Guschanski, K., Leuenberger, C., Bergl, R. a. and Vigilant, L. Historical sampling reveals dramatic demographic changes in western gorilla populations. *BMC evolutionary biology*, 11(1):85, 2011.

Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J. and Feldman, M. W. Recent common ancestry of human y chromosomes: evidence from dna sequence data. *Proceedings of the National Academy of Sciences*, 97(13):7360–7365, 2000.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O. et al. The genetic structure and history of africans and african americans. *science*, 324(5930):1035–1044, 2009.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. P. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., Kauffman, E., Bonné-Tamir, B., Bertranpetit, J., Francalacci, P. et al. Y chromosome sequence variation and the history of human populations. *Nature genetics*, 26(3):358–361, 2000.

Veeramah, K. R. and Hammer, M. F. The impact of whole-genome sequencing on the reconstruction of human population history. *Nature Reviews Genetics*, 15(3):149–162, 2014.

Veeramah, K. R., Wegmann, D., Woerner, A., Mendez, F. L., Watkins, J. C., Destro-Bisol, G., Soodyall, H., Louie, L. and Hammer, M. F. An early divergence of khoesan ancestors from those of other modern humans is supported by an abc-based analysis of autosomal resequencing data. *Molecular biology and evolution*, page msr212, 2011.

Veeramah, K. R., Woerner, A. E., Johnstone, L., Gut, I., Gut, M., Marques-Bonet, T., Carbone, L., Wall, J. D. and Hammer, M. F. Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. *Genetics*, 200(1):295–308, 2015.

Vilà, C., Savolainen, P., Maldonado, J. E., Amorim, I. R., Rice, J. E., Honeycutt, R. L., Crandall, K. A., Lundeberg, J. and Wayne, R. K. Multiple and ancient origins of the domestic dog. *Science*, 276(5319):1687–1689, 1997.

VonHoldt, B. M., Pollinger, J. P., Lohmueller, K. E., Han, E., Parker, H. G., Quignon, P., Degenhardt, J. D., Boyko, A. R., Earl, D. A., Auton, A., Reynolds, A., Bryc, K., Brisbin, A., Knowles, J. C., Mosher, D. S., Spady, T. C., Elkahloun, A., Geffen, E., Pilot, M., Jedrzejewski, W., Greco, C., Randi, E., Bannasch, D., Wilton, A., Shearman, J., Musiani, M., Cargill, M., Jones, P. G., Qian, Z., Huang, W., Ding, Z.-L., Zhang, Y.-p., Bustamante, C. D., Ostrander, E. A., Novembre, J. and Wayne, R. K. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature*, 464(7290):898–902, 2010.

Wakeley, J. and Wakeley, J. *Coalescent theory: an introduction*. 575: 519.2 WAK. 2009.

Walker, A., Smith, S., Smith, S. et al. Mitochondrial dna and human evolution. *Nature*, 325:1–5, 1987.

Wang, G.-D., Zhai, W., Yang, H.-C., Wang, L., Zhong, L., Liu, Y.-H., Fan, R.-X., Yin, T.-T., Zhu, C.-L., Poyarkov, A. D. et al. Out of southern east asia: the natural history of domestic dogs across the world. *Cell research*, 26(1):21–33, 2016.

Wegmann, D., Leuenberger, C. and Excoffier, L. Efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. *Genetics*, 182(4):1207–1218, 2009.

Wilton, P. R., Carmi, S. and Hobolth, A. The smc' is a highly accurate approximation to the ancestral recombination graph. *Genetics*, 200(1):343–355, 2015.

CHAPTER II

Resolving haplotypes using fosmid pool sequencing

2.1 Introduction

DNA resequencing technologies have made it possible to identify genetic variation across thousands of individuals. However, most resequencing studies of this kind are "phase-insensitive", providing a mixed readout of diploid genomes that neglects the haplotype configuration of two homologous chromosomes. Haplotype information is key to understanding the relationships between genetic variation and phenotype, such as how cis-acting eQTL affect gene expression and whether combinations of heterozygous variants lead to additional combined phenotypic effects (reviewed in Tewhey et al. (2011)). Haplotypes are also informative for inferring genetic ancestry (Lawson et al., 2012; Brisbin et al., 2012; Sohn et al., 2012; Price et al., 2009) and reconstructing population history (Harris and Nielsen, 2013; Palamara et al., 2012; Hellenthal et al., 2014). Recent studies have also identified haplotypes of Neanderthal ancestry and examined how archaic introgression shaped our genomes (Sankararaman et al., 2014; Prüfer et al., 2014).

Haplotypes are often inferred by statistical methods that utilize population genotype data to model the haplotype pairs of an individual as an imperfect mosaic of other haplotypes (Browning and Browning, 2011). However, this approach is more applicable to common SNPs rather than rare

The work in Chapter II is published in 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68-74. (Supplemental Section 6.3). Fosmid pool libraries were made by Elzbieta Sliwerskas and Sarah Emery or individual-specific variants. Trio-based phasing is more accurate but sometimes unavailable and is uncertain for the phase of variants when all individuals are heterozygous. Several experimental phasing methods are now available. Although full genome sequencing of individual haploid cells, such as sperm, is now possible (Wang et al., 2012; Kirkness et al., 2013), most approaches obtain sequence information from individual haplotypes by physically separating DNA fragments (Dear and Cook, 1989). These approaches include use of clone end-sequence pairs (Lippert et al., 2002; Levy et al., 2007; Kim et al., 2007), analysis of clone pools (Burgtorf et al., 2003; Kitzman et al., 2011; Suk et al., 2011), or library creation from dilute pools of single molecules (Kaper et al., 2013; Peters et al., 2012; Kuleshov et al., 2014). Approaches based on other concepts, such as the spatial structure of chromosomes in the nucleus (Selvaraj et al., 2013), are also being developed.

Here, we construct long-range haplotypes by sequencing fosmid pools where each fosmid represents 35kb of haplotype-specific sequence (Kitzman et al., 2011). Heterozygous variant positions within overlapping clone fragments were then used to assemble contiguous haplotypes. Fosmid-based haplotyping can usually achieve an N50 block size greater than 300kb, depending on the density of heterozygous SNPs and the number of clones sequenced (Duitama et al., 2012). Additional information from SNPs phased by trio transmission can further link blocks together, producing near-to-complete haplotypes. The haplotype of a Gujarati Indian Individual (NA20847; GIH) was first resolved by using this method (Kitzman et al., 2011), followed by NA12878, a HapMap trio child from the CEU population (Duitama et al., 2012) and sample “Max Planck One” with European ancestry (Suk et al., 2011). When trio data is not available, LD patterns from population-level phasing can be used to phase local blocks relative to each other, thus constructing global haplotypes (Kuleshov et al., 2014).

In this chapter, we construct physically phased genomes of five individuals from diverse African populations (including Yoruba, Esan, Gambia, Massai and Mende). We reanalyze fosmid sequencing data for Gujarati, San and Mbuti populations. We assess the ability to correctly assemble SNP haplotypes using fosmid pool sequencing and compare the resulting data with statistically phased haplotypes.

2.2 Methods

2.2.1 Fosmid pool sequencing

Genomic DNA for fosmid library construction of samples NA19240, HG03428, HG02799, and HG03108 was purchased (Coriell, Camden, NJ) or isolated from cell lines (Coriell) using Gentra Puregene Blood Kit (Qiagen, Valencia, CA). Aliquots of 10 ug of DNA were sheared in 120 ul volumes on a Digilab Hydroshear for 60 cycles at a speed code of 16 or 20 cycles at a speed code of 20. Sheared DNA was loaded and ran on a pulse field gel at 200V for 26 hours with 0.5s-15s switching or a BioRad CHEF DR III (Hercules, CA) at 6 V/cm for 16 hours with 1-6s switching. DNA from 25kb-45kb was cut out of the gel and isolated by electroelution for 12 hours at 120 V or 3 hours at 150 V. After electroelution, DNA was isolated with Ampure XP beads, end-repaired with the Epicentre End-It kit and ligated to the Epicentre pCC1Fos fosmid arms. The resulting ligation was packaged and transfected into the Phage T-1 Resistant EPI300-T1 E. coli plating strain (Catalog Number CCFOS110). One hour after transfection, the resulting cells were split into the appropriate volumes to give pools of 1,500-3,000 cells per pool. Barcoded libraries for sequencing were constructed from mini-prepped DNA obtained from each pool using either the Illumina Nextera or Bioo NEXTFlex protocols. In addition, we created high-coverage ($>$ 20x) coverage of standard Illumina whole genome sequencing.

2.2.2 Haplotype reconstruction

For whole genome sequencing data, paired end reads were aligned to the reference genome assembly (GRCh37, with the pseudoautosomal regions of the Y chromosome masked to ‘N’) using BWA v0.5.9-r16 (Li and Durbin, 2009). PCR duplicates were removed by Picard v1.62. Reads in regions with known indels were locally realigned and base quality scores were recalibrated using

GATK (McKenna et al., 2010). We generated GVCF files (Genomic VCF) with a record for every position in the genome using GATK HaplotypeCaller v3.2-2. Variants were called using GenotypeGVCFs and filtered by applying Variant Quality Score Recalibration(VQSR) implemented in GATK to select a SNP set that included 99% of sites that intersect with the HapMap, 1000 Genomes and dbSNP training set (Table 2.5). We define callable regions as sites that are within half and 2 times of the average coverage and with genotype and mapping quality larger than 20. We kept variants that either passed VQSR filtering or were present in 1000 Genomes Phase I reference panel, which were the starting point for subsequent haplotype phasing. We applied the same procedure of variants calling and filtering for NA12878, NA20847, HGDP00456, HGDP01029, using available whole genome sequencing data.

Reads from each fosmid pool were mapped to reference assembly including the human genome (GRCh37/hg19), Epstein Barr virus, the E. coli genome and fosmid vector backbone using BWA v0.5.9-r16 (Li and Durbin, 2009). Candidate fosmid clones were identified by computing read-depth in 1k bp windows for each clone pool and merging consecutive windows allowing a maximum gap of 3 windows. Reads where one end mapped to the fosmid vector backbone and another end mapped to human genome, called anchoring reads, were used to better assign clone breakpoints. Observing anchoring reads in the middle of consecutive windows identified overlapping clones. Overlapping clones were excluded from downstream analysis. Potential clones were further filtered by length (10kb to 50kb) and read depth (above 0.25). Each clone pool was separately genotyped at heterozygous SNPs called using whole genome shotgun sequencing data. Clones covering one or more heterozygous SNP positions identified in the whole genome sequencing data were used to resolve haplotype in next stage. Small proportions of clones (8.1% for NA19240) were genotyped as heterozygous and were excluded from further analysis.

We applied ReFHap (Duitama et al., 2010), an efficient algorithm for Single Individual Haplotyping. This algorithm considers clone fragments that contain at least one heterozygous SNP and seeks to bipartite clone fragments into two sets that maximize the difference between them. The algorithm first builds a graph where fragments are linked upon sharing positions and a score is assigned on the edge indicating how different two fragments are. RefHap then applies a heuristic algorithm to find the bipartition maximizing the cut, which can also be formulated into an NP-hard Max-CUT problem. Finally, the algorithm generates a consensus haplotype within one partition and flips every allele to get the other haplotype. However, if a site is observed equally in each partition, it will remain undecided resulting in gaps in the haplotype. We calculated the minimum error correction (MEC) value, namely the number of errors to make when comparing output haplotype with each individual clone fragment. A smaller MEC value locally indicates greater support for the consensus haplotype from RefHap.

For NA19240, HG02799, HG03108 and NA21302, we used phase-determined SNPs from trio genotyping available from HapMap and AffyMetrix to guide paternal and maternal allele assignment within blocks. We determined paternal and maternal allele (transimitted or non-transmitted for individuals that are not child in the trio) based on the majority of phased SNP assignments, found switch errors and corrected part of switch errors only if the increase in MEC value is less than 50 after correction. For NA20847, HG03428, HGDP01029, HGDP00456, phase-determined SNPs from trio data are unavailable. For these samples, we applied Prism (Kuleshov et al., 2014), a statistical phasing algorithm designed to assemble local blocks into long global haplotype contigs. This method is an extension to the Li and Stephens's model (Li and Stephens, 2003) that utilizes a reference panel of phased haplotypes and a genetic map of the genome but added an additional parameter of the phase of each block into the hidden Markov model to enforce the locally phased structure at the global phasing level. We grouped local blocks into windows with size smaller than 1Mbp and with at least 2 local blocks. Each window overlapped by 1 local block, which was used to link adjacent window together.

For sample NA12878, we used the phased SNP haplotypes constructed by fosmid pool sequencing from a previous study (Duitama et al., 2012). We obtained callable regions and high-confidence snp call sets from the sequencing results of 1000 Genomes Pilot Project (Consortium et al., 2010a) to construct full haplotypes.

2.3 Results

2.3.1 Haplotype reconstruction

We performed fosmid pool sequencing on individuals from five African populations (NA19240 from Yoruba, HG02799 from Gambia, HG03108 from Esan, HG03428 from Mende, NA21302 from Maasai, Table 2.1). In total across all pools, each genome was covered by an average of 6-17 clones and a median sequence coverage ranging 16.9-24.8x (Table 2.4). We identified high confidence variant calls and generated a callability mask for each individual based on conventional whole genome sequencing, which served as a starting point for haplotype construction. We utilized the ReFHap algorithm (Duitama et al., 2010), previously demonstrated to have superior performance on this type of data, and obtained haplotype-resolved blocks. We additionally analyzed available fosmid pool sequencing data for HGDP01029 from San, HGDP00456 from Mbuti (Prüfer et al., 2014) and NA20847 for Gujarati (Kitzman et al., 2011) using the same pipeline, and obtained phased haplotypes for NA12878 from a previous study based on AB SoliD Sequencing of fosmid pools (Duitama et al., 2012). The effect of increased clone counts on phased block size is dramatic: when doubling the number of fosmid clones, the N50 of phased blocks tripled, achieving over 1Mbp for four of the African samples (Figure 2.1). Over 98% of heterozygous SNPs were phased into haplotype-resolved blocks for African individuals except NA19240, for which 93% of SNPs were phased, HGDP01029, 87%, and HGDP00456, 79% (Figure 2.1, Table 2.2).

Although SNPs within each block are phased, the relationships between blocks cannot be directly established due to the absence of linking fosmid clones. We utilized two approaches to overcome this limitation. For samples that are members of genotyped trios (Consortium et al., 2010b), we utilized SNP transmission patterns to link adjacent blocks together producing near-to-complete haplotypes, encompassing over 97% of total heterozygous SNPs for HG02799, HG03108, NA21302, and 92.7% for NA19240. Comparison with deterministically phased SNPs identified potential switch errors due to insufficient clone support within our inferred haplotypes, which we corrected prior to subsequent analysis (Table 2.2). Switch errors are counted as the number of switches required to obtain the same haplotype phase when comparing the inferred haplotype phase with true haplotype phase (Browning and Browning, 2011). Examination of the errors indicates that most of the switch errors are due to insufficient clone support when linking variants together as ReFHap will assemble variants even when there is a single clone overlapping two variants (Figure 2.2). We find 99.66% concordance between the fosmid-phased SNPs for NA19240 and heterozygous SNPs phased based on transmission from this sequenced trio (Consortium et al., 2010a). We further compared our phased haplotypes for NA19240 to the sequence of 33 fosmid clones from the same individual (Kidd et al., 2008), observing a differences at 5 of the 1,013 heterozygous sites (0.5%) encompassed by the 33 clones (Table 2.6). In total, the aligned clones encompass 1,102,213 bp excluding alignment gaps, and have 51 single nucleotide differences in comparison with our data. If we assume that all of these differences are errors in our inferred sequences, this suggests that our haplotypes have an overall sequence error rate of less than 0.005% or a Phred24 quality greater than Q40.

For individuals HG03428 (MSL), NA20847 (GIH), HGDP01029 (San), and HGDP00456 (Mbuti), trio data is unavailable. For these samples, we assigned 80%-98% of SNPs to a parental allele using Prism (Kuleshov et al., 2014), a statistical phasing algorithm designed to assemble short local blocks into longer global haplotype contigs. To evaluate how well Prism performs in this context of large haplotype-block assignment, we applied Prism to NA19240 and HG02799 and compared the assignment of local blocks with our assignment based on trio phase-determined SNPs. For NA19240, 6575 out of 13591 blocks (47.6%) were assigned differently, affecting 45.88% of total heterozygous SNPs. For HG02799, 1214 out of 2810 blocks (43.2%) were assigned differently, affecting 41.82% of total heterozygous SNPs. This results in mean inter-switch distance 2335 kbp and mean incorrectly phased haplotype length of 1967 kbp, with a 0.03% switch error rate.

2.3.2 Comparison with statistical phasing

We compared 1000 Genomes phase3 haplotypes (Consortium et al., 2015) with haplotypes obtained using fosmid pool sequencing. Switch error is an inconsistency between an assembled haplotype and the real haplotype between two contiguous variants. Switch error rate is switch error normalized by number of variants for comparison. Overall, the haplotype concordance between 1000 Genomes phase 3 haplotypes and physically phased haplotypes are quite high, 96.41% in average, with switch error rate around 0.56% and mean inter-switch distance 191.7 kbp. Among switch errors, 85.7% are flip errors, namely individual alleles appearing on the opposite haplotype, indicating overall high quality of long range haplotypes, a result of a multi-stage phasing process that utilized a haplotype scaffold of trio-genotyped SNPs.

We also quantified the performance of statistical phasing algorithms compared with physically phased haplotypes. We applied ShapeIT (Delaneau et al., 2008) with the 1000 Genomes Phase1 reference panel (Consortium et al., 2012) (contained 1092 individuals, 14 populations) and Phase3 reference panel (Consortium et al., 2015) (contained 2504 individuals, 27 populations) separately to statistically phase each individual and calculated the haplotype concordance, switch error rate, flip error rate, mean inter-switch distance and mean length of incorrectly phased haplotype relative to the fosmid pool-derived haplotypes (Figure 2.3, Table 2.3). For haplotypes phased using 1000 Genomes Phase1 reference panel, the average switch error rate is 2.52%, half of which are flip errors, namely single alleles appearing on the opposite haplotype. The mean length of incorrectly phased haplotypes is 49.7 kbp, with a mean inter-switch distance of 62.5 kbp and an overall haplotype concordance is 62.57%. For HG03428, NA20847, HGDP01029 and HDP00456, since we statistically assigned and assembled local blocks into long global haplotypes, the comparisons of haplotypes are restrained to within blocks. This accounts for the relative high haplotype concordance and shorter length of incorrectly phased haplotype for those individuals. Compared with 1000 Genomes Phase1 reference panel, haplotypes phased by using Phase3 reference panel have higher concordance rate (72.79%), longer mean length of incorrectly phased haplotype (108.0 kbp) and mean inter-switch distance (184.2 kbp), but similar level of switch error rate (2.04%) and flip error rate (1.12%).

2.4 Discussion

Haplotype information is essential for population genetics analysis, such as ancestry mapping, population structure inference and detection of signals of natural selection. However, statistical phasing generally produce switch errors that may affect population genetics analysis listed above. Therefore, it's important to quantify phasing accuracy and how it affects downstream analysis. We found that 1000 Genomes Phase3 generated high quality haplotypes, especially a result of a multi-stage phasing process that utilized a haplotype scaffold of trio-genotyped SNPs (Consortium et al., 2015). However, when applying statistical phasing algorithm on single individual using existing reference panel, the average switch error rate is still around 2%, with half of them flip errors, indicating the potential space for improvement.

When constructing global haplotypes for individuals without trio phasing data available, we applied Prism to statistically link blocks together. Prism was designed to link much shorter phased segments into longer blocks (Kuleshov et al., 2014). The performance of Prism was not fully evaluated. When applied to our phased haplotype blocks, we found that around 40% of blocks were assigned wrongly compared to assignment based on trio data, resulting in switch errors every 2 Mbp. We also compared the haplotypes of NA12878 phased by Kuleshov et al. (2014) with the haplotypes phased by fosmid pool sequencing and found only 54.4% concordance, 0.47% switch error rate, 0.15% flip error rate and mean inter-switch distance 297k, mean length of incorrectly phased haplotypes 249k. This suggests that dilution haplotyping effectively produces locally phased haplotypes and the globally constructed haplotypes using Prism produce switch errors similar to all other statistical phasing algorithms. However, such combined method produces less switch errors and longer incorrectly phased haplotypes compared to using statistical phasing algorithm alone.

Figures

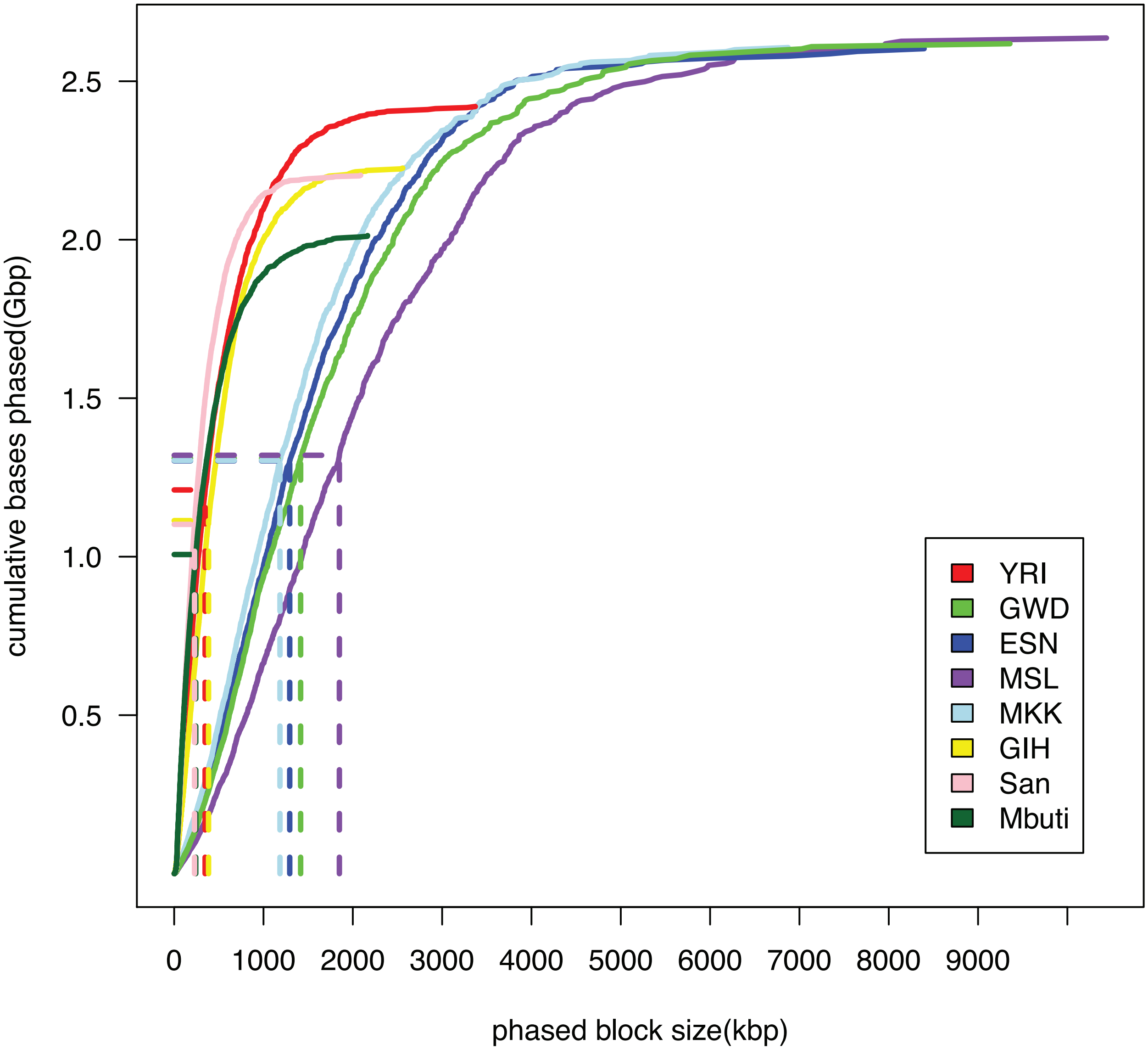

Figure 2.1: **Haplotype assembly results.** The relationship of block size and the cumulative length of haplotype assembly. Dashed lines correspond to half of the assembled sequence length and N50 of phased block size.

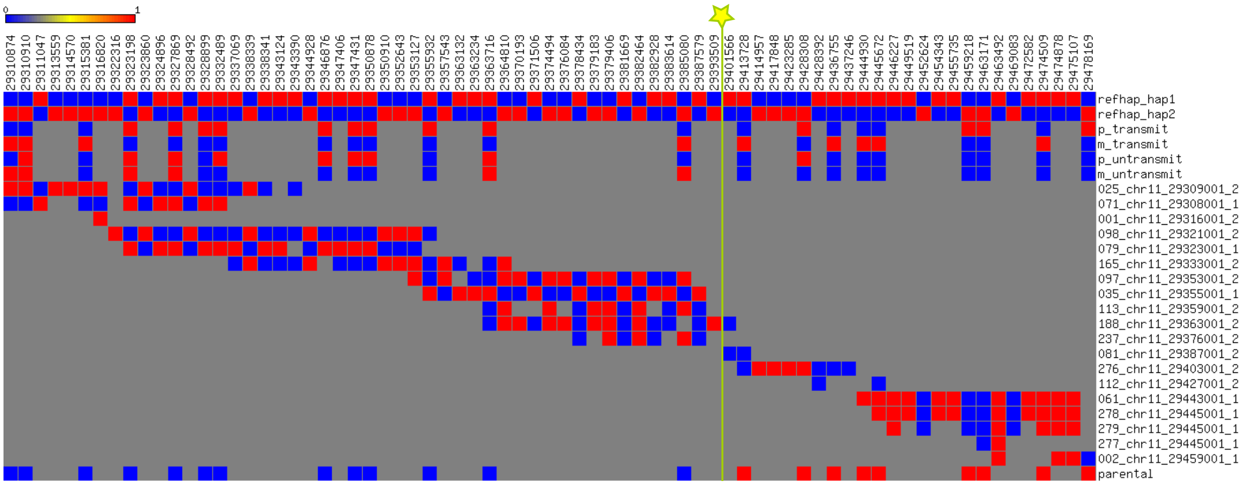

Figure 2.2: **Illustration of ReFHap's phasing result and a switch error.** Each column corresponds to a SNP position, with blue indicating the reference allele and red the alternative. The first two rows are the haplotype prediction by ReFHap, followed by four rows showing HapMap phase based on trio transmission. This is followed by 12 rows depicting clone genotypes. The last row indicates the parental allele assigned for RefHap haplotype based on HapMap phasing. In the last row, blue indicates paternal allele and red indicates maternal allele. The line with a star shows where the switch error occurred.

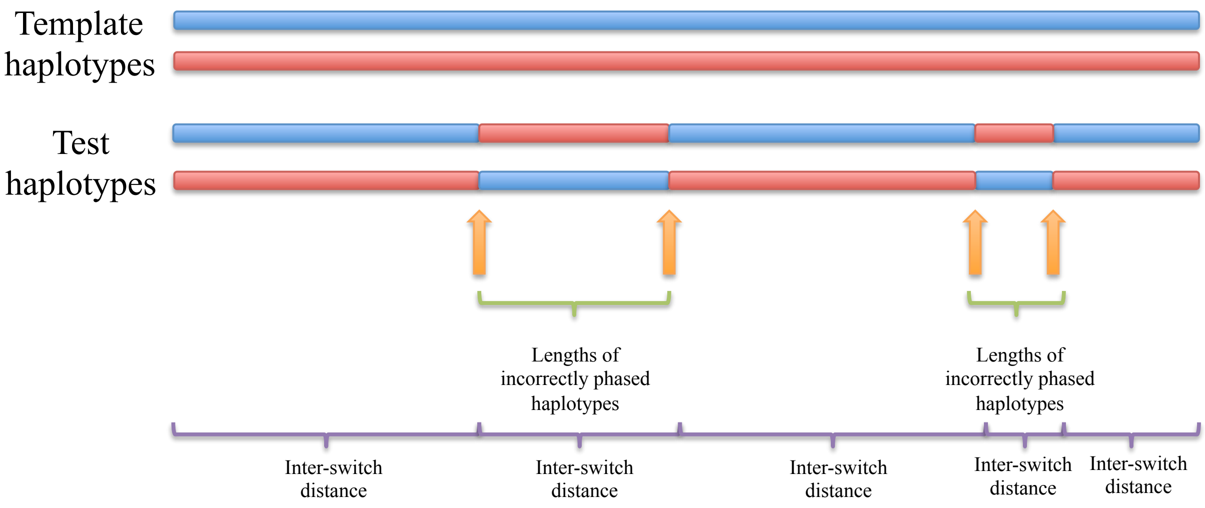

Figure 2.3: **Illustration of the metrics used to quantify phasing errors.** We illustrate switch error(green bracket), inter-switch distance(purple bracket) and length of incorrectly phased haplotypes (green bracket) when comparing test haplotypes with template haplotypes.

Tables

Table 2.1: **Summary of population geographic information and presence in HapMap or 1000 Genomes Project.**

Population	Short code	Sample ID	Geography	HapMap	1000 Genomes
Yoruba	YRI	NA19240	Yoruba in Ibadan, Nigeria	Yes	Yes, Phase1
Mende	MSL	HG03428	Mende in Sierra Leone	No	Yes, Phase3
Gambian	GWD	HG02799	Gambian in Western Division, The Gambia	No	Yes, Phase3
Esan	ESN	HG03108	Esan in Nigeria	No	Yes, Phase3
Massai	MKK	NA21302	Maasai in Kinyawa, Kenya	Yes, HapMap	No
Gujarati	GIH	NA20847	Gujarati Indians in Houston, Texas	Yes, HapMap3	No
San	San	HGDP01029	Indigenous hunter-gatherer people of Namibia	No	No
Mubti	Mbuti	HGDP00456	Indigenous pygmy groups in the Congo region of Africa	No	No

Table 2.2: **Phasing statistics from fosmid pool sequencing.** We resolved haplotypes using fosmid pool sequencing. MEC is the number of entries to correct when resolving haplotypes. Switch errors are counted as the number of switches required to obtain the same haplotype phase when comparing inferred haplotype phase with true haplotype phase. Switch error rate is switch error normalized by number of variants for comparison. For samples labeled with *, we applied Prism to link adjacent block together.

Population	Sample	#clones after filter	#blocks	N50 (kbp)	MEC value	#SNPs to be phased	% phased SNPs within blocks	# blocks assigned parental allele	% SNPs assigned parental allele	switch error	switch error corrected	switch error rate
YRI	NA19240	521,783	16,334	347	37,143	2,588,454	92.94%	15,171	92.74%	586	421	0.09%
GWD	HG02799	1,141,020	5,236	1416	82,387	2,780,269	99.00%	3,041	98.38%	327	146	0.16%
ESN	HG03108	1,058,027	5,416	1294	77,499	2,756,725	99.07%	3,323	98.45%	258	115	0.13%
MKK	NA21302	892,863	6,097	1416	175,935	2,736,727	98.60%	3,751	97.96%	336	265	0.10%
MSL	HG03428*	1,424,234	4,390	1849	167,294	2,775,099	99.30%	3,549	97.90%	-	-	-
GIH	NA20847*	571,419	16,838	385	44,870	1,680,704	93.37%	13,319	90.98%	-	-	-
San	HGDP01029*	358,759	17,695	228	27,712	2,623,001	87.10%	16,516	89.15%	-	-	-
Mbuti	HGDP00456*	381,075	18,465	242	28,629	2,517,569	78.70%	17,385	81.72%	-	-	-

Table 2.3: **Comparison between statistical phasing.** We calculated haplotype concordance, switch error rate, flip error rate, mean inter-switch distance, mean length of incorrectly phased haplotype when comparing haplotypes resolved by fosmid pool sequencing with haplotypes statistically phased by either 1000 Genomes Phase1 or Phase3 reference panel. * indicates that trio data was unavailable to link blocks together and phasing comparison analysis was limited to comparisons within RefHap blocks.

Individual	Haplotype Concordance	Switch Error Rate	Flip Error Rate	Mean inter-switch distance(kbp)	Mean length of incorrectly phased haplotype(kbp)
fosmid phased haplotypes vs 1000 Genomes haplotypes					
NA19240	99.40%	0.34%	0.31%	2406.6	27.4
HG02799	98.44%	0.69%	0.64%	1010.1	29.6
HG03108	99.01%	0.50%	0.47%	1875.1	21.6
NA12878	98.59%	0.74%	0.62%	846.6	19.8
HG03428*	90.17%	0.55%	0.43%	149.3	70.0
NA20847*	92.89%	0.56%	0.39%	85.1	54.8
Average	**96.42%**	**0.56%**	**0.48%**	**1062.1**	**37.2**
fosmid phased haplotypes vs Shapeit phased haplotypes using 1000 Genomes Phase1 reference Panel					
NA19240	54.60%	1.33%	0.60%	84.6	69.6
HG02799	52.46%	1.84%	0.79%	52.2	43.3
HG03108	53.62%	1.05%	0.47%	94.1	78.8
NA12878	53.18%	0.87%	0.32%	170.0	144
NA21302	52.00%	2.32%	1.02%	43.6	37.6
HG03428*	70.01%	1.88%	0.95%	42.6	28.5
NA20847*	79.30%	1.83%	0.97%	46.5	29.5
HGDP01029*	69.83%	6.87%	3.50%	12.5	7.3
HGDP00456*	78.09%	4.68%	2.70%	16.1	8.8
Average	**62.57%**	**2.52%**	**1.26%**	**62.5**	**49.7**
fosmid phased haplotypes vs Shapeit phased haplotypes using 1000 Genomes Phase3 reference Panel					
NA19240	68.00%	0.33%	0.21%	480.5	293.6
HG02799	77.10%	0.63%	0.27%	296.5	124.4
HG03108	69.40%	0.42%	0.27%	346.5	208.5
NA12878	58.90%	0.67%	0.32%	264.4	204.4
NA21302	53.10%	2.44%	1.08%	41.2	32.9
HG03428*	89.70%	0.66%	0.50%	132.2	56.1
NA20847*	91.50%	1.00%	0.73%	70	36.9
HGDP01029*	69.97%	7.17%	3.77%	12	6.9
HGDP00456*	77.47%	5.08%	2.97%	14.9	8.1
Average	**72.79%**	**2.04%**	**1.12%**	**184.2**	**108.0**
fosmid phased haplotypes: assign parental alleles using trio data vs using Prism					
NA19240	54.12%	0.05%	0.00%	1242.6	1115.0
HG02799	58.18%	0.02%	0.00%	3427.3	2821.9
Average	**56.15%**	**0.03%**	**0.00%**	**2335.0**	**1968.5**

Table 2.4: **Summary of clone statistics of fosmid pool sequencing.**

Sample	Population	SRA Accession	# of fosmid pools	Mean fosmids per pool	Median insert length(kbp)	Median 1kbp read depth per clone	Fosmid clone coverage of genome	Median 1kb read depth of genome
NA19240	YRI	SRS628777	288	1825	34	1.95	5.8	17.9
NA20847	GIH	SRA026360	115	4969	39	1.10	7.5	14.4
HG03428	MSL	SRS722908	440	3237	37	1.03	17.0	24.8
HG02799	GWD	SRS722940	288	3962	36	0.95	14.0	18.1
HG03108	ESN	SRS722941	288	3674	35	1.20	12.7	19.2
NA21302	MKK	SRS722942	288	3100	35	1.34	10.2	16.9
HGDP01029	San	PMID:24352235	192	1868	36	-	4.6	-
HGDP00456	Mbuti	PMID:24352235	192	1984	37	-	5.0	-

Table 2.5: **Summary of variant calling for whole genome sequencing.**

Sample	Population	SRA Accession	coverage	# called sites	# heterozygous sites pre-filter	# heterozygous sites post-filter
NA19240	YRI	SRS628777, SRR496444, SRR496446	20	2,397,179,422	2,991,938	2,588,454
HG02799	GWD	SRS722940	26	2,582,596,514	3,247,760	2,780,269
HG03108	ESN	SRS722941	25	2,573,126,220	3,234,813	2,756,725
HG03428	MSL	SRS722908	26	2,548,867,098	3,372,407	2,775,099
NA21302	MKK	SRS722942	26	2,588,203,996	3,176,366	2,736,727
NA20847	GIH	SRP000806, SRP048601, SRR072881	24	2,013,696,898	1,791,428	1,636,344
NA12878	CEU	1000 Genomes High coverage bam	42	2,464,361,297	-	1,843,256
HGDP01029	San	PMID:24352235	34	2,473,030,394	3,140,371	2,623,001
HGDP00456	Mbuti	PMID:24352235	24	2,481,367,128	2,936,250	2,517,569

Table 2.6: **Comparison of fosmid-resolved haplotype for NA19240 with Sanger sequenced fosmid clones.** Comparison of our haplotypes with the sequence of 33 fosmid clones from the same individual that were previously sequenced using standard capillary sequencing (hap1 refers to paternal allele, hap2 refers to maternal allele).

clone_name	chr	pos1	pos2	strand	# het SNP	Het mismatch	All mismatch	Length of callable sites	error rate	haplotype assignment
AC203596	20	60332560	60367311	-	55	1	6	32288	0.018	hap2
AC208180	7	109124048	109164730	-	13	0	22	34982	0.000	hap1
AC203618	14	24625251	24665998	+	38	0	0	38494	0.000	hap1
AC203625	3	13175578	13207238	+	49	0	0	30273	0.000	hap2
AC203613	17	42388348	42422435	+	9	0	0	30723	0.000	hap2
AC209301	5	103498132	103533083	+	38	0	4	29924	0.000	hap2
AC211777	2	131597914	131644819	+	50	0	1	38297	0.000	hap2
AC207436	20	61771673	61805848	-	20	0	0	31984	0.000	hap2
AC203629	2	27532275	27572319	+	26	0	0	38173	0.000	hap1
AC203601	13	84295142	84330569	+	11	0	8	31632	0.000	hap1
AC214990	20	1835027	1870122	+	33	0	0	34441	0.000	hap2
AC203623	13	50529642	50571393	-	19	0	0	38426	0.000	hap2
AC209312	19	11076696	11117117	+	6	0	0	38301	0.000	hap1
AC204964	20	36206659	36240122	-	51	0	0	32552	0.000	hap2
AC203663	12	120664025	120704371	-	11	2	0	39060	0.182	hap1
AC203633	15	83719658	83761946	+	27	2	2	38954	0.074	hap2
AC207998	5	42518686	42548692	+	15	0	0	28690	0.000	hap1
AC204962	10	73293352	73327357	-	3	0	0	33625	0.000	hap2
AC203585	12	127683544	127718788	-	24	0	0	32465	0.000	hap1
AC207584	22	24202000	24237415	-	46	0	1	33114	0.000	hap1
AC203595	17	15129257	15154073	-	24	0	0	23449	0.000	hap1
AC204968	12	108020551	108054912	-	45	0	0	32583	0.000	hap1
AC214217	20	34745579	34780635	-	14	0	0	34239	0.000	hap1
AC203614	13	27271316	27306555	-	28	0	0	32255	0.000	hap1
AC226164	17	8937695	8978811	+	48	0	0	39232	0.000	hap2
AC203609	17	9512088	9545453	+	22	0	0	25789	0.000	hap2
AC213115	2	5679730	5713339	-	19	0	0	31868	0.000	hap1
AC210876	17	3897067	3932032	-	43	0	1	33305	0.000	hap2
AC215711	8	142371608	142412790	-	55	0	0	34903	0.000	hap2
AC207992	4	162086171	162119199	+	29	0	1	29492	0.000	hap2
AC204957	7	135193786	135227398	+	22	0	0	32251	0.000	hap2
AC209156	12	104071670	104106822	+	13	0	0	33732	0.000	hap1
AC208068	16	85067617	85102315	-	107	0	0	32717	0.000	hap2

Bibliography

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J. G. and Bustamante, C. D. Pcadmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human biology*, 84(4):343, 2012.

Browning, S. R. and Browning, B. L. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

Burgtorf, C., Kepper, P., Hoehe, M., Schmitt, C., Reinhardt, R., Lehrach, H. and Sauer, S. Clone-based systematic haplotyping (csh): a procedure for physical haplotyping of whole genomes. *Genome research*, 13(12):2717–2724, 2003.

Consortium, . G. P. et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010a.

Consortium, . G. P. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

Consortium, . G. P. et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

Consortium, I. H. . et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010b.

Dear, P. and Cook, P. Happy mapping: a propoaal for linkage mapping the human genome. *Nucleic acids research*, 17(17):6795–6807, 1989.

Delaneau, O., Coulonges, C. and Zagury, J.-F. Shape-it: new rapid and accurate algorithm for haplotype inference. *BMC bioinformatics*, 9(1):1, 2008.

Duitama, J., Huebsch, T., McEwen, G., Suk, E.-K. and Hoehe, M. R. Refhap: a reliable and fast algorithm for single individual haplotyping. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pages 160–169. ACM, 2010.

Duitama, J., McEwen, G. K., Huebsch, T., Palczewski, S., Schulz, S., Verstrepen, K., Suk, E.-K. and Hoehe, M. R. Fosmid-based whole genome haplotyping of a hapmap trio child: evaluation of single individual haplotyping techniques. *Nucleic acids research*, 40(5):2041–2053, 2012.

Harris, K. and Nielsen, R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet*, 9(6):e1003521, 2013.

Hellenthal, G., Busby, G. B., Band, G., Wilson, J. F., Capelli, C., Falush, D. and Myers, S. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, 2014.

Kaper, F., Swamy, S., Klotzle, B., Munchel, S., Cottrell, J., Bibikova, M., Chuang, H.-Y., Kruglyak, S., Ronaghi, M., Eberle, M. A. et al. Whole-genome haplotyping by dilution, amplification, and sequencing. *Proceedings of the National Academy of Sciences*, 110(14):5552–5557, 2013.

Kidd, J. M., Cheng, Z., Graves, T., Fulton, B., Wilson, R. K. and Eichler, E. E. Haplotype sorting using human fosmid clone end-sequence pairs. *Genome research*, 18(12):2016–2023, 2008.

Kim, J. H., Waterman, M. S. and Li, L. M. Diploid genome reconstruction of ciona intestinalis and comparative analysis with ciona savignyi. *Genome research*, 17(7):1101–1110, 2007.

Kirkness, E. F., Grindberg, R. V., Yee-Greenbaum, J., Marshall, C. R., Scherer, S. W., Lasken, R. S. and Venter, J. C. Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome research*, 23(5):826–832, 2013.

Kitzman, J. O., MacKenzie, A. P., Adey, A., Hiatt, J. B., Patwardhan, R. P., Sudmant, P. H., Ng,

S. B., Alkan, C., Qiu, R., Eichler, E. E. et al. Haplotype-resolved genome sequencing of a gujarati indian individual. *Nature biotechnology*, 29(1):59–63, 2011.

Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M. and Snyder, M. Whole-genome haplotyping using long reads and statistical methods. *Nature biotechnology*, 32(3):261, 2014.

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

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G. et al. The diploid genome sequence of an individual human. *PLoS Biol*, 5(10):e254, 2007.

Li, H. and Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

Li, N. and Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

Lippert, R., Schwartz, R., Lancia, G. and Istrail, S. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Briefings in bioinformatics*, 3(1):23–31, 2002.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

Palamara, P. F., Lencz, T., Darvasi, A. and Pe'er, I. Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, 91(5):809–822, 2012.

Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y. T., Haas, J. et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, 487(7406):190–195, 2012.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D. and Myers, S. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, 5(6):e1000519, 2009.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C. et al. The complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505(7481):43–49, 2014.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N. and Reich, D. The genomic landscape of neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–357, 2014.

Selvaraj, S., Dixon, J. R., Bansal, V. and Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature biotechnology*, 31(12):1111–1118, 2013.

Sohn, K.-A., Ghahramani, Z. and Xing, E. P. Robust estimation of local genetic ancestry in admixed populations using a nonparametric bayesian approach. *Genetics*, 191(4):1295–1308, 2012.

Suk, E.-K., McEwen, G. K., Duitama, J., Nowick, K., Schulz, S., Palczewski, S., Schreiber, S., Holloway, D. T., McLaughlin, S., Peckham, H. et al. A comprehensively molecular haplotype-resolved genome of a european individual. *Genome research*, 21(10):1672–1685, 2011.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. and Schork, N. J. The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215–223, 2011.

Wang, J., Fan, H. C., Behr, B. and Quake, S. R. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell*, 150(2):402–412, 2012.

CHAPTER III

Human separation history inference using physically phased genomes

3.1 Introduction

Haplotypes contain rich information about population history and are shaped by population size, natural selection, and recombination (Veeramah and Hammer, 2014; Schraiber and Akey, 2015). Due to historic recombination events there are hundreds of thousands of pairs of loci along a chromosome that have distinct histories. Recent methodological advances permit the estimation of a detailed population demographic history from a single or several whole genome sequences based on the distribution of coalescent times across the genome. For example, Li and Durbin (2011) developed the Pairwise Sequentially Markovian Coalescent model (PSMC) to reconstruct the distribution of the time since the most recent common ancestor (TMRCA) between the two alleles of an individual and infer population size changes over time. Typically, these TMRCA values are calculated using the two haploid genomes that compose the diploid genome of a single sample (Li and Durbin, 2011). When PSMC is applied to two haplotypes obtained from different populations, the inferred TMRCA distribution is informative about the timing of population splits since the time after which nearly no coalescence events occur is a good estimate for the population split time. One key question regarding human population history is the timing of population splits

The work in Chapter III is submitted as **Shiya Song**, Elzbieta Sliwerskas, Sarah Emery, Jeffrey M. Kidd. Modeling human population separation history using physically phased genomes. (*In review*) and the dynamics of separation between Africans and non-Africans, which has a great influence on modern genetic diversity. Li and Durbin (2011) paired X chromosomes from African and non-African males and suggested that the two groups remained as one population until 60-80 kyrs ago with substantial genetic exchange up until 20-40 kyrs ago (assuming a mutation rate of $2.5 \times 10^{-8)}$ bp per generation and 25 years as generation time, estimates which approximately double when assuming a mutation rate $1.25 \times 10^{-8)}$ bp per generation and 30 years as generation time (Schiffels and Durbin, 2014). Subsequently, PSMC applied to pseudo-diploid sequences was used to date the divergence time between non-human primate subspecies (Prado-Martinez et al., 2013). However, PSMC curves themselves provide only a qualitative measure of population separation and estimating split times is complicated by the presence of migration (Pritchard, 2011). MSMC (Schiffels and Durbin, 2014) extends PSMC to multiple individuals, focusing on the first coalescence event for any pair of haplotypes. With multiple haplotypes from different populations, MSMC calculates the ratio between cross-population and within-population coalescence rates, termed the 'relative cross coalescence rate', a value reflecting population separation history. Schiffels and Durbin (2014) applied MSMC on statistically phased genomes (two or four haplotypes per population) and suggested that African and non-African populations exhibited a slow, gradual separation beginning earlier than 200,000 years ago and lasting until about 40,000 years ago, while the median point of such divergence was around 60,000 – 80,000 years ago. The midpoint of the relative cross-coalescence decay curve has been used as an estimate of population separation time (Schiffels and Durbin, 2014; Pagani et al., 2015). Although useful, this approach does not generate parametric estimates for population history under standard models. As none of these methods to infer population separation history were applied on physically phased genomes, it is unclear how phasing errors and missing data affect this type of analysis.

In Chapter II, we constructed physically phased genomes of nine individuals from diverse African populations (including Yoruba, Esan, Gambia, Massai, Mende, San, Mbuti) and non-

African populations (CEU and Gujarati). In this chapter, we further assess the impact of phasing error on MSMC's estimates of population split times using physically phased genomes vs. statistically phased genomes. We extend the current PSMC method to model population splits. We apply an approximate Bayesian computation (ABC) method to obtain posterior estimates of split time and migration rate by fitting the inferred TMRCA distribution obtained from PSMC on pseudo-diploid genomes to a standard Isolation-with-Migration model. Additionally, we assess the sensitivity of existing methods to missing data and phasing errors from statistically phased haplotypes.

3.2 Methods

3.2.1 MSMC analysis

We applied the Multiple sequentially Markovian coalescent (MSMC) model on four haplotypes, two haplotypes per individual each population. We used 'fixedRecombination' and 'skipAmbigous' for inference of population separation. MSMC analysis yields inferred cross-population and within-population coalescence rates. We calculated the relative cross coalescence rate (RCCR) by dividing the cross-population coalescence rate by the average of within-population coalescence rate and plotted it as a function of time. We also applied MSMC on individual diploid genomes, which is very similar to PSMC, with subtle differences due to the underlying model SMC' (Marjoram and Wall, 2006) versus SMC (McVean and Cardin, 2005). In order to differentiate it from PSMC, we refer to such analysis as PSMC'.

3.2.2 PSMC on pseudo-diploid genome

Pairwise sequentially Markovian coalescent (PSMC) inference was performed as previously described (Li and Durbin, 2011). PSMC builds a HMM to infer the local TMRCA based on the local density of heterozygotes. In the model, hidden states are discretized TMRCA values, and transitions represents ancestral recombination events. On autosomal data, we use the default setting with Tmax=15, n=64, and pattern '1*4+25*2+1*4+1*6'. When applying PSMC on a pseudo-diploid genome, there are four possible configurations of the two haplotypes, namely hap1-hap1, hap1-hap2, hap2-hap1, hap2-hap2. We applied PSMC to each possible configuration and took the average of the estimates. We obtained the inferred TMRCA distribution directly from PSMC output, the fifth column representing the fraction of the genome that coalesced in an indicated TMRCA bin.

3.2.3 ABC analysis

We implemented an ABC framework to estimate split time and migration rate given the inferred TMRCA distribution from PSMC output. We computed the coalescence time density of two chromosomes based on the Isolation-With-Migration model (Wang and Hey, 2010; Hobolth et al., 2011) and integrated coalescence time density on the 64 time intervals in which PSMC is parameterized. We use chi-square statistics calculated between the observed TMRCA distribution obtained from PSMC output and the computed one as the distance between estimates in the ABC framework.

We formulate the IM model as a continuous time Markov chain (Wang and Hey, 2010; Hobolth et al., 2011). The rate matrix Q is given by:

$$Q = \begin{pmatrix} . & 2m_1 & 0 & 2/\theta_1 & 0 \\ m_2 & . & m_1 & 0 & 0 \\ 0 & 2m_2 & . & 0 & 2/\theta_2 \\ 0 & 0 & 0 & . & m_1 \\ 0 & 0 & 0 & m_2 & . \end{pmatrix} \tag{3.1}$$

where the states are s_{11}(both gene are in population 1), s_{12} (one gene is in population 1 and the other is in population 2), s_{22} (both gene are in population 2), s_1 (the genes have coalesced and the single gene is in population 1), s_2 (the genes have coalesced and the single gene is in population 1), and θ_1 and θ_2 is the scaled population sizes, and m_1 and m_2 are the migration rates. Transition from one deme to another deme is quantified by migration rate and transition to coalescence event is quantified by the rate of coalescence. The density of coalescence time can be calculated as follows (Hobolth et al., 2011):

for t<T:

$$f(t) = (e^{Qt})_{S_{12}S_{11}} \times \frac{2}{\theta_1} + (e^{Qt})_{S_{12}S_{22}} \times \frac{2}{\theta_2} \tag{3.2}$$

for t>T:

$$f(t) = [(e^{Qt})_{S_{12}S_{11}} + (e^{Qt})_{S_{12}S_{12}} + (e^{Qt})_{S_{12}S_{22}}] \times \frac{2}{\theta_a(t)} \times \exp(-\int_T^t \frac{2}{\theta_a(t')} dt') \tag{3.3}$$

where T is the split time and $\theta_a(t)$ the ancestral population size. We use the ancestral population size inferred from PSMC of the pseudo-diploid genome as the ancestral population size, and use the inferred population size of each diploid genome (from PSMC) as the population size for each population after the split. For African populations, we assume constant population size after the split. For non-African populations, we assume that the population experienced a bottleneck event after the split and experienced population growth beginning 40 kyrs ago. For our ABC framework, the parameters of interests are T (split time) and m (migration rate after the split). We assumed a uniform prior for the split time and time when migration ends, and a uniform prior on migration rate in log10 scale, and applied an ABC method based on sequential Monte Carlo (Toni et al., 2009) (SMC) to the parameter estimation, since it can be easily run in parallel and is more efficient than an ABC rejection sampler. We drew a pool of 5,000 candidate parameter values (called particles) from the prior distribution. Instead of setting the final stringent cut-off ϵ (if the distance between summary statistics are lower than ϵ, we accept it), we gradually lowered the tolerance $\epsilon_1 > \epsilon_2 > \epsilon_3 >> 0$, thus the distributions gradually evolve toward the target posterior. The first pool was generated by sampling from the prior distribution. The particles that were accepted using the first threshold ϵ_1 were sampled by their weights and perturbed to get new particles. As the tolerance threshold lowered to the final cut-off, we obtained the target posterior distribution. In each iteration, we choose the threshold ϵ such that 20% of particles are accepted, achieving N=1000 accepted particles. The perturbation kernels for all parameters are uniform, $K = \sigma U(-1, 1)$, with ϵ equal to 20% of the difference between maximum and minimum values. We perform three iterations and summarized the mean, median and 95% HPD confidence interval for each parameter. For simulations, we generated 100 30Mb sequences of two individuals representing African and European samples and having split times ranging from 60 kyrs to 150 kyrs ago, with subsequent migration until 30 kyrs ago using MaCS (Chen et al., 2009).

3.3 Results

3.3.1 Demography inference using PSMC'

We applied PSMC', similar to PSMC but using the SMC' framework to perform demographic inference on nine individuals from nine populations. We assumed a human mutation rate of 1.25×10^{-8} bp per generation and 30 years as generation time, although results can be easily rescaled for comparison with other estimates (Schiffels and Durbin, 2014). The PSMC' curves of nine individuals revealed that all populations shared the same two-fold increase of ancestral population size prior to 300 kyrs ago, after which the inferred population size of the African populations began to differentiate from non-African populations and they all gradually experienced a population size reduction (Figure 3.1), although we note that the simulations indicate that such shifts in PSMC curves may overestimate the timing of population size changes (Prüfer et al., 2014). Non-African populations experienced a ten-fold reduction of population size, but undergone a rapid population growth after 30 kyrs ago. African populations experienced 2-4 fold reduction of population size and remained roughly constant population size after 70 kyrs ago, among which San (a subgroup of KhoeSan) and Mbuti (indigenous pygmy group), populations that maintain hunter-gather lifestyles, have the least bottleneck effect compared to other populations. Such observations were equivalent to previous PSMC analysis on diploid genomes after adjusting for differences in assumed mutation rate (Schiffels and Durbin, 2014).

3.3.2 Split time inferred using MSMC and the effect of phasing error

The multiple sequentially Markovian coalescent (MSMC) model extends PSMC to multiple individuals. When applying MSMC on multiple individuals from two populations, the ratio of cross-population coalescence rate to the average of within-population coalescence rate, namely relative cross coalescence rate, is informative about population separation. As population began to separate, relative cross coalescence rate dropped from one to zero. We applied the multiple sequentially Markovian coalescent (MSMC) model on four haplotypes, two haplotypes per individual each population and plotted the relative cross coalescence rate as a function of time (Figure 3.5). We summarized the time range when relative cross coalescence rate dropped to 75% and 25% as putative time range of the population split event, with 50% as the median point of the split event (Figure 3.2). We noticed that the more ancient the split event, the wider the resulting time interval you would get, a similar pattern observed using simulated data (Figure 3.6). The divergence of San from other populations was the most ancient, with putative time range 50 kyrs to 300 kyrs ago and the median point of the split range from 130 kyrs to 200 kyrs ago. The median point of divergence of San from other African populations was around 130 years ago, and 160 kyrs ago with CEU population. The median point of divergence of Mbuti from other African populations except San was around 80 kyrs to 100 kyrs ago, and 120 kyrs ago with CEU population. The effect of phasing error on the separation events among San and other populations, Mbuti and other populations are dramatic, all shifted towards present time, with time difference ranging from 20 kyrs to 70 kyrs. The most severe effect is for the separation among San and Mbuti where the inferred split time was shifted 70 kyrs later. The separation between west African populations and CEU population ranged from 120 kyrs to 50 kyrs ago, the median point of which was around 70 kyrs to 80 kyrs ago with little difference when using statistical phased haplotypes. The separation between MKK (Massai) and CEU population ranged from 40 kyrs to 70 kyrs ago, the median point of which was around 50 kyrs ago. The separation between GIH and CEU population probably occurred around 30 kyrs to 40 kyrs ago and the split time interval was shifted 20 kyrs later when using statistical phased haplotypes. The separation between west African and MKK (Massai) occurred around 20 kyrs to 40 kyrs ago, while the separation among west African populations happened around 10 kyrs to 15 kyrs ago. For the separation among Africans, the relative cross coalescence rate curve inferred using haplotypes phased by ShapeIT compared with those phased by fosmid data are quite similar. Overall, statistically phased haplotypes show a more recent separation time and a narrower time span, particularly for comparisons involving San or Mbuti samples, emphasizing the importance of correctly phased haplotypes on split time inference using MSMC, especially for more ancient split.

3.3.3 An ABC method to infer population split time using PSMC on pseudo-diploid genomes

PSMC applied to pseudo-diploid samples also provides information on population separation history. If population splits are total and sudden, no coalescent events between populations will occur after their separation. Thus, when applying PSMC on a pseudo-diploid individual where one chromosome comes from one population and the second chromosome comes from another population, the time when the PSMC estimate of Ne goes to infinity provides an estimate for the population split time (Li and Durbin, 2011). However, the inferred PSMC curve usually increases in a step-wise manner, making it difficult to determine the exact time of the split event. Subsequent migration after the split is a further confounding factor (Pritchard, 2011).

To better interpret pseudo-diploid PSMC curves (Figure 3.3 and Figure 3.7), we implemented an ABC framework to estimate the population split time and migration rate given the TMRCA distribution inferred from the PSMC output. We compared the observed TMRCA distribution with the analytical distribution determined by an Isolation-With-Migration model (Wang and Hey, 2010;

Hobolth et al., 2011) with the indicated values for split time and post-separation migration and applied an ABC method based on sequential Monte Carlo (Toni et al., 2009) (also abbreviated as ABC-SMC) to estimate the target posterior distribution of each parameter. We tested this approach using simulated data with a split time ranging from 60 kyrs to 150 kyrs ago, with subsequent migration continuing until 30 kyrs ago (Figure 3.8). For each split-time, we considered three levels of symmetrical migration: 2×10^{-5}, 10×10^{-5}, 40×10^{-5}. For small levels of migration, the inferred split is quite accurate, with the mean value of the posterior distribution centered on the true value. However, for larger migration rates the inferred split-time tends to be smaller than the true value. This bias is exacerbated with subsequent iterations of ABC sampling. The magnitude of the inferred migration rate is reasonably accurate, as observed in the log10 scale.

An additional complication in the application of this method to real data is the treatment of unphased sites, which generally impact less than 10% of SNPs in each comparison (Table 3.1). Using our simulations, we evaluated three methods for processing unphased SNPs: 1) randomly assigning the phase, 2) marking unphased sites along with all homozygous segments ending in an unphased heterozygous site as missing data (as recommended for MSMC) (Schiffels and Durbin, 2014) and 3) marking only unphased SNPs as missing data. Even with 10% of unphased sites, the third method results in a PSMC curve similar to the original, while the first two methods give PSMC curves shifted to an earlier increased effective population size, which may result in an earlier inferred split time (Figure 3.9). We therefore applied the third method to unphased SNPs in our analysis.

We applied our ABC method to date the split-times among African and European populations (Table 3.3,Figure 3.4, Figure 3.10). We find that the San population separated from the other samples the earliest, around 120 kyrs to 140 kyrs ago, with subsequent migration rate around $10\ 15 \times 10^{-5}$ until 30-40 kyrs ago, an estimate that is more recent than that obtained from MSMC analysis (the median point of divergence using MSMC of San from other African populations was around 130 kyrs ago, and 160 kyrs ago with CEU population). The separation between west African and CEU populations occurred 70-80 kyrs ago with migration at a rate of $8-40\times10^{-5}$ until 30 kyrs ago, while Maasai separated from the CEU population around 50 kyrs ago with a greater amount of gene flow until present, with migration rate on the magnitude of 10^{-3}. The separation between west African and MKK population occurred around 36 kyrs to 40 kyrs ago, also with a great amount of gene flow until present, with migration rate on magnitude of 10^{-3}. The separation between CEU and GIH occurred around 36 kyrs to 38 kyrs ago, with ongoing migration on the magnitude of 10^{-3} until present. Comparisons with statistically-phased data suggest that the impact of phasing error on our PSMC-ABC method is less dramatic than for MSMC analysis, however when using haplotypes phased by SHAPEIT, a larger proportion of the genome coalesced 50,000 years ago than when fosmid-phased haplotypes are used (Figure 3.11, Figure 3.12). This may result in larger amounts of inferred gene flow when using statistically phased data.

3.4 Discussion

The utility of phase-resolved genome sequence data in the interpretation of variants impacting gene expression, transcription factor binding, human disease, and genome assembly has motivated the development of multiple approaches for determining phase. Here, we focus on samples phased using fosmid-based dilution haplotyping, and analyze a diverse set of eight phase-resolved human genomes. As expected, we find that phase results improve with increasing number of sequenced clones. We also demonstrate that statistical phasing performs well using existing reference panels, particularly when the panel captures population variation form the studied individuals. Nonetheless, the resulting phase-errors are sufficient to impact inference of population history using the MSMC model. We find that the statistically phased haplotypes show a more recent inferred population split time, perhaps due to phasing bias that make haplotypes appear more similar than they truly are. This effect is particularly noticeable for comparisons involving more deeply diverged population samples that are not well-phased using existing reference panels.

Existing PSMC and MSMC approaches represent important methodological advances and have had a clear impact on the inference of population history using individual genome sequences. However, these approaches provide only a qualitative sense of population separation history. Here, we describe the fitting of a standard Isolation with Migration model to cross-population TMRCA distributions inferred from PSMC. This allows the acquisition of parameter estimates under standard models widely utilized in population genetic inference. However, as expected, multiple combination of split time and migration rate are sometimes indistinguishable, highlighting the difficulty of inferring split times with the presence of migration (Pritchard, 2011). This is partly due to the limitations of discretizing time and the poor resolution for recent history when given two haplotypes. Additionally, we find very high levels of migration for recent population splits (MKK and CEU, GIH and CEU, YRI and MKK), values which might be over-estimated because of the high uncertainty for estimates of recent population history.

The split times inferred using our ABC method are generally concordant with the time when relative cross-coalescence rate dropped to 50% as inferred using MSMC, however our method provides a narrower range while quantifying the level of subsequent migration (Table 3.3, Figure 3.2). Utilizing this approach with fully phased haplotypes from nine populations, we provide additional estimates of key population separation in human population history. With fully phased haplotypes from nine populations, we provide a picture of human separation history within the last 200,000 years. The separation among hunter-gather San population with all other populations happened the earliest. Gronau et al. (2011) applied Bayesian coalescence based model (G-PhoCS) on 37,574 1kb ‘neutral loci’, and revealed that the separation between San and other populations happened around 218,000 years ago when adjusted to our mutation rate and generation time. Veeramah et al. (2011) applied Approximate Bayesian Computation method on 40 nongenic 2kb regions, and suggested San ancestors diverged around 220,000 years ago with an interval between 104,000 to 374,000 years ago. Excoffier et al. (2013) applied a simulation-based framework to infer demographic parameters from the site frequency spectrum, and inferred that San ancestors diverged around 156,000 years ago with an interval between 60,000 to 230,000 years ago. From our MSMC results, we inferred that San ancestors diverged with other African populations around 130,000-140,000 years ago when we took the time when the inferred relative cross coalescence rate dropped to 0.5 and the potential range was around 60,000 to 200,000 years ago. We tried not to over-interpret the relative cross coalescence rate curve because the curve doesn't reach 1 at 300,000 years ago for the split between African and non-Africans. From our modified PSMC approach, San diverged around 120,000 to 140,000 years ago. For the separation among west African and non-African populations, Gronau et al. (2011) inferred that west African diverged around 79,000 years ago, Harris and Nielsen (2013) using IBS sharing around 85,600 years ago, and Gravel et al using allele frequency spectrum around 115,500 years ago. Both MSMC and modified PSMC approach estimated that that west African diverged around 70,000 to 80,000 years ago. For the separation among European and Asian populations, Gronau et al. (2011) inferred that the split was around 60,000 years ago, and Gravel (2012) around 45,200 years ago. Both MSMC and modified PSMC approach pointed to around 40,000 years ago. Overall, our estimates are broadly consistent with other contemporary methods (Table 3.2) and our estimates reconciles the timing of the most recent common ancestor of African and non-African mitochondrial DNA, around 78,300 years ago (62,400-94,900 years ago) and the timing of the MRCA for all modern humans at 157,000 years ago (134,000-188,000 years ago) (Fu et al., 2013).

Similar to previous results (Schiffels and Durbin, 2014), the separation history between CEU and MKK populations was different from that observed between CEU and LWK (Luhya, another east African population). Two pulses of admixture have been estimated in the ancestors of the MKK, occurring 8 and 88 generations ago (Pagani et al., 2015; Pickrell et al., 2014). Since the impact of long segments of shared ancestry due to recent admixture is unclear, we masked out regions of recent European ancestry in our MKK sample using RFMix (Maples et al., 2013) (Figure 3.13) and found that the MSMC curves are not altered when recent segments of European ancestry are masked (Figure 3.14). Although such ancestral masking becomes increasingly imperfect for older admixture events, this suggests that long segments of shared ancestry due to recent admixture do not explain the latter divergence of Massai population compared to other African populations and supports a more complex ancient history for the Massai.

When constructing global haplotypes for individuals without trio phasing data available, we applied Prism to statistically link blocks together. Prism was designed to link much shorter phased segments into longer blocks. When applied to our phased haplotype blocks, we found that around 40% of blocks were assigned incorrectly, resulting in switch errors every 2 Mbp. However, we found very similar MSMC curve using global haplotypes constructed by Prism with those constructed with trio phasing data (Figure 3.15), indicating long switch errors have little effect on such inference. This is reassuring since we are using Prism to construct global haplotypes for four individuals; but, the inferred split times involving the San and Mbuti populations are still likely underestimated.

Our results indicate that the separation of the studied human populations was a gradual event, with substantial genetic exchange continuing after an initial split, a finding consistent with hypotheses of long-standing ancient population structure in Africa (reviewed in Harding and McVean (2004); Henn et al. (2012)). We provide a comparison of PSMC and MSMC based methods with other contemporary methods on inferring population separation history and our results emphasize the importance of accurately phased haplotypes on MSMC analyses, especially for more ancient splits.

Figures

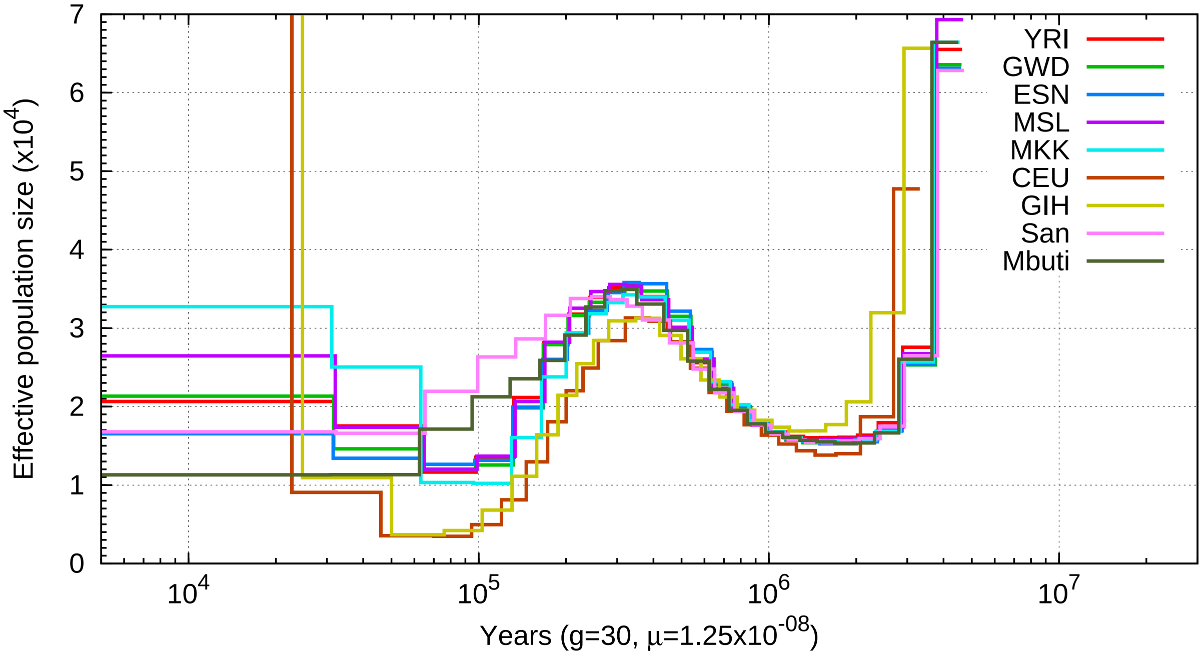

Figure 3.1: **PSMC' inferred population history**. Population sizes inferred from the autosomes of nine individuals from nine populations are shown.

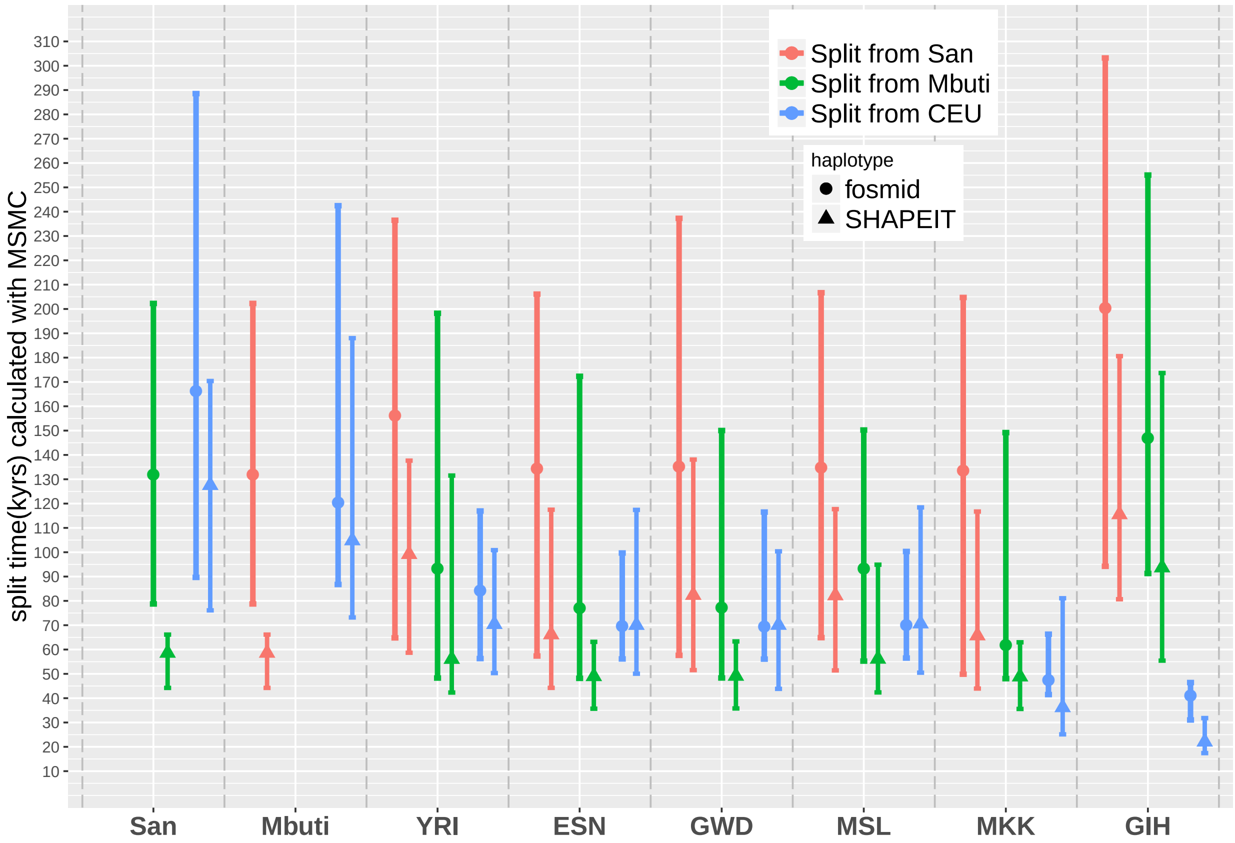

Figure 3.2: **MSMC inferred split times.** Circles or triangles represent the time when the relative cross-coalescence rate dropped to 50%, with lines representing the time when relative cross-coalescence rate reached 25% and 75%. Inferred split times were inferred using haplotypes phased by the fosmid pools approach (circle) or SHAPEIT (triangle).

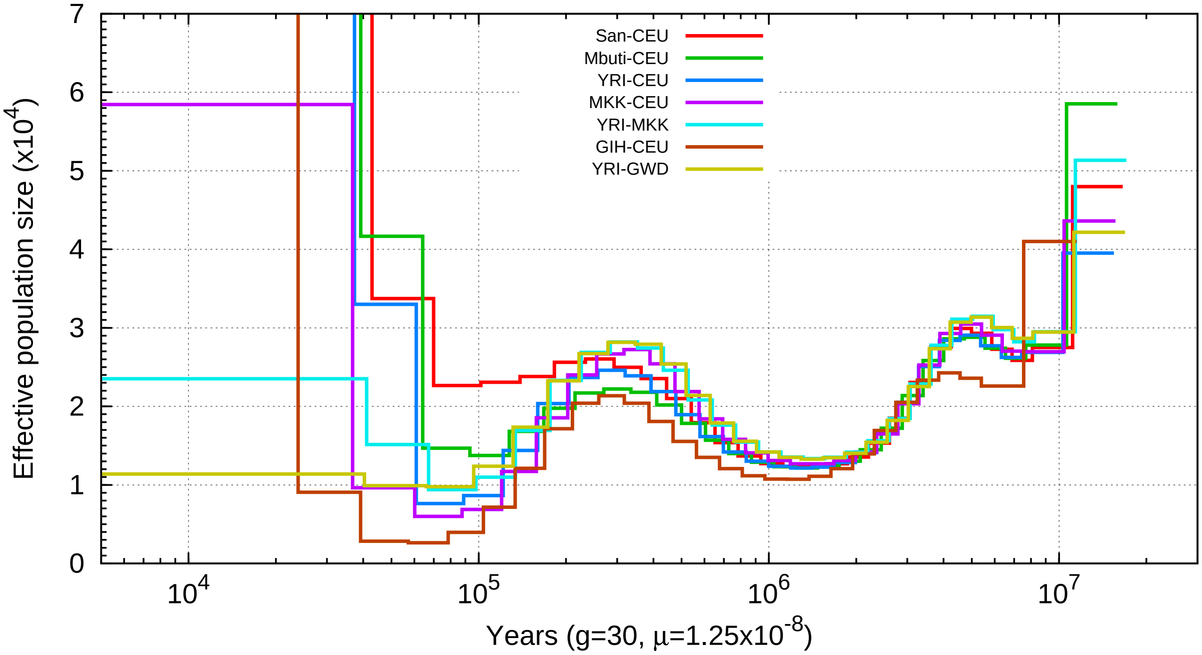

Figure 3.3: **PSMC on pseudo-diploid genomes.** Population sizes inferred from combined autosomes, with one haplotype chosen from each population are shown. Plotted curves are the average results obtained from four possible global haplotype configuration, namely hap1-hap1,hap1-hap2,hap2-hap1,hap2-hap2. Haplotypes were constructed using the fosmid pool approach.

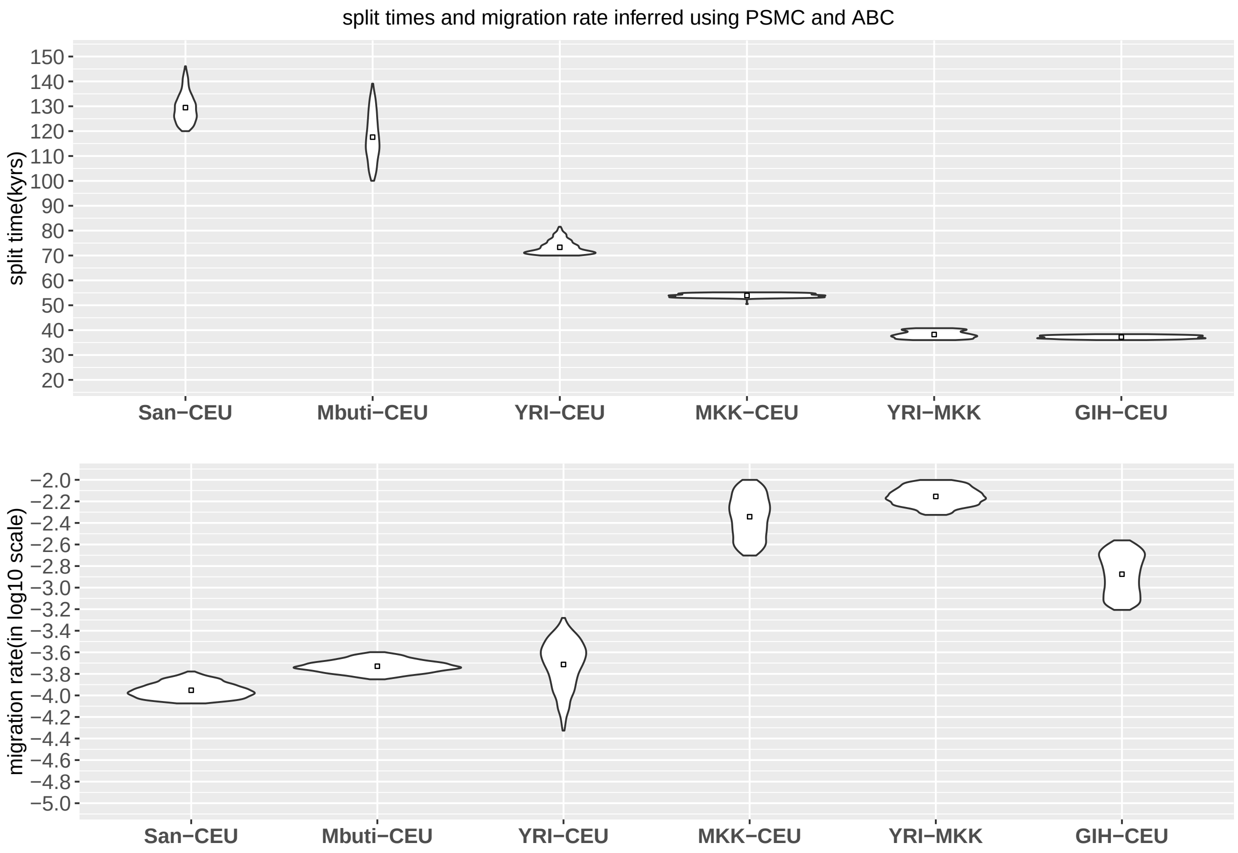

Figure 3.4: **Split times and migration rate inferred using PSMC and ABC.** We implemented ABC-SMC framework to estimate split time (A) and migration rate (B) given the inferred TMRCA distribution obtained from PSMC output. The posterior distribution of last iteration (N=1000 particles) and the mean value is shown.

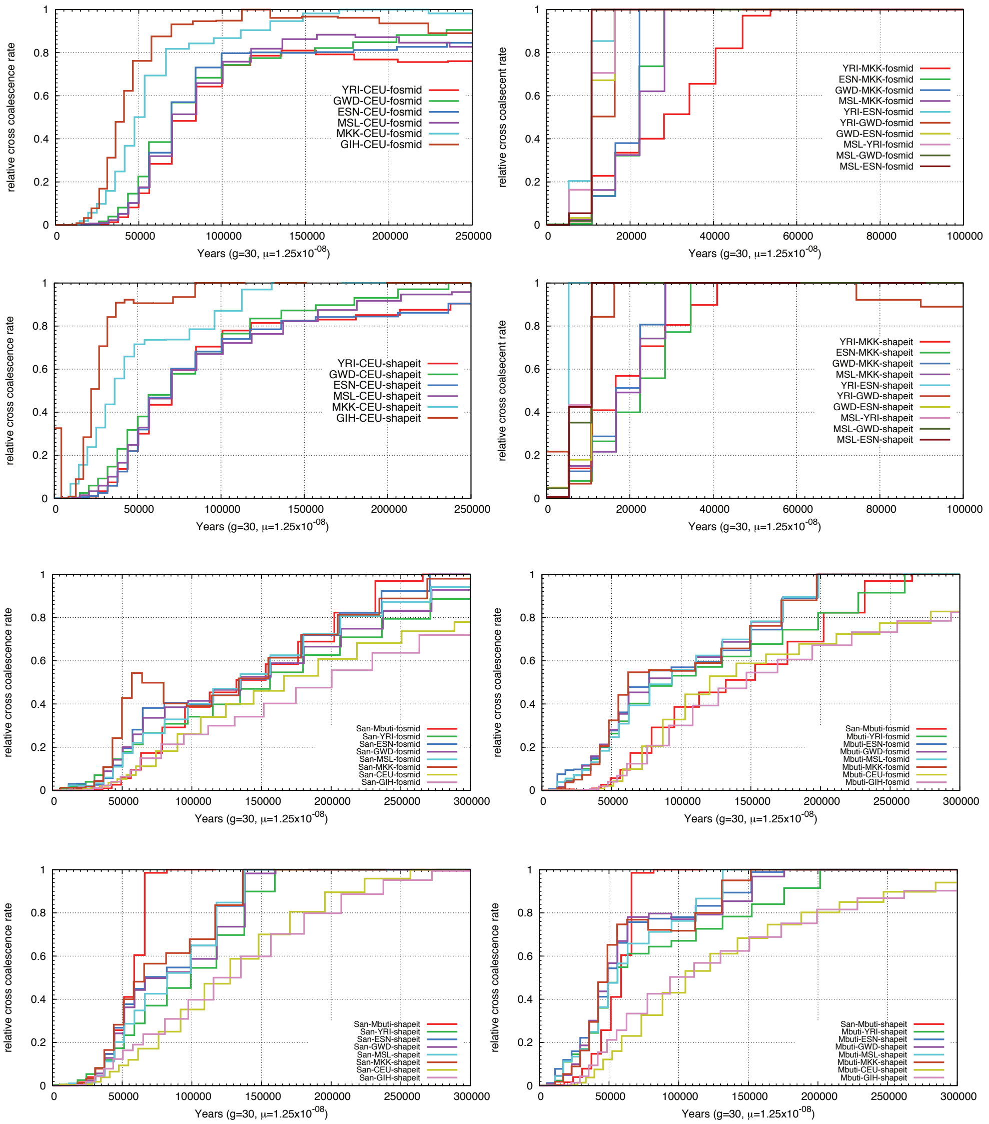

Figure 3.5: **Relative cross coalescence rate inferred using MSMC.** We applied msmc on four haplotypes, two from each population. We compared the relative cross coalescence curve using physically phased haplotypes vs haplotypes phased using SHAPEIT with 1000 Genomes Phase1 panel.

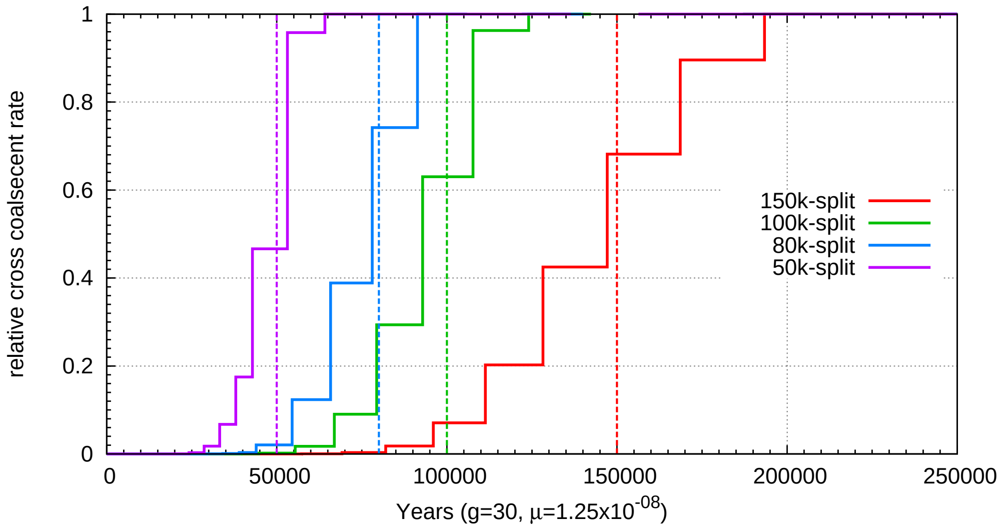

Figure 3.6: **Relative cross coalescence rate inferred using MSMC on simulated data.** We performed simulation using MaCS (100 30M sequences for each individual) with a clean population split at 50 kyrs, 80 kyrs, 100 kyrs and 150 kyrs ago. We applied msmc on simulated sequences and plotted the relative cross coalescence curve.

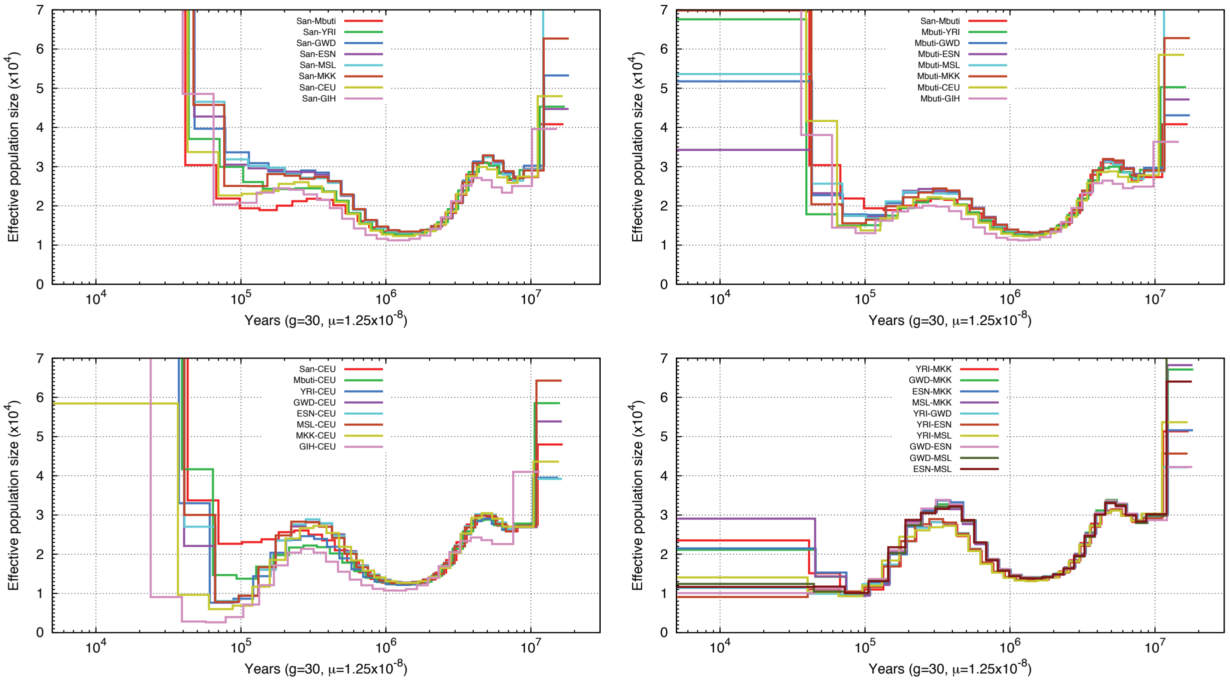

Figure 3.7: **PSMC on pseudo-diploid genomes.** Population sizes inferred from combined autosomes, one haplotype from each population are shown. Sizes are the average from 4 haplotype configuration, namely hap1-hap1,hap1-hap2,hap2-hap1,hap2-hap2. Haplotypes are constructed using fosmid pool approach.

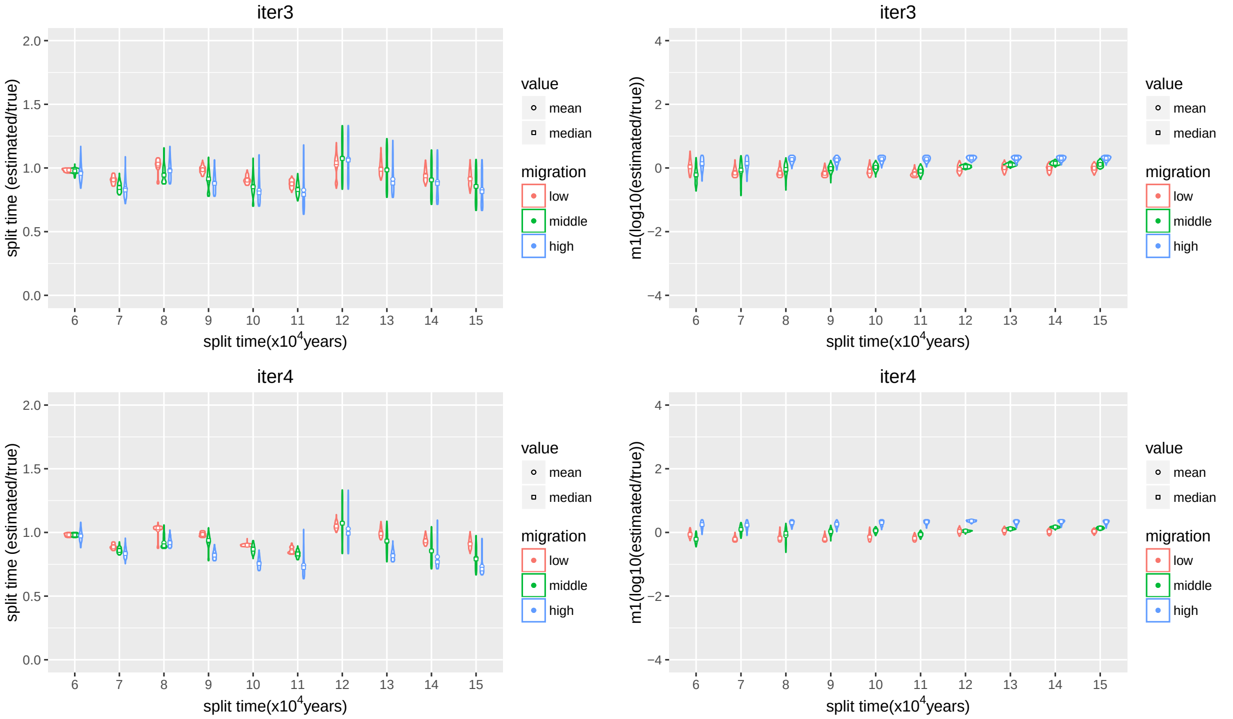

Figure 3.8: **Simulation results on inferring split time and migration using the combined approach of PSMC and ABC.** We tested our approach using simulated data of two individuals representing African and European population and having a split time from 60 kyrs to 150 kyrs ago, with subsequent migration until 30 kyrs ago. We tested three level of migration rate, 2×10^{-5} (low), $10 \times 10^{-5}(middle)$, 40×10^{-5}(high) and plotted the posterior distribution, mean and median value of split time (estimated/true) and migration rate (log10(estimated/true)) of iteration3 and iteration4 of our ABC approach.

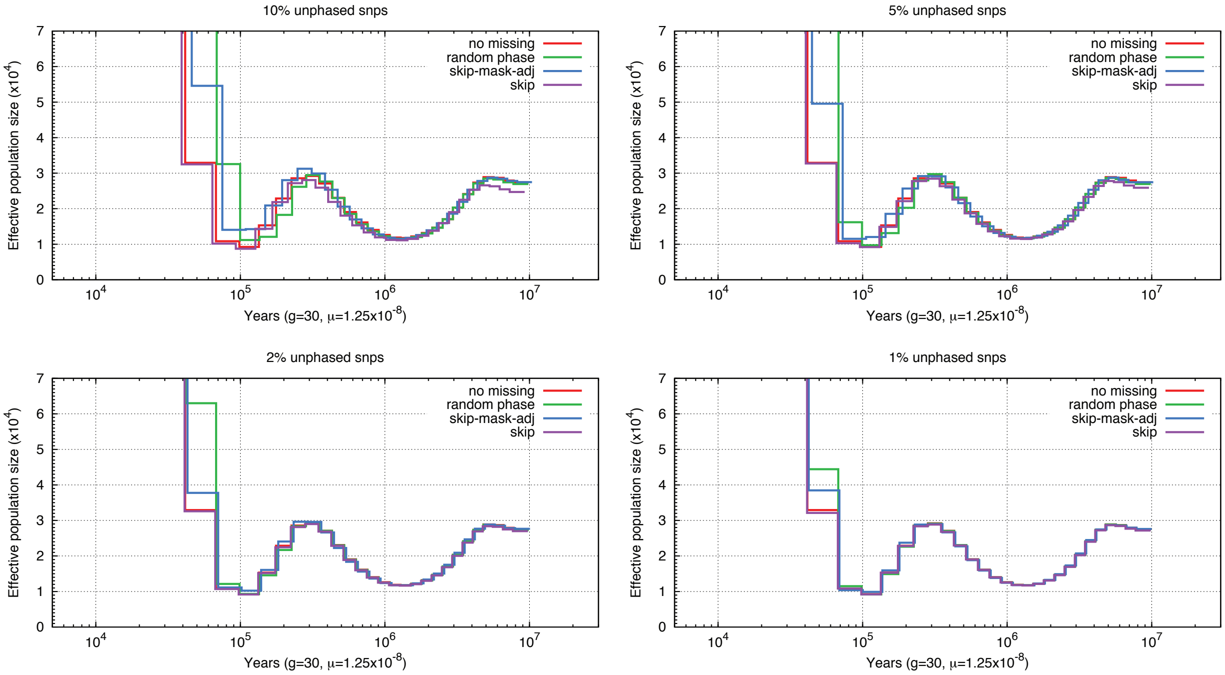

Figure 3.9: **Simulation results on different approaches to deal with unphased SNPs.** We simulated sequences with different levels of unphased snps (1%, 2%, 5%, 10%) and evaluated three different methods to deal with unphased snps, 1) randomly assigning the phase (green lines), 2) marking unphased snps as missing data and removing all blocks of homozygous calls that ended in an unphased heterozygous site (blue) and 3) merely marking unphased snps as missing data (purple).

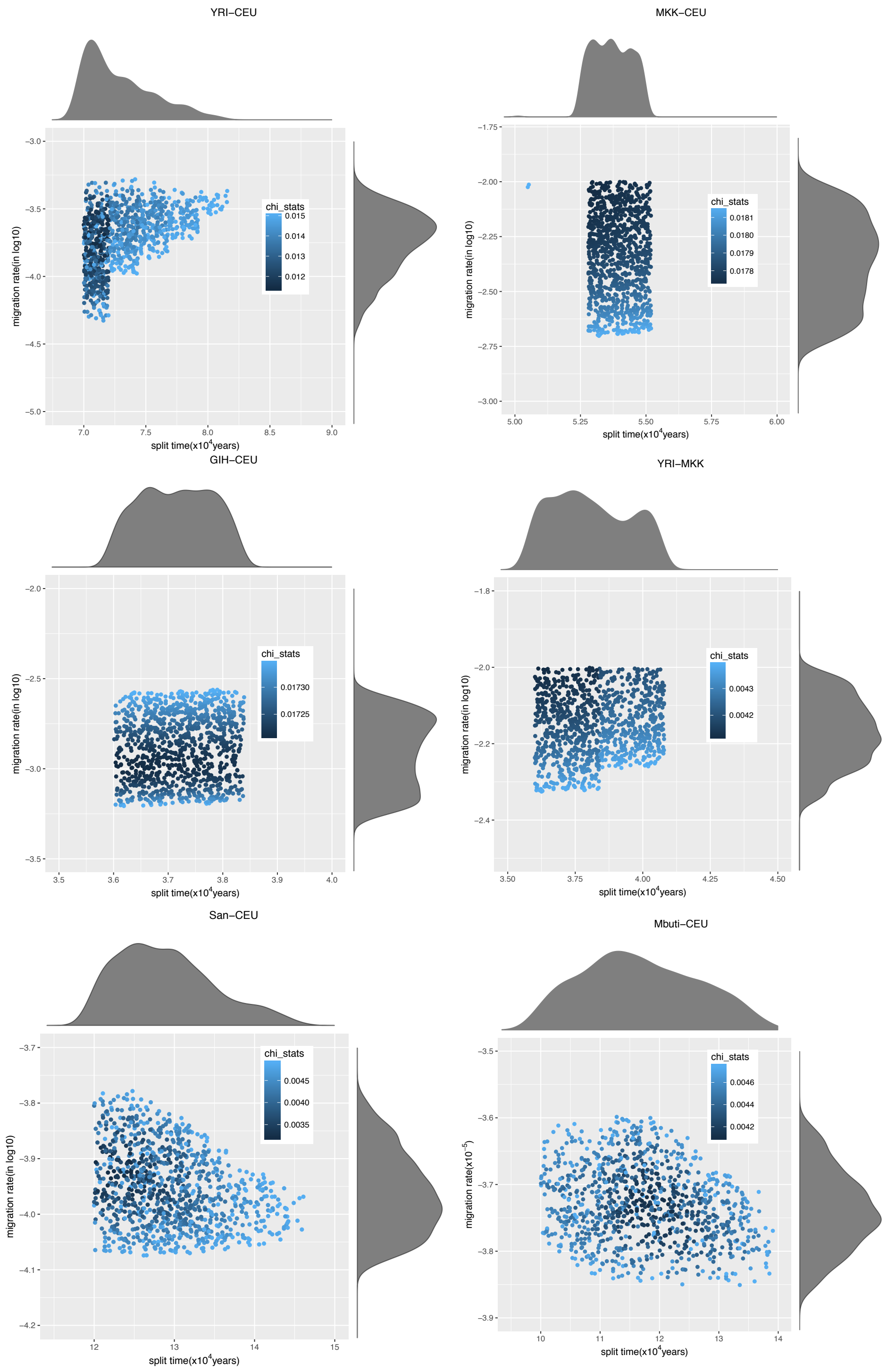

Figure 3.10: **Posterior distribution of split time and migration rate inferred using ABC.** We applied ABC-SMC to infer split time and migration rate based on the inferred TMRCA distribution obtained from PSMC. For each pair of populations, we plotted the posterior distribution of split time and migration rate. The color represents chi square distance between the TMRCA distribution from the observed data and the model.

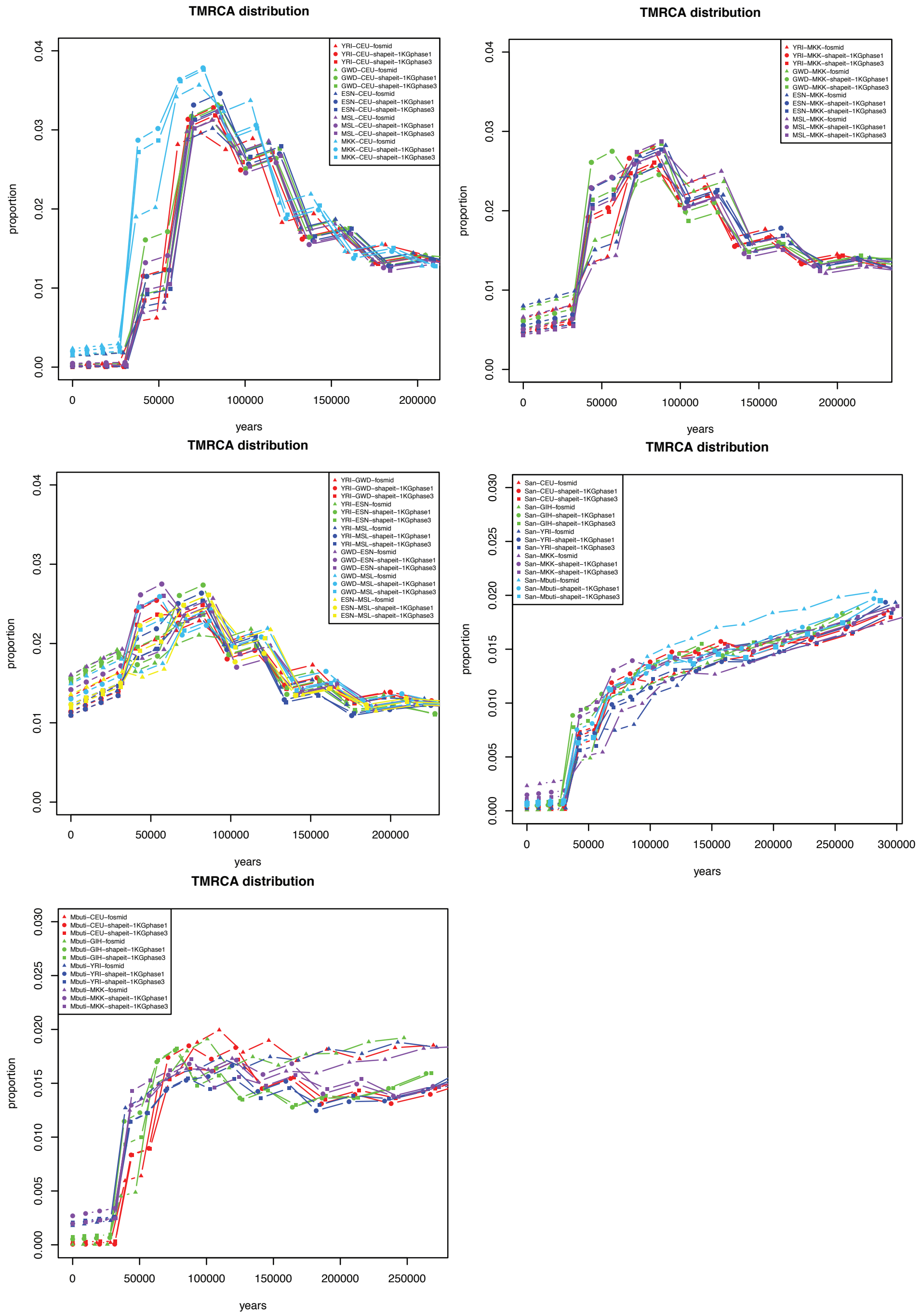

Figure 3.11: **TMRCA distribution inferred using PSMC.** The figure shows the left tail of TMRCA distribution inferred using PSMC on pseudo-diploid individuals for comparisons involving CEU, MKK, GWD, San, and Mbuti. Each plot shows the TMRCA distribution inferred using haplotypes phased using fosmid data (triangle) and phased using SHAPEIT with 1000 Genomes Phase1 (circle) and Phase3 (square) reference panels.

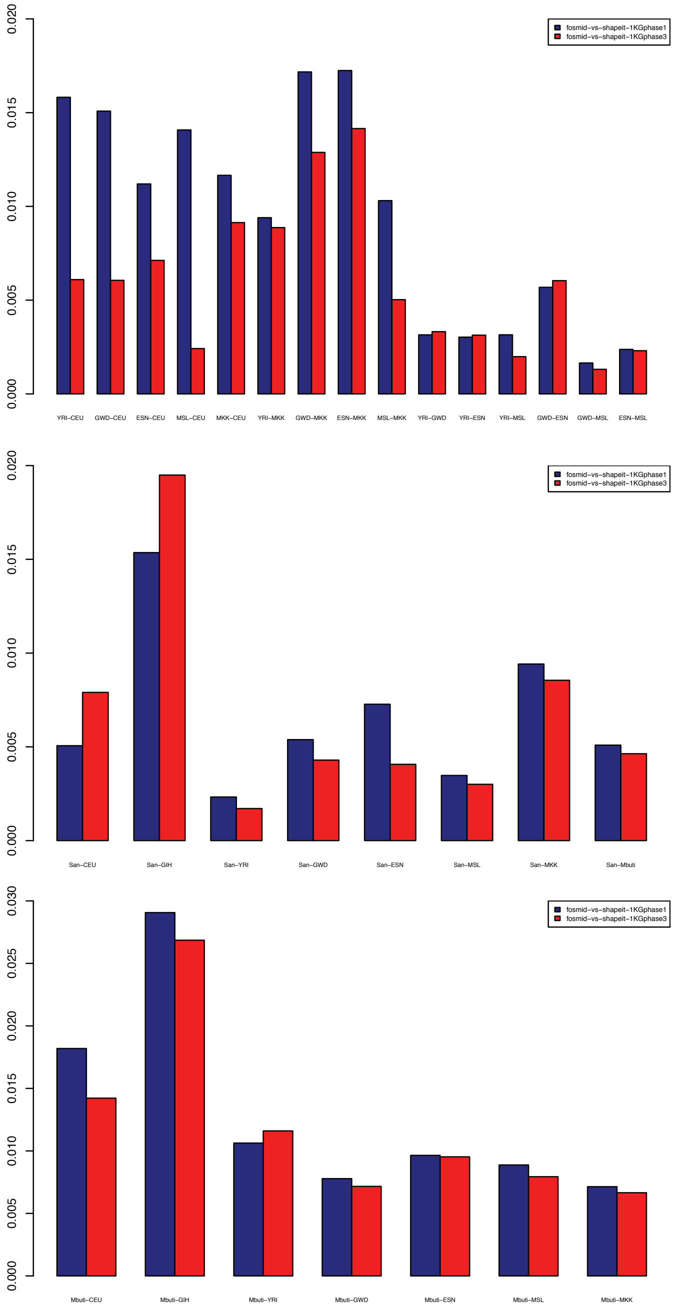

Figure 3.12: **Chi-square distance between TMRCA distributions using different haplotypes.** We plotted the chi-square distance between TMRCA distributions obtained using different haplotypes phased using fosmid data and phased using SHAPEIT with 1000 Genomes Phase1 (blue) and Phase3 (red) reference panels.

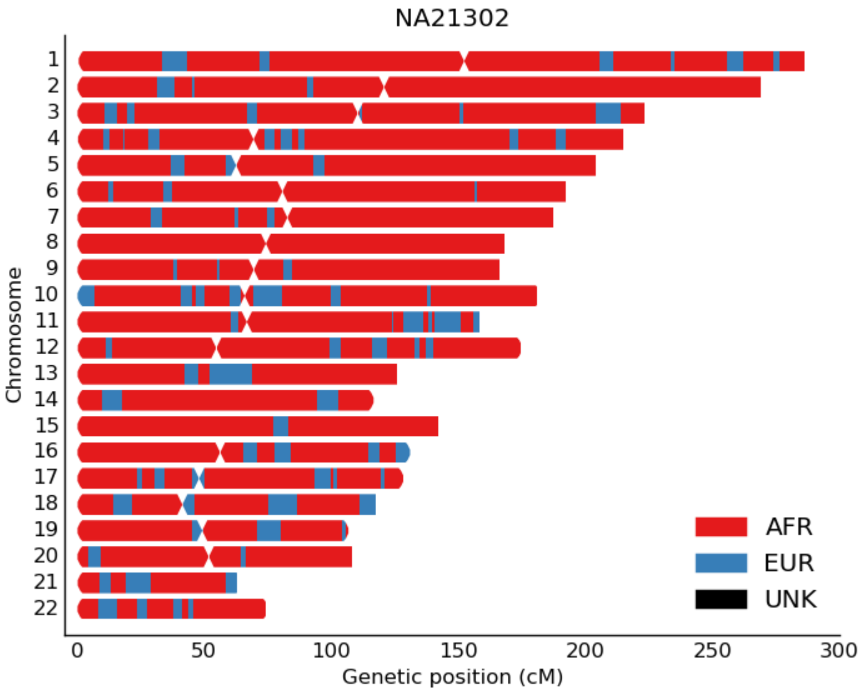

Figure 3.13: **Recent European ancestry inferred by RFMix in Massai individual NA21302.** The genomic locations of European ancestry (colored blue) in Massai individual NA21302 are shown.

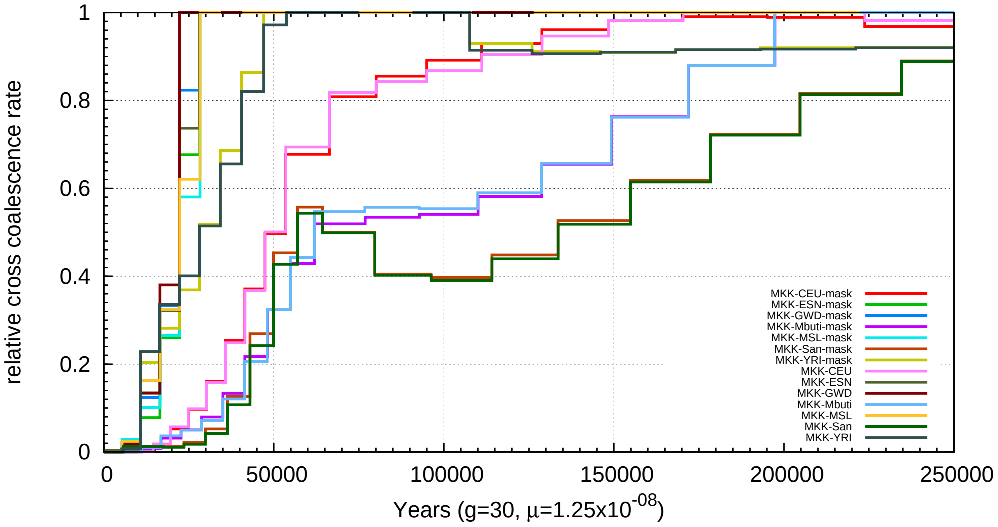

Figure 3.14: **Relative cross coalescence rate inferred using MSMC with and without masking out European ancestry from Massai individual.** We applied MSMC on MKK and every other population with and without masking European ancestry from the Massai individual.

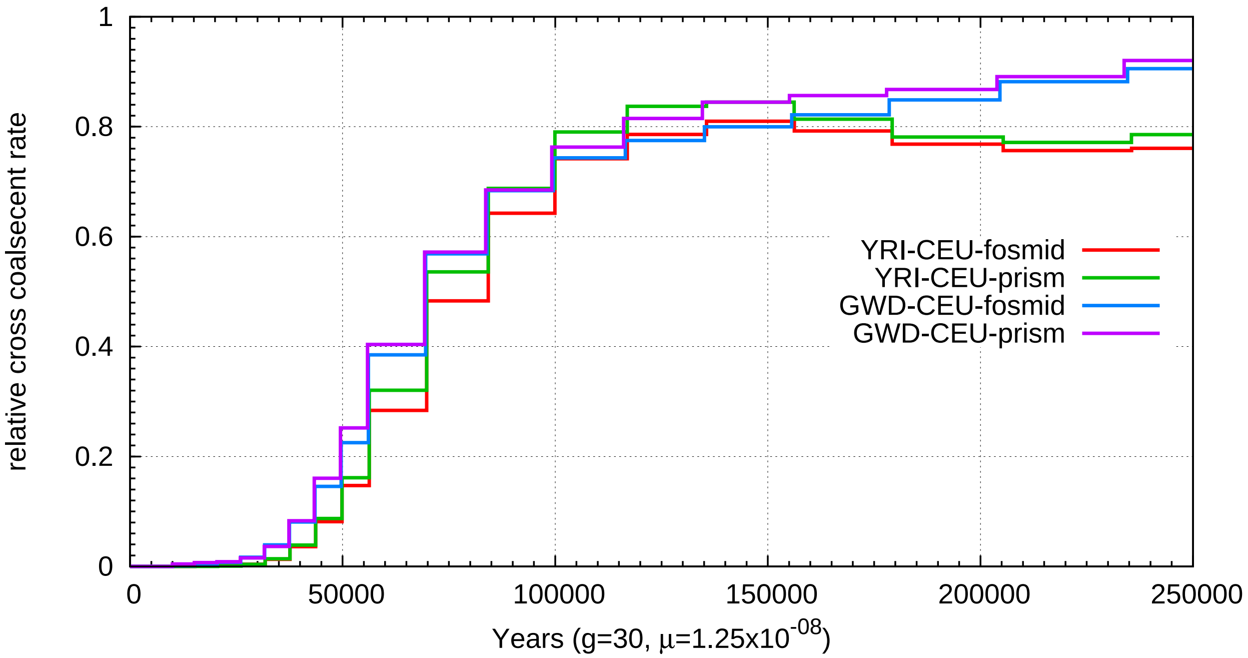

Figure 3.15: **Comparison of relative cross coalescence rate inferred using MSMC on fosmid haplotypes constructed by trio phasing data or by Prism.** We compared the relative cross coalescence curve for YRI-CEU and GWD-CEU, where the global haplotypes for YRI and GWD are either constructed by trio phasing data or by Prism.

Tables

Table 3.1: **Percentage of unphased SNPs.** We summarized the proportion of unphased SNPs for each population combination. For ‘Shapeit’ we refer to applying SHAPEIT with 1000 Genomes Phase I reference panel.

	Phasing method	Unphased	Total heterozygous	% unphased
YRI-CEU	Fosmid-Fosmid	416,956	4,123,413	10.11%
	Shapeit-Shapeit	106,327	4,123,413	2.58%
MKK-CEU	Fosmid-Fosmid	311,519	4,157,467	7.49%
	Shapeit-Shapeit	155,041	4,157,467	3.73%
GWD-CEU	Fosmid-Fosmid	302,959	4,360,333	6.95%
	Shapeit-Shapeit	118,494	4,360,333	2.72%
ESN-CEU	Fosmid-Fosmid	303,306	4,360,684	6.96%
	Shapeit-Shapeit	99,700	4,360,684	2.29%
MSL-CEU	Fosmid-Fosmid	310,165	4,329,461	7.16%
	Shapeit-Shapeit	142,952	4,329,461	3.30%
GIH-CEU	Fosmid-Fosmid	328,041	2,876,160	11.41%
	Shapeit-Shapeit	109,671	2,876,160	3.81%
YRI-MKK	Fosmid-Fosmid	210,859	4,596,417	4.59%
	Shapeit-Shapeit	168,853	4,596,417	3.67%

Table 3.2: **Split time estimation from previous studies.** Reported estimates are adjusted by using the same mutation rate 1.25×10^{-8} bp/generation and generation time 30 years. Estimates are in 1000 years

Paper	Method	Mutation	San & Others		African & Non-African		European-Asian	
			Original	Adjusted	Original	Adjusted	Original	Adjusted
Gravel et al. 2011	diffusion, AFS	$\mu = 2.36 \times 10^{-8}$, G=25			51	115.5	23	45.2
Gronau et al. 2011	Bayesian coalescent based	HC divergence 6.5Mya, $\mu = 2 \times 10^{-8}$	131 (127-135)	218 (214-226)	47 (44-49)	79 (74-82)	36 (34-38)	60 (57-65)
Harris et al. 2013	IBS sharing				55	85.6		
Excoffier et al, 2013	fastsimcoal2	$\mu = 2.5 \times 10^{-8}$,G=25	65 (25-96)	156 (60-230)				
Veeramah et al. 2011	ABC	HC divergence 6 Mya, $\mu = 2.5 \times 10^{-8}$	110 (52-187)	220 (104-374)				

Table 3.3: **Posterior estimates of split time and migration rate using IM model.** We report the mean, median and 95% credible intervals for the posterior distribution. Migration rate are in log10 scale. We set migration continuing to the present for recent separations.

	Migration End (in kyrs)				MigrationRrate (in log10 scale)				Migration End (in kyrs)			
	Mean	Median	95% lower HPD	95% higher HPD	Mean	Median	95% lower HPD	95% higher HPD	Mean	Median	95% lower HPD	95% higher HPD
YRI-CEU	73.3	72.5	70.2	81.6	-3.71	-3.7	-4.12	-3.41	27.2	30.1	6.2	38.7
MKK-CEU	53.9	53.9	52.9	55.1	-2.34	-2.22	-2.66	-2.04	-	-	-	-
GIH-CEU	37.2	37.2	36.2	38.2	-2.88	-2.87	-3.17	-2.6	-	-	-	-
San-CEU	129.5	128.8	121.3	140.9	-3.95	-3.96	-4.07	-3.83	37.2	37.1	33.4	41.5
Mbuti-CEU	117.6	116.9	103.1	139.1	-3.73	-3.73	-3.82	-3.63	34.6	34.4	30.6	39.1
YRI-MKK	38.2	38.1	36.2	40.6	-2.15	-2.16	-2.3	-2	-	-	-	-

Bibliography

Chen, G. K., Marjoram, P. and Wall, J. D. Fast and flexible simulation of dna sequence data. *Genome research*, 19(1):136–142, 2009.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. and Foll, M. Robust demographic inference from genomic and snp data. *PLoS Genet*, 9(10):e1003905, 2013.

Fu, Q., Mittnik, A., Johnson, P. L., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J. et al. A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology*, 23(7):553–559, 2013.

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

Harding, R. M. and McVean, G. A structured ancestral population for the evolution of modern humans. *Current opinion in genetics & development*, 14(6):667–674, 2004.

Harris, K. and Nielsen, R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet*, 9(6):e1003521, 2013.

Henn, B. M., Cavalli-Sforza, L. L. and Feldman, M. W. The great human expansion. *Proceedings of the National Academy of Sciences*, 109(44):17758–17764, 2012.

Hobolth, A., Andersen, L. N. and Mailund, T. On computing the coalescence time density in an isolation-with-migration model with few samples. *Genetics*, 187(4):1241–1243, 2011.

Li, H. and Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

Maples, B. K., Gravel, S., Kenny, E. E. and Bustamante, C. D. Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2):278–288, 2013.

Marjoram, P. and Wall, J. D. Fast" coalescent" simulation. *BMC genetics*, 7(1):1, 2006.

McVean, G. A. and Cardin, N. J. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1459):1387–1393, 2005.

Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., Xue, Y., Haber, M., Ekong, R., Oljira, T. et al. Tracing the route of modern humans out of africa by using 225 human genome sequences from ethiopians and egyptians. *The American Journal of Human Genetics*, 96(6):986–991, 2015.

Pickrell, J. K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B. and Reich, D. Ancient west eurasian ancestry in southern and eastern africa. *Proceedings of the National Academy of Sciences*, 111(7):2632–2637, 2014.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G. et al. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, 2013.

Pritchard, J. K. Whole-genome sequencing data offer insights into human demography. *Nature*, 201(1), 2011.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C. et al. The complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505(7481):43–49, 2014.

Schiffels, S. and Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919–925, 2014.

Schraiber, J. G. and Akey, J. M. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 2015.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. P. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

Veeramah, K. R. and Hammer, M. F. The impact of whole-genome sequencing on the reconstruction of human population history. *Nature Reviews Genetics*, 15(3):149–162, 2014.

Veeramah, K. R., Wegmann, D., Woerner, A., Mendez, F. L., Watkins, J. C., Destro-Bisol, G., Soodyall, H., Louie, L. and Hammer, M. F. An early divergence of khoesan ancestors from those of other modern humans is supported by an abc-based analysis of autosomal resequencing data. *Molecular biology and evolution*, page msr212, 2011.

Wang, Y. and Hey, J. Estimating divergence parameters with small samples from a large number of loci. *Genetics*, 184(2):363–379, 2010.

CHAPTER IV

Demographic history inference on gorilla species

4.1 Introduction

The Gorilla genus consists of two morphologically distinguishable species, western (*Gorilla gorilla*) and eastern (*Gorilla beringei*) gorillas (Grubb et al., 2003), each of which is divided into two recognized subspecies. Eastern gorilla populations occur in lowlands and highlands in the Democratic Republic of Congo, Uganda and Rwanda; while western gorilla populations reside primarily in Cameroon, Equatorial Guinea, Gabon, Congo, and the Central African Republic. Western gorillas include western lowland gorillas (*Gorilla gorilla gorilla*), the subspecies with the largest population size and Cross River gorillas (*G. gorilla diehli*), of which only a few hundred individuals remain. Eastern gorillas are composed of eastern lowland gorillas (*G. beringei graueri*) and mountain gorillas (*G. beringei beringei*), which are found today in only two small isolated subpopulations. Both species of gorilla are considered threatened on the IUCN Red List of Threatened Species (IUCN 2013); western gorillas are classified as critically endangered and eastern gorillas are classified as endangered. Recent census estimates indicate a rapid recent population size contraction in gorillas due to multiple factors including: outbreaks of the Ebola virus

The work presented in Chapter IV is published as Kimberly F. McManus*, Joanna L. Kelley*, **Shiya Song***, Krishna Veeramah, August E. Woerner, Laurie S. Stevison, Oliver A. Ryder, Great Ape Genome Diversity Consortium, Jeff M. Kidd, Jeff Wall, Carlos D. Bustamante, and Michael Hammer. (2015). Inference of Gorilla Demographic and Selective History from Whole-Genome Sequence Data. *Molecular biology and evolution*, 32(3), 600-612. (*equal contribution). The SFS analyses were performed by Kimberly McManus and Joanna Kelley.

(Le Gouar et al., 2009), the bushmeat trade, habitat loss and fragmentation (Walsh et al., 2003; Anthony et al., 2007).

Gorillas are human's closet living relatives other than chimpanzees and are as well important for the study of human evolution. The Gorilla reference genome assembly was first released in 2012, generated from a single female western lowland gorilla (Scally et al., 2012). Application of the CoalHMM coalescent inference model (Hobolth et al., 2007) on five-way whole genome alignment of human, chimpanzee, gorilla, orangutan and macaque revealed that the divergence time between human and chimpanzee is around 5.5 to 7 Myr ago, and the divergence time between human, chimpanzee and gorilla is 8.5 to 12 Myr ago considering possible variation of mutation rates. Prado-Martinez et al. (2013) sequenced 27 gorillas in high coverage, together with other individuals representing all six great ape species, providing a comprehensive catalog of great ape genetic diversity.

Research on demographic events experienced by gorillas can provide insights to the evolutionary forces that have uniquely influenced patterns of gorilla morphological and genetic variation. There has been considerable effort to estimate split times and population sizes for western and eastern gorillas (Ackermann and Bishop, 2010; Prado-Martinez et al., 2013; Scally et al., 2012; Thalmann et al., 2011; 2007). These studies make use of disparate data sets and modeling assumptions, particularly in terms of the treatment of gene flow subsequent to initial population separations. Based on eight microsatellites, Thalmann et al. (2011) estimate that the separation of Cross River and western lowland gorilla populations occurred 17.8 kyrs ago, followed by a comparatively high level of gene flow. On the other hand, Prado-Martinez et al. (2013) estimated this population divergence time at 114 kyrs ago based on a modified PSMC approach (Note: the above mentioned values have been adjusted to match the mutation rate used in this study where appropriate). The random phasing procedure applied in the modified PSMC approach may not be appropriate for such recent population split times. Moreover, estimates of the separation of eastern gorillas from the western lowland/Cross River gorillas range from about 100 kyrs to 450 kyrs ago, with varying degrees, lengths and directions of gene flow (Ackermann and Bishop, 2010; Becquet and Przeworski, 2007; Mailund et al., 2012; Prado-Martinez et al., 2013; Scally et al., 2012; Thalmann et al., 2007). Additionally, previous studies suggest substructure within the western lowland gorilla species (Scally et al., 2012).

In this study, we use a Bayesian coalescent approach to infer divergence times, rates of gene flow, and effective population sizes based on medium to high coverage whole genome data from three gorilla subspecies: western lowland, Cross River and eastern lowland. We also use a diffusion approximation approach to infer temporal changes in western lowland gorilla effective population size.

4.2 Methods

4.2.1 Whole genome sequences of gorilla

We obtained 17 whole genome sequences of gorillas, including 14 western lowland gorillas, 2 eastern lowland gorillas and a Cross River gorilla. DNA from those individuals was mostly obtained from blood from wild-caught zoo specimens (Prado-Martinez et al., 2013). All samples were sequenced on an Illumina sequencing platform (HiSeq 2000) with data production at three different sequencing centers; samples were sequenced to 12.7–42.1 coverage. Samples were collected under the supervision of ethical committees and CITES permissions were obtained as necessary. Sequence reads are available from the SRA under accession SRP018689.

4.2.2 Genotype calling

Sequences were mapped to gorGor3 and filtered as detailed in Prado-Martinez et al. (2013). Variants were identified in three pools of samples: the 14 western lowland gorillas, the 2 eastern gorillas, and the 1 Cross River gorilla sample. To compare variant calls among sample sets, we generated genome masks that identified all sites that were callable across all samples. Filters were calibrated such that we captured 90 % of sites that passed the VQSR procedure (Prado-Martinez et al., 2013). For western lowland gorillas, the filters correspond to a total sample read depth (DP) ≥ 95 and ≤ 307, mapping quality (MQ) ≥ 39 and percent of reads with mapping quality 0 (MQ0fraction) ≤ 3. For eastern lowland gorilla, the criteria were DP ≥ 12 and ≤ 37, MQ ≥ 33, and MQ0fraction ≤ 4. For the Cross River gorilla, the criteria were DP ≥ 5 and ≤ 24, MQ ≥ 38 and MQ0fraction ≤ 0. For each sample set, we additionally removed sites within 5 bp of called indels, and removed all positions overlapping with segmental duplications (Sudmant et al., 2013). For G-PhoCS analysis, variants were identified for each sample independently using BSNP to avoid bias induced by the reference genome and from population level genotype calling.

4.2.3 Population structure

Inference of population structure and principle components analysis, which require a set of independent SNPs, were conducted on 10% thinned data when comparing all three subspecies, using ADMIXTURE (Alexander et al., 2009) and smartpca (Patterson et al., 2006), respectively. PCA of three species was conducted on the intersect of the 8x data in western lowland, Cross River, and eastern lowland gorillas. When considering only the western lowland gorillas, data was pruned for linkage (using plink –indep 50 5 2). We performed 10 independent ADMIXTURE runs for each tested value of K.

4.2.4 Demographic analysis using G-PhoCS

G-PhoCS utilizes input alignments from multiple independent “neutral loci” in which recombination within loci occurred at negligible rate but recombination between loci were sufficient to assume that genealogies are approximately uncorrelated (Gronau et al., 2011). Assuming that parameters of recombination are broadly consistent among primates, we adopted the 37,574 neutral loci previously identified by Gronau et al. (2011) for the human genome (build NCBI36), lifted-over these loci to gorGor3.1 (Ensembl release 64) and then applied a series of filters to obtain a new set of “neutral loci” for the gorilla genome. Specifically, we removed regions without conserved synteny in human-gorilla alignments, recent transposable elements annotated by RepeatMasker with ¡=20% divergence, exons of protein-coding genes, conserved noncoding elements according to phastCons, and recent segmental duplications in gorilla. This resulted in 26,248 loci, with size of 1 kb and interlocus distance of 50 kb. We called genotypes from the whole genome data at these neutral loci using BSNP (Gronau et al., 2011), setting –P flat, which assumes uniform prior distribution to determine genotype calls for each individual without bias introduced by the reference genome. For each locus, we also masked simple repeats, positions within 3 bp of an insertion/deletion, positions with less than 5 reads, and CpG sites. Finally we used MUSCLE (Edgar, 2004) to make alignments of each inferred sequence. After removing loci with completely missing data (all Ns) in at least one individual, we obtained a final set of 25,573 neutral loci for input to G-PhoCS.

We applied G-PhoCS to different combinations of samples. These combinations always included both eastern lowland gorillas, the single Cross River gorilla but contained different combinations of two western lowland gorillas. An aligned human reference genome was included as an outgroup. We first evaluated four alternative scenarios: no gene flow between gorillas species and bi-directional gene flow between any two gorillas species. For each case, we ran G-PhoCS for 50,000 iterations and found that this was sufficient to establish convergence for the no gene flow and bi-directional gene flow models between western and Cross River gorilla. We reran the analysis with bi-directional gene flow between western and eastern gorilla, and allowed two migration band parameters, one from western lowland to eastern lowland gorilla and another from Cross River to eastern lowland gorilla. We found that 300,000 iterations were sufficient to establish convergence for parameters of interests and we set the burn-in as the first two-thirds of iterations. The raw estimates by G-PhoCS are ratios between model parameters. Using humans as an outgroup, we calibrated the model based on the average genomic divergence time between human and gorilla, denoted T_{div}. We assume a range of $T_{div} = 8.0 - 12.0$ Mya. An average mutation rate was calculated by $\mu = \tau_{div}/T_{div}$. Effective population sizes were calibrated by a factor of $(4 \times 19 \times \mu)^{-1}$, assuming an average gorilla generation time of 19 years (Langergraber et al., 2012). We also calculated estimates of expected number of migrants per generation, given by $m_{AB} \times \theta_B$ and the total migration rate, given by $m_{AB} \times \tau_{AB}$.

4.2.5 Demographic Inference of Western Lowland Gorilla

The western lowland gorilla single population demographic model was inferred through a diffusion approximation approach implemented in the software (Gutenkunst et al., 2009). This approach calculates the log likelihood of the model fit based on a comparison between the observed and expected site frequency spectrum (SFS). Five demographic models were evaluated: a standard neutral model, an exponential growth model, a model of a bottleneck followed by exponential growth, a two epoch model and a three epoch model. We evaluated results with all SNPs that passed the filters and had at least 8x coverage, as well as a subset of these SNPs thinned to 100kb. For each model, ten independent runs were performed and the model and associated parameters that maximized the likelihood were chosen. To convert parameter output to years and effective population sizes, we assumed a mutation rate of 1.1×10^{-8} per generation (Roach et al., 2010; Consortium et al., 2010) and a generation time of 19 years (Langergraber et al., 2012). Confidence intervals for each parameter were determined through bootstrapping the input SNPs in blocks of 500 kb 1000 times.

For analysis of western lowland gorillas we used all sites from the genomic data with at least 8x coverage in all samples. The unfolded (polarized) SFS was determined using humans as an outgroup and ancestral misidentification was corrected using the method developed in (Hernandez et al., 2007), which is implemented in $\partial a \partial i$. Briefly, this approach infers the unfolded SFS through a context dependent mutation model. It considers the trinucleotide sequence context of each SNP in gorillas and the outgroup, the great ape transition rate matrix for each nucleotide (as in (Hwang and Green, 2004), provided by Hwang DG, unpublished), the proportion of each trinucleotide sequence in the gorilla sequence data, and the gorilla-outgroup divergence (empirically estimated at 1.60% and 1.51% from the complete and 8x filtered sequence data, respectively).

4.3 Results

4.3.1 Gorilla Population Structure

Whole genome sequence data from 17 gorillas, including 14 western lowland gorillas, 2 eastern lowland gorillas, and 1 Cross River gorilla were aligned to the gorGor3 (Ensembl release 62) reference genome and processed with filtering as previously described (Prado-Martinez et al., 2013). We limited our analysis to samples without evidence of inter and intra-species sequence contamination and characterized patterns of genetic variation based on SNP genotypes obtained using the GATK Unified Genotyper (McKenna et al., 2010), limited to sites with at least eight-fold (8x) coverage in all samples. Across the autosomes, we observe that eastern lowland gorillas have the lowest heterozygosity ($5.62 - 5.69 \times 10^{-4}$) of all the groups studied here followed by the single Cross River sample (9.09×10^{-4}) and 14 western lowland gorillas ($1.2 - 1.6 \times 10^{-3}$) (Figure 4.1. We used principal components analysis (PCA) (Patterson et al., 2006) and ADMIXTURE (Alexander et al., 2009) to further explore relationships among the samples. As previously observed (Prado-Martinez et al., 2013), when considering all samples together, PC1 shows clear separation of eastern and western gorillas with western lowland and Cross River gorillas arrayed along PC2 (Figure 4.2). PCA performed on only the western lowland gorilla samples does not reveal clear population clusters, although the individuals are somewhat ordered by sample geography (Figure 4.3). Results from ADMIXTURE, a model-based clustering algorithm allowing for mixed ancestry, support the existence of two clusters dividing eastern and western lowland gorillas (Figure 4.4). When applied to only the 14 western lowland gorilla samples, we observe that K=1 has the lowest cross-validation (CV) error (Figure 4.5).

4.3.2 Demographic inference on gorilla

We applied Generalized Phylogenetic Coalescent Sampler (G-PhoCS), a Bayesian coalescent-based approach, to infer ancestral population sizes, divergence times, and rates of gene flow (Gronau et al., 2011) amongst the three gorilla subspecies. This inference is based on genealogies inferred at many independent and neutrally evolving loci across the autosomal genome. To avoid bias caused by the alleles represented in the reference genome, which is derived from a western gorilla, we used BSNP (Gronau et al., 2011), a reference-genome–free Bayesian genotype inference algorithm, to perform variant calling separately for each sample. Based on the BSNP output, we produced diploid sequence alignments of two eastern lowland gorillas, nine western lowland gorillas, one Cross River gorilla, and the human reference genome at 25,573 "neutral loci" with size approximately 1 kilobase (kb) and an interlocus distance of approximately 50 kb. The neutral loci were chosen based on the positions of putatively neutral loci previously utilized for humans (Gronau et al., 2011), but further filtered to remove loci that intersected with exons, conserved elements, recent transposable elements, and recent segmental duplications in the gorilla genome.

For many of the analyses presented here, we used a four-population phylogeny as inferred by TreeMix (Pickrell and Pritchard, 2012), with eastern and western gorilla ancestors separating first, followed by western lowland and Cross River gorilla (Figure 4.6, Figure 4.7). We first evaluated four alternative scenarios, having either no gene flow between any gorillas (Figure 4.7, scenario 1) or bi- directional gene flow between any two gorilla species (Figure 4.7, scenarios 2, 3, and 4). In G- PhoCS, gene flow is modeled using migration bands of constant migration rate between two lineages over the entire time period of their existence. We utilized several combinations of western lowland gorilla samples, always including two eastern lowland gorillas, two western lowland gorillas, one Cross River gorilla, and one human. We initially ran G-PhoCS for 50,000 iterations and monitored convergence using Tracer (Rambaut and Drummond, 2013). Estimates of population split times are sensitive to model assumptions, particularly gene flow. Our G-PhoCS analysis finds no evidence of migration events between western lowland and Cross River gorillas (Figure 4.8, scenario 2). We do observe evidence of gene flow from western lowland gorilla to eastern lowland gorilla with mean total migration rate 0.3 (95% CI: 0.240-0.356), equivalent to 0.37 migrants per generation (95% CI: 0.312-0.433) (Figure 4.8, scenario 4). We also observe a small signal of gene flow from Cross River gorilla to eastern lowland gorilla (Figure 4.8, scenario 3); however, 50,000 iterations were not sufficient for convergence. To further explore these results, we tested a scenario with two migration bands: one from western lowland to eastern lowland gorilla and another from Cross River to eastern lowland gorilla (Figure 4.7, scenario 5), and extended the number of iterations to 300,000 to allow the posterior estimates to fully converge (Figure 4.9). Setting an additional migration band from Cross River to eastern lowland gorilla makes little difference because migration from western lowland to eastern lowland gorilla has the strongest migration signal (Figure 4.10). The estimated migration rate from Cross River to eastern lowland gorilla is 0.004 (95% CI: 0.000-0.018), equivalent to 0.019 migrants per generation (95% CI: 0.000-0.071). By using this setting (Figure 4.7, scenario 5), we estimate the split time between western lowland gorilla and Cross River gorilla to be 68 thousand years ago, and the split time between eastern lowland and western ancestral gorilla to be 261 thousand years ago when assuming a human and gorilla divergence time of 12 million years ago (Scally et al., 2012) (Table 4.1). We also observed a decrease of western gorilla population size and a decrease of eastern gorilla population size after their initial split and a six-fold difference between current eastern and western gorilla population size. The relative population sizes of the gorilla populations are rather robust to the chronological human/gorilla split time used for calibration, though the actual estimated size and chronological date of the split times are sensitive to the split time assumptions as many calculations are pegged to the calibration date (Table 4.1).

4.3.3 Western Gorilla Demographic Inference

We additionally inferred the fine-scale population history of western lowland gorillas using the genome-wide site frequency spectrum (SFS) obtained from 14 individuals (Gutenkunst et al., 2009). We utilized a diffusion approximation for demographic inference ($\partial a \partial i$) on the unfolded SFS based on 4,554,752 SNPs only considering sites where all samples had at least 8x coverage. Variants were polarized to ancestral and derived alleles based on human out-group sequences, and we implemented a context-dependent correction for ancestral misidentification (Hernandez et al., 2007). Five demographic models were fit using $\partial a \partial i$ and inferring the best-fit demographic model requires us to assess whether the improvement in fit afforded by additional parameters needed in more complex models are justified (Table 4.2). While the bottleneck followed by exponential growth model and the three-epochs models have similar fits, the three-epochs has the best fit; moreover, the model selection is robust when SNPs are thinned to 100 kb. Our results suggest an ancient expansion followed by a more recent drastic, 5.6-fold, population contraction is the best model for the data. Specifically, assuming a mutation rate of 1.1×10^{-8} per base pair per generation (Roach et al., 2010) and generation time of 19 years (Langergraber et al., 2012), the best-fit model is a three-epoch model that has an ancestral effective population size of 31,800 (95% CI: 30,690-32,582) (Table 4.2). The first size change event occurred 969,000 years ago (95% CI: 764,074-1,221,403) and increased the effective population size to 44,200 (95% CI: 42,424-46,403) individuals. The second size change event occurred 22,800 years ago (95% CI: 16,457-30,178) and decreased the effective population size to 7,900 (95% CI: 6,433-9,240) individuals (Figure 4.11).

4.4 Discussion

Several other studies have made use of disparate data sets and modeling assumptions to estimate population split times, sizes, and levels of gene flow for different gorilla species. The estimates described in this manuscript are broadly consistent with previous studies, but there are some differences (Table 4.3). Our estimate of 68 kyrs for the Cross River–western lowland split is intermediate between the previous estimates; however, we do not find support for gene flow between these two groups in our G-PhoCS analysis. Two main caveats apply to this analysis. First, in our G-PhoCS model estimates of gene flow and population split-time are confounded since western lowland and Cross River gorilla are sister species. Second, immediately following a gene flow event the variance in individual ancestry proportions across a population is large, with ancestry proportions becoming more uniform over time (Gravel, 2012). Since our analysis utilized a single Cross River sample, by chance we may have missed signals associated with very recent gene flow.

We estimate that the separation of eastern gorillas from the western lowland/Cross River ancestor occurred 261 kyrs ago, with subsequent gene flow from both western lowland and Cross River populations to the eastern gorillas. This value is similar to the 214 kyrs split time inferred by the modified PSMC approach. Scally et al. (2012), based on a model of symmetric gene flow, estimated a separation time of 429 kyrs ago. Mailund et al. (2012) arrives at a broadly similar estimate based on a coal-HMM, and estimates gene flow continuing until 150 kyrs ago. We note that our analysis indicates that the direction of gene flow was from western lowland and Cross River to eastern gorillas, with a higher rate from western lowland than from Cross River gorilla. However, Thalmann et al. (2007) find evidence for gene flow from eastern to western gorillas. Alternatively, Ackermann and Bishop (2010) find support for a western to eastern gene flow in morphological and molecular data. One way to assess evidence for gene flow is through the use of D-statistics, which provide a formal test for the fit of an unrooted tree to the data (Patterson et al., 2012). Excessive allele sharing not accounted for by the population tree is evidence in support of gene flow among the studied populations. The D-statistics calculated in Prado-Martinez et al. (2013) suggest that Cross River gorillas are genetically closer to eastern gorillas than western lowland gorillas are to eastern gorillas, which would not be predicted by the gene flow values we infer. We further explored this apparent contradiction by calculating D-statistics for additional samples from Prado-Martinez et al.

(2013) and using variants identified by BSNP based on mapping to the gorilla reference genome (Table 4.4). The western lowland gorilla sample A934_Delphi is not included in this study since it contains low-level contamination from a bonobo (Prado-Martinez et al., 2013). Consistent with this potential contamination, A934_Delphi shows an extreme value for the D-statistic relative to other western gorillas; however, significant statistics are also obtained when using other samples (Table 4.4). We do not observe significant D-statistics for genotypes calculated from reads mapped to the gorilla reference genome using BSNP (Table 4.4). Additional Cross River samples, as well as new analytic approaches that take advantage of the additional information contained in physically phased genome sequences (Schiffels and Durbin, 2014) may shed further light on patterns of gene flow among extant gorilla species.

Given the availability of 14 western lowland gorilla samples, we estimated a single-population demographic history using $\partial a \partial i$. Due to limited sample size, our model does not incorporate other subspecies/species. Our analysis indicates that western lowland gorillas have undergone a small, ancient population size expansion event 970 kyrs ago followed by a drastic size reduction 23 kyrs ago. These results are broadly concordant with previous estimates of temporal population size change in gorillas based on the PSMC model (Prado-Martinez et al., 2013) (Figure 4.12), especially given that it is known that PSMC tends to smooth instantaneous size changes. We note that the ancient increase predates our estimation for the separation of eastern and western gorillas, and the recent size decrease post-dates our estimation of Cross River – western lowland separation. The underlying causes of these effective population size changes are unclear. Previous studies note glacial and inter-glacial oscillations during the last two million years may have had an effect on gorilla population size and structure (Thalmann et al., 2007). For example, during the Last Glacial Maximum, rainforest cover was greatly diminished, especially in West Africa where a few refugia were surrounded by tropical grassland (Jolly et al., 1997). Previous studies suggest substructure within the western lowland gorilla species (Nsubuga et al., 2010; Fünfstück et al., 2014; Scally et al., 2012), but our results support the use of a one-population model of western lowland gorillas (though there may be some subtle isolation-by-distance or demic structure). Earlier studies that involved analysis of SSR motifs (DNA microsatellites) provided some indications of substructure within western lowland gorillas (Nsubuga et al., 2010; Fünfstück et al., 2014). While a slower evolving set of markers, such as SNPs, can identify expansion from a common ancestor and imply demographic changes over tens of thousands of generations, more rapidly evolving microsatellite loci can reveal more recent aspects of gene flow and population substructure. Our PCA and ADMIXTURE analysis support grouping of samples into one population for analysis. Additionally, models inferred separately on subsets of the data yielded concordant results.

In addition to the inferred decline in gorilla effective population size, census estimates note that the gorilla population has declined by more than 60% in the past 20-25 years, prompting their "critically endangered" conservation status (IUCN 2013). This decrease is thought to be due predominantly to Ebola outbreaks and commercial hunting (Walsh et al., 2003; Le Gouar et al., 2009). This sharp decline is much too recent to be observed in our analysis given the dataset available.

Figures

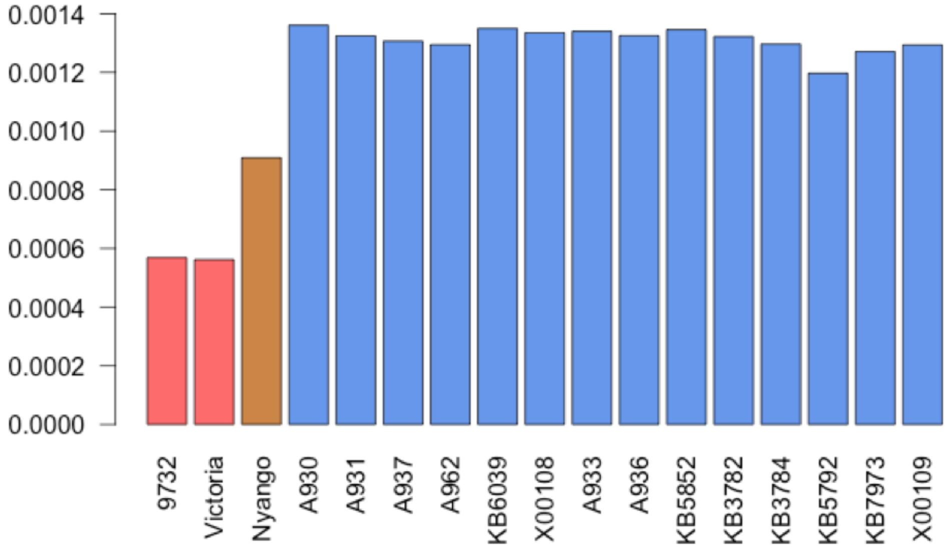

Figure 4.1: **Genome wide heterozygosity values.** Heterozygosity was calculated along the autosomes for each sample based on the 8x coverage masks. Samples are colored based on species: eastern lowland (red), Cross River (yellow), and western lowland (blue).

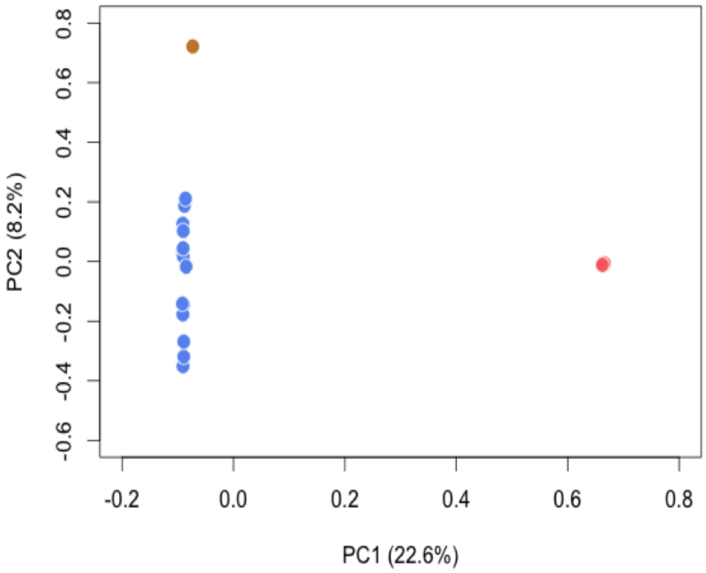

Figure 4.2: **PCA of three gorilla species.** Colors indicate species: Cross River (brown), western lowland (blue), and eastern lowland (red). Data was thinned to include 20% of the total number of intersecting SNPs. Percentages indicate the percent of variance explain by each principle component.

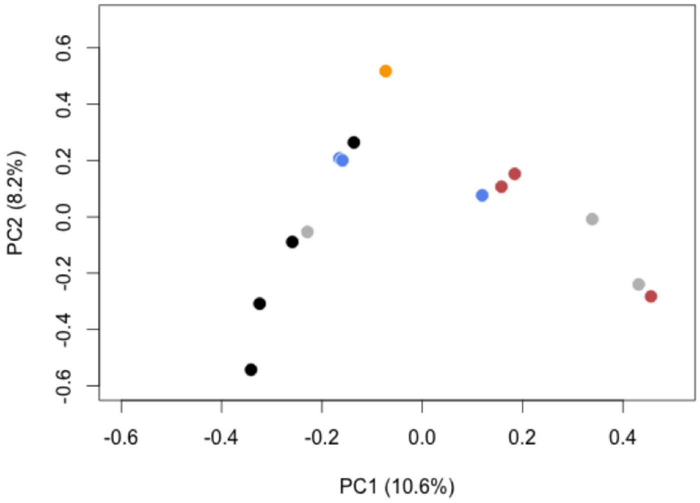

Figure 4.3: **PCA of western lowland gorilla.** PCA of 14 western lowland gorillas based on 8x sequencing data. Colors indicate geographical origin: Equatorial Guinea (orange), Cameroon (blue), Congo (red), captive born (black), unknown origin (grey). Percentages indicate the percent of variance explained by each principle component.

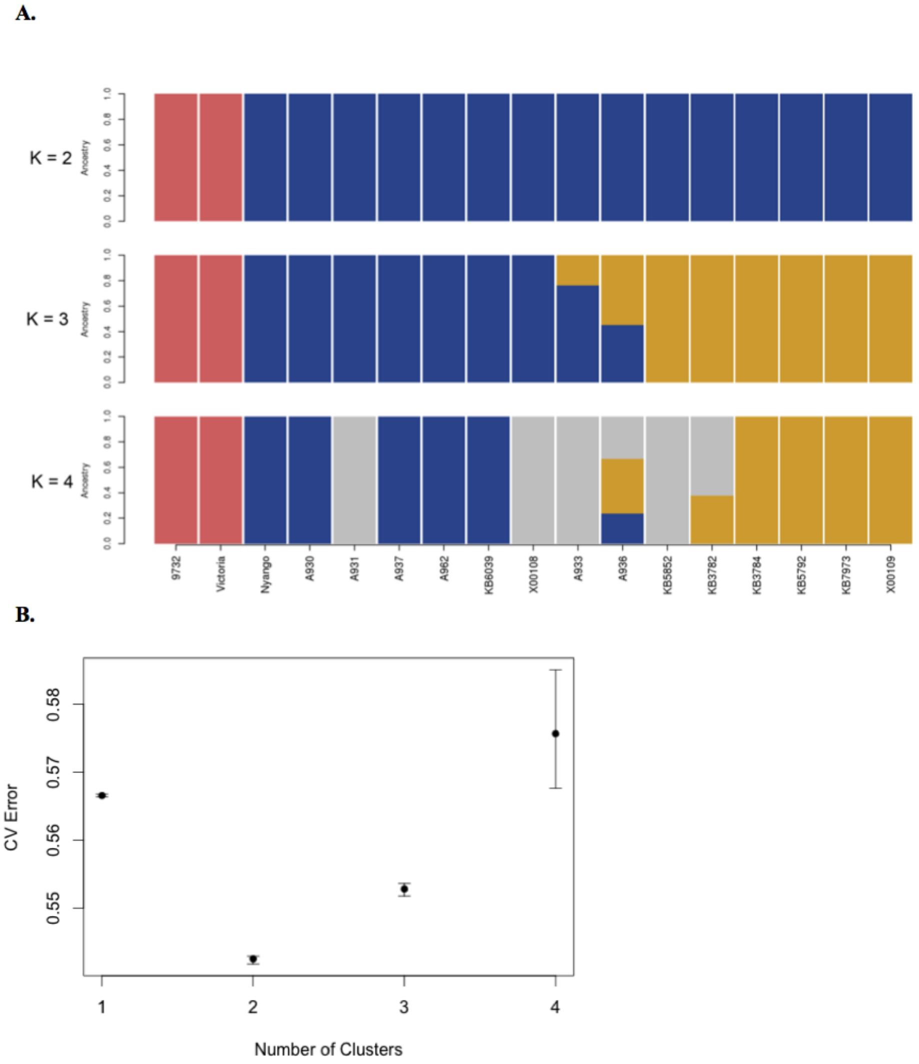

Figure 4.4: **Admixture results for the three gorilla subspecies.** A: Population membership inferred from ADMIXTURE with 2 - 4 populations. B: Cross Validation Error from ADMIXTURE analyses on three gorilla subspecies. CV error is estimated for 10 independent runs, and the results from run with the CV error is shown in A. (data generated by Kimberly F. McManus)

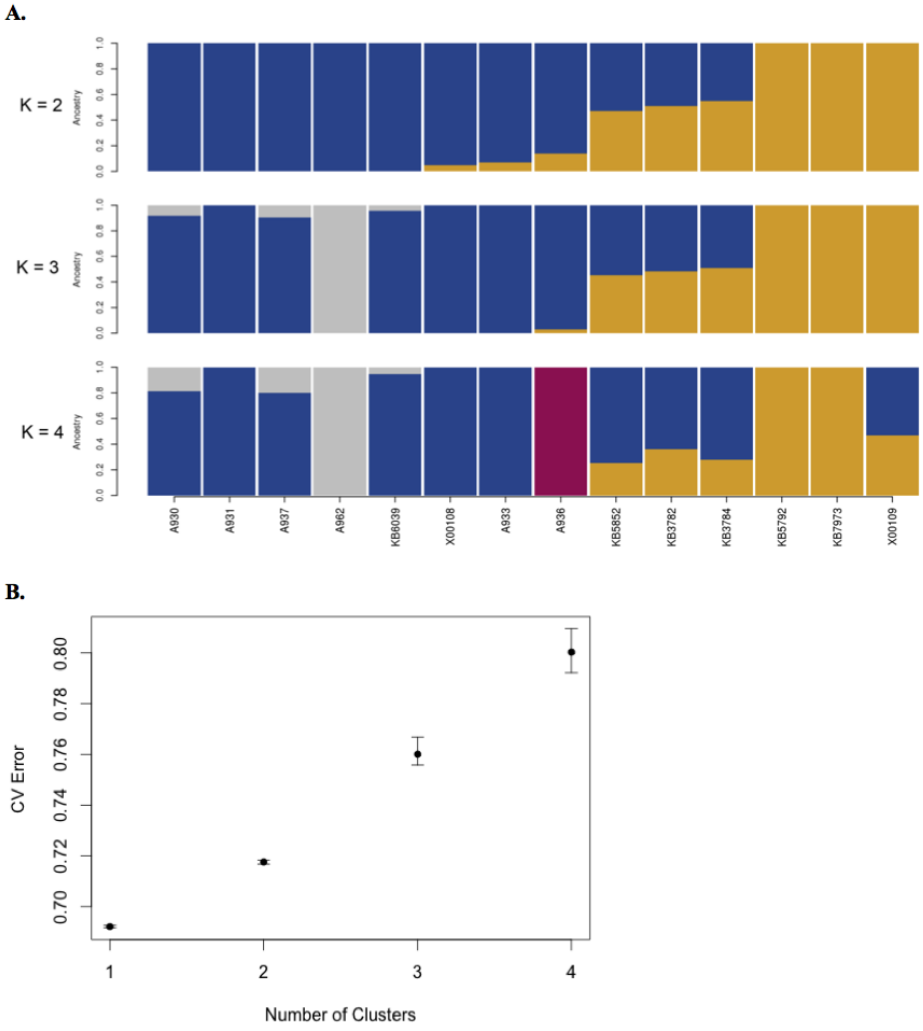

Figure 4.5: **Admixture results for 14 western lowland gorillas** A: Population membership inferred from ADMIXTURE with 2 - 4 populations. B: Cross Validation Error from ADMIXTURE analyses on three gorilla subspecies. CV error is estimated for 10 independent runs, and the results from the runs with the lowest error is shown in A. (data generated by Kimberly F. McManus)

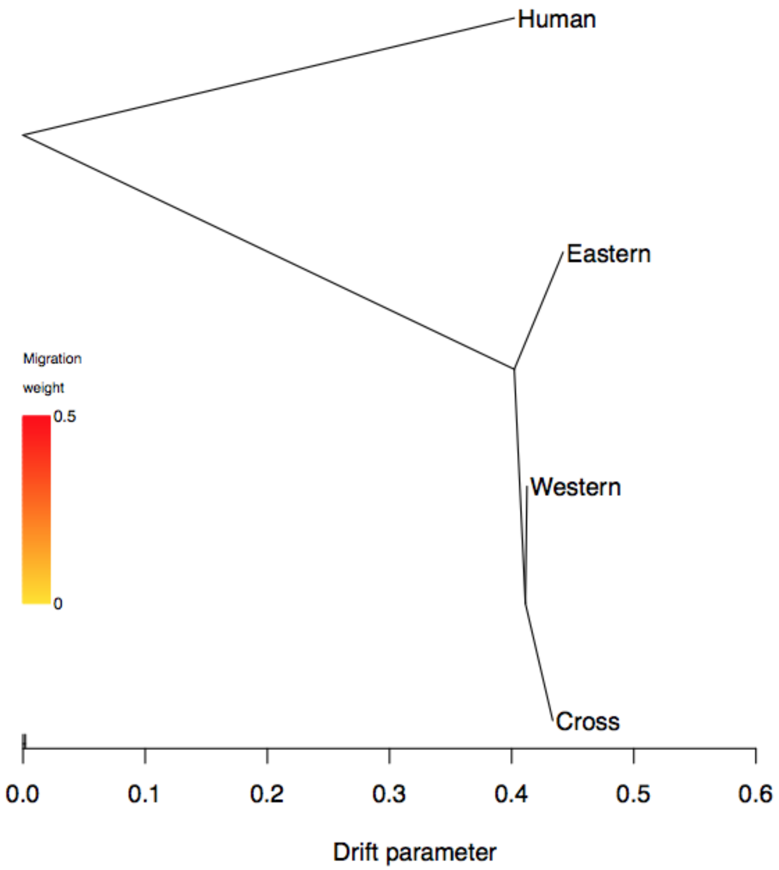

Figure 4.6: **Four-population phylogeny as inferred by TreeMix** The tree was inferred using reference-free genotypes obtained using BSNP at putatively unlinked neutrally evolving loci. (data generated by Kimberly F. McManus)

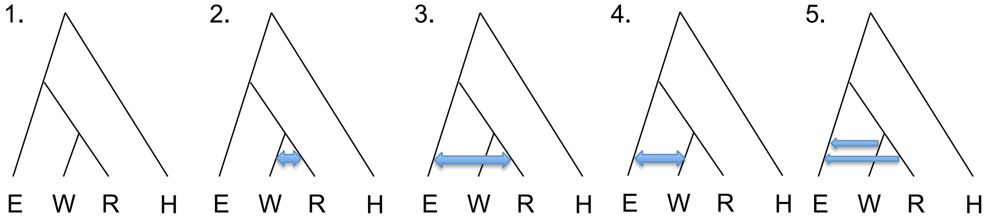

Figure 4.7: **Phylogeny and explored migration bands for G-PhoCS analysis.** We used the indicated phylogeny for eastern lowland (E), western lowland (W), Cross River (R) gorilla species and human (H), and tested the indicated migration scenarios. Scenario 1: no migration. Scenario 2: bi-directional migration between western and Cross River gorilla. Scenario 3: bi-directional migration between Cross River and eastern gorilla. Scenario 4: bi- directional migration between western and eastern gorilla. Scenario 5: migration from western to eastern gorilla and from Cross River to eastern gorilla.

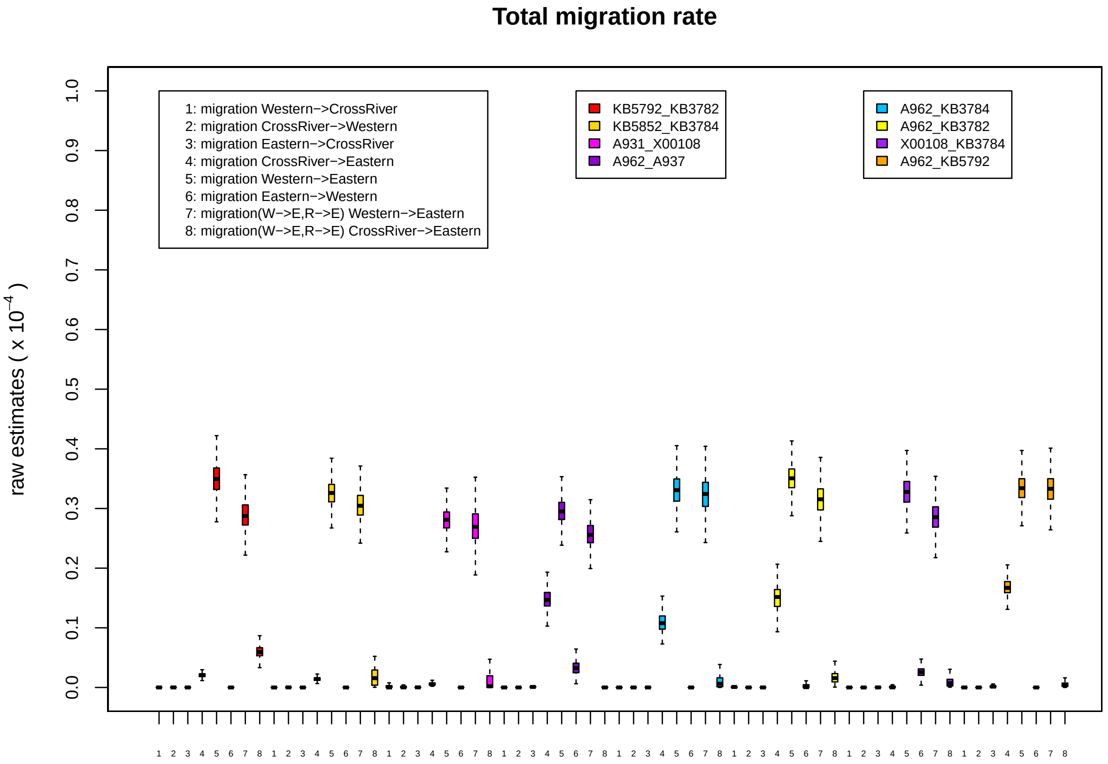

Figure 4.8: **Total migration rate inferred by G-PhoCS, given by $m_{AB} \times \tau_{AB}$.** Along the x-axis values from different migration bands associated with different scenarios (Figure 1) are depicted. The first six migration rate estimates come from a bi-directional migration setting. The last setting is with two migration bands, one from western lowland to eastern lowland gorilla and another from Cross River to eastern lowland gorilla. A strong signal of migration of western to eastern lowland gorillas is observed. Colors correspond to different combinations of western gorilla samples included in the analysis.

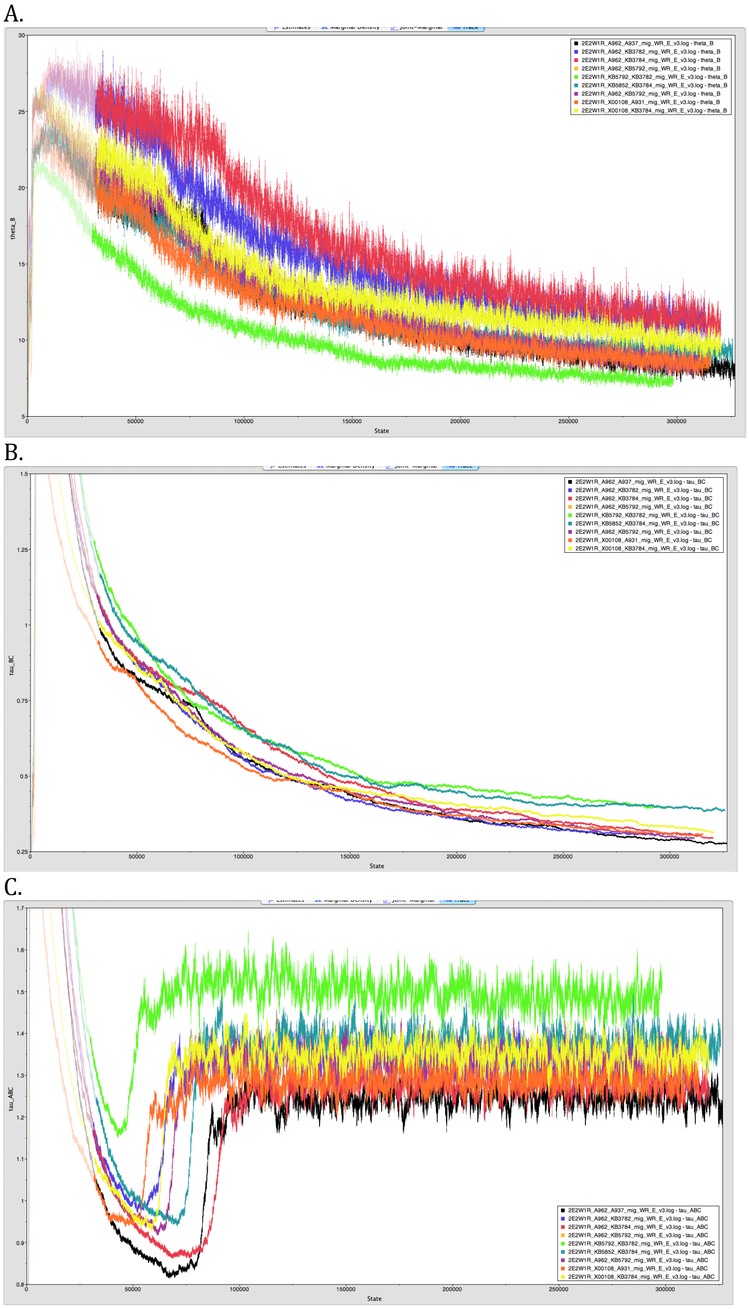

Figure 4.9: **MCMC traces for three demographic parameters inferred using G-PhoCS.** Each trace plot (obtained using Tracer) displays 9 combinations of samples including 2 eastern gorillas, 2 western gorillas and 1 cross river gorilla with migration setting scenario5 (Figure1). A. Western gorilla population size. B. Western-cross river gorilla split time C. Eastern-Western-Cross River ancestral split time.

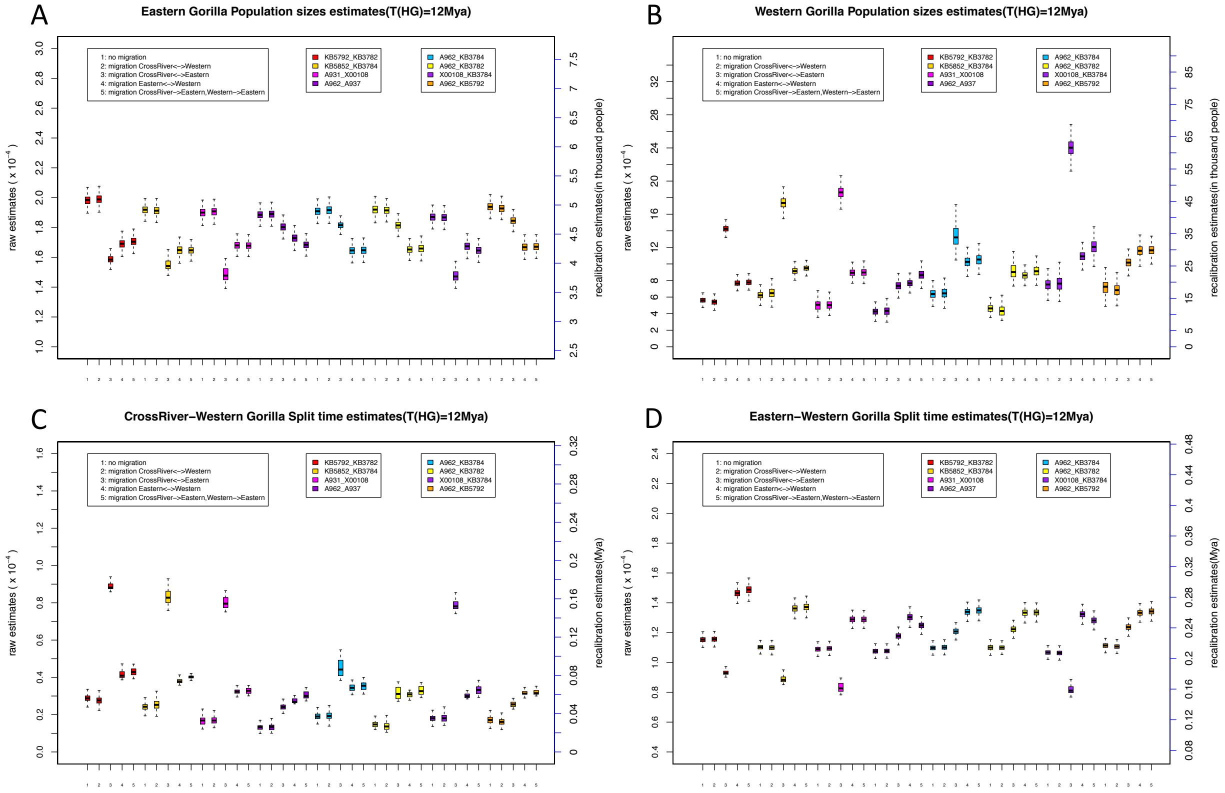

Figure 4.10: **Demographic parameters inferred by G-PhoCS.** We include 2 eastern gorilla individuals (Victoria, 9732), different combinations of Western individuals, 1 Cross River gorilla (B646) and always include human reference as an outgroup. Values along the x-axis correspond to five different scenarios depicted in Figure 1. Raw estimates are depicted on the left axis scale, and recalibrated values (effective population size) assuming a human-gorilla divergence time of 12 Mya are depicted on the right axis. A. Eastern gorilla population size B. Western gorilla population size. C. Western-cross river gorilla split time D. Eastern-Western-Cross River ancestral split time.

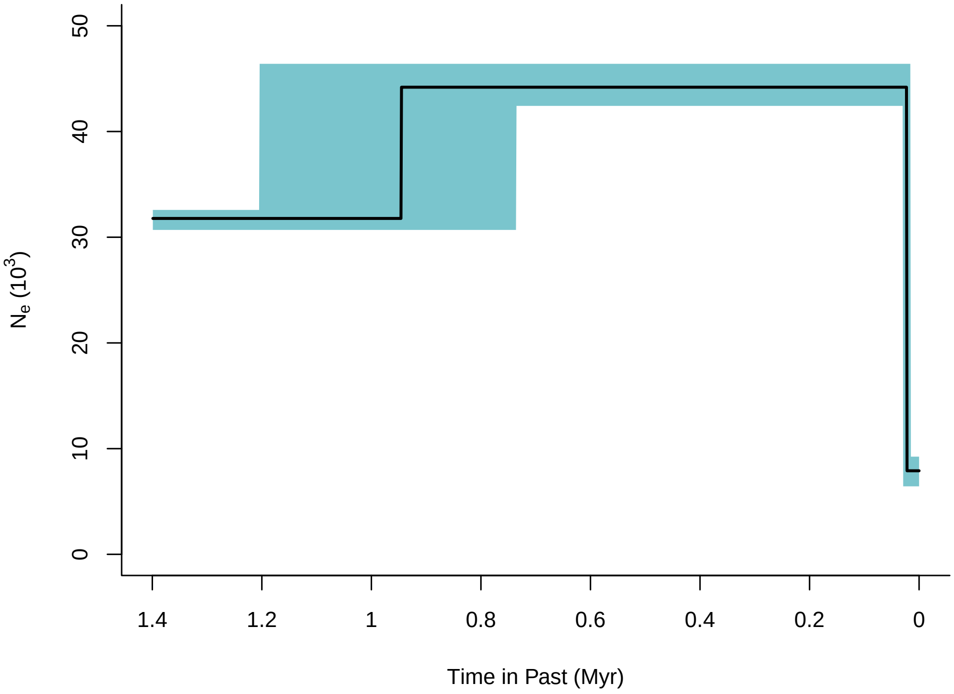

Figure 4.11: **Inferred best-fit demographic model of western lowland gorillas.** Shading represents confidence intervals determined by bootstrapping. Fitted parameters are depicted assuming a mutation rate of 1.1×10^{-8} per base pair per generation and a generation time of 19 years. (data generated by Kimberly F. McManus)

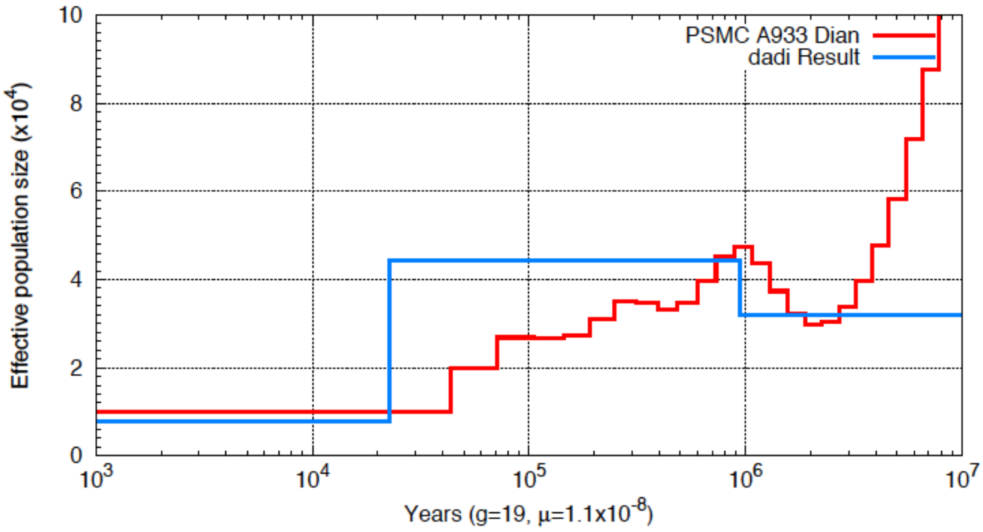

Figure 4.12: **Comparison of $\partial a \partial i$ and PSMC Models.** The red line indicates population history obtained for a western lowland gorilla using PSMC based on mapping to the gorilla genome and scaled using generation time and mutation rates employed in this paper. The blue line corresponds to the best-fit three-epoch model inferred for western lowland gorillas using $\partial a \partial i$.

Tables

Table 4.1: **Gorilla population history estimates.** Population history estimates by using G-PhoCS when assuming a range of human-gorilla divergence time (8, 10, and 12 Mya). We assumed migration events from western lowland to eastern lowland gorilla and from Cross River to eastern lowland gorilla (Figure 1 scenario 5). Values in parentheses correspond to 95% credible intervals.

	Human-Gorilla Divergence Time (Mya)		
	8	10	12
Mutation rate per generation	1.461	1.169	0.974
without CpG ($\times 10^{-8}$)	(1.456-1.466)	(1.165-1.173)	(0.970-0.978)
Eastern Gorilla population size	2.853	3.566	4.280
($\times 10^{3}$)	(2.755-2.956)	(3.443 -3.696)	(4.132-4.435)
Western Gorilla population size	16.774	20.967	25.161
($\times 10^{3}$)	(13.114 -21.439)	(16.393-26.798)	(19.672-32.158)
Cross River Gorilla population size	2.054	2.567	3.080
($\times 10^{3}$)	(2.352-2.755)	(2.940-3.443)	(3.529-4.132)
Western-Cross River ancestral	20.462	25.578	30.693
population size ($\times 10^{3}$)	(17.294-24.191)	(21.617-30239)	(25.940-36.287)
Gorilla ancestral population size	26.500	33.126	39.751
($\times 10^{3}$)	(25.829-26.965)	(32.286-33.706)	(38.743-40.447)
Human-Gorilla ancestral	45.472	56.840	68.208
population size($\times 10^{3}$)	(44.349-46.608)	(55.437-58.259)	(66.524-69.911)
Western – Cross River split time	0.046	0.057	0.068
(Mya)	(0.038-0.056)	(0.048-0.070)	(0.057-0.084)
Eastern-Western split time	0.174	0.218	0.261
(Mya)	(0.161-0.194)	(0.201-0.243)	(0.242-0.292)

Table 4.2: **Western lowland gorillas demographic history estimates from $\partial a \partial i$.** With P1 first population size change, T1 length of bottleneck, P2 second size change, and T2 time of second size change. For the conversion, a mutation rate of 1.1×10^{-8} mutations per base pair per generation and a 19-year generation time were used. The total number of callable sites is 812,645,853. (data generated by Kimberly F. McManus)

Demographic Model	θ	P1	T1	P2	T2	Log-likelihood	AIC
Standard Neutral	1,167,204					-60,420	120,840
Exponential Growth	1,299,805	0.09	0.009			-6,222	12,448
	(36,352)	(3,272)	(12,432)				
Bottleneck then Exponential Growth	1,181,405	39.54	0.33	0.32	-578	1,162	
	(33,040)	(1,306,416)	(10,903)	(401,771)			
Two Epochs	1,297,300	3.4e-13	1.2e-14			-5,654	11,312
	(36,282)	(0)	(0)				
Three Epochs	1,136,249	1.391	0.785	0.249	0.019	-473	954
	(31,777)	(44,190)	(946,129)	(7,905)	(22,842)		

Table 4.3: **Comparison of published estimates of gorilla population split times.** Values from each study are adjusted to match the generation time and mutation rates employed in this study. A mutation rate adjustment is not performed for Thalmann et al (2011), which is based on microsatellites. We note that Thalmann et al (2007) report a range of split time values under several different models, with population split times ranging from 78 kya (without subsequent gene flow) to 1.6 mya, with some models including Eastern to Western migration.

Paper	Method	Data	Cross River-Western		Western-Eastern		Migration
			Original	Adjusted	Original	Adjusted	
Thalmann et al, 2011	ABC	8 microsatellite loci	17.8				4.5 migrants per generation, symmetric
Becquet et al, 2007	MIMMAR	15 loci			92	171	M=0.87, symmetric
Scally et al, 2012	Custom-IM	Genome wide, reduced representation			500	429	0.2 migrants per generation, symmetric
Prado-Martinez et al, 2013	PSMC	Genome wide	80	114	150	214	none
Mailund et al, 2012	CoalHMM	95 loci (10Mbp interval between locus)				450	gene flow ended 150kyr ago
Thalmann et al, 2007	IM	16 loci			78*	107	more gene flow from eastern to western gorilla after initial split at around 0.9-1.6 mya, no gene flow until 78 kyr ago
This study	G-PHOCS	25,573 loci		68		261	Western to Eastern, 0.37 migrants per generation

Table 4.4: **Comparison of D-statistics.** D-statistics were calculated for combinations of human, eastern lowland, western lowland, and Cross River gorilla samples based on published hg18 SNP call sets (A) and SNPs defined using BSNP based on mapping to the gorilla reference genome (B). All analysis was limited to the autosomes. For hg18 SNP call set, additional values are shown when analysis is restricted to sites passing the all-sample coverage level 5 and 8 masks defined in Prado-Martinez et al. For BSNP call sets, either all sites (limited to sample genotypes with at least 5 reads), sites located within callable genome mask defined for the analysis of 14 western gorilla samples using read depth and mapping quality criteria, or all sites located within defined syntenic regions of alignment between gorGor3.1 and hg19 were used. For BSNP call sets, the human outgroup allele was determined based on the 'net.axt' alignments available from the UCSC genome browser.

SNPs from hg18 mapping in Prado-Martinez et al				All		Coverage 5 Mask		Coverage 8 Mask	
Human	Eastern	Western	Cross River	D	Z	D	Z	D	Z
hg18	9732_Mkubwa	A934_Delphi	B646_Nyango	0.031	8.656	0.058	16.098	0.058	15.353
hg18	9732_Mkubwa	A962_Amani	B646_Nyango	0.016	4.372	0.019	5.063	0.017	4.357
hg18	9732_Mkubwa	A931_Banjo	B646_Nyango	0.029	7.734	0.025	6.31	0.022	5.173
hg18	9732_Mkubwa	KB3784_Dolly	B646_Nyango	0.028	7.312	0.031	8.071	0.029	7.161
hg18	Victoria	A934_Delphi	B646_Nyango	0.03	8.094	0.057	15.352	0.056	14.317
hg18	Victoria	A962_Amani	B646_Nyango	0.018	4.787	0.021	5.581	0.019	4.824
hg18	Victoria	A931_Banjo	B646_Nyango	0.028	7.685	0.025	6.251	0.021	4.99
hg18	Victoria	KB3784_Dolly	B646_Nyango	0.024	6.405	0.028	7.069	0.026	6.305
SNPs using BSNP based on gorGor3 Mapping				All (DP5)		Western Callable Mask		Syntenic Regions	
Human	Eastern	Western	Cross River	D	Z	D	Z	D	Z
hg19	9732_Mkubwa	A962_Amani	B646_Nyango	0.004	1.21	0	-0.047	0.003	0.673
hg19	9732_Mkubwa	KB3784_Dolly	B646_Nyango	0.007	2.142	0.01	2.551	0.007	2.015
hg19	Victoria	A962_Amani	B646_Nyango	0.003	0.978	0.001	0.199	0.003	0.756
hg19	Victoria	KB3784_Dolly	B646_Nyango	0.002	0.61	0.005	1.258	0.001	0.375

Bibliography

Ackermann, R. R. and Bishop, J. M. Morphological and molecular evidence reveals recent hybridization between gorilla taxa. *Evolution*, 64(1):271–290, 2010.

Alexander, D. H., Novembre, J. and Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

Anthony, N. M., Johnson-Bawe, M., Jeffery, K., Clifford, S. L., Abernethy, K. A., Tutin, C. E., Lahm, S. A., White, L. J., Utley, J. F., Wickings, E. J. et al. The role of pleistocene refugia and rivers in shaping gorilla genetic diversity in central africa. *Proceedings of the National Academy of Sciences*, 104(51):20432–20436, 2007.

Becquet, C. and Przeworski, M. A new approach to estimate parameters of speciation models with application to apes. *Genome research*, 17(10):1505–1519, 2007.

Consortium, . G. P. et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

Edgar, R. C. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

Fünfstück, T., Arandjelovic, M., Morgan, D. B., Sanz, C., Breuer, T., Stokes, E. J., Reed, P., Olson, S. H., Cameron, K., Ondzie, A. et al. The genetic population structure of wild western lowland gorillas (gorilla gorilla gorilla) living in continuous rain forest. *American journal of primatology*, 76(9):868–878, 2014.

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

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. and Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics*, 43(10):1031–1034, 2011.

Grubb, P., Butynski, T. M., Oates, J. F., Bearder, S. K., Disotell, T. R., Groves, C. P. and Struhsaker, T. T. Assessment of the diversity of african primates. *International Journal of Primatology*, 24(6):1301–1357, 2003.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. and Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet*, 5(10):e1000695, 2009.

Hernandez, R. D., Williamson, S. H., Zhu, L. and Bustamante, C. D. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher gc-content in humans. *Molecular biology and evolution*, 24(10):2196–2202, 2007.

Hobolth, A., Christensen, O. F., Mailund, T. and Schierup, M. H. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS Genet*, 3(2):e7, 2007.

Hwang, D. G. and Green, P. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(39):13994–14001, 2004.

Jolly, D., Taylor, D., Marchant, R., Hamilton, A., Bonnefille, R., Buchet, G. and Riollet, G. Vegetation dynamics in central africa since 18,000 yr bp: pollen records from the interlacustrine highlands of burundi, rwanda and western uganda. *Journal of Biogeography*, 24(4):492–512, 1997.

Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., Inoue, E., Inoue-Muruyama, M., Mitani, J. C., Muller, M. N. et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences*, 109(39):15716–15721, 2012.

Le Gouar, P. J., Vallet, D., David, L., Bermejo, M., Gatti, S., Levréro, F., Petit, E. J. and Ménard, N. How ebola impacts genetics of western lowland gorilla populations. *PLoS One*, 4(12):e8375, 2009.

Mailund, T., Halager, A. E., Westergaard, M., Dutheil, J. Y., Munch, K., Andersen, L. N., Lunter, G., Prüfer, K., Scally, A., Hobolth, A. et al. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genet*, 8(12):e1003125, 2012.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

Nsubuga, A. M., Holzman, J., Chemnick, L. G. and Ryder, O. A. The cryptic genetic structure of the north american captive gorilla population. *Conservation Genetics*, 11(1):161–172, 2010.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. and Reich, D. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

Patterson, N., Price, A. L. and Reich, D. Population structure and eigenanalysis. *PLoS genet*, 2(12):e190, 2006.

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

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G. et al. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, 2013.

Rambaut, A. and Drummond, A. Tracer v1. 5 available from http://beast. bio. ed. ac. uk/tracer. 2013.

Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M. et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639, 2010.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T. et al. Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388):169–175, 2012.

Schiffels, S. and Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919–925, 2014.

Sudmant, P. H., Huddleston, J., Catacchio, C. R., Malig, M., Hillier, L. W., Baker, C., Mohajeri, K., Kondova, I., Bontrop, R. E., Persengiev, S. et al. Evolution and diversity of copy number variation in the great ape lineage. *Genome research*, 23(9):1373–1382, 2013.

Thalmann, O., Fischer, A., Lankester, F., Pääbo, S. and Vigilant, L. The complex evolutionary history of gorillas: insights from genomic data. *Molecular biology and evolution*, 24(1):146–158, 2007.

Thalmann, O., Wegmann, D., Spitzner, M., Arandjelovic, M., Guschanski, K., Leuenberger, C., Bergl, R. a. and Vigilant, L. Historical sampling reveals dramatic demographic changes in western gorilla populations. *BMC evolutionary biology*, 11(1):85, 2011.

Walsh, P. D., Abernethy, K. A., Bermejo, M., Beyers, R., De Wachter, P., Akou, M. E., Huijbregts, B., Mambounga, D. I., Toham, A. K., Kilbourn, A. M. et al. Catastrophic ape decline in western equatorial africa. *Nature*, 422(6932):611–614, 2003.

CHAPTER V

Evolutionary and demographic history of canine species

5.1 Introduction

The origins and dynamics of dog (*Canis familiaris*) domestication have long been an interesting and controversial questions for both geneticists and archaeologists. As human's first domesticated friend, understanding dog domestication is crucial to our understanding of human evolution and interactions of animals. Although it is widely accepted that dogs were domesticated from a gray wolf (*Canis lupus*) (Vilà et al., 1997), the location and timing of such domestication and the specific wolf population that dogs are domesticated from are still largely debated (Perri, 2016; Horard-Herbin et al., 2014). Savolainen et al. (2002) used markers on mtDNA to suggest that dogs originated from Southeast Asia 15,000 years ago. Since the first whole genome sequence of dog is released (Lindblad-Toh et al., 2005), the analysis of whole genome sequencing data have suggested different potential regions of dog domestication. VonHoldt et al. (2010) showed that dog breeds shared a higher proportion of multi-locus haplotypes unique to grey wolves from the Middle East, suggesting that they are a dominant source of genetic diversity. Shannon et al. (2015) proposed Central Asia as a domestication origin based on LD-decay curves: LD is lowest in Afghanistan and Central Asia at short inter-SNP distances. Wang et al. (2016) included dogs from southern East

The work in Chapter V is submitted as Laura R. Botigué*, **Shiya Song***, Amelie Scheu*, Shyamalika Gopalan, Amanda L. Pendleton, Matthew Oetjens, Angela Taravella, Timo Seregély, Andrea Zeeb-Lanz, Rose- Marie Arbogast, Dean Bobo, Kevin Daly, Martina Unterländer, Joachim Burger, Jeffrey M. Kidd, Krishna R. Veeramah.(2016). Genomic analysis of Neolithic dogs from Central Europe. (*In review*). (*equal contribution)

Asia and suggested they have higher genetic diversity compared to other populations and are the basal group relating to gray wolves, indicating southern East Asian origin of domestic dogs 33,000 years ago. The analysis of complete mitochondrial genomes of ancient canids suggest that dogs are domesticated in Europe since all modern dogs are phylogenetically most closely related to either ancient or modern canids of Europe (Thalmann et al., 2013). This study and the analysis of an ancient Siberian wolf genome (Skoglund et al., 2015) demonstrate that analyzing ancient as well as contemporary canids genomes are crucial to integrate both genetic and paleogenetics evidence.

Europe has been a critically important region in the history and evolution of dogs, with most modern breeds having a common European ancestry (Parker et al., 2004). Furthermore, the oldest remains that can be unequivocally attributed to domestic dogs (Canis lupus familiaris) are found on the continent, including an Upper Paleolithic 14,700-year-old jaw-bone from the Bonn Oberkassel site in Germany (Benecke, 1987), although older specimens from Siberia and the Near East have been proposed (Perri, 2016).

Intriguingly, multiple studies have found evidence of a striking prehistoric turnover of mtDNA lineages in the European continent sometime between the Late Neolithic and today, with haplogroup C, which appears in almost all Neolithic dogs but less than 10% of modern dogs, being replaced by haplogroup A in most of Europe (Thalmann et al., 2013; Deguilloux et al., 2009). Frantz et al. (2016) argue that this matrilineal turnover was a consequence of a major population replacement by analyzing genomic data from modern dogs as well as a Late Neolithic (5,000 years) Irish dog, Newgrange (henceforth referred to as NGD). Placing their results within the context of existing archaeological data, they also posited a novel dual origin for domestication. However, NGD shares ancestry primarily with modern European dogs, implying the proposed population replacement had largely occurred by this time.

The characterization of samples from earlier in the Neolithic and from continental Europe are therefore necessary to examine whether and to what extent a large-scale replacement occurred on the continent during this period, which would be evidenced by the presence of a distinct ancestry not present in modern dog genomes, as opposed to genomic continuity from the Early Neolithic to today. This, in turn, is key for understanding human-dog interactions during the major replacement of indigenous Paleolithic hunter-gatherers by Neolithic farmers from Anatolia (Bramanti et al., 2009; Hofmanová et al., 2016), and the subsequent migrations from the Eastern European steppe (Haak et al., 2015), as well as for disentangling the process of dog domestication.

We present analysis of 9x coverage whole genomes of two dog samples from Germany dating to the Early and End Neolithic and demonstrate genetic continuity throughout this era as well as substantial shared ancestry with modern European dogs. However, we also find that the End Neolithic sample possesses an additional ancestry component consistent with admixture from a population of dogs found further east which may have migrated concomitant with people originating from the steppe associated with Late Neolithic and Early Bronze age cultures, such as Yamanya and Corded Ware. We also show that most autosomal haplotypes previously associated with domestication were already established in our Neolithic dogs, though adaptation to a starch-rich diet likely occurred later. Our results are consistent with a single domestication process that occurred between 20,000-40,000 years ago.

5.2 Methods

5.2.1 Sequencing of two ancient dog genomes

HXH: A single petrous bone was identified in the internal ditch structure of Herxheim, an Early Neolithic site in Germany discovered in 1996 containing archaeological material from the Linearbandkeramik culture. Herxheim contains a significant amount of faunal remains, including >250 remains from dogs that constitute the largest bone series of Early Neolithic dogs in western Europe. A 14C dating of 5223-5040 cal. BCE (95.4 %) was estimated for the bone (Mams-25941: 6186+/-30, calibrated with OxCal 4.2 (Ramsey et al., 2009) using the IntCal13 calibration curve

(Reimer et al., 2013)). The petrous part of the temporal bone of sample HXH was prepared in clean-room facilities dedicated to ancient DNA in Trinity College Dublin (Ireland). DNA extraction was performed using a Silica column method as described in MacHugh et al. (2000). Two genomic libraries were prepared as described in Gamba et al. (2014). Screening of one library via an Illumina MiSeq run and mapping against various reference genomes demonstrated that reads for this sample mapped almost exclusively to the CanFam3 genome, revealing that it was a canid. Blank controls were utilized throughout.

CTC: The entire cranium of a dog was found in the Kirschbaumhöhle (Cherry Tree Cave) in the Franconian Alb, Germany. The cave was discovered in 2010 and contains human and animal remains from at least six prehistoric periods. CTC was an adult dog demonstrating morphological similarity to the so-called Torfhund (*Canis familiaris palustris*), and was found close to two human skulls dated to the early End Neolithic (2,800 - 2,600 BCE cal.). A 14C dating of 2,900-2,632 cal. BCE (95.4 %) was estimated for the cranium (Erl-18378: 4194+/-45, calibrated with OxCal 4.2 using the IntCal13 calibration curve). Sample preparation was conducted in dedicated ancient DNA facilities of the Palaeogenetics Group at Johannes Gutenberg-University Mainz under strict rules for contamination prevention as described in Bramanti et al. (2009). DNA was extracted independently twice from the petrous bone using a phenol-chloroform protocol31. A total of four double indexed genomic libraries were prepared as described in Hofmanová et al. (2016). One library was screened for endogenous DNA content via Illumina MiSeq sequencing, with 61.5% of reads mapping to CanFam3. Blank controls were utilized throughout.

Combinations of various genomic libraries from each ancient sample (CTC and HXH) were sequenced on two lanes of an Illumina HiSeq 2500 1TB at the New York Genome Center (NYGC) using the High Output Run mode to produce 2x125 bp paired-end reads. Reads were trimmed, merged and filtered using a modified version of the ancient DNA protocol described by Kircher (2012). Merged reads were then mapped using BWA aln (Li and Durbin, 2010) (maximum edit distance=1%, number of gap opens=2, l=16500) to a modified version of the CanFam3.1 reference genome containing a Y chromosome. Read groups were added using the PICARD tool AddOrReplaceReadGroups and duplicate reads were removed using MarkDuplicates. Finally, InDels were realigned using the GATK (McKenna et al., 2010) RealignerTargetCreator and IndelRealigner tools to produce a finished BAM file for each sample.

The proportion of reads mapping to the reference genome was >67% for both samples, confirming high endogenous DNA content identified during screening. MapDamage (Ginolhac et al., 2011) analysis demonstrated that both samples possessed characteristics typical of ancient DNA (Briggs et al., 2007) such as high numbers of 5' C>T and 3' G>A changes at the end of fragments (Figure 5.2), while fragment length was also small (mean 60-70bp). When examining a subset of 1 million random reads, only 3% of reads mapped to the hg19 reference genome for both samples, and almost all of these reads also mapped to CanFam3.1, which indicated very low levels of human contamination in our data. Therefore, we conclude that both our samples appear to contain substantial authentic canine ancient DNA. The mean coverage for both samples was 9x. Additionally, the mean coverage for the X and Y chromosomes was 5x for both samples, indicating they are males.

As well as the two ancient samples generated in this study, we also reanalyzed the ancient Irish Newgrange dog (henceforth known as NGD) described in Frantz et al. (2016). A BAM file containing only single ended unique mapped reads was provided by the authors of that study. In order to map to our modified version of CanFam3.1, reads were converted to fastq files using the bamtofastq function in Bedtools. All subsequent processing, beginning with mapping using BWA aln, were performed as described above for CTC and HXH.

5.2.2 Genotype calling for ancient samples and contemporary samples

We utilized distinct schemes to ascertain variants depending on the analysis being conducted. In all cases involving an ancient sample, we used a custom genotype caller implemented in Python (code available at https://github.com/kveeramah/aDNA_GenoCaller) that incorporates DNA damage patterns estimated from MapDamage Ginolhac et al. (2011) using the likelihood model described in Briggs et al. (2007). Briefly, damage patterns with respect to read position are fit with a Weibull distribution of the form $a \times \exp(-(x^c) \times b)$, where x is the proportion of damaged C>T or G>A bases at a particular position along the read (unlike Briggs et al. (2007), we find a slightly better fit with a Weibull than when assuming exponential decay). We then calculated the likelihood of each possible diploid genotype using a model that incorporates the possibility of both sequencing error and post-mortem damage (see Table S13 in Hofmanová et al. (2016) for the likelihood expression for each possible allele, which can then be averaged for two alleles to obtain the likelihood for a particular genotype).

However, rather than simply reporting the best likelihood we also incorporated additional hard filtering steps to produce final genotype calls. Firstly, any site with less than 7x coverage was reported as missing. In addition, any position where the highest likelihood is a heterozygote must have a minimum Phred-scaled genotype quality of 30 or the next highest homozygote likelihood genotype was chosen instead. We found that this practice eliminated many false positives that are likely the result of post-mortem damage, resulting in much more balanced numbers of C>T vs T>C and G>A vs A>G heterozygous reference to alternate allele changes compared to when using the standard GATK Unified Genotyper caller (Figure 5.3) (however, we note that when a site is already known to be segregating in other dogs or wolves, our algorithm and GATK Unified Genotyper are almost completely concordant for sites with >7x coverage). Base calls with a quality score less than 15 and reads with a mapping quality less than 15 were not included during genotype calling. Base calls with a quality score greater than 40 (which can occur during paired-end read merging) were adjusted to 40.

In addition to the three ancient samples, we examined whole genome sequence data from 96 modern canids. Additional genomes were generated using Illumina sequencing for a Great Dane and Iberian wolf. We also posted sequencing reads to the SRA for a Portuguese village dog, Chinese Mongolian Shepherd village dog and Afghan Hound. All remaining genome data were acquired from previously published datasets deposited on SRA. As above, all dog genome sequence data was aligned against a modified version of CanFam3.1 reference genome with unplaced contig sequences combined into a single chromosome sequence (separated by 200 'N' characters) and including a representation of the non-pseudo autosomal Y chromosome sequence 30 using BWA (Li and Durbin, 2010). PCR duplicates were removed by Picard v1.62 (http://broadinstitute.github.io/picard), reads in regions with candidate indels were locally realigned and base quality scores were recalibrated using GATK v3.4 (McKenna et al., 2010), resulting in a dataset that exhibited mean autosomal coverages of 5.53-44.74x. We generated GVCF files (Genomic VCF) with a record for every position in the genome using GATK HaplotypeCaller (GATK v3.4) (McKenna et al., 2010). These data were supplemented with genotype data for six canids from Freedman et al. (2014)(Basenji, Dingo, Golden jackal, Croatian wolf, Israeli wolf, and Chinese wolf).

We generated three different call sets with different ascertainment schemes. In each case, variant sites on the autosomes were identified in specific sets of individuals and then genotyped across all sequenced samples. Call set 1 includes variants discovered in 89 contemporary genomes for which we processed the sequencing reads, the 3 ancient genomes and 6 genomes obtained from the Freedman et al. (2014) call set based on Illumina and SOLiD data. This call set represents the most comprehensive set of variants, but may show biases due to differences in coverage among sample sets. Call set 2 only includes variants discovered in the three ancient genomes. Call set 3 only includes sites discovered as variable in New World wolves, and is the primary call set utilized for most analyses.

Call set 1-Comprehensive variants: This call set aims to include all variants from the 89 genomes (with bam files), 3 ancient genomes (with bam files) and Freedman's 6 genomes (with vcf files from Freedman). We first generated call sets from the three groups separately, took the union of the call sets, genotyped all the variants in each group, and then applied filtering accordingly. 1). We applied the GATK HaplotypeCaller to call variants (SNPs and indels) from the 89 genomes for which we had BAM files together. We applied a hard filter to remove sites that are within 5bp of an indel, and with MQ<25, QD<10, qual<33, mean DP<mean_read_depth/2, or mean DP>mean_read_depth x 2. The mean read depth for 89 genomes together is 879X. This set of variants contained 18.7 M SNPs. For the three ancient genomes, we applied the ancient DNA caller to discover variants in each ancient sample , using DP7, MQ15, BQ15, GQ30 as cut-off as described above. The variants from each ancient genome were merged, resulting in 5.8M SNPs. For Freedman's genomes, we took variants found in the vcf files from Freedman et al. 2). We took the union of these call sets and genotyped each variant in each group. For the 89 genomes, we genotyped those variants from 89 genomes together and applied a hard filter to remove sites with mean DP<mean_read_depth/2, mean DP>mean_read_depth x 2, or MQ<25. For the ancient genomes, we used the ancient DNA caller to genotype each variant, using DP7, MQ15, BQ15 as a cut-off. For Freedman's samples, we directly obtained genotype calls from Freedman's emit-all vcf files. The comprehensive SNP call set contains 24M SNPs. We additionally genotyped these SNPs in the Andean fox. 3). After removing sites with at least one missing genotype, we ended up with a final set of 7.4M SNPs.

Call set 2-Ascertained in ancient genomes: This call set only includes sites discovered in the three ancient genomes. 1). We applied the ancient DNA caller to discover variants in each ancient genome, using DP7, MQ15, BQ15, GQ30 as cut-off as described above. The variants from each ancient genome were merged, resulting in 5.8M SNPs. 2). We applied the GATK UnifiedGenotyper to genotype these variants in other contemporary dog genomes and applied the aDNA caller to genotype those variants using DP7, MQ15, BQ15 as cut-off in each ancient genome. We directly obtained genotype calls from Freedman's emit-all VCF files. We additionally genotyped these SNPs in the Andean fox. 3). After merging VCF files and restricting to no missing genotypes, we ended up with a final set of 1.9M SNPs

Call set 3-Ascertained in New World wolves: This call set is designed to include only sites that are variables in New World wolves, a group that is sister to Old World wolves and dogs (Fan et al., 2016). 1). We chose three New World wolves (glw, ysa, mxa), each with 20X coverage. We applied HaplotypeCaller implemented in GATK to call variants (SNPs and indels) from the three genomes together and applied a hard filter to remove sites that are within 5bp of indel, MQ<25, QD<10, qual<33, mean DP<mean_read_depth/2, or mean DP>mean_read_depth x 2. The mean read depth for the three genomes together is 76X. We additionally only keep variants with an alternative allele count of 1-5, resulting in 8.4M SNPs. 2).We applied GATK UnifiedGenotyper to genotype those variants in other contemporary dog genomes and applied the aDNA caller using DP7, MQ15, BQ15 as cut-offs to genotype those variants in each ancient genome. We directly obtained genotype calls from Freedman's emit-all vcf files and only retained variants that passed Freedman's Genome Filter (repeat divergence greater than or equal to 25, no CNV, MA, CpG). We additionally genotyped those SNPs in the Andean fox. 3). After merging VCF files and restricting to no missing genotypes, we ended up with 1.8M SNPs. 4). Ignoring the impact of recurrent mutation and post-divergence gene flow, this call set includes only mutations that either are private to New World wolves or occurred in the ancestral population of New World wolves, Old World wolves and dogs and thus have an essentially unbiased ascertainment with respect to Eurasian dog and wolf populations.

We also examined available SNP array data. Canine SNP array datasets were obtained from Shannon et al. (2015) and Pilot et al. (2015). Genotypes were also supplemented by data from the six canids reported in Freedman et al. (2014). We used GATK GenotypeCaller (McKenna et al.,

2010) to obtain genotype calls of SNP array loci from the whole genome vcf files (see methods above). For genomes from Freedman et al. (2014) genotype calls were obtained directly from emit-all vcf files. We only included calls that passed sample filters (GQ greater than 20, DP less than 2*genome-wide depth, DL=0, and DV greater than 5). After removing sites with more than 5% missing data across individuals our final SNP array dataset consisted of genotypes at 128,743 autosomal SNPs.

5.2.3 Mitochondrial Analysis

Ancient sample mtDNA consensus sequences were aligned to the canid alignment from Thalmann et al. (2013); Duleba et al. (2015), which contain whole mtDNA genomes for both modern and ancient canids. For measurements of nucleotide diversity and phylogenetic reconstruction, we only used positions where genotypes were called across all samples. After removing missing sites, our mtDNA alignment consisted of 14,936 nucleotide positions and 616 variant sites. A NJ tree was built with a TN93 substitution model (500 bootstraps) using MEGA 6.06 (Tamura et al., 2013).

5.2.4 Population Structure

Principle component analysis was performed on both the SNP array data set and all genome SNP Call sets using smartpca, part of the EIGENSOFT package version 3.0 (Patterson et al., 2006). Both diploid and pseudo-haploid genotype calls with and without C<->T and G<->A SNPs (the most likely sites to undergo post-mortem damage) were used to construct the PCA, but little difference was found amongst these analyses.

SpaceMix (Bradburd et al., 2016) was used to create a geogenetic map and infer potential long-distance admixture events across this map using the SNP array data, allowing only SNPs separated by at least 100kb and no more than five individuals per population. The method provides a similar description of isolation-by-distance-based population structure as PCA but also has the added advantage of visualizing major deviations of increased covariance that are likely to reflect long-distance admixture events. SpaceMix requires fitting a complex parameter space and thus utilizes an Markov Chain Monte Carlo (MCMC) search to find the most likely model parameters and also estimate confidence intervals. We ran 10 initial burn-ins of 100,000 generations to identify the best starting position for the MCMC chain, followed by a single long run for 10,000,000 generations, which appeared to be sufficient for the chain to stabilize, sampling every 1,000 generations. We performed eight separate iterations of this process and present results from the run with the highest peak probability. Latitude and Longitude coordinates for each population's country of origin was also provided to inform priors for the fitting of geogenetic space.

ADMIXTURE v.1.22 (Alexander et al., 2009) was used to perform an unsupervised clustering analysis on the SNP array data for the ancient dogs and a subset of 105 modern dogs that provided a global representation of dog structure, while NGSadmix(Skotte et al., 2013) was used to perform a similar analysis for the genome SNP data while taking into account genotype uncertainty by examining genotype likelihoods. Cross validation was performed for the ADMIXTURE analysis to identify the most appropriate number of clusters, K.

Neighbor-joining trees were constructed for the whole genome SNP set ascertained in the New World wolves (Call set 3) using the ape R package (Paradis et al., 2004) using distance matrices based on the metric of sequence divergence from Gronau et al. (2011). One hundred bootstrap replicates were generated by dividing the genome into 5 cM windows and sampling with replacement in order to determine node support.

5.2.5 f-statistics analysis

We estimated the relative genetic similarity of the three ancient dogs to each other and various modern dog and wolf populations by calculating an outgroup f_3-statistic (Patterson et al., 2012) of the form $F_3(C : A, B)$. Assuming a simple three population model with no post-divergence admixture, where population C is an outgroup to A and B, the value of this statistic will reflect the amount of shared drift between A and B relative to C. If one population (e.g. B) is kept constant, then introducing different populations to represent A will provide relative estimates of genetic similarity from B. We set either CTC, HXH or NGD as population B and applied this test to both the SNP array data (with Jackal set as an outgroup and population A set as a Old World village dog or Wolf population, with individuals from the same population merged) and the genome data (Call set 3 with Andean fox set as an outgroup and population A set as individual Village dog samples, plus Basenji and Dingo). f_3 statistics were estimated using the qp3pop program (modified to allow more than 22 autosomes) found within the Admixtools software package. Standard errors were estimated using a weighted block jackknife as previously described in Patterson et al. (2012).

We calculated D-statistics of the form D(A,B,Wolf,Outgroup) using the qpDstat function in Admixtools (Patterson et al., 2012) to examine potential differences in dog-wolf admixture between the three ancient dogs.We applied the f_4-ratio test (qpF4ratio in AdmixTools package) to estimate ancestry proportions in an admixed population for our ancient and modern dogs. Consider a topology where X is admixed from two source populations that are ancestral to B and C population, with proportions α and $1 - \alpha$, respectively. The admixture proportion α is calculated as the ratio between $f_4(A, O; X, C)$ and $f_4(A, O; B, C)$. Population A is more closely related to population B, and population O is an outgroup.

Both MixMapper (Lipson et al., 2013) and ADMIXTUREGRAPH (Patterson et al., 2012) were used to perform model-based inference of specific admixture events involving the three ancient dogs for both the SNP array and whole genome SNP datasets. Significance was assessed using a weighted block jackknife procedure. Genetic map positions for each SNP used in these analyses were inferred from Auton et al. (2013).

MixMapper (Lipson et al., 2013) provides a method for inferring admixture events within the context of a population tree based on f-statistics. The approach is similar in principle to TreeMix in that a bifurcating tree is first fit to allele frequency correlations and migration is then inferred on top of the tree if it improves the fit of the data. We applied the MixMapper approach to both the SNP array data and genome SNP data (Call set 3). First, we used the f_3 statistic to determine which populations were likely to be admixed. Again, this test statistic is of the form $F_3(C : A, B)$. However, when C is not chosen specifically to be an outgroup, any phylogeny where population C is admixed between two populations descended from A and B will have some portion where allele frequencies are negatively correlated, which can result in negative f_3 values. Z-scores for significant negative f_3 statistics were again calculated using a weighted block jackknife. All combinations of three way phylogenies were examined for the three ancient dogs, Old World modern village and free-breeding dogs, Dingo, Basenji, wolves (Old and New World), coyotes, golden jackal and Andean fox, with individuals grouped by population.

ADMIXTUREGRAPH (implemented in qpgraph in the Admixtools package) performs an approximate likelihood maximization to best fit f_2, f_3, f_4 statistics for all combinations of populations considered in a user-defined model of population demography that includes population split and admixture. MixMapper can be considered an automated version of ADMIXTUREGRAPH. However ADMIXTUREGRAPH can incorporate more complex models with multiple admixture events and ghost populations, providing the opportunity to fit all f-statistics (MixMapper only minimizes these within the context of one or two admixture events on a scaffold tree). Therefore we drew upon the various demographic inferences from the previous analyses to use ADMIXTUREGRAPH to find a model of population demography for our ancient and modern canids. We restrict our analyses to the genome data (Call set 3) and utilize the Andean fox. We consider a f-statistic to be fit if the inferred value in the model is within three standard errors of the estimated value from the data ($|Z - score|$>3). Note that occasionally best fit branch lengths between two nodes will have a value of 0. While these could be collapsed into descendent nodes, we generally keep them in the graph to clearly demonstrated the topology of the model (unless the admixture graph becomes to visually unwieldy), as they have no effect on the overall statistical fit of the model.

5.2.6 G-PhoCS and HXH divergence estimation

G-PhoCS (Gronau et al., 2011) is a full-likelihood-based method that uses independent loci in order to perform Bayesian coalescent-based inference of divergence times (τ), population diversity (θ) and, if specified, migration bands. To apply this method to our canid data and estimate when our ancient samples (and in particular our oldest sample, HXH) may have diverged from modern European samples, we performed a LiftOver on the 16,434 "neutral loci" (interspersed genomic segments of 1kb length) previously identified by Freedman et al. (2014). We then generated alignments at these loci that included the golden jackal, multiple Old World wolves (Israeli, Croatian, Chinese and Indian), several village (each from South China, India and Portugal) and breed dogs (Boxer and Dingo), and the two ancient samples. Genotypes for all samples were called from recalibrated BAM files using reads with mapping quality $\geq$ 15 and base quality $\geq$ 15 for modern samples, and calls with genotype qualities less than 30 and less than 7x coverage were marked as 'N'. Ancient samples were called as described above (see 5.2.2). We masked sites for all samples if any sample appeared to have a 'CpG' dinucleotide sequence and removed loci if any sample had complete missing data. As we tested many different combinations of samples and as G-PhoCS is computationally intensive, we often randomly selected 5,000 loci to perform the analysis. The final results presented in Figure 5.10 however are based on the full data set (16,434 loci), as well as select analyses where indicated. When the two ancient samples were included, we set the age of HXH to be 7,000 years old and the age of CTC to be 5,000 years old assuming the mutation rate 1×10^{-8} bp/generation and generation time = 3 years. For each run we ran 500,000 MCMC iterations and used Tracer (Rambaut and Drummond, 2013) to examine the chain convergence.

We built a NJ tree based on the neutral loci alignments on selected samples using the same method as described in 5.2.4. This tree and the global NJ tree were largely concordant (Figure 5.11). We ran G-PhoCS on the following tree structure, (((((Boxer, Village_Europe), Village_India), Village_ChinaS), ((wolfY, IsraeliWolf), ChineseWolf)),GoldenJackal), with wolfY either being the Croatian or Indian wolf. We then ran G-PhoCS with each ancient sample added separately, with HXH or CTC or NGD sister to the ancestral of Boxer and Village_Europe. For the migration bands, we set Village_ChinaS dog with the Chinese wolf, and other modern village dogs with the Israeli wolf, Village_India dog with the Indian wolf, the Israeli wolf with the golden jackal, and dog/wolf ancestors with golden jackal. Evidence for these migration bands either come from significant D statistics calculated by Admixtools (Table 5.2) or were previously identified by Freedman et al. (2014). For ancient samples, we added migration between South China village dogs and HXH, Indian village dogs and CTC, South China village dogs and NGD, and each wolf with the ancient samples. We found that 500,000 iterations were sufficient to establish convergence for all parameters and we sampled the the last 200,000 iterations to estimate the posterior distribution and calculate the 95% CI for each parameter.

Estimating the divergence time of HXH and NGD using G-PhoCS could potentially lead to large biases due to false positive singleton variants observed in HXH that are caused by post-mortem damage and the somewhat lower coverage of our ancient samples. Therefore we devised a method that would be robust to these issues. In particular, we utilized demographic parameters estimated by G-PhoCS (τ and θ) that describe the relationship between European, Indian and East Asian village dogs, and then inferred the HXH/NGD divergence time by using coalescent theory to predict the ratio of shared derived sites between European village dogs and HXH/NGD versus European and Indian Village dogs. We give the expectation considering a simple bifurcating tree for Europe, HXH/NGD and India, and the expectation assuming East Asian admixture into HXH/NGD.

We assume the population tree shown in Figure 5.8. Our main problem is false positives that appear as mutations on the branch leading to HXH/NGD after diverging from Europe. In order to limit this effect, we will condition on whether derived mutations (identified via an outgroup) present on a European chromosome but absent on an Indian chromosome are also present on one of the HXH/NGD chromosomes or not (110 sites). In addition, as HXH/NGD and Europe are sister clades versus India, there should be more such sites than those where mutations are shared between a European and Indian chromosome, but not a HXH/NGD chromosome (101 sites). This ratio should be approximately equal to the ratio of genealogies with these two topologies across the genome. To infer this we need to consider two types of parameters, the amount of population diversity, θ and the divergence time measured in expected numbers of mutations, τ. With these we can use coalescent theory to predict the relative number of 110 versus 101 genealogies.

There are two scenarios that would lead to 110 sites. In the first there is a coalescent event, C, between a HXH/NGD and European lineage between τ_1 and τ_2, which will depend on θ_1. Using standard coalescent theory 50, this will occur with the following probability:

$$P(C_{\tau_1<t<\tau_2}) = 1 - \exp^{-\frac{2}{\theta_1}(\tau_2-\tau_1)} \tag{5.1}$$

In addition, no coalescence could occur during the period, after which there is 1/3 chance that the next coalescent event will be between a HXH/NGD and European chromosomes (versus an European and Indian chromosome or a HXH/NGD and Indian chromosome). Thus the probability of observing a 110 compatible genealogy is:

$$P(110) = (1 - \exp^{-\frac{2}{\theta_1}(\tau_2-\tau_1)}) + (\exp^{-\frac{2}{\theta_1}(\tau_2-\tau_1)}/3) \tag{5.2}$$

Similarly, the probability of observing a 101 compatible genealogy is:

$$P(101) = (\exp^{-\frac{2}{\theta_1}(\tau_2-\tau_1)}/3) \tag{5.3}$$

The expected ratio is then simply P(110)/ P(101).

We have found that HXH/NGD demonstrate evidence of small but statistically significant admixture with a population resembling modern Southeast Asian dogs, which may decrease the number of occasions where HXH/NGD and Europe chromosomes coalesce before an Indian chromosome compared to the expectation above. We denote this admixture fraction, α, and $\beta = 1 - \alpha$. The probability of observing a 110 now depends on whether the chromosome chosen in HXH/NGD traces its ancestry through East Asia or not. If it does not, then the probability of P(110) is simply the same as above multiplied by β. If it does, then the probability of a 110 compatible genealogy depends on whether the European and Indian chromosome coalesced between 2 and 3, which will depend on θ_2 (i.e. essentially the reverse situation to before). Thus, the new total probability of a 110 genealogy is:

$$P(110) = \beta[(1 - \exp^{-\frac{2}{\theta_1}(\tau_2 - \tau_1)}) + (\exp^{-\frac{2}{\theta_1}(\tau_2 - \tau_1)}/3)] + \alpha[\exp^{-\frac{2}{\theta_2}(\tau_3 - \tau_2)}/3] \tag{5.4}$$

$$P(101) = \alpha[\exp^{-\frac{2}{\theta_1}(\tau_2 - \tau_1)}/3] + \beta[(1 - \exp^{-\frac{2}{\theta_2}(\tau_3 - \tau_2)}) + (\exp^{-\frac{2}{\theta_2}(\tau_3 - \tau_2)}/3)] \tag{5.5}$$

Using the whole genome SNP set, we computed the observed ratio of the number of SNPs where i) a European village dog and HXH have the derived allele and an Indian sample has the ancestral allele versus ii) a European and Indian village dog have the derived allele and HXH has the ancestral allele. We took the allele with the highest probability from one sample of each population. We tried using either golden jackal or andean fox to determine the ancestral allele. However, we found estimates in the former to be somewhat higher than in the latter. The golden jackal is known to be admixed with wolf populations (Freedman et al., 2014) while recurrent mutation on the longer andean fox lineage may cause underestimation. Therefore, we use the sites where the golden jackal and andean fox are concordant as the ancestral allele. We then find the value of τ_1 from our equations that is compatible with the observed 110/101 ratio. The two major advantages of our estimate are that a) it only depends on sites for which there is already evidence of a mutation in other higher coverage modern dogs (i.e. our genotype calling in ancient samples is likely to be much more accurate in such situations), and b) it uses only a single chromosome from each population (which can be randomly picked), and thus does not require calling heterozygotes accurately (i.e. it should not be sensitive to the lower coverage of our ancient samples). When comparing this expectation to real data it is is also assumed that sites (i.e. genealogies) are independent. To take into account dependence amongst linked sites we utilize a weighted block jackknife (Busing et al., 1999) using windows of 10cM to estimate confidence intervals.

We are interested in estimating the HXH/NGD-Europe split time, so τ_1. We sampled several parameters, namely Ne for European/Boxer ancestral population (θ_1) (we assume the Ne of the Boxer-European ancestral population is the same as that of the HXH/NGD-European ancestral population), Ne for European/Indian ancestral population (θ_2), time of divergence for Europe and Boxer (τ_0), time of divergence for Europe/India (τ_2) and time of divergence for Europe-India/Asia (τ_3) based on the estimates from G-PhoCS analysis, and also the percentage of HXH/NGD that is made up of Asian admixture (α) from the f_4-ratio analysis (Table 5.4). We then sampled τ_1 from a uniform distribution of (τ_0,τ_2). We computed the ratio P(110)/P(101) using the analytical formula 5.4 and 5.5 explained above and kept 1000 τ_1 estimates if the ratio fell into the confidence interval.

5.2.7 Analysis of loci associated with the domestication process

Coordinates of thirty putative “domestication loci” were obtained from Axelsson et al. (2013) and lifted over from CanFam2.0 to CanFam3.1 coordinates. Call set 1 SNPs within each window were extracted from the ancient samples and our genome sequence dataset. Eigenstrat genotype file formats were generated per window using convertf from the EIGENSOFT package (Price et al., 2006) and custom scripts were used to convert the genotype files into matrix formats for visualization using matrix2png (Pavlidis and Noble, 2003) using a filtered subset of SNPs (minor allele frequencies between 0.05 and 0.49) for easing visualization of the matrices. Altogether, the haplotypes of the three ancient samples were classified as either dog or wolf-like for 18 matrices that showed clear distinction between dog and wild canid haplotypes based on reference allele frequencies calculated per window.

Genomic copy-number was estimated from read depth as previously described (Sudmant et al., 2010; Alkan et al., 2009). Reads were split into non-overlapping 36-bp fragments and mapped to a repeat-masked version of the CanFam3.1 reference using mrsFAST (Hach et al., 2010), returning all read placements with two or fewer substitutions. Raw read depths were tabulated at each position and a loess correction for local GC content was calculated utilizing control regions not previously identified as copy number variable. The mean depth in 3kb windows was then calculated and converted to estimated copy-number based on the depth in the autosomal control regions.

5.3 Results

5.3.1 Sequencing of two Neolithic canid genomes from Germany

The older specimen (and the oldest whole nuclear genome sequenced dog to date), which we refer to hereafter as HXH, was found at the Early Neolithic site of Herxheim and is dated to 5,223-5,040 cal. BCE (7,000 years old) (Figure 5.1A). The younger specimen, which we refer to hereafter as CTC, was found in the Kirschbaumhöhle (Cherry Tree Cave) and is dated to 2,900-2,632 cal. BCE (4,700 years old), which corresponds to the End Neolithic period in Central Europe15 (Figure 5.1B). Both samples demonstrated substitution patterns and fragment lengths consistent with substantial post-mortem damage of ancient DNA (Figure 5.2). We used the model described by Hofmanová et al. (2016) to account for this post-mortem damage in downstream genotype calling, resulting in substantial reductions in C>T and G>A transitions compared with standard ancient DNA-non-aware genotype callers. In order to better understand how these Neolithic dogs are genetically related to modern dogs, we analyzed them within the context of a comprehensive collection of 5,649 canids, including breed dogs, village dogs and wolves previously genotyped at 128,743 SNPs (Shannon et al., 2015; Pilot et al., 2015), as well as 99 canid whole genomes sequenced at medium to high coverage (6-45x), including NGD. After exploring different ascertainment schemes for the whole genome data, we performed downstream analyses using 1,815,911 variants ascertained in New World wolves, as such mutations must either be private to this clade or arose in the Gray Wolf (*Canis lupus*) ancestral population, and thus will be essentially unbiased with regard to their ascertainment in Old World wolves and dogs.

5.3.2 Mitochondrial similarity amongst German Paleolithic and Neolithic dogs

The average sequencing depth for mtDNA was 179x, 208x and 170x in the CTC, HXH and NGD samples, respectively. A comprehensive pairwise comparison of nucleotide differences across the dataset showed that HXH and CTC mitochondrial sequences are more similar to each other (n = 5 differences) than to any other ancient canid or modern dog. Given the geographic proximity of the excavation sites of the HXH and CTC, a strong relationship between the mitochondrial haplotypes is not surprising. Furthermore, the low count of pairwise differences between the HXH and CTC suggests that random ancient artifacts are not a significant contribution to our call set of mitochondrial variants. HXH and CTC show higher sequence identities to NGD ancient dog than to any Thalmann sample: HXH has a slightly higher identity to NGD (n = 13 substitutions) than CTC dog (n = 18 substitutions). Interestingly, the most similar haplogroup in the Thalmann et al. (2013) dataset was the ancient dog-like sample from Germany (Germany 12.5 kya), which differed from the HXH and CTC dogs by 17 and 22 sites, respectively. A NJ tree built with a TN93 substitution model (500 bootstraps) of our alignment revealed that like other European Neolithic dogs examined previously, CTC and HXH mtDNA haplotypes are members of the C clade of modern dogs (Figure 5.4A). They are sister clades within sub-haplogroup C1, together with NGD and the Upper Paleolithic 12,500 year old Kartstein Cave dog (also from Germany). We note that Bonn Oberkassel also falls in the same haplogroup subclade (though analysis of this sample is complicated by low mtDNA sequence coverage), pointing to some degree of matrilineal continuity in Europe over at least 10,000 years, ranging from the late Paleolithic to almost the entire Neolithic.

Interestingly, the clustering of all four ancient European samples based on the entire mtDNA sequence (versus just hypervariable segments where such discrimination is not possible (Deguilloux et al., 2009; Frantz et al., 2016)) reveals an older, sub-haplogroup that is sister to the progenitor of both the C1b and C1a sub-haplogroups and appears private to the ancient European dogs in our mitochondrial dataset.

5.3.3 Genetic clustering of the European Neolithic dogs

In order to determine which modern dog population showed greatest genetic similarities with the ancient samples, we constructed a neighbor-joining (NJ) tree using the whole genome dataset. We found that dogs were sister to the Old World wolves, consistent with previous findings. Sister to the group with mostly southeast Asian dogs, there are three major clades, a clade containing Indian village dogs branched out first, then two sister clades, one containing Sub-Saharan, Egypt and Qatar village dogs and the other containing ancient dogs, European breeds, Portuguese and Lebanon village dogs. Among the third clade, CTC branches out first, followed by NGD and HXH, which forms a clade and sister to other European breeds and Portuguese, Lebanon village dogs (Figure 5.4B).

Additionally, principle component analysis (PCA) using both the SNP array and whole genome data demonstrated that all three ancient samples fell within the range of modern dog variation (Figure 5.5A,B). Modern dogs generally orientate towards five main cores: European dogs (from which most modern breeds are derived), African village dogs, Middle Eastern (Lebanon, Qatar, Turkey, Saudi Arabia, Armenia, Iraq) village dogs, Indian village dogs and Southeast Asian (Vietnam, Indonesia, Thailand, China) village dogs. We note that the position of NGD is radically different to that reported by Frantz et al. (2016) due to a technical error they made. When using the larger SNP array reference dataset, HXH and NGD are the ancient samples found closest to the major European cluster, with both lying adjacent to the cluster of Pacific Island dogs that are thought to be derived almost completely from European dogs (Shannon et al., 2015). CTC is located next to village dogs from Afghanistan, a known admixed population also inferred to have a major European-like ancestry component (Shannon et al., 2015). The pattern that CTC falls along the Indian axis and HXH and NGD fall nearest the European core but along the Southeast Asian axis is seen across all SNP call sets (three callsets with different ascertainment schemes of whole genome data and SNP array data).

In addition, an f_3-outgroup analysis performed on both the SNP array and whole genome datasets (with the Golden jackal and Andean fox set as outgroups, respectively) demonstrated that all three Neolithic European samples share most of their ancestry with modern dogs from Europe (Figure 5.6).

5.3.4 Signatures of admixture within canids

Our results are consistent with continuity of a European-like genetic ancestry from modern dogs through the entire Neolithic period, and, based on mtDNA from Bonn Oberkassel, perhaps even into the Upper Paleolithic. However, the slightly displaced position of the ancient samples from the European cluster in the PCAs (particularly for CTC) suggests a complex history possibly involving ancestry from other sources. Unsupervised clustering analyses performed with ADMIXTURE (SNP array data, Figure 5.7) and NGSadmix (whole genome data, Figure 5.5C) revealed that, unlike contemporary European village dogs, all three ancient genomes possess a significant ancestry component that is present in modern Southeast Asian dogs (though to a lesser extent for NGD). This component appears only in a minority of modern European village dogs at very low levels. Furthermore, CTC harbors an additional component that is found predominantly in modern Indian village dogs.

In order to formally model these potential admixture events we applied the tree-based framework, MixMapper (Lipson et al., 2013) to both the SNP array and whole genome data. This approach interrogates every pair of branches in a scaffold tree to infer putative sources of admixture for non-scaffold target samples (in this case HXH, CTC and NGD) via the fitting of f-statistics observed in the data. We inferred that HXH and NGD were formed by an admixture event involving the ancestors of modern European and Southeast Asian dogs. An f_4-ratio test estimated 15% and 8% Southeast Asian-like gene flow into HXH and NGD, respectively (Table 5.1, Figure 5.9). Analysis with ADMIXTUREGRAPH (Patterson et al., 2012), a method related to MixMapper that examines a manually defined demographic history, demonstrated a perfect fit for the observed f-statistics under this model.

In order to disentangle the more complex admixture patterns observed in CTC, we first sought to understand its relationship with HXH given that both samples come from the same geographic region. Interestingly, we found an indication of possible genetic continuity between both samples with our f_3-outgroup analysis, which revealed that CTC had greater affinity with HXH than with any modern canid or with NGD (Figure 5.6B). We therefore performed a MixMapper analysis where HXH was set as one of the sources of admixture for CTC. This analysis identified a population ancestral to modern Indian village dogs as the second source of admixture for CTC, supporting the pattern identified in the unsupervised clustering analyses. An f_4-ratio test estimated 19% Indian admixture in CTC (Table 5.1).

However, given that HXH and modern European dogs share substantial genetic ancestry, it is possible that the observed European-like component in CTC is derived from a different lineage from HXH (for example via a distinct European-Indian admixed population that migrated into Germany sometime during the Neolithic). Therefore, we used ADMIXTUREGRAPH to compare a model of canid demography where a) CTC descended from the same population as HXH followed by admixture with an Indian-like population versus b) both ancient samples being descended from independently diverged European lineages. The model of CTC being a descendant of HXH (Figure 5.10A) provides a much better fit to the data, pointing to general continuity amongst German dogs during the Neolithic, along with admixture towards the latter end of this era with an outside source similar to modern Indian village dogs.

Many modern dog populations in Central Asia (such as Afghanistan) and the Middle East exhibit patterns of admixture that appear similar to those seen in CTC, as shown in our unsupervised clustering analyses (Figure 5.7). MixMapper analysis showed that when including HXH in the scaffold tree, CTC draws its European-like component exclusively from this Early Neolithic German population (consistent with the ADMIXTUREGRAPH results above). However, modern Afghanistan dogs generally demonstrate inferred ancestry from modern European village dogs. This suggests that modern Afghanistan village dogs and CTC are the result of independent admixture events, which in turn implies that dog gene flow across Eurasia has been occurring for thousands of years.

5.3.5 Canid demographic history

The distinct genetic makeup of the European Neolithic dogs from that of modern European dogs indicates that while ancient and contemporary populations share substantial genomic ancestry, some degree of population structure was present in the continent. Neolithic dogs would thus represent a now extinct branch that is somewhat diverged from the modern European clade. In addition, our best fit model of modern and ancient canid demography using ADMIXTUREGRAPH involved a topology that would be consistent with a single dog lineage diverging from wolves (Figure 5.10A). Therefore, we attempted to infer the divergence time of HXH and NGD from modern European dogs after divergence of the Indian lineage, that according to the NJ tree analysis is the sister clade of the Western Eurasian branch (Figure 5.4B, Figure 5.11) (we note that this is a simplistic bifurcating model of what may have been more complex European geographic structuring).

We firstly inferred divergence time among contemporary wolves and dogs using G-Phocs. When assuming the slower mutation rate, μ, of 4×10^{-9} bp/generation, examining only modern village dogs resulted in an estimate of the Asian and non-Asian dog divergence time of 17,500 to 23,900 years ago and dog-wolf divergence time approximately 36,900 to 41,500 years ago (Figure 5.10B). As seen previously in Freedman et al. (2014); Fan et al. (2016), we observed that wolves appeared to diverge rapidly (within the space of 1,000 years). The branching of the main dog lineages occurred over a much longer period of time: after the initial Asian-non-Asian dog divergence, the divergence between Indian and European dogs occurred around 13,700 to 17,900 years ago, while the divergence between European village dogs and modern breeds were around 4,300 to 9,300 years ago. We found that using the Indian or Croatian wolves generally gave similar results to each other (Figure 5.12). We compared our divergence time estimates with the ones in previous studies using G-PhoCS (Table 5.3). We found that dog-wolf divergence time is similar to Freedman's estimates when using the same mutation rate; however our dog divergence time is younger than the Freedman et al. (2014)'s estimate(33,000 years) but similar to the Wang et al. (2016)'s estimate (24,000 years). This discrepancy appears to be result of the sample used to represent the southeast Asian lineage. Freedman et al. (2014) used Dingo, while our study and Wang et al. (2016) used village dogs from South China. When we changed the village dog from South China to Dingo, we observed higher dog divergence time estimates while other estimates remain the same (Figure 5.16). We primarily emphasize result using the village dogs versus the Dingo because the latter are generally considered to be only semi-domesticated.

When adding either of the ancient dogs, we found that the divergence time between European dogs and CTC was 18,000 years and HXH more than 30,000 years, much older than the estimated European-Indian dog split inferred from using only modern samples (Figure 5.12). These in turn led to a larger European-Indian dog split (similar to the divergence time between European and ancient dog) and dog wolf divergence time almost double the original (60,000-80,000 years). Though our genotype calling did substantially lower the number of false positive due to post-mortem damage, when examining the number of private variants for our two ancient samples at these loci, we found a slight excess compared to the modern European village dogs. Thus we anticipate that false positive singleton variants due to the post-mortem damage and lower coverage of the two ancient samples may be artificially elongating branch lengths in the G-PhoCS analysis. We also tried adding NGD to the G-PhoCS analysis, which was sequenced to 28X. The divergence time for European and NGD was 20,000 years ago, and the European-Indian dog 23,000 years ago, much older than the estimated European-Indian dog split inferred from using only modern samples. Although NGD has higher coverage and better genotype calls, false positive singleton variants due to the post-mortem damage are likely still affecting G-PhoCS results (Figure 5.15). As seen in Figure 5.3, even with this high coverage using standard genotype callers will still substantially overestimate C to T and G to A mutations, while our aDNA genotype caller may still not fully capture all damage despite clearly improving the overall false positive rate. Therefore we devised a new method for estimating the HXH-European and NGD-European split time using the G-PhoCS results for only the modern samples as a baseline that would be robust to this signal (see 5.2.6).

We found that the effective population size of village dogs is 5 to 10 folds higher than that of Boxer. The effective population size of Israeli wolf is the highest among all wolves and jackal (Figure 5.13). We also inferred the total migration rate in our analysis, calculated by multiplying migration rate with the time that both population exists during the migration period (Figure 5.14). Total migration rate can be viewed as the probability that a lineage in the target population will migrate into the source population. We found that there was significant non-zero migration from Israeli wolf to Boxer, European village dog and Indian village dog. We also found that the total migration rate from the Israeli wolf to Indian village dogs was around 0.47 when utilizing Croatian wolf as WolfY, much higher than estimated migration to other dogs. However, when utilizing the Indian wolf as WolfY, the total migration rate from the Israeli wolf to Indian village dogs was reduced to 0.02, while total migration rate from Indian wolves to Indian village dogs was

0.34, suggesting this was the more likely source of wolf admixture in Indian village dogs. We hypothesize that this signal is of similar origin to the high migration rate (0.12-0.24) observed in (Freedman et al., 2014), between the Israeli wolf to Basenji. We also found significant non-zero migration between the Chinese wolf and village dogs from South China, the Israeli wolf to the golden jackal and a dog/wolf ancestors to the golden jackal, all of which again are concordant with the results from Freedman et al. (2014).

We also performed a G-PhoCS analysis using the same phylogeny but without any migration band setting. We found that the divergence time among wolves and the divergence time of dog/wolf ancestral population were smaller when migrations between wolves and dogs are neglected. However, the divergence time among dogs were not affected with/without migration (Figure 5.16).

We developed a coalescent-based numerical approach to estimate the divergence time between ancient samples and European populations that is robust to post-mortem DNA damage in order to avoid biases resulting from the use of ancient samples (5.2.6). We first calculated the relative amount of derived allele sharing exclusive to European village dogs and HXH/NGD versus that exclusive to European and Indian village dogs. We then estimated the HXH/NGD versus Europe divergence time in units of expected mutations per site by fitting this observed value to a theoretical expectation obtained using a) demographic parameters inferred by a coalescent-based G-PhoCS analysis of the model that used a subset of eight high coverage modern canid genomes (a mix of a Golden jackal, three wolves, three village dogs, and a boxer) characterized at 15,000 1kb neutral autosomal loci and b) Southeast Asian admixture proportions estimated with the f4-ratio test. The range of sampled parameters and the confidence interval of relative amount of derived allele sharing exclusive to European village dogs and HXH/NGD versus that exclusive to European and Indian village dogs are summarized in Table 5.4 and Table 5.5.

Using this approach, the mean value of divergence time between HXH and European village dogs was estimated as 1.4×10^{-5} and the 95% confidence interval as 9.87×10^{-6} to 1.81×10^{-5}

(Figure 5.18A). If we assume that this divergence time is older than 7,000 years ago, then μ has an upper bound with mean value 6.0×10^{-9} per generation and 95% CI of 4.23×10^{-9} to 7.8×10^{-9} (Figure 5.18C), which is consistent with the $\mu = 4 \times 10^{-9}$ per generation suggested by Skoglund et al. (2015). When using the $\mu = 4 \times 10^{-9}$ rate, the mean divergence time between HXH and European village dogs is 10,542 years ago, with a 95% CI of 7,406 to 13,573 years ago. We also tested our method by replacing HXH with Boxer and estimated the divergence time between Boxer and European to be 8.84×10^{-6} (6630 years ago) as the mean value and 4.1×10^{-6} to 1.41×10^{-5} as the 95% CI, while the mean G-PhoCS estimate of τ_0 was 8.5×10^{-6} (6,375 years ago) with a 95% CI of 5.76×10^{-6} - 1.24×10^{-5} (Figure 5.18B). We also estimated the divergence time between NGD and European village dogs, with mean divergence time 10,971 years ago with a 95% CI of 8,399 to 13,654 years ago, similar to the divergence time between HXH and European village dogs.

In summary (Figure 5.10), we calibrated the divergence time using the mutation rate from Skoglund et al. (2015) (4×10^{-9} per base per generation, 3 years per generation), estimated that modern European and Indian village dogs diverged 13,700-17,900 years, both of which diverged from Southeast Asian dogs 17,500-23,900 years ago as a basal dog divergence event. We note that when correcting for the choice of mutation rate used for calibration, this “east-west” divergence time is compatible with that reported by Wang et al. (2016), but is considerably older than the one recently reported by Frantz et al. (2016). We also estimated the dog-wolf divergence time to be 36,900-41,500 years ago, which is consistent with predictions from the ancient Taimyr wolf genome (Skoglund et al., 2015). Our results therefore provide an upper and lower bound (20,000-40,000 years) for the start of dog domestication, as this process must have occurred subsequent to the dog-wolf divergence and prior to Southeast Asian divergence. We note that the slower mutation rate from Freedman et al. (2014)(1×10^{-8} per generation) is not compatible with the age of HXH (estimated divergence time of 4,000 years, compared to 14C age of 7,000 years).

5.3.6 Functional variants associated with the domestication process

As a result of domestication from wolves, dog genomes have significantly differentiated from their wild counterparts, especially in and around genes believed to have contributed to their domestication and that were selected by humans (Axelsson et al., 2013). To ascertain the extent to which the Neolithic dog genomes display selection at putative "domestication variants" compared to modern dogs, we analyzed their haplotypes at eighteen candidate domestication and selection associated loci (Axelsson et al., 2013). HXH appeared homozygous for the dog-like haplotype at all but one locus, and thus was often indistinguishable from most modern dogs. The younger NGD appeared dog-like at all but two loci. CTC, however, was heterozygous for the wolf-like haplotype at six loci, perhaps because of increased wolf ancestry as revealed by NGSadmix analyses of the whole genome data (Figure 5.5C) and SpaceMix analyses of the SNP array data (Figure 5.8).

The Neolithic saw drastic changes in human culture and behavior, including the advent of agriculture, resulting in a shift toward more starch-rich diets. Elevated *AMY2B* copy-number, which is associated with increased starch metabolism efficiency, has traditionally been suggested to be a strong candidate feature of domestication, even though *AMY2B* copy-number is known to vary widely in diverse collections of wolves and breed dogs (Freedman et al., 2014; Arendt et al., 2014; Reiter et al., 2016). Although the dog haplotype is present in all three Neolithic samples (Figure 5.19A), none show evidence for the extreme copy number expansion of *AMY2B* (Figure 5.19B). Based on read depth, we estimate that CTC and HXH carry two copies of the *AMY2B* gene while NGD carries three copies due to a larger encompassing segmental duplication. Since all ancient dogs lack the tandem copy-number expansion of *AMY2B*, the selective sweep associated with this locus must have occurred after the Neolithic. A similar pattern has recently been observed in humans, where alleles associated with lactase persistence in Europe rise to significant frequencies during the Bronze Age, i.e. 3,000 years after the introduction of milkable livestock.

5.4 Discussion

The admixture events observed in European Neolithic dogs but not in modern dogs from the same region suggest some degree of population structure in the continent during that period. This is further supported by HXH and NGD carrying Southeast Asian ancestry but lacking ancestry shared between modern Indian village dogs and CTC, even though NGD and CTC are almost contemporaneous (4,800 vs. 4,700 years old) . It is likely that a different subpopulation from a structured European Neolithic population eventually became dominant in modern European dogs, which may have also provided an opportunity for the observed mtDNA turnover from haplogroup C to A, especially if this subpopulation also passed through a strong bottleneck or mtDNA was also structured in the continent. Additionally, the age of the samples provide a time frame, between 7,000 and 5,000 years ago, for CTC to obtain its additional ancestry component. Considering that CTC shows similar admixture patterns to Central Asian and Middle Eastern modern dog populations, as seen in the PCA (Figure 5.5) and ADMIXTURE (Figure 5.7) analysis, and that the cranium was found next to two individuals associated with the Neolithic Corded Ware Culture, we speculate that the Indian-like gene flow may have been acquired by admixture with incoming populations of dogs that accompanied steppe people migrating from the East.

We do not find evidence of a remnant European Paleolithic dog population contributing to the genetic architecture of ancient dogs from either the Early or End Neolithic and therefore our results do not support the hypothesis of a population replacement during the Neolithic. Instead, we find that NGD is genetically very similar to HXH, with substantial proportions of modern European-like ancestry.

Frantz et al. (2016) used an MSMC approach to estimate the split time between Asian and non-Asian dogs. However, MSMC requires phased haplotypes from all samples. Statistical phasing errors in human haplotypes result in a more recent split time estimates. The performance of statistical phasing on dog genomes is not quantified but the size of the dog reference panel is less than 10% of the size available for human data. Our study and Freedman et al. (2014) and Wang et al. (2016) obtained Asian vs non-Asian dog divergence time estimates over 20,000 years ago when using the same mutation rate (4×10^{-9}/generation) and generation time (3 years) (Table 5.3). Since the earliest dog fossil remains are dated to be 15,000 years old in Western Eurasia (Europe and the Near East), the divergence time between Asian and non-Asian dogs occurring 20,000 years ago does not support two domestication processes.

In addition, potential genotyping errors due to post-mortem damage may also cause biases when utilizing the NGD to estimate a divergence time with Asian dogs, which Frantz et al. (2016) find has a slightly older divergence compared to using modern European dogs and suggest is evidence of ancestry from the remnant European Paleolithic dog population. Fig S26 in Frantz et al. (2016) notably demonstrates that the Tv/Ti heterozygote ratio for NGD is lower than all but one of their contemporary canid genomes, suggesting even with base quality recalibration, false positive transitions (C to T and G to A) due to post-mortem damage may still be prominent in the inferred sequence for the NGD, which would likely lead to increased divergence times by artificially elongating the branch lengths for NGD.

Given our older estimate of divergence time between Southeast Asian and European dog populations compared to Frantz et al. (2016) and our best model of population demography (Figure 5.10), our results can most parsimoniously be explained by a single domestication process for dogs. If there was a replacement of a highly distinct European dog lineage with independent domestication origins to that of dogs in Asia, it must have occurred prior to the Neolithic (and perhaps much earlier given the matrilineal continuity between HXH, CTC and Bonn Oberkassel). However, currently we find no genetic evidence that supports the recently proposed hypothesis of a dual domestication processes for dogs. Clearly, analysis of additional spatiotemporally diverse ancient genomes is required to better understand the possibly complex admixture processes that occurred during this period, while older specimens, particularly from Asia, will be necessary to resolve the history of dog domestication.

Figures

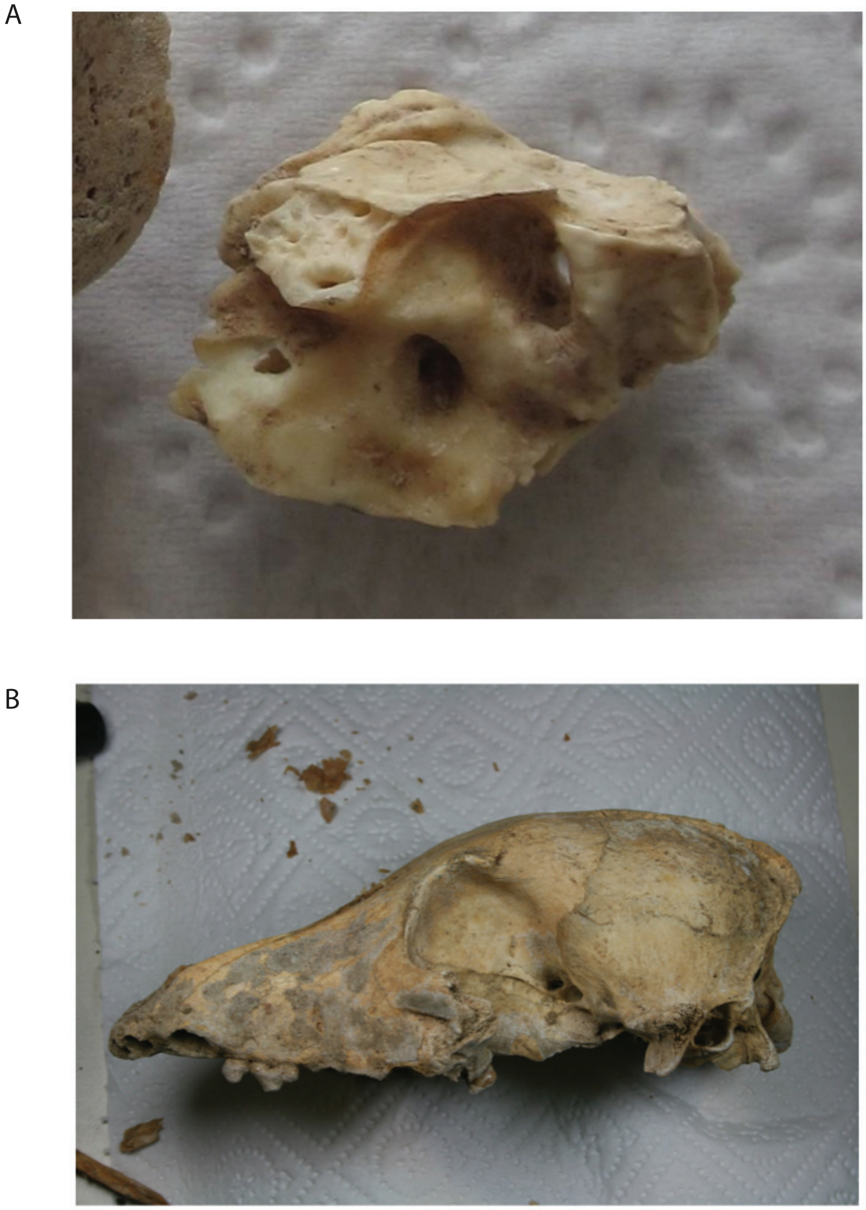

Figure 5.1: **Archaeological remains.** A. HXH petrous bone. B. CTC cranium. (data generated by Amelie Scheu)

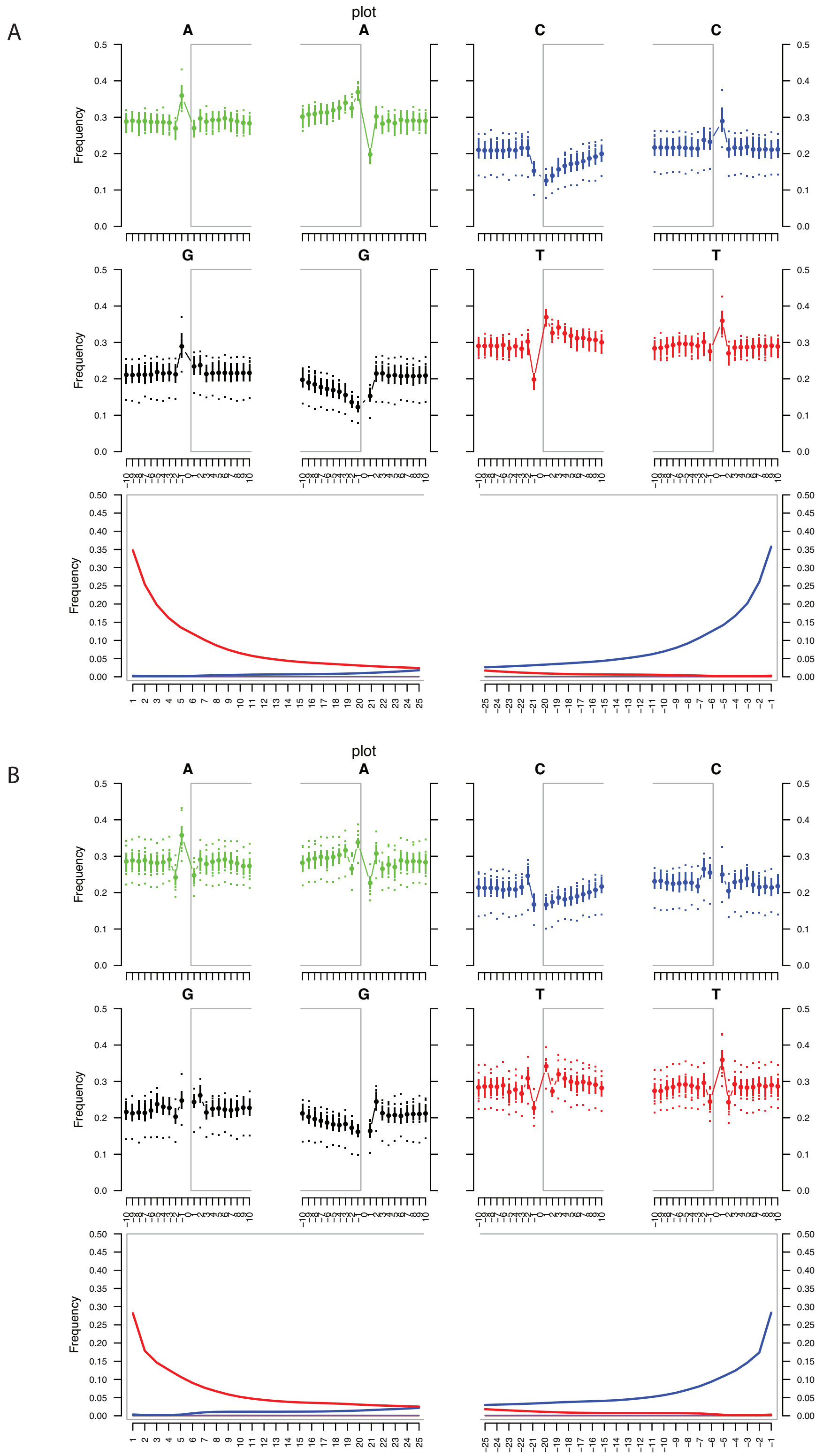

Figure 5.2: **MapDamage analysis showing the deamination changes typical of ancient DNA.** A. HXH. B. CTC.

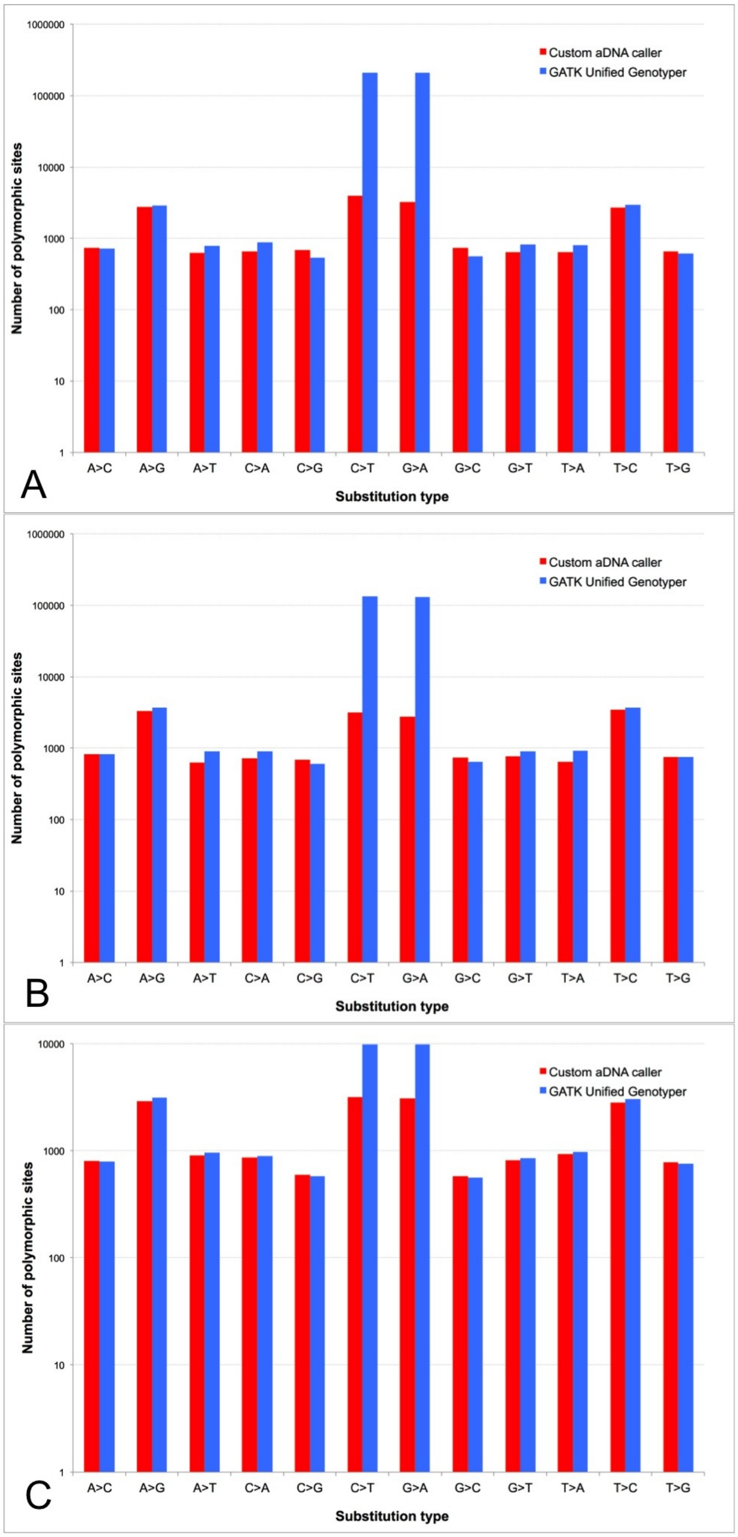

Figure 5.3: **Performance of ancient genotype caller.** Comparison of base substitution count with Weibull-based caller (red) and the GATKUnified Genotyper caller (blue). a) HXH, b) CTC, c) NGD. Note Y axis is log-scaled. (data generated by Krishna R. Veeramah)

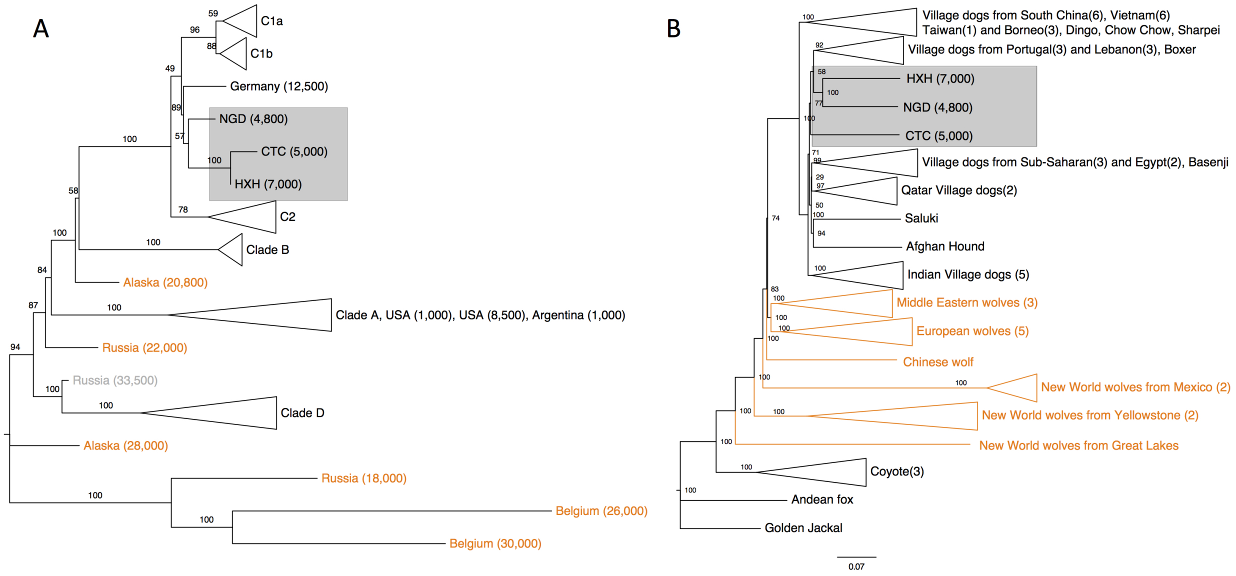

Figure 5.4: **Phylogeny of ancient and contemporary canids.** A). Phylogeny based on mtDNA. Age of the samples is indicated in parentheses, wolf samples shown in orange. B) Neighbor-joining tree based on pairwise sequence divergence from whole genome data. (figureA generated by Matthew Oetjens)

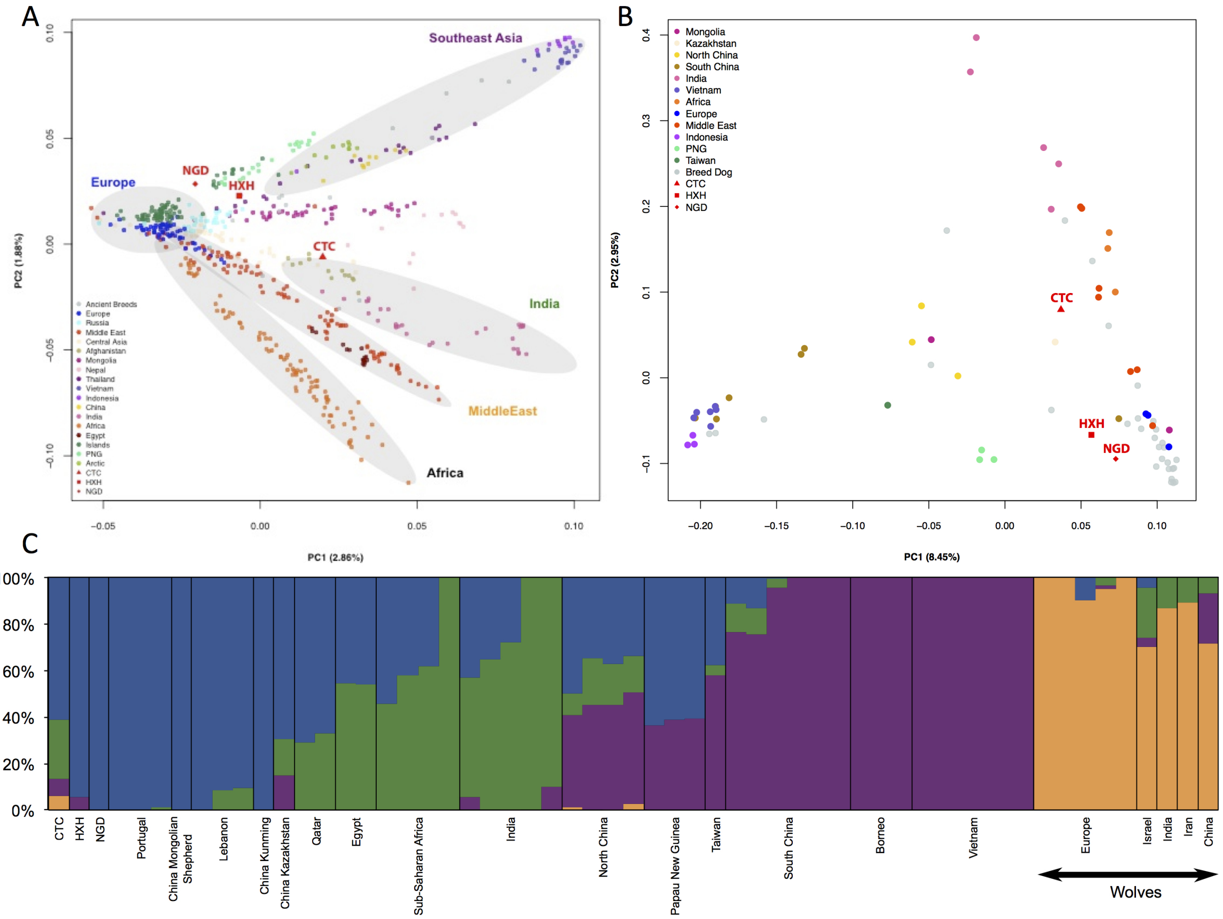

Figure 5.5: **PCA and population structure between ancient and contemporary canids.** A). PCA of village dogs, with breed dogs and ancient dogs projected onto the PC space using SNP array data. B). PCA of village dogs, breed dogs and ancient dogs using whole genome SNP data ascertained in the New World wolves. C). NGSadmix clustering for K=4 for village dogs, ancient dogs and Old World wolves based on the whole genome SNP data. (data generated by Angela M. Taravella and Krishna R. Veeramah)

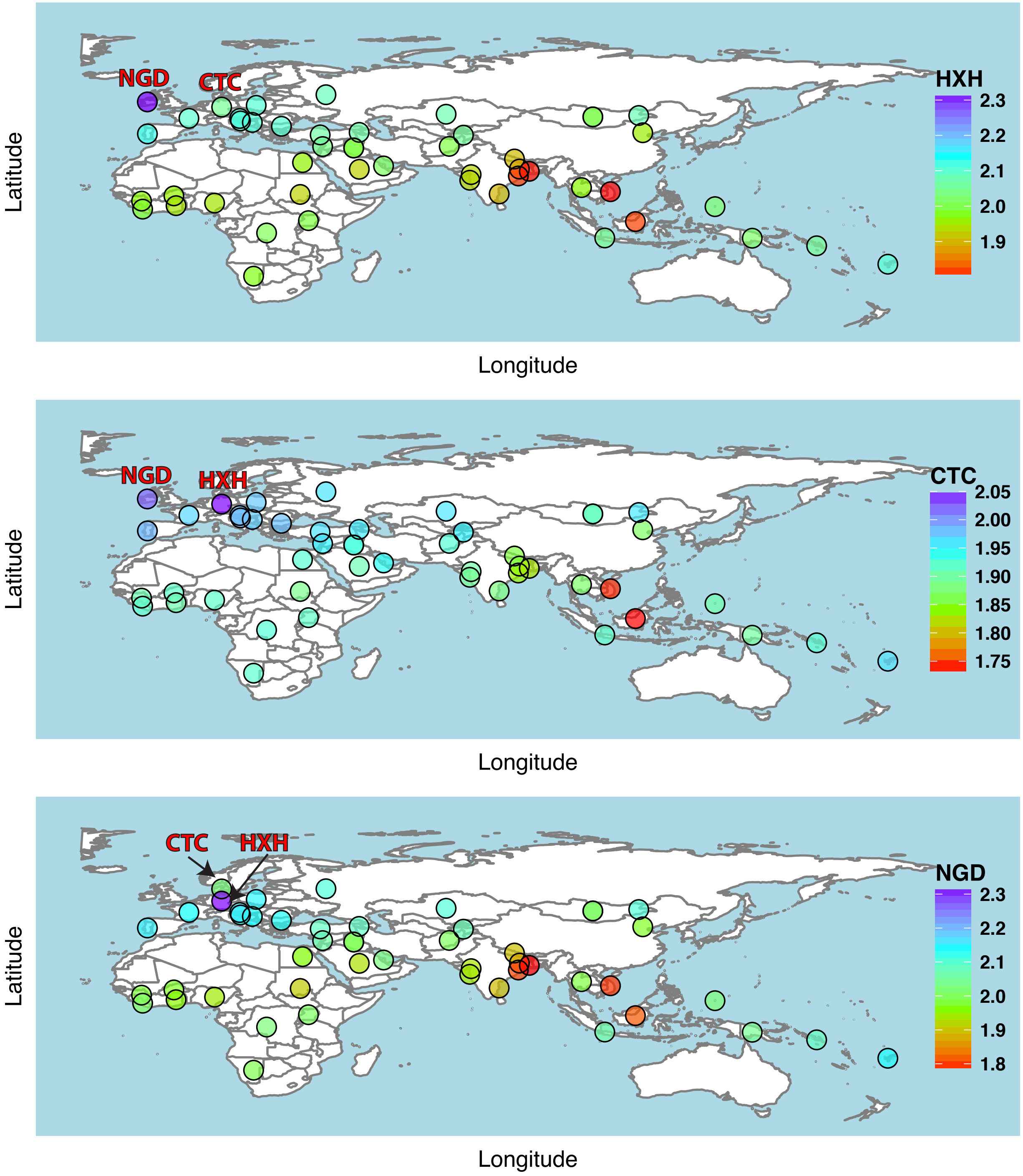

Figure 5.6: **Genetic affinity of ancient samples.** Heat map of outgroup f_3-statistics of the form f_3(Golden Jackal; Ancient, X) based on SNP array genotype data. A). HXH shows greatest similarity with NGD and modern European village dogs, and is most distant to East Asian and Indian village dogs. B). CTC shares the most genetic similarity with HXH, followed by NGD and other European dogs. In addition, CTC shows greater similarity to village dogs from India (particularly unadmixed populations in the east) than HXH does.

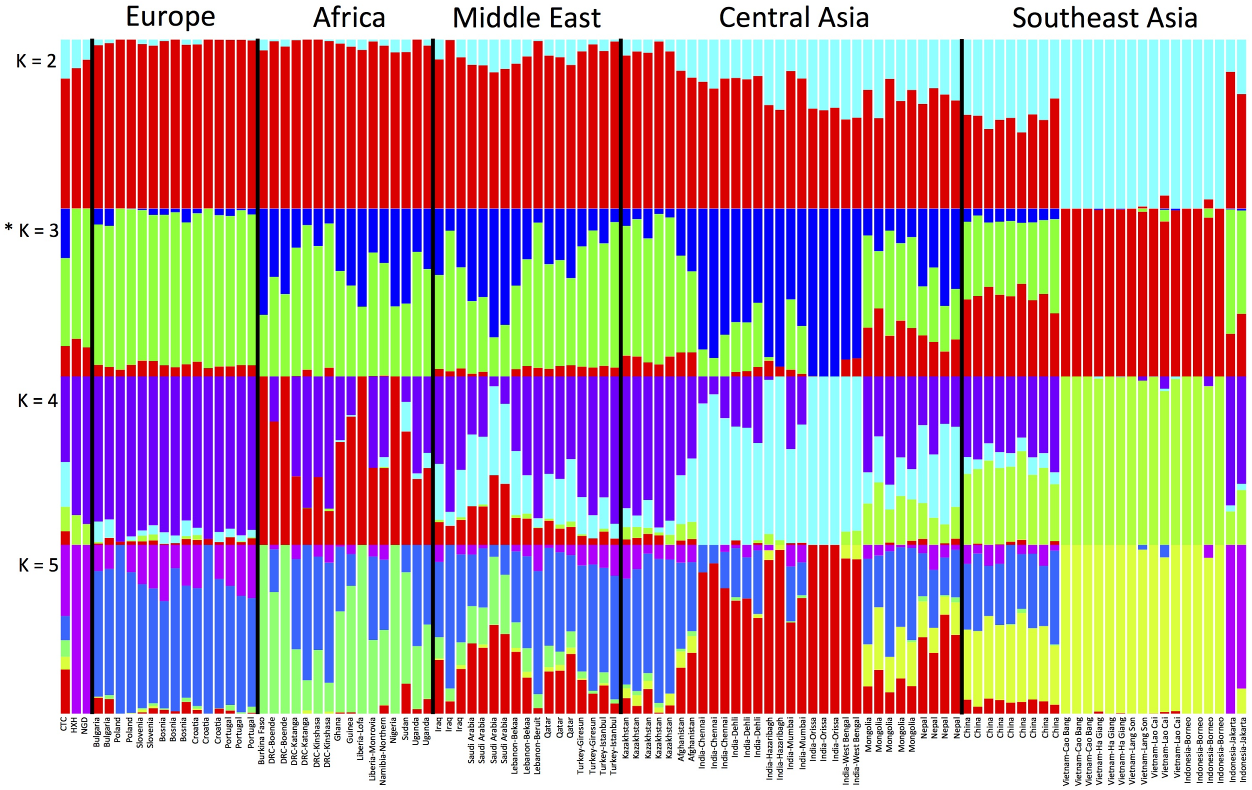

Figure 5.7: **ADMIXTURE analysis based on SNP array data.** ADMIXTURE analysis for K = 2 through 5 for a global representation of village dogs, CTC, HXH and NGD. Vertical lines represent individual dogs. (data generated by Angela M. Taravella)

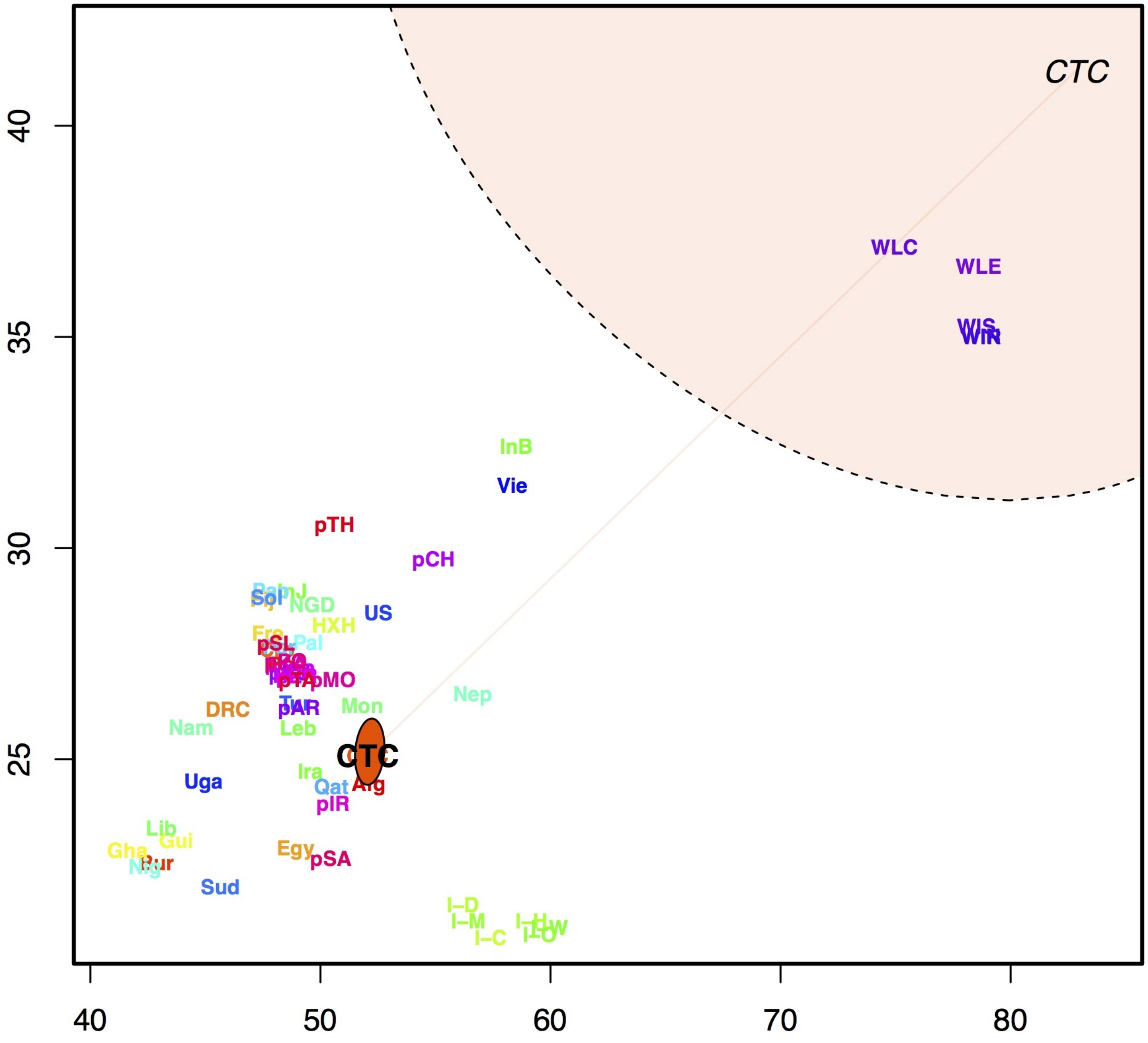

Figure 5.8: **CTC-Wolf admixture as inferred by SpaceMix**. CTC in bold type reflects the samples geogenetic position (95% CI solid orange ellipse), CTC in italics reflects the geogenetic position of the proposed source of admixture into CTC, with an estimated value of 9% (95% CI transparent orange ellipse). (data generated by Krishna R. Veeramah)

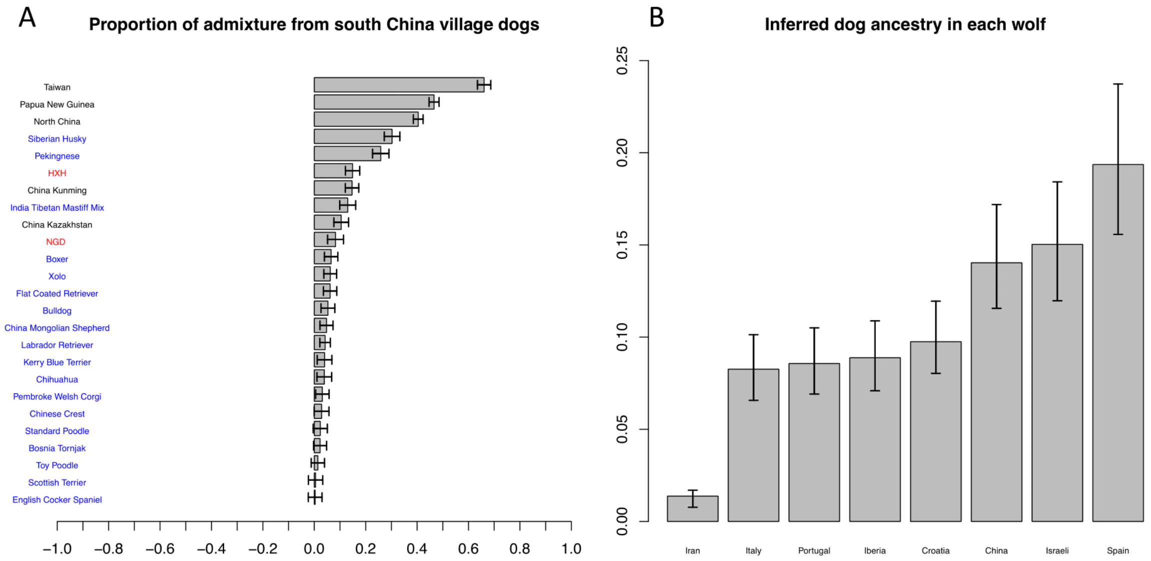

Figure 5.9: **Inferred admixture proportion from f_4-ratio tests.** A. Proportion of admixture from south China village dogs in other dogs. Breeds were colored blue, village dogs black, ancient dogs red. B. The inferred proportion of dog ancestry in each wolf. The boxplot shows minimum, mean and maximum.

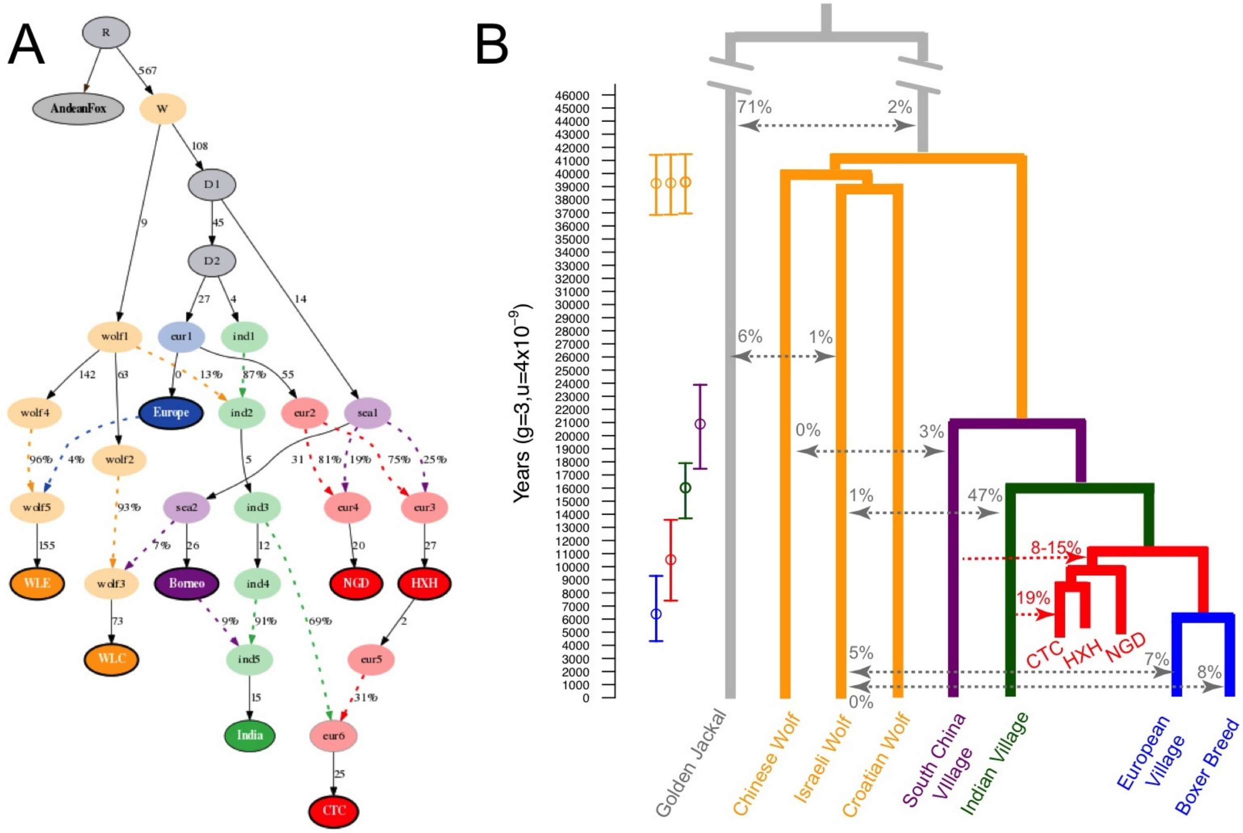

Figure 5.10: **Demographic model regarding ancient and contemporary dogs and wolves.** A) The best model fit to both modern and ancient canid data using ADMIXTUREGRAPH. This model had four f_4-statistic outliers. Branches indicated by solid black lines, admixture indicated by coloured dashed lines. Sampled individuals/populations indicated by solid circles with bold outline. B) Divergence times of contemporary dogs and wolves were inferred using G-PhoCS. Mean estimates are indicated by squares with ranges correspond to 95% Bayesian credible intervals. Migration bands are shown in grey with associated value representing the inferred total migration rates (the probability that a lineage in the target population will migrate into the source population). The divergence time for HXH and NGD and modern European dogs is inferred using a numerical approach. The proportion of Indian village dog ancestry in CTC and South China village dog ancestry in HXH and NGD are inferred by f_4 ratio test, shown in red. (figureA generated by Krishna R. Veeramah)

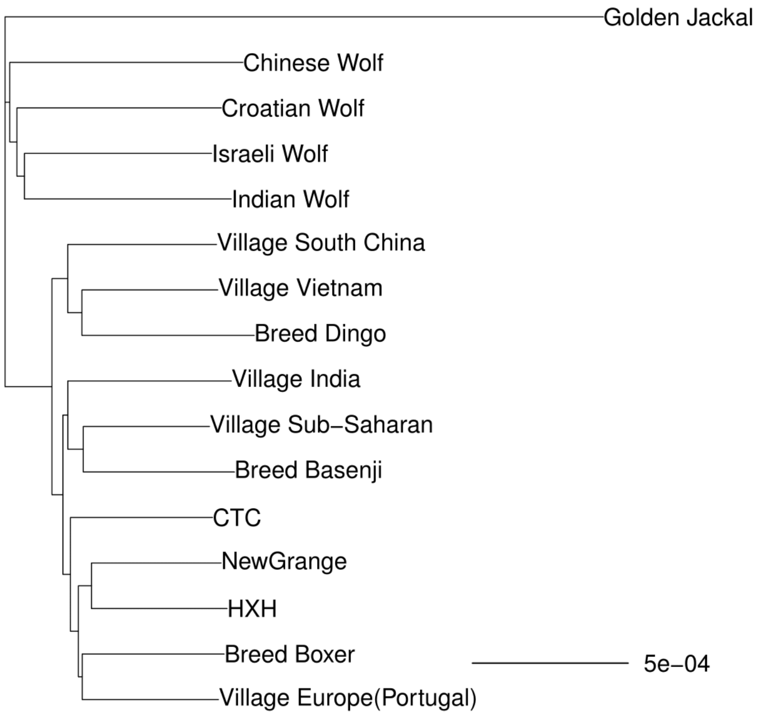

Figure 5.11: **NJ tree based on 16,434 “neutral” loci of selected samples used in G-PhoCS.**

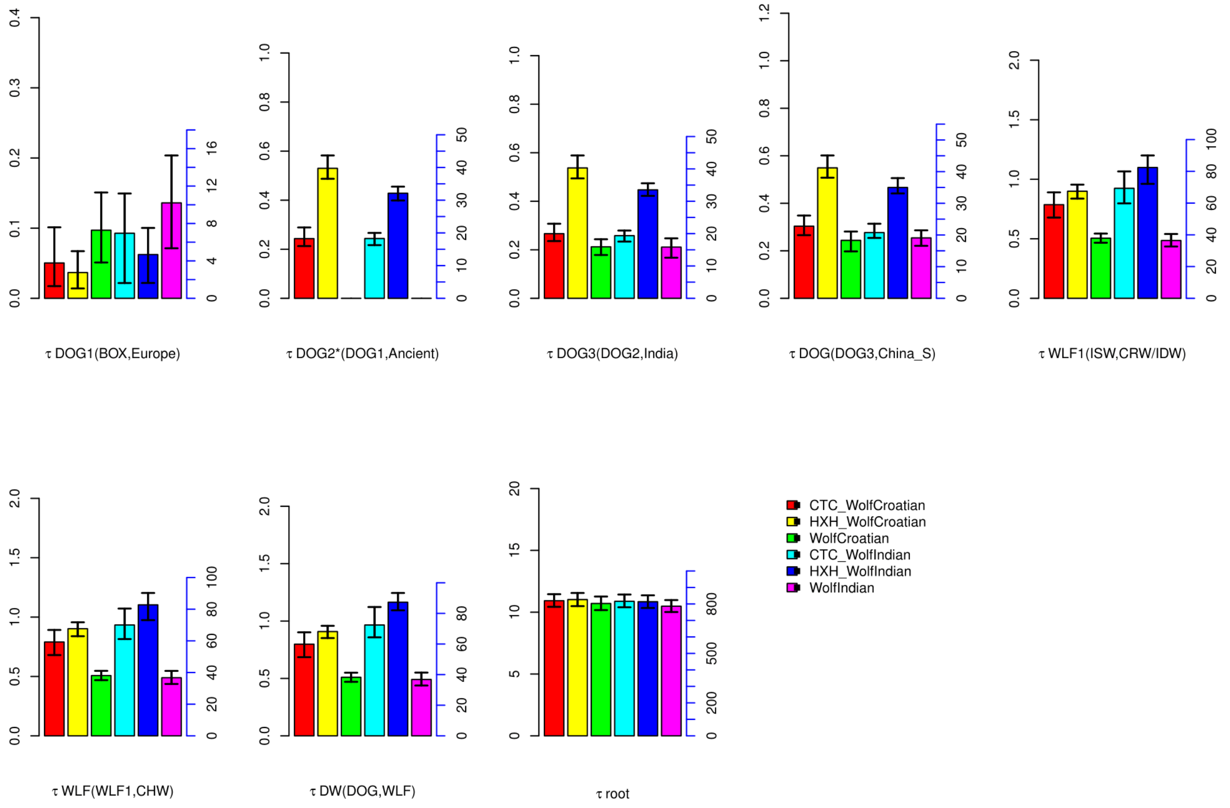

Figure 5.12: **Divergence time estimates in G-PhoCS analysis when including and excluding ancient samples.** We use the tree of the form (((((Boxer, Village_Europe), Village_India), Village_ChinaS), ((wolfY, IsraeliWolf), ChineseWolf)), GoldenJackal), with or without one ancient sample sister to the ancestral of Boxer and Village_Europe: (1) wofY as CroatianWolf, with CTC; (2) wofY as CroatianWolf, with HXH; (3) wofY as CroatianWolf, no ancient sample; (4) wofY as IndianWolf, with CTC; (5) wofY as IndianWolf, with HXH; (6) wofY as IndianWolf, no ancient sample. Raw estimates on the left axis (scaled up by 1e04) and calibrated estimates on the right axis (in 1,000 years). This analysis used 5000 randomly selected loci.

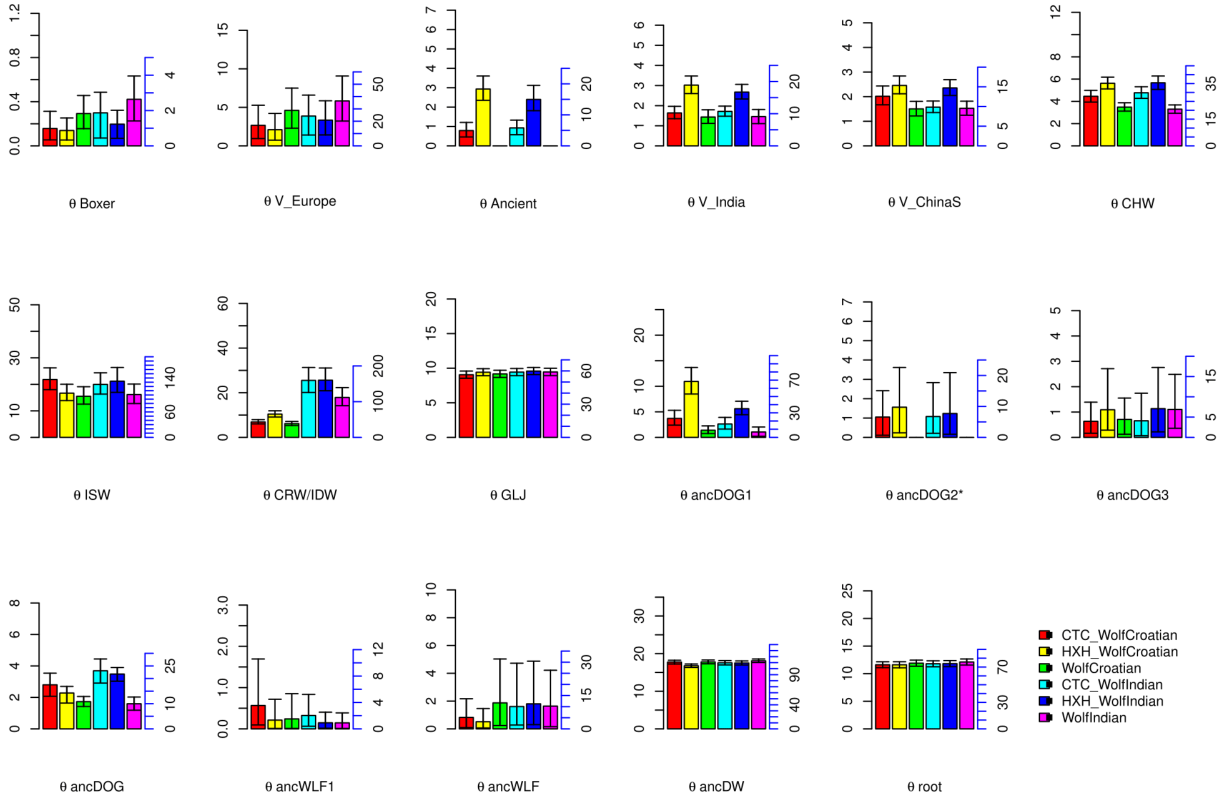

Figure 5.13: **Population size estimates in G-PhoCS analysis when including and excluding ancient samples.** We use the tree of the form (((((Boxer, Village_Europe), Village_India), Village_ChinaS), ((wolfY, IsraeliWolf), ChineseWolf)), GoldenJackal), with or without one ancient sample sister to the ancestral of Boxer and Village_Europe: (1) wofY as CroatianWolf, with CTC; (2) wofY as CroatianWolf, with HXH; (3) wofY as CroatianWolf, no ancient sample; (4) wofY as IndianWolf, with CTC; (5) wofY as IndianWolf, with HXH; (6) wofY as IndianWolf, no ancient sample. Raw estimates on the left axis (scaled up by 1e04) and calibrated estimates on the right axis (in 1,000 years). This analysis used 5000 randomly selected loci.

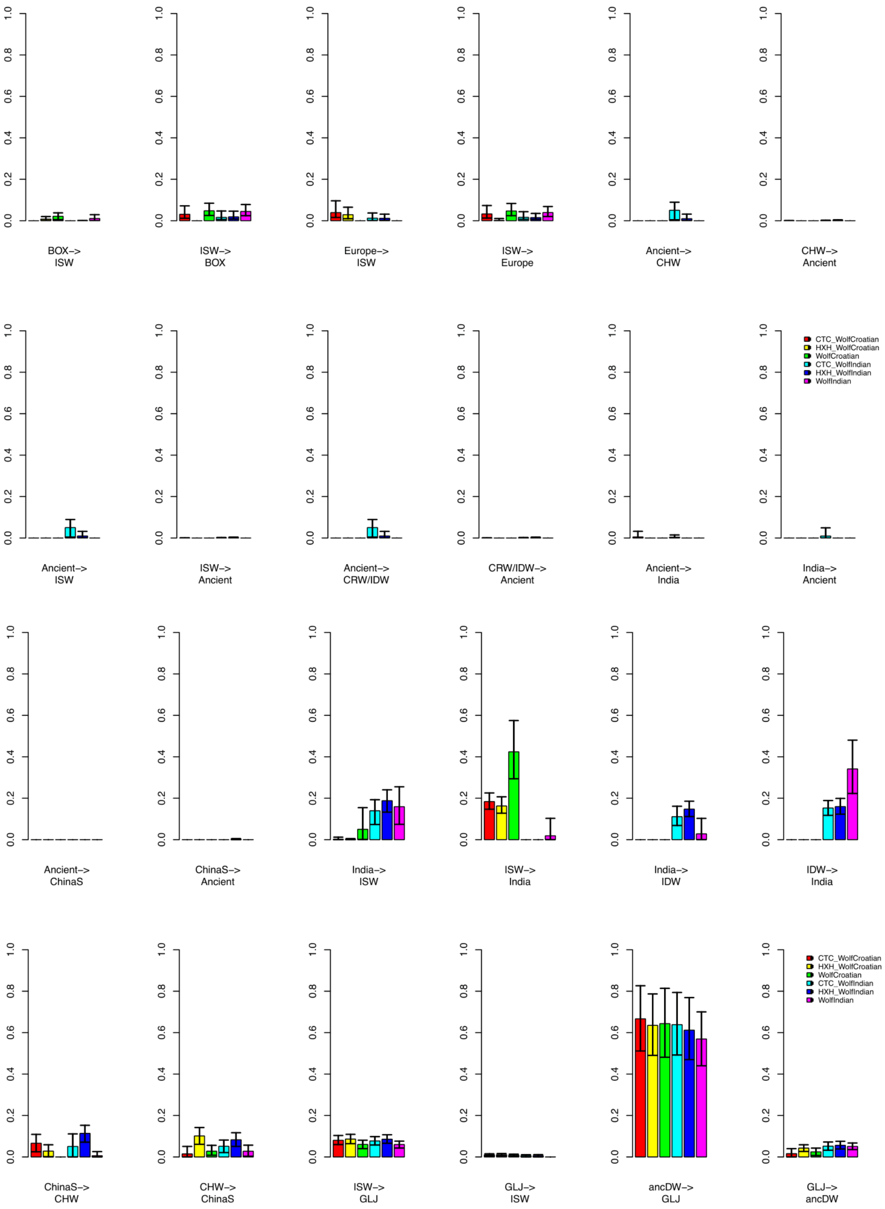

Figure 5.14: **Total migration rate estimates in G-PhoCS analysis when including and excluding ancient samples.** We use the tree of the form (((((Boxer, Village_Europe), Village_India), Village_ChinaS), ((wolfY, IsraeliWolf), ChineseWolf)), GoldenJackal), with or without one ancient sample sister to the ancestral of Boxer and Village_Europe: (1) wofY as CroatianWolf, with CTC; (2) wofY as CroatianWolf, with HXH; (3) wofY as CroatianWolf, no ancient sample; (4) wofY as IndianWolf, with CTC; (5) wofY as IndianWolf, with HXH; (6) wofY as IndianWolf, no ancient sample. Raw estimates on the left axis (scaled up by 1e04) and calibrated estimates on the right axis (in 1,000 years). This analysis used 5000 randomly selected loci.

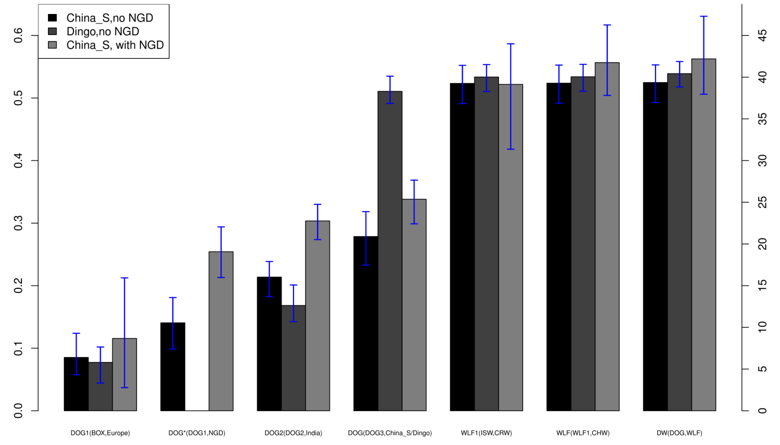

Figure 5.15: **Comparison of divergence time estimates with/without NGD in G-PhoCS analysis.** We use the tree of the form (((((Boxer, Village_Europe), Village_India), Village_ChinaS or Dingo), ((CroatianWolf, IsraeliWolf), ChineseWolf)), GoldenJackal), with or without one ancient sample sister to the ancestral of Boxer and Village_Europe. The divergence time for (DOG1,NGD) when NGD is not in G-PhoCS was estimated using a numerical approach. Axis on the left are raw estimates, scaled up by 1e04. Axis on the right are recalibrated estimates in thousand years, assuming mutation rate 4×10^{-9} per/generation, generation time 3 years.

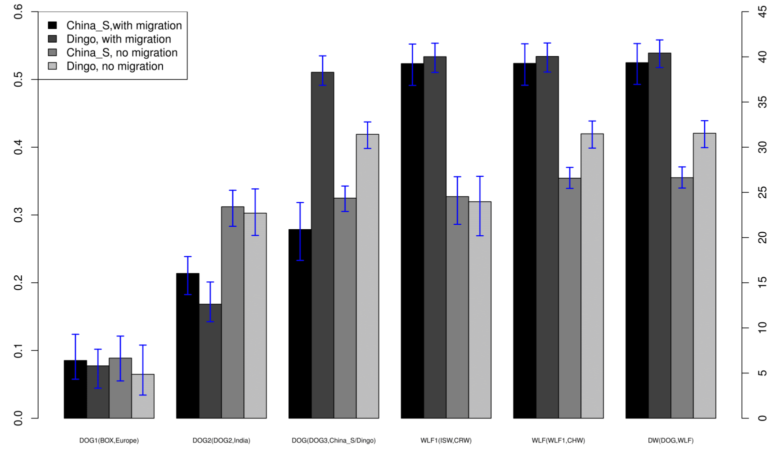

Figure 5.16: **Comparison of divergence time estimates with/without migration setting in G-PhoCS analysis.** We use the tree of the form (((((Boxer, Village_Europe), Village_India), Village_ChinaS or Dingo), ((CroatianWolf, IsraeliWolf), ChineseWolf)), GoldenJackal), with or without migration setting. Axis on the left are raw estimates, scaled up by 1e04. Axis on the right are recalibrated estimates in thousand years, assuming mutation rate 4×10^{-9} per/generation, generation time 3 years. This analysis used 16,434 loci.

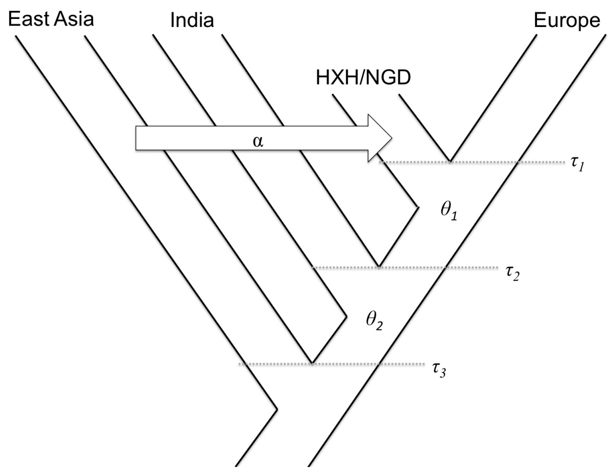

Figure 5.17: **Illustration of the tree structure used for estimating HXH/Europe divergence time.**

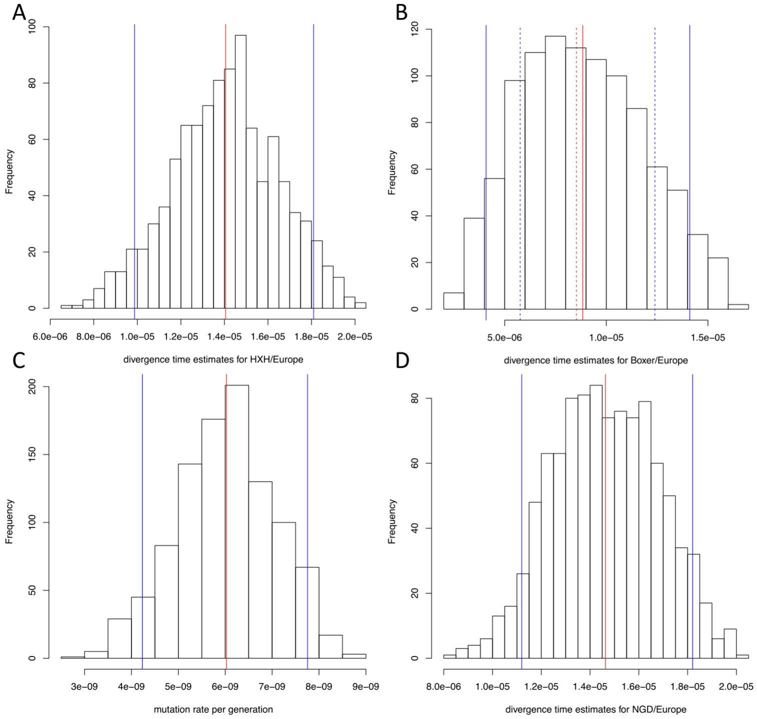

Figure 5.18: **Results for numerical analysis**. Red lines indicate mean value and blue as confidence intervals. A. Inferred divergence time for HXH/Europe. B. Inferred divergence time for Boxer/Europe using this method compared to G-PhoCS estimates. dashed lines are the G-PhoCS estimates. C. Inferred upper boundaries for mutation rate when assuming the divergence time between HXH/Europe must be later than 7,000 years ago. D. Inferred divergence time for NGD/Europe.

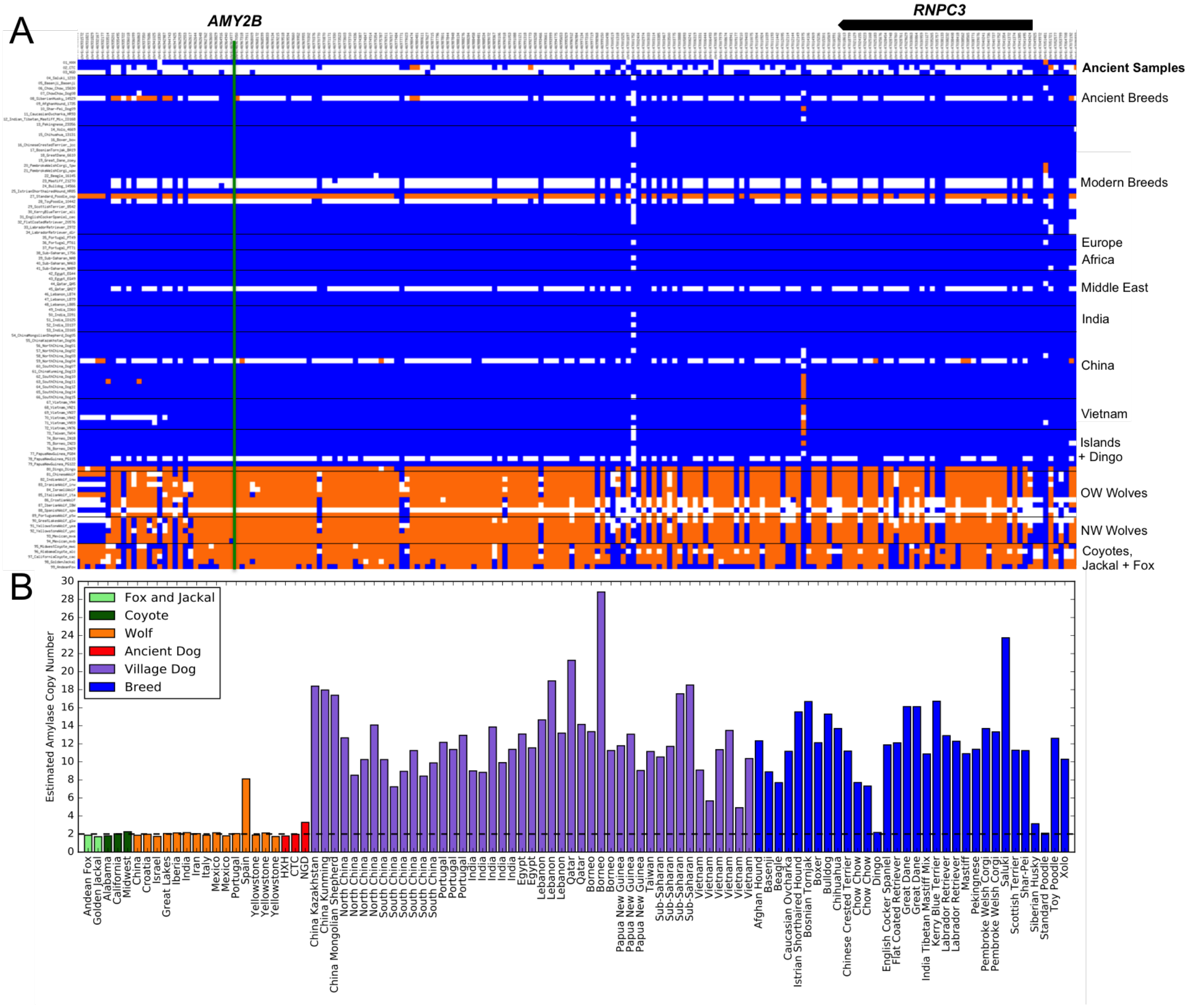

Figure 5.19: **Haplotype and copy-number variation at the *AMY2B* locus.** A). Genotype matrix of selected sites within FST-derived domestication locus 12 (chr6: 46854109-47454177)24. SNP genotypes are represented as either homozygous for the reference allele (0/0; blue), heterozygous (0/1; white), or homozygous (1/1; orange) for the alternate allele. The positions of AMY2B (green line) and RNPC3 (model above) are indicated. B). Read-depth based estimation of AMY2B copy number for the Andean fox (light green), golden jackal (light green), coyotes (dark green), wolves (orange), ancient samples (red), village dogs (purple), and breed dogs (blue). (data generated by Amanda L. Pendleton)

Tables

Table 5.1: f_4**-ratio test results** f_4-ratio test on the form of the ratio between $f_4(A, O; X, C)$ and $f_4(A, O; B, C)$.

A	B	X	C	O	alpha	std.err	Z
Vietnam	South China	HXH	Portugal	Andean fox	0.1489	0.0279	5.342
South China	Vietnam	HXH	Portugal	Andean fox	0.1607	0.0306	5.25
China	Borneo	HXH	Portugal	Andean fox	0.1755	0.0334	5.247
Vietnam	Borneo	HXH	Portugal	Andean fox	0.1307	0.0247	5.287
Vietnam	South China	NGD	Portugal	Andean fox	0.0827	0.031	2.665
Basenji	India	CTC	Portugal	Andean fox	-1.2248	0.4393	-2.788
Basenji	India	CTC	HXH	Andean fox	0.1852	0.0977	1.896
Lebanon	Portugal	CTC	India	Andean fox	0.1816	0.0369	4.922
Lebanon	HXH	CTC	India	Andean fox	0.2436	0.0489	4.987
Portugal	HXH	CTC	India	Andean fox	0.3083	0.0402	7.663
South China	Portugal	wolf_Israeli	wolf_India	Andean fox	0.1547	0.0147	10.501
South China	HXH	wolf_Israeli	wolf_India	Andean fox	0.1487	0.014	10.62
South China	CTC	wolf_Israeli	wolf_India	Andean fox	0.1652	0.0154	10.716
Vietnam	South China	wolf_China	wolf_India	Andean fox	0.1176	0.0098	12.028
Vietnam	HXH	wolf_China	wolf_India	Andean fox	0.1448	0.012	12.07
Vietnam	CTC	wolf_China	wolfIndia	Andean fox	0.1608	0.0132	12.213

Table 5.2: **D-statistics supporting the migration band setting used in G-PhoCS.**

Pop A	Pop B	Pop C	Outgroup	f_4	Z-score
wolf_Israeli	wolf_China	Boxer	Andean fox	0.0389	6.628
wolf_Israeli	wolf_Croatia	Boxer	Andean fox	0.0368	6.39
wolf_China	wolf_Croatia	South China	Andean fox	0.0139	4.079
wolf_Israeli	wolf_China	India	Andean fox	0.0366	7.474
wolf_Israeli	wolf_Croatia	India	Andean fox	0.0355	7.278
wolf_Israeli	wolf_India	India	Andean fox	0.06	11.115
wolf_Israeli	wolf_China	Portugal	Andean fox	0.0378	6.744
wolf_Israeli	wolf_Croatia	Portugal	Andean fox	0.0363	6.602
Boxer	South China	wolf_China	Andean fox	-0.0129	-3.89
India	South China	wolf_China	Andean fox	-0.017	-7.902
Portugal	South China	wolf_China	Andean fox	-0.0143	-5.187
Boxer	South China	wolf_Israeli	Andean fox	0.0342	9.086
India	South China	wolf_Israeli	Andean fox	0.0258	10.603
Portugal	South China	wolf_Israeli	Andean fox	0.0311	9.812

Table 5.3: **Comparison of divergence time estimates from previous studies using G-PhoCS.**

	dog/wolf divergence (kyrs)		dog divergence (kyrs)	
Studies	$\mu = 1 \times 10^{-8}$	$\mu = 4 \times 10^{-9}$	$\mu = 1 \times 10^{-8}$	$\mu = 4 \times 10^{-9}$
Freedman et al	15	37.5	12.9	32.3
Wang et al	24.6	61.5	9.6	24
Fan et al	11.7	29.3	-	-
Our study	15.7	39.4	8.4	20.9

Table 5.4: **The range of parameters sampled for estimating HXH/European divergence time**. θ and τ estimates were based on G-PhoCS results and alpha value was based on f_4-ratio analysis.

Parameter	Uniform range
θ_1	$(0.86 \times 10^{-4}, 1.95 \times 10^{-4})$
θ_2	$(0.25 \times 10^{-4}, 2.34 \times 10^{-4})$
τ_0	$(5.76 \times 10^{-6}, 1.24 \times 10^{-5})$
τ_1	(τ_0, τ_2)
τ_2	$(1.83 \times 10^{-5}, 2.39 \times 10^{-5})$
τ_3	$(2.33 \times 10^{-5}, 3.18 \times 10^{-5})$
α	(0.12,0.17) for HXH (0.05,0.11) for NGD 0 for Boxer

Table 5.5: **Jackknife estimates of P(110)/P(101).** Jackknife estimates and confidence interval of the ratio of the number of SNPs where i) a European village dog and dogX have the derived allele and an Indian sample has the ancestral allele versus ii) a European and Indian village dog have the derived allele and dogX has the ancestral allele.

dogX	Jackknife Estimates	Standard Deviation	Confidence Interval
Boxer	1.533886	0.020438	1.513448-1.554324
HXH	1.201485	0.015546	1.185939-1.217031
NGD	1.212794	0.01781	1.194984-1.230604

Bibliography

Alexander, D. H., Novembre, J. and Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J. O., Baker, C., Malig, M., Mutlu, O. et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*, 41(10):1061–1067, 2009.

Arendt, M., Fall, T., Lindblad-Toh, K. and Axelsson, E. Amylase activity is associated with amy2b copy numbers in dog: implications for dog domestication, diet and diabetes. *Animal genetics*, 45(5):716–722, 2014.

Auton, A., Li, Y. R., Kidd, J., Oliveira, K., Nadel, J., Holloway, J. K., Hayward, J. J., Cohen, P. E., Greally, J. M., Wang, J. et al. Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genet*, 9(12):e1003984, 2013.

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M. T., Perloski, M., Liberg, O., Arnemo, J. M., Hedhammar, Å. and Lindblad-Toh, K. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, 495(7441):360–364, 2013.

Benecke, N. Studies on early dog remains from northern europe. *Journal of Archaeological Science*, 14(1):31–49, 1987.

Bradburd, G. S., Ralph, P. L. and Coop, G. M. A spatial framework for understanding population structure and admixture. *PLoS Genet*, 12(1):e1005703, 2016.

Bramanti, B., Thomas, M., Haak, W., Unterlaender, M., Jores, P., Tambets, K., Antanaitis-Jacobs, I., Haidle, M., Jankauskas, R., Kind, C.-J. et al. Genetic discontinuity between local hunter-gatherers and central europe's first farmers. *science*, 326(5949):137–140, 2009.

Briggs, A. W., Stenzel, U., Johnson, P. L., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M. et al. Patterns of damage in genomic dna sequences from a neandertal. *Proceedings of the National Academy of Sciences*, 104(37):14616–14621, 2007.

Busing, F. M., Meijer, E. and Van Der Leeden, R. Delete-m jackknife for unequal m. *Statistics and Computing*, 9(1):3–8, 1999.

Deguilloux, M., Moquel, J., Pemonge, M. and Colombeau, G. Ancient dna supports lineage replacement in european dog gene pool: insight into neolithic southeast france. *Journal of Archaeological Science*, 36(2):513–519, 2009.

Duleba, A., Skonieczna, K., Bogdanowicz, W., Malyarchuk, B. and Grzybowski, T. Complete mitochondrial genome database and standardized classification system for canis lupus familiaris. *Forensic Science International: Genetics*, 19:123–129, 2015.

Fan, Z., Silva, P., Gronau, I., Wang, S., Armero, A. S., Schweizer, R. M., Ramirez, O., Pollinger, J., Galaverni, M., Del-Vecchyo, D. O. et al. Worldwide patterns of genomic variation and admixture in gray wolves. *Genome research*, 26(2):163–173, 2016.

Frantz, L. A., Mullin, V. E., Pionnier-Capitan, M., Lebrasseur, O., Ollivier, M., Perri, A., Linderholm, A., Mattiangeli, V., Teasdale, M. D., Dimopoulos, E. A. et al. Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*, 352(6290):1228–1231, 2016.

Freedman, A. H., Gronau, I., Schweizer, R. M., Ortega-Del Vecchyo, D., Han, E., Silva, P. M., Galaverni, M., Fan, Z., Marx, P., Lorente-Galdos, B. et al. Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet*, 10(1):e1004016, 2014.

Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., Mattiangeli, V., Domboróczki, L., Kővári, I., Pap, I., Anders, A. et al. Genome flux and stasis in a five millennium transect of european prehistory. *Nature communications*, 5, 2014.

Ginolhac, A., Rasmussen, M., Gilbert, M. T. P., Willerslev, E. and Orlando, L. mapdamage: testing for damage patterns in ancient dna sequences. *Bioinformatics*, 27(15):2153–2155, 2011.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. and Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics*, 43(10):1031–1034, 2011.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K. et al. Massive migration from the steppe was a source for indo-european languages in europe. *Nature*, 522(7555):207–211, 2015.

Hach, F., Hormozdiari, F., Alkan, C., Hormozdiari, F., Birol, I., Eichler, E. E. and Sahinalp, S. C. mrsfast: a cache-oblivious algorithm for short-read mapping. *Nature methods*, 7(8):576–577, 2010.

Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez-del Molino, D., van Dorp, L., López, S., Kousathanas, A., Link, V. et al. Early farmers from across europe directly descended from neolithic aegeans. *Proceedings of the National Academy of Sciences*, page 201523951, 2016.

Horard-Herbin, M.-P., Tresset, A. and Vigne, J.-D. Domestication and uses of the dog in western europe from the paleolithic to the iron age. *Animal Frontiers*, 4(3):23–31, 2014.

Kircher, M. Analysis of high-throughput ancient dna sequencing data. *Ancient DNA: methods and protocols*, pages 197–228, 2012.

Li, H. and Durbin, R. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–819, 2005.

Lipson, M., Loh, P.-R., Levin, A., Reich, D., Patterson, N. and Berger, B. Efficient moment-based inference of admixture parameters and sources of gene flow. *Molecular biology and evolution*, 30(8):1788–1802, 2013.

MacHugh, D. E., Edwards, C. J., Bailey, J., Bancroft, D. R. and Bradley, D. G. The extraction and analysis of ancient dna from bone and teeth: a survey of current methodologies. *Ancient Biomolecules*, 3(2):81–103, 2000.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

Paradis, E., Claude, J. and Strimmer, K. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290, 2004.

Parker, H. G., Kim, L. V., Sutter, N. B., Carlson, S., Lorentzen, T. D., Malek, T. B., Johnson, G. S., DeFrance, H. B., Ostrander, E. A. and Kruglyak, L. Genetic structure of the purebred domestic dog. *science*, 304(5674):1160–1164, 2004.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. and Reich, D. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

Patterson, N., Price, A. L. and Reich, D. Population structure and eigenanalysis. *PLoS genet*, 2(12):e190, 2006.

Pavlidis, P. and Noble, W. S. Matrix2png: a utility for visualizing matrix data. *Bioinformatics*, 19(2):295–296, 2003.

Perri, A. A wolf in dog's clothing: Initial dog domestication and pleistocene wolf variation. *Journal of Archaeological Science*, 68:1–4, 2016.

Pilot, M., Malewski, T., Moura, A. E., Grzybowski, T., Oleński, K., Ruść, A., Kamiński, S., Fadel, F. R., Mills, D. S., Alagaili, A. N. et al. On the origin of mongrels: evolutionary history of free-breeding dogs in eurasia. In *Proc. R. Soc. B*, volume 282, page 20152189. The Royal Society, 2015.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

Rambaut, A. and Drummond, A. Tracer v1. 5 available from http://beast. bio. ed. ac. uk/tracer. 2013.

Ramsey, C. B. et al. Bayesian analysis of radiocarbon dates. *Radiocarbon*, 51(1):337–360, 2009.

Reimer, P. J., Bard, E., Bayliss, A., Beck, J. W., Blackwell, P. G., Bronk Ramsey, C., Buck, C. E., Cheng, H., Edwards, R. L., Friedrich, M. et al. Intcal13 and marine13 radiocarbon age calibration curves 0-50,000 years cal bp. 2013.

Reiter, T., Jagoda, E. and Capellini, T. D. Dietary variation and evolution of gene copy number among dog breeds. *PloS one*, 11(2):e0148899, 2016.

Savolainen, P., Zhang, Y.-p., Luo, J., Lundeberg, J. and Leitner, T. Genetic evidence for an east asian origin of domestic dogs. *Science*, 298(5598):1610–1613, 2002.

Shannon, L. M., Boyko, R. H., Castelhano, M., Corey, E., Hayward, J. J., McLean, C., White, M. E., Said, M. A., Anita, B. A., Bondjengo, N. I. et al. Genetic structure in village dogs reveals a central asian domestication origin. *Proceedings of the National Academy of Sciences*, 112(44):13639–13644, 2015.

Skoglund, P., Ersmark, E., Palkopoulou, E. and Dalén, L. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology*, 25(11):1515–1519, 2015.

Skotte, L., Korneliussen, T. S. and Albrechtsen, A. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3):693–702, 2013.

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., Eichler, E. E. et al. Diversity of human copy number variation and multicopy genes. *Science*, 330(6004):641–646, 2010.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. Mega6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution*, 30(12):2725–2729, 2013.

Thalmann, O., Shapiro, B., Cui, P., Schuenemann, V. J., Sawyer, S. K., Greenfield, D., Germonpré, M., Sablin, M., López-Giráldez, F., Domingo-Roura, X. et al. Complete mitochondrial genomes of ancient canids suggest a european origin of domestic dogs. *Science*, 342(6160):871–874, 2013.

Vilà, C., Savolainen, P., Maldonado, J. E., Amorim, I. R., Rice, J. E., Honeycutt, R. L., Crandall, K. A., Lundeberg, J. and Wayne, R. K. Multiple and ancient origins of the domestic dog. *Science*, 276(5319):1687–1689, 1997.

VonHoldt, B. M., Pollinger, J. P., Lohmueller, K. E., Han, E., Parker, H. G., Quignon, P., Degenhardt, J. D., Boyko, A. R., Earl, D. A., Auton, A., Reynolds, A., Bryc, K., Brisbin, A., Knowles,

J. C., Mosher, D. S., Spady, T. C., Elkahloun, A., Geffen, E., Pilot, M., Jedrzejewski, W., Greco, C., Randi, E., Bannasch, D., Wilton, A., Shearman, J., Musiani, M., Cargill, M., Jones, P. G., Qian, Z., Huang, W., Ding, Z.-L., Zhang, Y.-p., Bustamante, C. D., Ostrander, E. A., Novembre, J. and Wayne, R. K. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature*, 464(7290):898–902, 2010.

Wang, G.-D., Zhai, W., Yang, H.-C., Wang, L., Zhong, L., Liu, Y.-H., Fan, R.-X., Yin, T.-T., Zhu, C.-L., Poyarkov, A. D. et al. Out of southern east asia: the natural history of domestic dogs across the world. *Cell research*, 26(1):21–33, 2016.

CHAPTER VI

Conclusions and Future Work

6.1 Conclusions

Recent advances in next generation sequencing technologies and statistical methods can be applied to answer the questions of population history for human and other species. The availability of numerous whole genome sequences poses both challenges and opportunities to upgrade our understanding of the past. In this dissertation, I have contributed to both methods development and intensive data analysis to understand the demographic and population separation history of human, gorilla and canine species.

In Chapter II, we analyzed fosmid pool sequencing data and resolved phased haplotypes of nine individuals from diverse populations (including Yoruba, Esan, Gambia, Massai, Mende, San, Mbuti, CEU, Gujarati). We physically phased 98% of heterozygous SNPs into haplotype-resolved blocks, obtaining a block N50 of 1 Mbp. We compared 1000 Genomes phase3 haplotypes (Consortium et al., 2015) with haplotypes obtained using fosmid pool sequencing and found 96.41% concordance in average, with switch error rate around 0.56%, indicating overall high quality of long range haplotypes, a result of a multi-stage phasing process that utilized a haplotype scaffold of trio-genotyped SNPs. However, when comparing haplotypes phased by ShapeIT (Delaneau et al., 2008) using 1000 Genomes Phase1 reference panel, the haplotype concordance dropped to 62.57% with average switch error rate 2.52%, and half of them flip errors, indicating the potential space for improvement. The level of switch errors existed in statistical phasing needs to be considered when applying statistically phased haplotypes to infer population history.

In Chapter III, we analyzed population size and separation history using the Pairwise Sequentially Markovian Coalescent (PSMC) and Multiple Sequentially Markovian Coalescent (MSMC) models on physically phased haplotypes (Li and Durbin, 2011; Schiffels and Durbin, 2014). Previous studies are only based on statistically phased haplotypes and the effect of switch errors on such inference is largely unknown. We find that statistically phased haplotypes yield a more recent split-time estimation compared with experimentally phased haplotypes, probably due to switch errors that make haplotypes appear more similar. To better interpret patterns of cross-population coalescence and provide quantitative measure of split times, we implemented an approximate Bayesian computation (ABC) approach (Toni et al., 2009) to estimate population split times and migration rates by fitting the distribution of coalescent times inferred between two haplotypes, one from each population, to a standard Isolation-with-Migration model. This extends PSMC model to model population split in a quantitative way. We inferred that the separation between hunter-gather populations and other populations happened around 120,000 to 140,000 years ago with gene flow continuing until 30,000 to 40,000 years ago; separation between west African and out of African populations happened around 70,000 to 80,000 years ago, while the separation between Massai and out of African populations happened around 50,000 years ago. We provide a comparison of PSMC and MSMC based methods with other contemporary methods on inferring population separation history and our results emphasize the importance of accurately phased haplotypes on MSMC analyses, especially for more ancient splits.

In Chapter IV, we analyzed medium to high coverage whole-genome sequences from 14 western lowland gorillas (*Gorilla gorilla gorilla*), 2 eastern lowland gorillas (*G. beringei graueri*), and a single Cross River individual (*G. gorilla diehli*). We infer that the ancestors of western and eastern lowland gorillas diverged from a common ancestor approximately 261 kyrs ago, and that the ancestors of the Cross River population diverged from the western lowland gorilla lineage approximately 68 kyrs ago. Using a diffusion approximation approach to model the genome-wide site frequency spectrum, we infer a history of western lowland gorillas that includes an ancestral population expansion of 1.4-fold around 970 kyrs ago and a recent 5.6-fold contraction in population size around 23 kyrs ago. The latter may correspond to a major reduction in African equatorial forests around the Last Glacial Maximum.

In Chapter V, we analyzed the whole genomes of an Early and End Neolithic dog from Germany, and another End Neolithic dog from Ireland. We found that ancient dogs demonstrate genomic continuity with each other and predominantly share ancestry with modern European dogs, contradicting a Late Neolithic population replacement previously suggested by analysis of mitochondrial DNA (Frantz et al., 2016). We also found that all ancient dogs possess ancestry from southeast asian dogs and that the end Neolithic sample from Germany possesses additional ancestry found in modern Indian dogs, indicating historical migration events. We applied G-PhoCS to infer a complete demographic model for dogs and wolves and further developed a numerical method to date the divergence time between ancient dogs and modern European dogs. By calibrating the mutation rate using our oldest dog, we narrowed the timing of dog domestication to a window of 20-40 kyrs ago. Given our older estimates of divergence time between Southeast Asian and European dogs compared to Frantz et al. (2016) and our best model of population demography, our results can most parsimoniously be explained by a single domestication process for dogs. Interestingly, we find no evidence of increased copy number for the *AMY2B* gene that has been hypothesized to have arisen as an adaptation to a diet rich in starch during the Neolithic, indicating that the selective sweep associated with this locus must have occurred after the Neolithic.

6.2 Future directions

There are potential space for improvement when applying statistical and computational method on whole genome sequences to infer population history.

In Chapter II, we applied fosmid pool sequencing data to phase SNPs and quantified the level of switch errors using statistical phasing approach. The same strategy can be applied to phase indels, namely short insertions and deletions. It's more difficult to accurately detect indels, let alone determining the phase of indels. Fosmid pool sequencing data can be used to test the accuracy of indel discovery as well as phasing performance. Besides, fosmid pool sequencing data are also valuable in validating structural variants, such as deletions, novel insertions and inversions. Deletions can be verified by observing gaps in computed read depth of 1kb windows and either shared across all pools as homozygous deletions or subset of pools as hemizygous deletion. Inversions can also be found by abnormal read pairs spanning clone boundaries. Fosmid pool sequencing data can be used to verify structural variants and differentiating them to be homozygous or hemizygous. Figure 6.1 shows an example of deletion and inversion events in NA19240 individual. Globally phased haplotypes data are crucial in population history inference.

In Chapter III, we implemented an Approximate Bayesian Computation approach to fit the inferred TMRCA distribution from PSMC to a standard isolation with migration model. This provides a quantified measure of split time and migration rate when applying PSMC on pseudo-diploid genomes to study population separation. However, we haven't explored the relationship of the MSMC model to Island-Migration models. Ideally, the coalescence rates within and between populations can also be parameterized to a standard Isolation with Migration model. Such parameterization is more straight-forward than looking at the relative cross coalescence curve to decide the time of population separation events. Moreover, there are many potential uses of phased haplotypes in demographic inference, such as studying more recent events, incorporating more complex models, studying introgression (Harris and Nielsen, 2013; Sheehan et al., 2013). Besides, other new methodologies are emerging. (Rasmussen et al., 2014) developed a highly efficient Markov chain Monte Carlo samplers for 'ancestral recombination graph' (ARG). This algorithm has been used to find signatures of natural selection. Incorporating this sampler into a full phylogenetic demographic model, such as the one used by G-PhoCS (Gronau et al., 2011) can generalize this fully Bayesian method to allow recombination and make full use of complete genome sequences.

In Chapter IV, we analyzed three gorilla subspecies, western lowland gorillas (*Gorilla gorilla gorilla*), Cross River gorillas (*G. gorilla diehli*), and eastern lowland gorillas (*G. beringei graueri*). Xue et al. (2015) performed whole genome sequencing on a mountain gorilla (*G. beringei beringei*)). Mountain gorillas are an endangered great ape subspecies and we know little about their evolutionary past. It would be interesting to know when two Eastern gorilla subspecies (eastern lowland gorilla and mountain gorilla) diverged and the migration pattern between mountain gorilla and other subspecies.

In Chapter V, we performed comprehensive population genetics analysis on two ancient dogs from the Neolithic that we sequenced and an additional Irish late Neolithic dog from Frantz et al. (2016). As more and more ancient dog genomes becoming available, the joint efforts from geneticists and archaeologists will continue to shed light on dog evolutionary history. Here we emphasize the correct handling of ancient genomic sequences to mitigate the effect of post-mortem damage, as well as integrating a collection of contemporary samples that well represent canine genetic diversity. Besides, biased sampling of canine individuals and ascertainment bias from SNP array based analysis may influence results and interpretation, as seen in the contentions on domestication origin between Boyko and Zhang's group (Wang et al., 2016; Shannon et al., 2016). Further analysis should incorporate model testing, such as the application of Approximate Bayesian Computation to confirm two domestication events of pigs (Frantz et al., 2015). Similar analysis can also be applied to test the two domestication model of dogs proposed by Frantz et al. (2016). However, it's hard to differentiating between 'dogs are domesticated from one extinct wolf population and then diverged' vs 'two populations of dogs are domesticated from two extinct wolf populations'. An ideal case would be sampling the appropriate ancient wolf population that dogs are domesticated from. Besides, it's interesting to examine positive selection on the dog lineage early in the domestication by performing Fst analysis on wolves and village dogs. Three ancient dogs at varying stages of time will help us roughly determine time point for certain advantageous mutation.

In summary, both methodology advances and proper sampling of populations of interest are crucial in demographic and population separation history inference.

Figures

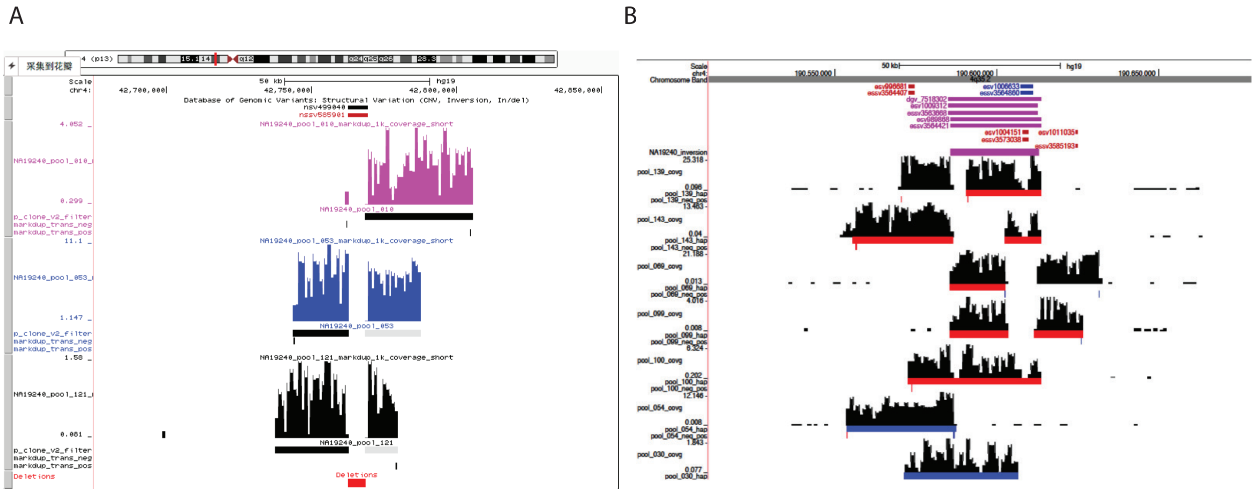

Figure 6.1: **Structural Variants discovered using fosmid pool sequencing data** A. Deletions can be discovered by observing gaps in computed read depth of 1kb window. B. Inversions can also be found by abnormal read pairs spanning clone boundaries.

Bibliography

Delaneau, O., Coulonges, C. and Zagury, J.-F. Shape-it: new rapid and accurate algorithm for haplotype inference. *BMC bioinformatics*, 9(1):1, 2008.

Frantz, L. A., Mullin, V. E., Pionnier-Capitan, M., Lebrasseur, O., Ollivier, M., Perri, A., Linderholm, A., Mattiangeli, V., Teasdale, M. D., Dimopoulos, E. A. et al. Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*, 352(6290):1228–1231, 2016.

Frantz, L. A., Schraiber, J. G., Madsen, O., Megens, H.-J., Cagan, A., Bosse, M., Paudel, Y., Crooijmans, R. P., Larson, G. and Groenen, M. A. Evidence of long-term gene flow and selection during domestication from analyses of eurasian wild and domestic pig genomes. *Nature genetics*, 47(10):1141–1148, 2015.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. and Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics*, 43(10):1031–1034, 2011.

Harris, K. and Nielsen, R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet*, 9(6):e1003521, 2013.

Li, H. and Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

Rasmussen, M. D., Hubisz, M. J., Gronau, I. and Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet*, 10(5):e1004342, 2014.

Schiffels, S. and Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919–925, 2014.

Shannon, L. M., Boyko, R. H., Castelhano, M., Corey, E., Hayward, J. J., McLean, C., White, M. E., Said, M. R. A., Anita, B. A., Bondjengo, N. I. et al. Reply to wang et al.: Sequencing datasets do not refute central asian domestication origin of dogs. *Proceedings of the National Academy of Sciences*, 113(19):E2556–E2557, 2016.

Sheehan, S., Harris, K. and Song, Y. S. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. *Genetics*, 194(3):647–662, 2013.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. P. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

Wang, G.-D., Peng, M.-S., Yang, H.-C., Savolainen, P. and Zhang, Y.-P. Questioning the evidence for a central asian domestication origin of dogs. *Proceedings of the National Academy of Sciences*, page 201600225, 2016.

Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., Frandsen, P., Chen, Y., Yngvadottir, B., Cooper, D. N. et al. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, 348(6231):242–245, 2015.