Bayesian maximum *a posteriori* algorithms for modern and ancient DNA

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet

Informatik

vorgelegt von

M. Math. Gabriel Renaud
geboren am 19. November 1980 in Montreal
Leipzig, den 1. Juli 2015

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Christophe Dessimoz (London, England)
2. Professor Dr. Peter F. Stadler (Leipzig)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 21.1.2016 mit dem Gesamtprädikat *magna cum laude*.

Renaud, Gabriel

Bayesian maximum a posteriori algorithms for modern and ancient DNA
Max-Planck Institute for Evolutionary Anthropology,
Leipzig University, Germany,
Dissertation 2015

222 pages, 118 references, 79 figures, 12 tables

I hereby declare that I am the sole author of this thesis.

I authorize the University of Leipzig and the Max Planck Society to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the University of Leipzig and the Max Planck Society to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Large portions of the text were taken from peer-reviewed articles of which I am the first author.

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

.....
(Ort, Datum)

.....
(Unterschrift)

Abstract

When DNA is sequenced, nucleotide calls are produced along with their individual error probabilities, which are usually reported in the form of a per-base quality score. However, these quality scores have not generally been incorporated into probabilistic models as there is typically a poor correlation between the predicted and observed error rates. Computational tools aimed at sequence analysis have therefore used arbitrary cutoffs on quality scores which often unnecessarily reduce the amount of data that can be analyzed. A different approach involves recalibration of those quality scores using known genomic variants to measure empirical error rates. However, for this heuristic to work, an adequate characterization of the variants present in a population must be available - which means that this approach is not possible for a wide range of species.

This thesis develops methods to directly produce error probabilities that are representative of their empirical error rates for raw sequencing data. These can then be incorporated into Bayesian maximum *a posteriori* algorithms to make highly accurate inferences about the likelihood of the model that gave rise to this observed data. First, an algorithm to produce highly accurate nucleotide basecalls along with calibrated error probabilities is presented. Using the resulting data, individual reads can be robustly assigned to their samples of origin and ancient DNA fragments can be inferred even at high error rates. For archaic hominin samples, the number of DNA fragments from present-day human contamination can also be accurately quantified.

The novel algorithms developed during the course of this thesis provide an alternative approach to working with Illumina sequence data. They also provide a demonstrable improvement over existing computational methods for basecalling, inferring ancient DNA fragments, demultiplexing, and estimating present-day human contamination along with reconstruction of mitochondrial genomes in ancient hominins.

Summary

In recent years, next-generation sequencing (NGS) platforms, combined with novel experimental protocols, have enabled the large-scale sequencing of ancient DNA (aDNA) molecules from fossils and forensic samples. The sequencing of aDNA has enabled researchers to gain insight into the history of extinct species. Despite its potential, aDNA has a few unique properties that distinguishes it from modern DNA. For instance, DNA molecules tend to undergo chemical damage which causes spurious signals of base substitution when the DNA is sequenced. Also, because of degradation, aDNA molecules generally tend to be shorter than the read length achieved by most sequencers. As a consequence, DNA adapters used by the NGS platforms are sequenced along with the aDNA fragment. Specialized programs are required to remove these adapters and infer the original aDNA fragment. Furthermore, contaminating DNA from present-day humans will be sequenced along with the endogenous DNA of the fossil. Present-day human contamination in aDNA samples from archaic humans is especially pernicious as it is highly similar to the endogenous DNA. Both endogenous and present-day human contamination will therefore align equally well to the human genome reference. This mixture of fragments makes the determination of the genetic make-up of the archaic individual particularly challenging.

In the past decade, NGS platforms have mostly supplanted the more classical chain-termination sequencing, which had been the most widely used method for reading DNA since the 1970s. As of writing, the Illumina sequencing technology is the most used NGS platform for both ancient and modern DNA. Although NGS platforms have the ability to sequence simultaneously millions of DNA reads, they tend to have a higher error rate than the chain-termination method. To quantify the probability of error, sequencers report a per-base quality score which represents the error probability on a logarithmic scale. However, many algorithms for computational biology do not make use of the quality scores produced by the sequencer. Furthermore, there is evidence that the quality scores produced by Illumina sequencers suffer from a lack of correlation between observed and predicted error rates.

This thesis demonstrates that if quality scores that are representative of observed error rates can be produced, Bayesian maximum *a posteriori* (MAP) algorithms can be used to make highly accurate and robust inferences from observed sequencing data without the need for arbitrary base quality cutoffs. Such MAP algorithms describe a set of computational techniques that use Bayes' theorem to estimate the most likely values for unobserved parameters given observed data and a prior probability distribution over the parameters one seeks to estimate. In this thesis, Bayesian algorithms were applied to the problem of reconstructing aDNA fragments from sequencing reads, assigning sequencing reads to their sample of origin and estimating the amount of present-day human contamination in aDNA extracted from archaic human fossils. To obtain accurate estimates, these algorithms require error probabilities for individually sequenced DNA bases that

are representative of their empirical error rates. This thesis also presents an algorithm to produce representative error rates for sequences generated using the Illumina platform.

Firstly, this thesis shows that this error can be accurately predicted using the distance to the decision hyperplane produced by support vector machines aimed at predicting individual bases from raw Illumina nucleotide intensity data. By using logistic regressions on those decision boundaries, accurate quality scores can be produced. Furthermore, the predicted bases using these support vector machines are more accurate than those obtained using the default software provided by Illumina.

Secondly, using those scores, highly accurate inferences of the original unobserved aDNA fragment from the sequence data can be made using a Bayesian MAP algorithm. As this algorithm does not require arbitrary cutoffs, the resulting reconstructed aDNA fragments are highly accurate at various levels of sequencing error. This algorithm outperforms current approaches in terms of accuracy on both simulated and empirical data.

Thirdly, using calibrated quality scores, individual sequencing reads can be assigned to the most likely sample of origin, a problem referred to as demultiplexing. This computational step is required when different samples have been pooled together and sequenced simultaneously. A Bayesian algorithm computes the posterior probability for each putative sample and assigns a sequencing read to the most likely one. This algorithm also quantifies the uncertainty of this assignment. This uncertainty score is directly correlated with the rate of erroneous assignments. Unlike current computational methods to demultiplex sequencing data, this novel algorithm can confidently assign individual sequencing reads to samples in spite of missing data and high error rates.

Finally, this thesis shows that highly accurate estimates of the rate of present-day human contamination in aDNA from archaic humans can be obtained using Bayesian MAP methods. Obtaining accurate estimates of the rate of present-day human contamination for aDNA datasets helps researchers in prioritizing samples. However, obtaining accurate present-day human contamination estimates for DNA extracted from ancient samples is particularly difficult due to the previously mentioned properties of aDNA. The Bayesian approach to estimating contamination presented in this thesis is more accurate than current approaches and is robust to uncertainties that are characteristic of aDNA.

The four applications described in this thesis are in active use and represent, as of writing, the most accurate algorithms for the respective problems they aim to solve. The applications described in this thesis demonstrate the advantages of using Bayesian and maximum-likelihood methods for the analysis of sequencing data.

Acknowledgments

First and foremost, I would like to thank my supervisor Janet Kelso for her supervision and support during the 4 years of my Ph.D. I also have to highlight the incredible contribution and mentorship I received from my colleague and friend Udo Stenzel.

Furthermore, I would like to thank my fellow colleagues from my institute who took time to teach me new concepts without which this thesis would not have been written. Thank you Matthias, Jesse, Susanna, Marie, Petra, Viviane, Mateja and Isabelle for your expertise in ancient DNA. Thank you to Aida, César, Felix, Fernando, Jin, João, Leonardo, Monty and Sergi for your knowledge in population genetics. I am indebted to Ana for her extensive knowledge of mitochondrial analyses. I would like to acknowledge the help of Tomislav Maričić and Victor Wiebe for their help in sequencing a dataset to test my algorithms. I would also like to thank Joana for pushing me to write an mitochondrial consensus caller aimed at ancient DNA. I would also like to thank Alex, Martin and Christoph for providing valuable insight. I want to also thank members of my group, Diana, Johann, Kay, Miguelito and Martin for their useful input about my work. I also would like to thank Peter Stadler from the University of Leipzig for his help with this thesis. Last but certainly not least, the genetics department would grind to a halt without Ines and Rigo for their help with system's administration and Viola for virtually every administrative tasks. I am grateful to all three for their outstanding diligence and dedication to duty.

From a personal side, I would like to thank my mom, my family and friends back home in North America who provided support and encouragements. Thank you Amneris, Andrew, Angèle, Carla, Cecilia, Daniel, Ester, Geneviève, Kareen, Lisa, Joan, Mike, Marc-André, Marie-Hélène, Nam, Robert, Vicky, Yonaira, and others ! My friends in Germany, Aida, Ammie, Anahita, Cleve, Lele, Ole, Raphaella, Steffi, Thomas and others, thank you for being there !

My time as a Ph.D student has been somewhat tarnished by the passing away of my officemate Ivan "Vano" Nasizde who suddenly died in front of my eyes in the office we shared. I wish he could have been there to read this acknowledgment.

Finally, I also would like to thank the Max Planck Society and the Natural Science and Engineering Research Council of Canada for providing financial support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The goal of this chapter is to provide an overview of the material necessary to understand this thesis. Two main concepts need to be covered. First, the computational challenges in the analysis of modern and ancient DNA are presented (pages 1-14). An introduction to Bayesian maximum *a posteriori* algorithms follows (pages 14-21). The chapter then shows how the Bayesian principle can be applied to DNA sequencing (see page 21) and finally presents the how the thesis is organized (page 24).

1.1 Sequencing modern and ancient DNA

1.1.1 DNA sequencing

DNA sequencing to study evolution

Living organisms store the information necessary for their maintenance and reproduction as deoxyribonucleic acid (DNA). DNA is formed from a backbone on which 4 different chemical side chains are attached: adenine (A), cytosine (C), guanine (G) and thymine (T). This backbone, with its side chains, pairs with another complementary copy to form a double strand. When left in a solvent, this double strand forms a double helical structure [112]. This information encodes the necessary instructions for protein synthesis and regulation [17]. It also encodes the information for non-regulatory ribonucleic acid (RNA), like ribosomal RNA for instance. This information is copied during cell division. For eukaryotic cells, the DNA is divided into structures called chromosomes which are located in the nucleus. The entire DNA contained in those chromosomes is called the nuclear genome. The mitochondrion is an extra-nuclear organelle that is maternally inherited and involved in cellular respiration. This organelle has its own genome however, it is composed of a single circular molecule which is much smaller than the nuclear genome.

During reproduction, organisms pass on this information to their offsprings [114]. However, this copy mechanism is not flawless and errors occur during the transmission of this DNA information [80]. While having an error at a given position in the genome does not follow predictable deterministic patterns, the process by which these variants are retained or eliminated is far from stochastic [31]. As the DNA of an organism has a direct influence on its observable characteristics or traits, certain novel variants will grant certain organisms a greater reproductive success and will rise in frequency in future generations. Other mutations may harm an organism's reproductive success or its own survival and will have greater difficulty finding its way to subsequent generations [18].

Due to those copying errors, DNA accumulates mutations as it is passed on from generation to generation [1]. Mutations can occur in the DNA passed on to one offspring but not necessarily to the other. Due to this independence of errors, individual mutations will begin to accumulate independently in the descendants of a common ancestor. Hence, given a constant mutation rate for each descendant lineage, each descendant will be, on average, equally distant to its common ancestor [52]. Each individual descendant will be closer in terms of molecular similarity to their common ancestor than they are to each other. This reasoning can be applied recursively on the common ancestor of these two sequences and another common ancestor with a third sequence. This history where each pair of sequences have a common ancestor which itself has one with a third sequence can be presented as a tree where the internal nodes represent common ancestors and, the terminal ones, the extant sequences [28].

While the history of a species is not known *a priori*, Felsenstein showed [29] that, given a model that quantifies the probability of a mutation occurring, the probability of observing a set of sequences given a putative tree can be computed. Furthermore, given that information and in the absence of prior knowledge, the most likely tree given a set of sequences can also be selected. This approach, generally called a maximum-likelihood approach to phylogenetics, enables researchers to reconstruct the evolutionary history of organisms given DNA sequences (see Figure 1.1).

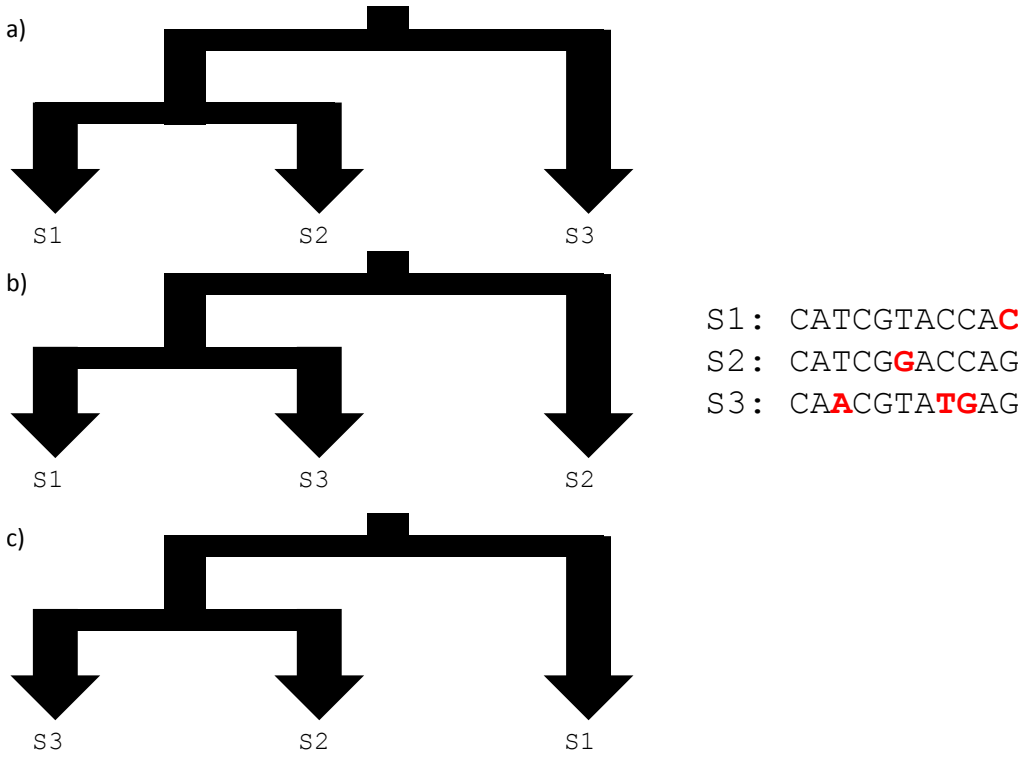


Figure 1.1: Representation of the maximum-likelihood principle for various phylogenetic reconstructions. a) tree where s3 is the outgroup, there are two mismatches in the (s1,s2) subtree and 3 in the long branch leading to the outgroup. b) tree where s2 is the outgroup and c) where s1 is the outgroup. For the last two trees, there are 4 mismatches in the short branch in the smaller subtree. The tree with the highest likelihood would be a).

DNA sequencing technologies

The first DNA sequencing technology [101] provided few sequences at a time. Developed by Frederick Sanger, this DNA sequencing method was called chain-termination sequencing or, more colloquially, as "Sanger sequencing". Improvements in the form of capillary sequencing automated the process by which DNA was sequenced and increases in throughput ensued. The DNA reads that can be obtained with this technology are around 1000 basepairs (bp) in length and have a very low error rate (from 0.001% [26] to less than 1% [51]). However, the number of reads obtained using this technology is very limited (order of 384 per sequencing run in up to 24 runs a day).

Since 2005, new technologies like 454 [69], ABI SOLiD [105], Ion Torrent Personal Genome Machine [99] and Illumina [6] have revolutionized the field of DNA sequencing by drastically increasing the number of bases being sequenced at a time. Although the

sequences generally do not have the same high quality as those generated by Sanger sequencing, the amount of data they make available allows research groups to sequence full genomes with unprecedented speed. These next-generation sequencing (NGS) technologies. They are able to produce millions of reads in a single sequencing run. However, these reads typically have a much higher error rate than Sanger. For instance, nucleotide miscalls ranging from 0.01% to 1.0% and a high rate of incorrectly inserted or deleted bases were reported for the 454 sequencing technology [104]. For the Illumina sequencing technology, error rates of 0.1% for single nucleotides and very low rates of inserted/deleted bases were reported [79]. Illumina and 454 sequencers normally produce read lengths between 70-120 basepairs (bp).

Due to the higher error rates for NGS technologies, research groups often view Sanger sequencing as the gold standard [73] and continue to validate variants using Sanger sequencing [116]. As of this writing, several technologies are on the market and many others are being developed. However, Illumina is currently the most widely used NGS platform [1].

To account for the higher error rates present in NGS technologies, computational methods like maximum-likelihood and Bayesian maximum *a posteriori* are ideally suited as they allow uncertainty to be directly modeled instead of discarding data. The algorithms that are described in this thesis are based on those computational approaches and were applied to the analysis of data generated using the Illumina sequencing technology. With the exception of the basecalling algorithm described in Chapter 2 these approaches can be generalized to any sequencing technology where the error rate is accurately quantified. Two out of the four applications described in this thesis are aimed at ancient DNA, which is DNA material recovered from either fossils or forensic samples, shortened as aDNA. The Illumina sequencing technology is briefly described and followed by introducing aDNA sequencing and sample multiplexing.

1.1.2 Illumina sequencing technology

Chemical principle

In 2006, Illumina purchased Solexa, a company that commercialized a platform which used the reversible terminator sequencing technology [6]. The first generation of the Genome Analyzer commercialized by Solexa could produce about 40 million (M) reads with a read length of 36bp. Two years later, the second generation of the Genome Analyzer was able to produce 125M reads of 2x75bp in length in each run. A new platform, the Illumina HiSeq 2000 which enabled the generation of 1.5 billion (G) reads of 2x100bp each was released in 2010. As of 2015, the HiSeq X Ten can produce 3G reads of 2x150bp each.

[1] http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_ign.pdf accessed: 06/24/15

This technology operates by the measuring the fluorescence obtained by laser excitement of fluorescently marked nucleotides paired to a sequencing template. Since multiple templates can be targeted in parallel and independently, this reaction can take place simultaneously for each template. The reason for the high throughput is due to the massive number of reactions taking place in parallel.

For the subsequent sequencing steps to work, the fragments must be sufficiently short to allow amplification. Hence, prior to sequencing, DNA must be fragmented to a certain size. The fragments go through a ligation procedure to attach sequencing adapters on both ends (see Figure 1.2A). These adapters bind to complementary sequences on the Illumina flowcell. In Chapter 4, one of those adapters will be referred to as the "p7" adapter and the other, the "p5" adapter. The flowcell is the solid surface on which sequencing reactions are carried out. The flowcell is covered with a lawn of complementary adapters (see Figure 1.2B). Once the adapter ligation is complete, the sequences are added to the flowcell and hybridize with random anchoring sequences on the flowcell surface (see Figure 1.2C). A flowcell is typically divided into lanes (typically 8 for an Illumina's HiSeq sequencer), each lane is further divided into 2 surfaces (top and bottom). Each surface is divided into laser swaths (currently 3) and each swath is divided into tiles (16 for a HiSeq). Anchored sequences are amplified to form clusters of identical sequences. For Illumina sequencing technology, this amplification uses solid-state synthesis as molecules are bound to the flowcell. This amplification process is referred to as cluster amplification (see Figure 1.2D). This process can suffer incorrect replication of the template where nucleotide replication errors occurred early on in the cluster amplification thus causing divergent nucleotides to be present for a given cycle. Another problem that can arise is that different fragments can bind in the same vicinity and form a population of mixed sequences that contribute to the same cluster.

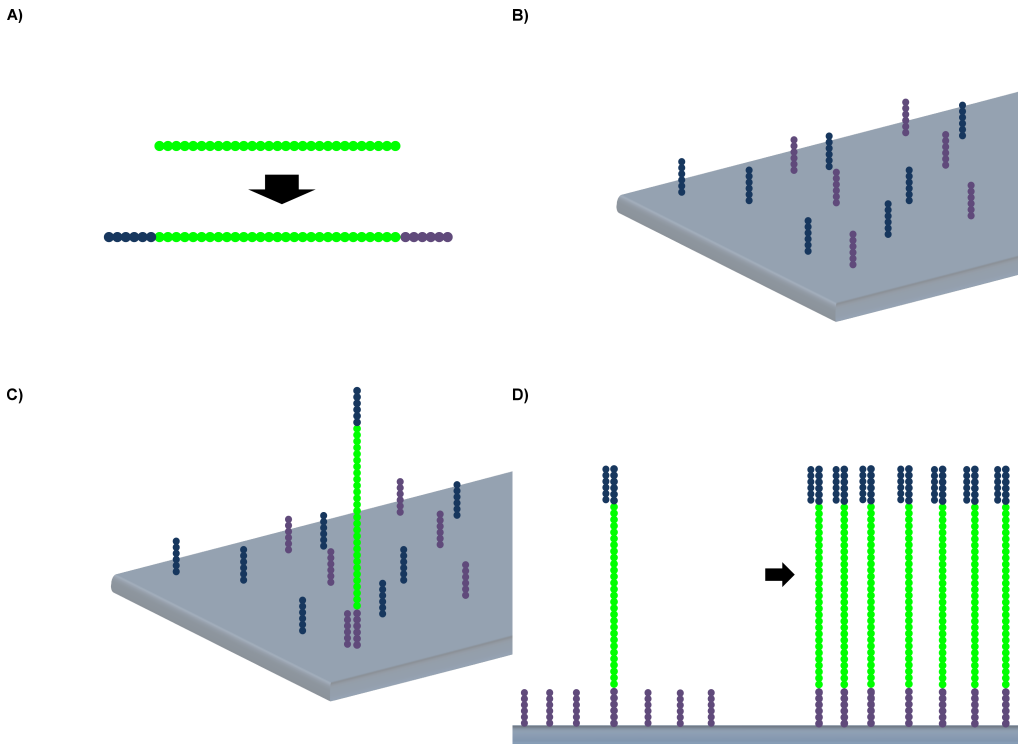


Figure 1.2: Schematic representation of Illumina's sequencing technology. A) The original fragment (green) is obtained from the DNA template. Two adapters, represented in purple and blue, are ligated to the original fragment. B) A flowcell has a lawn of DNA probes complementary to the ones ligated to the fragment. C) Once the fragment with the ligated adapters is added on the flowcell, they hybridize to the probes on the flowcell surface and are amplified. D) To obtain a sufficiently strong fluorescent signal to be able to read individual nucleotides, individual sequences need to be amplified into clusters of the same template. This process is referred to as cluster amplification.

Sequencing cycles

Sequencing primers specifically bind to one of the adapters (see Step 1) on Figure 1.3). Subsequently, dideoxynucleoside triphosphate (ddNTP) with terminators to avoid extension are added. These terminators contain a fluorescently marked label. Since, in theory, extension is not possible, every sequence in the cluster has the same nucleotide at the same position on their respective sequence.

Once the process of ddNTP integration is finished, it is followed by imaging. A series of lasers and filters are then used to produce intensities for all 4 possible channels. After this is completed for the first nucleotide adjacent to the primer, the fluorescently marked terminator is cleaved thus allowing for the incorporation of additional nucleotides (see

Step 2) on Figure 1.3). A second round of ddNTPs with fluorescently marked terminators are added to read the second base (see Step 3) on Figure 1.3). This process of adding ddNTP, measurements with lasers, cleaving the terminators is referred to as a cycle. As of this writing, Illumina HiSeq instruments generally perform around 100 cycles and MiSeq platforms support up to 300 cycles. The concatenation of all the bases obtained at each cycle for a given cluster is called a "read". The raw data captured from the Illumina measurement hardware is stored as large high definition images. The first steps of the internal analysis done by the Illumina sequencer are to identify the individual clusters and measure the intensities for each cluster, for each channel and for each cycle. These inferred intensities from the images will be simply referred to as the "intensities" in the rest of the document.

Once the final cycle is completed, sequences ligated to the flowcell are then bound from the other terminal end to the flowcell and reverse complemented. This process referred to as paired-end turnaround allows for another round of cycles to be performed on the other end. Data produced using this strategy is referred to as paired-end as both pairs stem from the end of the same DNA molecule. The two reads stemming from the same cluster are called "paired-end reads". The first read obtained before the paired-end turnaround is called the "forward read". The second one obtained after the paired-end turnaround is called the "reverse read".

To illustrate, say a sequence of 250 bp in length is subjected to 100 cycles of paired-end sequencing. It will yield the sequence of the first 100 bp of the molecule and the first 100 bp of its reverse complement which corresponds to the last 100 bases of the original molecule. However, 50 bases in the intervening portion will remain unsequenced.

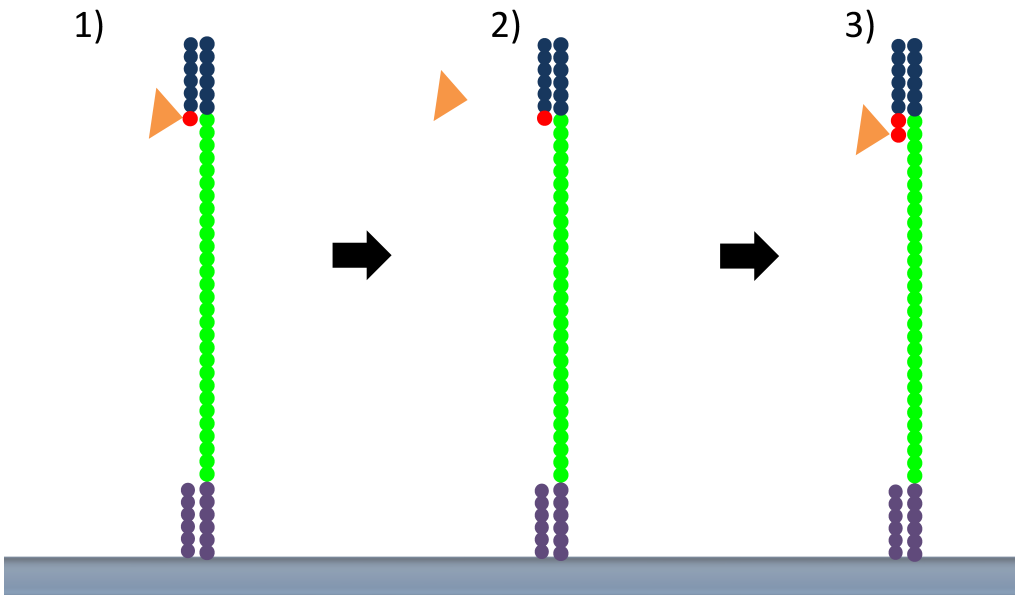


Figure 1.3: Representation a sequencing cycle for the Illumina sequencing technology. 1) A sequencing adapter is added and the first cycle begins where ddNTPs are added with a fluorescently marked terminator that prevents extension. 2) The first base is read and the terminator cleaved. 3) A second cycle begins where new ddNTPs are added and subsequently read.

Basecalling

The process by which individual bases along with their respective error probabilities are produced from these intensities is called basecalling. While identifying the bases from individual intensities might seem straightforward several factors make this process arduous:

1. **Cross-talk**: different nucleotides share the same laser and can only be distinguished using a filter. The fluorescent groups for ddNTPs complementary to A and C are excited by the same laser and a filter is needed to distinguish them (see Figure 1.4). The same holds for ddNTPs complementary to G and T. An A will cause a strong A signal and a medium C signal while a C will cause a low A signal and a medium C signal. This means that certain signals are often difficult to interpret accurately (e.g. a medium signal in both the A and C channels).

2. **Phasing**: the accumulation of fluorescent signal in different channels due to accelerated or delayed incorporation of ddNTPs. The process of hybridization of ddNTPs to the template is not flawless in each cycle. A fraction of sequences of a given

cluster will not have any correctly incorporated ddNTPs. This results in a subset of molecules in each cluster lagging behind in the integration of ddNTPs. This leads to some molecules in the cluster being out of phase and giving the incorrect signal. While this effect is hardly noticeable in early cycles, in later cycles the added contribution of the lagging ddNTPs dominates a large fraction of the signal (see Figure 1.4). **Pre-phasing** refers to the opposite effect where ddNTPs are running ahead of the current cycle and are also creating incorrect fluorescent signals.

3. **Fading**: the waning of the intensity of the fluorescence. As fewer ddNTPs are hybridized on the template, the overall intensity of the fluorescence for the correct nucleotide tends to fade (see Figure 1.4).

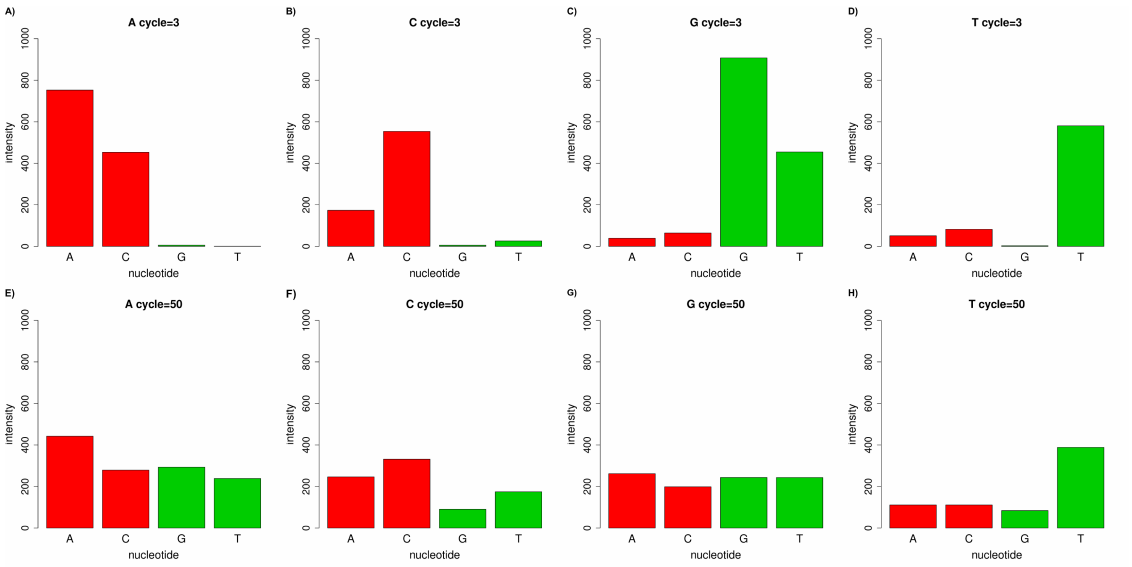


Figure 1.4: Empirical intensities for all 4 nucleotides for cycle 3 and 50 for a sequencing run from a Genome Analyzer II. For cycle 3, the raw intensities for all 4 nucleotides for all 4 channels are plotted (plots A through D). The same intensities were plotted at cycle 50 (plots E through H).

Due to these artifacts, identifying the bases in later cycles or due to weak or mixed signals is a difficult computational problem. By default, Illumina uses a software called Bustard part of the CASAVA standard pipeline. Bustard has a model for the cross-talk matrix, phasing and fading, it then estimates parameters for those based on sequenced data and calls the bases and their respective quality scores given the raw intensities.

Sequencing errors

These three phenomena make basecalling a difficult computational problem. A basecaller must not simply produce the most likely base given the intensities but also quantify the uncertainty of the call. To inform downstream processes of this uncertainty, an estimate of the probability of sequencing error is produced with each individual base. As probabilities are often small fractional numbers (e.g. 0.0001), they are therefore hard to represent in the output of a basecalling program. A potential solution is to encode each base on a logarithmic scale. The most commonly used logarithmic scale is the PHRED scale (see [27]). Let ϵ be the error probability, the quality score q on the PHRED score (ϵ_{phred}) is represented by:

$$q = -10 \cdot \log_{10}(\epsilon) \tag{1.1}$$

The expression above is often rounded to the nearest integer. To save disk space, a single character for each base can be used to represent this error probability using the ASCII table which establishes a correspondence between characters and their numerical representation. Often, an offset to the first non-space character in the table, the "!", which corresponds to the 33rd character is used. Therefore, the "!" character represents a quality score of 0 on the PHRED scale and represents the baseline. For example, the "+" sign is 10 characters above the "!" on the ASCII table and represents a quality score of 10 on the PHRED scale. Trivially, the original error probability can be computing using the following expression:

$$\epsilon = 10^{\frac{q}{-10}} \tag{1.2}$$

It is worth noting that while there is a non-enumerable infinity of error probability which ranges between 0 and 1, there is a finite number of characters on the ASCII table which leads to loss of information.

Multiplexing

The high throughput of NGS platforms can be beneficial for genome resequencing or *de novo* sequencing as the length of an entire eukaryotic genome requires large amounts of data to provide decent coverage. For smaller loci (e.g. targeted capture or mitochondrial sequencing) however, such high-throughput provide more coverage than is required. Methods have therefore been developed to sequence multiple samples on the same sequencing run, a process referred to as multiplexing. One possible technique is to add one unique, very short nucleotide sequence per sample in the middle of the priming adapters defined earlier. This unique sequence is referred to as an "index", sometimes referred

to as "barcode". In this thesis, the use of the word index over barcode is deliberate as barcode can also refer to the species identification barcodes [45].

Once sequencing of the forward read is completed, the sequence immediately adjacent to the index is used as primer and the bases of the index are read (see Figure 1.5). Illumina makes available a list of 96 indices, each of 7 base pairs in length, such that the edit distance between any possible pair is at least 2. However, this strategy is no longer viable for 100 samples as indices would have to be used twice. To address this problem, an alternative strategy is to sequence a second index on the remaining adapter. Instead of selecting a unique index for each sample, unique pairs of indices have to be chosen. This increases the number of samples that can be pooled to 9,216.

The computational process of identifying the sample of origin for each sequenced read is referred to as demultiplexing (see Figure 1.5).

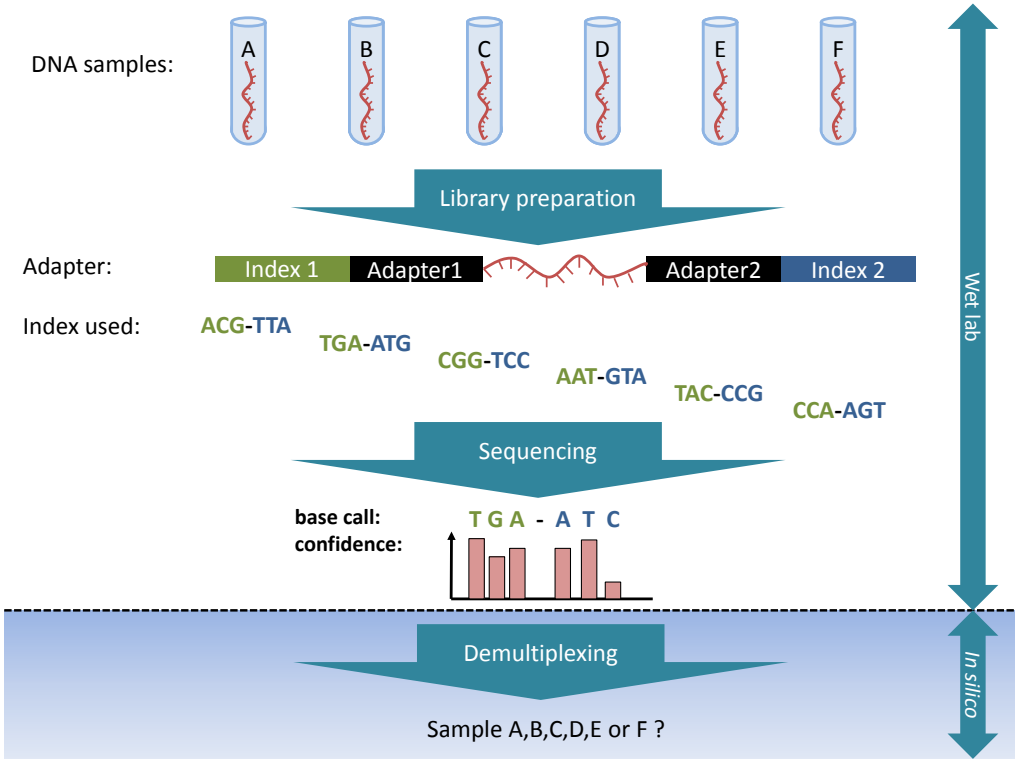


Figure 1.5: Schematic representation of Illumina's multiplexing strategy. For a double-indexing protocol, each sample is assigned a unique pair of indices on its adapters. The indices are sequenced at a different stage than the forward and reversed portions of the fragment. Once basecalling is complete, the *in silico* task of identifying the original sample for a given observation is called demultiplexing.

1.1.3 Sequencing ancient DNA

The extraction of DNA from forensic samples or fossils requires specialized wet-lab techniques. The Illumina sequencing technology has been instrumental in enabling the production of large aDNA datasets. This subsection discusses the characteristics of aDNA. A number of features distinguish the type of sequencing data obtained from modern and ancient samples. This section aims at presenting 3 different features of aDNA: fragment length, post-mortem damage, and contamination.

Fragment length

As previously mentioned, DNA molecules extracted from living organisms are naturally too long to be directly sequenced. A process referred to as shearing is used to stochastically break the long molecules to allow Illumina sequencing. This is achieved using exposure to ultrasounds which can be targeted for the desired average fragment size [22].

However, the DNA fragments extracted from ancient material such as bones, teeth or hair are typically highly degraded with average molecules lengths ranging from 20-50 nucleotides depending on the age and preservation of the sample [90].

When sequencing ancient DNA, the read lengths therefore often exceed the length of the ancient molecule. This may result in the inclusion of all or part of the adapter sequence in the sequencing reads. For instance, using 100 cycles on a fragment of 70 bp will result in sequencing the original fragment for the first 70 cycles and sequencing the adapter sequence for the last 30 cycles. If the paired-end mode is used, then the 70 bases of the fragment will be sequenced twice (see Figure 3.1 in Chapter 3 for an illustration). For aDNA, the step where adapters are removed is therefore crucial.

Due to issues specific to ancient DNA, throughout this thesis, the word "fragment" will refer to the original DNA sequence between the adapters. The word "read", will refer to the raw sequence produced by the sequencer which, in the case of ancient DNA, could be a concatenation of the fragment and the adapter. The word "sequence" will refer to any generic DNA sequence.

Post-mortem damage

Apart from degradation resulting in short fragment lengths, some cytosines will tend to lose their amino groups due to hydrolysis [7]. This transforms some cytosines into uracils, a base normally associated with RNA. During DNA sequencing, uracils are read as thymine. Greater rates of deamination have been reported at the end of fragments potentially as a result of single-stranded overhangs [7, 39]. As hydrolysis requires contact with the solvent, it is likely the overhangs are more exposed than the rest of the molecule.

Deaminated sites may affect downstream analyses as they can appear as mutations or as heterozygous (C/T) sites [85, 92]. However, despite its negative impact on downstream analyses, deamination can be useful to determine whether the DNA extracted actually stemmed from the fossil in question and not solely from potential contaminants. There are sometimes doubts in the scientific field about whether the sequenced fragments genuinely came from the DNA of the fossil [14, 41, 88]. As deamination rarely occurs in DNA molecules that are less than a century [102], this deamination signal can also be used to ascertain the authenticity of ancient DNA [84, 106].

Contamination

Bacterial and fungal contamination When extracting DNA from ancient hominin remains, microbial DNA often forms the bulk of all recoverable fragments [87, 117]. The sequences of the aDNA fragments are generally aligned to a particular reference genome. However, due to their high divergence to a mammalian reference sequence, for example, these fragments will not align at the same rate as the endogenous material. Due to this, the total number of sequences aligning has often been used to estimate the percentage of endogenous DNA [42]. These bacterial and fungal sequences will also align mostly due to random similarly between their DNA and the genome reference for the endogenous sample. As the length of two random sequences increases, the lesser the probability that they are identical [8]. Since, as previously mentioned, aDNA fragments can be very short and due to the volume of microbial DNA, such bacterial DNA fragments will find themselves aligned to a mammalian reference at a greater rate in aDNA than in modern DNA sequencing data.

Present-day human contamination In addition to the DNA from bacteria and fungi, contaminating DNA from individuals that handled the ancient sample, is sequenced along with the ancient material [3]. When working with archaic humans such as Neanderthals, aDNA fragments are typically aligned to the human genome. While bacterial sequences do not typically align to the human reference genome, present-day human contaminants will align together with the endogenous DNA fragments.

The presence of contaminant fragments affects both consensus calling and genotyping (determining the most likely genotype given the sequencing data), and the resulting errors may influence comparisons to present-day humans including the calculations of genotype likelihoods, divergence times, parameters of population demography and phylogenetic reconstructions [85, 111].

Such contaminating fragments from present-day humans are less likely to be deaminated than the endogenous DNA [102]. Further, ancient fragments tend to shorter than modern contaminating DNA fragments due to degradation of aDNA [39, 42, 59].

1.2 Bayesian maximum *a posteriori* methods

1.2.1 Bayes' theorem

The probability of an event, denoted $P(event)$ is a real number between 0 and 1 that quantifies the chance that this event has or will happen. For instance, the probability that a fair, 6 sided, dice yields 4 at a given cast is $\frac{1}{6}$. However, let us assume that someone replaces the dice with a flawed one where the odd numbers $(1, 3, 5)$ have been respectively replaced with $(2, 4, 6)$. This flawed dice has sides with duplicated even numbers $(2, 2, 4, 4, 6, 6)$. For such a dice, the probability of obtaining 4 is no longer $\frac{1}{6}$ but rather $\frac{1}{3}$ as, two events with probability $\frac{1}{6}$ can yield a 4.

It is now trivial to write the probability of obtaining 4 given that either one of the dices was chosen. Such probability is referred to as a conditional probability. To define the probability of seeing 4 given that either one of the dices has been cast using standard notation, the following expression is used:

$$P(4|fair) = \frac{1}{6}$$

$$P(4|flawed) = \frac{1}{3}$$

While this is intuitive, asking the question the other way around: "What is the probability that the flawed dice was chosen given the fact that 4 was observed ?" is less trivial. To derive this quantity $P(flawed|4)$, let us note that the conditional probability of seeing 4 given that the dice is flawed is the fraction of the probability space defined by the probability of picking the flawed dice, that is occupied by the event of both picking the flawed dice and obtaining 4:

$$P(4|flawed) = \frac{P(4 \cap flawed)}{P(flawed)}$$

The quantity above is of course $\frac{1}{3}$. The quantity $P(flawed|4)$ follows the same logic. The fraction of the event "throw yielded 4" occupied by the event "throw yielded 4 and dice was flawed" is essentially:

$$P(flawed|4) = \frac{P(4 \cap flawed)}{P(4)}$$

Since $P(4 \cap flawed)$ is present in both equations, they can be equated to yield Bayes' theorem:

$$P(flawed|4) = \frac{P(4|flawed)P(flawed)}{P(4)}$$

The sought quantity is therefore the fraction of the probability space of the event "throw yielded 4" occupied by the product of the probability of seeing 4 given that the flawed dice was picked times the probability that the flawed dice was picked to begin with. If each dice was equally likely to have been picked, the probability of obtaining 4 can be computed by adding the probability of either one of the dices yielding a 4. The probability that the flawed dice was thrown given that the throw yielded a 4 is:

$$\begin{aligned}
P(flawed|4) &= \frac{P(4|flawed)P(flawed)}{P(4)} && (1.3)\\
&= \frac{P(4|flawed)P(flawed)}{P(4|flawed)P(flawed) + P(4|fair)P(fair)} && (1.4)\\
&= \frac{\frac{1}{3}\frac{1}{2}}{\frac{1}{3}\frac{1}{2} + \frac{1}{6}\frac{1}{2}} && (1.5)\\
&= \frac{\frac{2}{12}}{\frac{2}{12} + \frac{1}{12}} && (1.6)\\
&= \frac{\frac{2}{12}}{\frac{3}{12}} && (1.7)\\
&= \frac{2}{3} && (1.8)\\
& && (1.9)
\end{aligned}$$

Given that an even number was obtained and either one of the dices could have been used, there is a $\frac{2}{3}$ chance that the flawed one was the dice that was used. The prior probability of the flawed dice being the correct one ($P(flawed)$) is called the "prior". Conditioning on a certain dice having been picked (ex: flawed), the probability of seeing the data (ex: 4) which is denoted as $P(4|flawed)$ in the equations above, is the "likelihood". The denominator $P(4)$ is the probability of seeing a 4 given any possible model and is called the "evidence". The final probability $P(flawed|4)$ is called the "posterior".

Generally, Bayes' theorem can be therefore written as:

$$posterior = \frac{likelihood \cdot prior}{evidence} \tag{1.10}$$

Notice that in the example above, the prior of the flawed dice was $\frac{1}{2}$ but after seeing the evidence of the data, the posterior became $\frac{2}{3}$. Notice that the probability of the fair dice being the one that was originally used for the same observation is $\frac{1}{3}$. It can be

concluded that, after having seen an even number, it is twice as likely that the flawed dice had been used. If the most probable dice given the observation is needed without needing a precise probability for it, the following can be written:

$$posterior \propto likelihood \cdot prior \tag{1.11}$$

1.2.2 Bayesian theory for the sciences

> *"We can always prove any definite theory wrong. Notice however we never prove it right"*
>
> - Richard Feynman,

The simple example of the Bayesian theory has a greater conceptual underpinning. The dices represent unattainable mental representations of reality or a "model". The most likely model among the two (flawed and fair dice) that could have given rise to a particular observation (a cast of 4) was selected. Furthermore, the confidence in that model being the correct one was computed, the flawed model was twice as likely as the fair one. As the quote above illustrates, science cannot prove that a particular theory is correct. The only remaining strategy is to compute the probability that a certain model is the correct one. In contrast, it also cannot prove that any theory is "wrong" in absolute terms, it can only say that the probability of such a model is so unlikely that it can be safely discarded.

Such Bayesian framework allows scientists to measure the probability of a certain model being the correct one given data. More specifically, the parameters that will yield the highest posterior probabilities are desired hence the name Bayesian maximum *a posteriori* (MAP). For the inferences of parameter values given a model, Bayesian methods have an edge over statistical or frequentist methods as they do not require quality cutoffs as uncertainty can be directly built within the model. Current modern and aDNA sequencing data have various sources of uncertainty when it comes to determining the pre-mortem DNA sequence from the original host:

- sequencing errors
- erroneous mappings
- deamination of certain bases, particularly at the end
- contaminant from present-day humans for ancient early modern humans and archaic hominins

A simple example of inferring the probability of generating "head" for a biased coin with observations with some introduced errors is presented. First, how a frequentist would approach the problem is described then a MAP approach is presented.

Frequentist approaches

Suppose there is a biased coin toss that yields head with probability h and tails with probability $1-h$ where h is not necessarily $\frac{1}{2}$. Given a set of coin tosses, an estimate of the value of h can be obtained using:

$$\hat{h} \approx \frac{\#head}{\#head + \#tail} \tag{1.12}$$

Bayesian approaches

In a Bayesian framework, the posterior for a particular value of h given the data is proportional to:

$$\binom{\#head + \#tail}{\#head} (h^{\#head} \cdot (1-h)^{\#tail}) \cdot P(h) \tag{1.13}$$

Maximum-likelihood methods, in contrast, try to find h with the maximum value for the likelihood term in equation 1.13 by removing $P(h)$. Bayesian methods however, apply a probability on the prior belief ($P(h)$) in the model. Finding the value of h that yields the highest posterior (denoted $\hat{h}$), assuming a uniform prior for h, will yield:

$$\hat{h} = \frac{\#head}{\#head + \#tail} \tag{1.14}$$

Thus far, Bayesian methods may seem unimpressive as the most likely value of h can be obtained using a simple counting method. However, in science, there is always a need to include the error in any measurement. Statistical methods can account for the number of observations by using confidence intervals. Despite this, they cannot properly account for uncertainty in the data i.e. erroneous observations. If the confidence in the measurement is properly quantified, a Bayesian approach can account for it by including the uncertainty into the computation of the posterior.

To illustrate, let us suppose that the process of reading the result of the coin toss is imperfect and that uncertainty has been quantified. For each ith observation, let ϵ_i be the probability of error ($0 \leq \epsilon_i \leq 1$). At an error probability of 0, the measure of the coin toss is perfect. At an error probability of 1, the coin toss can be read either as head or tail with equal probability.

To account for error, a frequentist approach would be to set a quality cutoff and compute the value of h using equation 1.12. The problem however, is that the estimate of h varies considerably given the cutoff used.

A Bayesian approach would be to modify equation 1.13 to account for error. For N independent coin tosses $t = t_1, t_2, ..., t_N$ where t_i is either head (H) or tail (T), the posterior probability of observing the data is equal to the following expression:

$$\prod_{i=1}^{N} P_{obs}(t_i) \tag{1.15}$$

where the probability of observing t_i is determined by considering that the original (i.e. before an error could have occurred) ith coin toss, denoted T_i, could have been either head or tail. This gives us:

$$P_{obs}(t_i) = P_{obs}(t_i|T_i = H)P(T_i = H) + P_{obs}(t_i|T_i = T)P(T_i = T) \tag{1.16}$$

The likelihood of observing the data depends on whether the observed coin toss matches the model (T_i). If both match, then either no error has occurred or, an error did occur, with probability ϵ, and t_i was obtained by chance ($\frac{1}{2}$). If they do not match, then the only possibility is that an error has occurred:

$$P(t_i|T_i) = \begin{cases} (1-\epsilon) + \frac{\epsilon}{2} & \text{if } t_i = T_i \\ \frac{\epsilon}{2} & \text{if } t_i \neq T_i \end{cases} \tag{1.17}$$

Finally, the prior probability ($P(T_i)$ in equation 1.16, are defined as follow:

$$P(T_{ki}) = \begin{cases} h & \text{if } P(T_i = H) \\ 1-h & \text{if } P(T_i = T) \end{cases} \tag{1.18}$$

If $\epsilon = 0$ and errors are impossible, equation 1.15 becomes equation 1.13 without the binomial which corresponds to the likelihood of a specific observation. If $\epsilon = 1$ and the entire observation corresponds to errors, equation 1.15 simply becomes $\frac{1}{2^N}\ \forall\ h$. Hence, when error is very high, nothing is learned from the observed data and every h is equally likely to have generated the observation. Had a prior distribution on h been applied, the posterior probability distribution would have matched the prior distribution, again indicating that the only source of information comes from the prior, not the observed data.

Comparing methods for parameter inference

To compare both methods, $n = 1000$ coin tosses can be generated along with errors with a simulated probability of generating head of $h = 0.2$. An error means that the coin toss is uninformative and either head or tail was generated with equal probability ($\frac{1}{2}$). Data has the following format:

```
call Q_PHRED error
T 10.9594 0.0801794
T 14.5165 0.0353464
H 1.27187 0.746127
T 8.00988 0.158129
T 10.9735 0.0799184
T 1.51651 0.705259
...
```

where the first column is the call, the second the error probability on a unrounded PHRED scale and the third column, the raw error probability.

The frequentist approach was used using various quality cutoffs. The estimate for h was obtained using equation 1.12. Confidence intervals were computed using a binomial proportion confidence interval:

$$z \cdot \sqrt{\frac{\hat{h}(1-\hat{h})}{n}} \tag{1.19}$$

where z is 1.96 representing a 95% confidence interval.

The resulting estimates using the frequentist approach were plotted in Figure 1.6. There are two problems with this approach. First, using liberal cutoffs, the estimate is skewed towards $\frac{1}{2}$ due to the inclusion of low quality bases. Second, at stricter cutoffs, the estimate has such a broad confidence interval as to make difficult any conclusion as to the value of h. A problem with frequentist approaches is to determine which cutoffs are appropriate. Furthermore, comparing two different datasets with drastically different error rates will become arduous.

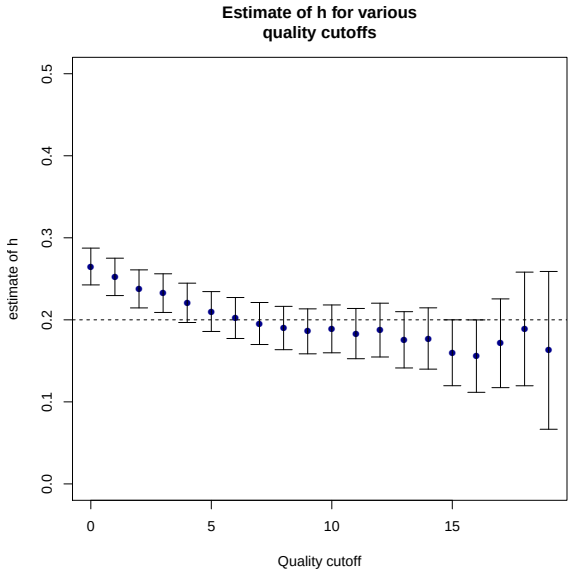


Figure 1.6: Estimates the probability of a coin of generating head (h) provided by a counting and cutoff method. Using various cutoffs, the estimate for h is given by equation 1.12. The whiskers represent a 95% confidence interval using a binomial proportion confidence interval.

For the Bayesian approach, the posterior probability for the dataset was plotted in Figure 1.7. The advantage of this approach is that the posterior probability peaks at 0.2. This simple approach also obviates the need for cutoffs thus allowing datasets with different error rates to be compared.

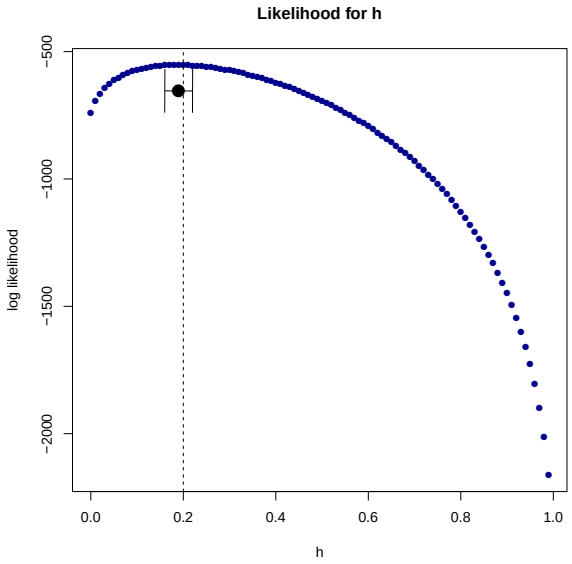


Figure 1.7: Distribution of the logarithm of the posterior probability for h, the probability of generating head, provided by a Bayesian method. The black dot represents the highest posterior probability and the whiskers represent the 95% confidence interval for this distribution.

1.2.3 Probabilistic approach to inference for sequencing

Let Ω be the set of all DNA bases ($\Omega = \{A, C, G, T\}$). Let $b \in \Omega$ be a certain base on a read and let ϵ be its error rate. In a Bayesian framework, the likelihood of the data given the model is often desired, $P(data|model)$. The model here is the actual base that was in the original fragment that went on the flowcell during sequencing. Let $b' \in \Omega$ be this specific base. There are two ways to construe error probabilities:

1. An error will generate any base but the correct one
2. An error will generate any base at random

Under model 1, the probability of observing b given that b' was the correct base is given by:

$$P(b|b') = \begin{cases} 1 - \epsilon & \text{if } b = b' \\ \epsilon \cdot \frac{1}{3} & \text{if } b \neq b' \end{cases} \tag{1.20}$$

however, under model 2, equation 1.20 becomes:

$$P(b|b') = \begin{cases} 1 - \epsilon + \frac{\epsilon}{4} & \text{if } b = b' \\ \epsilon \cdot \frac{1}{4} & \text{if } b \neq b' \end{cases} \tag{1.21}$$

Please note that, regardless of the model used, the probabilities for every case sum up to 1. There are advantages and disadvantages for each. One advantage of model 2, is the ability to have error probabilities that reach 1 thus making every nucleotide equally likely to have given rise to b. In model 1, the maximum error probability is 0.75. At low error rates, the choice of either one has a significant impact. Despite this, at higher rates of base quality, both models produce almost identical base probabilities.

However, it was reported that not every base substitution equally likely in Illumina sequencing data [81]. One distinct advantage of model 1 is the ability to incorporate empirical base substitution rates instead of using substitutions with equal probabilities ($\frac{\epsilon}{3}$). Using this, equation 1.20 becomes:

$$P(b|b') = \begin{cases} 1 - \epsilon & \text{if } b = b' \\ \epsilon \cdot P(b' \to b) & \text{if } b \neq b' \end{cases} \tag{1.22}$$

where $P(b' \to b)$ is the probability that b' gets observed as b in the final read. In equation 1.20, that probability was simply $\frac{1}{3}$. However, using empirical data, better estimates can be used. If, for instance, an A is more likely to be observed as a C than a G given a sequencing error, $P(A \to C)$ will be greater than $P(A \to G)$.

A disadvantage of model 2 is the inability to quantify the rate of nucleotides that, given that a sequencing error has occurred, get read as the original nucleotide by chance. Model 1 can incorporate such empirical substitution rates. Therefore, model 1 (equation 1.20 and 1.22) is used throughout this thesis. It is worth noting that $P(b'|b)$, the probability that b' is the correct base can also be inferred using Bayes' rule:

$$P(b'|b) = \frac{P(b|b') \cdot P(b')}{P(b)} \tag{1.23}$$

The probability of seeing the original base, b', as well as the probability of seeing base b in the read, are considered to be both equally likely ($\frac{1}{4}$). One could add the *a priori* probability of seeing particular bases however, that would necessitate the characterization of the organism being sequenced in advance which is often not feasible. Assuming $P(b') = P(b) = \frac{1}{4}$, equation 1.23 becomes simply:

$$P(b'|b) = P(b|b') \tag{1.24}$$

This allows us to quantify:

- The probability of observing a particular base in the read given a base in the original sequence bound to the flowcell or,
- The probability of observing a particular base in original sequence bound to the flowcell given a base in the read

However, this quantification is predicated on the principle that the predicted error probability is indicative of the actual one.

The need for cutoff-free methods for sequencing

As shown in section 1.2.2, a Bayesian method that incorporates the uncertainty within the model allows us to make inferences without the use of arbitrary cutoffs. This section motivates why such approaches are highly suitable for next-generation sequencing.

Sequencing groups routinely find that error rates differ between sequencing runs, however, quality can also differ from the location of the clusters on the flowcell. To illustrate this, the expectancy of the number of mismatches was computed for an Illumina MiSeq run with a flowcell with 2 surfaces, 2 swaths each and, in turn, 16 tiles each. The expected number of mismatches was computed for a sequence of length L using the following expression:

$$\frac{\sum_{l=1}^{L} 10^{\frac{-q_l}{10}}}{L} \tag{1.25}$$

where q_l is the reported error rate on the PHRED scale (see equation 1.1).

The heatmap for the expected number of mismatches for each combination of surface/swath/tile was plotted (see Figure 1.8). The error rate increases as a function of the tile number for both surfaces. This figure also shows that surface 1 generally has a worse error rate than surface 2. Other groups have also observed this discrepancy between error rates for various parts of the flowcell [79].

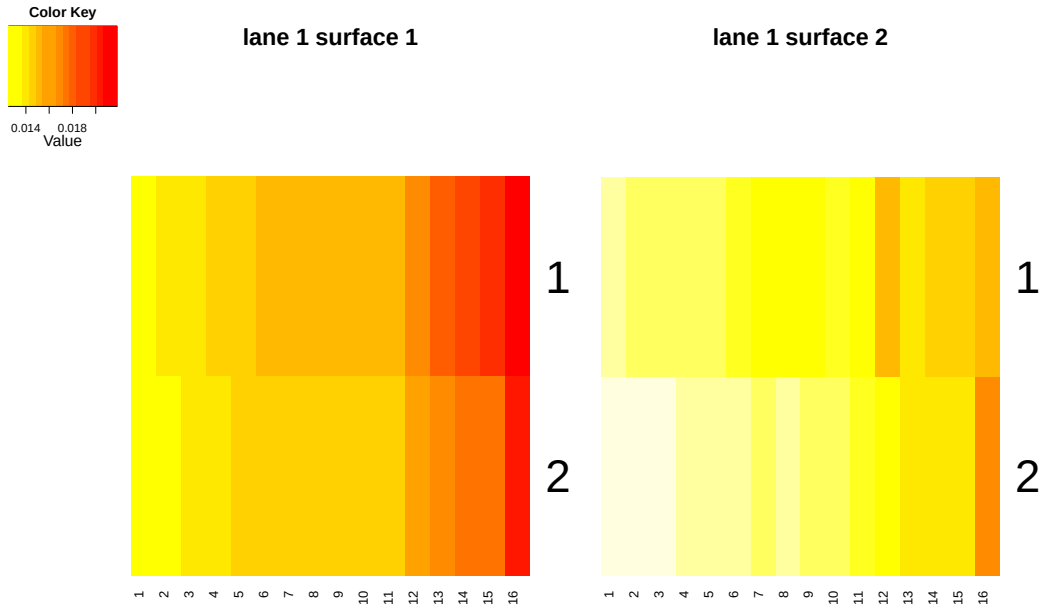


Figure 1.8: Heatmap of the expected mismatch rate for an Illumina MiSeq run from 2014 with 2 surfaces (left and right), 2 swaths (y-axis) and 16 tiles (x-axis). The expected number of mismatches varies depending on the location of the clusters on the flowcell thus making the use of general quality cutoffs difficult.

These results illustrate how it is disadvantageous to use a single quality threshold for the entire flowcell as error rates differ within the same flowcell.

1.3 Overview of the thesis

This thesis is built on the principle that, given representative error probabilities for individual sequenced DNA bases, Bayesian models can be built to make inference using the data. In greater detail, the work in this thesis presents the following algorithms:

1. **freeIbis**: Representative error probabilities for Illumina sequencing data can be predicted using logistic regression on distances to the decision boundaries from support vector machines. (Chapter 2, page 27)
   reference:
   Gabriel Renaud, Martin Kircher, Udo Stenzel, and Janet Kelso. freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. *Bioinformatics*, 29(9):1208–1209, 2013.

2. **leeHom**: Using 1, Bayesian models can be used to infer the most probable DNA fragment that gave rise to sequencing reads. (Chapter 3, page: 47)
   reference:
   Gabriel Renaud, Udo Stenzel, and Janet Kelso. leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Research*, 42(18):e141, 2014

3. **deML**: Using 1, Maximum-likelihood methods can be used to infer the most likely sample of origin for each individual read from multiplexed sequencing runs. (Chapter 4, page: 67)
   reference:
   Gabriel Renaud, Udo Stenzel, Tomislav Maricic, Victor Wiebe, and Janet Kelso. deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics*, 31(5):770–772, 2015

4. **schmutzi**: Using 1, 2 and 3, the resulting data can be used to infer the endogenous mitochondrial genome and quantify present-day human contamination rates for aDNA datasets. (Chapter 5, page: 91)
   reference:
   Gabriel Renaud, Viviane Slon, Ana T. Duggan, and Janet Kelso. schmutzi: Contamination estimate and endogenous mitochondrial consensus calling for ancient DNA. submitted, 2015

A diagram represents how these programs rely on the output of freeIbis (see Figure 1.9).

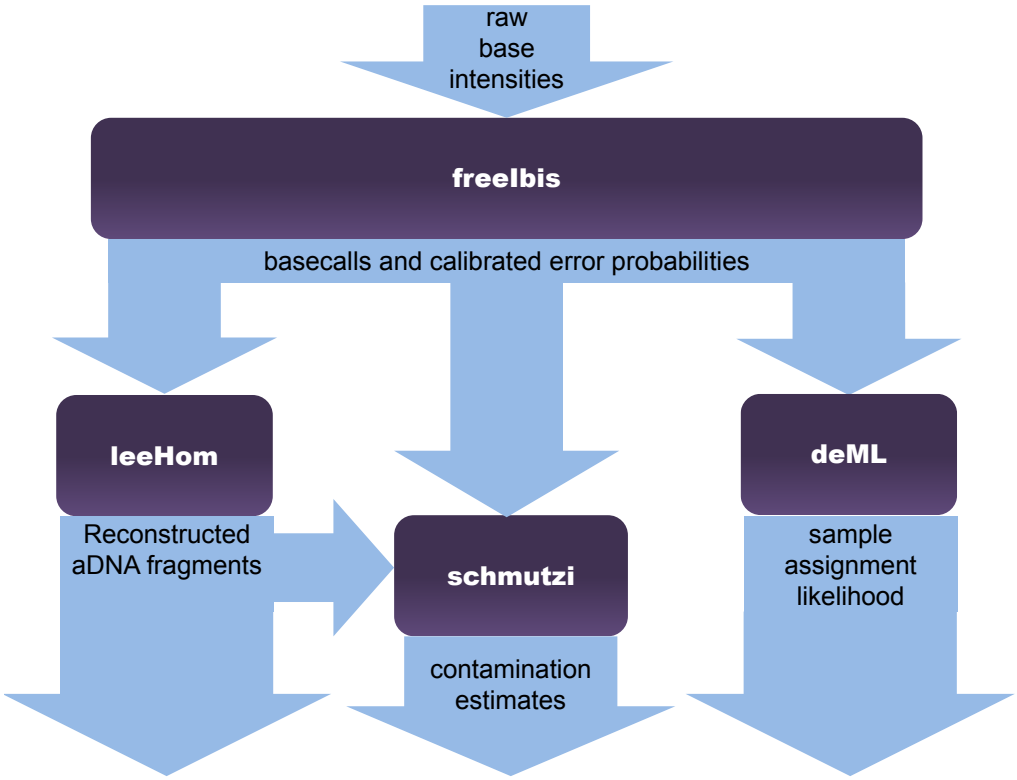


Figure 1.9: Representation of how freeIbis, leeHom, deML and schmutzi depend on each other. All of the three downstream programs, leeHom, deML and schmutzi require freeIbis' output to compute Bayesian probabilities. Finally, schmutzi also needs leeHom's reconstructed aDNA fragments to compute present-day human contamination estimates.

Previous work done in the field for each individual problem is presented at the beginning of each chapter. The description of how to obtain such accuracy for error probabilities is presented (Chapter 2). Using those error probabilities, how to reconstruct aDNA fragments via MAP methods is explained (Chapter 3). A maximum-likelihood algorithm to robustly assign individual reads to the most likely sample of origin is then presented (Chapter 4). Finally, methods for inferring the endogenous mitochondrial genome despite heavy present-day human DNA contamination and algorithms to quantify this contamination for ancient hominin samples are described (Chapter 5). An overall conclusion about the applicability of the principles that undergird this thesis is found in the final chapter (Chapter 6) and is followed by appendices.

Chapter 2

Basecalling with calibrated quality scores

This chapter presents freeIbis, a third-party basecalling algorithm for Illumina sequencers that produces calibrated quality scores.

2.1 Background

A crucial step in the Illumina sequencing pipeline is basecalling: the generation of individual nucleotide sequences and associated quality scores, which measure the probability of a sequencing error, from raw intensities. To evaluate the applicability of a basecalling program for practical purposes, 3 aspects must be taken into account:

1. Accuracy of the nucleotide sequences being produced
2. Quality scores that adequately represent the probability of error
3. Reasonable runtime

The default basecaller provided by Illumina, Bustard, develops a model from the raw intensities and uses it to perform basecalling. This model estimates the quantities defined in section 1.1.2 namely, the cross-talk matrix, phasing and pre-phasing. It then applies corrections to the measured intensities for cross-talk and then corrects for phasing and pre-phasing [1].

[1] http://supportres.illumina.com/documents/myillumina/ec3129a6-b41f-4d98-963f-668391997f1a/olb_194_userguide_15009920d.pdf accessed: 06/30/15

In an attempt to either provide more accurate sequences or quality scores than the default Bustard basecaller, third-party basecallers have been described in the literature. Since a lesser number of basecalling errors have the potential to produce more accurate and simpler downstream analyses, designing improved basecallers has been an active research field [61]. Certain of these alternative basecallers reported achieving a better performance than Bustard [113]. These basecallers can be divided into those that apply an unsupervised modeling strategy like Bustard and those that rely on supervised learning approaches or intermediate approaches (Altacyclic [25]).

2.1.1 Unsupervised modeling approaches

The aim of unsupervised modeling methods for basecalling is to infer parameters for the cross-talk matrix, pre-phasing, phasing and fading (see Section 1.1.2). This can be achieved by applying a mathematical model that includes each parameter. Using empirical intensities, the basecaller finds values for these internal parameters. The advantage of such methods is that they do not require control data like PhiX for training.

Different strategies have been suggested to model Illumina sequencing data. BayesCall[49] tries to incorporate cycle-dependent parameters into a Bayesian framework and predicts the bases with the highest posterior probability (see naiveBayescall [48] or [19] for faster implementations). To improve on this, All your Base (AYB) [72] tries to consider the entire sequence to infer model parameters rather than a few neighboring cycles. Furthermore, under the assumption that neighboring clusters on the flowcell might have parameters that are more similar, AYB allows for basecalling on a per-tile basis. Finally, BlindCall (see [115]) treated the problem of basecalling as a blind deconvolution problem, which aims at recovering a latent signal given an observed one [62].

There are two types of model-based algorithms, those that build a single model for the entire flowcell and those, like AYB, that use a single model per tile. Currently, basecallers that use a different model for each tile perhaps have the potential to be the most accurate in terms of the sequence they produce as they can capture local effects on the flowcell (see results in Section 2.4.3 later in this chapter). A downside however, is that they are often too slow to suit the needs of even small or mid-size sequencing centers.

2.1.2 Supervised learning approaches

The supervised learning strategy uses as training data a small viral genome that has been previously sequenced. The virus used is the PhiX174 (shortened to "PhiX" throughout the thesis) GenBank ID: J02482.1 but where the reference sequence was provided by Illumina Inc. Such data establishes a correspondence between the intensities and the nucleotides they represent. A downside to this is the requirement that a sufficient number of control sequences must be available for every sequencing run.

Altacyclic [25] used a combination of SVMs and internal parameters to achieve basecalling. In the original publication, the authors reported improved base accuracy over the default Illumina basecalls. The Bioinformatics group at the Max Planck Institute for Evolutionary Anthropology (MPI-EVA) has previously developed Ibis [55], a basecaller that uses a multiclass support vector machine (SVM) with a linear kernel to predict the most likely nucleotides. Briefly, an SVM with a linear kernel seeks a fit a hyperplane between labeled multidimensional training data such that this hyperplane divides data points with different labels. This hyperplane is fitted given a cost function that penalizes misclassifications. In the case of basecalling, Ibis seeks to fit a hyperplane using label intensities with their respective nucleotides determined from the PhiX genome (see Figure 2.1).

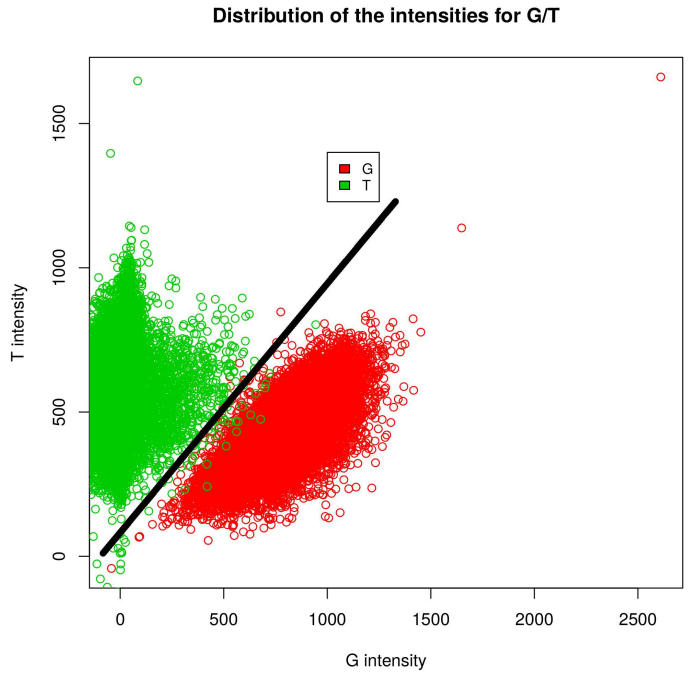


Figure 2.1: The distribution of the G and T intensities from a Genome Analyzer II for nucleotides identified as either G or T using the PhiX genome. The G nucleotides have high intensities for both G and T due to cross-talk and the T nucleotides have low G and high T intensities. A black line represents a putative hyperplane that separates the space into a G and T prediction given intensities.

The quality scores were produced by assuming that the distances to the decision boundaries were normally distributed. Therefore, the probability of error of a given prediction could be inferred by the relative position of its distance to its decision boundary to the overall distribution of such distances.

While Ibis produced nucleotides that were more accurate than the default Illumina basecaller, certain drawbacks remained:

1. The SVM library used (SVM-Light [46]) had a non-commercial license thus preventing uses in industry.
2. Bases on the PhiX genome reference showed signs of divergence. This meant that a majority of empirical bases supported a different nucleotide than the one from the reference. This meant that erroneous training examples were given to Ibis.
3. The quality scores produced using the heuristic described above do not correlate well with the observed error rate. This lack of correlation was observed by both the Bioinformatics group at the MPI-EVA and another research group [72] independently.

The last point is particularly important as downstream analyses require quality scores to distinguish sequencing errors from genuine mutations. To illustrate this, PhiX sequencing data basecalled with the latest version of Ibis was aligned to its genome. For each possible quality score, the fraction of mismatches over the total was measured. For a predicted error rate of 1 in 10,000, one mismatch per 10,000 aligned nucleotides, on average, should also be observed in the aligned data. However, the results show that the predicted error rates produced by Ibis do not correlate well with their observed ones (see Figure 2.2). Moreover, they are not strictly increasing (i.e. a nucleotide with a predicted error rate of 1/10,000 has a greater error probability than a nucleotide with a predicted error rate of 1/1,000).

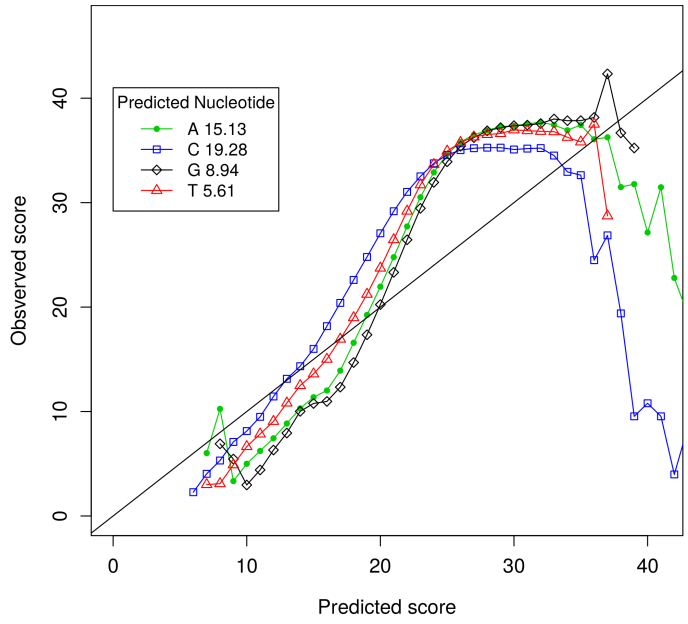


Figure 2.2: The predicted versus observed error probabilities from a Genome Analyzer II for sequences aligned to the PhiX reference genome. The 4 color represents the predicted nucleotide. The diagonal line represents a perfect prediction. While the predicted quality scores do not correlate well with observed ones, the same quality scores for different nucleotides have different observed error rates. Also, nucleotides with a lower predicted error rate have a higher observed error rates. The numbers in the legend represent the root-mean-square error of the predicted error rates to the observed ones.

Another reason for the continued development of Ibis is that, despite the high basecalling accuracy, tiled-based predictions are often impractically slow for everyday use (see results in Section 2.4.3 later in this chapter). To illustrate this, BlindCall[115] reports a minimum runtime of 4-8 minutes for a single tile, which is faster than the time the authors obtained for AYB. However, for an entire HiSeq run, basecalling 8 lanes X 2 surfaces X 3 swaths X 16 tiles would require a minimum of 2.1-4.2 days. As reported in the original Ibis publication, inferring a single model over the entire flowcell offers superior accuracy than simply using the Bustard basecalls and sequencing centers can do so at a feasible runtime.

2.1.3 Quality score recalibration

The Genome Analysis Toolkit (GATK) [74] is a series of programs aimed mostly at genotyping. As part of their recommended steps prior to genotyping, the GATK development team advises users to recalibrate the raw quality scores from Illumina sequencers

to address the problem of the imperfect correlation between predicted and observed error rates.

The software package provides subprograms to recalibrate the quality scores for aligned sequencing data. Briefly, it uses known single nucleotide polymorphisms (SNPs) to measure empirical error rates and thus modifies the quality scores from the input BAM. While this approach might yield good results for whole genome resequencing of humans from Eurasia, it might cause biases for certain highly divergent African populations where the extent of mutations has not been characterized. Further, this approach is not applicable for aDNA stemming from archaic humans like Neanderthals. Also, this approach is not applicable for non-human species as a well-characterized set of variants is needed from the same species as the sample for this recalibration strategy to work.

A basecaller that produces sequences whose accuracy ideally exceeds that of the default basecaller from the vendor, at a reasonable runtime and directly produces calibrated quality scores was needed.

2.2 Introduction

To address the issues presented in section 2.1, an update to the Ibis basecaller is introduced, rechristened freeIbis. FreeIbis replaces the SVM library with a restricted license by LIBOCAS [32] which is released under the GNU Public License. Results show LIBOCAS offers an optimal performance in terms of basecalling and that freeIbis outperforms the previous version of the basecaller, Ibis, in terms of sequence accuracy.

How the decision score of the SVM corresponded to the observed error rate was measured. A function approximating this distribution is then used to assign quality scores for individual bases. The resulting scores show a high level of correlation between their observed error rate and the predicted one, thus obviating the need for quality score recalibration as a post-processing step [74]. The newest versions of freeIbis and Ibis were compared against the default basecaller for two Genome Analyzer II (GA) runs, a HiSeq run and a MiSeq run. On a set of DNA sequences genotyped using both Sanger and Illumina sequencing technologies, freeIbis provides an improvement in genotype accuracy over the default Illumina basecaller.

2.3 Methods

This section presents how various SVM libraries were tested for their accuracy in terms of basecalling (see Section 2.3.3), a strategy to eliminated mislabeled training examples (see Section 2.3.2) and how calibrated quality scores can be produced (see Section 2.3.3).

2.3.1 Testing SVM libraries

To evaluate the performance and accuracy of various SVM libraries for basecalling, a small dataset containing 51 cycles from a PhiX control lane sequenced on a GAII using an earlier version of the chemistry was used. The objective is not to produce usable sequences but rather, to evaluate how different libraries perform on training and testing datasets representing actual intensities. It is worth noting that, in previous versions of the Illumina chemistry, an accumulation of the ddNTP for the T base induced spurious high-intensity values for the T channel which was mostly noticeable in later cycles. This dataset is therefore befitting for the purpose of testing machine learning techniques as predicting bases in later cycles becomes difficult due to increased phasing. To account for phasing, the intensities of the previous and following cycle from the same cluster are used as additional features.

The performance and accuracy was compared for LIBOCAS v0.93[33], LIBLINEAR v1.8[12] (with option 4 for the multi-class SVM) and SHOGUN v0.10.0 [108] (with LIBSVM_MULTICLASS) against SVM-Light v2.12 [46], the SVM library used in Ibis. The results indicated that LIBOCAS offered overall superior accuracy at a lesser required training time than remaining methods (see section 2.4.1). Similar results regarding the accuracy of LIBOCAS compared to other methods for different classification problems were reported by the LIBOCAS authors. This library was therefore embedded into the base-calling software.

2.3.2 Masking divergent positions on the PhiX

The control reads, generally the PhiX phage, provide the SVM library with training data to find an optimal set of hyperplanes to divide the feature space into the various labels which are, in freeIbis' case, the 4 different base pairs. These hyperplanes are subsequently used to assign a label to the intensities given as input. The optimal hyperplanes are derived by determining parameters that minimize a cost function which penalizes misclassified training examples proportionally to their respective decision score. Therefore, numerous mislabeled training examples (i.e. set of intensities for an adenine labeled as a thymine) could have an influence on the set of hyperplanes determined by the algorithm. A manual look at the training data revealed that these artifacts were in fact present and were probably the results of two distinct causes: genuine sequencing errors and divergent bases in the control genome population. The former would cause mismatches to

the reference that would be scattered across the genome while the latter, would create mismatches at specific bases with a clear bias for a given nucleotide. The mismatches for each base of the PhiX genome as a fraction of the coverage were plotted (see Figure 2.3) and, to distinguish between random sequencing errors and actual divergent positions, the plots were separated according to the substituted base pair on the read. These effects can be hard to disentangle however, both cause the introduction of mislabeled examples in the training and testing datasets. As the mismatches were concentrated around certain positions and a clear substitution bias for certain nucleotides were observed, a masking procedure for these positions on the genome of the organism used as control was implemented. Thus, any training or testing example created from a position having more than 10% of its coverage represented as a mismatch to a given nucleotide was removed. Sequence patterns indicating systematic sequencing errors (SSE) described by Nakamura *et al.*[81] were not disproportionately found around these masked positions.

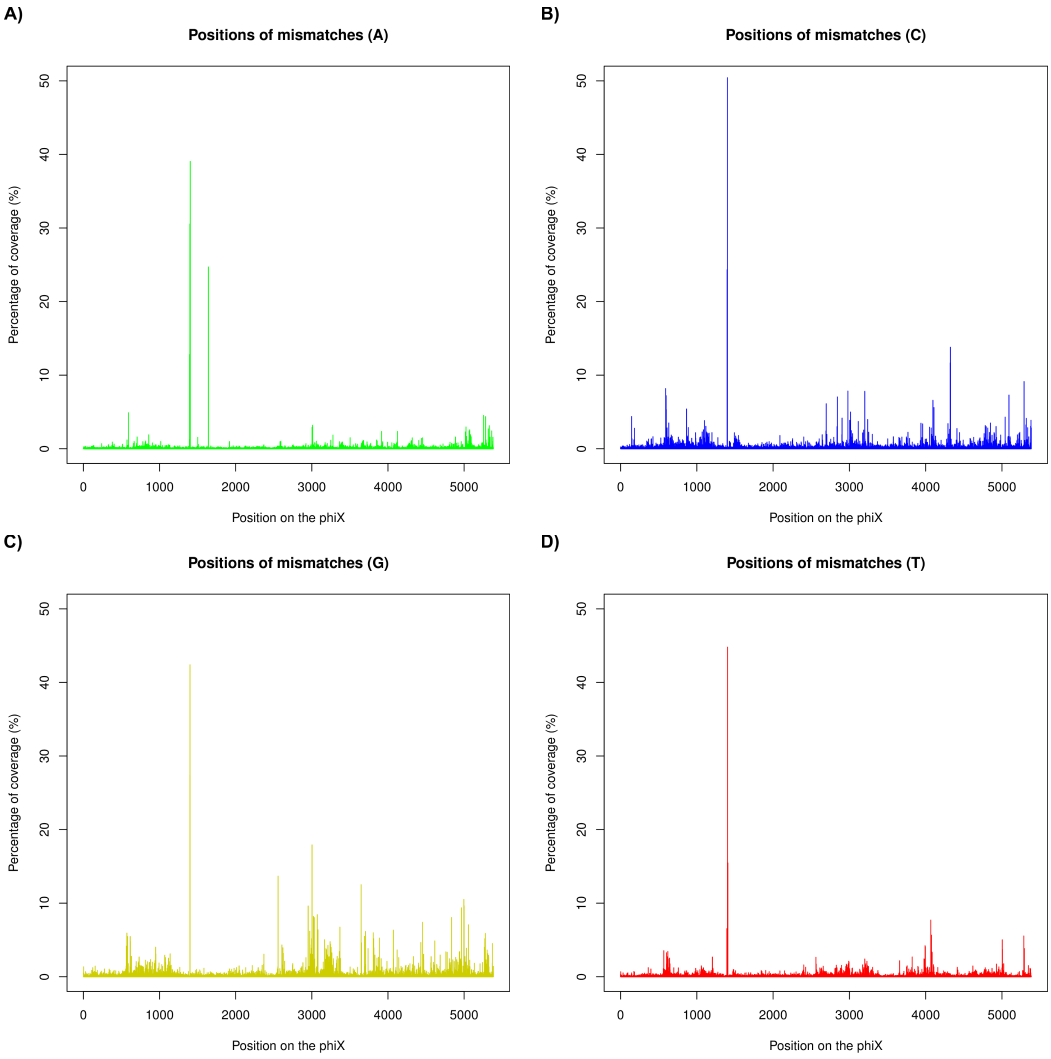


Figure 2.3: The distribution on the PhiX genome of mismatches as a ratio of coverage for control reads for a GAII run. The plots indicate the mismatches from the reference to: A (subfigure A)), C (subfigure B)), G (subfigure C)), T (subfigure D)

2.3.3 Quality Score Calibration

As the divergent bases on the PhiX were masked, whether the posterior probabilities of the SVM corresponded with the observed error rate was evaluated. However, standard implementations of the SVM algorithm do not output posterior probabilities but decisions values for each hyperplane. A key observation is that the rate of misclassifications decreases as the distance to the hyperplane increases (see schematic representation in

Figure 2.1).

The previous version of the basecaller (Ibis) sought to predict the quality based on empirically observed SVM decision scores and misclassification rates. However, the new version offers the possibility of calibrating SVM decision scores to observed errors using a logistic regression whose confidence probability score is, in turn, correlated to sequencing quality. This calibration is computed on every cycle and each nucleotide position within the sequence reads.

The library used for support vector machines produces, along with class assignments, decision score values associated for each nucleotide (δ_A, δ_C, δ_G, δ_T). A method for calibrating these values into actual posterior probabilities was described by [86] which proposes that this posterior probability can be modeled using a logistic function:

$$p_{error} = \frac{1}{1 + e^{-z}} \tag{2.1}$$

where z is defined by :

$$z = b + a_A \cdot \delta_A + a_C \cdot \delta_C + a_G \cdot \delta_G + a_t \cdot \delta_T \tag{2.2}$$

The parameters b, a_A, a_C, a_G, a_t from equation 2.2 are obtained using the logistic regression function from LIBLINEAR v1.8[12]. For the ith base being predicted, freeIbis uses as input its decision scores (δ_{Ai}, δ_{Ci}, δ_{Gi}, δ_{Ti}) from the SVM and the empirical errors e_i such that: $e_i = 0$ if the predicted base matches the template and $e_i = 1$ otherwise. This allows freeIbis to compute the set of z_i values for each predicted base.

To compute the error rate for an average value of z, overlapping windows of n observations are used where the average error is simply $\frac{\sum_{i \in n} e_i}{n}$. The size of the window is adjusted to encompass a fixed number of mismatches such that windows at high error rates are smaller than those at low error rates. The result is a set of average z values ($\hat{z}_i$) and estimated error rate $\hat{\epsilon}_i$ such that $0 < \hat{\epsilon}_i < 1$.

However, as quality scores are computed on a PHRED scale defined as $-10 \cdot \log_{10}(p_{error})$ and since p_{error} is modeled using the logistic function, plotting the input of the linear function z from equation 2.2 against the logarithm of the observed error rate (i.e. ϵ_i on a PHRED scale) would be expected to follow a somewhat linear relationship (see Figure 2.4).

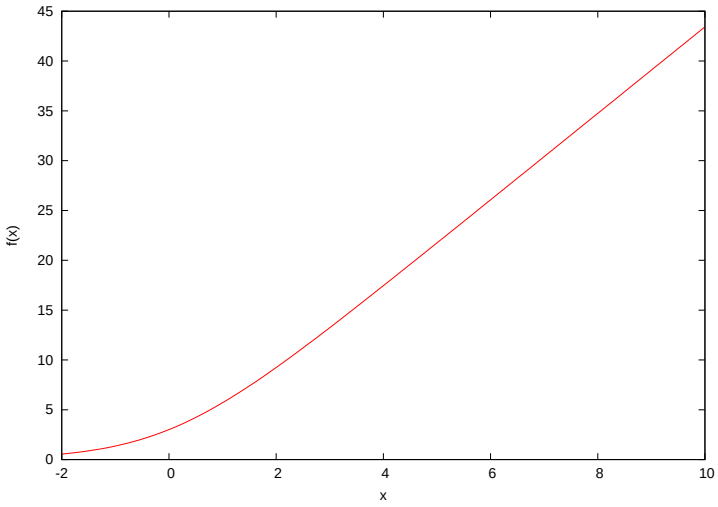


Figure 2.4: The plot of the error probability on a PHRED scale when a logistic function is used to represent such a probability ($f(x) = -10 \cdot \log_{10}(1 - \frac{1}{1+e^{-x}})$). At values of x greater than 4, the log of the logistic function behaves like a linear function.

However, it was empirically determined that, despite this relationship being linear for the earlier scores, it reaches a quality score plateau induced by the background error rate of the procedure (see Figure 2.5). For high quality sequencing runs (e.g. a HiSeq with recent chemistry and normal cluster density) this plateau usually hovers around 40.

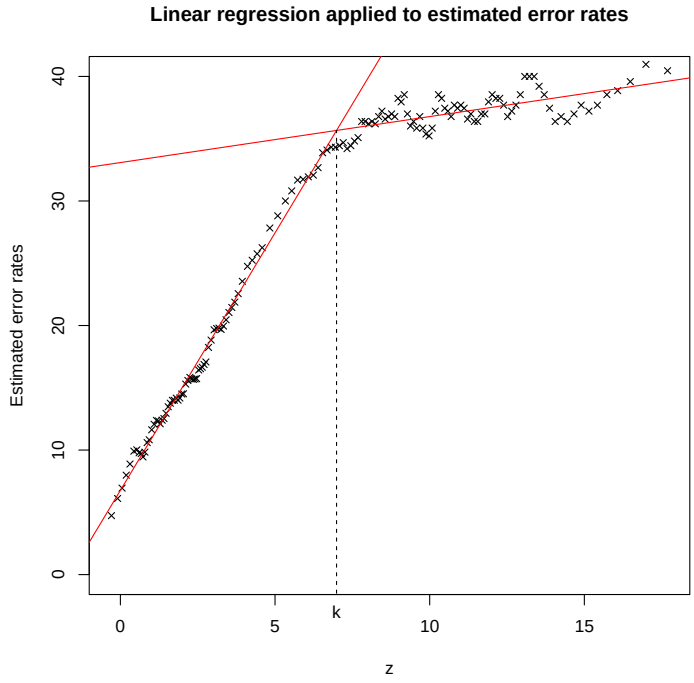


Figure 2.5: Estimate of the error rate for control reads as a function of the input of the logistic function. A linear relationship would be expected between both variables, however, a plateau after reaching error rates of 40 is often seen thus the need to model this relationship using a piecewise linear regression. The observed error rate can be computed by sorting observations according to the value of the logistic function and computing the ratio of mismatches to observations for a given window. This process can be repeated using multiple windows to obtain estimates for various values of the logistic function. The value k represents the boundary of the two subdomains of the piecewise linear functions which are represented in red.

To model this distribution, a piecewise linear regression was used:

$$h(\hat{z}_i) = \begin{cases} \theta_0 + \theta_1 \hat{z}_i & : \hat{z}_i \leq k \\ \theta_2 + \theta_3 \hat{z}_i & : \hat{z}_i > k \end{cases}$$

where both lines intersect at k and both slopes (θ_1, θ_3) are positive. The following expression:

$$\sum_i (h(\hat{z}_i) - (-10 \cdot \log_{10}(\hat{\epsilon}_i)))^2 \tag{2.3}$$

was used as the cost function. The result of this regression can be seen as red lines in Figure 2.5. One equation models the ascending quality scores due to higher SVM classification confidence while the other models the plateau which increases at a much lower rate than the former. This equation is subsequently used on reads to predict the quality score for each base. The resulting scores show a high correlation to their respective error rate. Even if a lane containing control reads is not used for this calibration, the high concordance between predicted and observed quality scores still holds.

2.4 Results

The effect on prediction accuracy of changing the SVM library for LIBOCAS is presented (see Section 2.4.1), followed by the effect of masking divergent bases on the PhiX genome (see Section 2.4.2). The accuracy of the individual bases (see Section 2.4.3) and the quality scores (see page 42) is presented. The effect of having improved basecalling accuracy on genotyping follows (see Section 2.4.5) and finally, the results obtained when using freeIbis on a problematic sequencing run is presented (see Section 2.4.6).

2.4.1 Effect of the SVM library

Out of the SVM libraries described in the methods, LIBOCAS outperforms the other libraries that were tested for both metrics: accuracy (see Figure 2.6) at a reasonable runtime (see Figure 2.7). However, it should be pointed out that the training time for LIBOCAS had to be set manually. Even at a training of only 5 seconds, LIBOCAS outperforms other libraries in terms of accuracy. As previously mentioned, for those reasons, LIBOCAS was chosen as the SVM library for the new version of the software.

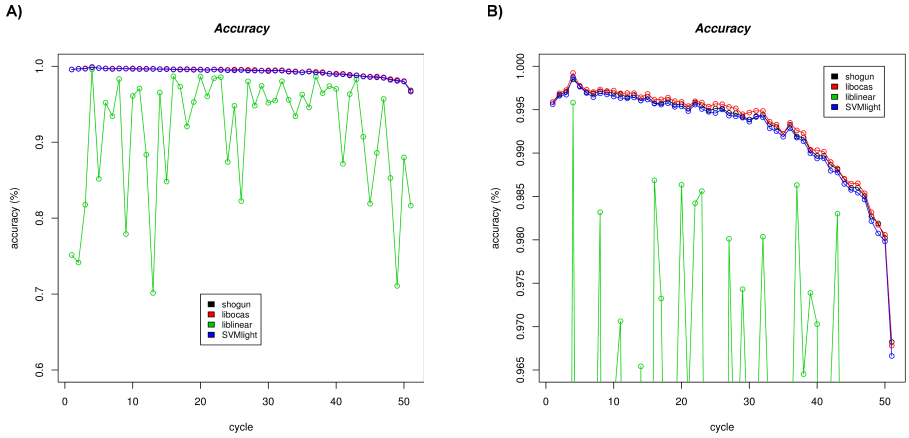


Figure 2.6: A) The prediction accuracy of 4 different SVM libraries on predicting the bases in a test set made from half of the control reads from 51 cycles from a control lane on a GAII. B) A close-up of figure A), showing that LIBOCAS generally outperforms the other libraries that were tested in terms of accuracy.

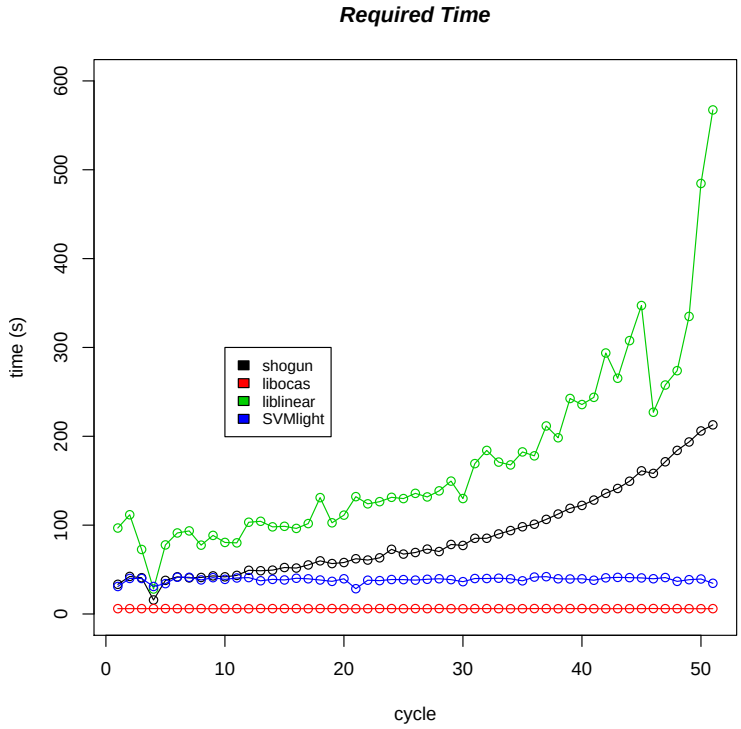


Figure 2.7: The training time required by each library. The training time for LIBOCAS was set at 5 seconds while the remaining libraries were allowed to reach their default convergence criterion.

2.4.2 Effect of masking positions on the PhiX genome

FreeIbis was compared with the masking of divergent positions on the PhiX genome disabled, thus changing only the SVM library for LIBOCAS, to the most recent version of Ibis . The aforementioned run containing 200,000 sequences from a PhiX control lane with a high thymine retention [55] was used. The reads produced by both versions were aligned back to the PhiX genome and the number of sequences mapped and average edit distance was computed.

Since the introduction of incorrectly labeled training examples could influence the quality of the SVM model, whether the masking procedure would have an effect on the number of mapped reads was evaluated. The mapping statistics confirmed that masking divergent bases on the PhiX genome improves the final sequence accuracy (170,572 sequences mapped) compared to not masking any bases (170,220) or masking random bases (170,225) (see Table 2.1).

Method	Mapped	Perfectly mapped (%)
Ibis	169,500 (84.75%)	124,260 (73.31%)
freeIbis (random masking)	170,225 (85.11%)	125,471 (73.71%)
freeIbis (no masking)	170,220 (85.11%)	125,457 (73.70%)
freeIbis (with masking)	170,572 (85.29%)	125,870 (73.79%)

Table 2.1: Accuracy for Ibis, freeIbis with masking disabled and by masking both random and divergent bases on the PhiX genome.

2.4.3 Base prediction accuracy

FreeIbis was tested on a recent paired-end GAIIx run from mid-2011 from the sequencing center at the Max Planck Institute for Evolutionary Anthropology with 2x126 cycles and a single index of 7 nucleotides. This multiplexed run had both human DNA as target, and PhiX as control and was basecalled using Ibis, freeIbis as well as naiveBayesCall version 0.3 and All your base (AYB) version 2.08. Their performance was evaluated in terms of sequence accuracy, the number of sequences mapped and edit distance to the reference, as well as runtime (see Table 2.2 and see Table A.1 in the Appendix for various other Illumina runs). FreeIbis provides more high quality base calls, leading to an increased number of reads being mapped to the reference with a lower edit distance than is the case for other basecallers.

Basecaller	Training Time	Calling Time	Mapped (%)	Edit distance
Bustard			583,348,201 (83.93%)	1.379
naiveBayesCall	591h	658h	578,957,145 (83.34%)	1.496
AYB	394h		593,183,967 (85.52%)	1.076
Ibis	19.4h	13.2h	592,929,953 (85.31%)	1.167
freeIbis	21.3h	12.2h	594,095,219 (85.48%)	1.145

Table 2.2: Accuracy for each basecaller on an Illumina GAIIx data set (2x126 cycles with 366,135,257 clusters). The human sequences were mapped to the hg19 version of the human genome. The number of mapped sequences and the average number of mismatches for those, were tallied for each method. Time trials were conducted on a machine with 74GB of RAM and using 8 of the 12 Intel Xeon cores running at 2.27GHz. The percentage mapped is relative to sequences assigned to the read group of interest.

2.4.4 Quality score accuracy

The predicted versus observed quality scores were plotted for Bustard and for freeIbis (see Figure 2.8). The sequences for a GAII run used for comparison were produced using

Bustard Off-Line Basecaller (OLB v.1.9.3). Results show that freeIbis offers improved accuracy and calibrated quality scores for various sequencing runs, including one run on a HiSeq and another on a MiSeq, (see Figures B.1 through B.8 in the Appendix) and outperforms Bustard on runs with unusually high error rates (see the following section 2.4.6).

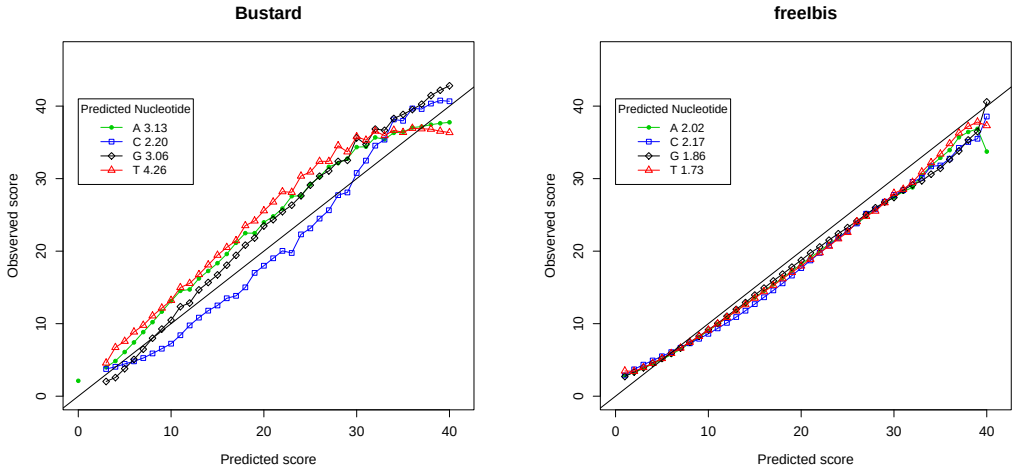


Figure 2.8: Plot of the predicted versus observed base quality score for control reads on an Illumina GAII. Ideally the base qualities should follow the diagonal line. The root mean square error (RMSE) shows that quality scores predicted using freeIbis have a greater correlation to their observed error rates.

Genotype calls obtained using Sanger sequencing was compared to the genotype calls from the same sequencing data but by using 3 different basecallers (Ibis, freeIbis and Bustard). Our results show that freeIbis offers improved genotyping accuracy (see section 2.4.5).

2.4.5 Comparing influence on genotype

To evaluate whether the new quality scores combined with the increased accuracy in basecalling would have any effect on the genotyping, the SNPs obtained from sequences basecalled using Ibis, freeIbis and the default basecaller provided by Illumina (Bustard) was compared against genotype data from Sanger sequencing. Three different Illumina GAII runs from 2011 were basecalled using the 3 aforementioned basecallers. The data was demultiplexed, stripped of sequencing adapters using an in-house sequencing pipeline [53]. The Sanger genotyped data came from different individuals. From this panel of

various individuals, 10 individuals were selected for comparison by the completeness of the genotyping obtained using the Sanger reads.

The data stemming from 49 genomic regions with a total length of 93kb (average: 1.9kb) from extant humans samples was mapped against the hg19 version of the human genome using BWA v.0.5.10 [63]. The resulting data was genotyped using GATK v.1.3-14 [74] (using option EMIT_ALL_SITES) after duplicate marking and removal using Picard v. 1.56 (http://picard.sourceforge.net) and indel realignment, again using GATK. Given a general genotype quality cutoff value, the number of true positives, where Sanger and Illumina agreed, false positives (i.e. Illumina SNP but no Sanger), false negatives (SNP detected in Sanger but no alternative allele in Illumina) and true negatives were tabulated. To avoid any potential omissions in the Sanger sequenced data, only SNPs not found in dbSNP and with no clear sign of strand bias were tabulated as a false positive.

When comparing to the previous version of the software, the resulting genotyping accuracy (Table A.2 in Appendix) presents less false positives at low quality but freeIbis produces more accurate calls and better accuracy at higher genotype quality. This is due to the distribution of the quality scores (see Figure B.6 in Appendix) between both basecallers as Ibis produces quality scores in the 20-30 range whereas freeIbis is able to confidently call bases at higher quality scores. At any genotype quality cutoff, freeIbis produces more accurate calls and fewer erroneous ones than Bustard. Furthermore, the average genotype quality for all positions for freeIbis (58.98) is higher than Ibis' (58.77) or Bustard's (58.77).

2.4.6 On problematic data

To evaluated whether freeIbis would still have the robustness to improve the accuracy of a problematic dataset, freeIbis was compared to Bustard on a run with a high error rate sequenced on an Illumina GAIIx from the sequencing facilities at the Max Planck Institute for Evolutionary Anthropology. The high error rate was due to an overloading of the flowcell thus making it arduous for the sequencer to delineate the different sequence clusters. This run was basecalled both with freeIbis and Bustard and the error rates for sequences identified as controls were compared. Across lanes, the edit distance for reads basecalled with freeIbis had a lower edit distance to their reference (see Table A.3 in the Appendix) and a greater percentage of sequences were mapped overall (see Figure 2.9).

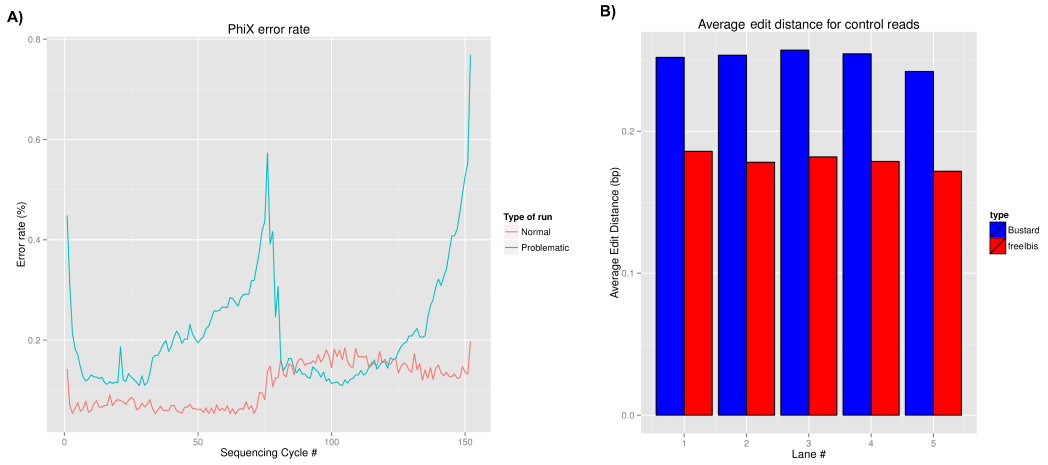


Figure 2.9: The error rate of control sequences for a problematic sequencing run (Illumina GAIIx 2x76bp) with a very high error rate A) compared to a different run (Illumina MiSeq 2x76bp) with a standard error rate. Although the error rate for control reads usually increases at the end due to increased phasing, it reaches for this particular run one error in 200 bases. B) The edit distance of these control reads to their reference genome reveals that despite the increased error rate, freeIbis performs better than Bustard in terms of edit distance. For comparison purposes, the edit distance for the aforementioned MiSeq run with a standard error rate was 0.101632 thus revealing the problematic nature of this dataset.

2.5 Conclusion

FreeIbis provides substantial improvements in sequence accuracy, quality score calibration and genotyping accuracy over Bustard, and is more computationally efficient than equally accurate model-based methods such as AYB.

Chapter 3

Bayesian ancient DNA fragment reconstruction

This chapter introduces, leeHom, a Bayesian algorithm to infer aDNA fragments from sequencing data.

3.1 Background

As alluded to in section 1.1.3, DNA molecules extracted from ancient samples are often short due to the degradation of DNA after the death of the organism and average length rarely exceeds 100 bp [102].

As a consequence the length of reads obtained from the Illumina sequencer often exceeds the length of the DNA molecule, and the read therefore contains both the sequence of the original molecule and also part of the adapter sequence (see Figure 3.1). For paired-end reads that exceed the molecule length, both the forward and reverse reads will have the sequence of the same original molecule before showing residual adapter sequence. Similarly, molecules that are shorter than the sum of the forward and reverse read length are expected to show identical bases at the ends of both reads since the same part of the molecule is read twice. Merging of identical sequences is also expected to reduce sequencing error due to the repeated observation of the same base. Since residual adapter sequences in the reads interfere with mapping and assembly, it is necessary to trim reads up to the start of the original molecule.

Several algorithms have been implemented to trim adapter sequences (see [70], [57] and http://code.google.com/p/ea-utils/wiki/FastqMultx) and to merge overlapping paired-end sequences (see [53],[65] and https://github.com/jstjohn/SeqPrep). However, these algorithms use cutoffs for detecting adapters and merging reads and need to be adapted to

varying rates of sequencing errors. For instance, both cutadapt and AdaptorRemoval have default thresholds for the rate of mismatches and minimum overlap length. More liberal cutoffs can lead to a greater number of false positives. Other algorithms [71, 118, 66] have been designed to merge overlapping pairs but do not provide the likelihood of seeing the adapter at the end of both reads. Furthermore, sequencing centers often give end-users sequencing data with the adapters already trimmed. Including the quality of similarity to the sequencing adapters in the computation to reconstruct very short molecules therefore becomes impossible.

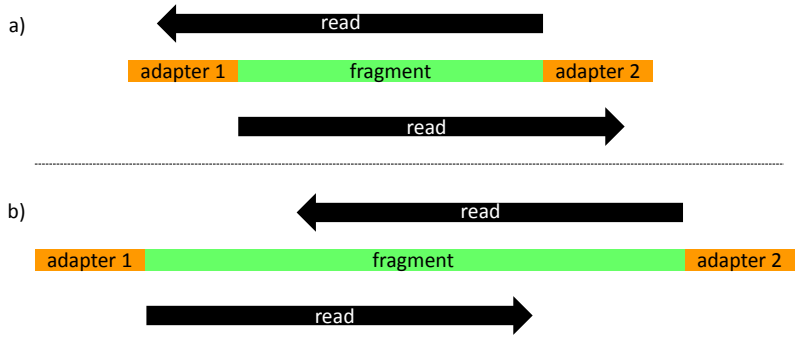


Figure 3.1: Schematic representation of paired-end sequencing for very short molecules. a) When the molecule is shorter than the read length, both reads will run into the adapters and the remaining part will completely overlap. b) If the sequence is longer but still not longer than twice the read length, adapter sequences will be absent but a partial overlap can be observed between the end of the sequences.

The field of bioinformatics applied to NGS and aDNA required an algorithm that:

- Uses quality scores to distinguish genuine mismatches from common sequencing errors. Furthermore, quality scores should be used for the consensus call of overlapping portions of paired-end data
- Does not have cutoffs for mismatches to the adapter or percentage overlap as those cannot be tailored for the entire flowcell which has divergent rates of errors (see section 1.2.3)
- Does not separate the process of adapter trimming and sequence overlap as those are linked tasks

3.2 Introduction

This chapter presents a new Bayesian maximum *a posteriori* based trimming and merging algorithm, leeHom, that is particularly useful for aDNA and other cases where short

molecules are sequenced. Instead of separating the processes of adapter trimming and merging, leeHom combines both steps into a single probabilistic model. Briefly, leeHom computes the probability of observing the reads given a certain original molecule length and returns the most likely sequence. This algorithm is highly robust to sequencing error, produces few false positives and is able to handle common sequencing problems such as missing cycles. The algorithm was tested on a set of simulated aDNA sequences where the original molecule sequence was known, and on Neanderthal sequence data. Results show that leeHom outperforms currently available software in speed and accuracy for both simulated and real ancient DNA data, and that it is suitable for processing large volumes of sequence data. It can take unaligned BAM or fastq files as input and requires the sequence of the adapters be provided. leeHom is released under the GPLv3 and is freely available from: https://bioinf.eva.mpg.de/leehom/

3.3 Methods

First, the algorithm for computing the likelihood for various fragment lengths is presented (see Section 3.3.1), following by the computation of the posterior probability for overlapping portions of the sequences (see Section 3.3.2) and finally, strategies to test the algorithm are presented (see Section 3.3.3).

3.3.1 Computation of the likelihood for a given fragment length

The algorithm described here relies on computing the probability of observing both pairs of reads assuming that the original sequence is of a certain length. A similar maximum-likelihood approach for paired-end reads was used in the literature for assembling 16S rRNA or PCR product flanked by primer sequences using partially overlapping paired-end reads (see [71]). Apart from computing the likelihood of all possible overlapping sequence lengths, the likelihood of stemming from non-overlapping pairs is also computed, thus removing the need for hard cutoffs. Furthermore, a probabilistic prior of seeing a sequence of a certain length can be added.

Given that paired-end reads r_1 and r_2 have been sequenced, it is assumed that if the original sequence was shorter than the read length, each read will have, at the end, the sequences of the adapters: a_1 and a_2 respectively. Let $l_1 = length(r_1)$ and $l_2 = length(r_2)$. The probability of observing this data given that it is assumed that the original sequence was of length i, denoted as $P(r_1, r_2, a_1, a_2|i)$, can be computed using the following formula:

$$P(a_1 \approx r_1[i-1..]) \cdot P(r_1[1..i-1] \approx \overline{r_2}[1..i-1]) \cdot P(a_2 \approx \overline{r_2}[i-1..]) \tag{3.1}$$

where P represents the probability, $\overline{r_2}$ is the reverse complement of r_2, the $[i..]$ and $[1..i-1]$ operators denote the suffix starting at position i and the prefix ending before position i, respectively and where an end index greater than the start one represents an empty string. The first and last terms correspond to the probability of observing $r_1[i..]$ and $\overline{r_2}[i..]$ given that the templates were a_1 and a_2 respectively. The middle term corresponds to the probability of observing the stretches $r_1[1..i-1]$ and $\overline{r_2}[1..i-1]$ given that they stemmed from a common sequence. The specific equations for those two probability functions are defined in greater detail below. This probability is computed for every $i \in 0...l_1 + l_2$. The posterior probability of any length being i given the data can be described using the following expression:

$$P(i|r_1, r_2, a_1, a_2) \propto P(r_1, r_2, a_1, a_2|i) \cdot P(i) \tag{3.2}$$

The prior on the sequence length i is defined using the probability density function of the log-normal distribution given by:

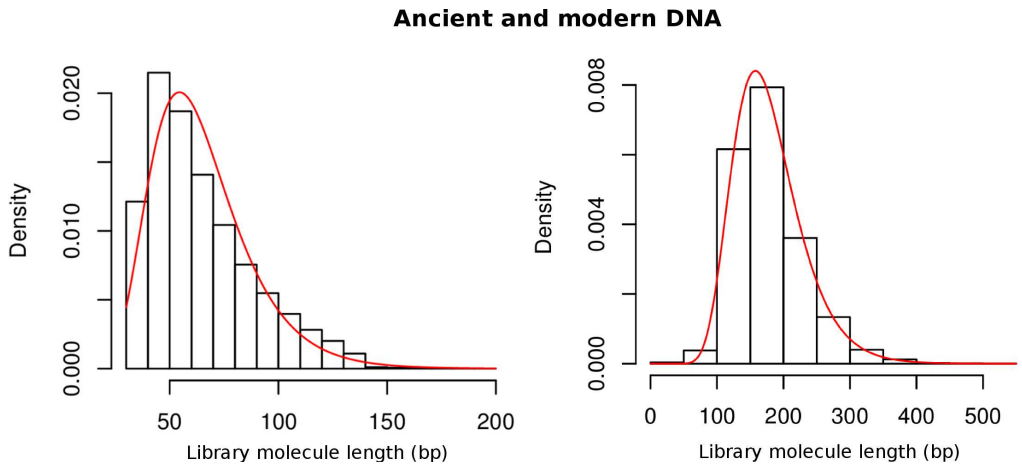


Figure 3.2: Empirical (black) and theoretical (red) fragment length distributions of ancient and modern DNA libraries. Presented is the output of the maximum-likelihood fit from the Fitdistrplus R package using a log-normal distribution for an ancient DNA library (left) and a modern DNA one (right). Ancient DNA tends to be of shorter length and of much narrower variance than modern DNA.

$$P(i) = \frac{1}{i\sqrt{2\pi}\sigma} \exp^{-\frac{(\log(i)-\mu)^2}{2\sigma^2}} \tag{3.3}$$

The term above models the likelihood of seeing that particular sequence size given a prior belief on the sequence size distribution. To find the most suitable distribution to model the length of DNA sequences, various heavy-tail distribution were compared using the maximum-likelihood fit from the Fitdistrplus R package (http://cran.r-project.org/web/packages/fitdistrplus/) and the one maximizing the likelihood of the fit was log-normal (data not shown). To illustrate how the shape of the prior changes from modern to ancient DNA sequences, the log-normal distribution for both a modern and ancient DNA dataset was computed (see Figure 3.2). Users also have the option of using a uniform prior on the sequence length if the size distribution of the sequences is unknown.

leeHom aims at finding the original sequence length i_{max} that maximizes the posterior of observation of r_1 and r_2:

$$i_{max} = argmax_{i \in \{0 \ldots l_1+l_2\}} P(i|r_1, r_2, a_1, a_2)$$

and returns the most likely bases for the sequence of length i_{max}.

To compute $P(a_1 \approx r_1[i..])$ and $P(a_2 \approx r_2[i..])$, a string comparison is used that disallows insertions/deletions while tolerating mismatches. The probability of seeing a substring

of a read $r[i..]$ given that an adapter a was the template is given by the product of the likelihood for each base :

$$P(a \approx r[i-1..]) = \prod_{k=i-1}^{k=length(r)} P_{match}(a[k-i+1], r[k]) \tag{3.4}$$

where P_{match} is the likelihood of a match for two bases. Let $q[i]$ be the quality score associated with base $r[i]$, the probability of sequencing error (see Section 1.1.2 for further explanation about quality scores on the PHRED scale) for a given quality score $q[k]$ is defined as follows:

$$P_e(q[k]) = 10^{\frac{-q[k]}{10}} \tag{3.5}$$

Therefore, the probability of observing $r[k+i]$ given that the correct nucleotide is $a[k]$ is computed as follows:

$$P_{match}(a[k], r[k+i]) = \begin{cases} 1 - P_e(q[k]) & \text{if } a[k] = r[k+i] \\ P_e(q_k) \cdot \frac{1}{3} & \text{if } a[k] \neq r[k+i] \\ \frac{1}{4} & \text{if } k > length(a) \end{cases} \tag{3.6}$$

Equation 3.6 assumes that the probability of error given a certain sequenced base represents the probability of miscalling the base to any other base with equal probability (see Section 1.2.3).

The likelihood of the overlap $P(r_1[1..i-1] \approx \overline{r_2}[1..i-1])$, is defined as the probability of having seen both substrings given that they stemmed from the same DNA sequence. Assuming that each base is independent of the remaining ones, the likelihood for each base can therefore be multiplied as such:

$$P(r_1[1..i-1] \approx \overline{r_2}[1..i-1]) = \prod_{k=1}^{k=i-1} P_{match}(r_1[k], \overline{r_2}[k]) \tag{3.7}$$

Given that the strings r_1 and $\overline{r_2}$ have the associated quality scores q_1 and $\overline{q_2}$, the likelihood of two bases from two different reads stemming from the same original base is given by marginalizing the probabilities for each potential nucleotide that could have been this original nucleotide multiplied by the respective probability of observation of the two sequenced nucleotides :

$$P_{match}(r_1[k], \overline{r_2}[k]) = \sum_{n \in \{A,C,G,T\}} P_{obs}(n) \cdot P_{obs}(r_1, \overline{r_2}|n) \tag{3.8}$$

where $P_{obs}(n)$ representing the likelihood of observing nucleotide n in the original overlapping sequence, approximated to $\frac{1}{4}$ for $\forall n \in \{A, C, G, T\}$. The second term ($P_{obs}(r_1, \overline{r_2}|n)$) can be quantified as follows:

$$\begin{cases} (1 - P_e(q_1[k])) & \cdot & (1 - P_e(q_2[k])) & \text{if } r_1[k] = \overline{r_2}[k] \wedge r_1[k] = n \\ (1 - P_e(q_1[k])) & \cdot & (P_e(q_2[k])) \cdot \frac{1}{3} & \text{if } r_1[k] \neq \overline{r_2}[k] \wedge r_1[k] = n \\ (P_e(q_1[k])) \cdot \frac{1}{3} & \cdot & (1 - P_e(q_2[k])) & \text{if } r_1[k] \neq \overline{r_2}[k] \wedge \overline{r_2}[k] = n \\ (P_e(q_1[k])) \cdot \frac{1}{3} & \cdot & (P_e(q_2[k])) \cdot \frac{1}{3} & \text{if } r_1[k] \neq n \wedge \overline{r_2}[k] \neq n \end{cases} \tag{3.9}$$

Again, it is assumed that a sequencing error is equally likely to produce any nucleotide besides the correct one.

Once again, leeHom aims at finding the sequence length i that maximizes equation 3.1. However, different values of i can be equally likely to have occurred. To avoid incorrect reconstructions due to multiple sequence length that are equally likely, the program avoids reconstructing sequences where the ratio of the likelihoods of the second most likely sequence length to the most likely one exceeds 1 in 20. This ensures that the most likely sequence has to be several folds more likely than the second-best option. As mentioned before, the likelihood of having no overlap and therefore having a sequence length exceeding twice the read length is also computed as follows:

$$\int_{l_1+l_2}^{\infty} \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}} \cdot \prod_{l_1+l_2} P_{obs}(n) \tag{3.10}$$

where $P_{obs}(n)$ is defined as in equation 3.8. The prior ($\int_{l_1+l_2}^{\infty} \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}}$) on the sequence length represents the probability of generating a sequence longer than l_1+l_2 and can be interpreted as $1 - cdf(l_1 + l_2)$ where $cdf()$ is the cumulative distribution function for the aforementioned log-normal distribution. The resulting value is compared with the likelihood values, defined by equation 3.1, for all sequence lengths : $i \in \{0...l_1 + l_2\}$.

3.3.2 Consensus of overlapping regions

Once the most likely sequence length has been computed, the remaining task is to assemble the sequence using the information provided by r_1 and r_2. If a base has been covered in only one read, it is reported along with the original quality score. However, if the base is covered by both reads, a consensus base with its associated quality score is produced. Again, a principle of independent observations with quantified error probabilities given by the quality scores to produce both quantities is assumed.

Let two sequenced bases b_1 and b_2 with quality scores on the PHRED scale q_1 and q_2 respectively. For any given nucleotide $n \in \{A, C, G, T\}$ that is believed to be the actual base, the probability of observing b_1 can be computed by the following:

$$p(b_1|n) = \begin{cases} 1 - P_e(q_1) & \text{if } b_1 = n \\ \frac{P_e(q_1)}{3} & \text{if } b_1 \neq n \end{cases} \tag{3.11}$$

Assuming that both bases b_1 and b_2 represent independent observations, the probability of n given b_1 and b_2 can be defined as :

$$P(b_1, b_2|n) = P(b_1|n) \cdot P(b_2|n) \tag{3.12}$$

For calling the consensus base, the likelihood of a nucleotide n given the observation b_1 and b_2 needs to be computed. This can be computed using Bayes' rule:

$$P(n|b_1, b_2) = \frac{P(b_1, b_2|n) \cdot P(n)}{P(b_1, b_2)} \tag{3.13}$$

The probability of having observed b_1 and b_2 can be computed by summing the probability of having generated both bases, given that it is assumed that they came from the same base. Since there are only four possibilities for this base, the following equation can be used:

$$P(b_1, b_2) = \sum_{m \in \{A,C,G,T\}} P_{obs}(m) \cdot P(b_1, b_2|m) \tag{3.14}$$

Where $P_{obs}(m)$ is the prior for that given nucleotide (see Section above) and $P(b_1, b_2|m)$ can be derived using equation 3.12 and 3.11. In resulting BAM files, the probability of error, which is the probability of not observing n given the two bases b_1 and b_2, is reported. Hence the following can be derived:

$$\begin{aligned} P(\neg n|b_1, b_2) &= 1 - P(n|b_1, b_2) && (3.15) \\ &= 1 - \frac{P(b_1, b_2|n) \cdot P(n)}{P(b_1, b_2)} && (3.16) \\ &= \frac{P(b_1, b_2) - P(b_1, b_2|n) \cdot P(n)}{P(b_1, b_2)} && (3.17) \\ & && (3.18) \end{aligned}$$

By substituting the result from equation 3.14 in the previous expression, $P(\neg n|b_1, b_2)$ becomes:

$$P(\neg n|b_1, b_2) = \frac{\sum_{m \in \{A,C,G,T\} \setminus n} P(b_1, b_2|m)}{\sum_{m \in \{A,C,G,T\}} P(b_1, b_2|m)}. \tag{3.19}$$

Finally, the most likely nucleotide is produced along with its associated quality score by taking the PHRED scaled quantity defined in equation 3.19.

3.3.3 aDNA sequencing data

Since sequencing error rates vary between sequencing runs and even vary within a run, such a complex error rate is difficult to model and an actual dataset would be needed to evaluate reconstruction accuracy. To benchmark the aforementioned programs on actual aDNA data, the first 10M reads from a paired-end Illumina HiSeq 2500 run from the Altai Neanderthal [89] were used as test dataset. Programs that trim and merge reads - MergeTrimReads, SeqPrep and AdaptorRemoval, - were used with default parameters for comparison. The resulting reconstructed sequences were mapped back to the human reference genome (1000 Genomes version hg19) using BWA 0.5.10 [63] with default parameters. The number of aligned sequences along with the number of sequences aligning with mapping quality greater than 30 were tallied for each algorithm.

A common feature of the programs being tested is the ability to merge overlapping stretches. To evaluate whether this strategy improved sequence accuracy compared to simply trimming the adapters and mapping both remaining reads separately, the same dataset was processed by cutadapt [70] and the resulting unmerged and paired reads were mapped with BWA. The number of mismatches per aligned basepair was computed for aligned reads which were merged by leeHom and simply left as trimmed paired reads by cutadapt.

3.4 Results

The distribution of the log-likelihood for modern versus aDNA sequencing data is presented (see Section 3.4.1) followed by leeHom's accuracy on simulated paired-end data (see Section 3.4.2, page 57) and on simulated single-end data (see Section 3.4.2, page 60). Finally, results on empirical data are presented (see Section 3.4.3).

3.4.1 Distribution of the log likelihood

To illustrate the differences in the likelihood landscape between actual modern and ancient DNA paired reads, the log-likelihood for different potential sequence lengths was plotted (see Figure 3.3). For aDNA pairs, there is a clear peak in log-likelihood around the length of the original fragment whereas modern DNA shows a more even probabilistic landscape. For the aDNA, the difference between the log of the most likely and the second most likely fragment length was 66.93 whereas that difference was 0.19 for the modern DNA pairs.

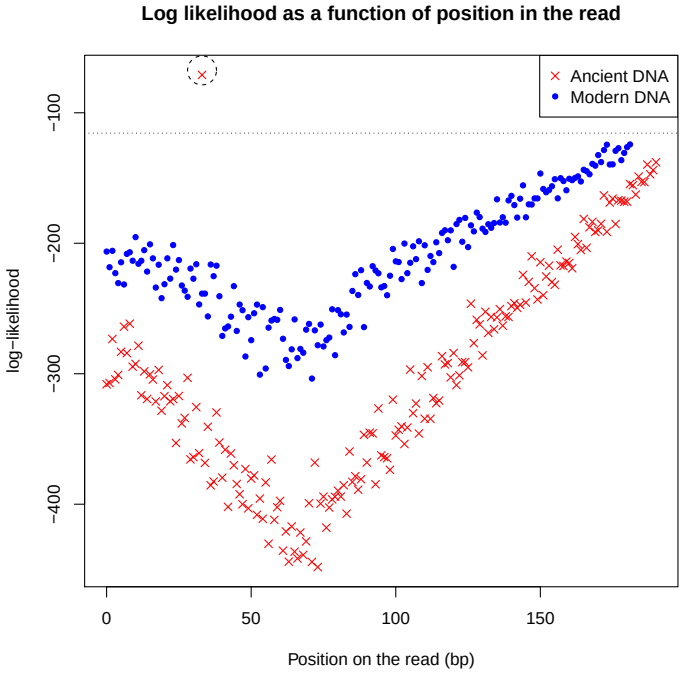


Figure 3.3: The log-likelihood for various possibilities of length of original sequence for an ancient and modern DNA read pair. The dotted line represents the likelihood that the reads do not merge and came from a sequence length greater than the longest possible overlap. For the aDNA read pairs, a certain length of sequence is more likely than the remaining possibilities. This does not occur for modern DNA read pairs due to the longer size of the original sequence.

3.4.2 Simulated data

Paired-end

Using simulated paired-end reads at different levels of error, the performance of leeHom was compared to MergeTrimReads (from [53]), SeqPrep (https://github.com/jstjohn/SeqPrep) and to AdaptorRemoval [65]. Briefly, sequences matching the read distribution of aDNA molecules generated for the Denisova genome project [77] were selected at random from the genome. Sequences with unresolved base pairs ("N") were removed. Reads of 100bp were simulated by either adding adapter sequences to the end of reads if the original sequence was shorter than the simulated read length or by simply taking the first hundred base pairs from each end. An Illumina error profile was used by aligning PhiX control sequences to the PhiX genome and building a frequency table for matches and types of

substitution. The frequency of quality scores associated with each was tallied. Errors were introduced at a certain rate and nucleotide substitutions were added with the associated quality score taken from the error profile. Errors were introduced for each base independently of each other. As the dataset contains the original sequence, both the number of molecules for which the sequence was reconstructed perfectly and the number of sequences with the correct length was assessed.

The number of perfectly reconstructed sequences versus the simulated error rate is plotted in figure 3.4. Clearly, the number of inferred sequences without any mismatches decreases both due to the increased difficulty of inferring the original sequence and the smaller number of sequences with no mismatches. The relative number of sequences with at least one mismatch was also plotted. This number tends to reach a plateau due to the absence of reads without any sequencing errors. Both in terms of perfectly reconstructed sequences and inexact matches, leeHom outperforms remaining algorithms especially at high error rates. Furthermore, in terms of reconstructed sequences with the correct length irrespective of the number of mismatches, leeHom also offers superior accuracy.

In terms of falsely merged reads, out of 931,767 paired-end reads, neither AdaptorRemoval nor SeqPrep generated any false positives. MergeTrimReads and leeHom generated respectively 11 and 22 false positives. Those reads were located in regions of genomic repeats. It should also be noted that using a prior in leeHom on the sequence length equal to the distribution used to generate the simulated reads eliminates these false positives.

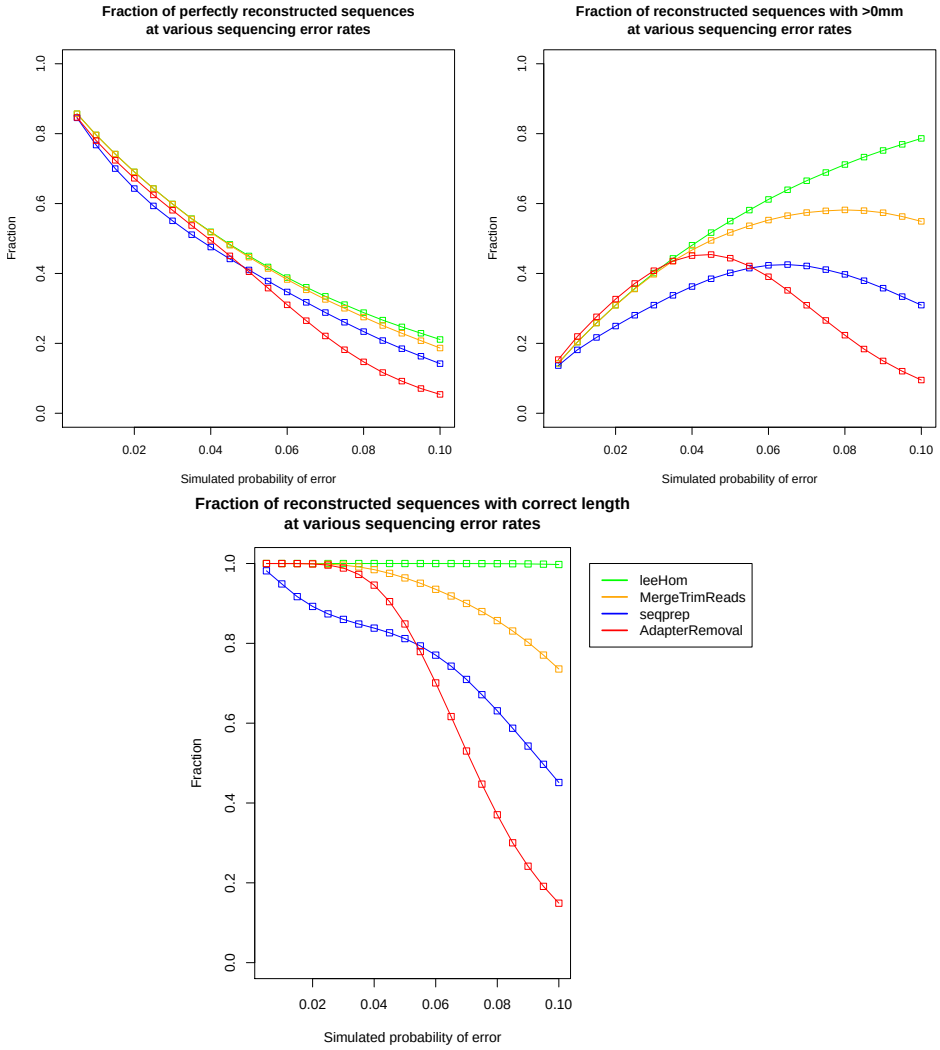


Figure 3.4: Comparison of the fraction for all input reads of reconstructed sequences as a function of simulated error rate for the output of leeHom and currently available software for sequence reconstruction based on paired-end reads. The number of perfectly reconstructed sequences (top left), the ones with a single mismatch (mm) to the original sequence (top right) and those with the correct length (bottom) are presented. Both in terms of perfectly reconstructed sequences and in terms of sequences with the correct length, leeHom outperforms other currently available algorithms.

Single-end

In a similar approach to the one taken for paired-end reads, the number of perfectly inferred sequences was tallied for single-end reads for various software packages that trim adapter sequences. The set of forward reads for the simulated aDNA sequences were used as a test set of single-end reads. Also, the number of sequences with imperfect matches to the original simulated sequence as well as the total number of sequences with correct length was computed.

The 4 aforementioned programs (leeHom, MergeTrimReads, AdaptorRemoval and cutadapt) were used to reconstruct the original sequences (see Figure 3.5). leeHom and AdaptorRemoval offer the greatest robustness to sequencing errors.

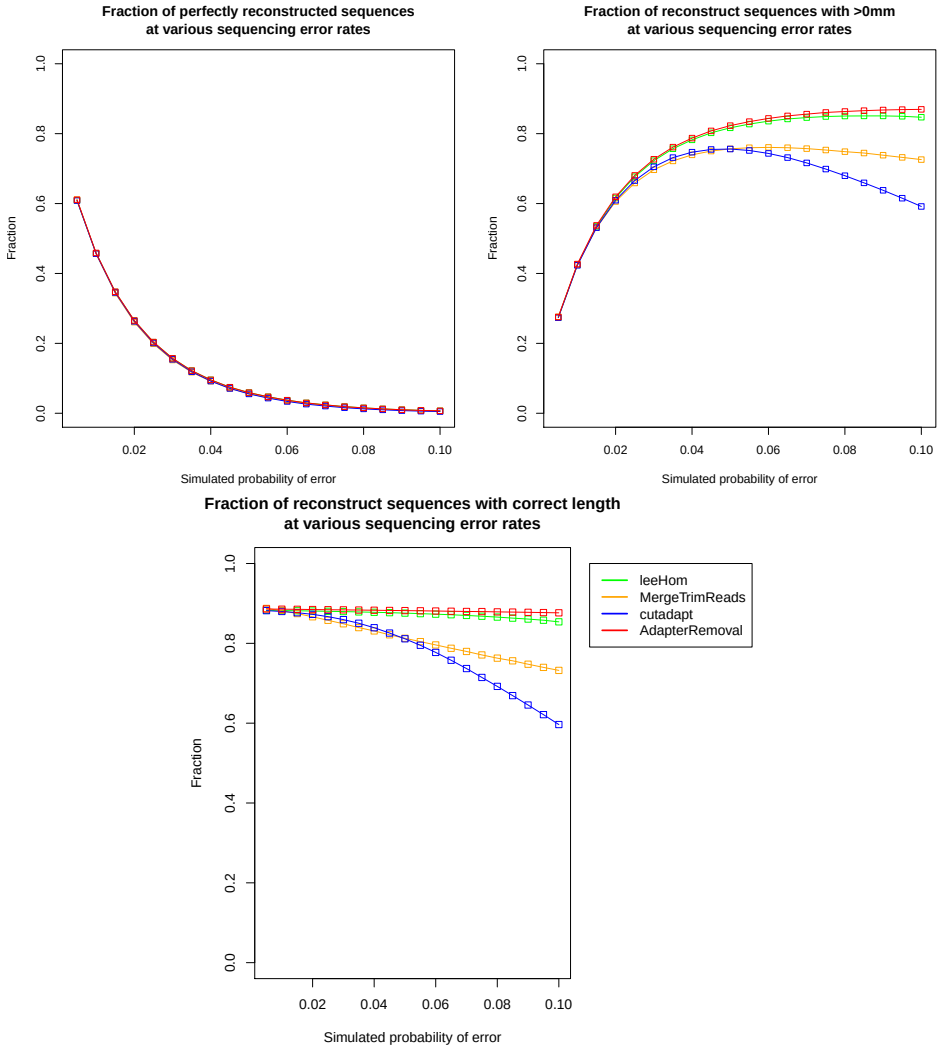


Figure 3.5: Accuracy of various programs for adapter removal on a simulated set of single-end aDNA reads. The number of sequences with no mismatches (top left), those with a single mismatch (mm) to the original sequence (top right) and those with the correct length (bottom) are presented. leeHom and AdaptorRemoval offer the most liberal trimming while cutadapt is the most conservative.

For a set of 931,767 reads taken at random from the human genome, the number of sequences that were trimmed was computed. The result reported in Table 3.1 show that leeHom without a prior on the sequence length generates few false positives but using a prior on the sequence length generates no false positives due to the low likelihood of observing such short sequences.

Although leeHom and AdaptorRemoval had the greatest robustness to sequencing errors, upon measuring the number of false positives on simulated modern DNA reads, leehom offered fewer false positives than AdaptorRemoval.

leeHom (prior)	leeHom	MergeTrimReads	AdaptorRemoval	cutadapt
0	102,964	359,103	164,744	25,553

Table 3.1: Number of false positives on a simulated set of 931,767 single-end modern DNA reads. A false positive is defined as any trimmed read as the simulated insert size (1000bp) should not yield any overlap with the adapters. As mentioned in figure 3.5, cutadapt is the most conservative while other tools tend to trim more liberally. However, cutadapt also has lower sensitivity at higher error rates for aDNA while leeHom offers higher accuracy for aDNA reads while yielding fewer false positives than MergeTrimReads and AdaptorRemoval.

3.4.3 Empirical sequencing data

On an empirical aDNA dataset of 10M paired-end reads from [89], the runtime as well as the number of inferred sequences mapping back to the genome was computed (see Table 3.2). Also, the number of sequences aligning with mapping quality of at least 30 was computed. Since an algorithm is unlikely to produce a sequence that aligns to the human genome by chance and even more unlikely to align with high mapping quality, the number of aligned sequences indicates the accuracy of the reconstruction. Both in terms of runtime and accuracy, leeHom outperforms currently available programs.

Trivially, it should be noted that the use of any of these tools is an improvement over aligning the raw sequences without any attempt at paired-end read merging since, out of 10M paired-end reads, only 1,506,567 (15.07%) reads align and among those, 1,338,397 (13.38%) have high mapping quality.

An assumption behind paired-end read merging for aDNA is the ability to cross-correct using double observations of the sequenced bases and quality scores. To test this hypothesis, the number of mismatches per aligned base was computed for both merged sequences produced by leeHom and trimmed reads produced by cutadapt (see Table 3.3). The number of mismatches per aligned nucleotide is lower in the merged reads produced by leeHom thus indicating the gain in accuracy is due to cross-correction.

	leeHom) (+prior)	leeHom (no prior)	MergeTrimReads	AdaptorRemoval	SeqPrep	cutadapt + FLASH
runtime (wallclock)	17m16s	16m51s	60m17s	27m37s	23m20s	4m32+5m37
runtime (CPU)	17m14s	16m49s	60m16s	28m16s	24m27s	6m00+5m15
Mapped	3,381,755	3,373,531	3,370,675	3,308,763	3,222,585	3,276,250
MQ30	2,814,558	2,806,692	2,803,915	2,758,884	2,743,703	2,744,661

Table 3.2: Runtime and accuracy for various adapter trimming and merging software packages. The runtime of different algorithms for sequence reconstruction was evaluated along with the number of produced sequences aligning to the human genome. In terms of aligned sequences both at minimum mapping quality 0 and 30, leeHom outperforms other algorithms especially if a prior on the sequence length is used. Also in terms of runtime, leeHom compares favorably to other programs. PEAR failed to run due to the amount of data even when increasing the amount of RAM (to 5GB). The time reported for FLASH is the time for cutadapt to run for both forward and reverse reads and for FLASH to run.

Approach	mismatches	aligned bases	mismatches per 1000 bases
leeHom	1,130,159	218,746,206	5.17
cutadapt	2,326,041	410,027,512	5.67

Table 3.3: Mismatches per aligned base for various aDNA reconstruction strategies. The number of mismatches and aligned bases for both merged sequences produced by leeHom and reads trimmed by cutadapt. This table presents the raw number of aligned nucleotides given that, for a given paired-end read, the merged sequenced produced by leeHom and the trimmed sequence produced by cutadapt were aligned to the human genome. By computing the number of mismatches per nucleotide, leeHom produces sequences that have greater affinity to the human reference due to cross-correction in the reconstructed sequence using the paired reads.

3.5 Conclusion

The tasks of stripping residual adapters and merging overlapping pairs are generally separated. Results show that considering both at once in a single model increases the number of sequences that can ultimately be mapped. Furthermore, the use of the prior distribution can help the program to break the probabilistic tie between corner cases. For instance, when a few bases of the adapters are seen, the decision to trim or not depends heavily on the prior probability for the length distribution. For very short aDNA molecules, trimming such bases might be beneficial whereas for longer molecules, resolved bases might be needlessly removed. A Bayesian approach given the distribution offers the possibility of a natural probabilistic transition from very short fragment size to longer ones without the use of arbitrary cutoffs. A prior on the distribution should therefore be used whenever there is data from the same library that provides information about the size distribution of the library inserts. If no previous data on insert-size is available, the default parameters should be used.

For other programs that use cutoffs to trim sequencing adapters and merge overlapping portions of the reads, stricter cutoffs can be used for high quality datasets as this will reduce the number of false positives. More liberal cutoffs should be used on datasets with higher errors as this will allow more sequences to be retrieved. However, as mentioned before, error rates vary between sequencing runs and often within a single sequencing run. Adapting the thresholds for the detection of the adapters and the overlap within a single sequencing run is generally infeasible. Probabilistic approaches obviate this need by returning the most likely model given the data at hand. As shown in the results section, this approach outperforms currently available algorithms especially at high error rates.

Since adapters are often simply trimmed and the reads left unmodified during standard processing, the value of merging overlapping parts for aDNA studies was evaluated. As

shown in the results section, the cross-correcting effect of having observed the same sequence twice reduces noise and mismatches to the reference.

leeHom can be used with a prior on the distribution of the molecule lengths. However, this information is not always available beforehand especially for newly sequenced libraries. Ideally, the step of trimming adapters and merging overlapping parts should be combined with mapping where the distribution of the original sequences could be empirically determined. Once sufficient confidence in the shape of the fragment size distribution is obtained, this could be used as prior for both aDNA and modern samples as a standalone tool. Furthermore, substitution rates to remaining nucleotides have been assumed to be equally likely which is empirically not the case [81]. More realistic substitution probabilities could be incorporated in the model. Also, leeHom assumes that quality scores are correlated positively with their observed error rate (see Chapter 2).

In summary, leeHom outperforms currently available algorithms for reconstruction of aDNA sequences from reads both in terms of accuracy and speed. The Bayesian MAP sequence reconstruction lowers error in aDNA, and other datasets with overlapping paired-end reads, thus leading to more accurate alignments.

Chapter 4

Maximum-likelihood demultiplexing

This chapter presents deML, a maximum-likelihood demultiplexing algorithm.

4.1 Background

As of this writing, an Illumina HiSeq 2500 can produce approximately 3 billion clusters, yielding a total of about 600G basepairs [1]. While this high throughput is beneficial for many applications, such as high-coverage whole genome sequencing, it may be economically disadvantageous for the sequencing of small numbers of loci. For example, sequencing a single mitochondrial genome, or a single amplification product will provide unnecessarily high coverage [68, 78]. However, it is possible to sequence a large number of samples in a single run by incorporating unique sequence indices for each sample [16]. In this strategy, referred to as multiplexing (see Section 1.1.2), each sample is assigned a unique short sequence (typically 7 bp in length) which is ligated onto the sequencing primer and sequenced along with the target DNA. Pooling multiple samples increases efficiency and lowers the cost per base. Using the standard Illumina protocol up to 96 samples can be multiplexed using one index per sample. Recently, it has been proposed that incorporating two indices into each read can provide a significant increase in the maximum number of pooled samples in a single run, and lead to more accurate assignment of reads to the sample of origin [54]. Such index sequences are usually designed to maximize the nucleotide differences between all pairs of indices in order to achieve accurate demultiplexing. Previous work has focused on the design of the indices *per se* as the main mean to optimize the accuracy of sample assignment [10, 11, 34].

[1] http://www.illumina.com/systems/hiseq_2500_1500/performance_specifications.ilmn accessed: 06/24/15

Once sequencing is complete, reads must be assigned *in silico* to the sample of origin, a process referred to as demultiplexing. The default demultiplexer provided by Illumina in the CASAVA package allows for 0 or 1 mismatches to the user-supplied reference indices. Different heuristics have been proposed to assign reads to their sample of origin [15, 21, 23, 93]. Once of such heuristics is deindexer (https://github.com/ws6/deindexer) which allows users to specify the number of mismatches they are willing to accept between the sequenced indices and the original reference indices.

Although these methods perform well for sequencing reads with high quality, poor demultiplexing remains a common reason for apparent low retrieval of sequences from a multiplexed run. This commonly occurs when increased error rates - especially during sequencing of the index - can lead to a higher number of mismatches thus hindering assignment to the appropriate sample. It may also occur when cycles in the index fail completely leading to unresolved bases in the index read. In the case of double indices, it is not uncommon for an entire index to fail thus leaving a single index to be used for demultiplexing. Indices are generally designed to maximize the pairwise distances to avoid misassignments. However, poorly designed indices can sometimes be used thus leading to a high probability of misassignment (i.e. a read is assigned to sampleX while it truly came from sampleY).

Due to such issues, the subfield of next-generation sequencing needed an algorithm that is:

- Able to quantify the confidence in the sample assignment
- Robust to poor resolution of the indices due to mismatches and missing cycles
- Able to handle a suboptimal set of index sequences
- Able to identify the reads for which the confidence in the assignment is high

4.2 Introduction

This chapter describes a novel maximum-likelihood approach called deML for demultiplexing samples based on the likelihood of assignment to a particular sample. A C++ implementation is available for free, licensed under the GPL and can be downloaded from http://bioinf.eva.mpg.de/deml/. deML can run on aligned or unaligned BAM files or FASTQ files [64, 13].

Briefly, deML computes the likelihood of a read to originate from each possible sample, assigns reads to the most likely sample of origin and computes the overall confidence in this assignment.

The confidence score given to a sample assignment reflects the probability of misassignment. This allows users to quantify the uncertainty in sample assignment given the observed indices. This approach can be used for demultiplexing sequences from any multiplexed sequence run, but is particularly useful for runs where the quality of the index sequence is low. The algorithm is also valuable when the initial index list was poorly designed as the assigned uncertainty also reflects the discriminative power of the selected index set. Poorly designed index sequence sets will therefore result in a lower sample assignment confidence.

The software was tested on a double-indexed Illumina MiSeq dataset of approximately 15M reads containing an estimated $\frac{2}{3}$ human PCR product and the remaining third, PhiX control sequence. Since there was a single PhiX sample, alignments to the PhiX and human genome reference can be used to determine the sample of origin and misassignment rates can be computed. There is a high correlation between the assignment quality scores and the observed misclassification rates. As error rates cannot be purposefully increased on Illumina sequencers, to evaluate how deML fares against heuristics that use fixed mismatches such as CASAVA on problematic datasets, simulations were used. Reads with perfect matches to a particular sample were taken and sequencing errors with reflective quality scores were added to the indices at various rates. The percentage of reads assigned back to the original sample was measured for both deML and a fixed mismatch approach. The methodology is highly robust to increased error rates compared to the default Illumina approach and allows the demultiplexing of a substantially greater number of reads at high error rates.

4.3 Methods

First, deML's algorithm is presented (see Section 4.3.1) followed by details about the sequencing test data (see Section 4.3.2).

4.3.1 Algorithm

DeML will compute the likelihood of assignment of a read to all potential samples of origin, assign each read to the most likely sample, and compute the uncertainty of the assignment.

If a run is double-indexed, with two indices of 7 basepairs each, there is a total of 14 nucleotides that were observed for the indices. Let $I = i_1, i_2, ..., i_{14}$ be the bases for a specific sample and $R = r_1, r_2, ..., r_{14}$ be the two sequenced indices with their respective quality scores $Q = q_1, q_2, ..., q_{14}$. Let m_i be a set of dummy variables which are equal to 1 if the corresponding bases between R and I match, or 0 otherwise. The likelihood of having sequenced the index given that it originates from a given sample, referred to as Z_0, is given by:

$$Z_0 = -10 \cdot \log_{10}[\prod_{i=1}^{14} m_i \cdot (1 - 10^{\frac{-q_i}{10}}) + (1 - m_i) \cdot 10^{\frac{-q_i}{10}}] \tag{4.1}$$

The Z_0 score is computed for each potential sample. Finally, the read is assigned to the most likely sample of origin. It can occur that a read is equally likely to belong to more than one sample. To quantify this uncertainty, the Z_1 score models the probability of misassignment. Let M be the number of potential samples of origin and let $Z_{0_1}, Z_{0_2}, ..., Z_{0_M}$ be the likelihood scores for each sample. Let t be the sample with the highest likelihood, the misassignment score is given by:

$$Z_1 = -10 \cdot \log_{10}[\frac{\sum_{i \in (1..M) \setminus t} 10^{\frac{-Z_{0_i}}{10}}}{\sum_{j \in (1..M)} 10^{\frac{-Z_{0_j}}{10}}}] \tag{4.2}$$

The overall algorithm can be described as follows:

Data: Set of reads to demultiplex R, set of samples S
Result: Sample ID for each read with probability scores
foreach *Read* $r \in R$ **do**
  foreach *Sample* $s \in S$ **do**
    Compute: Z_0 using equation 4.1 ;
  end
  Find sample $\hat{s}$ with max. Z_0 ;
  Assign r to $\hat{s}$;
  Compute : Z_1 using equation 4.2 ;
end

Algorithm 1: deML

In practice however, the likelihood of assignment to highly divergent indices will be negligible. Furthermore, their contribution to the sum in equation 4.2 will be equally negligible. To increase efficiency at an acceptable decrease in accuracy of the Z_0 and Z_1 scores, deML will only consider samples within a certain number of mismatches. The next section describes how deML efficiently searches for all indices in the user-provided list within a fixed maximal number of mismatches.

Sequence search

For a given observed sequenced index, deML needs to identify all possible index sequences from a user supplied list within a given number of mismatches. To achieve this in a timely fashion, deML builds a prefix tree of the user supplied indices which represent common prefixes as common paths in the tree (see Figure 4.1). The height of a given node directly indicates the position in the original index string.

An advantage of prefix trees is the ability to search with mismatches using recursive calls in the data structure. The call is launched on the root using the string to be searched and the tolerated number of mismatches. The recursive call is performed on the child nodes where the number of tolerated mismatches is decreased by one when the letter represented by the current node differs or, leaving the mismatch count as is otherwise. The query sequence is shortened by one after each function call. The recursion ends when the number of tolerated mismatches falls below zero or a leaf node is reached.

The overall prefix tree algorithm returns all possible indices within a fixed number of tolerated mismatches for downstream computations. Once all the indices have been identified, the likelihood of pertaining to each sample that has been detected is computed. As mentioned before, upon computing the sample assignment quality Z_1, the number of tolerated mismatches can be set to be lesser than the length of the indices as the contribution of the more divergent indices (edit distance exceeding the number of tolerated mismatches) can be generally considered negligible. As the likelihood of per-

taining to samples with low edit distance dwarfs the one to more divergent samples, their contribution can often be safely overlooked.

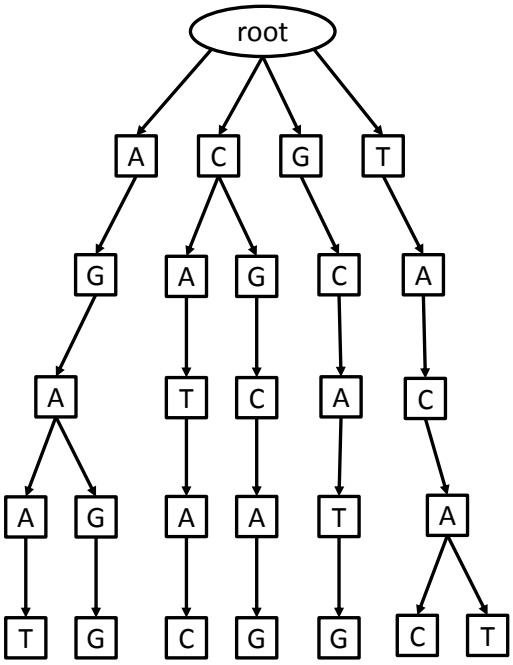


Figure 4.1: A prefix tree for the following sequences : AGAAT,AGAGG,CATAC,CGCAG,GCATG,TACAC,TACAT. The common prefixes become common paths in the tree.

It can happen however that the most likely sample based on the first index is not the same as the most likely sample based on the second index. The next section describes how to quantify the probability of mispairing.

Probability of mispairing

Due to shifting clusters or erroneous pairings of indices, in the case of double-indexed runs, a certain cluster on the flowcell can have a first index pertaining to a given sample and a second index belonging to another. In the Illumina sequencing technology, one of two adapters is referred to as the p7 adapter and the other is the p5 adapter (see Section 1.1.2). Let $r_{p7} = r_1, r_2, .., r_7$ and $r_{p5} = r_8, r_9, .., r_{14}$ be the two sequenced indices for the p7 and p5 adapters respectively. Let $I7_i$ and $I5_i$ be the sequences used for sample i which are considered to be the template. The probability ($P(r_{p7}|I7_i)$ of observing the sequenced data r_{p7} given that $I7_i$ was the template can be computed the same way as the

Z_0 score but using only the 7 available nucleotides. To quantify the risk of mispairing, the log odds ratio of the probability of mispairing to the sum of the probabilities for all pairs is used:

$$Z_2 = -10 \cdot \log_{10} \frac{\sum_{i \neq j} P(r_{p7}|I7_i) \cdot P(r_{p5}|I5_j)}{\sum_{i,j} P(r_{p7}|I7_i) \cdot P(r_{p5}|I5_j)} \tag{4.3}$$

However, the computation above is expensive. A potential way to speed it up is to consider certain terms as being negligible. If the best hit for both P7 and P5 stems from the same sample $\hat{i}$. Let the second best hit be $\hat{j}$ for each index. It can be assumed that remaining pairs are insignificant compared to the probability of pertaining to these two groups, equation 4.3 becomes:

$$Z_2 \approx -10 \cdot \log_{10} \frac{1}{2} \cdot \frac{P(r_{p7}|I7_{\hat{i}}) \cdot P(r_{p5}|I5_{\hat{j}}) + P(r_{p7}|I7_{\hat{j}}) \cdot P(r_{p5}|I5_{\hat{i}})}{P(r_{p7}|I7_{\hat{i}}) \cdot P(r_{p5}|I5_{\hat{i}})} \tag{4.4}$$

The scaling factor $\frac{1}{2}$ is due to the use of two terms in the numerator.

If the best hit for both P7 and P5 does not come from the same sample but rather two different ones, namely $\hat{i}$ and $\hat{j}$, equation 4.3 can be written as :

$$Z_2 \approx -10 \cdot \log_{10} \frac{2}{1} \cdot \frac{P(r_{p7}|I7_{\hat{i}}) \cdot P(r_{p5}|I5_{\hat{j}})}{P(r_{p7}|I7_{\hat{i}}) \cdot P(r_{p5}|I5_{\hat{i}}) + P(r_{p7}|I7_{\hat{j}}) \cdot P(r_{p5}|I5_{\hat{j}})} \tag{4.5}$$

The approximation for Z_2 from equations 4.4 and 4.5 is the one reported in practice by the software.

4.3.2 Empirical test data

To evaluate the correctness of the sample assignment based on the indices, double-indexed DNA libraries were produced from amplicons of a 245 bp region of chromosome 7 from 99 human samples and from PhiX DNA fragmented to 350 bp. More precisely, a 245 bases long fragment was amplified from human iPS cells digested in QuickExtract DNA Extraction Solution (epibio) using primers GGCTTAAGTCCTGCTGAGA and AGATAAATATAGAATAAAGCTCATGA. Each 25 l PCR reaction contained Phusion HF master mix (NEB) at 1X, each primer at 0.5 M, 0.024 l of template and the rest was water. The mixture was heated to 98C for 30 seconds, followed by 25 cycles of 98C for 10 seconds, 56C for 10 seconds and 72C for 10 seconds. PhiX DNA was fragmented with Covaris S2 with the 500 bases settings (duty cycle 5%; Intensity 3; cycle per burst 200; time 80s) which gave a fragments that had a mod length of 580 bases as judged by a 2100

Bioanalyzer (Agilent). Indexed Illumina libraries were prepared as described by [54] and indices are given in Table A.4 in the Appendix.

Double-indexing is increasingly used in applications requiring extremely accurate read assignment [54]. The reads were basecalled, demultiplexed using deML and mapped to both the human genome and the PhiX genome. The mapping of the forward and reverse reads indicates the sample of origin of the original cluster and were used to measure demultiplexing misassignments rates.

Using simulations, the robustness of deML read assignments was evaluated for datasets at various error rates. Indices with perfect matches to a known sample had sequencing errors added to them at various rates using an error profile derived from an Illumina MiSeq sequencing run. To provide a comparison to existing approaches with a variable number of allowed mismatches but no likelihood score to measure the rate of misassignments, deindexer (https://github.com/ws6/deindexer) was installed and tested. The number of sequences demultiplexed was computed for deML and for deindexer. The number of sequences with 0 or 1 mismatches was also measured as the standard Illumina demultiplexing approach (CASAVA) assigns sequences using this cutoff.

Accuracy of the quality scores for this dataset

DeML relies heavily on the base quality scores to compute the likelihood of pertaining to a given sample for a given read. In theory, the base quality scores reflect the probability of error and can therefore be used to accurately compute the probability of observing the bases from the read given a certain sequence template. It is therefore important for quality scores to accurately reflect the probability of error for a given base. The freeIbis basecaller (see Chapter 2) was used using quality score calibration. While plotting the resulting base quality scores against the observed error rate (see Figure 4.2), the ones predicted by freeIbis show significant improvement in terms of correlation upon comparison with the quality scores predicted by the default basecaller provided by the vendor, Bustard. However, results show that this algorithm is also useful for Bustard basecalled data since the correlation between the assigned confidence and observed false assignment rates as well as the robustness to error also hold true for Bustard basecalled data (see Section 4.4.8).

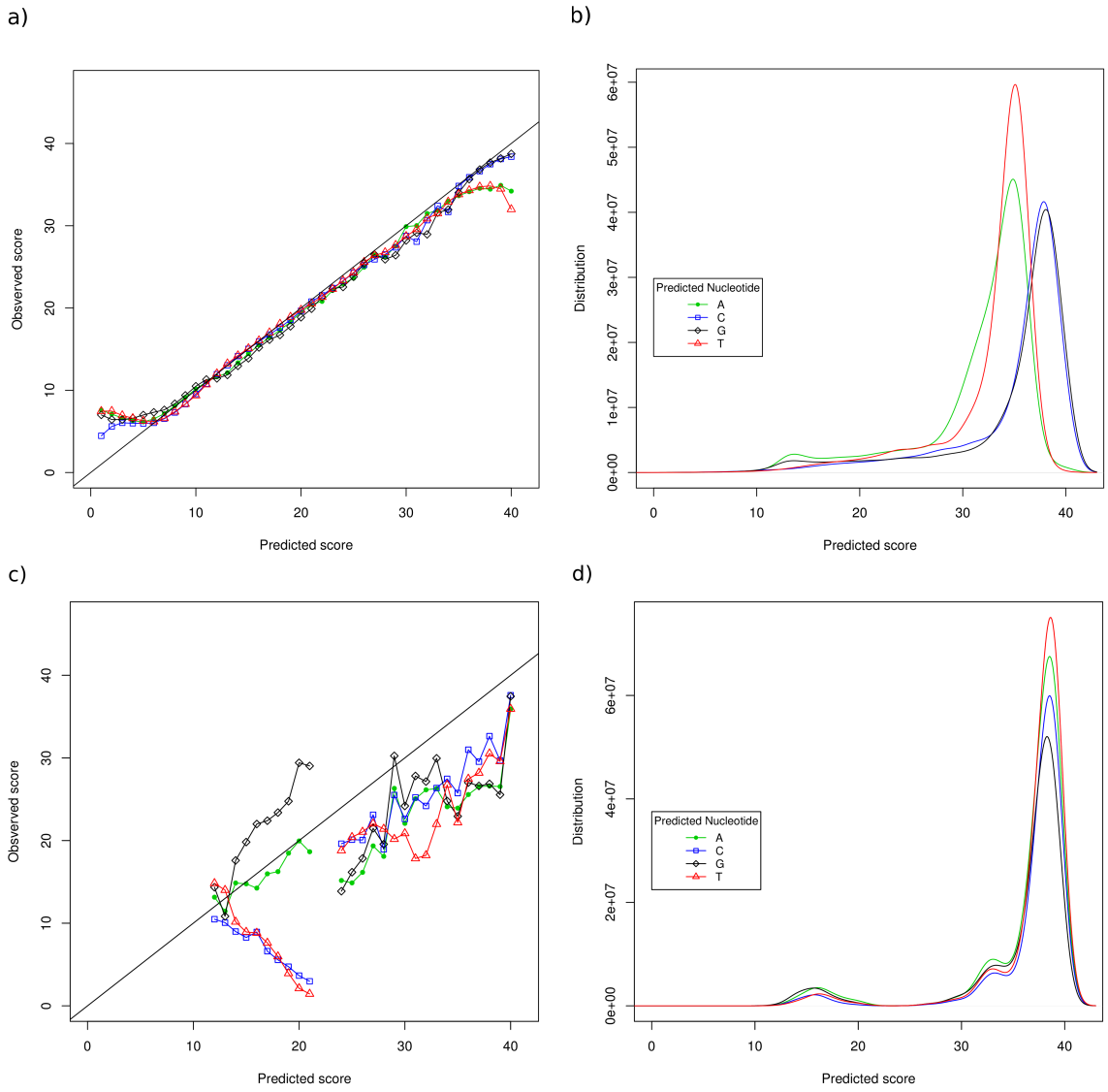


Figure 4.2: Quality scores for the MiSeq run described in this chapter. **(a)** The predicted versus observed quality scores for the sequenced bases basecalled using freeIbis. Ideally, the quality scores should follow the diagonal. **(b)** The distribution of the quality scores using a standard density plot for each nucleotide. **(c)** The predicted versus observed quality scores for the sequenced bases from the default basecaller. **(d)** The distribution of the quality scores provided by the default Illumina basecaller.

4.4 Results

The tally of how many reads map to the various targets is presented (see Section 4.4.1), followed by the distribution of the Z_0 and Z_1 scores for correct and incorrect assignments (see Section 4.4.2). The correlation of the Z_1 score and the false assignment rate is presented (see Section 4.4.3). Whether using both Z_0 and Z_1 scores together yielded better accuracy was evaluated (see Section 4.4.4) followed by deML's robustness to sequencing errors (see Section 4.4.5). Some results and discussion about discordant pairs (see Section 4.4.6) and the presence of a background error rate (see Section 4.4.7) follow. Finally, deML's accuracy when, instead of using freeIbis (see Chapter 2) the default Illumina basecalls are used, is presented (see Section 4.4.8).

4.4.1 Mapping statistics

The number of sequences for the empirical dataset mapping to the two reference genomes was evaluated. As the two potential templates of the sequencing run were human genome PCR product and PhiX controls, the fact that most reads map to either one of those two regions is expected. Each pair of reads was analyzed. A read can fall within 4 categories: PCR region, PhiX, mapped to a region outside the targets and unmapped. The tally (see Table 4.1) shows that most pairs are either both PCR product or both PhiX. Furthermore, a very small number (413) of read pairs had a discordant mapping for their forward and reverse reads where one mapped to the PCR product region and the other pair mapped to the PhiX genome. The possibility that these reads could be the result of shifting clusters is evaluated in section 4.4.6.

position of first mate	position of second mate	number
PCR product	PCR product	8,070,867
PCR product	PhiX	234
PCR product	outside targets	56
PCR product	unmapped	545,285
PhiX	PCR product	179
PhiX	PhiX	4,629,687
PhiX	outside targets	14
PhiX	unmapped	211,156
outside targets	PCR product	11
outside targets	PhiX	1
outside targets	outside target	10,084
outside targets	unmapped	24,496
unmapped	PCR product	241,960
unmapped	PhiX	66,132
unmapped	outside targets	22,470
unmapped	unmapped	1,368,465

Table 4.1: A tally for every possible combination of the forward and reverse read placement for the Illumina MiSeq presented in this thesis,

4.4.2 Distribution of the Z_0 and Z_1 scores

Out of the total of 15,245,844 clusters that were detected in the test dataset, 8,070,867 clusters had both forward and reverse reads aligning to the human control region and 4,629,687 to the PhiX. Using the sample assignment provided by deML for the reads mapping to the PhiX, the rate of false assignment was computed as a function of Z_0 and Z_1 scores. As expected, reads with a high likelihood of stemming from the PhiX control (Z_0) group and with a low likelihood of stemming from another sample (Z_1) were enriched for true assignments whereas misassignments were found at the other end of the distribution (see Figure 4.3).

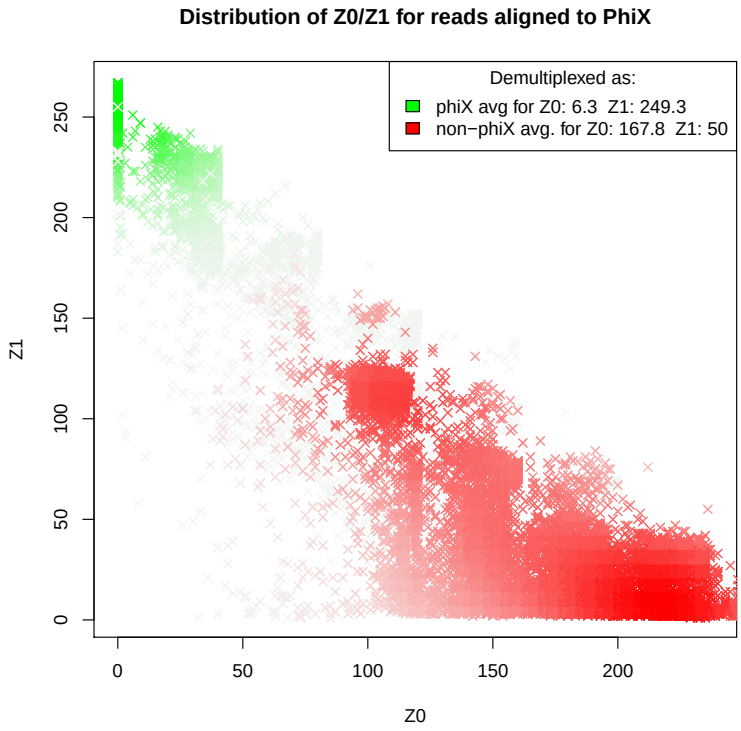


Figure 4.3: Distribution of true assignments (green) and false assignments (red) to the PhiX genome over their respective Z_0 and Z_1 score. The intensity of the color indicates the density of the data points for the given category.

The distribution of the Z_0 and Z_1 scores for true and false positives was evaluated. For reads aligning to the PhiX genome, it is assumed that reads demultiplexed as control are unequivocally true assignments, and human PCR samples are false assignments. For the Z_0 score (see Figure 4.4a), the majority of true assignments (green) have a high probability of pertaining to the sample to which they were assigned. False assignments (red) have on average a much lower probability of pertaining to the sample of origin as shown by the higher Z_0 score. The density of Z_1 score, the probability of pertaining to another sample than the most likely one, was also plotted (see Figure 4.4b). True assignments (green) have a lower probability of misassignment compared to actual misassignments (red).

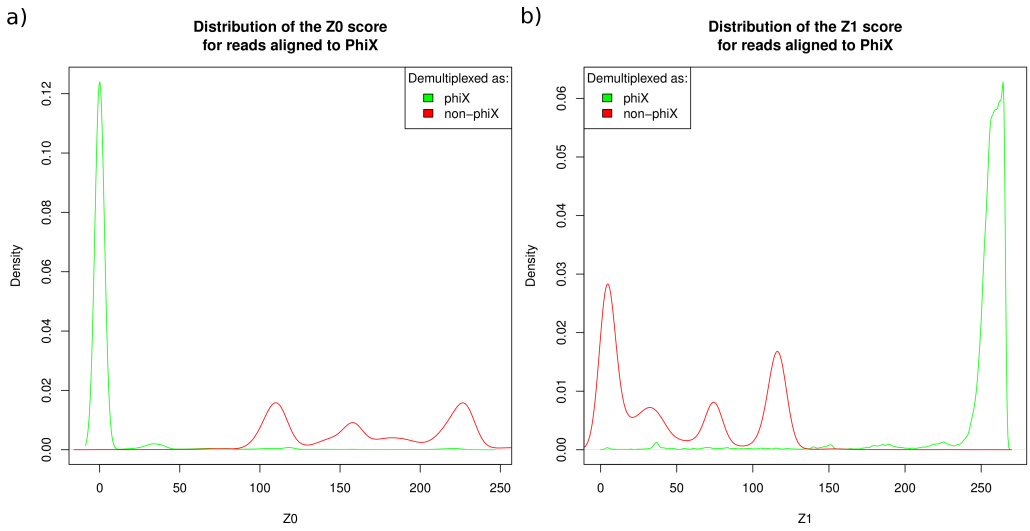


Figure 4.4: a) Distribution of the Z_0 score for reads aligning for the PhiX genome for reads either demultiplexed as control (green) or as human samples (red) b) Distribution for the same reads but for the Z_1 score.

4.4.3 Z_1 scores versus false assignment rates

As Z_1 measures the probability of misassignment on a PHRED scale given the potential index sequence set, the relationship between the misassignment rate on a log scale and the Z_1 score should be linear. For reads where both mates aligned to the PhiX, the misassignment rate was computed by considering any read pair not assigned by deML to the PhiX as a mislabeling. Since Z_1 can take many discrete values, the misassignment rate was plotted for multiple bins of Z_1 (see Figure 4.5).

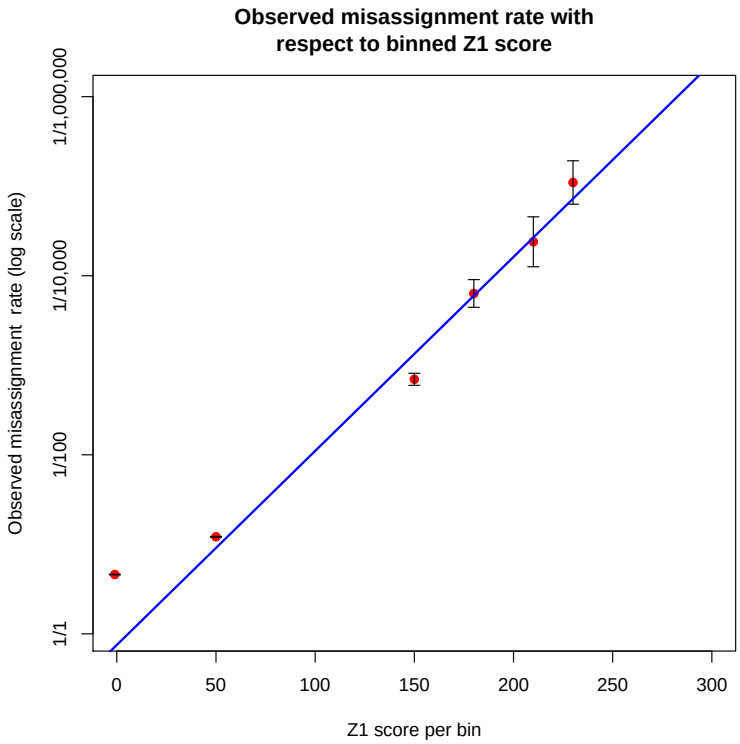


Figure 4.5: Correlation between the Z_1 score for reads aligned to the PhiX genome and the observed misassignment rate. Error bars were obtained using Wilson score intervals.

4.4.4 Predictive power of combined scores

To evaluate whether having both Z_0 and Z_1 scores has better predictive power than solely using a single one, a logistic regression was performed using each score individually and both at once. Using the PhiX data presented in Figure 4.5, positive and negative assignments were used as labels and the Z_0 and Z_1 scores were used as potential predictive values. The classification was performed using a logistic regression using the glm() function in R version 3.0.1. For all 3 set (Z_0, Z_1 and Z_0, Z_1 combined), the number of misclassifications was computed. The lowest number of misclassifications was obtained using both scores in conjunction (see Table 4.2).

predictor	misclassification out of 84,178
Z_0	1,751
Z_1	1,628
Z_0 and Z_1	1,606

Table 4.2: Predictive value of the Z_0, Z_1 and both scores used in conjunction to classify correct assignments from misassignments using the data presented in Figure 4.5

4.4.5 Robustness to sequencing errors

To evaluate the robustness of the demultiplexing to increased error rates, reads with perfect matches to an index sequence from the initial list were taken from the original set and mismatches were added using an Illumina error profile. This profile contains sequencer-specific nucleotide substitutions along with quality scores for those. The number of sequences with perfect, 1 mismatch and 2 or more mismatches to the original indices is presented. DeML retrieves more sequences and achieves a lower false discovery rate than currently available approaches (see Table 4.3 and Table A.5 in the Appendix). At higher error rates, the number of demultiplexed reads from the default software provided by the vendor decreases substantially as sequences with 1 mismatch or less are usually the only ones identified by the said software. The robustness of deML compared to a fixed-mismatch approach to increased simulated error rates was also plotted. The limits of heuristics using fixed-mismatches like CASAVA are plotted in Figure 4.6a. After a certain error rate, the number of sequences with either 0 or 1 mismatch decreases. However, results show that deML can confidently assign sequences even at very high error rates (see Figure 4.6b).

In conclusion, deML shows greater robustness to increased error rates while keeping a misassignment rate under 0.5% even at very high error rates for sequences meeting the default thresholds.

error per base	deML			deindexer			CASAVA	
	TP	FP	FDR	TP	FP	FDR	0 mm	1 mm
0.002408	12,374,119	1	0.00%	12,372,007	0	0.00%	11,962,540	405,318
0.101145	11,898,460	205	0.00%	9,784,321	146	0.00%	2,783,384	4,381,588
0.196708	9,779,898	2,761	0.03%	5,659,886	1,683	0.03%	577,456	1,978,848

Table 4.3: Number of sequences demultiplexed by deML and deindexer in terms of true positives (TP), false positives (FP) and false discovery rate (FDR) for 12,374,149 sequences. The leftmost column represents the average error rate per-base after simulated sequencing errors were added. The remaining columns present the number that could be identified using an approach allowing 1 mismatch (such as CASAVA).

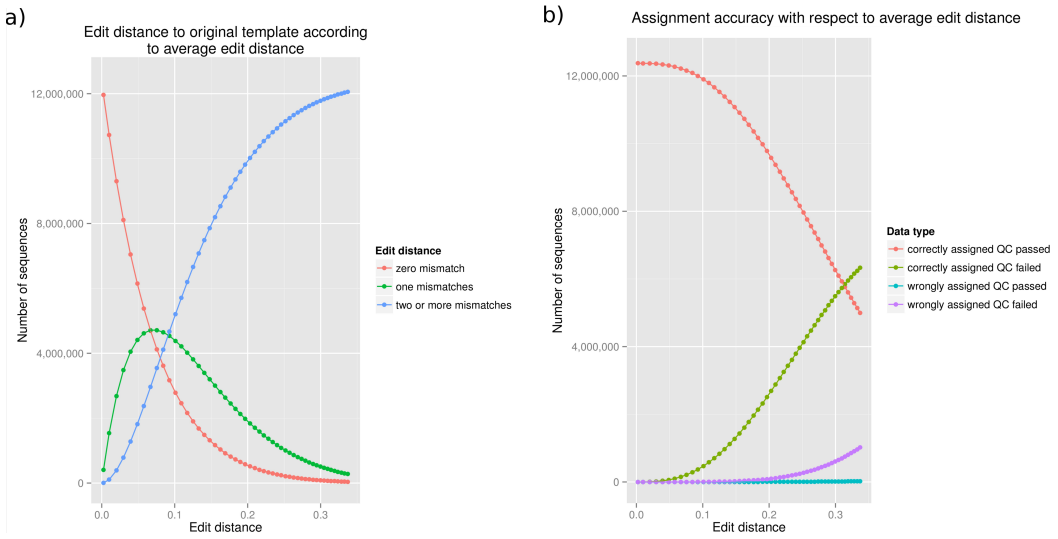


Figure 4.6: a) The edit distance of the simulated indices to the original index sequence as a function of the simulated edit distance to the original indices. This graph indicates the limits of heuristics using fixed-mismatches like CASAVA. b) For the same dataset, the number of sequences correctly assigned to the original sample for both the ones that passed quality threshold and those that did not. The number of incorrect assignments is also reported for both categories.

4.4.6 Discordant pairs

As mentioned in section 4.4.1, a small number (413) of read pairs exhibited unexpected mapping patterns (e.g. first mate mapping to PhiX and second one mapping to the PCR region). For those 413 clusters, the possibility that they might have been generated by shifting clusters was tested. These clusters should therefore have a high probability of error. For a read of length L and where q_l is the PHRED quality score for the base at position l, the expected number of mismatches to the reference is computed using the following expression:

$$\frac{\sum_{l=1}^{L} 10^{\frac{-q_l}{10}}}{L} \tag{4.6}$$

as previously defined in section 1.2.3. The expected number of mismatches for this subset was calculated to be 3.01 mismatches per 100 bases. To assess whether this number is higher in a statistically significant way, 10,000 subsets of 413 clusters were selected at random from the initial BAM file. The distribution of the expected number of mismatches for those randomized subsets were plotted (see Figure 4.7) against the same number of these discordant pairs. The expected number of mismatches is higher than any of the random subsets ($p<0.0001$).

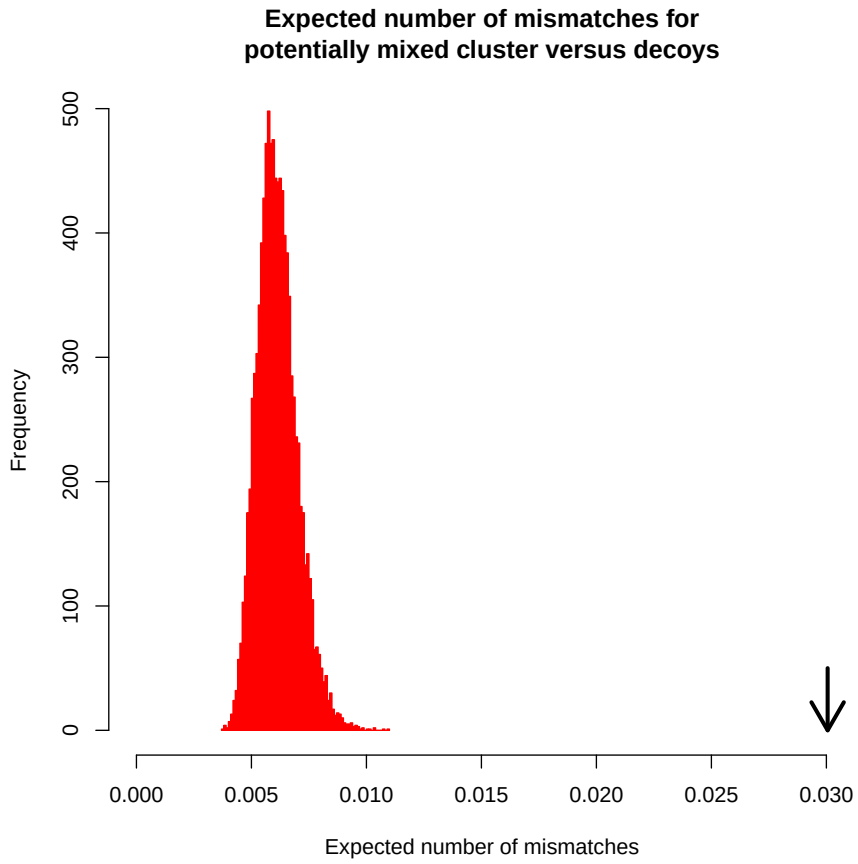


Figure 4.7: The expected number of mismatches for the 413 discordant pairs (e.g. one mate mapping to the PCR human target, the other mapping to the PhiX genome) is represented as a black arrow. The distribution of the expected number of mismatches for 10,000 subsets of 413 pairs taken at random is represented in red.

4.4.7 Background error rate

As mentioned in the discussion, if only clusters with a high probability of pertaining to their respective sample ($Z_0 = 0$) are considered, where both pairs map to the PhiX, the overwhelming majority were demultiplexed as PhiX. However, there were 9 clusters (18 sequences in total) which were assigned to the human PCR region samples. In theory, such sequences with indices matching perfectly the ones from samples pertaining to PCR regions yet where the forward and reverse read map to the PhiX control should not exist. To investigate whether mixed clusters could have produced such sequences, the expected number of mismatches for those 18 sequences was computed. The same quantity was computed for 10,000 independently sampled subsets of 18 sequences for the entire dataset.

A comparison reveals that those sequences do not have an expected number of mismatches above those of the background (see Figure 4.8) thus making the mixed cluster hypothesis unlikely.

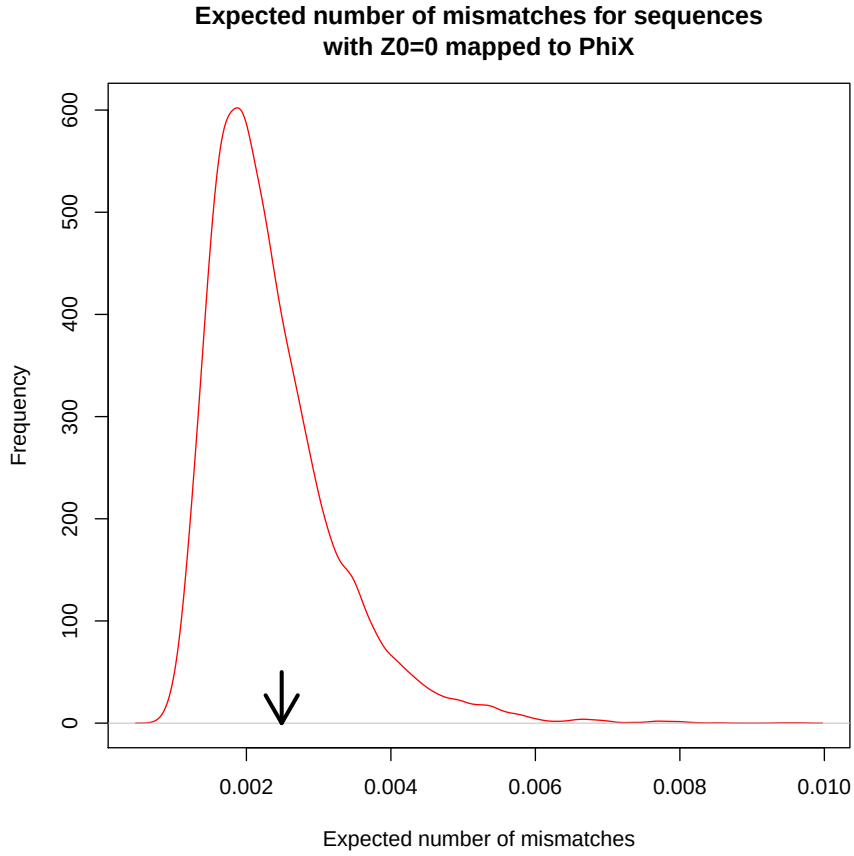


Figure 4.8: The distribution of the expected number of mismatches for 10,000 sets of 18 randomly chosen sequences mapping to the PhiX genome with $Z_0 = 0$ (red line) versus the ones not demultiplexed as PhiX but as human PCR region (black arrow).

4.4.8 Demultiplexing with default quality scores

For the MiSeq run used in this thesis, predicted quality scores produced by Bustard (the default Illumina basecaller, see Section 1.1.2) do not have a perfect correlation to their observed ones. Some groups rectify this discrepancy after basecalling using the Genome Analysis Toolkit (GATK) however, this is not feasible for index sequences (see [74] and section 2.1.3). As deML relies on quality scores, whether the algorithm would work equally well for sequence data produced by the default Illumina basecaller was

evaluated. More precisely, the correlation between the false assignment rate and the Z_0 and the Z_1 scores was evaluated. The same data was demultiplexed but instead of the freeIbis basecalls, the default Illumina basecalls were used. The distribution of the false assignments versus true ones were plotted (see Figure 4.9). Furthermore, the correlation between the misassignment rate and the Z_1 score was also measured (see Figure 4.10). In both cases, the correlation between both scores and the false assignment rate holds. This is a likely consequence of the fact that quality scores produced by Bustard, albeit not having a perfect correlation to their observed error rates, offer a reasonable approximation for the most part (quality scores between 30 and 40). Similarly to freeIbis, the quality scores at the lower end of the distribution (less than 20 on the PHRED scale) do not seem to correspond to their observed error rate. As a consequence, the first data point in Figure 4.10 does not seem to follow well the predicted linear relationship.

Whether deML would provide the same robustness for the data with simulated error rates with Bustard quality scores was also tested. In an approach identical to the one used for the freeIbis basecalled dataset, mismatches in the indices were added at various rates. The substitutions and quality scores for the mismatching bases were added using a Bustard error profile obtained from control sequences aligned to the PhiX. The number of sequences that could be demultiplexed by deML greatly exceeds the retrievable number of sequences using the default strategy of allowing 1 mismatch, especially at high error rates (see Figure 4.11). Similarly to results obtained on the data basecalled using freeIbis presented in Table A.5 in the Appendix, at the highest error rate, this set also had a low number of false assignments (8,469 sequences) out of those that passed the default quality thresholds (2,605,363 sequences) for a maximal observed false assignment rate of 0.33%.

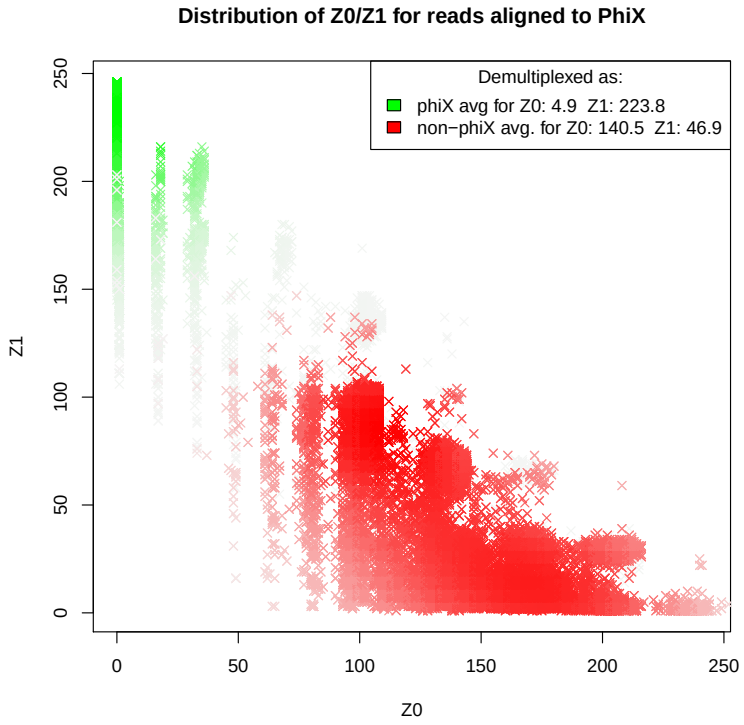


Figure 4.9: Distribution of true assignments (green) and false assignments (red) to the PhiX genome over their respective Z_0 and Z_1 score for reads from the Bustard basecaller. Like Figure 4.3, the intensity of the color indicates the density of the data points for the given category.

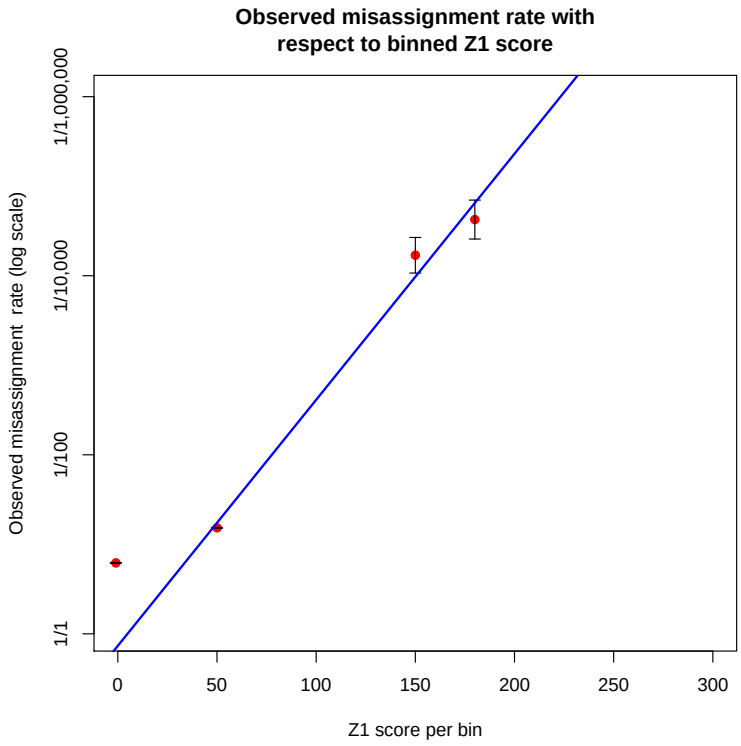


Figure 4.10: Correlation between the Z_1 score for reads aligned to the PhiX genome and the observed misassignment rate on a log scale for the Bustard basecalled reads. Like Figure 4.5, the line is a linear regression on all but the first data points. The size of the bins are the same as the ones used for Figure 4.5 except that no false assignments were seen for this dataset for a Z_1 score above 200 hence no data point was reported.

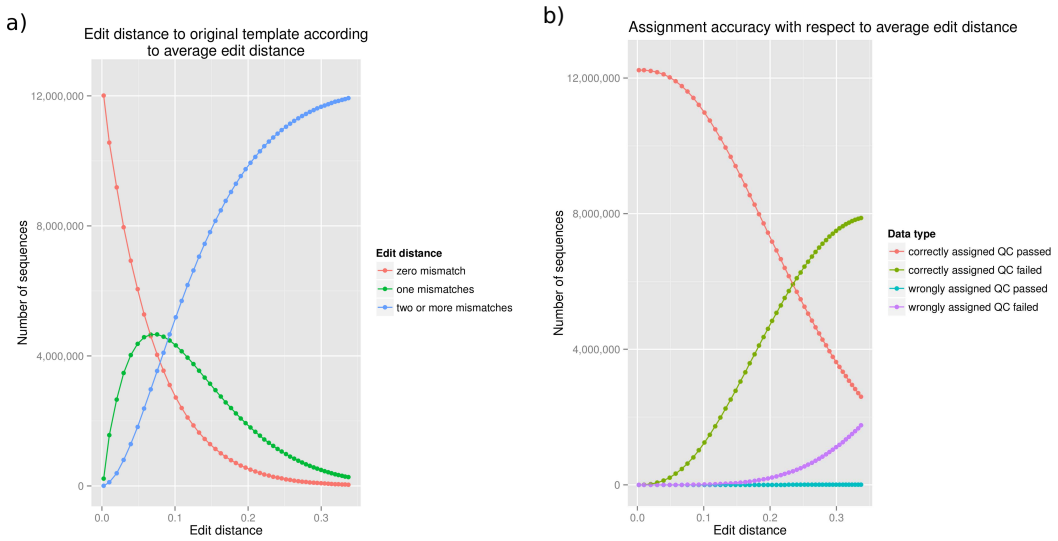


Figure 4.11: a) The edit distance of the simulated indices to the original index sequence as a function of the simulated edit distance to the original indices for Bustard basecalled data. This graph indicates the limits of heuristics using fixed-mismatches like CASAVA. b) For the same dataset, the number of sequences correctly assigned to the original sample for both the ones that passed quality threshold and those that did not. The number of incorrect assignments is also reported for both categories.

4.5 Conclusion

DeML is a maximum-likelihood approach that assigns each read from a multiplexed sequencing run to the most likely sample of origin and computes the confidence in this assignment using the likelihood of assignment to other possible samples. This confidence correlates positively with the rate of correct assignment.

DeML offers the possibility of demultiplexing problematic datasets and can confidently retrieve more sequences than the default Illumina pipeline for such sequencing runs. Such an approach allows users to specify the tolerated level of confidence for read assignment depending on the type of biological question being addressed.

Chapter 5

Endogenous genome inference and contamination estimates

This chapter introduces schmutzi, a MAP algorithm aimed at reconstructing the endogenous mitochondrial genome and estimating present-day human contamination for ancient human samples.

5.1 Background

When sequencing the mitochondrial genome of an archaic sample for which a reference genome is available, one crucial task is to determine the differences between the sample and the reference. Theses differences can either be single nucleotide substitutions or insertions/deletions. When analyzing data from ancient humans, the contribution from contaminating fragments from present-day humans can cause miscalls. The proportion of such fragments in the entire dataset should also be quantified.

This subsection discusses current methodology for two critical tasks in the analysis of mitochondrial data from archaic genomes: inferring the endogenous genome and quantifying present-day human contamination.

5.1.1 Endogenous genome inference

Previous approaches to reconstructing ancient mitochondrial genomes include the mapping iterative assembler (MIA) which iteratively calls a consensus from the DNA fragments [44]. When contamination is high (e.g. >30%), calling the consensus sequence of the endogenous mitochondrial genome without removing contaminant fragments is likely to result in an incorrect sequence (see Figure 5.1). Because ancient endogenous DNA

is more likely to be deaminated than the contaminant DNA from present-day humans [102], some studies have restricted the analyses to fragments carrying deaminated cytosines [76, 106]. However, using only deaminated fragments reduces the amount of data available for many ancient samples.

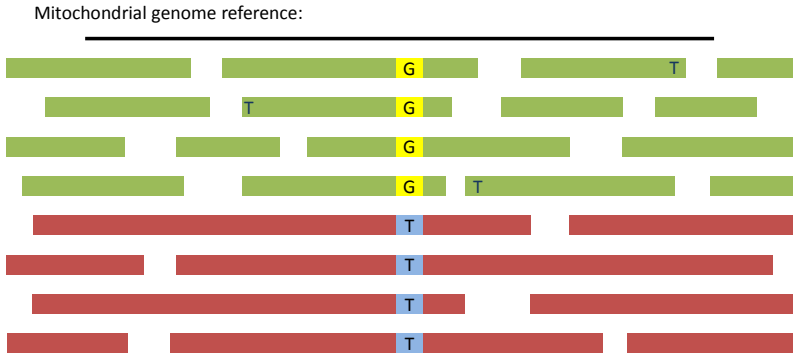


Figure 5.1: Schematic illustration of mitochondrial sequences from an ancient DNA library. When DNA from an ancient human sample is sequenced, DNA from the ancient human ("endogenous fragments" represented in green) as well as contaminant DNA fragments from the individuals that have handled the bone ("contaminating fragments" represented in red) are included. Because DNA undergoes deamination over time, endogenous fragments are likely to carry deaminated cytosines (represented as blue 'T's), particularly near the ends of the DNA fragments. schmutzi first identifies the endogenous fragments and, in a second step, uses these to quantify contamination. These steps are repeated until convergence is achieved and a single mitochondrial genome is identified.

5.1.2 Contamination estimates

Due to these issues, research groups have generally prioritized samples with low levels of present-day human contamination. To date, methods to quantify present-day human mitochondrial contamination have relied on the presence of fixed differences between the mitochondrial genomes of archaic and modern humans [43, 89]. This works well when analyzing the genomes of Neanderthals and Denisovans, but early modern human genomes typically carry too few fixed differences to permit a robust estimate of contamination. For early modern humans, various groups have therefore relied on sites in the ancient sample that differ from a large dataset of present-day human mitochondrial sequences [83]. Additionally, a maximum-likelihood approach which co-estimates sequencing error rates and contamination has been applied to sequences originating from both early modern humans and archaic humans [36]. Albeit not widely available for download, an implementation in R of this approach, called contamMix, has been distributed via email upon request by its author. Deamination patterns have also been used to estimate contamination from present-day humans in mitochondrial DNA [76]. Software tools are available to measure

overall deamination [47], to isolate deaminated fragments [106] and to perform nuclear contamination estimates based on the X-chromosome [58]. However, there is currently no software for estimating mitochondrial contamination, which has been thoroughly tested to ascertain its accuracy, available for download for the aDNA research community.

5.2 Introduction

To address the issues defined in 5.1, "schmutzi", a MAP approach was developed to assemble the endogenous mitochondrial genome while simultaneously estimating present-day human mitochondrial contamination in archaic and early modern human aDNA datasets. The approach to determine the endogenous mitochondrial genome sequence relies on distinguishing the endogenous and the contaminant nucleotides, given a prior on: contamination, deamination frequency and distribution of the length of the fragments. Contamination is estimated using single nucleotide differences between the endogenous mtDNA sequence and a database of potential contaminant mitochondrial genomes. The consensus calling and contamination estimation are run iteratively until a stable contamination rate estimate is reached.

schmutzi was tested on both simulated and empirical data. Results show that schmutzi outperforms currently available methods in terms of accuracy of the endogenous call and contamination estimate, particularly at high levels of contamination. An open-source implementation of schmutzi in C++ is released under the GPLv3.0 and is freely available together with the test datasets that were used from http://bioinfo.eva.mpg.de/schmutzi. On a desktop computer, schmutzi requires between 1 and 3 hours to reach convergence for approximately 1M fragments aligned to the mitochondrial reference genome. Faster runtimes (~ 30 minutes) can be achieved using multi-core systems.

5.3 Methods

The input for schmutzi is a set of aligned fragments ideally produced by freeIbis (see Chapter 2) followed by leeHom (see Chapter 3). Some recommendations regarding the alignment of aDNA fragments to the mitochondrial reference (see 5.3.1) are presented. The description of schmutzi's overall algorithm follows (see Section 5.3.2). More specifically, the contamination estimate based on deamination patterns is presented (see Section 5.3.3), followed by the algorithm for endogenous consensus calling (see Section 5.3.4). The algorithm for contamination estimates based on the differences between the predicted endogenous genome and a database of putative contaminants is then presented (see Section 5.3.5). A description of a previously published maximum-likelihood method follows (see Section 5.3.6). Observations about differences between the length of endogenous and contaminant fragments are then presented (see Section 5.3.7). A section discusses the database of putative mitochondrial contaminants (see Section 5.3.8). Finally, a description of the empirical test data used is presented (see Section 5.3.9).

5.3.1 Mitochondrial mapping strategies

This section presents recommendations when alignment aDNA to a mitochondrial genome reference. More specifically, this section covers the inclusion of the circularity of the mitochondrial genome (see Section 5.3.1.1) and the sensitivity of the aligner to highly divergent loci on the mitochondrial genome (see Section 5.3.1.2).

5.3.1.1 Handling circular references

Prior to running schmutzi, all fragments from both the contaminant and endogenous genomes must be aligned to a reference genome.

Most aligners for NGS data do not allow for circular reference genomes leading to spurious drops of coverage around the ends. To circumvent this, the first 1000 basepairs of the mitochondrial reference can be appended at the end and used as new reference. A script [1] folds alignments spanning the end of the mitochondrion back to the beginning of the reference.

To illustrate the corrective effect on coverage, a set of 1M fragments of 100 bp from the revised Cambridge Reference Sequence (rCRS) mitochondrion (GenBank: NC_012920) were simulated. Random coordinates were simulated using a uniform distribution and fragments were allowed to span the sequence junction as to reflect circularity.

Fragments were simulated using in-house programs [2]. The fragments were aligned to the default reference using BWA v0.5.10[63]. In a separate set, the fragments were

[1] https://github.com/udo-stenzel/biohazard/
[2] https://github.com/grenaud/simulateAncientDNA

aligned to the extended reference genome and fragments spanning the junction of the genome were folded back.

Figure 5.2 shows the coverage for the first and last bases of the mitochondrial reference. The advantage of accounting for circularity in mapping is seen by the more even coverage, compared to the alignment to the standard reference genome.

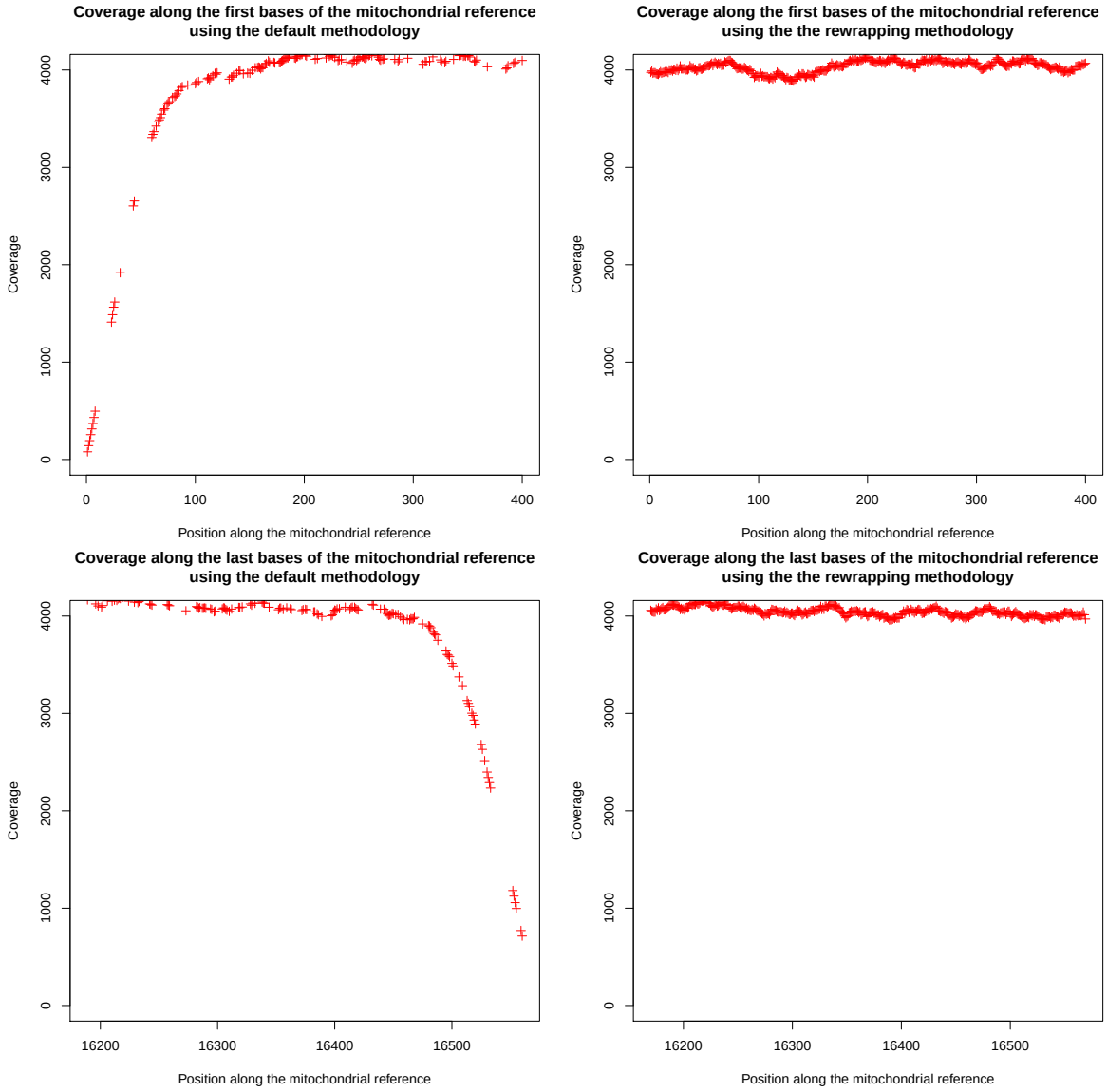


Figure 5.2: Coverage for the first 400 bases of the mitochondrial genome (top) and last 400 bases (bottom) for simulated short fragments from the rCRS reference. Without accounting for circularity (left) an artificial drop of coverage can be seen. However, if circularity is taken into account (right), the end of the sequence in the reference file does not influence coverage.

5.3.1.2 Mapping sensitivity

The lack of sensitivity of the aligner for highly divergent loci can create a bias towards having a greater proportion of contaminant fragments aligning than the average across the mitochondrial genome (see Figure 5.3). This is particularly true for highly divergent samples like the Denisovan mitochondrion[3] [59]. To illustrate this, fragments from the Denisovan mitochondrial genome were simulated. Its divergence against the human genome was plotted (see Figure 5.4). The regions of the mitochondrial genome with the highest divergence can be found around the displacement loop (D-loop) [2].

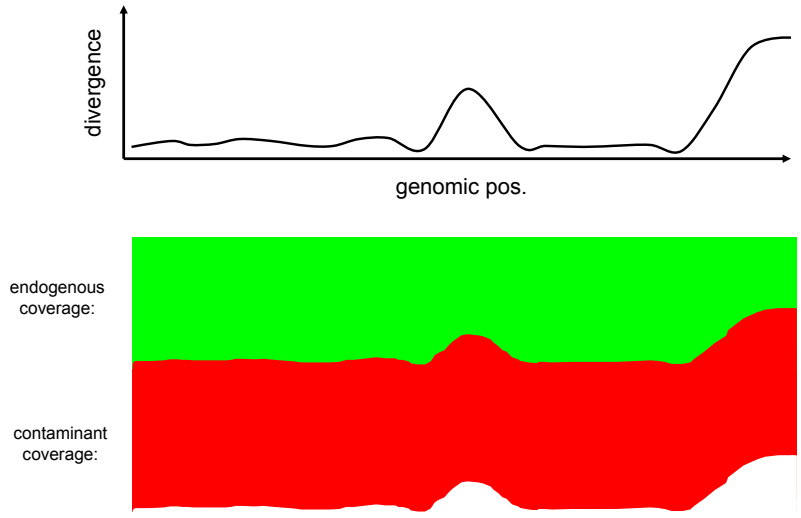


Figure 5.3: Schema of the effect of using a low sensitivity aligner to the human mitochondrial reference in regions of high divergence. The endogenous ancient DNA has higher divergence to the reference than the contaminant creating the possibility that the endogenous fragments will not align due to a higher edit distance. Although the distribution of the fragments will be representative of the contamination rate in regions of low divergence, contaminant fragments may overtake endogenous ones in regions of high divergence. A paucity of endogenous fragments could lead to an inability to call certain regions and an overestimate of the contamination rate.

[3] GenBank: FN673705.1

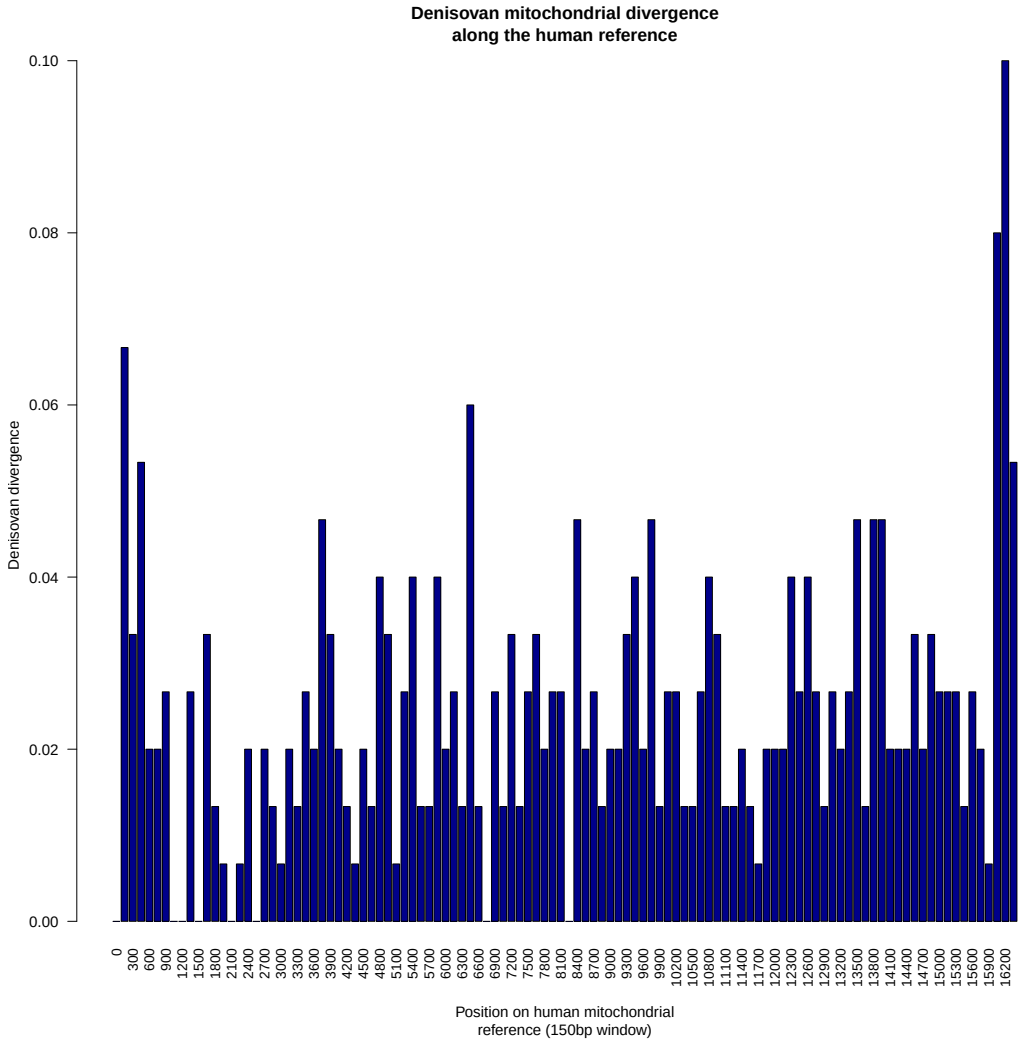


Figure 5.4: Divergence of the Denisovan mitochondrial genome when aligned to the human reference for windows of 150 basepairs. The most divergent portions of the genome are found in the vicinity of the D-loop.

To evaluate how currently used aligners would handle such a bias, aDNA fragments from the Denisovan mitochondrial genome were simulated again using the strategy described in the subsection above. The simulated length of the fragments was taken from empirical distributions (see Section 5.3.7). Deamination rates were added using the deamination rates from the single-stranded libraries from [60]. Sequencing errors were added along with representative quality scores using empirical rates obtained using Illumina

reads of PhiX control (see Section 3.4.2). The fragments were aligned to the extended human mitochondrial reference using both BWA v0.5.10 (with "-n 0.01 -o 2 -l 16500", optimized for increased sensitivity for ancient DNA [103]) and SHRIMP v2.2.3[20] ("-N 5 -o 1 –single-best-mapping –sam-unaligned –fastq –sam –qv-offset 33"). Again, fragments spanning the junction of the genome were wrapped back at the beginning. The impact of the mapping algorithm used on coverage for the endogenous and contamination is presented in Figure 5.5 which shows the correlation between divergence and coverage. When using BWA, even with parameters tailored for aDNA, a lesser number of fragments align to highly divergent loci. SHRIMP, a more sensitive aligner (see [100]) seems more robust to highly divergent loci. To avoid coverage biases between endogenous and exogenous material, a sensitive aligner is required to accurately quantify contamination.

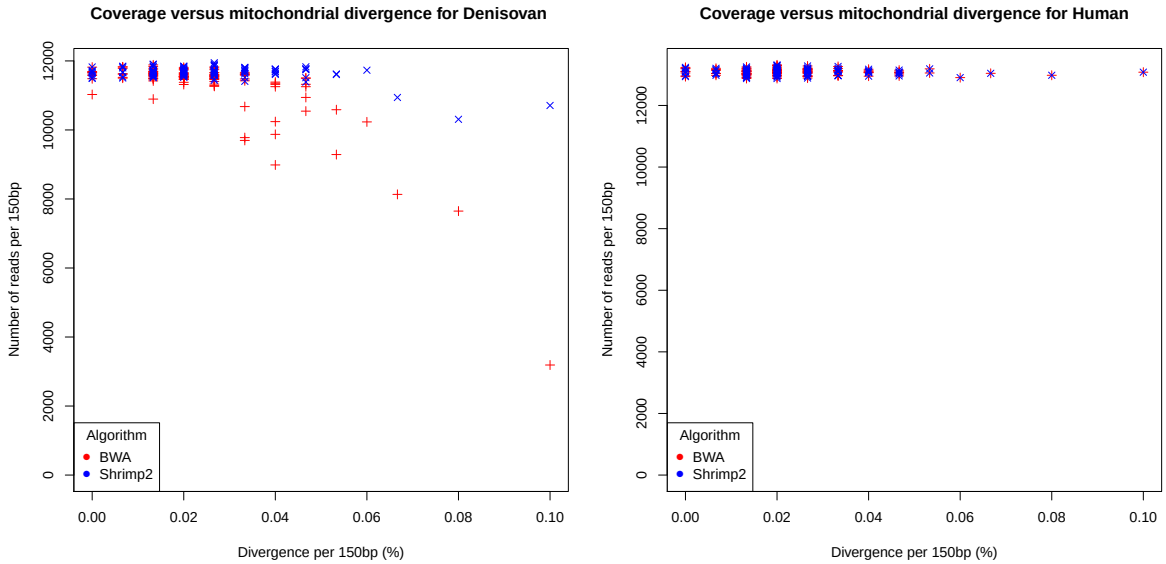


Figure 5.5: Effect of Denisovan mitochondrial divergence on coverage depending on the aligner. Certain mitochondrial loci of the Denisovan mitochondrial genome are highly divergent to the human reference. The coverage per region is presented both for simulated endogenous fragments from the Denisovan (left) and contaminant fragments (right). BWA (red) performs well at low divergence. At high levels of divergence, the fraction of the contaminant and endogenous fragments that align will not follow the average over the entire genome thus potentially leading to overestimates of contamination rates. SHRIMP (blue) has greater sensitivity to higher divergence and therefore this effect is less prominent.

5.3.2 Overview of schmutzi's algorithm

Schmutzi, iteratively calls (i) the endogenous mitochondrial consensus sequence and (ii) a contamination estimate, until a stable contamination rate is reached using two linked software programs: "endoCaller" and "mtCont" (see Figure 5.6).

The input for the consensus caller is: a set of fragments generated from the sequencing of an ancient DNA sample and aligned to a mitochondrial genome reference, a contamination prior and deamination rates of the potentially endogenous and potentially contaminating DNA fragments. The deamination rates and the prior for contamination are obtained from "contDeam", a sub-program of the schmutzi package (see Figure 5.6). This is an implementation of the methodology described in previous studies [76], but incorporates some additional information including base quality and mapping quality into a Bayesian framework. An underlying assumption is that the base qualities are reasonably representative of the sequencing error probability (see Chapter 2). The inputs for the contamination estimator "mtCont" are the same set of aligned fragments, the endogenous consensus sequence and a database of potential contaminant mitochondrial genomes.

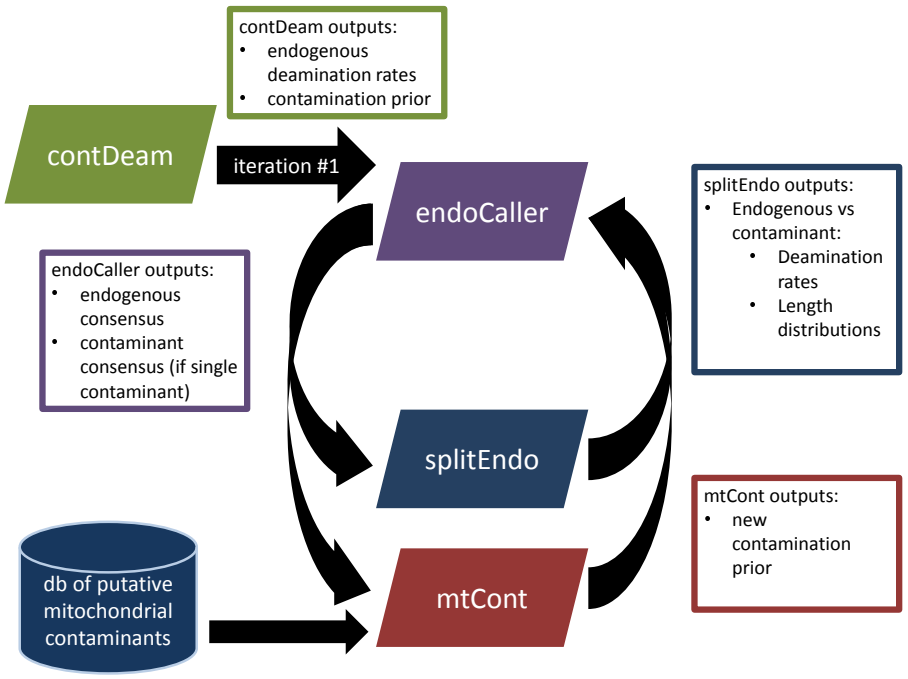


Figure 5.6: schmutzi's workflow. An initial contamination estimate is computed using the deamination rates of fragments by conditioning on the other end being deaminated and comparing these to the deamination rate of all fragments in the dataset (contDeam). This prior is provided to call an endogenous consensus (endoCaller). The consensus call is, in turn, used to re-estimate mitochondrial contamination (mtCont). Deamination rates and fragment length distributions are measured for fragments that support endogenous and contaminant mitochondrial genomes (splitEndo). The information from mtCont and splitEndo are used as input for re-calling the endogenous consensus (endoCaller). This cycle is repeated until a stable contamination rate is reached.

5.3.2.1 Simulated and empirical test data

The performance of schmutzi on simulated and empirical mitochondrial sequence data from both archaic humans and early modern humans was tested. Simulated mtDNA datasets with increasing levels of contamination were created by fragmenting and deaminating the mitochondrial genome sequences of a Denisovan (GenBank: FN673705.1) [77], a Neanderthal (GenBank: AM948965.1) [89] and an early modern human (Ust'-Ishim individual [35]) and adding increasing amounts of contamination from a single, randomly-selected present-day human mitochondrial genome (GenBank: KJ446110.1). Empirical deamination rates were used, more specifically, the data prepared using a double-stranded library preparation protocol (C $\rightarrow$ T at the 5' end and G $\rightarrow$ A at the 3' end, rates from [83]). The simulations were repeated by adding deamination rates from empirical data

prepared using a single-stranded library protocol (C $\rightarrow$ T at both ends, rates from [60]).

The empirical data included Illumina sequences from the same three ancient individuals as well as sequence data for two additional Neanderthal individuals from Mezmaiskaya [38] which were selected because of the high rate of present-day human contamination present in the sequencing libraries [89] (see Section 5.3.9 for greater detail about the empirical datasets).

The accuracy of the consensus sequence called by schmutzi was compared to the consensus sequences generated using a set of typical approaches that have been described in the literature: (i) MIA [44], (ii) PMDTools to identify deaminated reads followed by a haploid consensus call using htslib [106] and (iii) samtools mpileup[4] after removing deaminated reads [76]. Schmutzi's contamination estimates was also compared to the known contamination in the simulated sequence data, to the estimates based on diagnostic sites for the empirical data, and to the estimates obtained from the maximum-likelihood approach described in [35, 36]. This is currently the only published method that can estimate mitochondrial contamination for both early modern humans and archaic humans. In order to assess the robustness of schmutzi to varying coverage, fragments were downsampled from 1% to 50% of the data using a uniform probability distribution.

Firstly, how a reasonable contamination prior can be obtained using deamination patterns is presented. Details of the algorithm behind the endogenous consensus caller are then provided. Finally, how the contamination is estimated using the output of the endogenous consensus caller is shown.

5.3.3 Determining a contamination prior using deamination patterns

The first iteration of the endogenous genome inference needs a contamination prior that is ideally a reasonable approximation of the actual contamination rate. This first contamination estimate is computed by "contDeam" (see schematic in Figure 5.6). This program computes the likelihood of observing the aDNA fragments aligned to the reference genome given fixed endogenous deamination patterns and a prior on the rate of present-day human contamination. It then returns the contamination rate with the highest posterior probability. This contamination rate is the most likely value needed to explain the difference between deamination rates for fragments identified as endogenous and overall deamination rates for all the fragments of the entire dataset. An assumption is that only the endogenous DNA has the deamination patterns typical of ancient DNA and that contaminant fragments are not deaminated and will therefore only reduce overall deamination rates. Previous studies suggest that deamination is rare in contaminants younger than about 100 years old [102]. Having deaminated contaminant fragments may lead to underestimates. The extent of the potential underestimate is discussed at the end of this section.

[4] https://github.com/samtools/htslib

To identify the endogenous fragments and derive their deamination rate, there are two possible approaches. The first involves the separation of the endogenous and contaminant fragments using diagnostic positions on the mitochondrial genome. This is relatively straightforward when dealing with Neanderthal or Denisovan individuals, as their mitochondrial genome sequences fall outside of present-day human variation [42, 59]. For instance, there are 111 diagnostic positions on the mitochondrial genome sequence, at which 7 Neanderthals share the same base, which differs from 20 present-day humans.

However, when the endogenous sample is an early modern human and falls within present-day human variation, this approach lacks power due to the rarity of such diagnostic sites. A second strategy takes advantage of the observation that deamination at the 5' end of the fragment is independent of the deamination occurring at the 3' end and vice-versa. By conditioning on observing deamination at one end and measuring the rates of deamination at the other, an estimate of the deamination rates of the endogenous fragments can be obtained [76]. This second strategy requires an endogenous base to measure rates of deamination. The mitochondrial reference sequence is therefore used as the endogenous template. This assumption yields accurate results even for the highly divergent Denisovan mitochondrial genome. The contamination prior estimated by schmutzi uses this second approach by default. The estimate of endogenous deamination rate is calculated only once, when launching "contDeam". The contamination estimate obtained by "contDeam" is subsequently used as contamination prior for the first iteration (see Figure 5.6).

Let R be the set of all fragments and $R_j \in R$ be a particular aligned fragment of length l. The probability of observing this particular alignment to the reference genome is computed given two models: i) the null model where any divergence to the reference can be solely explained by sequencing error or ii) the deaminated model: where deamination and sequencing errors could have given rise to this particular alignment to the reference. For fragment R_j, let $\{r_1, ..., r_l\}$ be the individuals nucleotides and their respective error probabilities $\{\epsilon_1, ..., \epsilon_l\}$, both of which are provided by the basecaller. Let E denote the event that a sequencing error has occurred, D the event that deamination has occurred and let $\prime$ denote the complement of event (i.e. event has not occurred).

The likelihood of observing the base $r_i \in R_j$, aligned to the reference nucleotide n, is computed by assuming that nucleotide n was the endogenous template. The likelihood of observing r_i under the null model, denoted $P_n(r_i)$, is computed by taking into account two events, either a sequencing error has occurred or it has not:

$$P_n(r_i) = \begin{cases} (1 - \epsilon_i) \cdot P(n \to r_i | E\prime) & \text{if } n = r_i \\ \epsilon_i \cdot P(n \to r_i | E) & \text{if } n \neq r_i \end{cases} \tag{5.1}$$

where $P(n \to r_i | E\prime)$ is the probability that r_i is observed if n was the template. This quantity is 1 as both nucleotides are identical. The other term, $P(n \to r_i | E)$, is the probability of a substitution from nucleotide n to r_i given sequencing error. This term is approximately equal to $\frac{1}{3}$ but empirical substitution rates are used (see next subsection

on page 107 for details). Under the deaminated model, the probability of seeing base r_i (given the template n) denoted $P_d(r_i)$ is:

$$P_d(r_i) = \begin{cases} (1-\epsilon_i) \cdot P(n \rightarrow r_i | D\prime \cap E\prime) & \text{if } n = r_i \\ (1-\epsilon_i) \cdot P(n \rightarrow r_i | D) & \\ \quad + & \text{if } n \neq r_i \\ \epsilon_i \cdot P(n \rightarrow r_i | E) & \end{cases} \tag{5.2}$$

as three events need to be taken into account: i) $D\prime \cap E\prime$: absence of both sequencing error and deamination (if $n = r_i$) ii) D: either deamination or iii) E: error occurred and $n \neq r_i$. The probability of observing the data given that both deamination and a sequencing error have occurred ($D \cap E$) is currently ignored as it is very unlikely compared to the scenarios mentioned above. The probability of observing a substitution $n \rightarrow r_i$ given deamination ($P(n \rightarrow r_i|D)$) is computed using the endogenous deamination rates that were described earlier. The term $P(n \rightarrow r_i|D\prime \cap E\prime)$ is the probability that base r_i remains as is. This probability is obtained by subtracting from 1, the deamination probability of the remaining bases. For instance, if a given base has a deamination rate of 0.3, the probability that the base remains as is, given the absence of sequencing error, is 0.7.

Let C be the event that the fragment R_j was sampled from a contaminant mitochondrial genome and $C\prime$ be the event that it was sampled from the endogenous genome. A likelihood ratio of the null and deaminated models is computed, which is used to quantify $P(R_j|C\prime)$: the probability that fragment R_j was sampled from the endogenous mitochondrial genome. Assuming that every nucleotide r_i represents an independent observation, this likelihood ratio becomes:

$$P(R_j|C\prime) = \frac{\prod_{r_i \in R_j} P_d(r_i)}{\prod_{r_i \in R_j} P_n(r_i) + \prod_{r_i \in R_j} P_d(r_i)} \tag{5.3}$$

and $P(R_j|C)$ is simply $1 - P(R_j|C\prime)$. The overall probability of observing fragment R_j given its alignment to the reference is computed here. There are two events that could have occurred, either the fragment was sampled from the contaminant with probability denoted $c_{r_{deam}}$ or it was sampled from the endogenous genome with probability $1 - c_{r_{deam}}$. Using equation 5.3, the probability of observing R_j for a given contamination rate $c_{r_{deam}}$, denoted $P_{cont_{deam}}(R_j|c_{r_{deam}})$, is obtained by computing the following expression:

$$(1 - c_{r_{deam}}) \cdot P(R_j|C\prime) + c_{r_{deam}} \cdot P(R_j|C) \tag{5.4}$$

The probability of observing all the fragments in set R, assuming the reference as the template and the endogenous deamination rates that were initially computed, for a given contamination rate $c_{r_{deam}}$ is given by assuming that each fragment is an independent observation:

$$P_{cont_{deam}}(R|c_{r_{deam}}) = \prod_{R_i \in R} P_{cont_{deam}}(R_i|c_{r_{deam}}) \tag{5.5}$$

Finally, the posterior probability of the contamination rate is given by omitting the probability term for the data (P(R)), as it is independent of the contamination rate, and using a uniform prior for the contamination rate:

$$P_{cont_{deam}}(c_{r_{deam}}|R) \propto P_{cont_{deam}}(R|c_{r_{deam}}) \tag{5.6}$$

The contamination rate $\hat{c_{r_{deam}}}$ with the highest posterior probability is then produced:

$$\hat{c_{r_{deam}}} = \text{argmax}\, P_{cont_{deam}}(c_{r_{deam}}|R) \tag{5.7}$$

The overall algorithm can be described using the following pseudocode:

Data: Set of aDNA fragments R
Result: Most likely contamination rate $\hat{c_{r_{deam}}}$
Compute endogenous deamination rates ;
foreach $c_{r_{deam}} \in 0..1$ **do**
  Compute: Posterior probability $P_{cont_{deam}}(c_{r_{deam}}|R)$ given fixed endogenous deamination rates, using equation 5.6 ;
end
Find contamination rate $\hat{c_{r_{deam}}}$ with maximum $P_{cont_{deam}}(c_{r_{deam}}|R)$;

Algorithm 2: contDeam

One advantage of this approach is that it does not require the computation of the endogenous consensus. However, it also does not allow the user to identify the source of the contamination. Furthermore, it may underestimate contamination if the contaminant is deaminated (see Section 5.4.2.1.3 in the Results). The assumption that the mitochondrial genome reference sequence is the template does not seem to influence the final contamination estimate even for the highly divergent Denisovan mitochondrial genome (see Section 5.4.2.1 of the Results).

5.3.4 Mitochondrial consensus call

The first step of the iterative process is to call an initial consensus of the endogenous mitochondrial genome from mtDNA fragments aligned to a mitochondrial reference sequence ("endoCaller" in Figure 5.6).

The consensus call relies on computing the probability of observing the aligned aDNA data for a particular pair of endogenous and contaminant nucleotides at a specific site, given a fixed contamination prior and fixed deamination patterns. The endogenous consensus caller seeks to identify the pair of endogenous and contamination nucleotides with the highest posterior probability given the aligned aDNA fragments. Insertion/deletions at a given position are also considered. It is assumed that at any position there is a single nucleotide from the present-day human contaminant.

For a given position in the mitochondrial reference sequence, assuming a single contaminant, there are two bases to infer, b_e and b_c for the endogenous and contaminant

genome, respectively. Let R be the set of all fragments and $R_j \in R$ be a fragment of length l that overlaps the position. Let $\{r_1, ..., r_l\}$ be the individual nucleotides of the fragment R_j, as identified by the basecaller. The respective error probabilities $\{\epsilon_1, ..., \epsilon_l\}$ for each base are also provided by the basecaller.

For the position to be evaluated, let the nucleotide r_i be the base of fragment R_j that aligns at that specific position. Let ϵ_i be its error probability. It is assumed that the *a priori* probability that fragment R_j is endogenous is denoted by $P_{endo}(R_j)$. This quantity is computed by using both deamination patterns of the fragment and its length to derive a probability of that fragment being endogenous. The equations for this expression are described in greater detail at the end of this section.

In order to have observed the base r_i, there are two possibilities: the base came either from the contaminant with probability $1 - P_{endo}(R_j)$ or from the endogenous sample with probability $P_{endo}(R_j)$. It is assumed for now that the fragment was properly mapped, the final equation which considers either possibilities is presented on page 107 under the heading "Mapping". The probability of observing base r_i denoted by $P_{map}(r_i|b_e, b_c)$ is given by:

$$P_{endo}(R_j) \cdot P_e(r_i|b_e) + (1 - P_{endo}(R_j)) \cdot P_c(r_i|b_c) \tag{5.8}$$

The expression $P_e(r_i|b_e)$ is the probability of observing r_i given that the fragment is endogenous and b_e is the endogenous base. Let E denote the event that a sequencing error has occurred and let $E\prime$ denote the complement of the event or, in other words, that the sequencing was correct and no error has occurred. The quantity $P_e(r_i|b_e)$ is given by:

$$(1 - \epsilon_i) \cdot P_e(b_e \to r_i|E\prime) + \epsilon_i \cdot P_e(b_e \to r_i|E) \tag{5.9}$$

Given that the base is correct (i.e. without sequencing error) both r_i and b_e should be identical hence:

$$P_e(b_e \to r_i|E\prime) = \begin{cases} 1 & \text{if } b_e = r_i \\ 0 & \text{if } b_e \neq r_i \end{cases} \tag{5.10}$$

However, due to deamination, it is possible to have a substitution with the probability derived from the deamination profile entered as input. Let Ω be the set of all DNA bases ($\Omega = \{A, C, G, T\}$). Under the deamination model, the term $P_e(b_e \to r_i|E\prime)$ becomes:

$$\begin{cases} 1 - \sum\limits_{b'_e \in \Omega \setminus b} rate_{deam}(b_e \to b'_e) & \text{if } b_e = r_i \\ rate_{deam}(b_e \to r_i) & \text{if } b_e \neq r_i \end{cases} \tag{5.11}$$

where $rate_{deam}(b \to r_i)$ is the rate of nucleotide substitution from b to r_i due to deamination at that specific position of the fragment. As stated before, the deamination rates per base for each position of the fragment are entered as input and remain unchanged by "endoCaller". For sequencing errors, the probability of base substitution can be obtained using the assumption that any given nucleotide is equally likely to be miscalled as any of the remaining 3 nucleotides:

$$P_e(b_e \to r_i|E) = \frac{1}{3} \quad \forall b_e \neq r_i \tag{5.12}$$

However, studies on Illumina sequencing errors show that this assumption is often incorrect [81]. Using empirical nucleotide substitutions rates from an Illumina sequencing run (provided with the software package) is therefore recommended. The new error probability term becomes:

$$P_e(b_e \to r_i|E) = \frac{\#b_e \to r_i}{\sum\limits_{b'_e \in \Omega \setminus b_e} \#b_e \to b'_e} \tag{5.13}$$

where $\#x \to y$ represents the number of times a mismatch between the reference base x to an observed y occurred. These counts were determined using spiked-in control sequences aligned to the PhiX genome provided by Illumina Corp.

A similar computation is derived for the probability seeing r_i given that it was sampled the contaminant base b_c ($P_c(r_i|b_c)$). However, the deamination profile provided as input for the contaminant fragments are different from the endogenous ones and tend to be much lower (see end of Methods Section describing the test data for empirical deamination rates for both endogenous and contaminant fragments). The mitochondrial consensus caller "endoCaller" allows for deamination of the contaminant unlike "contDeam", which assumes that the contaminant fragments have little to no deamination.

5.3.4.1 Mapping

Thus far, it was assumed that the fragment R_j was correctly mapped. For fragments not properly mapped, it is assumed the probability of seeing the base r_i is independent of bases b_e and b_c and is simply the probability of observing r_i:

$$P_{mismap}(r_i|b) = P(r_i) = \frac{1}{4} \tag{5.14}$$

The probability of fragment R_j being incorrectly mapped is obtained using its mapping quality, and equations 5.8 and 5.14 are therefore combined into one to compute the final probability of observing the base r_i, denoted by $P(r_i|b_e, b_c)$:

$$(1 - m_{R_j}) \cdot P_{map}(r_i|b_e, b_c) + m_{R_j} \cdot P_{mismap}(r_i|b_e, b_c) \tag{5.15}$$

where m_{R_j} is the probability that the fragment R_j is mismapped.

5.3.4.2 Producing the most likely bases

The probability of observing the data given every endogenous and contaminant base has been described however, the posterior probability of the pair of bases given the data R is the quantity that is sought. It is assumed that every fragment R_j represents an independent observation and the likelihood of bases b_e and b_c given the data is considered

to be proportional to the probability of observing the data given the pair of nucleotides times a flat prior:

$$P(b_e, b_c|R) \propto \prod_{R_j \in R} P(R_j|b_e, b_c) \cdot \frac{1}{4^2} \tag{5.16}$$

Once the posterior probability for all pairs of nucleotides is computed, a sum of all the probabilities is performed to compute the likelihood of a given endogenous base:

$$P(b_e|R) = \sum_{b_c \in \Omega} P(b_e, b_c|R) \tag{5.17}$$

A marginalization over the endogenous base is used to call the contaminant base. Finally, the most likely endogenous nucleotide $\hat{b_e}$ is produced:

$$\hat{b_e} = \underset{b_e \in \Omega}{\operatorname{argmax}} P(b_e|R) \tag{5.18}$$

The probability of error on $\hat{b_e}$ is given by the ratio of the sum of the probabilities for all alternative bases except the most likely over the sum of the probabilities for all bases:

$$P(\neg\hat{b_e}|R) = \frac{\sum_{b_e \in \Omega \setminus \hat{b_e}} P(b_e|R)}{\sum_{b_e \in \Omega} P(b_e|R)} \tag{5.19}$$

An analogous computation is done to determine the contaminant base. The computation for insertions and deletions is similar (see Section 5.3.4.4 in the Methods).

5.3.4.3 Computation of $P_{endo}(R_j)$

For the probability that a given fragment R_j is endogenous, denoted as $P_{endo}(R_j)$, the model takes into consideration two factors: deamination patterns and the length of the fragments. Parameters for these two factors are introduced as input to the endogenous caller. Such parameters are re-estimated at each iteration using fragments that support an endogenous base versus a contaminant one ("splitEndo" in Figure 5.6). The "splitEndo" module will i) use the output of "endoCaller" from the previous iteration and separate fragments that support the endogenous or the contaminant base at positions where they differ ii) estimate deamination parameters and fit a log-normal distribution on each separated set of fragments independently.

Deamination rates are obtained by measuring rates of nucleotide substitution from the reference base at a given position in the fragment and the log-normal parameters are obtained by a maximum-likelihood fit using the fitdistrplus R package, similarly to Chapter 3. These estimates are fixed throughout a single iteration and get re-estimated by "splitEndo" in the following one.

Endogenous fragments tend to exhibit higher rates of deamination than contaminant fragments (see Section 5.3.9 in the Methods). In the previous section where "contDeam"

was described, likelihood ratios are computed, between a model which considers deamination and sequencing errors, and another model which solely uses sequencing errors to compute the probability of seeing a particular alignment given the reference as template. In this section, the possibility that the template might be a different base than the endogenous one is incorporated for greater accuracy. Let E denote the event that a sequencing error has occurred, D the event that deamination has occurred and let $\prime$ denote the complement of event (i.e. event has not occurred). First, the goal is to compute the probability of observing the base r_i, part of the fragment R_j, given that it originated from endogenous base b_e under a model where substitutions are solely due to sequencing errors. This term, denoted $P_n(r_i)$, is obtained similarly to equation 5.1 but by considering all 4 potential endogenous bases b_e as follows:

$$\sum_{b_e \in \Omega} (1 - P(\neg b_e)) \cdot P_n(r_i|b_e) \tag{5.20}$$

where $P_n(r_i|b_e)$ is equal to:

$$\begin{cases} (1-\epsilon_i) \cdot P(b_e \to r_i|E\prime) & \text{if } b_e = r_i \\ \epsilon_i \cdot P(b_e \to r_i|E) & \text{if } b_e \neq r_i \end{cases} \tag{5.21}$$

where $P(\neg b_e|R)$ is the probability of error for endogenous base b_e as defined in equation 5.19. The nucleotide substitution probabilities given either absence or presence of a sequencing error are computed as described in the "contDeam" section. Second, the probability of seeing base r_i given endogenous base b_e if any divergence is explained by either deamination or sequencing errors is computed. Similarly to equation 5.2, this probability denoted $P_d(r_i)$ is computed using this expression:

$$\sum_{b_e \in \Omega} (1 - P(\neg b_e)) \cdot P_d(r_i|b_e) \tag{5.22}$$

where $P_d(r_i|b_e)$ is equal to:

$$\begin{cases} (1-\epsilon_i) \cdot P(b_e \to r_i|D\prime \cap E\prime) & \text{if } b_e = r_i \\ (1-\epsilon_i) \cdot P(b_e \to r_i|D) & \\ \quad + & \text{if } b_e \neq r_i \\ \epsilon_i \cdot P(b_e \to r_i|E) & \end{cases} \tag{5.23}$$

Again, the substitution probabilities given either deamination or sequencing error are computed as described in the "contDeam" section.

The probability that the aligned fragment R_j was observed under a deamination and sequencing error model is computed, denoted $P(R_j \in deam)$, by taking the product for each base $r_1 ... r_l \in R_j$ of the term described by equation 5.22. The probability that aligned fragment R_j was observed under a sequencing error model, denoted $P(R_j \in null)$ uses the product of the term described by equation 5.20 where only sequencing errors are considered.

Finally, both probabilities are combined with the prior on a fragment being endogenous of $1 - c_{prior}$ as a likelihood ratio to obtain the probability, denoted $P_{deam}(R_j)$, that fragment R_j is deaminated:

$$\frac{(1 - c_{prior}) \cdot P(R_j \in deam)}{(1 - c_{prior}) \cdot P(R_j \in deam) + c_{prior} \cdot P(R_j \in null)} \tag{5.24}$$

Differences in fragment lengths between the endogenous and contaminant sequences can also be informative about contamination. Ancient fragments tend to shorter than modern contaminating DNA fragments due to degradation of ancient DNA [39, 42, 59, 90] (see Section 5.3.7 in the Methods). In Chapter 3, the distribution of the length of aDNA fragments was modeled using a single log-normal distribution (see also [96]). Here, the endogenous and contaminant fragment length distributions are modeled using two log-normal distributions and, using empirical distributions, four parameters are inferred: $\mu_{endo}, \sigma_{endo}, \mu_{cont}$ and σ_{cont} for the location and scale parameters of the endogenous and contaminant log-normal distributions, respectively. Again, these parameters are estimated by "splitEndo" at each iteration. The probability that the fragment R_j of length l pertains to the endogenous distribution is given by the probability density function for the log-normal distribution:

$$P(R_j \in endo_{dist}) = \frac{1}{l\sqrt{2\pi}\sigma_{endo}} e^{-\frac{(ln(l)-\mu_{endo})^2}{2\sigma_{endo}^2}} \tag{5.25}$$

The probability that the fragment is from the contaminant distribution ($P(R_j \in cont_{dist})$) is calculated the same way except using the location and scale for that distribution. The likelihood ratio of both terms is used to compute the probability, $P_{endo_{dist}}(R_j)$, that fragment R_j pertains to the endogenous distribution using the contamination prior:

$$\frac{(1 - c_{prior}) \cdot P(R_j \in endo_{dist})}{(1 - c_{prior}) \cdot P(R_j \in endo_{dist}) + c_{prior} \cdot P(R_j \in cont_{dist})} \tag{5.26}$$

Finally, the deamination and length probabilities are combined to compute the probability that a fragment is endogenous ($P_{endo}(R_j)$). The overall algorithm can be described using the following pseudocode (insertions and deletions are not represented for simplicity):

Data: Set of aDNA fragments R, deamination rates, endogenous/contaminant fragment length distribution, contamination prior
Result: Most likely endogenous b_e and contaminant b_c bases
foreach *position on the mitochondrial genome reference* **do**
  foreach *possible bases* $b_e, b_c \in \Omega^2$ **do**
    foreach *fragment* $R_j \in R$ **do**
      Find: base r_i from fragment R_j for that position ;
      Compute: probability of seeing r_i given b_e and b_c using equation 5.15 ;
    end
  end
  Produce : for that position, bases b_e and b_c with the highest posterior probability ;
end

Algorithm 3: endoCaller

5.3.4.4 Identifying endogenous insertions and deletions

For indels, two separate cases are considered:

- A deletion in the sample (which could also be an insertion in the reference)
- An insertion in the sample (which could also be a deletion in the reference)

Each case is described separately in the sections below. In both cases, it is not known *a priori* without using phylogenetic information in which lineage the indel occurred. The error rate of indels is considered to be a constant ϵ_{indel} for both cases. This constant is defined from the literature on sequencer-specific error rates [79]. Given than an indel was present in the fragment, it is considered to be present in the original fragment with probability $1 - \epsilon_{indel}$ and absent with probability ϵ_{indel}. As in the inference of a single nucleotide, the computation is different depending on whether a single contaminant is assumed or multiple ones.

5.3.4.4.1 Deletions

A deletion refers to missing nucleotides with respect to the reference in either the contaminant or the endogenous genome. This could be due to a deletion in the lineage leading to the sample or an insertion in the one leading to the reference.

Given that a deletion is observed, four different scenarios need to be considered:

- Both endogenous and contaminant genomes have the deletion

- Only the endogenous genome has the deletion
- Only the contaminant genome has the deletion
- Neither the contaminant nor the endogenous genome have the deletion and the observation was due to a sequencing error

The observation of a fragment with or without a deletion changes the likelihood for each possibility. For instance, for the first case, the observation of a fragment R_j with the deletion gives the following term in the product:

$$(1 - m_{R_J})[P_{endo}(R_J) \cdot (1 - \epsilon_{indel}) + (1 - P_{endo}(R_J)) \cdot (1 - \epsilon_{indel})] \tag{5.27}$$

where m_{R_J} is the probability that fragment R_j is mismapped and where $P_{endo}(R_J)$ is the probability that fragment R_j is endogenous (defined at the end of section 5.3.4, on page 110). As both genomes contain the deletion, the probability of observing the fragment R_j is the probability of having correctly detected the deletion in either of the two cases. If the fragment does not have the deletion, still under the assumption that both endogenous and contaminant genomes have the deletion, the term becomes the probability that either one contains an error:

$$(1 - m_{R_J})[P_{endo}(R_J) \cdot \epsilon_{indel} + (1 - P_{endo}(R_J)) \cdot \epsilon_{indel}] \tag{5.28}$$

as the fragments falsely called it in both cases. A similar computation is done for the remaining three possibilities but where the indel error term is used differently depending on which genome is believed to have the deletion. Finally, the possibility with the maximum posterior probability is used to produce both the endogenous and contaminant genomes. The error probability on that call is computed by the ratio of the sum of the probabilities for all three least likely scenarios over the sum of all probabilities.

5.3.4.4.2 Insertion

Insertions are produced in a manner similar to deletions. For the deletion case, the likelihood of a nucleotide being present or absent at a specific position is considered. In the case of insertions, the possibility of having various nucleotides being inserted at a specific position is considered. The likelihood for each putative inserted sequence and the absence of an insertion is calculated.

A bi-dimensional matrix is used for all possible insertions for both the endogenous and contaminant genomes. Each cell represents a specific model where either genomes could have a given insertion. The likelihood is computed using a product over all fragments using terms analogous to expressions 5.27 and 5.28 depending on which of the two genomes

has the insertion for that given model. Finally, the most likely model is retained. For calling the endogenous consensus genome, the error probability is marginalized over each possible contaminant insertion and vice-versa for the contaminant consensus calling.

5.3.4.5 Endogenous consensus calling with multiple contaminants

Multiple contaminants with equal contributions represent a more complex problem for consensus calling, compared to a single one (see Figure 5.7). Results show that schmutzi yields good results (a reliable consensus endogenous genome) at low contamination rates but not at higher ones (see Results section).

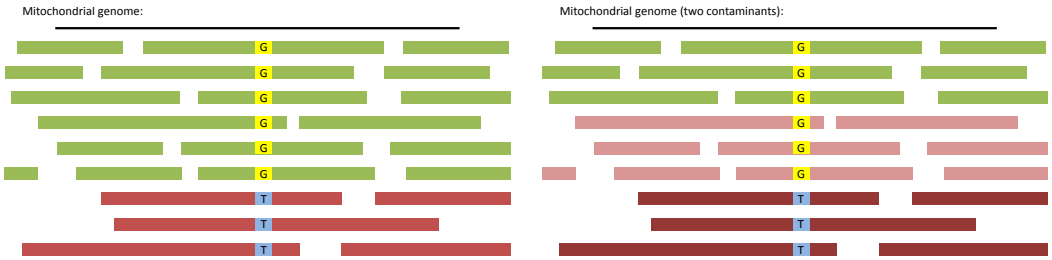


Figure 5.7: Schema of alignments to the mitochondrial genome where green lines represent endogenous fragments and red lines, the contamination. However, depending on whether there is a single source of mitochondrial contamination (left) or multiple ones (right), the distribution of the bases at a segregating site can change. Given that the contamination rate is $\frac{1}{3}$ for the single contaminant scenario, inferring the endogenous and contaminant bases is straightforward, as the relative number of each base follows the expected distribution. However, in the figure to the right, knowing that the contamination rate is $\frac{2}{3}$ does not translate into observing this fraction of a particular contaminant base.

5.3.4.5.1 Calling the endogenous nucleotide

Given that the fragment R_j was correctly mapped, it originated either from the endogenous or the contaminant genome. Let $P_{endo}(R_j)$ be the probability that the fragment came from the endogenous genome. The probability that the fragment stemmed from the contaminant genome is simply $1 - P_{endo}(R_j)$.

The probability of observing r_i on fragment R_j given that b_e is the putative endogenous base is computed. In the case where the fragment R_j is a contaminant fragment, no information can be obtained on the probability of observing base r_i given b_e hence the uniform prior for nucleotides is used:

$$P_{map}(r_i|b_e) = P_{endo}(R_j) \cdot P_e(r_i|b_e) + (1 - P_{endo}(R_j)) \cdot \frac{1}{4} \tag{5.29}$$

This term is similar to the way $P_{map}(r_i|b_e, b_c)$ is computed for the single contaminant case with the exception of the lack of a single contaminant base b_c. The remaining computations are identical to the case with a single contaminant.

5.3.4.5.2 Insertions

The likelihood of all observed insertions at a given position is computed, assuming that unobserved insertions have a negligible likelihood. The likelihood for not having an insertion is also considered. For a given insertion, if it is observed in a fragment, the term in the product becomes the following expression:

$$(1 - m_{R_J}) \cdot P_{endo}(R_J) \cdot (1 - \epsilon_{indel}) \tag{5.30}$$

where m_{R_J} is the probability of mismapping for that given fragment and $P_{endo}(R_J)$ is the probability that R_j is endogenous. However, for the remaining insertions, the following term is used:

$$(1 - m_{R_J}) \cdot P_{endo}(R_J) \cdot \epsilon_{indel} \tag{5.31}$$

The most likely insertion is produced and the error probability is defined as the ratio of the sum of the probabilities for possible insertions minus the most likely over the sum of all probabilities.

5.3.4.5.3 Deletion

The likelihood of two scenarios is considered: either the endogenous genome has a deletion or it does not. Again, using the assumption of independence of observation for each fragment, the likelihood for each fragment is multiplied independently for each of these two possibilities. For the former, where the endogenous genome has a deletion, for each fragment R_j with a deletion, the term in the product becomes the term defined in equation 5.30. For the second scenario, where the endogenous does not have the deletion and the fragment R_j has the deletion, the expression used is defined by equation 5.31. If fragment R_j does not have a deletion, the two previously defined terms are swapped for one another in the products. Finally, a deletion in the endogenous consensus is produced if the likelihood of such an event exceeds the likelihood of not having a deletion. The error probability is computed by taking the ratio of the second most likely scenario over the sum of the probabilities for both possibilities.

5.3.5 Mitochondrial contamination estimate

Once the endogenous base and its likelihood have been computed for a given site, a second program takes this information, together with the aligned BAM file of all fragments covering each site, and determines the most likely contaminating genome from the database of possible contaminants as well as the contamination rate ("mtCont" in Figure 5.6). This is achieved by determining the most likely contamination rate using sites where bases in the putative endogenous and contaminant genomes differ. Once this computation is finished for all mitochondrial genomes in the database, the genome with the highest likelihood of being the contaminant is identified (see details about the database of contaminants in section 5.3.8 in the Methods).

In the previous section, a fixed contamination prior was supplied to "endoCaller" and the most likely endogenous and contaminant base were inferred given the data. In this section, "mtCont" computes the most likely contamination rate given the data for fixed probabilities for the endogenous and contamination bases which are provided by "endoCaller". As in "endoCaller", the deamination rates are entered as input. The contamination estimate generated by "contDeam" at iteration #1 is re-calculated by "mtCont" in subsequent iterations (see Figure 5.6).

For a given position on the mitochondrion, let b_e be a possible base from the endogenous sample and c be a potential base from the contaminant. Let the contamination rate be c_r, defined as the probability of seeing a base from the contaminant at this given position. Therefore, the probability that the base is endogenous is $1 - c_r$. Similar to the terms used in the section above, let R_j be a fragment with mismapping probability m_{R_j} and let base r_i be its base at the position of interest. The probability of observing r_i given that either b_e or c could have given rise to it, denoted $P(r_i|b_e, c)$, is:

$$(1 - m_{R_j}) \cdot P_{map}(r_i|b_e, c) + m_{R_j} \cdot P_{mismap}(r_i|b_e, c) \tag{5.32}$$

where the probability of being mismapped is defined as in equation 5.14. If the fragment is properly mapped, it can either originate from the contaminant or the endogenous genome. By using the defined contamination rate, $P_{mapped}(r_i|b_e, c)$, the probability of observing r_i given that the fragment was correctly mapped, is quantified as:

$$(1 - c_r) \cdot P_e(r_i|b_e) + c_r \cdot P_c(r_i|c) \tag{5.33}$$

since the fragment was either sampled from the contaminant with probability c_r or from the endogenous base with probability $1 - c_r$. The probability of observing the base r_i given it came from either the endogenous material ($P_e(r_i|b_e, c)$) or the contamination ($P_c(r_i|b_e, c)$) considers sequencing errors and deamination rates. The precise terms for such quantities are derived as in equation 5.9. The only difference is that a deaminated substitution model is used for the endogenous base, but any different base for the contamination is not due to deamination but to sequencing errors.

Let Ω^2 be the set of all possible pairs of nucleotides. For a given contamination rate c_r, the probability ($P_{c_r}(r_i)$) of observing the base r_i is obtained by marginalizing over each possible contaminant and endogenous base:

$$\sum_{b_e,c\in\Omega^2} P(r_i|b_e,c)P(b_e,c) \tag{5.34}$$

where the term $P(r_i|b_e,c)$ is defined in equation 5.32. The combined probability of b_e being the endogenous and c being the contaminant base is given by: $P(b_e,c) = P(b_e)\cdot P(c)$. The prior on the endogenous base $P(b_e)$ is one minus the probability that b_e is not the endogenous base, a quantity defined by equation 5.19. The probability $P(c)$ is defined by the probability of having nucleotide c in the putative contaminant mitochondrion.

The total likelihood is obtained by the product of equation 5.34 for every fragment. This likelihood is computed for every contamination rate between 0% and 100% assuming a uniform prior on the contamination rate and, for each mitochondrial genome in the set of putative contaminants. Finally, the contaminant genome is determined and the contamination rate with the highest posterior probability, as well as a 95% confidence interval, is produced. The overall algorithm can be described using the following pseudocode:

Data: Set of aDNA fragments R, deamination rates, likelihood for endogenous b_e bases for the entire mtDNA genome, DB of human mitochondrial genomes
Result: Most likely contamination rate $\hat{c_r}$ and contaminant source
foreach *potential contaminant in the DB* **do**
  foreach *contamination rate* $c_r \in 0..1$ **do**
    foreach *position on the mitochondrial genome* **do**
      foreach *possible bases* $b_e, c \in \Omega^2$ **do**
        Find: prior $p(b_e)$ using the likelihood from the endogenous consensus ;
        Find: prior $p(c)$ using the current contaminant in the DB ;
        foreach *fragment* $R_j \in R$ **do**
          Find: base r_i from fragment R_j for that position ;
          Compute: probability of seeing r_i given b_e and c using equation 5.34 ;
        end
      end
    end
  end
  Keep : contamination rate with the highest posterior probability for this record in the DB ;
end
Return : $\hat{c_r}$ contamination rate with the highest posterior probability for all records in the DB ;

Algorithm 4: mtCont

5.3.6 Existing methods for mitochondrial contamination estimates

Although there have been descriptions of methods to estimate the contamination rate, there is currently no software implementation of an algorithm to estimate contamination for aDNA samples that is widely available for download. To provide a comparison to existing methods, the maximum-likelihood model described in [35, 36] was implemented and used on simulated datasets. The predicted contamination rate was compared to the simulated one.

Briefly, a rate of sequencing error denoted ϵ is estimated using monomorphic regions of a set of mitochondrial genomes. The fragments are aligned against the endogenous consensus call and a database of 311 potential contaminant mitochondria, as described in the original methodology. Since the simulations used a single contaminant, a single genome was used in the database. The method was run once using the closest mitochondrial genome in the database and once more using the same contaminant used in the simulations.

For a read R_i aligned to the endogenous genome, $M_{i,e}$ and $N_{i,e}$ were computed for the number of matches and mismatches respectively. The read is also re-aligned to the contaminant genome and the same analogous quantities, $M_{i,c}$ and $N_{i,c}$ are computed. Given the error rate ϵ, the number of matches and mismatches to the endogenous genome, the probability of observing R_i aligned to the endogenous mitochondrion was computed as:

$$\binom{M_{i,e} + N_{i,e}}{N_{i,e}} (1-\epsilon)^{M_{i,e}} (\epsilon)^{n_{i,e}} \tag{5.35}$$

In the original description, a vector of probabilities describes the probability that read R_j came from each possible contaminant genome in the database and the endogenous mitochondrial genome. This vector is used to compute the probability of observing read R_j. In the simulated datasets, as there are two genomes, this expression becomes:

$$(1-c) \cdot \binom{M_{i,e} + N_{i,e}}{N_{i,e}} (1-\epsilon)^{M_{i,e}} (\epsilon)^{n_{i,e}} + c \cdot \binom{M_{i,e} + N_{i,c}}{N_{i,c}} (1-\epsilon)^{M_{i,c}} (\epsilon)^{n_{i,c}} \tag{5.36}$$

where c is the predicted contamination rate. Finally, the most likely contamination rate given the data is produced by assuming that each fragment represents independent observations as described in [36]. As the method requires the endogenous consensus call, the mitochondrial genome produced by PMDtools and htslib was used as they represent the state of current methods. As the target contamination rate, the number of contaminant fragments over the total was used as the method operates on a per fragment basis rather than on per nucleotide basis.

5.3.7 Distribution of the endogenous and contaminant fragment size

It was previously suggested in the literature that endogenous and contaminant fragments might have different size distributions where the endogenous fragments are shorter than the contaminant fragments [42, 82]. To measure this, the fragments from the Sima de los Huesos hominin [76] that aligned to the mitochondrial genome were analyzed. As it was heavily contaminated, fragments could be separated into those supporting an endogenous or a contaminant base using diagnostic positions that supported an archaic hominin base or a present-day human one. The size distribution for both was plotted (see Figure 5.8).

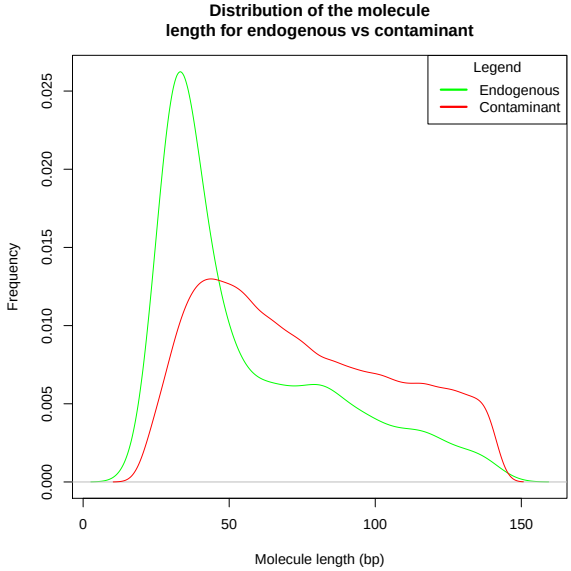


Figure 5.8: Size distribution of the endogenous (green) versus contaminant (red) fragments in the Sima de los Huesos sample.

5.3.8 Database of putative contaminants

To predict accurate rates of contamination, a database of human mitochondrial sequences that are representative of the natural diversity while restricting the total number of sequences due to the computational overhead was needed. As two nearly identical mitochondrial sequence will yield the same contamination rate, having the same sequence twice in the database will result in redundant computations. Every human mitochondrial sequence was downloaded from GenBank and a multiple sequence alignment was performed using mafft version 7.017b [50] due to its speed and multithreading options. All pairwise sequence distances were computed. The results were pruned according to a minimal pairwise edit distance as to have a non-redundant database of 197 records (see Table A.6, A.7 and A.8 in the Appendix).

Furthermore, the "–usepredC" option in the overall wrapper script allows the user to introduce the predicted contaminant as a database record. This option is recommended for cases where the contamination is very high thus allowing for adequate characterization of the contaminant mitochondrial genome, assuming that a single contamination source is responsible for most of the contaminating present-day human fragments. As this is not known in advance, it is recommended to run the wrapper once with this option and once without.

5.3.9 Empirical test data

Schmutzi was tested on simulated data and empirical ancient DNA datasets as well. Schmutzi's ability to call the endogenous mitochondrial genome and, assuming that the contamination stemmed from a single source, the accuracy in calling the contaminant's mitochondrial genome as well was tested. To show the program's robustness to contamination, highly contaminated samples were sought. Furthermore, whether the contamination estimate from schmutzi would be within the estimates obtained by simple contamination determination based on diagnostic positions was evaluated.

Schmutzi was tested on five heavily contaminated empirical aDNA datasets (see Table 5.3). The first three were subsets of the original empirical data that were rebasecalled using freeIbis (see Chapter 2) and processed using leeHom (see Chapter 3). The samples were from an Altai Neanderthal, a Denisovan and an early modern human from Ust'-ishim (see Table 5.3 for references and coverage). However, these samples are easy targets as they have low levels of present-day human contamination. Furthermore, they do not have a mitochondrial consensus sequence that was established on the same individual with a different library without any contamination to provide a standard for comparison. The other two datasets were from Mezmaiskaya samples with high amounts of present-day human contamination. These aDNA datasets (B9687 and B9688) described by [38] pertained to the same Mezmaiskaya Neanderthal individual described in [89]. The latter had the advantage of stemming from an extraction with low amounts of present-day human contamination (0.6%). Therefore, the endogenous consensus call should be identical to the Mezmaiskaya mitochondrial genome (GenBank: FM865411). The contaminating genome however, was not characterized. This thesis focuses on the analysis of these two aDNA datasets due to the availability of an independently inferred mitochondrial genome. High levels of present-day human contamination made them difficult datasets as well.

The total number of fragments and bases aligning to the mitochondrial reference was calculated (see Table 5.1). Using diagnostic positions for Neanderthal mitochondrial sequences, the number of contaminant and endogenous fragments was tallied (see Table 5.2). A contamination estimate could be computed by using the ratio of contaminant fragments over the sum of fragments that were flagged as either contaminant or endogenous. Furthermore, this estimate was recomputed by using the sum of the nucleotides instead of the number of fragments. This led to a different contamination estimate for the first sample as there is a difference in length between endogenous and contaminant fragments (see Section 5.3.7). Maximum-likelihood [29] phylogenetic inference was performed using phylip version 3.69 [30] with default parameters using the mitochondrial genomes enumerated in Table A.9 found in the Appendix. Multiple sequence alignments that were used as input were obtained from PRANK version 140603 [67].

sample ID	origin	total fragments	total bases	coverage
B9687	Mezmaiskaya	162,035	11,773,544	710.577
B9688	Mezmaiskaya	148,817	10,533,824	635.755

Table 5.1: Number of fragments, sum of all bases and coverage for the datasets from empirical samples

sample	endogenous		contaminant		contamination	
	fragments	base pairs	fragments	base pairs	rate per fragment	rate per base
B9687	30,876	2,443,418	23,598	1,989,785	0.433	0.449
B9688	25,437	1,971,127	24,083	1,972,954	0.486	0.500

Table 5.2: Tally of the fragments that support diagnostic positions in the archaic humans and *ad hoc* contamination estimate.

5.3.9.1 B9687

The details of the experimental procedures for the B9687 samples are found in [38]. Briefly, two extracts of the Mezmaiskaya 1 individual were prepared from 107 mg (extract ID: E734) and 90 mg (extract ID: E373) bone powder using the extraction protocol described in [98]. Sequencing libraries of the extracts were generated using single-stranded library preparation method [37] and double indexing was performed on the libraries [54]. All libraries were subsequently enriched for mitochondrial DNA using human mitochondrial DNA probes following the protocol detailed in [36].

For the B9687 sample, the coverage is the highest among the empirical samples at 710X (see Table 5.1). Aligned fragments were separated according to whether they stemmed from the endogenous (Neanderthal) or the contaminant (present-day human) mitochondrial genomes using 111 diagnostic positions (fixed sites between 7 Neanderthals and 21 present-day humans) on the mitochondrial reference. This separation into two sets was used to quantify contamination and yielded an estimate in the 43-45% range depending on the metric used (see Table 5.2). An analysis of the length distribution of the endogenous and contaminant fragments revealed an excess of fragments with approximately the same size as the sequencing read length (see Figure 5.9). After communication with the authors, this effect is unlikely to stem from library preparation but is more likely an artifact of the extraction procedure. Other libraries prepared using the same protocol does not show this enrichment of fragments with the same size as the length of sequencing. This entails that the use of length will not help the algorithm in gaining greater power to recognize the endogenous base. Deamination patterns were measured on both the fragments labeled as endogenous and those identified as contaminant (see Figure 5.10).

As the deamination rates of the endogenous fragments are several fold higher than the ones found for the contaminant ones, the algorithm can use this information to disentangle which base is likely to be endogenous and which is likely to be the contaminant one. Furthermore, the deamination rates for the contaminants are very low, enabling the possibility of getting an estimate of contamination based on deamination rates alone (see Section 5.3.3).

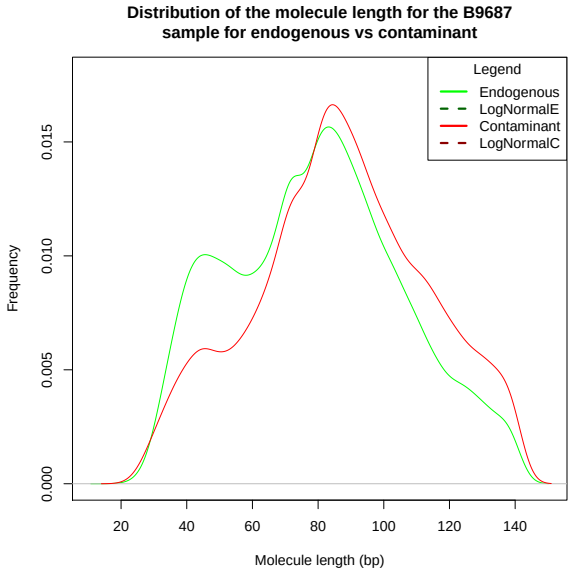


Figure 5.9: Density plot of the size of the fragments identified as endogenous (Neanderthal in green) and contaminant (present-day human in red) in the B9687 sample.

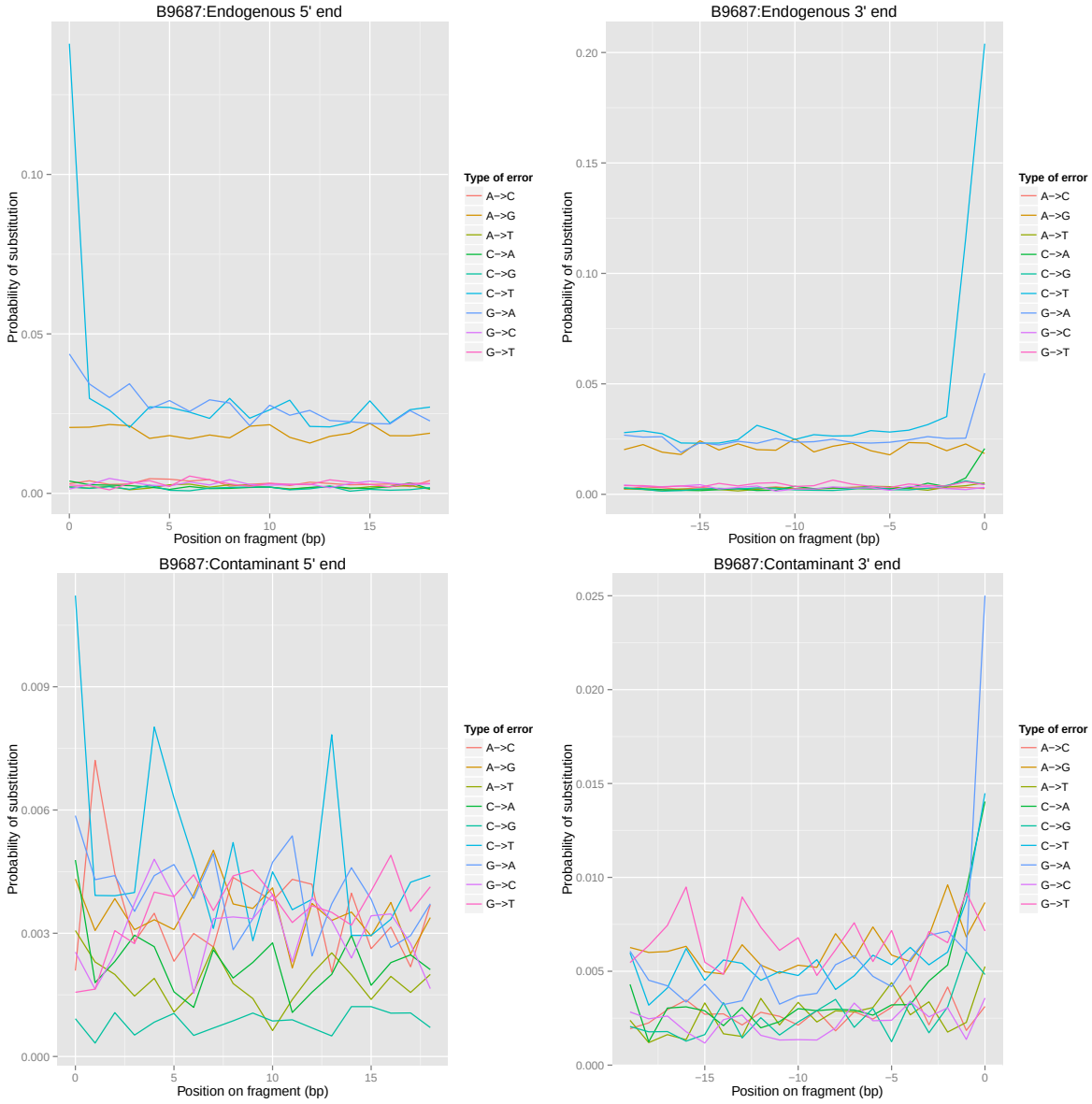


Figure 5.10: Deamination patterns for the fragments identified as endogenous and contaminant in the B9687 sample for the 5' end (left) and 3' end (right). The fragments were identified as either Neanderthal (top) or present-day human (bottom).

5.3.9.2 B9688

B9688 was the second sample described in [38]. It was sequenced in a similar way as B9687, however, coverage was slightly lower at 635X. Using the same diagnostic positions

as B9687, aligned fragments were split into two sets, those supporting a Neanderthal base and those supporting a present-day human one. Contamination estimates for this sample were between 48% and 50%, higher than the B9687 sample (see Table 5.2). A measure of fragment length revealed the same enrichment for fragments with the same length as the original read length previously seen in B9687 (see Figure 5.11). Contaminant fragments also showed very low rates of deamination (see Figure 5.12).

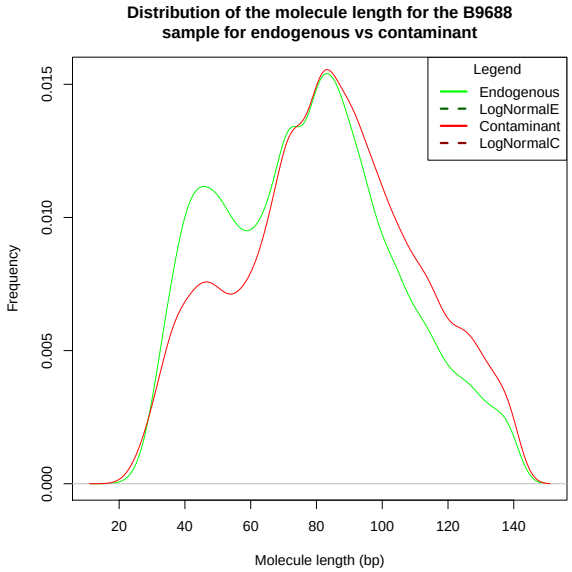


Figure 5.11: Density plot of the size of the fragments identified as endogenous (Neanderthal in green) and contaminant (present-day human in red) in the B9688 sample.

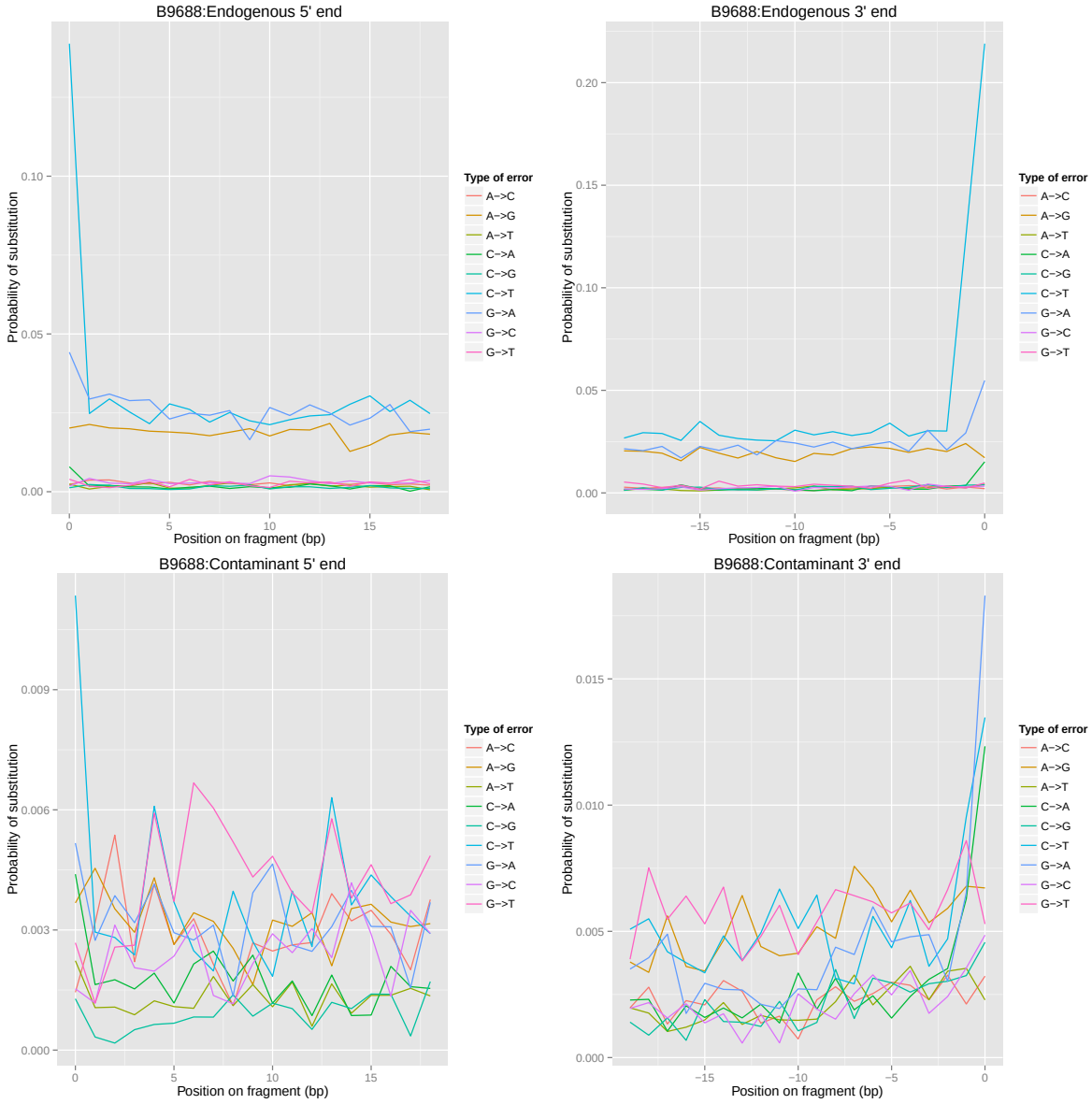


Figure 5.12: Deamination patterns for the fragments identified as endogenous and contaminant in the B9688 sample for the 5' end (left) and 3' end (right). The fragments were identified as either Neanderthal (top) or present-day human (bottom).

5.4 Results

This section describes the results obtained by schmutzi and existing approaches on simulated data (see Section 5.4.2 on page 134) and empirical data (see Section 5.4.1).

5.4.1 Empirical data

For the empirical data, contamination estimates computed on deamination patterns using "contDeam" is presented (see Section 5.4.1.1) followed by contamination estimates using divergence positions between the predicted endogenous genome and a database of putative contaminants using "mtCont" (see Section 5.4.1.2). This is followed by the accuracy of the endogenous consensus call (see Section 5.4.1.3) and the contaminant consensus call (see Section 5.4.1.4).

5.4.1.1 Contamination estimate based on deamination

For the Mezmaiskaya datasets, the maximum *a posteriori* estimates for contamination based on deamination alone were found at 51.0±0.5% and 44.5±0.5% for the B9687 and B9688 samples respectively. The posterior probability distribution was plotted for both samples (see Figure 5.13). In both cases, the true contamination rate is unknown but both estimates fall within a few percents of the ones presented in Table 5.2 that were measured using diagnostic positions, thus providing a reasonable initial estimate.

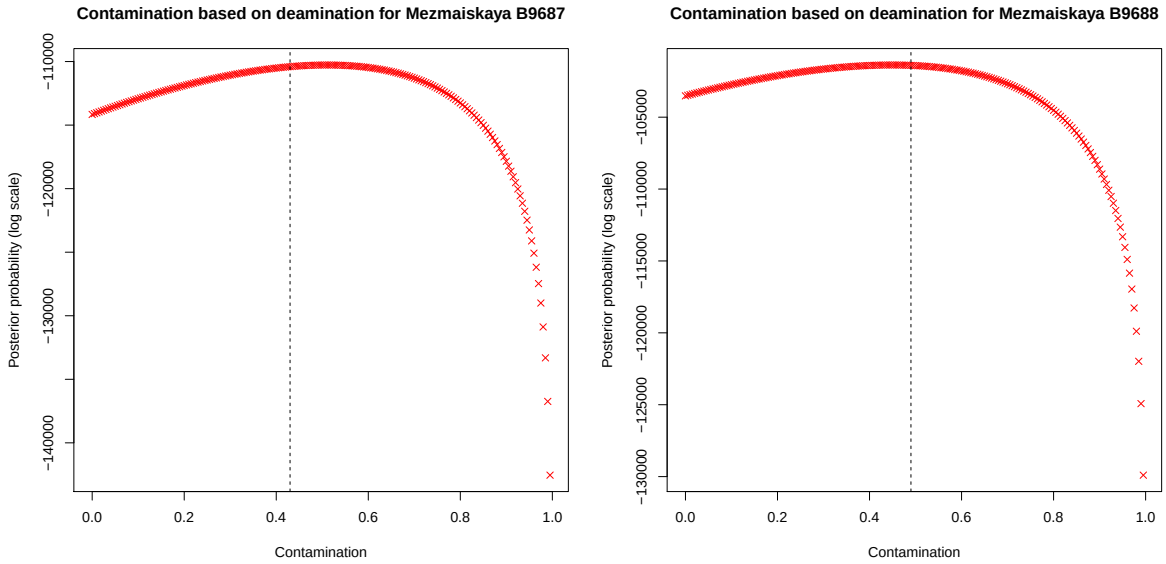


Figure 5.13: Distribution of the posterior probability for contamination rates as measured by endogenous deamination rates. For the two Mezmaiskaya samples B9687 (left) and B9688 (right), the fraction of contaminant fragments over the total sum is also represented (dotted line).

5.4.1.2 Contamination estimate based on divergent bases

Present-day human contamination was estimated for each of the five empirical datasets using schmutzi and contamMix (v1.0-10), an implementation from the authors of a previously described maximum-likelihood method for estimating mitochondrial contamination [35, 36] (see Section 5.3.6).

The correct contamination estimate was taken to be the one obtained from fragments aligned to sites in the reference mitochondrial genome where Neanderthals or Denisovans differ from 20 present-day humans ("diagnostic sites"). Since there are too few diagnostic sites, approach could not be used for the early modern human data.

For the Altai Neanderthal and Denisovan samples which have low contamination, both schmutzi and contamMix accurately estimate the contamination (see Figure 5.15A). However, for the highly contaminated Mezmaiskaya Neanderthal samples, schmutzi's contamination estimates are closer to the estimates provided using diagnostic positions (44.1±0.8 and 49.3±0.7 for Mezmaiskaya samples 1 and 2, respectively). For the Mezmaiskaya 1 for instance, using the 111 diagnostic sites, there were 2,443,418 individual bases supporting the Neanderthal base and 1,989,785 supporting the present-day human base, resulting in an estimated contamination of 44.9% (per nucleotide basis). The contamination estimates obtained using diagnostic positions are constant even when filtering for high base

quality and removing potentially deaminated bases. In comparison, the contamination estimate from schmutzi was 44±1% and the estimate from contamMix was 41.4±0.8%.

For both Mezmaiskaya datasets, which have higher contamination, a contamination rate of 43.0±1.0 and 48.0±1.0 was obtained using schmutzi without the inclusion of the predicted contaminant. In both cases, the contamination estimate increased by exactly 1% if the predicted contaminant was used in the database of contaminants (option "–usepredC", see Section 5.3.8). These estimates are closer to the expected ones presented in Table 5.2 and fall within the lower and upper bounds. The posterior probability distribution shows the peak estimate close to the one obtained using diagnostic positions (see Figure 5.14).

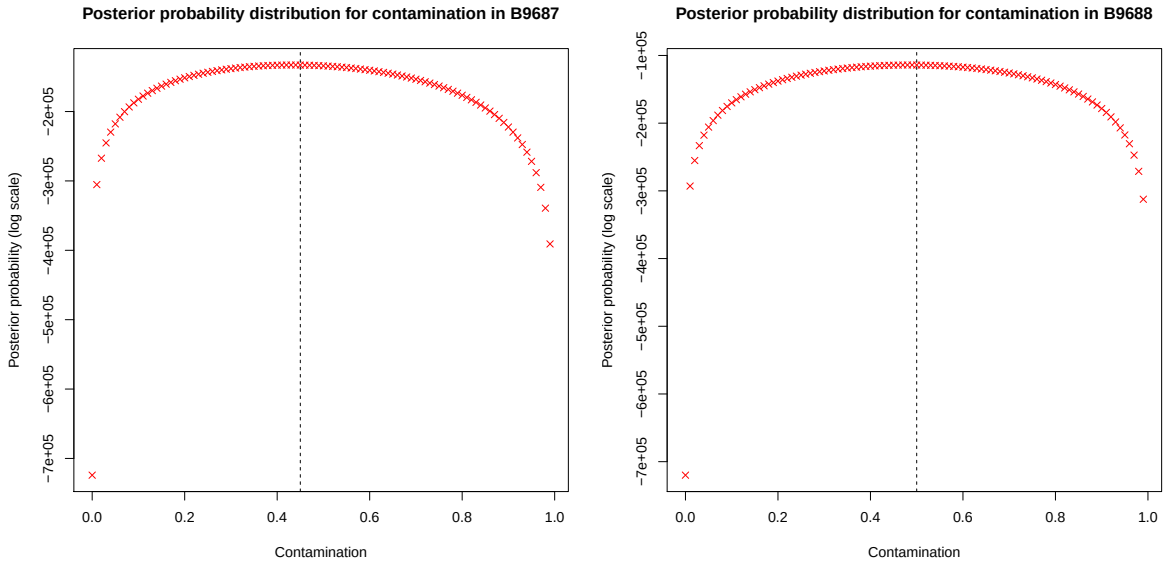


Figure 5.14: Distribution of the posterior probability for contamination as measured by the endogenous genome and the database of putative contaminants. For the two Mezmaiskaya samples B9687 (left) and B9688 (right), the fraction of contaminant bases over the total sum is also represented (dotted line).

5.4.1.3 Endogenous mitochondrion consensus call

Because not all features of empirical ancient DNA datasets can be accurately simulated, schmutzi was also tested on the 5 empirical datasets described in Table 5.3. Only a subset of the original data was used here. The accuracy of the endogenous consensus sequences called using schmutzi was compared to the published mitochondrial genomes and to the consensus sequence called using htslib. For htslib, the quality scores of potentially deaminated bases was reduced to avoid incorrect calls at deaminated sites, similarly to the procedure used in [60, 89].

sample ID	mtDNA coverage (X)	deamination rates (%)		present-day contamination	library ID and reference
		5'	3'		
Altai Neanderthal	1076	5.7	28.4	low (~ 1%)	L9198 from [89]
Denisovan	258	14.8	33.9	low (~ 1%)	B1108 from [77]
Ust'-ishim	124	2.7	3.4	low (~ 1%)	B3899 from [35]
Mezmaikaya Neanderthal B9687	711	8.8 (17.3)	13.3 (25.8)	high (~ 40-50%)	B9687 from [38]
Mezmaikaya Neanderthal B9688	636	8.5 (15.0)	12.7 (24.1)	high (~ 40-50%)	B9688 from [38]

Table 5.3: Empirical mitochondrial datasets. The number in parentheses represent the deamination rates when conditioning on the other end of the fragment being deaminated for heavily contaminated samples.

At contamination rates less than 5%, the consensus sequences called with htslib were highly similar (between 1 and 5 mismatches) to the published mitochondrial genome sequences (see Figure 5.15A). In all cases, schmutzi's prediction was identical to the published reference except for the Denisovan genome where there was an overprediction of one low-quality cytosine in a large 6 basepairs insert adjacent to the poly-cytosine stretch (position 5894-5899 on the rCRS).

However, at higher contamination rates (>40%), the consensus sequence becomes increasingly inaccurate when called with htslib. In contrast, the consensus sequence produced by schmutzi is robust to higher contamination (40-50%). For the highly contaminated Mezmaiskaya samples, the effect of using only deaminated fragments to generate the consensus using htslib was assessed. This approach has been previously used and substantially reduces the amount of contamination. Indeed, the consensus obtained using htslib and only deaminated fragments improves the accuracy of the consensus sequence (see Figure 5.15A). Despite this, the consensus sequence produced by schmutzi is still more accurate in all but one case which was influenced by capture bias (see paragraphs below and section 5.5).

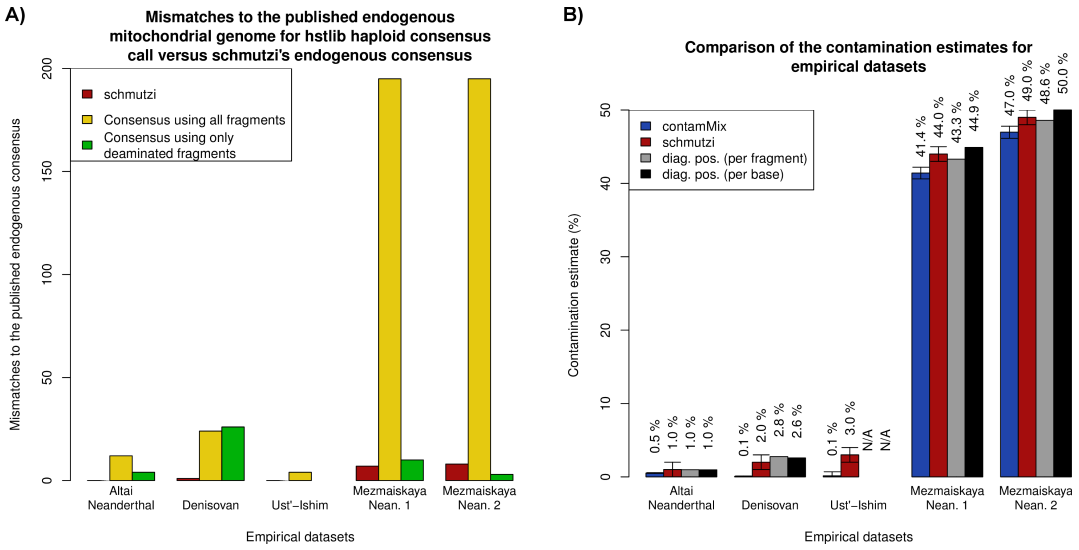


Figure 5.15: Consensus call and contamination estimate accuracy for empirical datasets. (A) The htslib consensus call (yellow) and the schmutzi consensus call (red) were performed on a subset of the data from 3 Neanderthals, 1 Denisovan and 1 early modern human (EMH). The number of mismatches between the mitochondrial consensus sequence and the published mitochondrial genome from the same individual was calculated. (B) Contamination was estimated using schmutzi (red) and contamMix v.1.0-10 (blue) and compared to the contamination computed using diagnostic positions (gray per fragment and black per base). For the two Mezmaiskaya individuals, the endogenous genome used for comparison was obtained using another library with low levels of contamination from the same individual.

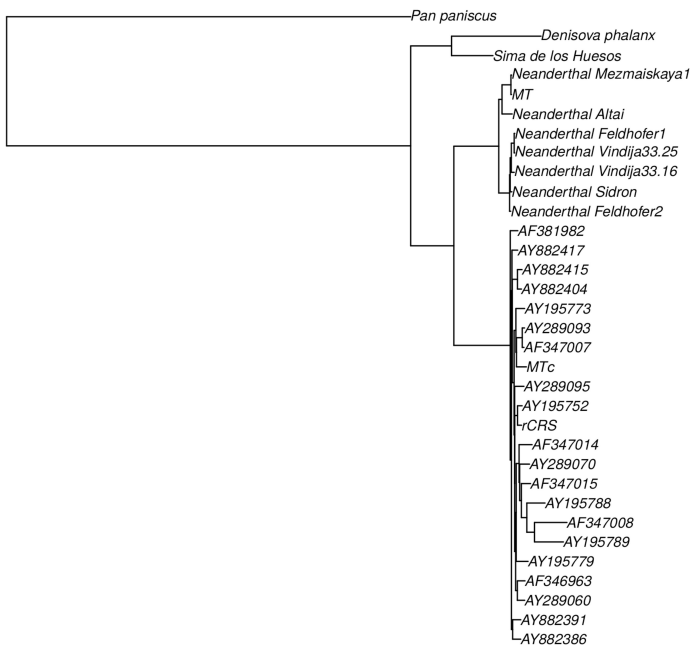


Figure 5.16: Phylogenetic placement of Mezmaiskaya 1 (library ID B9687) using a maximum-likelihood tree showing the placement of the mitochondrial genome of Mezmaiskaya 1 (labeled "MT" in the tree) and the inferred contaminant (labeled "MTc" in the tree), compared to 20 present-day humans and 9 archaic humans.

The sequence inferred for the mitochondrial genome of the Neanderthal from Mezmaiskaya 1 (library ID: B9687) which was generated from the same individual for which a high-quality mitochondrial genome from a library with low contamination is available (GenBank: FM865411) was examined in more detail. The contaminating mitochondrial sequence is not known.

To verify whether the inferred endogenous and contaminant genomes would respectively fall within the predicted archaic and present-day human clades, a maximum-likelihood tree was constructed using the mitochondrial genomes from 20 present-day humans and nine archaic hominins enumerated in Table A.9 in the Appendix (see Figure 5.16B for the tree obtained using the high-quality bases, $\geq$200 PHRED scale, for both the inferred endogenous and contaminant genomes for the Mezmaiskaya 1 and Figure B.16 in the Appendix for the remaining trees). The Mezmaiskaya B9687 and B9688 samples cluster with the Mezmaiskaya genome. The contaminant genomes all fall within human variation except the Mezmaiskaya B9687 without any quality filters applied where the contaminant mitochondrion falls outside of all human variations. This is due to low-quality bases as a reiteration of the phylogenetic reconstruction using only high-quality bases resulted in

an inferred contaminant mitochondrion which falls within the variation of extant humans (see Figure B.16 in the Appendix). Furthermore, the likelihood of the tree increases as only high-quality bases are retained. Attempts to assign the inferred contaminants to known haplogroups are presented in the following subsection on page 134.

As the Mezmaiskaya mitochondrial genome had been previously sequenced using data with low present-day human contamination, the endogenous consensus call for both Mezmaiskaya datasets could be compared to this mitochondrial genome.

5.4.1.3.1 Mezmaiskaya 1

Under the assumption that the sequence from GenBank is without errors, the endogenous genome inferred by schmutzi should match perfectly this reference sequence. The inferred endogenous sequence differed by 9 of the 16608 bases. These differences fall in the D-loop which is typically quite divergent. It could be that the incorrect identification of these 9 bases may arise from an ascertainment bias due to the mitochondrial capture of the Mezmaiskaya sample using probes designed with the human mitochondrial sequence. Indeed, in this region, the endogenous bases were significantly under-represented compared to the contaminant (75% rather than the average of 50% for the whole mitochondrial genome). However, these bases tend to have low consensus base quality, which implies that the consensus calls at these positions is unreliable. Filtering for consensus base quality $\geq$200 (PHRED scale) reduces the number from 9 mismatches to 1. This single mismatch is in the poly-C region (position: 16,184) which is routinely removed in downstream analyses [4, 24].

5.4.1.3.2 Mezmaiskaya 2

An alignment of the unfiltered predicted Mezmaiskaya B9688 genome to the original mitochondrion from the same individual revealed a total of ten mismatches, two of which had very low-quality score (5.09715 and 83.9567 respectively) while the eight remaining mismatches were all concentrated in a range of 60 basepairs at the end of the mitochondrial reference (positions 16129-16190 on the mitochondrial reference). Using a filter for high-quality bases (Q$\geq$200) eliminated the first two miscalls in that loci but left six mismatches in the aforementioned locus of 60 bases on the mitochondrial genome. A closer look revealed a high level of divergence of the Mezmaiskaya mitochondrial genome to the human reference and a drop of coverage in that area. At position 16,139 on the rCRS, for instance, total coverage was 431X and where the contaminant base had 327X coverage thus 75.9% of the fragments. In contrast, the genome-wide mean coverage was 636X and the contamination rate was 48-50%. To verify whether this was due to a bias caused by the short-read aligner, the fragments were re-aligned to the Mezmaiskaya mitochondrial genome. The results (data not shown) revealed the same drop in coverage in the same area. Communication with the authors involved in generating the original data revealed that, like Mezmaiskaya 1, a mitochondrial capture was performed using a tiled array only

with the human base on the probes. Therefore, this artifact was likely due to capture bias which is currently not modeled.

5.4.1.4 Contaminant mitochondrion consensus call

As previously mentioned, since there are no tools to call the contaminant mitochondrial genome and since the contaminant was not previously characterized, the inferred contaminant genomes could not be compared to a known sequence. However, as there is a finite set of mitochondrial haplotypes among present-day humans, the predicted contaminant sequence can be compared to existing haplotypes to determine whether it falls within a given haplogroup (i.e. the diagnostic positions for this haplogroup are found). The most likely haplogroup as determined by haplogrep [56, 110] and the calls produced by schmutzi at the diagnostic positions for the most likely haplogroup were evaluated.

For both Mezmaiskaya samples, the most likely haplogroup as determined by haplogrep was T2b3, a haplogroup predominantly found in Eurasia [5]. All but one of the 33 diagnostic positions were found in the predicted contaminant for the B9687 sample (see Table A.22 in the Appendix). The single mismatch had low-quality relative to the other diagnostic positions. The other Mezmaiskaya sample, B9688, had no mismatches for all of the 33 diagnostic positions (see Table A.23 in the Appendix).

5.4.2 Simulated data

This section presents results on simulated data in terms of the accuracy of the contamination estimates (pages 134-145) and in terms of the accuracy of the endogenous and contaminant consensus call (pages 145-149). The contamination estimates based on deamination patterns computed using "contDeam" is presented (see Section 5.4.2.1) for both full datasets (see Section 5.4.2.1.1) and subsampled datasets to measure the impact of low coverage (see Section 5.4.2.1.2). This is followed by a short discussion about biases affecting "contDeam" (see Section 5.4.2.1.3). Contamination estimates using divergence positions between the predicted endogenous genome and a database of putative contaminants using "mtCont" is presented (see Section 5.4.2.2) for both full datasets (see Section 5.4.2.2.1) and subsampled datasets (see Section 5.4.2.2.2). Comparison to existing methods for contamination estimates follows (see Section 5.4.2.3). Results are presented for the consensus call for the endogenous mitochondrial genome (see Section 5.4.2.4) and the contaminant one (see Section 5.4.2.5).

5.4.2.1 Contamination estimate based on deamination

The correlation of the contamination estimates obtained using endogenous deamination patterns was compared to the simulated ones. This is the contamination estimate provided to the endogenous caller for the first iteration. The correlation between simulated

and predicted contamination rates was measured for full datasets with 1M fragments. The robustness to low coverage was also measured by subsampling the set containing 1M fragments with 40% contamination. The target contamination rate for the simulations was calculated as the fraction of fragments pertaining to the contaminant over the total.

5.4.2.1.1 Full datasets

Schmutzi was run on the simulated datasets with 1M fragments to estimate contamination based on deamination patterns alone. The software was run for both categories of sets: one category where the endogenous genome had a double-stranded type of damage pattern, and the other where a single-stranded damage profile was used.

Results show that, regardless of the simulated DNA library-preparation protocol, the algorithm produces an estimate that is close to the simulated rate (see Figure B.9 in the Appendix). Furthermore, these estimates are robust to lower or higher divergence of the contaminant genome to the endogenous one, as this relationship is not *a priori* needed for this approach to produce an estimate.

5.4.2.1.2 Subsampled datasets

To evaluate the robustness of the contamination estimate based on deamination patterns to lower coverage, the dataset with 1M fragments and 40% contamination from the previous section was subsampled at various fractions ranging from 0.01 to 0.5. The algorithm to predict contamination based on deamination patterns was run on those and the correlation to the original contamination rate was plotted (see Figure B.10 in the Appendix). Results show that for the contamination estimate to be stable, a minimal mitochondrial coverage of about 100X to 250X is needed, which, depending on the size of the aDNA fragments, represents approximately 50k to 100k mapped fragments. The simulated type of library protocol or the type of endogenous genome used does not seem to affect the prediction.

5.4.2.1.3 Biases affecting the prior contamination estimate based on deamination

The contamination prior obtained for the first iteration relies on measuring deamination patterns for the endogenous fragments versus the entire dataset. One of the approaches to infer endogenous deamination rates is by conditioning on one end of the fragment being deaminated and measuring deamination rates for the other end, a previously described methodology by [76]. However, while this approach is used to obtain an initial mitochondrial contamination estimate, it can be used for contamination estimates in itself under the following assumptions:

- There is a sufficient number of fragments to allow estimates of deamination rates

- Deamination rates of the endogenous fragments are sufficiently high. Having endogenous fragments with no deamination patterns will not yield an accurate estimate
- The aDNA fragments from the present-day humans that contaminate the sample are not themselves deaminated
- The rates of deamination of the 5' end of the fragment are independent of the rates of deamination of the 3' end and vice-versa

5.4.2.1.3.1 Impact of low deamination rates

To measure the impact of having low deamination rates, simulations were repeated by adding various rates of deamination for the endogenous fragments for simulated datasets with 50% contamination. Schmutzi's contamination estimate based on deamination patterns was run on the simulations. Results presented in Table A.19 in the Appendix show that a minimum deamination rate of 5-10% on at least one end of the fragment is required to have a contamination estimate within 2-3% of the simulated contamination rate if 1M fragments are used. When a small number of fragments are used (100k), higher rates (40% and above) of deamination are required to obtain a reliable contamination estimate. At intermediate data sizes (500k), rates of deamination upwards of 15% are needed to obtain a reliable contamination estimate.

5.4.2.1.4 Impact of deamination for contaminating fragments

To measure the impact of various rates of deamination of the contaminant fragments, in addition to deamination to the endogenous fragments, deamination was also added to the simulated contaminant fragments. A contamination rate of 50% was used for the simulation sets of 1M fragments for various rates of deamination for both the endogenous and contaminant fragments. Schmutzi's contamination estimate was used on those datasets. Results show (see Table A.20 in the Appendix) that even a small amount of deamination for the contaminant can lead to an underestimate. This effect less pervasive if the endogenous fragments have high levels of deamination or if the contaminant has very low levels of deamination.

5.4.2.1.5 Independence tests for deamination on each end

The contamination estimate based on deamination relies on measuring endogenous deamination rates and plotting the posterior probability for a non-informative contamination prior. Diagnostic positions cannot always be used for measuring endogenous deamination rates for aDNA data. Therefore, the algorithm needs to condition on having one deaminated base on one end to measure endogenous deamination rates on the other and

vice-versa. One underlying assumption is that deamination on one end is independent of deamination the other end. Whether deamination rates on either end of the fragment were independent of deamination on the other end was evaluated. Deamination rates were measured on the 5' end conditioning on whether the 3' end was deaminated (C→T) or not (C→C). The converse was also measured. Subsets of the Altai Neanderthal [89], the Denisovan individual [77], the Loschbour individual [60], the Afontova Gora and Mal'ta genomes [91] (see Table A.21 in the Appendix) were evaluated. A χ^2 test was used on a two by two contingency table with one degree of freedom to test whether deamination on one end was independent of deamination on the other end. For all samples, except the Altai, the p-value was not sufficiently low to the point of concluding that deamination on one end is linked to deamination on the other. However, it should be noted that this is an assumption used by the algorithm and, if this assumption is incorrect and endogenous deamination rates are overestimated, an overestimate of the actual contamination rate will ensue.

5.4.2.2 Contamination estimate based on divergent bases

Once the endogenous consensus call is completed, contamination rates can be computed using this consensus and a set of putative mitochondrial contaminants. This process is repeated until a stable contamination rate is reached and the final rate is produced. To evaluate the range of contamination and coverage over which schmutzi can be used, the three simulated datasets were evaluated with increasing levels of contamination and at varying coverage. Similarly to the section above, the correlation between this final contamination rate and the predefined target contamination rate used in the simulated data was measured. The target contamination rate for the simulations was calculated as the fraction of bases pertaining to the contaminant over the total sum.

5.4.2.2.1 Full datasets

As mentioned in the methods, users can run the prediction with or without the inclusion of the predicted contaminant as a record in the database of putative contaminants. Schmutzi was run on the six types of previously described datasets of 1M fragments (see Section 5.3.2.1).

Schmutzi was run once with the inclusion of the predicted contaminant (see Section 5.4.2.2.1.2) and once again without this option (see Section 5.4.2.2.1.1).

5.4.2.2.1.1 Full datasets: Using the records in the database only

Using solely the records in the database described in Section 5.3.8, the contamination rate was computed once the algorithm reached convergence.

This option always results in an underestimate of the true contamination rate as some sites on the mitochondrial genome will not be considered due to natural divergence between the actual contaminant and the closest record in the database. The correlation was plotted between the simulated contamination rate and the predicted one (see Figure 5.17 for the single-stranded data and Figure B.11 in the Appendix for the double-stranded data).

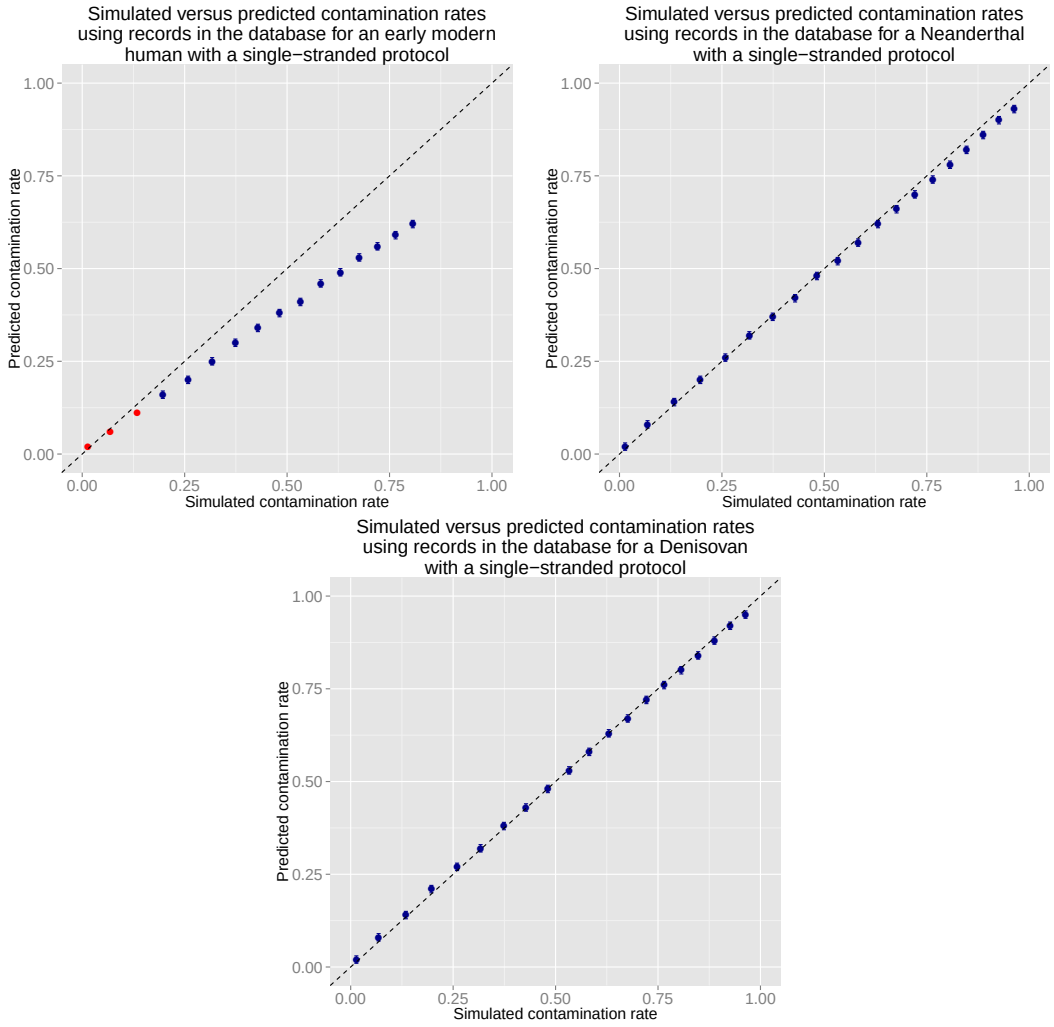


Figure 5.17: Simulated versus measured contamination rates using the contaminants in the database only. Several sets contained simulated aDNA fragments from a mitochondrial genome belonging to either an early modern human (top left), a Neanderthal (top right) and a Denisovan (bottom). All simulated sets had damage patterns associated with a single-stranded DNA protocol. The double-stranded figures can be found in Figure B.11 in the Appendix. A contaminating present-day human was pooled together at various rates to simulate contamination. The dotted black line represents a perfect prediction, blue dots are the predicted rates of contamination by schmutzi once convergence was achieved. The red dots represent sets for which the algorithm stopped prematurely due to lack of information about the contaminant fragments. The black whiskers represent the 95% confidence interval on contamination.

For both archaic hominins, due to the large numbers of segregating sites compared to the contamination source, the effect of this underestimate is minimal as the contamination estimate is highly correlated with the simulated one. For the EMH, due to the smaller divergence between the contaminant and endogenous genomes, very few sites are considered and the effect of the underestimate is more prominent, especially at higher contamination rates. The following section shows that these more difficult targets can be predicted by including the inferred contaminant in the database of putative contaminant genomes.

5.4.2.2.1.2 Full datasets: Including the predicted contaminant

The program was re-run on the same datasets used in the previous section with the inclusion of the predicted contaminant.

The correlation between the simulated contamination rate and the predicted one was plotted (see Figure 5.18 for the single-stranded data and Figure B.12 in the Appendix for the double-stranded data). The program performed well for both archaic hominin genomes, similarly to the previous section, as high divergence between the contaminant and the endogenous genomes provide an easy target for contamination estimates. For the EMH, the underestimate seen in the previous section is corrected for using the predicted contaminant as information. However, this approach does not perform well at very low levels of contamination, as adequate characterization of the contaminant genome is not feasible.

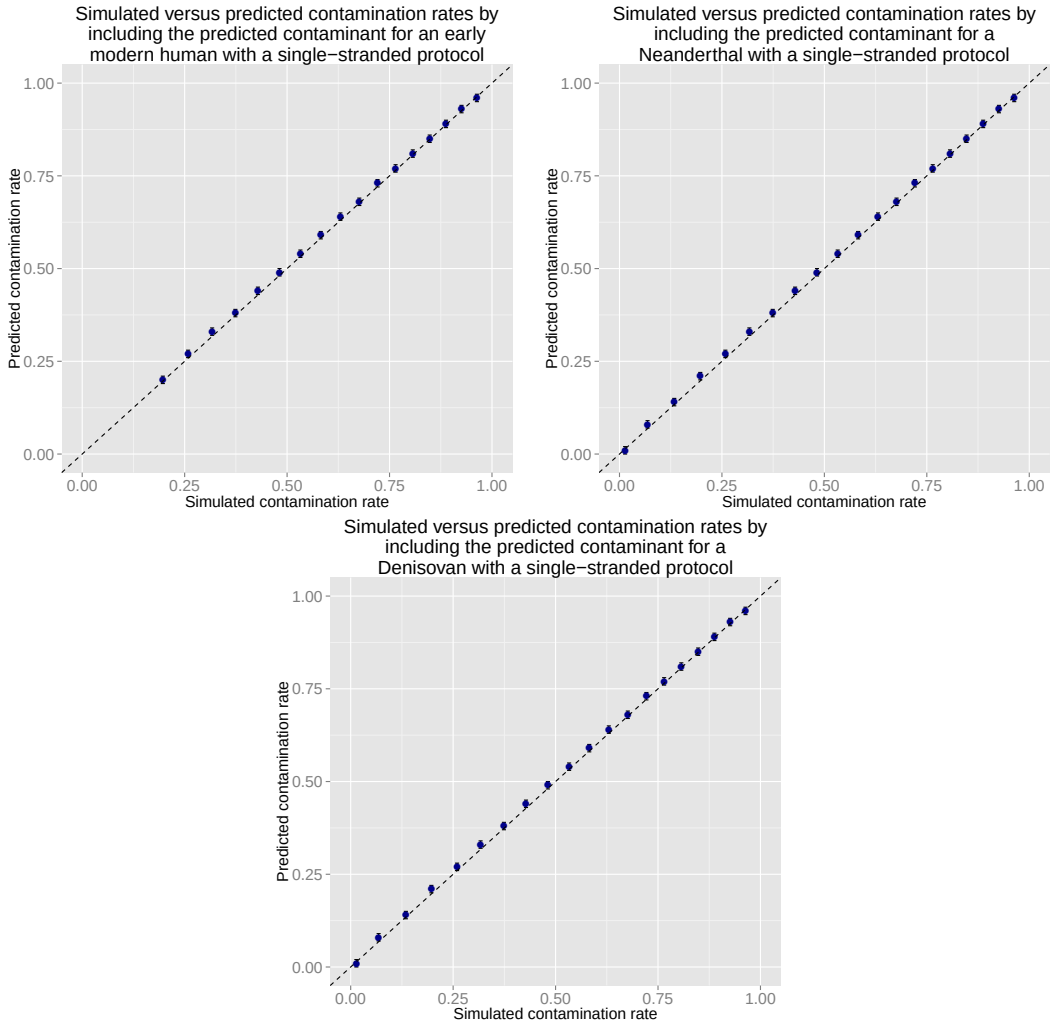


Figure 5.18: Simulated versus measured contamination rates using the contaminants in the database with the inclusion of the predicted contamiant. Sets contained simulated aDNA fragments from a mitochondrial genome belonging to either an early modern human (top left), a Neanderthal (top right) and a Denisovan (bottom). All simulated sets had damage patterns associated with a single-stranded DNA protocol. The double-stranded figures can be found in Figure B.12 in the Appendix. Like in Figure 5.17, the dotted black line represents a perfect prediction, blue dots are the predicted rates of contamination by schmutzi once convergence was achieved. The missing dots for the EMH represent sets for which the algorithm did not converge due to the inability of predicting the contaminant. The black whiskers represent the 95% confidence interval on contamination.

5.4.2.2.2 Subsampled datasets

To measure the robustness of the algorithm to low coverage samples, the dataset with ~ 47% contamination rate was subsampled at rates ranging from 0.5 down to 0.01. The rate of ~ 47% was chosen as it makes the use of currently available tools difficult. Furthermore, at this level of contamination, there is an almost even number of endogenous and contaminant bases thus making the inference of each one relatively difficult for the model.

Two distinct approaches were taken when re-running schmutzi on the resulting datasets. The first involved the default behavior of predicting the endogenous genome and scanning the contaminant database to estimate contamination rates. However, for very low coverage samples, getting an accurate resolution of the endogenous mitochondrial genome is often difficult. Sometimes, investigators have access to an endogenous mitochondrial genome that can serve as a close proxy (e.g. a different Neanderthal for a Neanderthal sample, a mitochondrial genome from the same haplogroup for early modern humans) to determine whether this low coverage sample is heavily contaminated. This is useful to prioritize which extractions are the most promising and should be further sequenced. The second approach therefore involved taking the endogenous consensus from the original high coverage dataset and using it as the endogenous genome. This latter approach has the advantage of being highly robust to low coverage but requires a well-characterized endogenous genome or a very close proxy.

5.4.2.2.2.1 Subsampled datasets: Using a consensus from the dataset itself

Using the default methodology, contamination rates were inferred from the predicted endogenous genome and putative contaminants in the database. As the simulated contamination rate was known, the final contamination rates were plotted as a function of coverage (see Figure 5.19 for the single-stranded data and Figure B.13 in the Appendix for the double-stranded data). For both archaics, the contamination estimate is reliable from about 100X or 200X coverage for the single-stranded and double-stranded rates of deamination respectively. For the EMH, the contamination estimate remains an underestimate since schmutzi does not use the predicted contaminant as a putative source of contamination.

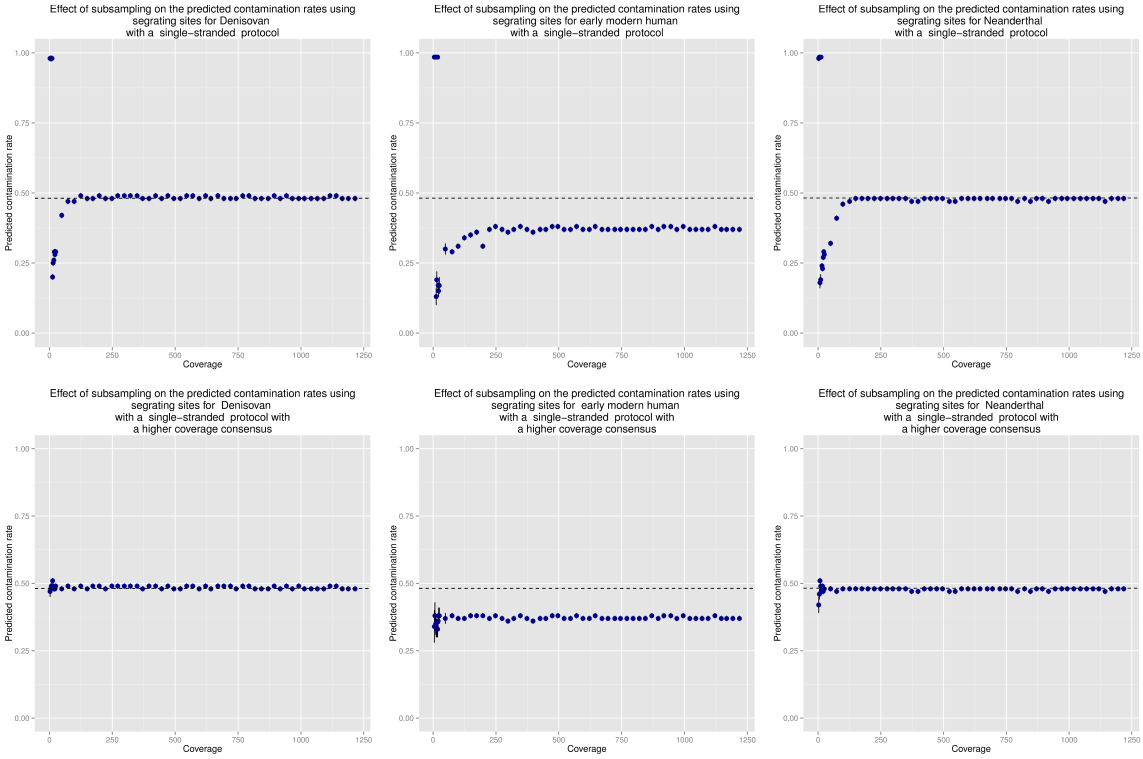


Figure 5.19: Robustness of the contamination estimate to lower coverage. The simulated dataset with a contamination rate of ~ 47% and single-stranded deamination patterns was subsampled at various coverages from 0 - 1250x. Top: Contamination rates were estimated across a range of coverages in simulated data for a Neanderthal, a Denisovan and an early modern human (Ust'-Ishim). Bottom: When a good quality proxy mtDNA sequence from a closely related individual is used as the endogenous genome, robust estimates can be made down to 15x coverage. For the early modern human, the contamination estimate provided was computed using the database alone and not the prediction of the contaminant genome thus leading to underestimates (see Table 5.4 for effect of using the predicted contaminant in the contamination estimate).

5.4.2.2.2.2 Subsampled datasets: Using a consensus from a higher quality source

In the previous section, it was shown that schmutzi performed well at coverage levels that are routinely seen in aDNA projects due to the relative abundance of the mitochondrial DNA compared to nuclear [42]. However, in certain studies, the relative amount of non-bacterial DNA is relatively small leading to extracts yielding low coverage across the mitochondrial genome (e.g. less than 15X). In those cases, neither approach to estimate

contamination by deamination patterns or by endogenous consensus calling followed by comparison to a database yielded accurate estimates.

A hurdle in predicting contamination using low coverage samples is the inability to accurately call the endogenous mitochondrial genome. However, it is possible that researchers have access to a higher quality mitochondrial genome from the same individual (obtained using mitochondrial capture for example) and wish to prioritize which extractions are most promising to fully sequence the nuclear genome. It is also possible to determine from which clade or haplogroup the individual being sequenced belongs to therefore providing a close proxy. Results show that if a research group has access to a high-quality mitochondrial genome from a close proxy, contamination can be estimated even at low coverage. This approach can be useful if a group prepared a new library from a Neanderthal extract and wishes to estimate contamination despite low coverage across the mitochondrial genome. Knowing that the sample pertains to a Neanderthal entails that a high-quality mitochondrial genome from a different Neanderthal can be used as substitute. The contamination rate could therefore be estimated for the new low coverage library. Schmutzi's contamination estimator was supplied with the endogenous genome predicted from the original 1M fragment datasets. Results show that the estimates are accurate even for very low coverage samples (see Figure B.14 in the Appendix).

For both archaics hominins, the estimate is close to the actual simulated rate even at low coverage. For the EMH, the underestimate due to the exclusion of the contaminant is still noticeable however, the estimate offers greater robustness to low coverage rates compared to simply estimating contamination using the endogenous consensus from the sample itself.

5.4.2.3 Comparison to existing methods

Using the maximum-likelihood method previously described in the literature [36], a contamination estimate for each simulated set of 1M fragment was computed. Results measure the correlation between the simulated contamination rate and the one obtained using this method (see Figure B.15 in the Appendix). As the contaminating mitochondrial genome was known, the program was run once where this genome was used as the contamination source. The program was run again using the closest mitochondrial genome to the contaminant one in the 311 database records provided in the original description of the method.

One issue with this maximum-likelihood method is the inability to quantify the three main sources of uncertainty: sequencing errors, deamination and mismappings. The result is an estimate that misses the simulated contamination rate at lower and higher levels of contamination. An underestimate of the error rate leads to an overestimate of the contamination rate and vice-versa. Mitigating measures against deamination can be taken like trimming the ends of fragments or restricting the analysis to transversions

only. However these approaches suffer from residual deamination in the middle of the fragments and reduction of ascertainment power respectively. The impact of mismappings could be mitigated by filtering for fragments with high mapping quality but this does not guarantee that every fragment is correctly mapped to its original position.

5.4.2.3.1 Using the implementation from the author

To test the accuracy of the algorithm to existing software implementations, schmutzi and contamMix were run on a simulated dataset of 1M fragments with double-stranded deamination patterns. The endogenous mitochondrial genome used was an early modern human with 50% present-day human contamination. Results show that schmutzi's algorithm offers superior accuracy compared to this existing method for estimating early modern human contamination (see Table 5.4). Results for the maximum-likelihood method used by contamMix for the remaining samples are presented in this section on page 144.

contamination estimate method	contamination estimate	runtime
Target contamination rate: 50% (fragment basis)		
contamMix 1.0-10	54.9±0.7 %	4 days
schmutzi ("contDeam")	49.0 ±0.5 %	68s
Target contamination rate: 58.2% (nucleotide basis)		
schmutzi ("mtCont" without the predicted contaminant)	32.0 ±1.0 %	183m
schmutzi ("mtCont" with the predicted contaminant)	60.0 ±1.0 %	200m

Table 5.4: Accuracy of contamination estimates on a simulated early modern human with double-stranded deamination patterns and high present-day modern human contamination. Three CPU cores were used for every program. The programs "contamMix" and "contDeam" estimate contamination on a per fragment basis while "mtCont" estimates contamination on per nucleotide basis. The contamination on a per nucleotide basis is higher due to the longer average length of contaminating fragments.

5.4.2.4 Endogenous mitochondrion consensus call

Schmutzi was run on the datasets created for three archaic genomes, each with increasing levels of present-day human contamination. For the simulated data, the accuracy of the consensus call for both the endogenous and the contaminant genomes was evaluated as the original mitochondrial genomes were known (see Figure 5.20).

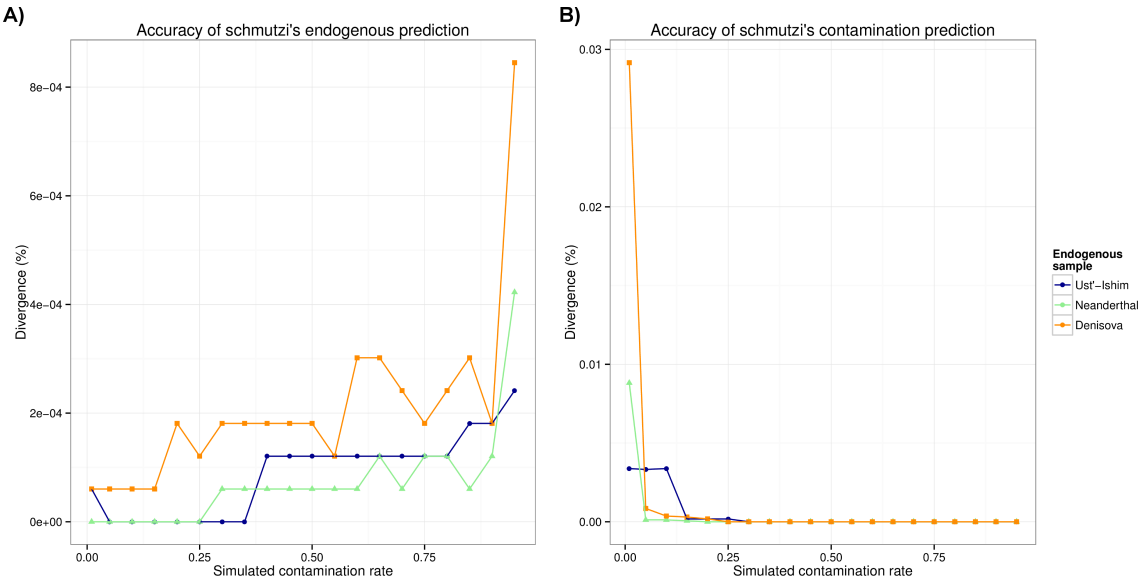


Figure 5.20: Accuracy of the ancient (A) and present-day human contaminant (B) mitochondrial consensus sequences produced by schmutzi on simulated data for an early modern human, a Neanderthal and a Denisovan mitochondrial genome. An error is defined as either a mismatch or an indel between the predicted endogenous sequence and the original mitochondrial sequence used for simulations. As contamination increases, inference of the endogenous mitochondrial genome becomes more difficult (A). In contrast, the prediction of the contaminant genome becomes more accurate at higher levels of present-day human contamination (B).

As detailed in section 5.3.2.1, fragments from three endogenous mitochondrial genomes (Denisovan, Neanderthal and early modern human (EMH)) were each blended independently with the fragments of a contaminant individual at various rates. Two sets were created, one where the endogenous fragments had a damage pattern consistent with a double-stranded library and the other set, with a single-stranded library. For the former, the endogenous consensus was also computed using PMDtools and htslib to provide a comparison. The mitochondrial consensus sequences are called for each sample after processing the data using PMDTools (using the parameters "-a" to adjust quality scores and the recommended PMD score threshold of 3) to identify deaminated reads and then calling the consensus with htslib (default parameters and haploid model).

The approach to infer the endogenous consensus of keeping fragments with signs of deamination, masking the deaminated bases and calling a consensus using "samtools mpileup", an approach used to reconstruct the heavily contaminated Sima de los Huesos mitochondrial sequence [76], was also evaluated. For each set, schmutzi was run using default parameters and the edit distance of the predicted endogenous genome to its respective reference was computed. Furthermore, the edit distance of the predicted contaminant

to the original contaminant mitochondrion was computed. Schmutzi was also run using the multiple contaminant option (described in section 5.3.4.5) and the accuracy of its predicted endogenous genome was evaluated. All the data presented in the remaining tables were computed without using any filters on the resulting predictions as to accurately represent error rates. Practically speaking, users are encouraged to retain only high-quality predictions for downstream analyses.

Schmutzi produced a consensus for both the endogenous and contaminant genomes that is very robust to high levels of contamination (see Figure 5.20). Results show that the endogenous consensus is accurate for up to 50% present-day human contamination for the double-stranded simulations and up to 70% for the single-stranded ones. This is due to higher levels of deamination in the single-stranded simulations resulting in better ascertainment of the endogenous base. The sequence identity of the consensus sequences generated by schmutzi to the published reference sequences of either the Denisovan, Neanderthal or early modern human are presented in Tables A.10-A.15 in the Appendix.

As schmutzi's algorithm relies on computing the length and deamination patterns of the endogenous and contaminant fragments, a paucity of contaminant fragments at low contamination rates can result in the program stopping after the first iteration. In the tables of the Appendix, a † symbol on a data point indicates that the algorithm did not continue on to the second iteration due to a lack of detected contaminant at low simulated contamination rates, thus the results presented are the ones from the first iteration. At high contamination rates, as the prediction of the endogenous genome becomes more arduous, the endogenous consensus genome will contain more bases from the contaminant, thus leading to an underestimation of contamination, which can, in turn, lead to the algorithm not converging. As mentioned in the software manual, a corrective measure can be performed by using the predicted contaminant genome as a putative contaminant source via the "–usepredC" option. Sets marked with a ‡ indicate that the predicted contaminant was used as a contaminant source. For very hard targets (e.g., EHM with around 90% contamination), the workflow provided by the wrapper script diverges even with the option of using the contaminant source. For such hard targets, manual intervention would be required and data that caused this type of problem is marked with an $*$ in subsequent tables.

The improvement obtained by schmutzi over approaches that use only deaminated reads from highly contaminated samples results from the inclusion of length and the observed ratio of endogenous and contaminant bases. Iteration increases the accuracy of the endogenous consensus call. The initial call for the Neanderthal dataset with a simulated contamination rate of 58% had 7 mismatches to its original reference while only a single mismatch remained after convergence.

For the simulated EMH, the endogenous genome predicted by schmutzi is identical to the simulated data up to a contamination rate of 35% for the double-stranded data and up to a contamination rate of 40% for single-stranded libraries (see Tables A.10 and A.11 in

the Appendix). As single-stranded data had greater rates of deamination, there is more power to accurately predict the contaminant and endogenous bases. As the contamination increases, more indels and mismatches appear.

The consensus made on the deaminated fragments using PMDtools and htslib has two indels compared to the endogenous genome, both of which are located in a region of two consecutive insertions in the EMH mitochondrion genome. The same comparison to the endogenous genome predicted by schmutzi was made using the data simulated under a single-stranded protocol (see Table A.13 in the Appendix). The algorithm was able to perfectly predict the endogenous genome up to a contamination of 25% and with a single mismatch up to a contamination rate of 90%. That single mismatch occurred in a region of high divergence of the Neanderthal genome and had a low-quality score relative to the neighboring bases.

For the Denisovan simulations, the edit distance of the predicted endogenous genome to the Denisovan one was computed (see Table A.15 and A.14 in the Appendix). This dataset had the highest divergence to the human reference. A single recurrent error was present even at low contamination around base 302 on the rCRS due to high divergence creating ambiguous short-read alignments. However, re-running schmutzi's endogenous caller using the predicted endogenous genome as reference successfully removes this single mismatch (data not shown). Despite this, the algorithm was more robust to high contamination than the current approach of isolating deaminated fragments and calling a consensus.

For all three types of endogenous genomes, at low levels of contamination (up to 10%), schmutzi did not go forward after the first iteration due to the lack of contaminating fragments. However, in all cases, the endogenous genome called after the first iteration gave an inference of sufficient quality with no or very few mismatches to the original genome.

Simulations show that the multiple contaminant option works well at very low rates of contamination, but does not at medium or high rates (see Tables A.10 through A.15 in the Appendix).

For calling the endogenous mitochondrial genome consensus, the mapping iterative assembler (MIA) was originally developed for reconstructing the Neanderthal mitochondrial genome [44]. MIA has been used for reconstructing the mitochondrial genome for multiple samples [9, 36, 60]. The latest version of MIA[5] was used on the simulated datasets and the distance to the original endogenous genome was computed (see Table A.16 in the Appendix). Results show that present-day human contamination quickly overruns the consensus call. This effect limits the applicability of a straightforward consensus call to samples with low rates of present-day human contamination.

[5]URL: https://github.com/udo-stenzel/mapping-iterative-assembler version:5a7fb5afad735da7b8297381648049985c599874

5.4.2.5 Contaminant mitochondrion consensus call

As previously mentioned, no currently available tool enables users to the call the contaminant mitochondrial genome. However, schmutzi's consensus call for the contaminant genome was compared to the original contaminant genome used by computing the edit distance as a metric (see Table A.17 and A.18 in the Appendix).

The accuracy of the contaminant genome inferred from the simulated datasets increased as the amount of contamination increased (see Figure 5.20). At very low rates of contamination, schmutzi is unable to call the contaminant mitochondrial genome. For contamination rates of about 20% and higher, the prediction of the contaminant genome is nearly perfect regardless of which endogenous genome was used.

5.5 Conclusion

aDNA analyses have typically decoupled reconstruction of the endogenous mitochondrial genome from quantification and characterization of present-day human contamination. Since these two tasks are interdependent, consensus calling and contamination estimation should be performed iteratively to achieve the most accurate results. Current approaches to determining the endogenous mtDNA sequence are very dependent on the amount of contamination. In samples with low present-day human contamination, a consensus sequence is usually called using all sequences, whereas for highly contaminated samples, only deaminated fragments are used. However, there is no clear contamination cut-off to determine which strategy should be used. Schmutzi can be applied to samples with either low or high levels of contamination thereby obviating this decision.

Results on empirical and simulated datasets were presented, demonstrating that schmutzi outperforms a number of existing approaches to consensus sequence calling and contamination estimation over a wide range of contamination rates and coverages. Simulations were conducted using empirical fragment length distributions and deamination rates. It is trivial to see that higher deamination rates can enable end-users to infer with greater confidence the endogenous sequence of even highly contaminated samples. The absence of deamination will yield incorrect estimates of contamination. Since deamination is also used to identify the endogenous and contaminant bases, an absence of deamination is also likely to lead to an incorrect endogenous consensus call at high levels of contamination. It is important to note that the number of parameters and their range hinders us from making simple, general statements about the amount of coverage or extent of deamination required for accurate estimates of present-day human contamination or accurate inference of the endogenous genome sequence.

Although many groups have implemented *ad hoc* methods to assess contamination, there are few available software implementations. Schmutzi was compared to contamMix, a

previously used maximum-likelihood method described in [36]. The predicted contamination rates produced by schmutzi are more accurate than those produced by this method on simulated data (see Section 5.4.2.3). Although the true contamination rate is not known for most ancient datasets, it was shown that the estimates are also consistent with contamination measured in empirical datasets using methods relying on diagnostic positions. While the approach of taking diagnostic positions is suitable for archaic humans like Neanderthals, it is not readily applicable to early modern humans who have few fixed differences to present-day humans. Schmutzi's modeling of mismatches due to deamination, sequencing errors and mismapping results in greater accuracy than simply estimating a single error parameter.

The endogenous consensus call shows a significant dependence on the prior, which is calculated based on the deamination patterns only for the first iteration ("contDeam"). This is interpreted as evidence that a reasonable estimate for contamination can be obtained from deamination. For "contDeam", the impact on the final estimate due to biases like insufficient deamination and having deamination for contaminant fragments was evaluated (see Section 5.4.2.1.3). It was noted that the contamination estimate improves incrementally during iterations of consensus calling and contamination estimation, suggesting that additional information is available in the mitochondrial endogenous consensus. This is particularly useful for low coverage samples.

Schmutzi accurately infers the endogenous ancient genome sequence from unfiltered ancient sequence data. This is of particular importance in cases where contamination is high. Interestingly, schmutzi is also more accurate than approaches that reduce contamination by using only deaminated fragments to call the consensus. Such approaches substantially reduce the number of fragments available for calling the consensus, which may explain why schmutzi is marginally better at determining the consensus sequence than these approaches.

Although schmutzi performs well for both simulated and empirical data, a few artifacts are not currently modeled in the software. First, it is possible that there are multiple present-day human contaminants. At low contamination rates with multiple contaminants, schmutzi will underestimate the contamination, but the inference of the endogenous consensus sequence should not be affected. However, at high contamination rates, multiple contaminants make the inference of the endogenous sequence and estimation of the contamination extremely difficult, since the endogenous and contaminant alleles do not follow the expected distributions. Second, the inclusion of misaligned microbial sequences and mitochondrial heteroplasmy are also not currently considered in the computation, though the empirical data suggest that schmutzi is not particularly sensitive to these. Lastly, the use of target enrichment approaches with DNA probes that are closer to the contaminant than to the endogenous sequence may cause differences in allele sampling, and may lead to incorrect consensus calls (see Figure 5.15 and see Section 5.4.1.3 for further discussion about capture bias).

Schmutzi is sensitive to the divergence between the actual contaminant and the closest record in the database of putative contaminants. If this divergence is very large (e.g. more than 30 mismatches), contamination will be underestimated.

When contamination rates are high, the predicted contaminant can be inferred at high resolution. This enables the program to use this predicted contaminant as a database record for the quantification of mitochondrial contamination. This is not feasible at low contamination rates, where the prediction of the contaminant mtDNA is poor. The method does not currently use phylogenetic information to infer the endogenous and contaminant sequences. Although the approach works well empirically, the use of phylogenetic information could provide additional power for obtaining contamination estimates in very low coverage samples.

In conclusion, an algorithm that infers the endogenous mitochondrial genome sequence from an ancient DNA sample, even in the presence of high contamination, has been described. This method was applied to the reconstruction of mitochondrial genomes for archaic and early modern humans and it was shown that it is possible to accurately quantify contamination from present-day individuals.

Chapter 6

Conclusion

> *He [Ludwig Wittgenstein] once greeted me with the question: "Why do people say that it was natural to think that the sun went round the earth rather than that the earth turned on its axis?" I replied: "I suppose, because it looked as if the sun went round the earth." "Well," he asked, "what would it have looked like if it had looked as if the earth turned on its axis?"*
>
> - Elizabeth Anscombe, *An Introduction to Wittgenstein's Tractatus [75]*

As the quote illustrates, a common error in science is to infer a model simply based on observations. The Bayesian principle turns the question on its head by asking, what is the probability that a specific model could have given rise to a particular observation. In return, the most likely model and its parameters can be inferred by choosing the one that maximizes the probability of having produced the observation. In the end, the posterior probability of a Bayesian approach is not a result in itself, it is a result only when compared to the posterior probability of other models.

This thesis used the Bayesian principle to compute the probability that a certain model could have generated a particular observation of NGS data. Generally speaking, this approach obviates the need for arbitrary cutoffs and ensures that samples are comparable across sequencing runs. As mentioned in the introduction, for Illumina sequencing data, the error rate often varies even within the same sequencing run.

In chapter 3, the most likely DNA fragment was produced given observed sequencing reads. This algorithm currently represents the most accurate method of inferring the original aDNA fragments from sequencing data. This was verified both on empirical data and on simulated data where the error rate was increased at various levels. Our approach is also faster than other existing computational tools aimed at the task of inferring aDNA fragments.

The problem of demultiplexing, i.e. assigning reads to their sample of origin, was tackled in chapter 4. A Bayesian approach with a non-informative prior was used to identify

the sample of origin and compute the confidence in that assignment. This strategy outperforms approaches using fixed mismatches in terms of assigning reads to samples, especially at high sequencing error rates. The Bayesian approach described in chapter 4 produces confidence scores in the sample assignment that measure the probability of a misassignment.

A Bayesian approach was also used for the problem of inferring the mitochondrial genome of the endogenous and contaminant individuals in aDNA samples from archaic humans. This algorithm was presented in chapter 5. Again, the bases that maximize the posterior probability are produced. The mitochondrial genomes being predicted are more accurate than the ones inferred by current computational methods. Once the endogenous genome is obtained, an accurate estimate of mitochondrial contamination from present-day humans can be inferred. The Bayesian algorithm for estimating contamination is faster and more accurate than previously described heuristics.

The three Bayesian algorithms which aim to solve the problem of aDNA fragment inference, demultiplexing and contamination estimates represent, as of this writing, the most accurate, fastest and most mathematically sound way to solve their respective problem. However, these algorithms rely heavily on the ability to accurately quantify uncertainty for sequencing data. To this effect, chapter 2 describes a basecalling algorithm that produces very accurate basecalls along with predicted error probabilities that are highly correlated with observed error rates. Furthermore, our implementation performs this basecalling at a reasonable runtime thus allowing for its routine use in sequencing centers. These programs have been used for various biological projects both here at the Max Planck Institute [35, 60, 76, 107, 109] and elsewhere [40, 115].

Although there is always room for improvement both in terms of the algorithms and their implementation, it is a step towards crafting computational methods that are tested and mathematically sound. The thesis demonstrated the advantage of Bayesian MAP algorithms over threshold-based heuristics both in terms of simplicity and results. Methods with cutoffs for input quality require the added labor of testing and measuring the effect of various values for cutoffs on estimated parameters. Furthermore, this effort needs to be duplicated as the quality of the input data changes. This thesis showed how Bayesian MAP algorithms can simplify NGS analyses, particularly for aDNA, and how estimating the probability of a sequencing error upstream obviates the need for downstream recalibration.

For the field of aDNA, there is a need to create, test and implement algorithms that incorporate uncertainties due to sequencing errors, low-coverage, DNA fragmentation, deamination and contamination into a probabilistic model. Reproducible and open science is a goal worth striving for as the methods section is often overlooked by biology-avid readers in large scale papers. As data quality varies from one project to another, Bayesian methods offer unbiased estimates in a manner than threshold-based methods cannot.

Appendices

Appendix A

Basecaller	Mapped	(%)	Average edit distance
Genome Analyzer II 2009 (2x76 cycles)			
Bustard	101,487,701	(68.12%)	0.7757
naiveBayesCall	100,003,123	(67.12%)	0.8426
AYB	103,156,920	(69.23%)	0.6422
Ibis	101,093,708	(67.85%)	0.7702
freeIbis (SVM)	102,091,337	(68.52%)	0.7205
HiSeq 2010 (1x100 cycles)			
Bustard	420,538,284	(83.71%)	0.8589
naiveBayesCall	423,616,381	(84,32%)	0.6962
AYB	431,426,132	(85.88%)	0.5148
Ibis	424,975,034	(84.59%)	0.7507
freeIbis (SVM)	426,560,342	(84.91%)	0.7449
Genome Analyzer II 2011 (2x126+7 (index) cycles)			[1] [2]
Bustard	583,348,201	(83.93%)	1.3792
naiveBayesCall	578,957,145	(83.34%)	1.4960
AYB	593,183,967	(85.52%)	1.0755
Ibis	592,929,953	(85.31%)	1.1670
freeIbis (SVM)	594,095,219	(85.48%)	1.1450
MiSeq (control sequences) 2012 (2x128+2x7 (indices) cycles)			
Bustard	273,642	(95.43%)	0.1844
Ibis	275,224	(95.41%)	0.1715
freeIbis (SVM)	278,773	(95.24%)	0.1673

Table A.1: Sequence accuracy according to basecaller for all 4 platforms from different years. Every run, with the exception of the MiSeq one, used modern human DNA as sample. The number of sequences that could be mapped back to the hg19 version of the human genome with the average edit distance to the reference is therefore reported. The percentage next to the total number of mapped sequences represent the fraction of the sequences pertaining to non-control lanes or, in the case of the 2 most recent runs, as a fraction of total number of sequences demultiplexed as belonging to the target sample. The average edit distance was computed using the NM field in the resulting BAM alignment. For every platform and various versions of the Illumina chemistry, freeIbis offers a significant improvement over Bustard in terms of sequence accuracy for having a greater number of mapped reads and a lower edit distance. For every run, the training was performed on the PhiX control sequences. For all but one of the runs, the reported number of aligned reads represents the number of human sequence reads aligning to the human genome reference. In the case of the MiSeq run, which used an ancient human DNA library with a low amount of endogenous human DNA, and which therefore had a small number of human sequences, the number of control reads aligning to the PhiX reference (provided by Illumina Inc.) is reported instead.

Basecaller:	Bustard				Ibis				freeIbis			
Genotype Quality	true pos.	false pos.	false neg.	true neg.	true pos.	false pos.	false neg.	true neg.	true pos.	false pos.	false neg.	true neg.
10	376	68	7	552,369	376	47	7	552,573	376	59	7	552,638
20	372	68	6	515,725	372	47	6	515,310	372	58	6	515,912
30	365	68	6	478,482	363	47	6	478,332	364	58	6	478,983
40	354	45	6	420,207	353	25	6	419,846	352	29	6	421,347
50	345	22	6	369,542	342	16	6	368,744	344	20	6	370,400
60	331	15	6	317,630	328	11	6	317,306	334	10	6	319,287
70	304	10	5	252,353	305	8	5	253,612	308	8	5	255,200
80	291	7	5	208,036	286	5	5	210,037	290	5	5	211,351
90	278	5	4	170,717	274	3	4	172,157	276	4	4	173,548

Table A.2: Genotype prediction accuracy according to basecaller at various genotype quality cutoffs . The accuracy of calling the genotype is present for 10 individuals, which were genotyped using Sanger sequencing depending on the basecaller used for positive calls (pos.) and negative calls (.neg). At low genotype quality cutoffs, the previous version of the software minimizes the number of false positives due to the distribution of the quality scores. At higher genotype quality cutoff levels, Ibis fails to produce a large number of correctly predicted sites like freeIbis. However, at every genotype quality cutoffs, freeIbis offers more accurately predicted sites and fewer errors than the default basecaller.

Basecaller	lane	number mapped	percentage mapped
Bustard	1	700,491	86.29%
	2	713,303	86.80%
	3	705,662	86.39%
	4	708,157	86.33%
	5	716,212	86.71%
freeIbis	1	711,741	87.93%
	2	724,318	88.41%
	3	717,236	88.21%
	4	719,325	88.10%
	5	727,228	88.59%

Table A.3: Percentage of sequences mapped for each basecaller on a run with a high error rate. Percentage and number of mapped sequences identified as controls (for this multiplexed run, identified using the index sequences). Both in terms of number and percentages, sequences basecalled using freeIbis have a greater tendency to map than the ones called with the default basecaller provided by the vendor.

Name	P7 sequence	P5 sequence	P7 index	P5 index	Name	P7 sequence	P5 sequence	P7 index	P5 index
PCR1	AATTCAA	CATCCGG	341	33	PCR51	CCTAGGT	CGTATAT	303	91
PCR2	CGCGCAG	TCATGGT	342	34	PCR52	GGATCAA	GCTAATC	304	92
PCR3	AAGGTCT	AGAACCG	343	35	PCR53	GCAAGAT	GACTTCT	305	93
PCR4	ACTGGAC	TGGAATA	344	36	PCR54	ATGGAGA	GTACTAT	306	94
PCR5	AGCAGGT	CAGGAGG	345	37	PCR55	CTCGATG	CGAGATC	307	95
PCR6	GTACCGG	AATACCT	346	38	PCR56	GCTCGAA	CGCAGCC	308	96
PCR7	GGTCAAG	CGAATGC	347	39	PCR57	AGTCAGA	CATCCGG	349	33
PCR8	AATGATG	TTCGCAA	348	40	PCR58	AACTAGA	TCATGGT	350	34
PCR9	AGTCAGA	AATTCAA	349	41	PCR59	CTATGGC	AGAACCG	351	35
PCR10	AACTAGA	CGCGCAG	350	42	PCR60	CGACGGT	TGGAATA	352	36
PCR11	CTATGGC	AAGGTCT	351	43	PCR61	AACCAAG	CAGGAGG	353	37
PCR12	CGACGGT	ACTGGAC	352	44	PCR62	CGGCGTA	AATACCT	354	38
PCR13	AACCAAG	AGCAGGT	353	45	PCR63	GCAGTCC	CGAATGC	355	39
PCR14	CGGCGTA	GTACCGG	354	46	PCR64	CTCGCGC	TTCGCAA	356	40
PCR15	GCAGTCC	GGTCAAG	355	47	PCR65	CTGCGAC	AATTCAA	357	41
PCR16	CTCGCGC	AATGATG	356	48	PCR66	ACGTATG	CGCGCAG	358	42
PCR17	CTGCGAC	AGTCAGA	357	49	PCR67	ATACTGA	AAGGTCT	359	43
PCR18	ACGTATG	AACTAGA	358	50	PCR68	CAGGAGG	CCGATTG	337	29
PCR19	ATACTGA	CTATGGC	359	51	PCR69	AATACCT	ATGCCGC	338	30
PCR20	TACTTAG	CGACGGT	360	52	PCR70	CGAATGC	CAGTACT	339	31
PCR21	AAGCTAA	AACCAAG	361	53	PCR71	TTCGCAA	AATAGTA	340	32
PCR22	GACGGCG	CGGCGTA	362	54	PCR72	ACCAACT	TCGCAGG	309	1
PCR23	AGAAGAC	GCAGTCC	363	55	PCR73	CCGGTAC	CTCTGCA	310	2
PCR24	GTCCGGC	CTCGCGC	364	56	PCR74	AACTCCG	CCTAGGT	311	3
PCR25	TTCAACC	TCAGCTT	373	65	PCR75	TTGAAGT	GGATCAA	312	4
PCR26	TTAACTC	AGAGCGC	374	66	PCR76	ACTATCA	GCAAGAT	313	5
PCR27	TAGTCTA	GCCTACG	375	67	PCR77	TTGGATC	ATGGAGA	314	6
PCR28	TGCATGA	TAATCAT	376	68	PCR78	CGACCTG	CTCGATG	315	7
PCR29	AATAAGC	AACCTGC	377	69	PCR79	TAATGCG	GCTCGAA	316	8
PCR30	AGCCTTG	GACGATT	378	70	PCR80	AGGTACC	ACCAACT	317	9
PCR31	CCAACCT	TAGGCCG	379	71	PCR81	TGCGTCC	CCGGTAC	318	10
PCR32	GCAGAAG	GGCATAG	380	72	PCR82	GAATCTC	AACTCCG	319	11
PCR33	AGAATTA	TTCAACC	381	73	PCR83	CATGCTC	TTGAAGT	320	12
PCR34	CAGCATC	TTAACTC	382	74	PCR84	ACGCAAC	ACTATCA	321	13
PCR35	TTCTAGG	TAGTCTA	383	75	PCR85	GCATTGG	TTGGATC	322	14
PCR36	CCTCTAG	TGCATGA	384	76	PCR86	GATCTCG	CGACCTG	323	15
PCR37	CCGGATA	AATAAGC	385	77	PCR87	CAATATG	TAATGCG	324	16
PCR38	GCCGCCT	AGCCTTG	386	78	PCR88	TGACGTC	AGGTACC	325	17
PCR39	AACGACC	CCAACCT	387	79	PCR89	GATGCCA	TGCGTCC	326	18
PCR40	CCAGCGG	GCAGAAG	388	80	PCR90	CAATTAC	GAATCTC	327	19
PCR41	TAGTTCC	AGAATTA	389	81	PCR91	AGATAGG	CATGCTC	328	20
PCR42	TGGCAAT	CAGCATC	390	82	PCR92	CCGATTG	ACGCAAC	329	21
PCR43	CGTATAT	TTCTAGG	391	83	PCR93	ATGCCGC	GCATTGG	330	22
PCR44	GCTAATC	CCTCTAG	392	84	PCR94	CAGTACT	GATCTCG	331	23
PCR45	GACTTCT	CCGGATA	393	85	PCR95	AATAGTA	CAATATG	332	24
PCR46	GTACTAT	GCCGCCT	394	86	PCR96	CATCCGG	TGACGTC	333	25
PCR47	CGAGATC	AACGACC	395	87	PCR97	TCATGGT	GATGCCA	334	26
PCR48	CGCAGCC	CCAGCGG	396	88	PCR98	AGAACCG	CAATTAC	335	27
PCR49	TCGCAGG	TAGTTCC	301	89	PCR99	TGGAATA	AGATAGG	336	28
PCR50	CTCTGCA	TGGCAAT	302	90	PhiX	GACGATT	GACGGCG	370	62

Table A.4: Read groups used in Chapter 4 along with the sequence of the indices and Illumina index numbers.

avg. edit distance	correctly assigned QC passed	correctly assigned QC failed	wrongly assigned QC passed	wrongly assigned QC failed	0 mismatches	1 mismatch	2 or more mismatches	error rate for QC passed
0.002408	12,374,119	29	1	0	11,962,540	405,318	6,291	0.00%
0.010169	12,373,301	847	0	1	10,725,994	1,540,905	107,250	0.00%
0.020160	12,368,861	5,277	2	9	9,305,305	2,679,637	389,207	0.00%
0.029793	12,358,306	15,809	3	31	8,105,076	3,481,942	787,131	0.00%
0.039433	12,339,811	34,251	9	78	7,048,970	4,047,523	1,277,656	0.00%
0.048776	12,311,489	62,491	12	157	6,146,987	4,410,820	1,816,342	0.00%
0.057784	12,274,074	99,708	29	338	5,379,913	4,618,279	2,375,957	0.00%
0.066878	12,221,546	151,926	42	635	4,697,957	4,713,000	2,963,192	0.00%
0.075641	12,160,346	212,680	64	1,059	4,119,163	4,712,652	3,542,334	0.00%
0.084253	12,085,534	286,819	89	1,707	3,613,460	4,648,446	4,112,243	0.00%
0.092736	11,998,912	372,511	151	2,575	3,169,831	4,535,227	4,669,091	0.00%
0.101145	11,898,460	471,721	205	3,763	2,783,384	4,381,588	5,209,177	0.00%
0.109136	11,789,483	579,097	260	5,309	2,456,736	4,212,362	5,705,051	0.00%
0.117206	11,664,964	701,423	368	7,394	2,163,380	4,017,595	6,193,174	0.00%
0.125214	11,528,595	835,066	420	10,068	1,903,980	3,813,182	6,656,987	0.00%
0.132844	11,388,127	972,122	534	13,366	1,684,998	3,609,607	7,079,544	0.00%
0.140526	11,234,310	1,122,111	698	17,030	1,486,412	3,399,036	7,488,701	0.01%
0.147897	11,076,751	1,274,925	868	21,605	1,317,939	3,198,341	7,857,869	0.01%
0.155148	10,909,794	1,435,683	1,022	27,650	1,169,888	3,005,939	8,198,322	0.01%
0.162508	10,731,576	1,607,100	1,288	34,185	1,034,347	2,807,681	8,532,121	0.01%
0.169414	10,555,336	1,775,939	1,516	41,358	921,690	2,631,870	8,820,589	0.01%
0.176525	10,362,080	1,959,683	1,757	50,629	815,864	2,452,554	9,105,731	0.02%
0.183351	10,171,497	2,139,740	2,061	60,851	727,790	2,286,576	9,359,783	0.02%
0.190047	9,979,106	2,320,527	2,437	72,079	648,330	2,128,565	9,597,254	0.02%
0.196708	9,779,898	2,506,808	2,761	84,682	577,456	1,978,848	9,817,845	0.03%
0.203133	9,581,645	2,690,309	3,087	99,108	516,142	1,837,545	10,020,462	0.03%
0.209702	9,376,786	2,878,642	3,657	115,064	460,367	1,705,867	10,207,915	0.04%
0.215989	9,170,620	3,067,416	3,966	132,147	410,283	1,583,551	10,380,315	0.04%
0.222165	8,967,960	3,249,621	4,513	152,055	368,415	1,469,445	10,536,289	0.05%
0.228286	8,764,114	3,432,621	4,981	172,433	329,986	1,363,794	10,680,369	0.06%
0.234293	8,563,485	3,610,821	5,398	194,445	294,948	1,261,862	10,817,339	0.06%
0.240130	8,361,952	3,788,095	6,069	218,033	265,964	1,172,482	10,935,703	0.07%
0.245970	8,161,182	3,961,789	6,572	244,606	238,975	1,087,117	11,048,057	0.08%
0.251751	7,960,228	4,134,271	7,191	272,459	214,151	1,005,856	11,154,142	0.09%
0.257444	7,757,196	4,307,519	8,025	301,409	191,882	932,307	11,249,960	0.10%
0.263088	7,559,422	4,472,017	8,819	333,891	172,448	863,078	11,338,623	0.12%
0.268502	7,367,339	4,631,315	9,260	366,235	155,434	799,425	11,419,290	0.13%
0.273824	7,179,504	4,782,441	10,071	402,133	140,083	742,376	11,491,690	0.14%
0.279133	6,990,316	4,934,765	10,947	438,121	126,579	687,184	11,560,386	0.16%
0.284395	6,809,979	5,074,764	11,593	477,813	114,574	637,054	11,622,521	0.17%
0.289590	6,621,120	5,222,775	12,424	517,830	103,693	590,018	11,680,438	0.19%
0.294594	6,445,396	5,355,157	13,339	560,257	93,404	546,745	11,734,000	0.21%
0.299673	6,266,929	5,488,381	14,353	604,486	84,181	507,277	11,782,691	0.23%
0.304587	6,096,095	5,611,609	15,211	651,234	76,847	470,773	11,826,529	0.25%
0.309421	5,928,675	5,731,211	16,346	697,917	69,722	435,871	11,868,556	0.27%
0.314206	5,766,160	5,842,967	17,101	747,921	62,569	405,096	11,906,484	0.30%
0.318938	5,606,108	5,949,529	18,190	800,322	57,340	375,096	11,941,713	0.32%
0.323720	5,447,198	6,053,435	19,140	854,376	51,726	347,396	11,975,027	0.35%
0.328343	5,288,741	6,156,930	20,350	908,128	47,175	322,750	12,004,224	0.38%
0.332825	5,143,582	6,244,076	21,403	965,088	42,879	300,473	12,030,797	0.41%
0.337214	4,997,095	6,334,509	22,246	1,020,299	38,762	278,974	12,056,413	0.44%

Table A.5: Tally of the mismatches found in the indices at various levels of simulated error rates. The remaining columns present the number of correctly and incorrectly classified sequences by deML, both for those that passed and failed default thresholds. The columns with 0 and 1 mismatch represent the limit of the default software provided by Illumina.

Sample ID	HaploGrep Quality	Predicted Haplogroup	Sample ID	HaploGrep Quality	Predicted Haplogroup
JQ703873	94.1	A2i	GQ301880	87.1	M22b
KC711027	89.5	B2a1	FJ543105	96.9	M23
AP008393	95.1	B4c1b1a	KJ154498	96.5	M27a
DQ834259	63.3	B4c1b2c	KJ154685	86.5	M27b
AP008788	91.2	B4f	KJ154771	96.0	M27c
KF540901	83.2	B5a2a2	KJ154541	93.7	M28a
AP008273	96.7	B5b1a2	DQ137407	98.7	M29a
FJ951464	95.3	C5b1a	KC990685	75.0	M2a1a
FJ951600	85.9	D4	EU443449	98.5	M2b3a
FJ858886	89.6	D4b1	KC911426	91.1	M2c
FJ168748	97.0	D4h3a9	AY950293	96.7	M31a1b
FJ951465	100.0	D5a2a2	GQ389779	98.4	M32c
KJ154788	98.8	E1b1	HM030510	91.0	M33b1
KF849964	94.4	F1a1d	JX462713	97.0	M33d
KC252477	100.0	F3b1a	AY922304	98.3	M34a1a
KF451331	93.9	F4a2	FJ383405	89.4	M38b
KF148403	92.9	G2a1	KC990670	72.0	M42'74
HM454265	92.2	I1a	DQ404443	83.3	M42a
JQ797764	94.8	J1b1a2b	FJ380216	82.3	M42b
JQ797929	96.2	J2a2b	FJ383746	89.4	M42b1a
JQ702671	95.9	K1a1b1a	KC990667	63.0	M5
KJ185548	98.9	L0a1b1a1	JX289098	91.5	M50a2
EF184602	94.8	L0a2	GQ301882	97.2	M51a2
KC533465	87.7	L0a2a2a	FJ383491	94.0	M52b1a
KJ185995	84.5	L0a'b	FJ383439	94.6	M53b
EU092936	94.1	L0b	KC896622	97.3	M55
KF672800	89.8	L0b	FJ383762	87.9	M57a
KC346214	97.6	L0d1b2b2a	JX289110	81.9	M58
KC533490	94.4	L0d1c1a	DQ834260	79.4	M59
KC346193	98.9	L0d2a1c	KC505104	87.6	M59
KC345912	98.9	L0d2b1a1a	JQ446396	83.7	M5a1a
KC346210	97.3	L0d2c1a	KC990648	72.3	M5a2a1a
KC533475	98.8	L0d3b	FJ383550	84.1	M5b2b1

Table A.6: Mitochondrial sources of contamination provided with schmutzi.

Sample ID	HaploGrep Quality	Predicted Haplogroup	Sample ID	HaploGrep Quality	Predicted Haplogroup
EF184595	82.6	L0f	FJ544233	96.6	M62b2
EF184598	87.4	L0f	KC887484	96.9	M68a1a
EU092870	90.2	L0f1	HM596653	79.1	M69
KJ185400	88.1	L0f1	FJ383302	93.7	M6a1a
EF184596	85.3	L0f2a	GQ119039	94.2	M73a
EF184599	91.0	L0f2a	HM030520	88.0	M74b
EF184597	97.8	L0f2a1	HM030540	90.6	M75
EU092786	100.0	L0f2b	HM030525	81.1	M76
KC345794	100.0	L0k1a2	AP009443	98.0	M7a1b2
KM101649	98.4	L1b1a4	KF540526	99.0	M7b1a1i
KC533514	87.4	L1c1	KC252522	88.1	M8a3a
HM771141	98.1	L1c1a1a1a	JX289130	89.1	M91a
JX303768	90.5	L1c1a2	KC887486	100.0	M91b
KJ185481	95.0	L1c1b	JN048455	60.7	N10
JQ701901	92.3	L1c1c	HM030542	84.3	N10a
HM771166	91.9	L1c1d	HM030500	98.8	N10b
KJ185466	92.1	L1c2a2	KF540803	82.6	N11a
EU092718	94.3	L1c3a1a	GU733776	97.9	N11b
KC257334	97.1	L1c3b2	JN226143	85.9	N13
HM771117	97.8	L1c4a	JQ705527	94.9	N1a1a1a2
JX303797	98.5	L1c5	JQ704073	94.2	N1a3a1
EU273489	90.0	L1c6	EF661011	97.7	N1b1a2
JQ044836	90.4	L1c6	KC867135	99.7	N3a
EF184581	77.0	L2'3'4'6+	GU480021	76.4	N5a
JQ045090	99.2	L2a1f	KC505118	97.4	N7a1
KJ185427	88.4	L2a5	HM030548	89.2	N8
JQ701833	97.5	L2b1a3	AY289059	74.4	O
JQ044878	99.2	L2c2b2	KC993994	98.8	P10
KJ185421	98.1	L2d1a	EF061154	78.0	P3b1
KJ185902	95.6	L2e1a	EF061158	92.6	P4a
DQ341081	96.5	L3a1b	AY289064	80.3	P4b
JN655803	79.0	L3a+709	AY289053	73.8	P6
KJ185776	97.0	L3b1a1a	KF451181	99.1	Q1c

Table A.7: Mitochondrial sources of contamination provided with schmutzi (cont.).

Sample ID	HaploGrep Quality	Predicted Haplogroup	Sample ID	HaploGrep Quality	Predicted Haplogroup
DQ341074	94.0	L3c	KJ154822	95.0	Q2a
KC622102	100.0	L3d3a1a	AY289079	85.1	Q3a
JX303776	96.0	L3e1	JF824990	87.0	R11b
EU092895	92.5	L3e3b	AY714045	96.6	R1a1a
JN655842	90.1	L3f1a1	KC911319	77.7	R2
JQ045052	92.2	L3f1b1a	AY963584	93.6	R21
JN655832	83.7	L3f2a1	GU170818	86.7	R30a1b1
EU092877	84.6	L3f2b	AY714050	88.7	R30b1
AF347000	96.3	L3h1a2a1	FJ004826	98.2	R31a1
JN655838	94.8	L3h1b1a	AY714046	92.8	R31b
JN655820	97.7	L3h2	FJ004811	99.0	R7b1a1
JN655789	95.0	L3k1	JF742196	90.7	R8a1a1a2
JN655802	88.7	L3x2a	DQ404441	89.4	S1a
FJ460531	95.8	L4a1	AY289067	96.6	S3
JQ044848	93.6	L4b1a	JQ705673	94.5	T2e
EU092951	96.9	L4b2a2b	KC911502	90.2	U1a1a
EF556173	97.8	L5a1a	JQ705704	92.5	U1b1
KC911364	94.5	L5b1a	KC533515	96.4	U2a2
EU092802	93.8	L6a	KF450851	95.4	U2b2
FJ770941	82.5	M	JX984460	83.0	U2c1
KF451676	92.5	M10a1+16129	KC990647	62.3	U2c1
KC709481	92.8	M11c	JQ706067	92.4	U2d2
KJ446520	96.2	M12a1a2	KJ445816	91.0	U2e1h
KF451769	81.6	M13	JX153094	87.3	U3a2
FJ544230	95.6	M13a2	JQ704121	96.7	U4c1a
JX289092	90.0	M13c	GU296627	95.0	U5b2b1a2
EF495222	77.7	M14	KC152579	91.4	U6a5
GU810076	92.3	M17a	KC911508	92.3	U7a3
DQ779925	93.8	M1a3b	JX273294	85.5	U8b1a2
HM030505	96.9	M20	KC911536	83.4	U8b1a2
GQ119046	84.2	M21a	AY339492	96.6	W1a
JF739541	87.9	M21b1a	JN415482	90.1	X2b+226
JX289109	94.9	M21b2			

Table A.8: Mitochondrial sources of contamination provided with schmutzi (cont.).

type of sample	Genbank accession
Revised Cambridge Reference Sequence (rCRS)	NC_012920
present-day human	AF347008
present-day human	AY195788
present-day human	AF347015
present-day human	AF347014
present-day human	AY289070
present-day human	AF381982
present-day human	AY195773
present-day human	AY195779
present-day human	AY882391
present-day human	AY882415
present-day human	AY882404
present-day human	AF346963
present-day human	AY882386
present-day human	AY289093
present-day human	AF347007
present-day human	AY289095
present-day human	AY289060
present-day human	AY195752
present-day human	AY882417
present-day human	AY195789
Denisovan phalanx	NC_013993
Sima de los Huesos	NC_023100
Neanderthal Mezmaiskaya1	FM865411
Neanderthal Feldhofer1	FM865407
Neanderthal Feldhofer2	FM865408
Neanderthal Vindija33.25	FM865410
Neanderthal Vindija33.16	AM948965
Neanderthal Sidron	FM865409
Neanderthal Altai	KC879692
Pan paniscus	NC_001644

Table A.9: Description of samples used in the maximum likelihood tree with accession identifier

Contamination rate	schmutzi		PMDtools and hstlib
	Default parameters	Multiple Contaminants	
0.01	16570/0/0 †	16570/0/0	16566/1/3
0.05	16570/0/0 †	16570/0/0	16566/1/3
0.10	16570/0/0 †	16570/0/0	16566/1/3
0.15	16570/0/0	16570/0/0	16566/1/3
0.20	16570/0/0	16567/1/2	16566/1/3
0.25	16570/0/0	16567/1/2	16566/1/3
0.30	16570/0/0	16567/1/2	16566/1/3
0.35	16570/0/1	16567/1/3	16566/1/3
0.40	16569/1/1	16567/1/3	16566/1/3
0.45	16569/1/1	16547/21/3	16566/1/3
0.50	16569/1/1	16547/21/3	16566/1/3
0.55	16569/1/1	16547/21/3	16566/1/3
0.60	16570/0/2	16547/21/3	16566/1/3
0.65	16570/0/2	16547/21/3	16566/1/3
0.70	16570/0/2	16547/21/3	16566/1/3
0.75	16569/1/2 ‡	16547/21/3	16566/1/3
0.80	16569/1/2 ‡	16547/21/3	16566/1/3
0.85	NA/NA/NA *	16547/21/3	16564/3/3
0.90	NA/NA/NA *	16547/21/3	16564/3/3
0.95	NA/NA/NA *	16547/21/3	16560/7/3

Table A.10: Edit distance to the original endogenous genome using an early modern human genome and a double-stranded protocol. The original endogenous genome had 16547 matches, 21 mismatches and 3 indels to the contaminant. A † symbol on a data point indicates that the algorithm did not continue on to the second iteration due to a lack of detected contaminant at low simulated contamination rates, thus the results presented are the ones from the first iteration. Sets marked with a ‡ indicate that the predicted contaminant was used as a contaminant source. For very hard targets (e.g., EHM with around 90% contamination), the workflow provided by the wrapper script diverges even with the option of using the contaminant source. For such hard targets, manual intervention would be required and data that caused this type of problem are marked with an $*$.

Contamination rate	schmutzi		mpileup consensus on deaminated fragments
	Default parameters	Multiple Contaminants	
0.01	16570/0/1 †	16570/0/0	16567/1/2
0.05	16570/0/0 †	16570/0/0	16567/1/2
0.10	16570/0/0 †	16570/0/0	16567/1/2
0.15	16570/0/0	16570/0/0	16567/1/2
0.20	16570/0/0	16567/1/2	16567/1/2
0.25	16570/0/0	16567/1/2	16567/1/2
0.30	16570/0/0	16567/1/2	16567/1/2
0.35	16570/0/0	16567/1/3	16567/1/2
0.40	16569/1/1	16567/1/3	16567/1/2
0.45	16569/1/1	16548/20/3	16567/1/2
0.50	16570/0/2	16547/21/3	16567/1/2
0.55	16569/1/1	16547/21/3	16567/1/2
0.60	16570/0/2	16547/21/3	16566/2/2
0.65	16570/0/2	16547/21/3	16566/2/2
0.70	16570/0/2	16547/21/3	16562/6/2
0.75	16570/0/2	16547/21/3	16562/6/2
0.80	16570/0/2 ‡	16547/21/3	16561/7/2
0.85	16569/1/2 ‡	16547/21/3	16558/7/5
0.90	16569/1/2 ‡	16547/21/3	16561/7/2
0.95	16568/2/2 ‡	16547/21/3	16553/10/7

Table A.11: Edit distance to the original endogenous genome using an early modern human genome and a single-stranded protocol. The original endogenous genome had 16547 matches, 21 mismatches and 3 indels to the contaminant. A † symbol on a data point indicates that the algorithm did not continue on to the second iteration due to a lack of detected contaminant at low simulated contamination rates, thus the results presented are the ones from the first iteration. Sets marked with a ‡ indicate that the predicted contaminant was used as a contaminant source.

Contamination rate	schmutzi		PMDtools and hstlib
	Default parameters	Multiple Contaminants	
0.01	16565/0/0	16565/0/0	16561/2/6
0.05	16565/0/0 †	16565/0/0	16561/2/6
0.10	16565/0/0	16565/0/0	16561/2/6
0.15	16565/0/0	16565/0/0	16560/3/6
0.20	16565/0/0	16564/1/0	16560/3/6
0.25	16565/0/0	16562/2/1	16558/5/6
0.30	16564/1/0	16559/5/5	16558/5/6
0.35	16564/1/0	16550/3/28	16556/7/6
0.40	16564/1/0	16542/22/6	16555/8/6
0.45	16564/1/0	16355/209/6	16553/10/6
0.50	16563/2/0	16355/209/6	16553/10/6
0.55	16564/1/0	16355/209/6	16554/9/6
0.60	16563/2/0	16355/209/6	16551/12/6
0.65	16563/1/1	16355/209/6	16551/12/6
0.70	16562/1/2	16355/209/6	16548/15/6
0.75	16563/1/1	16355/209/6	16546/17/6
0.80	16561/2/2 ‡	16355/209/6	16545/18/6
0.85	16563/1/1 ‡	16355/209/6	16544/19/6
0.90	16561/3/1 ‡	16355/209/6	16539/24/6
0.95	16550/15/7 ‡	16355/209/6	16532/31/6

Table A.12: Edit distance to the original endogenous genome using a Neandertal genome and a double-stranded protocol. The original endogenous genome had 16355 matches, 209 mismatches and 6 indels to the contaminant. A † symbol on a data point indicates that the algorithm did not continue on to the second iteration due to a lack of detected contaminant at low simulated contamination rates, thus the results presented are the ones from the first iteration. Sets marked with a ‡ indicate that the predicted contaminant was used as a contaminant source.

Contamination rate	schmutzi		mpileup consensus on deaminated fragments
	Default parameters	Multiple Contaminants	
0.01	16565/0/0	16565/0/0	16565/0/0
0.05	16565/0/0	16565/0/0	16565/0/0
0.10	16565/0/0	16565/0/0	16565/0/0
0.15	16565/0/0	16564/1/0	16565/0/0
0.20	16565/0/0	16564/1/0	16565/0/0
0.25	16565/0/0	16562/2/1	16565/0/0
0.30	16564/1/0	16559/5/5	16565/0/0
0.35	16564/1/0	16550/3/28	16565/0/0
0.40	16564/1/0	16549/15/6	16565/0/0
0.45	16564/1/0	16356/208/6	16565/0/0
0.50	16564/1/0	16355/209/6	16565/0/0
0.55	16564/1/0	16355/209/6	16564/0/1
0.60	16564/1/0	16355/209/6	16563/1/1
0.65	16563/1/1	16355/209/6	16560/4/1
0.70	16564/1/0	16355/209/6	16556/8/1
0.75	16563/1/1	16355/209/6	16546/7/23
0.80	16563/1/1	16355/209/6	16548/15/6
0.85	16564/1/0	16355/209/6	16544/20/1
0.90	16563/2/0	16355/209/6	16536/25/4
0.95	16558/7/0	16355/209/6	16524/33/12

Table A.13: Edit distance to the original endogenous genome using a Neandertal genome and a single-stranded protocol. The original endogenous genome had 16355 matches, 209 mismatches and 6 indels to the contaminant.

Contamination rate	schmutzi		PMDtools and hstlib
	Default parameters	Multiple Contaminants	
0.01	16569/1/0	16569/1/0	16557/2/19
0.05	16569/1/0	16569/1/1	16557/2/19
0.10	16569/1/1	16566/2/2	16557/2/19
0.15	16569/1/1	16566/2/5	16557/2/19
0.20	16568/2/0	16559/3/11	16557/2/19
0.25	16567/3/0	16558/4/12	16557/2/19
0.30	16567/2/1	16554/8/14	16557/2/19
0.35	16567/2/1	16552/9/17	16555/4/19
0.40	16567/3/0	16515/46/17	16554/5/19
0.45	16567/3/2	16174/387/17	16554/5/19
0.50	16568/2/0	16174/387/17	16551/8/19
0.55	16568/2/2	16174/387/17	16550/9/19
0.60	16566/2/4	16174/387/17	16549/10/19
0.65	16566/2/4	16174/387/17	16547/12/19
0.70	16566/2/4	16174/387/17	16544/15/19
0.75	16566/2/4	16174/387/17	16541/18/19
0.80	16566/4/2	16174/387/17	16534/25/19
0.85	16567/3/4	16174/387/17	16532/27/19
0.90	16565/5/7 ‡	16174/387/17	16529/31/17
0.95	NA/NA/NA *	16174/387/17	16512/48/17

Table A.14: Edit distance to the original endogenous genome using a Denisovan genome and a double-stranded protocol. The original endogenous genome had 16174 matches, 387 mismatches and 17 indels to the contaminant. Sets marked with a ‡ indicate that the predicted contaminant was used as a contaminant source. For very hard targets (e.g. 95% contamination), the workflow provided by the wrapper script diverges even with the option of using the contaminant source. For such hard targets, manual intervention would be required and data that caused this type of problem are marked with an $*$.

Contamination rate	schmutzi		mpileup consensus on deaminated fragments
	Default parameters	Multiple Contaminants	
0.01	16569/1/0	16569/1/0	16560/1/9
0.05	16569/1/0	16569/1/0	16560/1/9
0.10	16569/1/0	16566/2/2	16560/1/9
0.15	16569/1/0	16566/2/5	16560/1/9
0.20	16567/3/0	16559/3/11	16560/2/8
0.25	16568/2/0	16559/3/14	16560/2/8
0.30	16567/3/0	16555/7/14	16559/3/8
0.35	16567/3/0	16552/9/17	16560/2/8
0.40	16567/3/0	16542/19/17	16566/1/3
0.45	16567/3/0	16174/387/17	16560/1/9
0.50	16567/2/1	16174/387/17	16560/2/8
0.55	16568/2/0	16174/387/17	16559/3/8
0.60	16567/3/2	16174/387/17	16561/1/8
0.65	16567/3/2	16174/387/17	16560/1/9
0.70	16568/2/2	16174/387/17	16557/5/8
0.75	16569/1/2	16174/387/17	16558/9/3
0.80	16568/2/2	16174/387/17	16549/13/8
0.85	16569/1/4	16174/387/17	16538/23/12
0.90	16569/1/2	16174/387/17	16524/37/12
0.95	16563/7/7	16174/387/17	16505/56/12

Table A.15: Edit distance to the original endogenous genome using a Denisovan genome and a single-stranded protocol. The original endogenous genome had 16174 matches, 387 mismatches and 17 indels to the contaminant.

Contamination rate	MIA (with -k 12)		
	EMH	Neanderthal	Denisovan
0.01	16570/0/0	16523/1/4	16560/0/5
0.05	16565/1/3	16549/10/4	16565/0/6
0.10	16565/1/3	16545/10/4	16565/0/6
0.15	16565/1/3	16528/10/4	16548/0/12
0.20	16562/1/3	16360/11/4	16211/0/14
0.25	16547/1/3	16355/13/5	16175/2/14
0.30	16547/1/3	16355/18/5	16175/5/15
0.35	16547/1/3	16355/18/6	16174/19/17
0.40	16547/1/3	16355/20/6	16174/27/17
0.45	16547/2/3	16355/29/6	16174/30/17
0.50	16547/3/3	16355/44/6	16174/41/17
0.55	16547/3/3	16355/58/6	16174/65/17
0.60	16547/3/3	16355/109/6	16174/106/17
0.65	16547/3/3	16355/195/6	16174/220/17
0.70	16547/16/3	16355/209/6	16174/386/17
0.75	16547/20/3	16355/209/6	16174/387/17
0.80	16547/21/3	16355/209/6	16174/387/17
0.85	16547/20/3	16355/209/6	16174/387/17
0.90	16547/21/3	16355/209/6	16174/387/17
0.95	16547/21/3	16355/209/6	16174/387/17

Table A.16: Edit distance of the consensus genome predicted using MIA to the original endogenous genome when using a double-strand protocol. The contaminant genome had 16547 matches, 21 mismatches and 3 indels to the early modern human genome, 16355 matches, 209 mismatches and 6 indels to the Neandertal genome and 16174 matches, 387 mismatches and 17 indels to the Denisova genome.

Contamination rate	schmutzi with default parameters		
	EMH	Neanderthal	Denisovan
0.01	16538/20/44 †	16442/127/40	16311/254/150
0.05	16537/21/34 †	16358/211/62 †	16563/2/13
0.10	16544/25/11 †	16567/0/2	16566/0/6
0.15	16568/1/2	16567/0/2	16566/3/2
0.20	16568/1/2	16569/0/0	16568/0/1
0.25	16568/1/2	16569/0/0	16569/0/0
0.30	16569/0/1	16569/0/0	16569/0/0
0.35	16569/0/0	16569/0/0	16569/0/0
0.40	16569/0/0	16569/0/0	16569/0/0
0.45	16569/0/0	16569/0/0	16569/0/0
0.50	16569/0/0	16569/0/0	16569/0/0
0.55	16569/0/0	16569/0/0	16569/0/0
0.60	16569/0/0	16569/0/0	16569/0/0
0.65	16569/0/0	16569/0/0	16569/0/0
0.70	16569/0/0	16569/0/0	16569/0/0
0.75	16569/0/0 ‡	16569/0/0	16569/0/0
0.80	16569/0/0 ‡	16569/0/0 ‡	16569/0/0
0.85	NA/NA/NA *	16569/0/0 ‡	16569/0/0
0.90	NA/NA/NA *	16569/0/0 ‡	16569/0/0 ‡
0.95	NA/NA/NA *	16569/0/0 ‡	NA/NA/NA *

Table A.17: Edit distance of the predicted contaminant genome to the original contaminant genome when using a double-strand protocol. The contaminant genome had 16547 matches, 21 mismatches and 3 indels to the early modern human (EMH) genome, 16355 matches, 209 mismatches and 6 indels to the Neandertal genome and 16174 matches, 387 mismatches and 17 indels to the Denisova genome. A † symbol on a data point indicates that the algorithm did not continue on to the second iteration due to a lack of detected contaminant at low simulated contamination rates, thus the results presented are the ones from the first iteration. Sets marked with a ‡ indicate that the predicted contaminant was used as a contaminant source. For very hard targets (e.g., EHM with around 90% contamination), the workflow provided by the wrapper script diverges even with the option of using the contaminant source. For such hard targets, manual intervention would be required and data that caused this type of problem are marked with an $*$.

Contamination rate	schmutzi with default parameters		
	EMH	Neanderthal	Denisovan
0.01	16537/21/35 †	16435/134/12	16196/369/114
0.05	16538/20/35 †	16567/0/2	16565/1/13
0.10	16537/21/35 †	16567/0/2	16566/0/6
0.15	16568/1/2	16568/0/1	16566/3/2
0.20	16568/1/2	16569/0/0	16566/3/0
0.25	16568/1/2	16569/0/0	16569/0/0
0.30	16569/0/0	16569/0/0	16569/0/0
0.35	16569/0/0	16569/0/0	16569/0/0
0.40	16569/0/0	16569/0/0	16569/0/0
0.45	16569/0/0	16569/0/0	16569/0/0
0.50	16569/0/0	16569/0/0	16569/0/0
0.55	16569/0/0	16569/0/0	16569/0/0
0.60	16569/0/0	16569/0/0	16569/0/0
0.65	16569/0/0	16569/0/0	16569/0/0
0.70	16569/0/0	16569/0/0	16569/0/0
0.75	16569/0/0	16569/0/0	16569/0/0
0.80	16569/0/0 ‡	16569/0/0	16569/0/0
0.85	16569/0/0 ‡	16569/0/0	16569/0/0
0.90	16569/0/0 ‡	16569/0/0	16569/0/0
0.95	16569/0/0 ‡	16569/0/0	16569/0/0

Table A.18: Edit distance of the predicted contaminant genome to the original contaminant genome when using a single-strand protocol. The contaminant genome had 16547 matches, 21 mismatches and 3 indels to the early modern human genome, 16355 matches, 209 mismatches and 6 indels to the Neandertal genome and 16174 matches, 387 mismatches and 17 indels to the Denisova genome. A † symbol on a data point indicates that the algorithm did not continue on to the second iteration due to a lack of detected contaminant at low simulated contamination rates, thus the results presented are the ones from the first iteration. Sets marked with a ‡ indicate that the predicted contaminant was used as a contaminant source.

deam. rates (%)	subsampling fraction									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
1	99.5±0.5	98.5±0.5	74.0±2.5	64.5±1.5	56.0±3.0	47.0±2.5	63.5±2.0	64.5±2.0	73.0±1.0	70.5±1.0
5	52.5±2.5	62.0±1.5	56.5±1.5	55.5±1.5	54.0±1.0	57.5±1.0	55.5±1.0	58.0±0.5	55.5±0.5	55.5±0.5
10	65.0±1.0	63.0±1.0	59.0±1.0	56.5±1.0	56.0±0.5	54.0±0.5	53.5±0.5	52.5±0.5	52.0±0.5	52.0±0.5
15	50.0±1.0	52.5±0.5	53.5±0.5	53.5±0.5	52.5±0.5	52.5±0.5	52.5±0.5	53.5±0.5	51.5±0.5	51.0±0.5
20	54.5±1.0	55.0±1.0	55.5±0.5	55.0±0.5	55.5±0.5	54.5±0.5	54.0±0.5	54.0±0.5	54.5±0.5	54.0±0.5
25	51.5±1.0	54.5±1.0	54.0±0.5	53.5±0.5	53.5±0.5	52.5±0.5	52.5±0.5	52.0±0.5	51.5±0.5	52.0±0.5
30	53.0±1.0	52.0±0.5	50.5±0.5	50.0±0.5	50.0±0.5	50.0±0.5	50.0±0.5	49.5±0.5	49.0±0.5	49.5±0.5
35	54.5±0.5	52.5±0.5	51.0±0.5	51.5±0.5	52.5±0.5	52.0±0.5	52.5±0.5	52.5±0.5	52.5±0.5	53.0±0.5
40	51.5±0.5	52.0±0.5	52.0±0.5	52.0±0.5	52.0±0.5	52.5±0.5	52.5±0.5	52.0±0.5	52.0±0.5	52.5±0.5
45	50.0±0.5	51.5±0.5	52.0±0.5	52.0±0.5	52.5±0.5	52.5±0.5	52.5±0.5	52.5±0.5	52.0±0.5	52.5±0.5
50	52.0±0.5	52.5±0.5	52.0±0.5	52.0±0.5	52.0±0.5	52.0±0.5	52.0±0.5	52.0±0.5	52.0±0.5	52.0±0.5
55	50.5±0.5	52.5±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.5±0.5
60	53.0±0.5	53.0±0.5	53.0±0.5	53.5±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5
65	53.5±0.5	53.0±0.5	52.0±0.5	52.5±0.5	52.5±0.5	52.5±0.5	52.0±0.5	52.5±0.5	52.5±0.5	52.5±0.5
70	53.5±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5
75	53.0±0.5	52.5±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5
80	53.5±0.5	54.0±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.0±0.5
85	52.5±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5	53.0±0.5
90	53.0±0.5	53.0±0.5	53.0±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.0±0.5	53.5±0.5
95	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5	53.5±0.5

Table A.19: Contamination estimate based on deamination patterns as a fraction of amount of data and deamination rates for datasets with a simulated contamination rate of 50%. The "subsampling fraction", is the fraction of fragments from the original 1M dataset that were used in the subsample.

endog. deam. rates (%)	contaminant deamination rates (%)									
	0	1	2	3	4	5	6	7	8	9
1	70.5±1.0	73.0±0.5	64.5±0.5	61.0±0.5	58.0±0.5	54.0±0.5	53.5±0.5	49.0±0.5	53.0±0.5	50.5±0.5
5	55.5±0.5	38.0±0.5	30.5±0.5	28.0±1.0	30.5±1.0	23.5±0.5	26.5±0.5	22.0±0.5	19.5±0.5	26.0±0.5
10	52.0±0.5	32.0±0.5	32.0±0.5	15.5±1.0	21.5±0.5	18.0±0.5	19.5±0.5	11.5±0.5	13.5±1.0	12.5±0.5
15	51.0±0.5	30.5±0.5	32.0±0.5	22.5±0.5	21.5±0.5	15.5±0.5	15.5±0.5	13.5±0.5	14.5±0.5	10.0±0.5
20	54.0±0.5	38.0±0.5	34.0±0.5	28.5±0.5	28.0±0.5	27.0±0.5	21.0±0.5	19.0±0.5	17.5±0.5	14.0±0.5
25	52.0±0.5	39.0±0.5	33.0±0.5	31.5±0.5	29.5±0.5	26.0±0.5	23.0±0.5	21.0±0.5	14.5±0.5	14.5±0.5
30	49.5±0.5	41.0±0.5	37.5±0.5	34.0±0.5	30.5±0.5	29.5±0.5	26.0±0.5	23.0±0.5	19.5±0.5	20.5±0.5
35	53.0±0.5	43.5±0.5	41.5±0.5	38.0±0.5	34.5±0.5	34.0±0.5	31.0±0.5	29.0±0.5	25.5±0.5	23.0±0.5
40	52.5±0.5	44.0±0.5	41.0±0.5	39.5±0.5	37.0±0.5	34.5±0.5	32.0±0.5	30.5±0.5	28.0±0.5	26.0±0.5
45	52.5±0.5	45.0±0.5	42.5±0.5	41.5±0.5	38.5±0.5	36.5±0.5	36.0±0.5	34.0±0.5	32.0±0.5	29.5±0.5
50	52.0±0.5	45.0±0.5	43.0±0.5	41.0±0.5	39.5±0.5	38.5±0.5	35.5±0.5	35.0±0.5	33.0±0.5	31.5±0.5
55	53.5±0.5	46.5±0.5	45.0±0.5	43.0±0.5	41.0±0.5	40.0±0.5	38.0±0.5	36.5±0.5	35.5±0.5	33.5±0.5
60	53.0±0.5	46.5±0.5	45.5±0.5	43.5±0.5	42.0±0.5	40.0±0.5	39.0±0.5	37.0±0.5	36.0±0.5	34.0±0.5
65	52.5±0.5	47.5±0.5	45.5±0.5	44.0±0.5	43.0±0.5	41.5±0.5	40.0±0.5	38.5±0.5	37.5±0.5	36.0±0.5
70	53.0±0.5	47.5±0.5	46.0±0.5	45.0±0.5	43.5±0.5	42.5±0.5	41.0±0.5	39.5±0.5	38.0±0.5	37.5±0.5
75	53.0±0.5	48.0±0.5	47.0±0.5	46.0±0.5	44.0±0.5	43.0±0.5	41.5±0.5	40.5±0.5	39.5±0.5	38.5±0.5
80	53.0±0.5	48.5±0.5	47.5±0.5	46.5±0.5	45.5±0.5	43.5±0.5	43.0±0.5	41.5±0.5	41.0±0.5	39.5±0.5
85	53.0±0.5	49.0±0.5	47.5±0.5	46.5±0.5	45.5±0.5	44.0±0.5	43.0±0.5	42.0±0.5	40.5±0.5	39.5±0.5
90	53.5±0.5	49.0±0.5	48.0±0.5	47.0±0.5	46.0±0.5	45.0±0.5	44.0±0.5	42.5±0.5	41.5±0.5	40.5±0.5
95	53.5±0.5	49.5±0.5	48.5±0.5	47.5±0.5	46.0±0.5	45.5±0.5	44.5±0.5	43.5±0.5	42.0±0.5	41.0±0.5

Table A.20: Effect of having deamination for contaminant fragments on the contamination estimate at various deamination rates for endogenous fragments. The original simulated contamination rate was 50%.

sample type	deamination rates for				χ^2 test				contDeam cont.
	5'end		3'end [3]		5'end		3'end		est. (%)
	3': C→T	3': C→C	5': C→T	5': C→C	χ^2	p-value	χ^2	p-value	patterns
Af. Gora	0.0510778	0.0424544	0.0385227	0.0319476	7.2993	0.006898	6.4212	0.01128	0.0±0.5
Altai Neand.	0.079841	0.0677685	0.305404	0.269203	37.0381	1.158e-09	33.8252	6.029e-09	11.5±0.5
Denisovan	0.0933588	0.0899034	0.515545	0.50517	2.5505	0.1103	2.6116	0.1061	2.5±1.0
Loschbour	0.0846782	0.0784448	0.372735	0.353486	5.0888	0.02408	5.5029	0.01898	4.5±1.0
Mal'ta	0.0297943	0.0291137	0.0337034	0.0329365	0.111	0.7391	0.2124	0.6449	0.0±0.5

Table A.21: Independence of deamination rates for 5' and 3' ends of aDNA fragments for various empirical datasets with low levels of present-day human contamination. Two by two contingency χ^2 tests were used with 1 degree of freedom. The absence of independence between deamination rates at both ends for the Altai Neanderthal leads to an overestimate of the endogenous deamination rate and, as a consequence, of contamination.

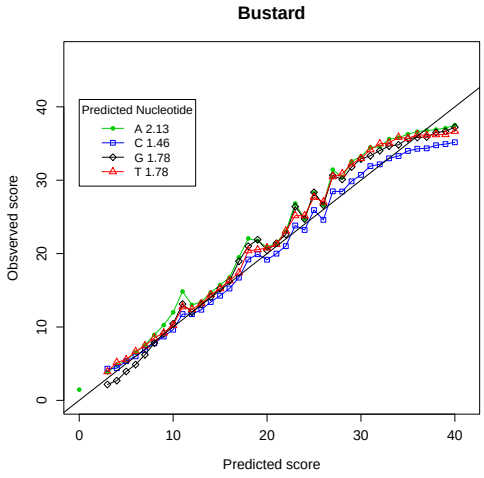
position	reference base	diagnostic base	predicted base with max. likelihood	quality on a PHRED scale	predicted = diagnostic ?
73	A	G	G	1126.34	yes
151	C	T	T	206.893	yes
263	A	G	G	664.649	yes
709	G	A	A	1073.22	yes
750	A	G	G	1293.59	yes
930	G	A	A	309.252	yes
1438	A	G	G	994.094	yes
1888	G	A	A	284.227	yes
2706	A	G	G	244.6	yes
4216	T	C	C	96.8476	yes
4769	A	G	G	709.473	yes
4917	A	G	G	252.591	yes
5147	G	A	A	241.557	yes
7028	C	T	T	1151.05	yes
8697	G	A	A	40.3619	yes
8860	A	G	G	1082.47	yes
10463	T	C	T	65.6473	no
10750	A	G	G	899.204	yes
11251	A	G	G	411.396	yes
11719	G	A	A	1096.27	yes
11812	A	G	G	305.569	yes
13368	G	A	A	412.61	yes
14233	A	G	G	232.08	yes
14766	C	T	T	935.827	yes
14905	G	A	A	396.984	yes
15326	A	G	G	1146.32	yes
15452	C	A	A	308.629	yes
15607	A	G	G	297.883	yes
15928	G	A	A	80.2745	yes
16126	T	C	C	28.9019	yes
16294	C	T	T	226.685	yes
16296	C	T	T	212.763	yes
16304	T	C	C	210.421	yes

Table A.22: Predicted contaminant from the Mezmaiskaya sample B9687 with the diagnostic positions for the T2b3 haplogroup. The base quality reported is from the output of schmutzi and is on a PHRED scale.

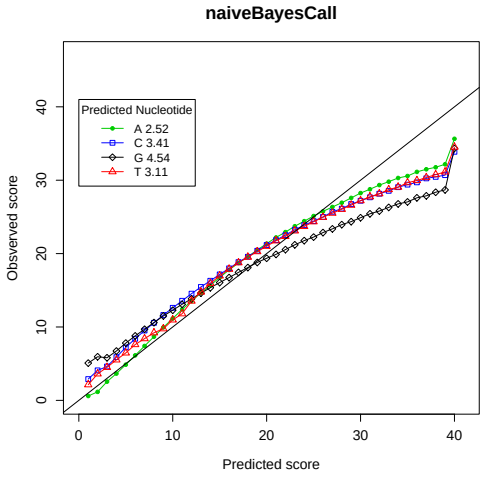
position	reference base	diagnostic base	predicted base with max. likelihood	quality on a PHRED scale	predicted = diagnostic ?
73	A	G	G	1085.01	yes
151	C	T	T	122.804	yes
263	A	G	G	892.348	yes
709	G	A	A	1216.99	yes
750	A	G	G	1490.97	yes
930	G	A	A	231.495	yes
1438	A	G	G	1173.56	yes
1888	G	A	A	201.053	yes
2706	A	G	G	232.934	yes
4216	T	C	C	71.3592	yes
4769	A	G	G	957.917	yes
4917	A	G	G	264.455	yes
5147	G	A	A	173.945	yes
7028	C	T	T	1407.42	yes
8697	G	A	A	75.387	yes
8860	A	G	G	1127.82	yes
10463	T	C	C	25.2927	yes
10750	A	G	G	968.847	yes
11251	A	G	G	202.335	yes
11719	G	A	A	1444.12	yes
11812	A	G	G	165.341	yes
13368	G	A	A	179.121	yes
14233	A	G	G	263.217	yes
14766	C	T	T	1117.02	yes
14905	G	A	A	312.261	yes
15326	A	G	G	1409.39	yes
15452	C	A	A	146.847	yes
15607	A	G	G	293.051	yes
15928	G	A	A	236.537	yes
16126	T	C	C	143.69	yes
16294	C	T	T	108.801	yes
16296	C	T	T	133.758	yes
16304	T	C	C	135.182	yes

Table A.23: Predicted contaminant from the Mezmaiskaya sample B9688 with the diagnostic positions for the T2b3 haplogroup. The base quality reported is from the output of schmutzi and is on a PHRED scale.

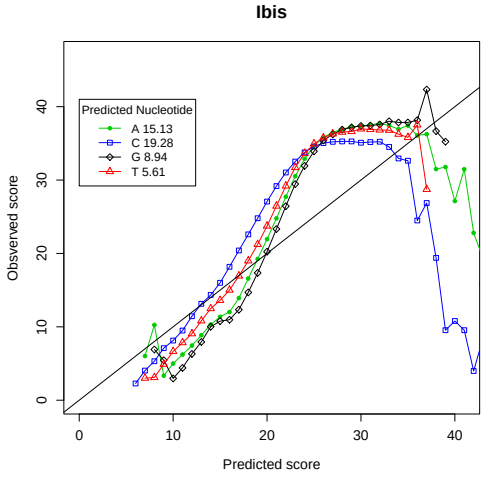
Appendix B


Bustard
Predicted Nucleotide
A 2.13
C 1.46
G 1.78
T 1.78
Obsverved score
Predicted score
0
10
20
30
40

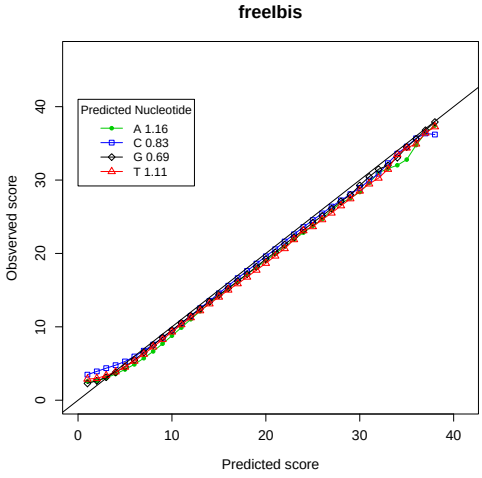
(a)


naiveBayesCall
Predicted Nucleotide
A 2.52
C 3.41
G 4.54
T 3.11
Obsverved score
Predicted score
0
10
20
30
40

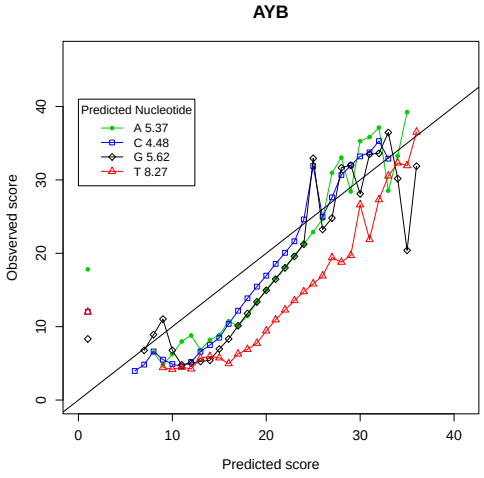
(b)


Ibis
Predicted Nucleotide
A 15.13
C 19.28
G 8.94
T 5.61
Obsverved score
Predicted score
0
10
20
30
40

(c)

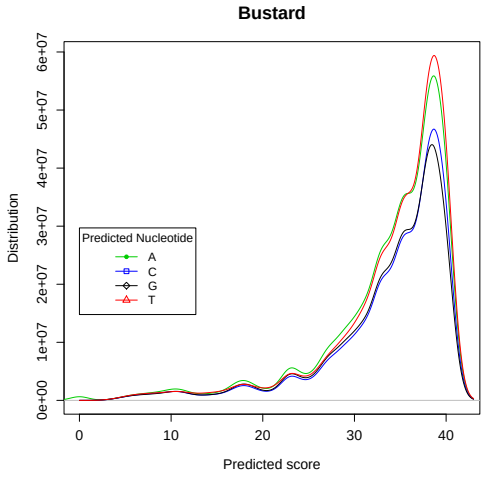

freelbis
Predicted Nucleotide
A 1.16
C 0.83
G 0.69
T 1.11
Obsverved score
Predicted score
0
10
20
30
40

(d)

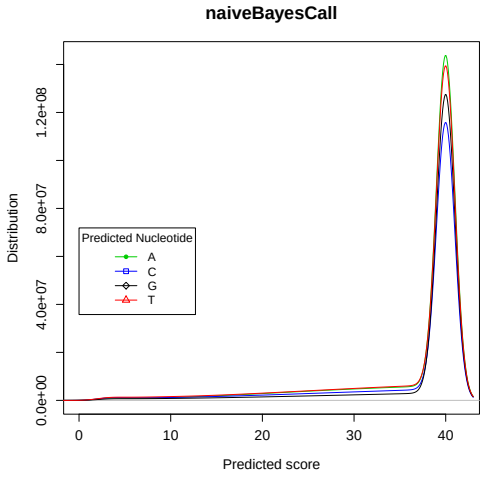

AYB
Predicted Nucleotide
A 5.37
C 4.48
G 5.62
T 8.27
Obsverved score
Predicted score
0
10
20
30
40

(e)

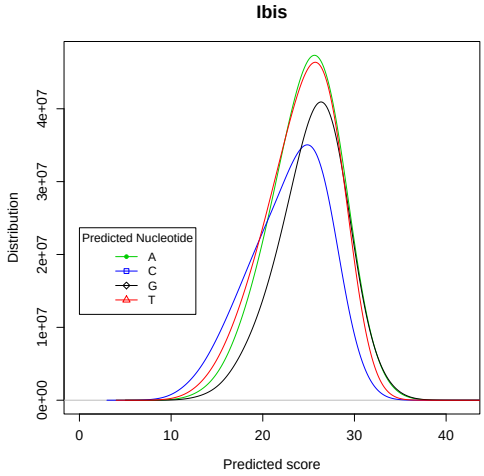
Figure B.1 *(preceding page)*: The observed versus predicted quality scores for each nucleotide for a Genome Analyzer II (2009) run along with the RMSE. The graphs represent Bustard (a), naiveBayesCall (b), Ibis (c), freeIbis with calibration (d) and AYB (e). AYB provides a separate tool to recalibrate the quality scores based on observed quality scores of clusters identified as controls. A downside of the freeIbis calibration method is, due to the shape of a the logarithm of the logistic function, an approximation using a linear function will underestimate data points around the origin and therefore, the actual error rate of bases with a low quality will be overstated (i.e. low quality bases have actually a higher observed quality score). This can be seen in (d) where low quality bases have a lower error rate than the predicted one and remain above the diagonal.


Bustard
Distribution
Predicted Nucleotide
A
C
G
T
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
0
10
20
30
40
Predicted score

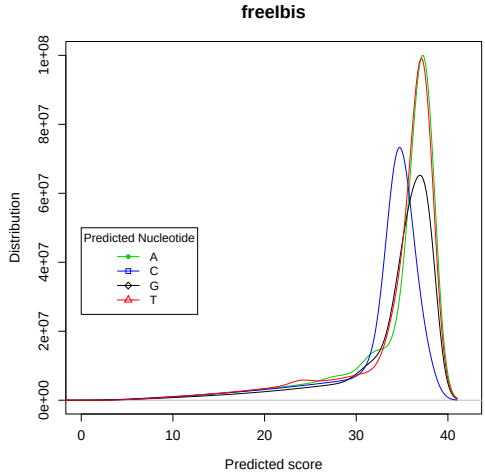
(a)


naiveBayesCall
Distribution
Predicted Nucleotide
A
C
G
T
0.0e+00
4.0e+07
8.0e+07
1.2e+08
0
10
20
30
40
Predicted score

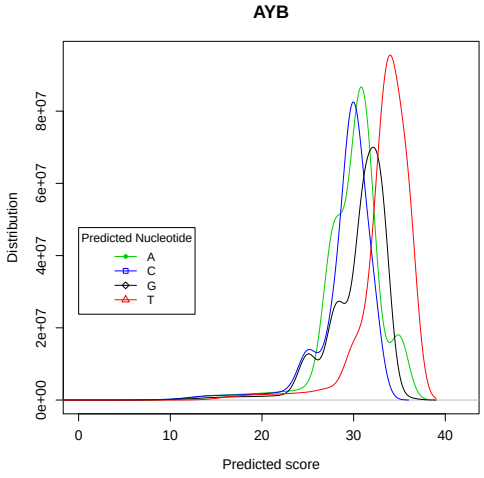
(b)


Ibis
Distribution
Predicted Nucleotide
A
C
G
T
0e+00
1e+07
2e+07
3e+07
4e+07
0
10
20
30
40
Predicted score

(c)

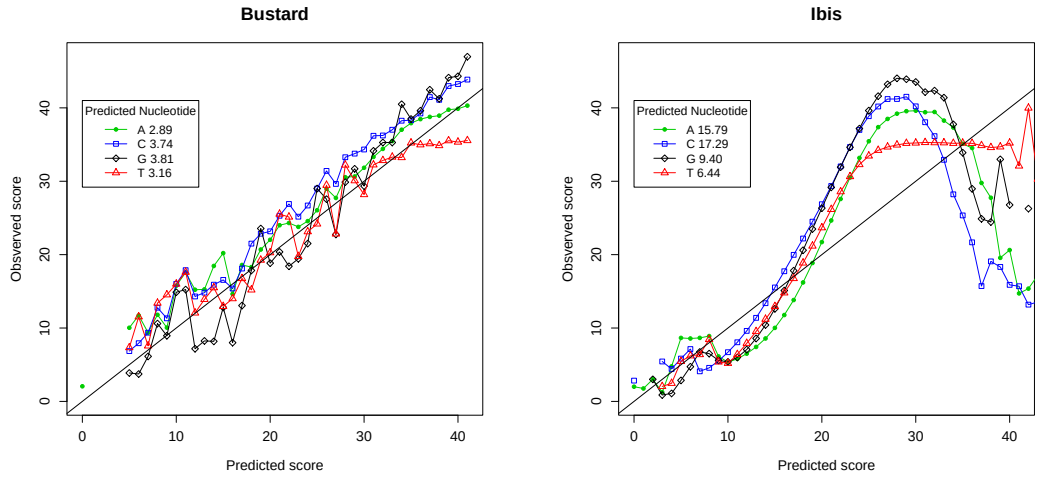

freelbis
Distribution
Predicted Nucleotide
A
C
G
T
0e+00
2e+07
4e+07
6e+07
8e+07
1e+08
0
10
20
30
40
Predicted score

(d)


AYB
Distribution
Predicted Nucleotide
A
C
G
T
0e+00
2e+07
4e+07
6e+07
8e+07
0
10
20
30
40
Predicted score

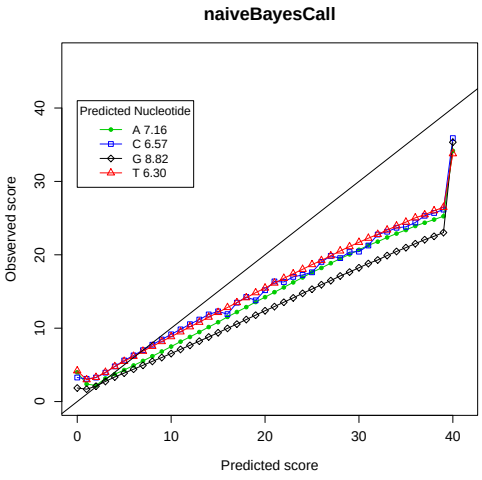
(e)

Figure B.2 *(preceding page)*: The distribution of the quality scores for each nucleotide for the Genome Analyzer II (2009) run for Bustard (a), naiveBayesCall (b), Ibis (c), freeIbis with calibration (d) and AYB (e).

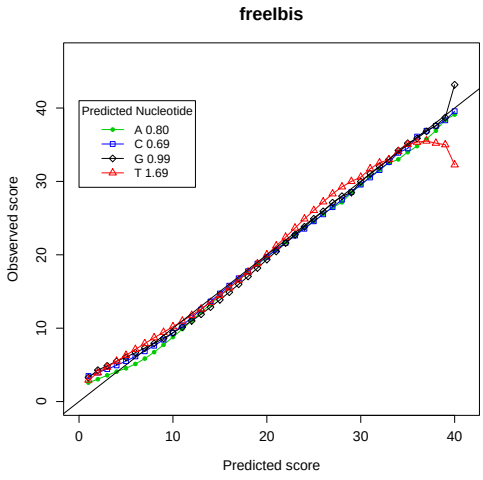


(a)

(b)

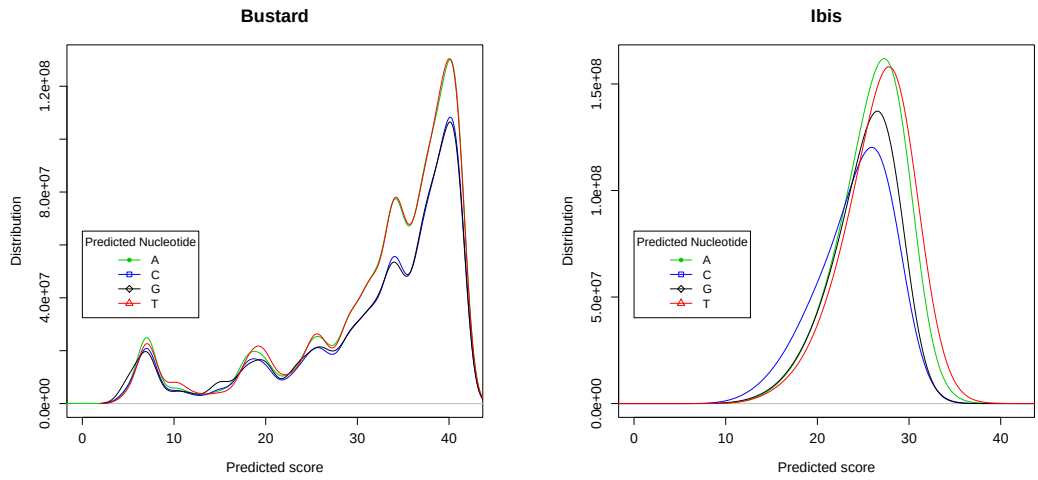


(c)



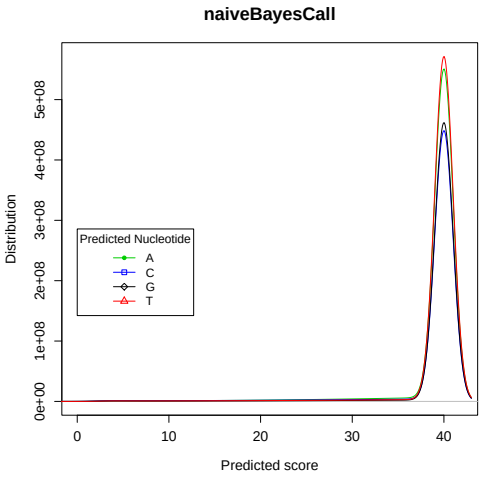
(d)

Figure B.3: The observed versus predicted quality scores for a HiSeq (2010) for each basecaller namely Bustard (a), naiveBayesCall (c), Ibis (b) and freeIbis with calibration (d). AYB was unable to produce sequences for this control lane.

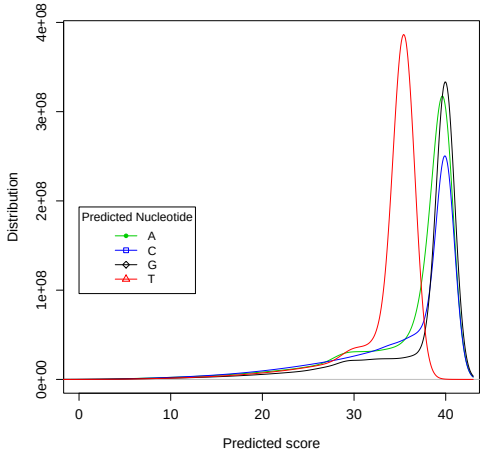


(a)

(b)

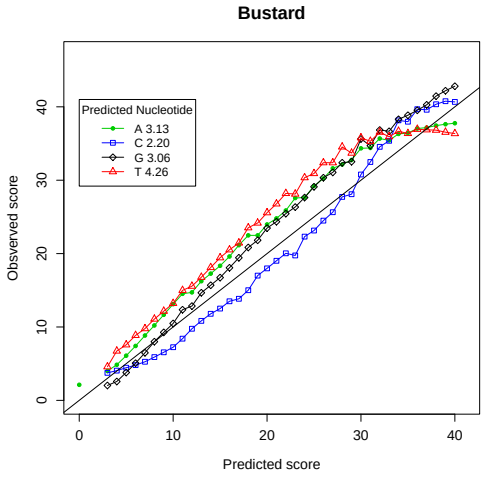


(c)

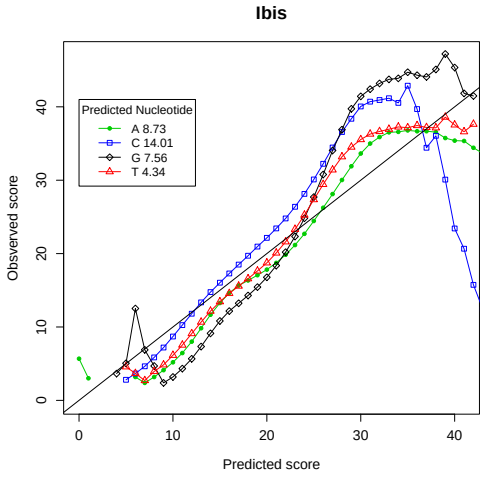


(d)

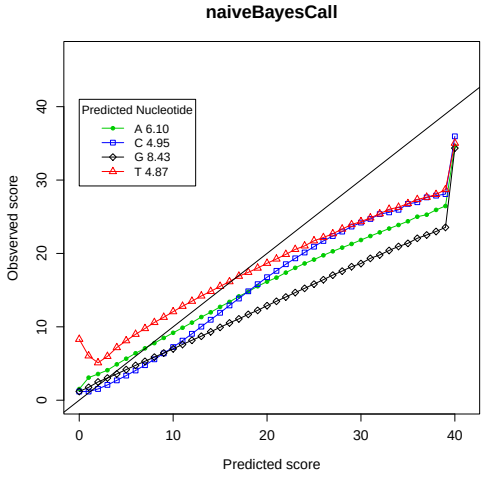
Figure B.4: The distribution of the predicted quality scores for a HiSeq (2010) for Bustard (a), naiveBayesCall (c), Ibis (b) and freeIbis with calibration (d). The skewed distribution of the T nucleotide in the calibrated scores in freeIbis can be explained due to a higher error rate for this given nucleotide. AYB was unable to produce sequences for this control lane.


Bustard
Predicted Nucleotide
A 3.13
C 2.20
G 3.06
T 4.26
Obsverved score
Predicted score
0
10
20
30
40

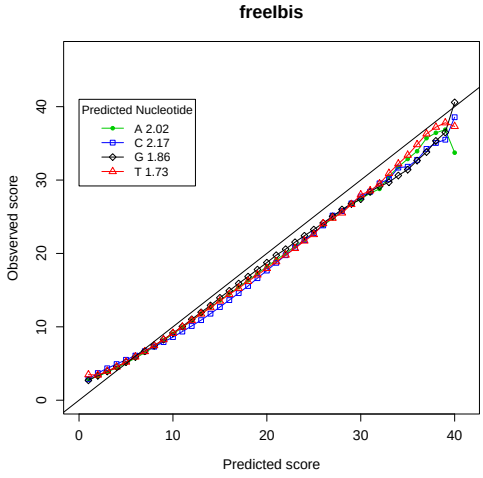
(a)


Ibis
Predicted Nucleotide
A 8.73
C 14.01
G 7.56
T 4.34
Obsverved score
Predicted score
0
10
20
30
40

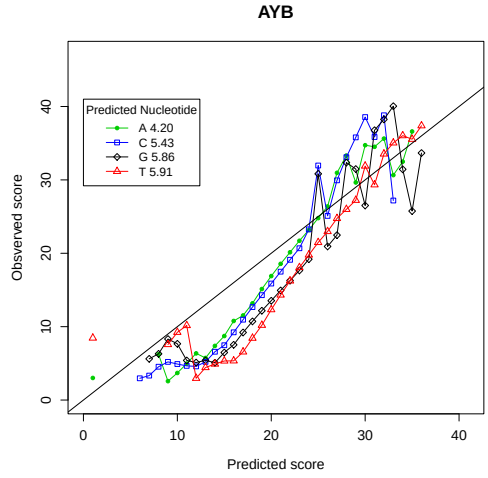
(b)


naiveBayesCall
Predicted Nucleotide
A 6.10
C 4.95
G 8.43
T 4.87
Obsverved score
Predicted score
0
10
20
30
40

(c)

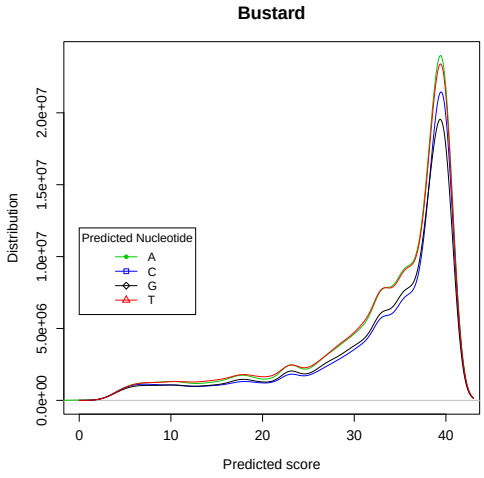

freelbis
Predicted Nucleotide
A 2.02
C 2.17
G 1.86
T 1.73
Obsverved score
Predicted score
0
10
20
30
40

(d)

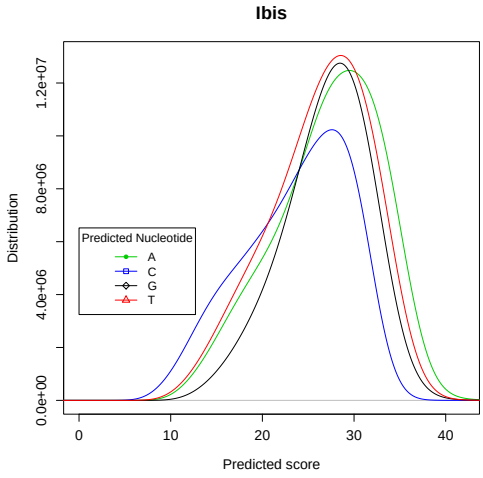

AYB
Predicted Nucleotide
A 4.20
C 5.43
G 5.86
T 5.91
Obsverved score
Predicted score
0
10
20
30
40

(e)

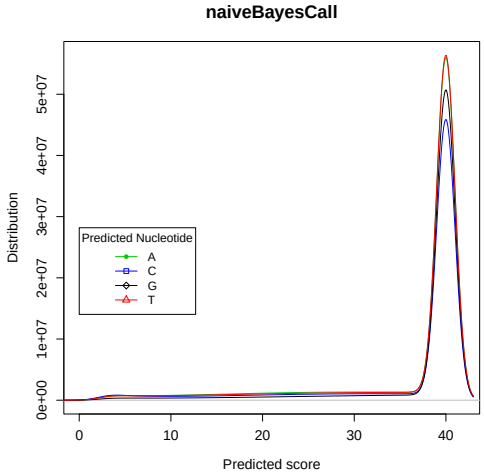
Figure B.5 *(preceding page)*: The observed versus predicted quality scores plots for Genome Analyzer II (2011) for Bustard (a), naiveBayesCall (c), Ibis without calibration (b), freeIbis with calibration (d) and AYB (e).


Bustard
Distribution
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
Predicted Nucleotide
A
C
G
T
0
10
20
30
40
Predicted score

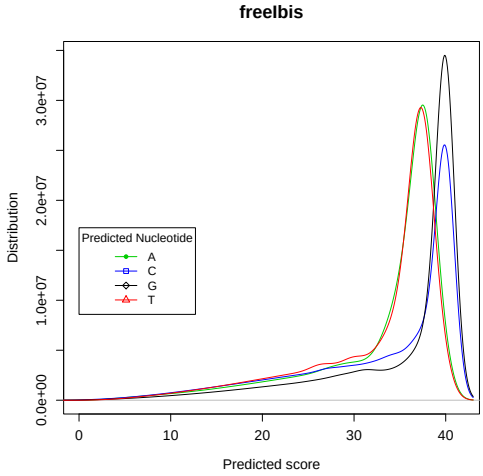
(a)


Ibis
Distribution
0.0e+00
4.0e+06
8.0e+06
1.2e+07
Predicted Nucleotide
A
C
G
T
0
10
20
30
40
Predicted score

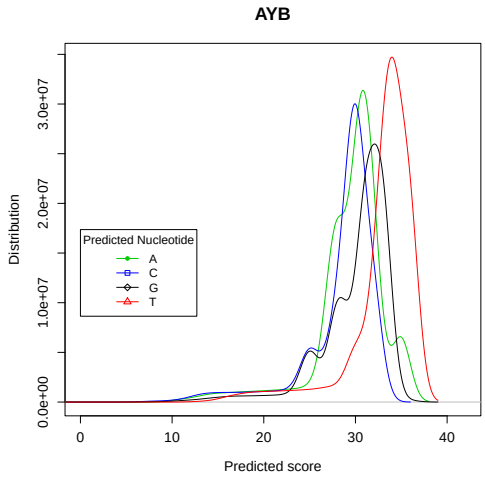
(b)


naiveBayesCall
Distribution
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
Predicted Nucleotide
A
C
G
T
0
10
20
30
40
Predicted score

(c)


freelbis
Distribution
0.0e+00
1.0e+07
2.0e+07
3.0e+07
Predicted Nucleotide
A
C
G
T
0
10
20
30
40
Predicted score

(d)


AYB
Distribution
0.0e+00
1.0e+07
2.0e+07
3.0e+07
Predicted Nucleotide
A
C
G
T
0
10
20
30
40
Predicted score

(e)

Figure B.6 *(preceding page)*: The distribution of predicted quality scores for a sequencing run on the Genome Analyzer II (2011) platform (Bustard (a), naiveBayesCall (c), Ibis without calibration (b), freeIbis with calibration (d) and AYB (e)).

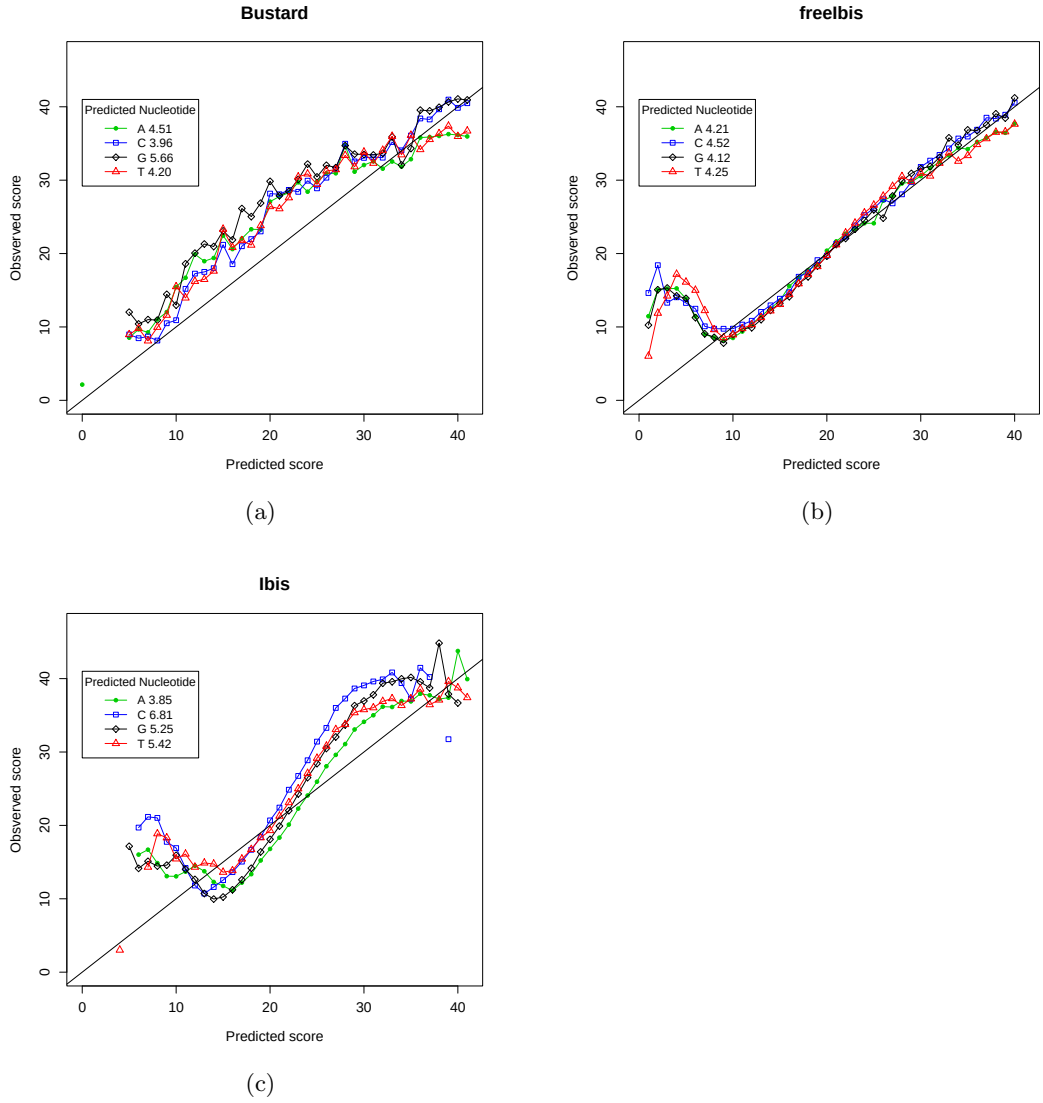


Figure B.7: Plots for the observed versus predicted quality scores for a sequencing run on the newest Illumina platform, the MiSeq (2012). The plots show the correlation for Bustard (a), Ibis without calibration (c) and freeIbis with calibration (b). Due to the paucity of control sequences needed to calibrate the quality scores, groupings of 5 consecutive cycles were used to measure the correlation between the SVM decision boundary distance and the observed error rate.

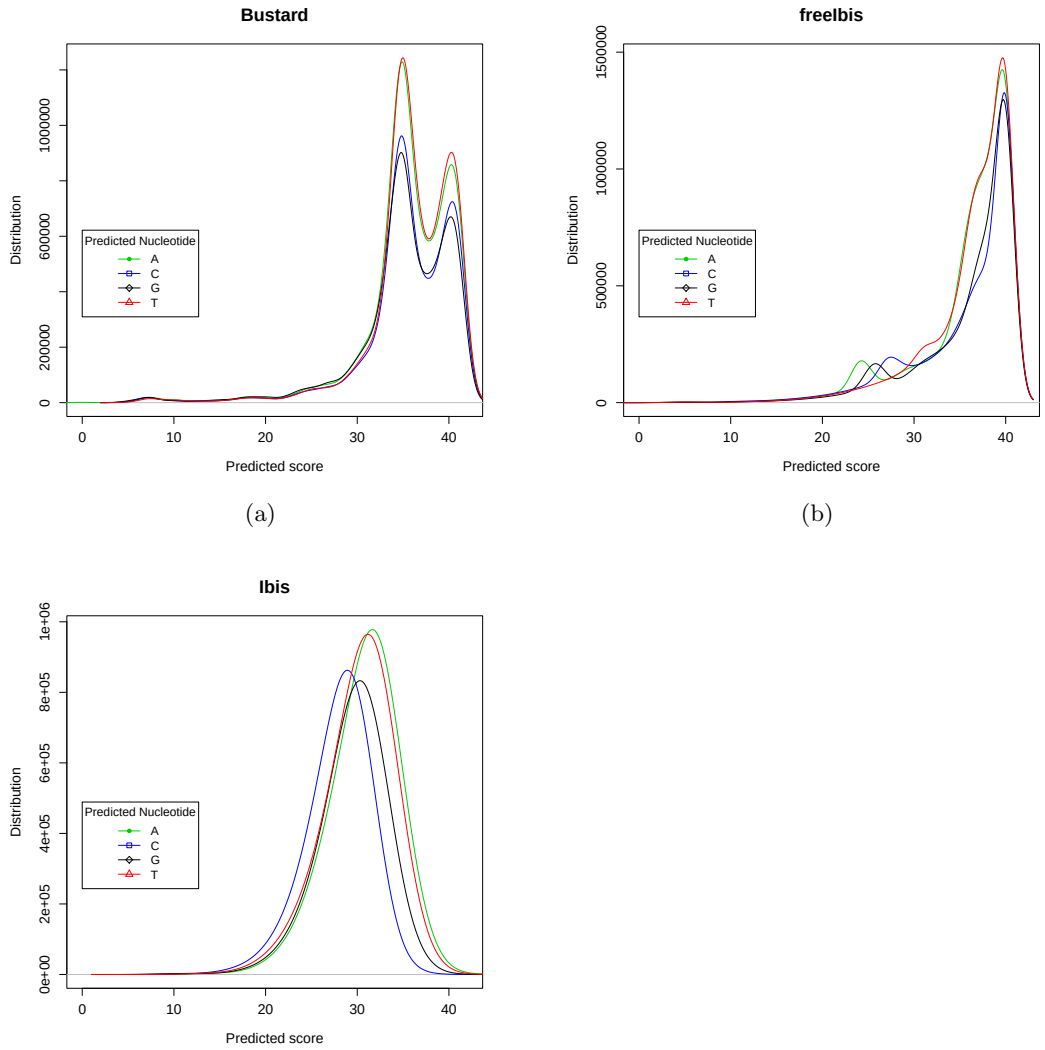
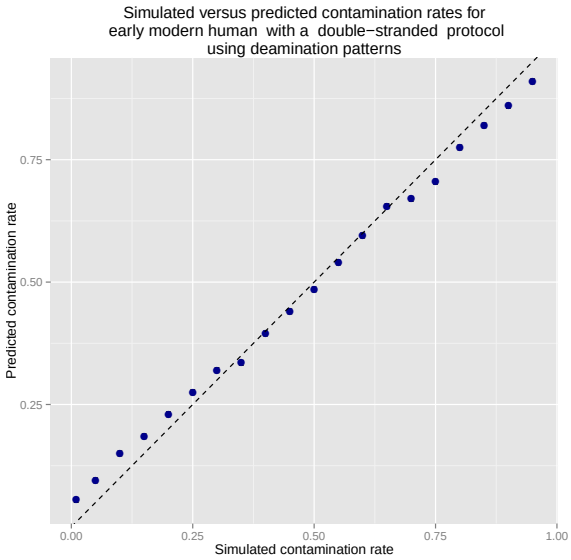
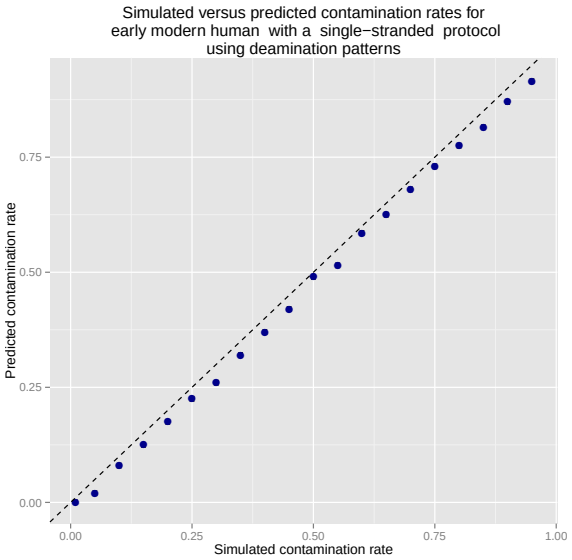
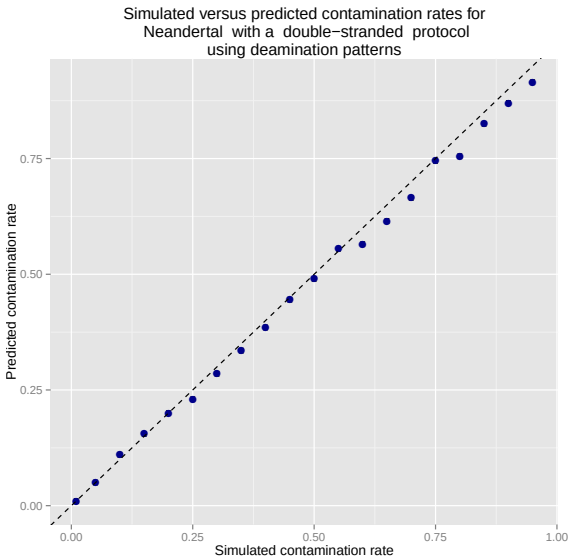
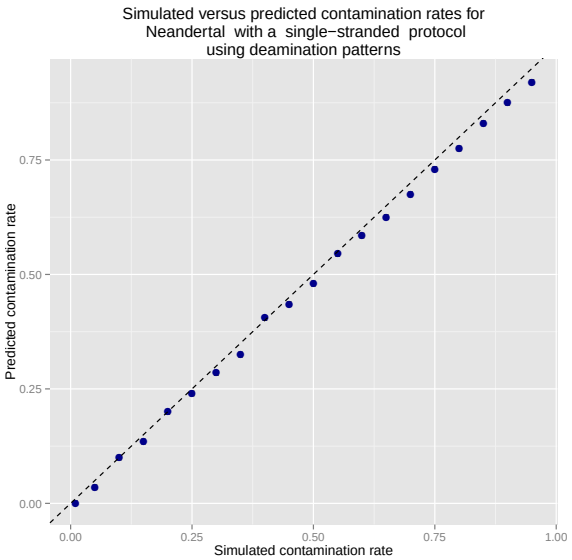


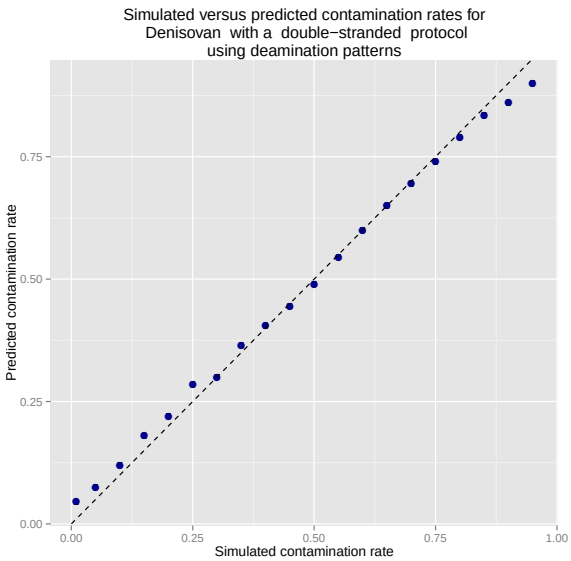
Figure B.8: Density plots of the predicted quality scores on a MiSeq (2012) for various basecallers (Bustard (a), Ibis without calibration (c) and freeIbis with calibration (b)).


Simulated versus predicted contamination rates for early modern human with a double−stranded protocol using deamination patterns
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates for early modern human with a single−stranded protocol using deamination patterns
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates for Neandertal with a double−stranded protocol using deamination patterns
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates for Neandertal with a single−stranded protocol using deamination patterns
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates for Denisovan with a double−stranded protocol using deamination patterns
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00

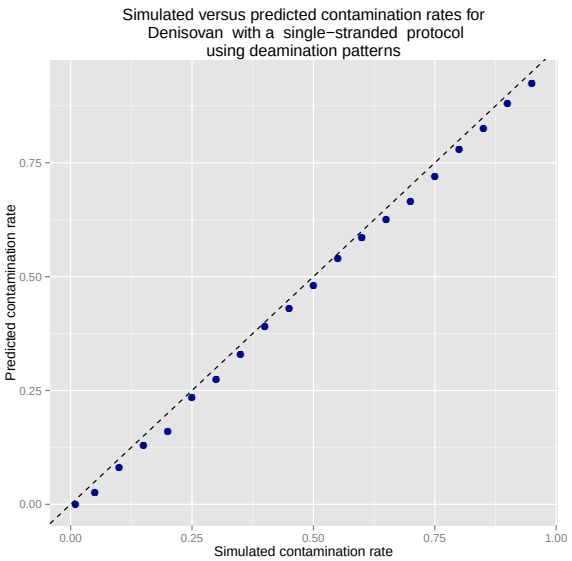
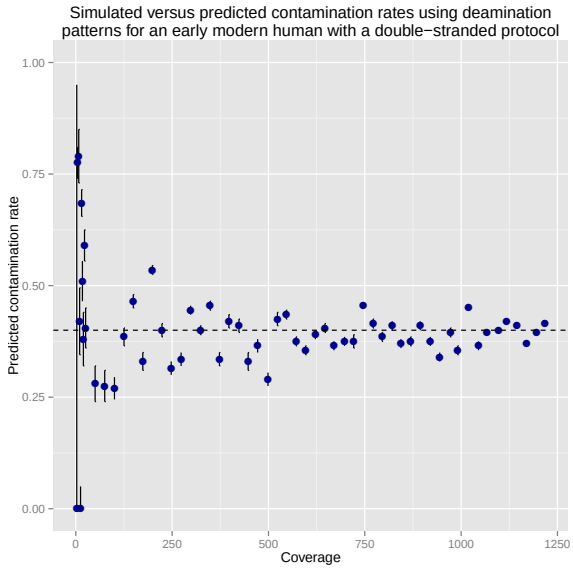
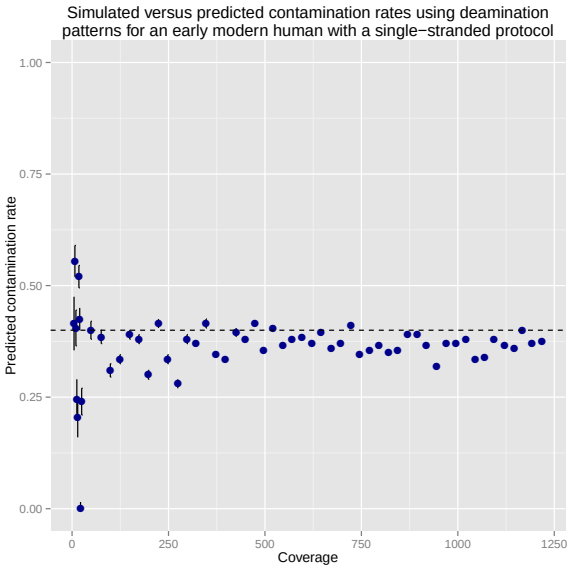
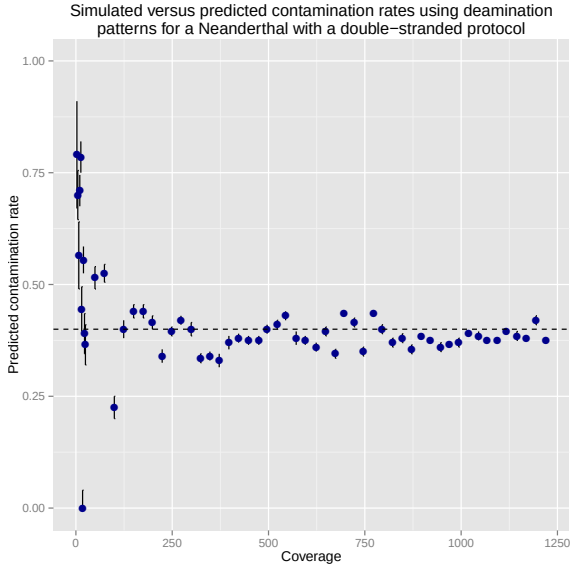
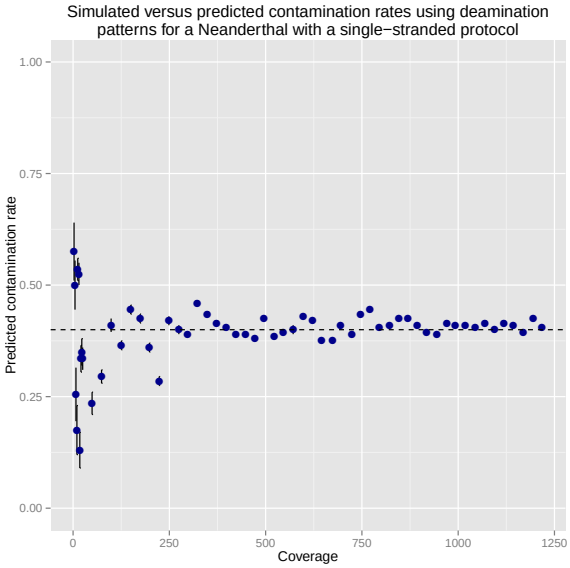

Simulated versus predicted contamination rates for Denisovan with a single−stranded protocol using deamination patterns
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00

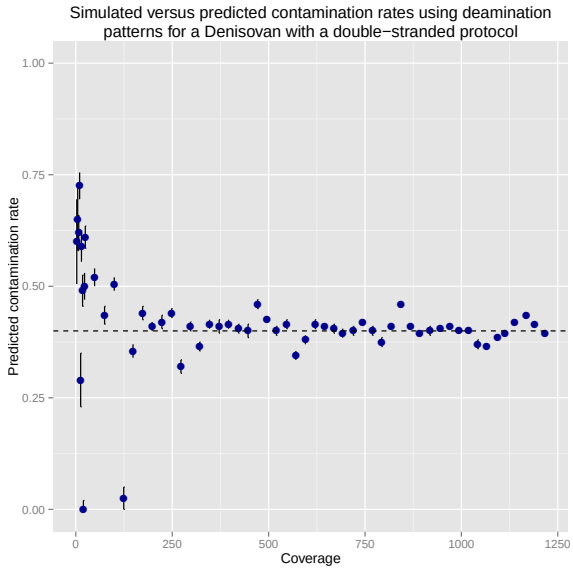
Figure B.9 *(preceding page)*: Simulated contamination rates versus predicted ones using deamination patterns alone. Schmutzi was tested on sets containing 1M simulated aDNA fragments using as endogenous genome an early modern humans (top), Neanderthals (middle) and a Denisovan(bottom). The program was tested both with simulated double-stranded (left) and single-stranded (right) protocols. The dotted black line represents a perfect prediction.


Simulated versus predicted contamination rates using deamination patterns for an early modern human with a double−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Simulated versus predicted contamination rates using deamination patterns for an early modern human with a single−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Simulated versus predicted contamination rates using deamination patterns for a Neanderthal with a double−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Simulated versus predicted contamination rates using deamination patterns for a Neanderthal with a single−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Simulated versus predicted contamination rates using deamination patterns for a Denisovan with a double−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage

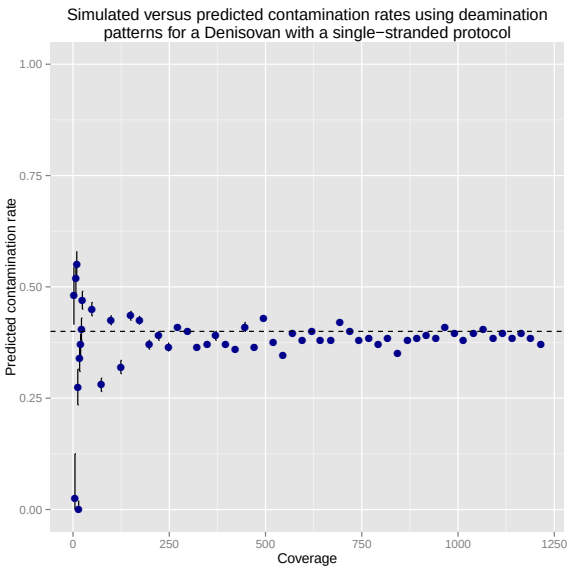
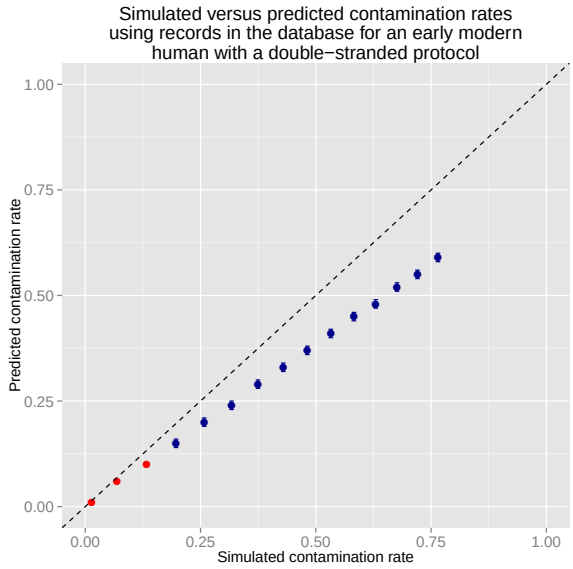
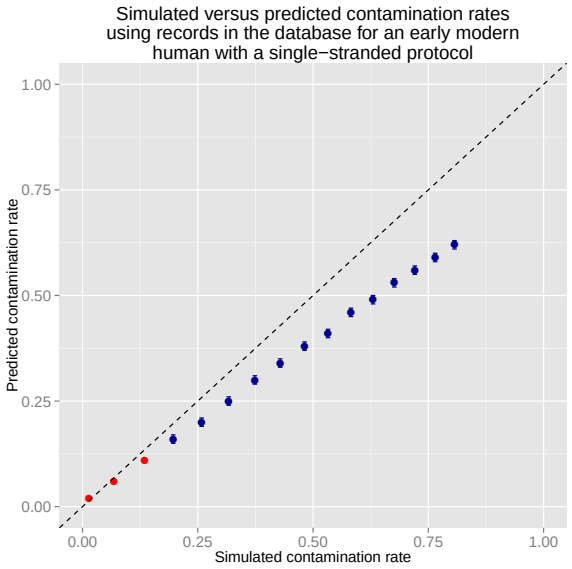
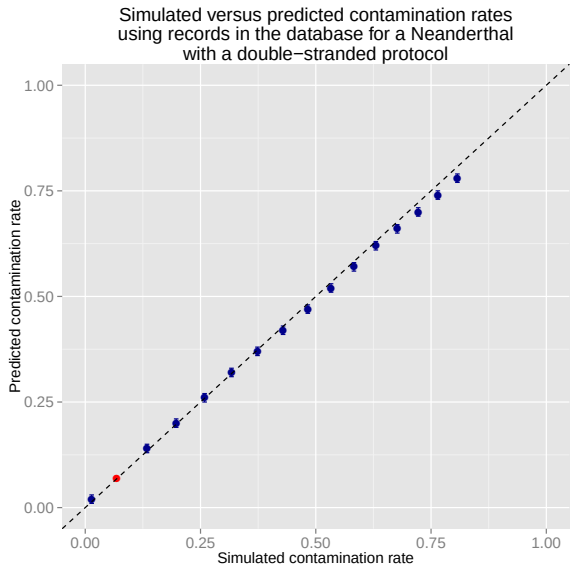
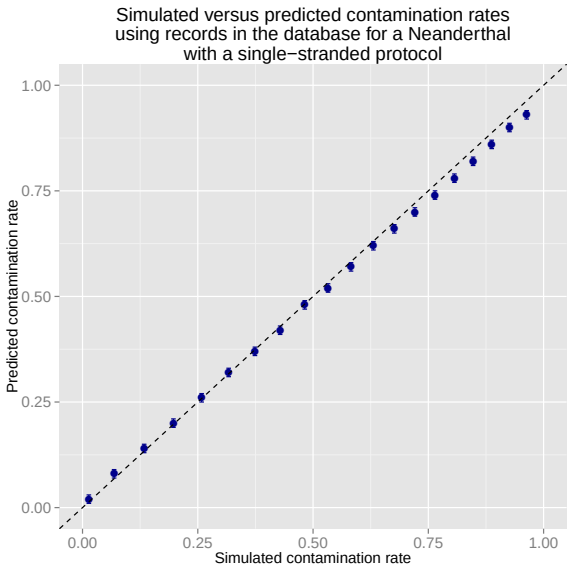

Simulated versus predicted contamination rates using deamination patterns for a Denisovan with a single−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage

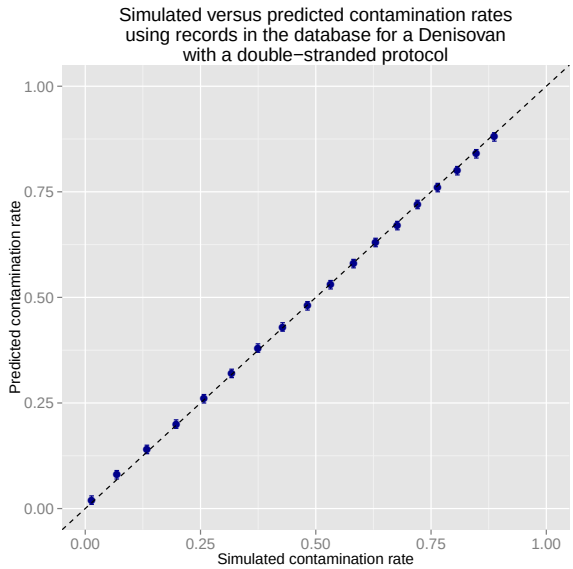
Figure B.10 *(preceding page)*: Simulated contamination rates using subsampled sets from a 1M fragment dataset where the original contamination rate was 40% (dotted black line) versus the predicted ones using deamination rates alone. The vertical black lines represent the boundaries of the 95% confidence interval.


Simulated versus predicted contamination rates using records in the database for an early modern human with a double−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates using records in the database for an early modern human with a single−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates using records in the database for a Neanderthal with a double−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates using records in the database for a Neanderthal with a single−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates using records in the database for a Denisovan with a double−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00

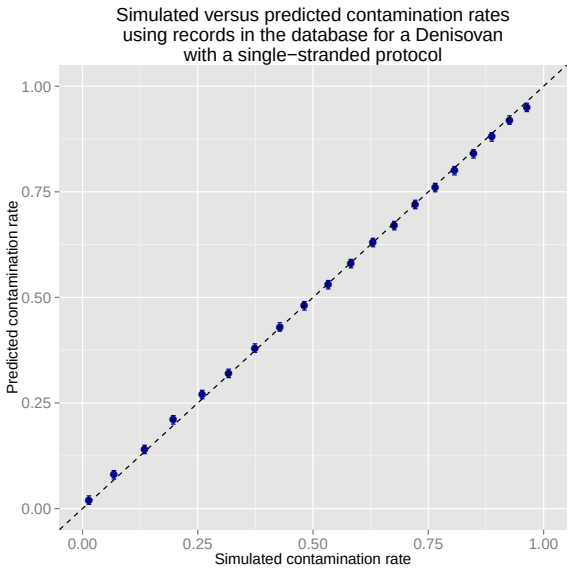
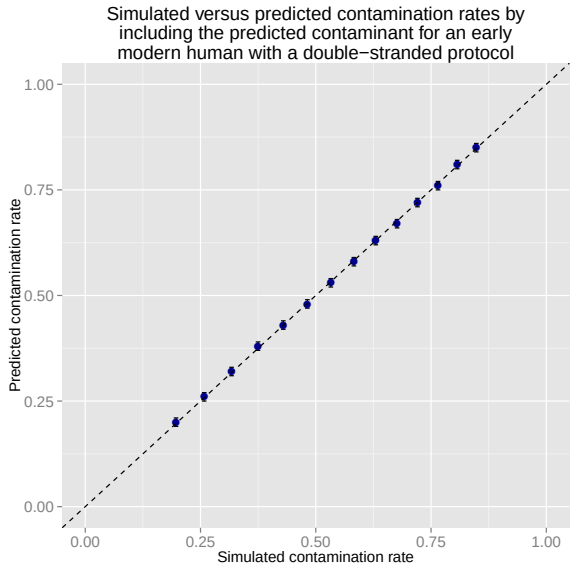
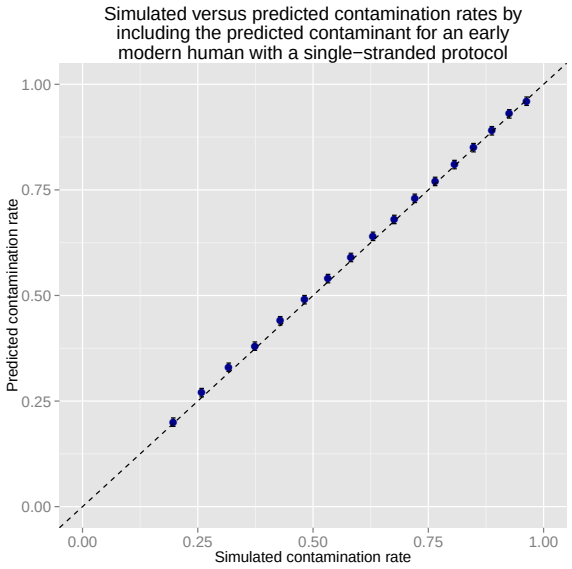
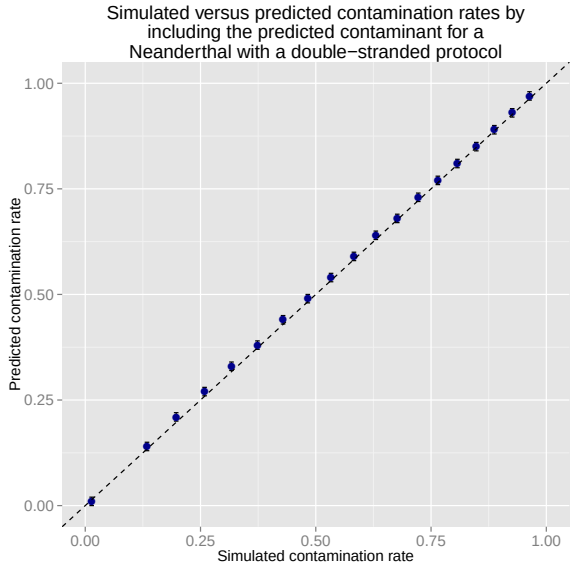
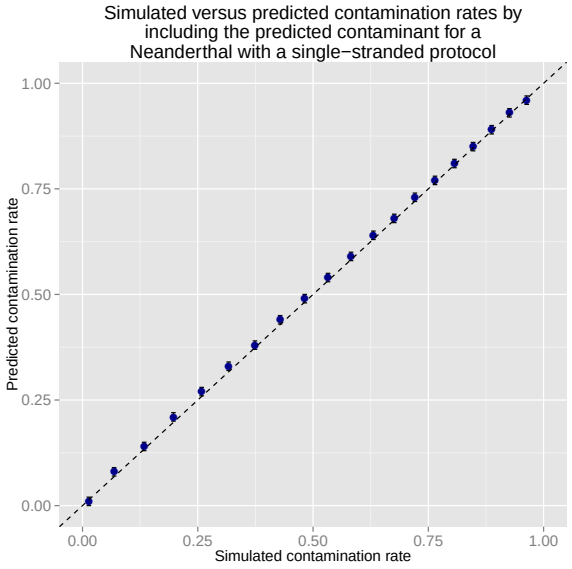

Simulated versus predicted contamination rates using records in the database for a Denisovan with a single−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00

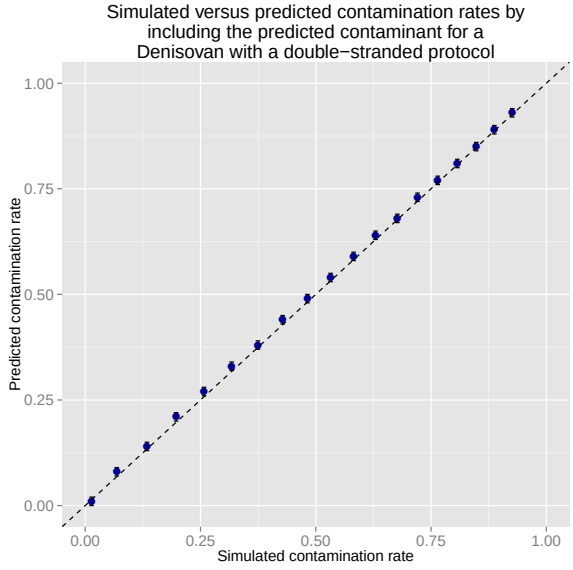
Figure B.11 *(preceding page)*: Simulated contamination rate versus the predicted one for datasets containing 1M fragments each. The endogenous genome used was either an early modern human (top), a Neanderthal (middle) or a Denisovan (bottom) and the simulated aDNA damage pattern was either double-stranded (left) or single-stranded (right). The data points where schmutzi stopped after the first iteration due to a lack of contaminant fragments to characterize are marked in red. As mentioned previously, for the EMH at high levels of contamination, the algorithm did not converge.


Simulated versus predicted contamination rates by including the predicted contaminant for an early modern human with a double−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates by including the predicted contaminant for an early modern human with a single−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates by including the predicted contaminant for a Neanderthal with a double−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates by including the predicted contaminant for a Neanderthal with a single−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates by including the predicted contaminant for a Denisovan with a double−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00

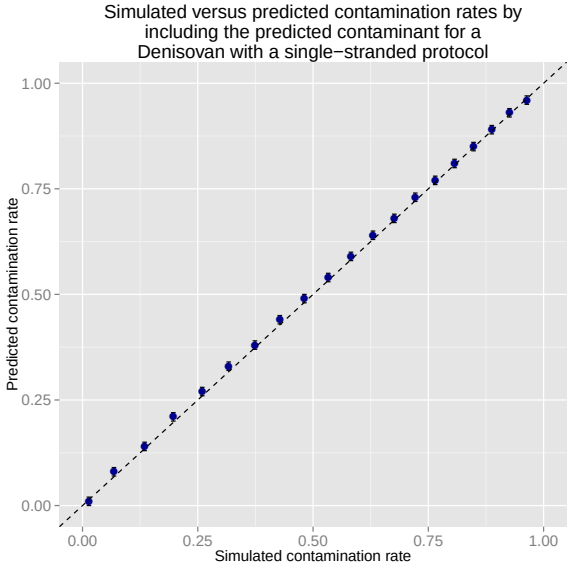
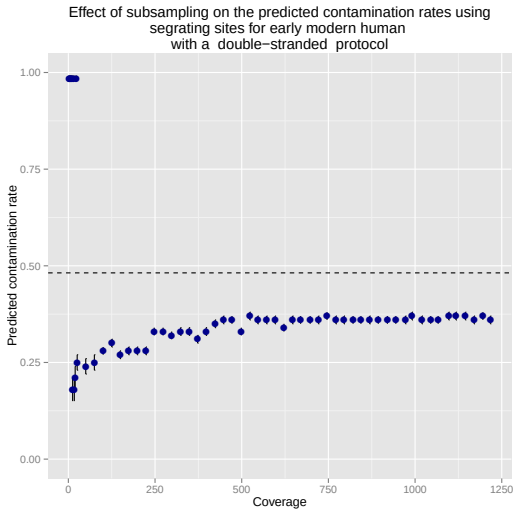
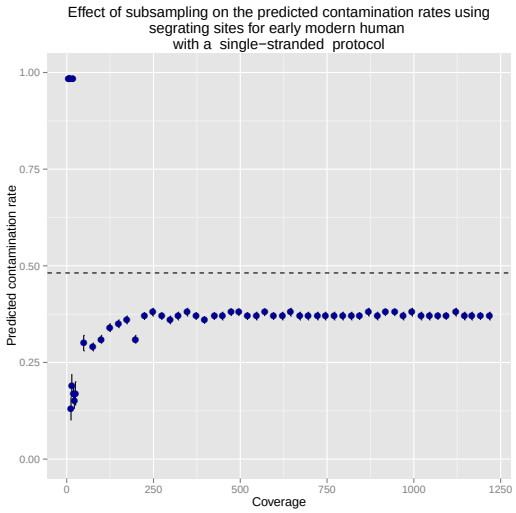
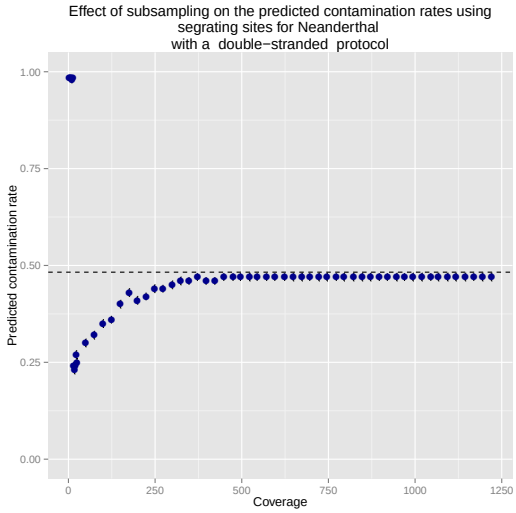
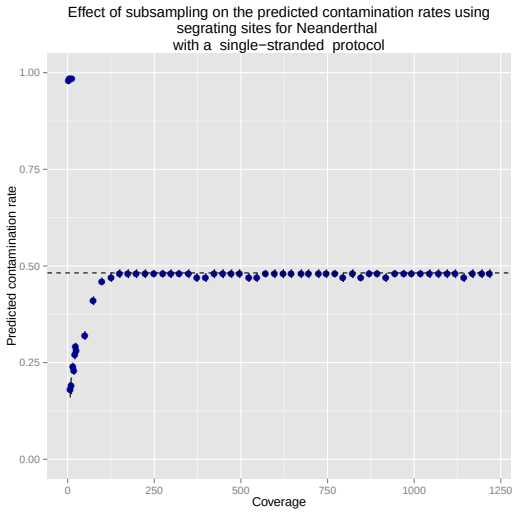

Simulated versus predicted contamination rates by including the predicted contaminant for a Denisovan with a single−stranded protocol
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00

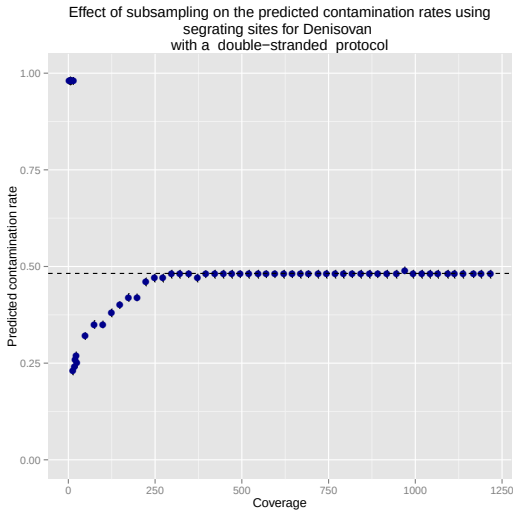
Figure B.12 *(preceding page)*: Simulated contamination rate versus the predicted one using the predicted contaminant as putative contaminant source for datasets containing 1M fragments each. The endogenous genome used was either an early modern human (top), a Neanderthal (middle) or a Denisovan (bottom) and the simulated aDNA damage pattern was double-stranded (left) or single-stranded (right). As for the previous graphs, the data points where schmutzi did not converge are omitted which mostly occur with an EMH as endogenous with either too little or too much contamination.


Effect of subsampling on the predicted contamination rates using segrating sites for early modern human with a double−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Effect of subsampling on the predicted contamination rates using segrating sites for early modern human with a single−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Effect of subsampling on the predicted contamination rates using segrating sites for Neanderthal with a double−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Effect of subsampling on the predicted contamination rates using segrating sites for Neanderthal with a single−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Effect of subsampling on the predicted contamination rates using segrating sites for Denisovan with a double−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage

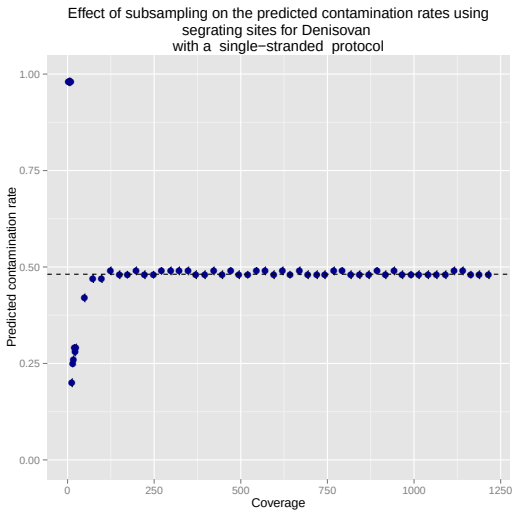
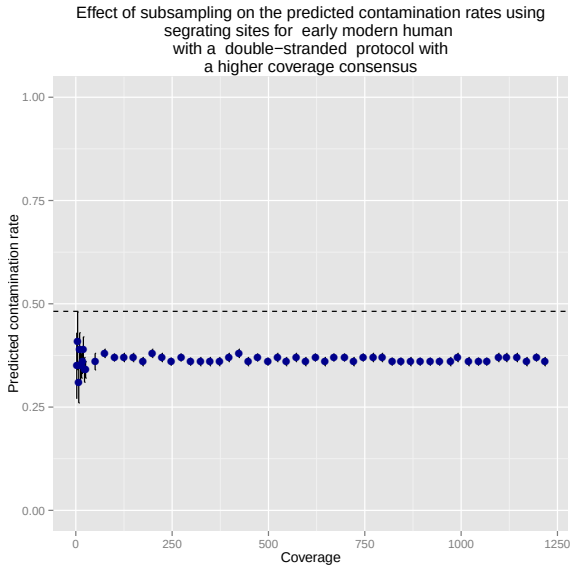
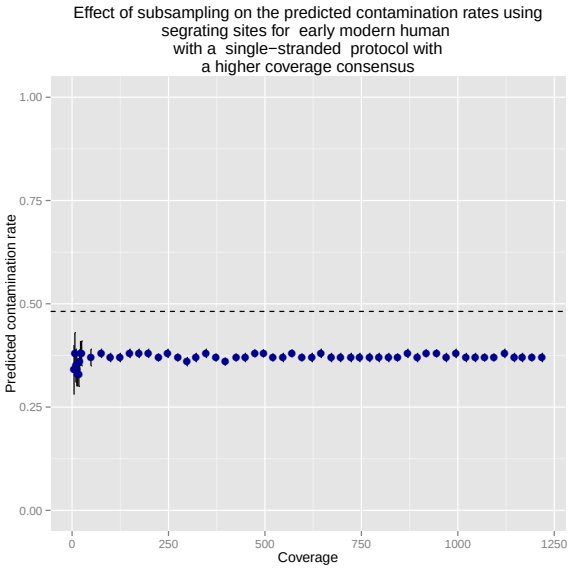
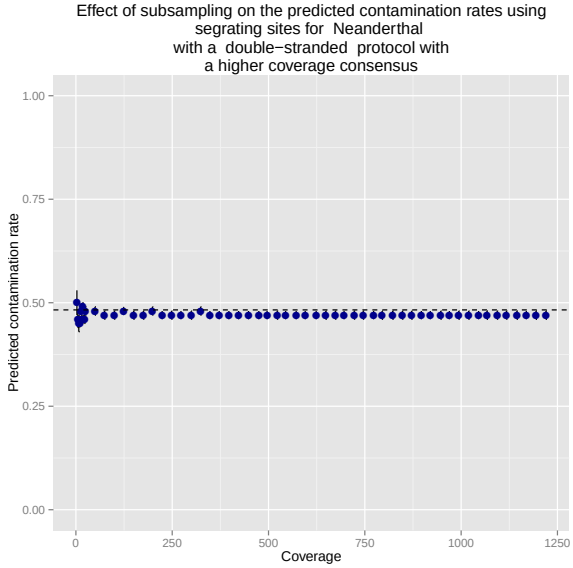
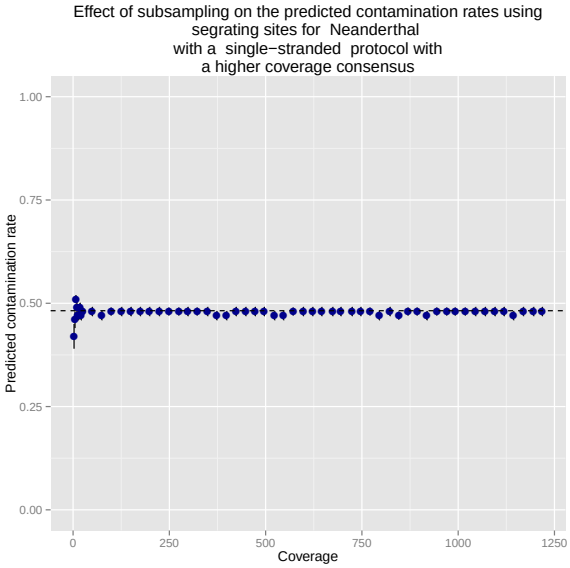

Effect of subsampling on the predicted contamination rates using segrating sites for Denisovan with a single−stranded protocol
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage

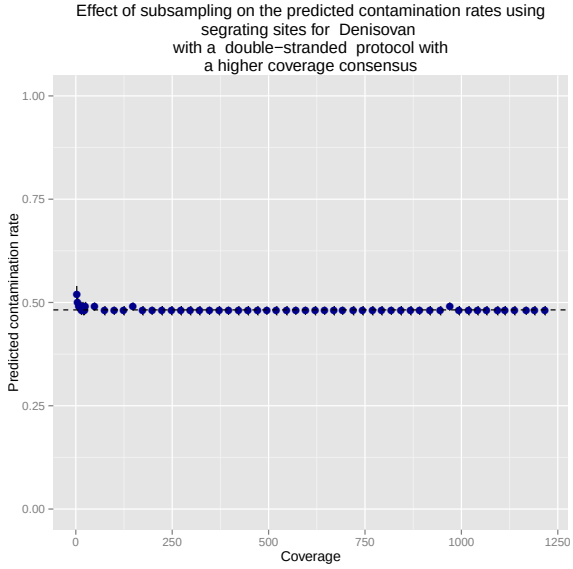
Figure B.13 *(preceding page)*: Predicted contamination rates at various coverages using schmutzi with default parameters. The endogenous genome used was either an early modern human (top), a Neanderthal (middle) or a Denisovan (bottom) and the simulated aDNA damage pattern was double-stranded (left) or single-stranded (right). The black dotted line corresponds to the simulated contamination rate.


Effect of subsampling on the predicted contamination rates using segrating sites for early modern human with a double−stranded protocol with a higher coverage consensus
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Effect of subsampling on the predicted contamination rates using segrating sites for early modern human with a single−stranded protocol with a higher coverage consensus
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Effect of subsampling on the predicted contamination rates using segrating sites for Neanderthal with a double−stranded protocol with a higher coverage consensus
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Effect of subsampling on the predicted contamination rates using segrating sites for Neanderthal with a single−stranded protocol with a higher coverage consensus
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage


Effect of subsampling on the predicted contamination rates using segrating sites for Denisovan with a double−stranded protocol with a higher coverage consensus
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage

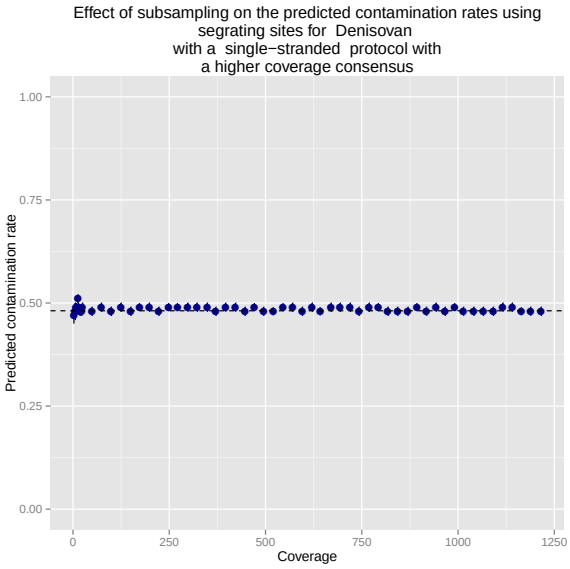
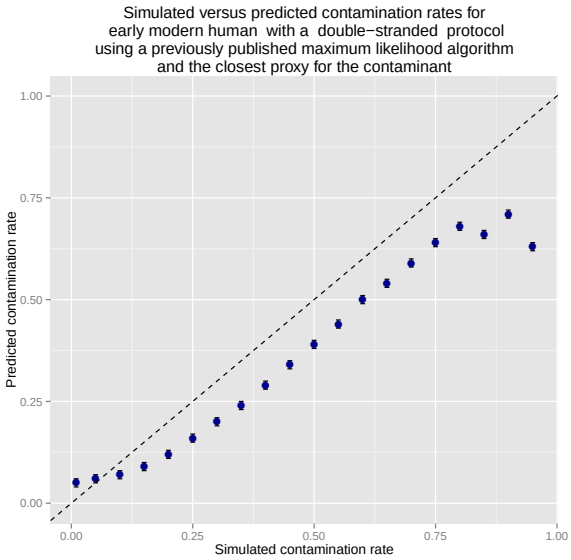
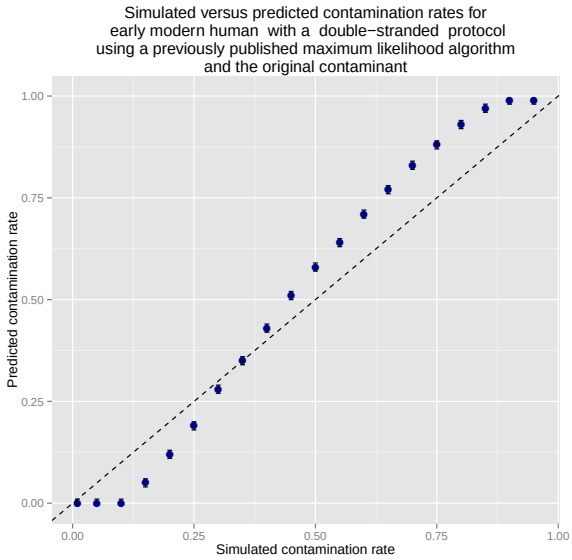
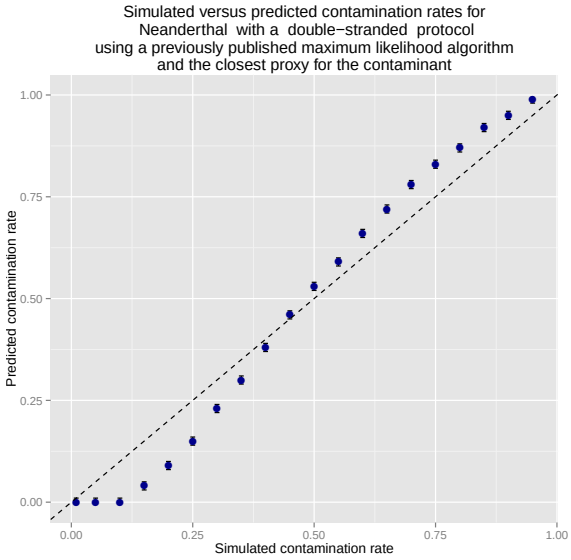
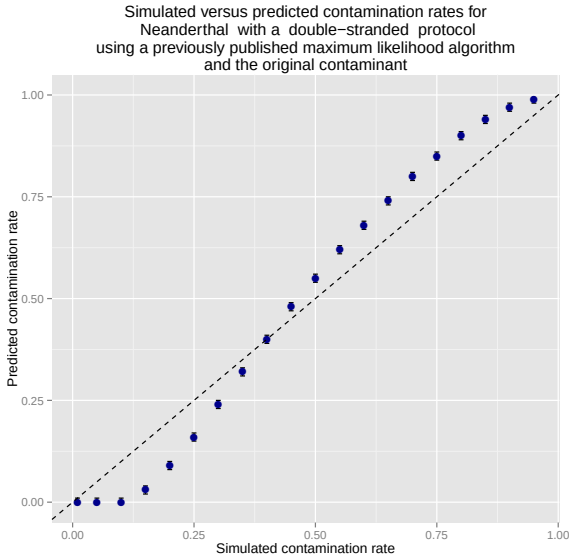

Effect of subsampling on the predicted contamination rates using segrating sites for Denisovan with a single−stranded protocol with a higher coverage consensus
Predicted contamination rate
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
Coverage

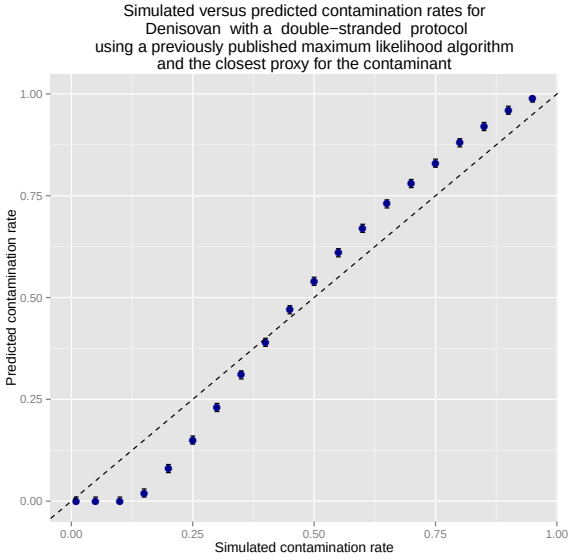
Figure B.14 *(preceding page)*: Predicted contamination rates at various rates of coverage using schmutzi with default parameters but with the endogenous genome inferred from the original set from which the fragments were subsampled. The endogenous genome used was either an early modern human (top), a Neanderthal (middle) or a Denisovan (bottom) and the simulated aDNA damage pattern was double-stranded (left) or single-stranded (right). The black dotted line corresponds to the simulated contamination rate.


Simulated versus predicted contamination rates for
early modern human with a double−stranded protocol
using a previously published maximum likelihood algorithm
and the closest proxy for the contaminant
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates for
early modern human with a double−stranded protocol
using a previously published maximum likelihood algorithm
and the original contaminant
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates for
Neanderthal with a double−stranded protocol
using a previously published maximum likelihood algorithm
and the closest proxy for the contaminant
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates for
Neanderthal with a double−stranded protocol
using a previously published maximum likelihood algorithm
and the original contaminant
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00


Simulated versus predicted contamination rates for
Denisovan with a double−stranded protocol
using a previously published maximum likelihood algorithm
and the closest proxy for the contaminant
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00

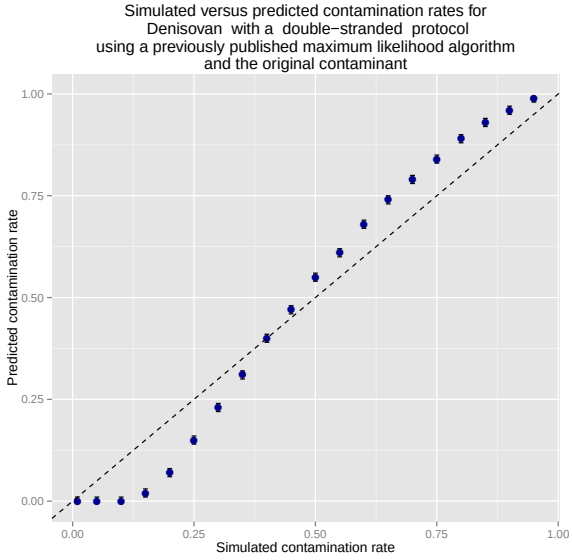

Simulated versus predicted contamination rates for
Denisovan with a double−stranded protocol
using a previously published maximum likelihood algorithm
and the original contaminant
Predicted contamination rate
Simulated contamination rate
0.00
0.25
0.50
0.75
1.00

Figure B.15 *(preceding page)*: Predicted contamination rates at various rates of coverage using a previously described maximum likelihood method. The method was tested on sets containing 1M simulated aDNA fragments using as endogenous genome an early modern humans (top), Neanderthals (middle) and a Denisovan (bottom). The method was used by including the closest record in the 311 mitochondrial genome database described in the method (left). To present the upper predictive limit, the actual contaminant used in the simulation was included (right). The dotted black line represents a perfect prediction.

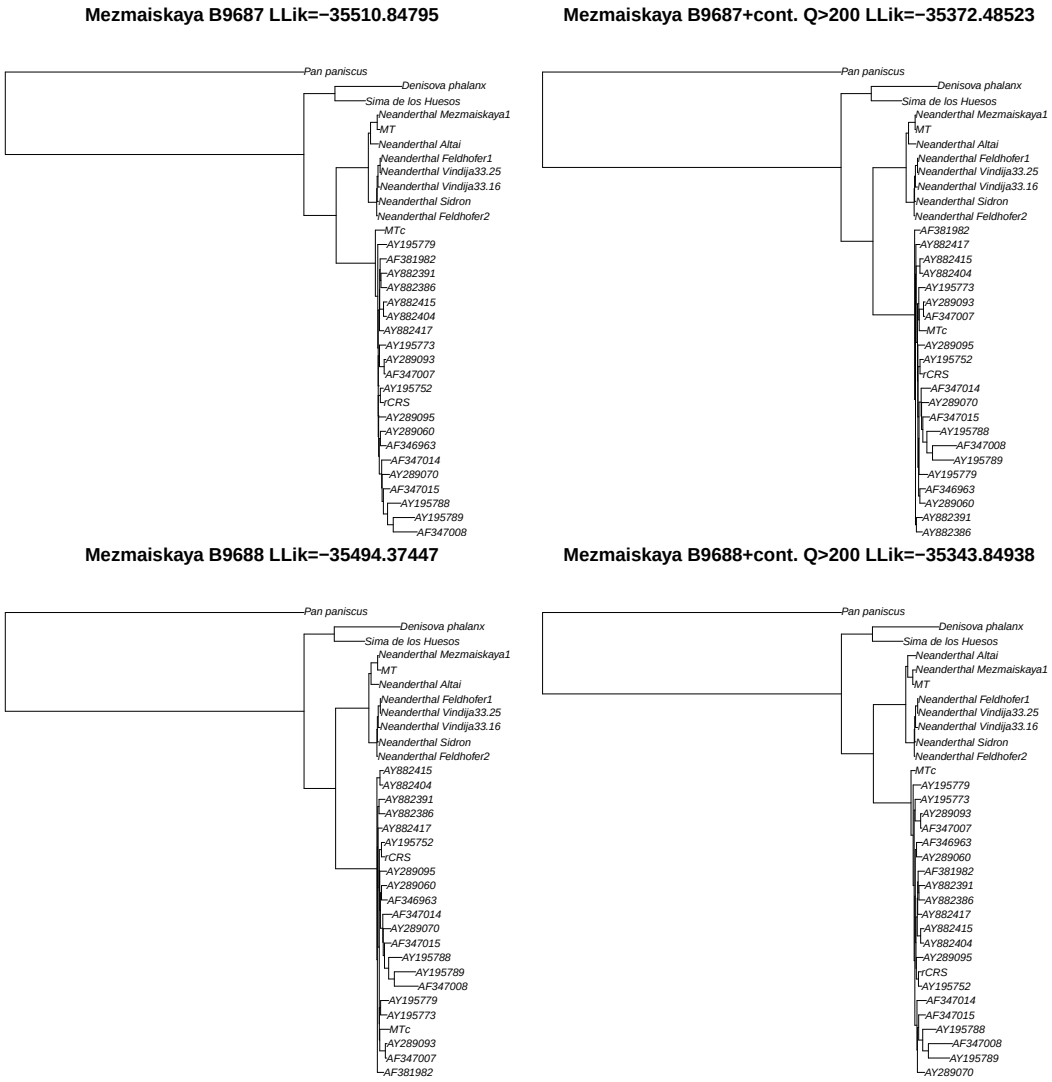


Figure B.16: Maximum likelihood trees for the Mezmaiskaya B9687 (top) and B9688 (bottom). The unfiltered data (left) and bases with quality greater than 200 on the PHRED scale (right) were plotted separately. The outgroup used is the bonobo mitochondrial genome.

Bibliography

[1] Edgar Altenburg and Hermann J. Muller. The Genetic Basis of Truncate Wing,-an Inconstant and Modifiable Character in Drosophila. *Genetics*, 5(1):1–59, 1920.

[2] Charles F. Aquadro and Barry D. Greenberg. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. *Genetics*, 103(2):287–312, 1983.

[3] Hans-Jürgen Bandelt. Mosaics of ancient mitochondrial DNA: positive indicators of nonauthenticity. *European Journal of Human Genetics*, 13(10):1106–1112, 2005.

[4] Chiara Barbieri, Mário Vicente, Sandra Oliveira, Koen Bostoen, Jorge Rocha, Mark Stoneking, and Brigitte Pakendorf. Migration and interaction in a contact zone: mtDNA variation among Bantu-speakers in southern Africa. *PLOS ONE*, 9(6):e99117, 2014.

[5] Doron M. Behar, Mannis van Oven, Saharon Rosset, Mait Metspalu, Eva-Liis Loogväli, Nuno M. Silva, Toomas Kivisild, Antonio Torroni, and Richard Villems. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *The American Journal of Human Genetics*, 90(4):675–684, 2012.

[6] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[7] Adrian W. Briggs, Udo Stenzel, Philip L.F. Johnson, Richard E. Green, Janet Kelso, Kay Prüfer, Matthias Meyer, Johannes Krause, Michael T. Ronan, Michael Lachmann, et al. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, 104(37):14616–14621, 2007.

[8] Lisa D. Brooks, B.S. Weir, and Henry E. Schaffer. The probabilities of similarities in DNA sequence comparisons. *Genomics*, 3(3):207–216, 1988.

[9] Hernán A. Burbano, Emily Hodges, Richard E. Green, Adrian W. Briggs, Johannes Krause, Matthias Meyer, Jeffrey M. Good, Tomislav Maricic, Philip L.F. Johnson, Zhenyu Xuan, et al. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science*, 328(5979):723–725, 2010.

[10] Tilo Buschmann and Leonid V. Bystrykh. Levenshtein error-correcting barcodes for multiplexed dna sequencing. *BMC Bioinformatics*, 14(1):272–272, 2013.

[11] Leonid V. Bystrykh. Generalized DNA barcode design based on Hamming codes. *PLOS ONE*, 7(5):e36852, 2012.

[12] Kai-wei Chang, Cho-jui Hsieh, Xiang-rui Wang, Chih-jen Lin, and Soeren Sonnenburg. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[13] Peter J.A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, 2010.

[14] Alan Cooper and Hendrik N. Poinar. Ancient DNA: do it right or not at all. *Science*, 289(5482):1139–1139, 2000.

[15] Paul Igor Costea, Joakim Lundeberg, and Pelin Akan. TagGD: Fast and Accurate Software for DNA Tag Generation and Demultiplexing. *PLOS ONE*, 8(3):e57521, 2013.

[16] David W. Craig, John V. Pearson, Szabolcs Szelinger, Aswin Sekar, Margot Redman, Jason J. Corneveaux, Traci L. Pawlowski, Trisha Laub, Gary Nunn, Dietrich A. Stephan, et al. Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods*, 5(10):887–893, 2008.

[17] Francis H. Crick. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138, 1958.

[18] Charles Darwin. *On the Origin of Species*. London: John Murray, 1859.

[19] Shreepriya Das and Haris Vikalo. Onlinecall: fast online parameter estimation and base calling for Illumina's next-generation sequencing. *Bioinformatics*, 28(13):1677–1683, 2012.

[20] Matei David, Misko Dzamba, Dan Lister, Lucian Ilie, and Michael Brudno. SHRiMP2: sensitive yet practical short read mapping. *Bioinformatics*, 27(7):1011–1012, 2011.

[21] Matthew Davis, Stijn van Dongen, Cei Abreu-Goodger, Nenad Bartonicek, and Anton J. Enright. Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods*, 63(1):41–49, 2013.

[22] Prescott L. Deininger. Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Analytical Biochemistry*, 129(1):216–223, 1983.

[23] Matthias Dodt, Johannes T. Roehr, Rina Ahmed, and Christoph Dieterich. FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology*, 1(3):895–905, 2012.

[24] Ana T. Duggan, Mark Whitten, Victor Wiebe, Michael Crawford, Anne Butthof, Victor Spitsyn, Sergey Makarov, Innokentiy Novgorodov, Vladimir Osakovsky, and Brigitte Pakendorf. Investigating the Prehistory of Tungusic Peoples of Siberia and the Amur-Ussuri Region with Complete mtDNA Genome Sequences and Y-chromosomal Markers. *PLOS ONE*, 8(12):e83570, 2013.

[25] Yaniv Erlich, Partha P. Mitra, W. Richard McCombie, Gregory J. Hannon, et al. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature Methods*, 5(8):679–682, 2008.

[26] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, 8(3):175–185, 1998.

[27] Brent Ewing and Phil Green. Base-calling of automated sequencer traces using phred. ii. Error probabilities. *Genome Research*, 8(3):186–194, 1998.

[28] James S. Farris. Estimating phylogenetic trees from distance matrices. *American Naturalist*, pages 645–668, 1972.

[29] Joseph Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[30] Joseph Felsenstein. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166, 1989.

[31] Ronald A. Fisher. *The Genetical Theory of Natural Selection*. Oxford University Press, US, 1930.

[32] V. Franc and S. Sonnenburg. Optimized cutting plane algorithm for large-scale risk minimization. *The Journal of Machine Learning Research*, 10:2157–2192, 2009.

[33] Vojtěch Franc and Soeren Sonnenburg. Optimized cutting plane algorithm for support vector machines. In *Proceedings of the 25th international conference on Machine learning*, pages 320–327. ACM, 2008.

[34] Daniel Frank. BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. *BMC Bioinformatics*, 10(1):362, 2009.

[35] Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M. Slepchenko, Aleksei A. Bondarev, Philip L.F. Johnson, Ayinuer Aximu-Petri, Kay Prüfer, Cesare de Filippo, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 514(7523):445–449, 2014.

[36] Qiaomei Fu, Alissa Mittnik, Philip L.F. Johnson, Kirsten Bos, Martina Lari, Ruth Bollongino, Chengkai Sun, Liane Giemsch, Ralf Schmitz, Joachim Burger, et al. A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology*, 23(7):553–559, 2013.

[37] Marie-Theres Gansauge and Matthias Meyer. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, 8(4):737–748, 2013.

[38] Marie-Theres Gansauge and Matthias Meyer. Selective enrichment of damaged DNA molecules for ancient genome sequencing. *Genome Research*, 24(9):1543–1549, 2014.

[39] Marc Garcia-Garcera, Elena Gigli, Federico Sanchez-Quinto, Oscar Ramirez, Francesc Calafell, Sergi Civit, and Carles Lalueza-Fox. Fragmentation of contaminant and endogenous DNA in ancient samples determined by shotgun sequencing; prospects for human palaeogenomics. *PLOS ONE*, 6(8):e24161, 2011.

[40] Michael Gerth, Marie-Theres Gansauge, Anne Weigert, and Christoph Bleidorn. Phylogenomic analyses uncover origin and spread of the Wolbachia pandemic. *Nature Communications*, 5, 2014.

[41] M. Thomas P. Gilbert, Hans-Jürgen Bandelt, Michael Hofreiter, and Ian Barnes. Assessing ancient DNA studies. *Trends in Ecology & Evolution*, 20(10):541–544, 2005.

[42] Richard E. Green, Adrian W. Briggs, Johannes Krause, Kay Prüfer, Hernán A. Burbano, Michael Siebauer, Michael Lachmann, and Svante Pääbo. The Neandertal genome and ancient DNA authenticity. *The EMBO Journal*, 28(17):2494–2502, 2009.

[43] Richard E. Green, Johannes Krause, Susan E. Ptak, Adrian W. Briggs, Michael T. Ronan, Jan F. Simons, Lei Du, Michael Egholm, Jonathan M. Rothberg, Maja Paunovic, et al. Analysis of one million base pairs of Neanderthal DNA. *Nature*, 444(7117):330–336, 2006.

[44] Richard E. Green, Anna-Sapfo Malaspinas, Johannes Krause, Adrian W. Briggs, Philip L.F. Johnson, Caroline Uhler, Matthias Meyer, Jeffrey M. Good, Tomislav Maricic, Udo Stenzel, et al. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134(3):416–426, 2008.

[45] Paul D.N. Hebert, Alina Cywinska, Shelley L. Ball, et al. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1512):313–321, 2003.

[46] Thorsten Joachims. Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight.joachims.org/, University of Dortmund*, 19, 1999.

[47] Hákon Jónsson, Aurélien Ginolhac, Mikkel Schubert, Philip L.F. Johnson, and Ludovic Orlando. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, 29(13):1682–1684, 2013.

[48] Wei-Chun Kao and Yun S Song. naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing. *Journal of Computational Biology*, 18(3):365–377, 2011.

[49] Wei-Chun Kao, Kristian Stevens, and Yun S. Song. Bayescall: A model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Research*, 19(10):1884–1895, 2009.

[50] Kazutaka Katoh and Daron M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

[51] Carlyn S. Keith, Danee O. Hoang, Bruce M. Barrett, Barry Feigelman, Mary C. Nelson, Henry Thai, and Chris Baysdorfer. Partial sequence analysis of 130 randomly selected maize cDNA clones. *Plant Physiology*, 101(1):329–332, 1993.

[52] Motoo Kimura. Evolutionary Rate at the Molecular Level. *Nature*, 217(5129):624–626, 1968.

[53] Martin Kircher. Analysis of high-throughput ancient DNA sequencing data. In *Ancient DNA*, pages 197–228. Springer, 2012.

[54] Martin Kircher, Susanna Sawyer, and Matthias Meyer. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, 40(1):e3–e3, 2012.

[55] Martin. Kircher, Udo Stenzel, and Janet Kelso. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biology*, 10(8):R83, 2009.

[56] Anita Kloss-Brandstätter, Dominic Pacher, Sebastian Schönherr, Hansi Weissensteiner, Robert Binna, Günther Specht, and Florian Kronenberg. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human Mutation*, 32(1):25–32, 2011.

[57] Yong Kong. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics*, 98(2):152–153, 2011.

[58] Thorfinn S. Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1):356, 2014.

[59] Johannes Krause, Qiaomei Fu, Jeffrey M. Good, Bence Viola, Michael V. Shunkov, Anatoli P. Derevianko, and Svante Pääbo. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, 464(7290):894–897, 2010.

[60] Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, Joshua G. Schraiber, Sergi Castellano, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, 2014.

[61] Christian Ledergerber and Christophe Dessimoz. Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics*, 12(5):489–497, 2011.

[62] Anat Levin, Yair Weiss, Fredo Durand, and William T. Freeman. Understanding blind deconvolution algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2354–2367, 2011.

[63] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[64] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[65] Stinus Lindgreen. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Research Notes*, 5(1):337, 2012.

[66] Binghang Liu, Jianying Yuan, Siu-Ming Yiu, Zhenyu Li, Yinlong Xie, Yanxiang Chen, Yujian Shi, Hao Zhang, Yingrui Li, Tak-Wah Lam, et al. COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics*, 28(22):2870–2874, 2012.

[67] Ari Löytynoja and Nick Goldman. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10557–10562, 2005.

[68] Lira Mamanova, Alison J. Coffey, Carol E. Scott, Iwanka Kozarewa, Emily H. Turner, Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J. Turner. Target-enrichment strategies for next-generation sequencing. *Nature Methods*, 7(2):111–118, 2010.

[69] Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju Chen, Zhoutao Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[70] Marcel Martin and Sven Rahmann. From cutadapt to sequencetools (sqt): a versatile toolset for sequencing projects. *EMBnet. journal*, 17(B):p–35, 2012.

[71] Andre P. Masella, Andrea K. Bartram, Jakub M. Truszkowski, Daniel G. Brown, and Josh D. Neufeld. PANDAseq: paired-end assembler for Illumina sequences. *BMC Bioinformatics*, 13(1):31, 2012.

[72] Tim Massingham and Nick Goldman. All Your Base: a fast and accurate probabilistic approach to base calling. *Genome Biology*, 13(2):R13, 2012.

[73] Clare M. McCourt, Darragh G. McArt, Ken Mills, Mark A. Catherwood, Perry Maxwell, David J. Waugh, Peter Hamilton, Joe M. O'Sullivan, and Manuel Salto-Tellez. Validation of next generation sequencing technologies in comparison to current diagnostic gold standards for BRAF, EGFR and KRAS mutational analysis. *PLOS ONE*, 8(7):e69604, 2013.

[74] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[75] Thomas Metzinger. *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books, 2010.

[76] Matthias Meyer, Qiaomei Fu, Ayinuer Aximu-Petri, Isabelle Glocke, Birgit Nickel, Juan-Luis Arsuaga, Ignacio Martínez, Ana Gracia, José María Bermúdez de Castro, Eudald Carbonell, et al. A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature*, 505(7483):403–406, 2014.

[77] Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G Schraiber, Flora Jay, Kay Prüfer, Cesare de Filippo, et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338(6104):222–226, 2012.

[78] Matthias Meyer, Udo Stenzel, Sean Myles, Kay Prüfer, and Michael Hofreiter. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research*, 35(15):e97, 2007.

[79] André E. Minoche, Juliane C. Dohm, Heinz Himmelbauer, et al. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, 12(11):R112, 2011.

[80] Hermann J. Muller. Evolution by mutation. *Bulletin of the American Mathematical Society*, 64(4):137–160, 1958.

[81] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C. Linak, Aki Hirai, Hiroki Takahashi, et al. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13):e90–e90, 2011.

[82] James P. Noonan, Graham Coop, Sridhar Kudaravalli, Doug Smith, Johannes Krause, Joe Alessi, Feng Chen, Darren Platt, Svante Pääbo, Jonathan K. Pritchard, et al. Sequencing and analysis of Neanderthal genomic DNA. *Science*, 314(5802):1113–1118, 2006.

[83] Iñigo Olalde, Morten E. Allentoft, Federico Sánchez-Quinto, Gabriel Santpere, Charleston W.K. Chiang, Michael DeGiorgio, Javier Prado-Martinez, Juan Antonio Rodríguez, Simon Rasmussen, Javier Quilez, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature*, 507(7491):225–228, 2014.

[84] Svante Pääbo, Hendrik Poinar, David Serre, Viviane Jaenicke-Després, Juliane Hebler, Nadin Rohland, Melanie Kuch, Johannes Krause, Linda Vigilant, and Michael Hofreiter. Genetic analyses from ancient DNA. *Annual. Reviews of Genetics*, 38:645–679, 2004.

[85] Matthew Parks and David Lambert. Impacts of low coverage depths and postmortem DNA damage on variant calling: a simulation study. *BMC Genomics*, 16(1):19, 2015.

[86] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[87] Hendrik N. Poinar, Carsten Schwarz, Ji Qi, Beth Shapiro, Ross D.E. MacPhee, Bernard Buigues, Alexei Tikhonov, Daniel H. Huson, Lynn P. Tomsho, Alexander Auch, et al. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, 311(5759):392–394, 2006.

[88] Kay Prüfer and Matthias Meyer. Comment on "Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans". *Science*, 347(6224):835–835, 2015.

[89] Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, et al. The complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505(7481):43–49, 2014.

[90] Kay Prüfer, Udo Stenzel, Michael Hofreiter, Svante Pääbo, Janet Kelso, and Richard E. Green. Computational challenges in the analysis of ancient DNA. *Genome Biology*, 11(5):R47, 2010.

[91] Maanasa Raghavan, Pontus Skoglund, Kelly E. Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen, Thomas W. Stafford Jr, Ludovic Orlando, Ene Metspalu, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, 2013.

[92] Andrew Rambaut, Simon Y.W. Ho, Alexei J. Drummond, and Beth Shapiro. Accommodating the effect of ancient DNA damage on inferences of demographic histories. *Molecular Biology and Evolution*, 26(2):245–248, 2009.

[93] Jeffrey G. Reid, Andrew Carroll, Narayanan Veeraraghavan, Mahmoud Dahdouli, Andreas Sundquist, Adam English, Matthew Bainbridge, Simon White, William Salerno, Christian Buhay, et al. Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics*, 15(1):30, 2014.

[94] Gabriel Renaud, Martin Kircher, Udo Stenzel, and Janet Kelso. freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. *Bioinformatics*, 29(9):1208–1209, 2013.

[95] Gabriel Renaud, Viviane Slon, Ana T. Duggan, and Janet Kelso. schmutzi: Contamination estimate and endogenous mitochondrial consensus calling for ancient DNA. submitted, 2015.

[96] Gabriel Renaud, Udo Stenzel, and Janet Kelso. leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Research*, 42(18):e141, 2014.

[97] Gabriel Renaud, Udo Stenzel, Tomislav Maricic, Victor Wiebe, and Janet Kelso. deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics*, 31(5):770–772, 2015.

[98] Nadin Rohland and Michael Hofreiter. Ancient DNA extraction from bones and teeth. *Nature Protocols*, 2(7):1756–1762, 2007.

[99] Jonathan M. Rothberg, Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, Kim Johnson, Mark J. Milgrew, Matthew Edwards, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011.

[100] Matthew Ruffalo, Thomas LaFramboise, and Mehmet Koyutürk. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20):2790–2796, 2011.

[101] Frederick Sanger, Steven Nicklen, and Alan R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[102] Susanna Sawyer, Johannes Krause, Katerina Guschanski, Vincent Savolainen, and Svante Pääbo. Temporal patterns of nucleotide misincorporations and dna fragmentation in ancient DNA. *PLOS ONE*, 7(3):e34131, 2012.

[103] Mikkel Schubert, Aurelien Ginolhac, Stinus Lindgreen, John F. Thompson, Khaled A.S. Al-Rasheid, Eske Willerslev, Anders Krogh, and Ludovic Orlando. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, 13(1):178, 2012.

[104] Wei Shao, Valerie F. Boltz, Jonathan E. Spindler, Mary F. Kearney, Frank Maldarelli, John W. Mellors, Claudia Stewart, Natalia Volfovsky, Alexander Levitsky, Robert M. Stephens, et al. Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of Low-frequency drug resistance mutations in HIV-1 DNA. *Retrovirology*, 10:18, 2013.

[105] Jay Shendure, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon, Abraham M. Rosenbaum, Michael D. Wang, Kun Zhang, Robi D. Mitra, and George M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, 2005.

[106] Pontus Skoglund, Bernd H. Northoff, Michael V. Shunkov, Anatoli P. Derevianko, Svante Pääbo, Johannes Krause, and Mattias Jakobsson. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, 111(6):2229–2234, 2014.

[107] Viviane Slon, Isabelle Glocke, Ran Barkai, Avi Gopher, Israel Hershkovitz, and Matthias Meyer. Mammalian mitochondrial capture, a tool for rapid screening of DNA preservation in faunal and undiagnostic remains, and its application to Middle Pleistocene specimens from Qesem cave (Israel). *Quaternary International*, 2015.

[108] Sören Sonnenburg, Gunnar Rätsch, Sebastian Henschel, Christian Widmer, Jonas Behr, Alexander Zien, Fabio de Bona, Alexander Binder, Christian Gehl, and

Vojtěch Franc. The SHOGUN machine learning toolbox. *The Journal of Machine Learning Research*, 99:1799–1802, 2010.

[109] Claudia Stäubert, Diana Le Duc, and Torsten Schöneberg. Examining the dynamic evolution of G protein-coupled receptors. In *G Protein-Coupled Receptor Genetics*, pages 23–43. Springer, 2014.

[110] Mannis Van Oven and Manfred Kayser. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, 30(2):E386–E394, 2009.

[111] Jeffrey D. Wall and Sung K. Kim. Inconsistencies in Neanderthal genomic DNA sequences. *PLOS Genetics*, 3(10):e175, 2007.

[112] James D. Watson and Francis H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[113] Nava Whiteford, Tom Skelly, Christina Curtis, Matt E. Ritchie, Andrea Löhr, Alexander Wait Zaranek, Irina Abnizova, and Clive Brown. Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics*, 25(17):2194–2199, 2009.

[114] Sewall Wright. Evolution in Mendelian Populations. *Genetics*, 16(2):97–159, 1931.

[115] Chengxi Ye, Chiaowen Hsiao, and Héctor Corrada Bravo. BlindCall: ultra-fast base-calling of high-throughput sequencing data by blind deconvolution. *Bioinformatics*, 30(9):1214–1219, 2014.

[116] Ming Yi, Yongmei Zhao, Li Jia, Mei He, Electron Kebebew, and Robert M Stephens. Performance comparison of SNP detection tools with illumina exome sequencing data–an assessment using both family pedigree information and sample-matched SNP array data. *Nucleic Acids Research*, 42(12):e101, 2014.

[117] Katarzyna Zaremba Niedźwiedzka and Siv G.E. Andersson. No ancient DNA damage in Actinobacteria from the Neanderthal bone. *PLOS ONE*, 8(5):e62799, 2013.

[118] Jiajie Zhang, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5):614–620, 2014.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die aufgeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien und erbrachten Dienstleistungen als solche gekennzeichnet.

Hiermit erkenne ich die Promotionsordnung der Fakultät für Mathematik und Informatik der Universität Leipzig vom 1 Juli 2015 an. Die eingereichte Arbeit wurde ich gleicher oder ähnlicher Form nicht einer anderen Prüfungsbehörde zum Zwecl einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt. Abbildungen, Tabellen und Texte dieser Arbeit wurden in Teilen bereits in Publikationen präsentiert von welchen ich einer der Hauptauthoren bin (siehe Referenzen [94, 95, 96, 97]).

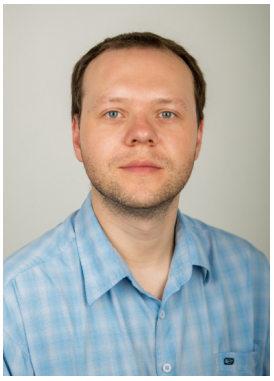
Leipzig, 30. Juni 2015

Gabriel Renaud

Gabriel Renaud

Leipzig, Germany
gabriel_renaud@eva.mpg.de

Mailing address:	Deutscher Platz 6 04103 Leipzig, Germany
Phone:	+49 176 3042 5860
Citizenship:	Canadian
Languages:	English (native) • French (native level) • Spanish (fluent) • Portuguese (fluent) • German (Intermediate) • Italian (intermediate) • Russian (basic) • Chinese Mandarin (basic)
Membership:	American Mathematical Society • International Society for Computational Biology • Brazilian Association for Bioinformatics and Computational Biology



Education

2011-2016 Ph.D, **Max Planck Institute For Evolutionary Anthropology**
Doctor rerum naturalium, Computer Science
Leipzig, Saxony, Germany | Advisor: Dr. Janet Kelso
Magna Cum Laude
Thesis: "Bayesian maximum *a posteriori* algorithms for modern and ancient DNA"

2003-2005 M.Math, **University of Waterloo**
Master of Mathematics, Computer Science
Waterloo, Ontario, Canada | Advisor: Dr. Brendan McConkey
Thesis: "Protein secondary structure prediction using inter-residue contacts"

2000-2003 B.Sc, **University of Montreal**
Bachelor of Science with Mention of Excellence, Theoretical Computer Science Option
Montreal, Quebec, Canada
Dean's Honor List
extra biochemistry classes from the University of Massachusetts, Boston, MA
GPA: 4.1/4.3

Work Experience

2009 to 2011 | Bioinformatician | National Institute for Cancer of Brazil, Rio de Janeiro, Brazil

Within the Bioinformatics and Computational Biology Lab of the National Institute for Cancer of Brazil (INCA) to provide bioinformatics expertise to various research projects.

- Participated in the Breast Cancer Sequencing Project, a consortium aiming at sequencing the genome of a patient with breast cancer:
  - Mapped RNA-Seq data from tumor tissues, performed transcript quantification
  - Designed and implemented a series of Perl modules and components in C++ to annotate SNPs
- Analysis of RNA-Seq data stemming from glioblastoma tissues using SOLiD reads in collaboration with Children's Memorial Hospital, Chicago, IL.

2005 to 2009 | Scientific Programmer | National Institutes of Health, Bethesda, MD, USA

NHGRI Bioinformatics Core to provide bioinformatics support to biological/medical research projects through the design of software tools.

- Developing new algorithms to suit the custom needs of researchers e.g.:
  - Designed and implemented in C++ two new algorithms for the genomic mapping of short tags
  - Developed a bioinformatics pipeline to map sequenced retroviral integrations
- Design and extension of software for numerous research projects in Perl and C/C++ ex:
  - Annotation of retroviral integration sites and generation of random decoys
  - Analysis of DNAse-chip data and generation of stochastic datasets to compute p-values

2003 to 2005 | Teaching Assistant | University of Waterloo, Waterloo, ON, Canada

Various topics including: Introduction to Java (1 term), Theory of Computation (1 term) and Algorithms (3 terms including 1 term as head TA).

- Explained course material to students to facilitate the learning process
- Maintained course website which allowed students to download relevant material and to retrieve their marks from the database, implemented in PHP
- Marked and proctored exams and assignments for evaluation purposes (180 students)
- Supervised a team of graduate students to plan and orchestrate teaching logistics

2002 | Research Assistant | University of Montreal, Montreal, QC, Canada

Summer internship in the Computational Linguistics Lab, funding was provided by the National Science and Engineering Research Council (NSERC).

- Upgraded and extended a Java based web-agent to automatically search for parallel texts to build a parallel corpora (e.g. English-Spanish) to train statistical translation engines
- Implemented distributed computing using RMI to exploit the existing network to increase efficiency

Winter 2000 | Chemistry Tutor | College Ahuntsic, Montreal, QC, Canada

- Tutored a first year student on a weekly basis to assist him with his chemistry course.
- Helped a student that had previously failed the class to obtain a mark above 80

Awards and Honors

2013-2015 **National Science and Engineering Research Council PGS D Scholarship**
NSERC Postgraduate Scholarship Doctoral (PGS D) Holder at a Foreign Institution. Highly competitive scholarship awarded to top-ranked applicants with an innovative research proposal to conduct research overseas. Value: $21,000 CAD per year

2008 **National Human Genome Research Institute (NHGRI) Merit Award**
For contribution to the Human Microbiome Project, part of the NIH Roadmap for medical research

2004-2005 **National Science and Engineering Research Council Graduate Scholarship**
Prestigious scholarship awarded to promising graduate students. Value: $17,500 CAD per year

2004 **Ontario Graduate Scholarship (OGS)**
Scholarship awarded to outstanding students. Value: $10,000 CAD per year

2003-2004 **University of Waterloo Graduate Scholarship**
Scholarship awarded to graduate students with first-class standing. Value: $10,000 CAD per year

2003 **Dean's Honor List and Mention of Excellence**
For maintaining an average of 4.1/4.3 during undergraduate studies

2003 **Computer Science Games at McGill University**
Part of a team that finished second out of 22 teams from different universities from Eastern Canada and the Northeastern United States

Journal Reviewing

- BMC Genomics
- Bioinformatics
- PLOS One
- BMC Bioinformatics

Publications

Manuscripts submitted or in preparation

- Fernando Racimo, **Gabriel Renaud**, and Montgomery Slatkin. Joint estimation of contamination, sequencing error and demography for nuclear dna from ancient humans. resubmitted to PLOS Genetics, 2016

Journal Articles

- **Gabriel Renaud**, Viviane Slon, Ana T. Duggan, and Janet Kelso. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biology*, 16:224, 2015
- Diana Le Duc, **Gabriel Renaud**, Arunkumar Krishnan, Markus Sällman Almén, Leon Huynen, Sonja J. Prohaska, Matthias Ongyerth, Bárbara D. Bitarello, Helgi B. Schiöth, Michael Hofreiter, Peter F. Stadler, Kay Prüfer, David Lambert, Janet Kelso, and Torsten Schöneberg. Kiwi genome provides insights into evolution of a nocturnal lifestyle. *Genome Biology*, 16:147, 2015
- **Gabriel Renaud***, Susanna Sawyer*, Bence Viola, Jean-Jacques Hublin, Marie Theres-Gansauge, Michail Shunkovc, Anatoly Dereviankoc, Kay Pruefer, Janet Kelso, and Svante Pääbo. Nuclear and mitochondrial DNA sequences from two Denisovan individuals. *Proceedings of the National Academy of Sciences*, 112(51):15696–15700, Dec 2015
- **Gabriel Renaud**, Udo Stenzel, Tomislav Maricic, Victor Wiebe, and Janet Kelso. deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics*, 31(5):770–772, Mar 2015
- **Gabriel Renaud***, Matthew C LaFave*, Jin Liang, Tyra G Wolfsberg, and Shawn M Burgess. trieFinder: an efficient program for annotating Digital Gene Expression (DGE) tags. *BMC Bioinformatics*, 15:329, 2014
- **Gabriel Renaud**, Udo Stenzel, and Janet Kelso. leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Research*, 42(18):e141, Oct 2014
- Iosif Lazaridis, Nick Patterson, Alissa Mittnik, **Gabriel Renaud**, Swapan Mallick, Karola Kirsanow, Peter H Sudmant, Joshua G Schraiber, Sergi Castellano, Mark Lipson, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, Sep 2014
- Sebastian Lippold, Hongyang Xu, Albert Ko, Mingkun Li, **Gabriel Renaud**, Anne Butthof, Roland Schröder, Mark Stoneking, et al. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. *Investigative Genetics*, 5:13, 2014
- Sergi Castellano, Genís Parra, Federico A Sánchez-Quinto, Fernando Racimo, Martin Kuhlwilm, Martin Kircher, Susanna Sawyer, Qiaomei Fu, Anja Heinze, Birgit Nickel, Jesse Dabney, Michael Siebauer, Louise White, Hernán A Burbano, **Gabriel Renaud**, Udo Stenzel, Carles Lalueza-Fox, et al. Patterns of coding variation in the complete exomes of three Neandertals. *Proceedings of the National Academy of Sciences*, 111(18):6666–6671, 2014
- Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, **Gabriel Renaud**, Peter H Sudmant, Cesare de Filippo, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49, 2014
- **Gabriel Renaud**, Martin Kircher, Udo Stenzel, and Janet Kelso. freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. *Bioinformatics*, 29(9):1208–1209, May 2013
- Jin Liang, Dongmei Wang, **Gabriel Renaud**, Tyra G Wolfsberg, Alexander F Wilson, and Shawn M Burgess. The stat3/socs3a pathway is a key regulator of hair cell regeneration in zebrafish. *The Journal of Neuroscience*, 32(31):10662–10673, Aug 2012
- Raphael Tavares, **Gabriel Renaud**, Paulo Sergio Lopes Oliveira, Carlos G Ferreira, Emmanuel Dias-Neto, and Fabio Passetti. Identical sequence patterns in the ends of exons and introns of human protein-coding genes. *Computational Biology and Chemistry*, 36:55–61, 2012

Publications (continued)

- Fabricio F Costa, Jared M Bischof, Elio F Vanin, Rishi R Lulla, Simone T Sredni, Veena Rajaram, Min Wang, **Gabriel Renaud**, Fabio Passetti, Tadanori Tomita, et al. Identification of differentially expressed microRNAs and other non-coding RNAs in ependymomas. *Cancer Research*, 72(8 Supplement):198–198, 2012
- Daphne W Bell, Nilabja Sikdar, Kyoo-young Lee, Jessica C Price, Raghunath Chatterjee, Hee-Dong Park, Jennifer Fox, Masamichi Ishiai, Meghan L Rudd, Lana M Pollock, Sarah K Fogoros, Hassan Mohamed, Christin L Hanigan, Suiyuan Zhang, Pedro Cruz, **Gabriel Renaud**, Nancy F Hansen, et al. Predisposition to cancer caused by genetic and functional defects of mammalian Atad5. *PLoS Genetics*, 7(8):e1002245, 2011
- Mariana Emerenciano, **Gabriel Renaud**, Mariana Sant'Ana, Caroline Barbieri, Fabio Passetti, Maria S Pombo-de Oliveira, and Brazilian Collaborative Study Group of Infant Acute Leukemia. Challenges in the use of NG2 antigen as a marker to predict MLL rearrangements in multi-center studies. *Leukemia Research*, 35(8):1001–1007, 2011
- **Gabriel Renaud**, Pedro Neves, Edson Luiz Folador, Carlos Gil Ferreira, and Fabio Passetti. Segtor: rapid annotation of genomic coordinates and single nucleotide variations using segment trees. *PLOS ONE*, 6(11):e26715, 2011
- Donna M Toleno, **Gabriel Renaud**, Tyra G Wolfsberg, Munirul Islam, Derek E Wildman, Kimberly D Siegmund, and Joseph G Hacia. Development and evaluation of new mask protocols for gene expression profiling in humans and chimpanzees. *BMC Bioinformatics*, 10(1):77, 2009
- Stacie K Loftus, Anthony Antonellis, Ivana Matera, **Gabriel Renaud**, Laura L Baxter, Duncan Reid, Tyra G Wolfsberg, Yidong Chen, ChenWei Wang, Megana K Prasad, et al. Gpnmb is a melanoblast-expressed, MITF-dependent gene. *Pigment Cell & Melanoma Research*, 22(1):99–110, 2009
- Tiffany C Scharschmidt, Karin List, Elizabeth A Grice, Roman Szabo, **Gabriel Renaud**, Chyi-Chia R Lee, Tyra G Wolfsberg, Thomas H Bugge, and Julia A Segre. Matriptase-deficient mice exhibit ichthyotic skin with a selective shift in skin microbiota. *Journal of Investigative Dermatology*, 129(10):2435–2442, 2009
- Yoo-Jin Kim, Yoon-Sang Kim, Andre Larochelle, **Gabriel Renaud**, Tyra G Wolfsberg, Rima Adler, Robert E Donahue, Peiman Hematti, Bum-Kee Hong, Jean Roayaei, et al. Sustained high-level polyclonal hematopoietic marking and transgene expression 4 years after autologous transplantation of rhesus macaques with SIV lentiviral vector–transduced CD34+ cells. *Blood*, 113(22):5434–5443, 2009
- Elizabeth A Grice, Heidi H Kong, **Gabriel Renaud**, Alice C Young, Gerard G Bouffard, Robert W Blakesley, Tyra G Wolfsberg, Maria L Turner, and Julia A Segre. A diversity profile of the human skin microbiota. *Genome Research*, 18(7):1043–1050, 2008
- Brian L Pike, Timothy C Greiner, Xiaoming Wang, Dennis D Weisenburger, Ya-Hsuan Hsu, **Gabriel Renaud**, Tyra G Wolfsberg, Myungjin Kim, Daniel J Weisenberger, Kimberly D Siegmund, et al. DNA methylation profiles in diffuse large B-cell lymphoma and their relationship to gene expression status. *Leukemia*, 22(5):1035–1043, 2008
- Anthony Antonellis, Jimmy L Huynh, Shih-Queen Lee-Lin, Ryan M Vinton, **Gabriel Renaud**, Stacie K Loftus, Gene Elliot, Tyra G Wolfsberg, Eric D Green, Andrew S McCallion, et al. Identification of neural crest and glial enhancers at the mouse Sox10 locus through transgenesis in zebrafish. *PLoS Genetics*, 4(9):e1000174, 2008
- Jingqiong Hu, **Gabriel Renaud**, Theotonius Golmes, Andrea Ferris, Paul C Hendrie, Robert E Donahue, Stephen H Hughes, Tyra G Wolfsberg, David W Russell, and Cynthia E Dunbar. Reduced genotoxicity of avian sarcoma leukosis virus vectors in rhesus long-term repopulating cells compared to standard murine retrovirus vectors. *Molecular Therapy*, 16(9):1617–1623, 2008
- Jingqiong Hu, Theotonius Gomes, Andrea Ferris, **Gabriel Renaud**, Paul C Hendrie, Allen E Krouse, Robert E Donahue, Tyra G Wolfsberg, Stephen H Hughes, David W Russell, et al. Distinctive integration profile of avian sarcoma leukosis virus vectors in rhesus long-term repopulating cells. *Blood*, 110(11):66A–66A, 2007

Publications (continued)

- Gregory E Crawford, Sean Davis, Peter C Scacheri, **Gabriel Renaud**, Mohamad J Halawi, Michael R Erdos, Roland Green, Paul S Meltzer, Tyra G Wolfsberg, and Francis S Collins. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature Methods*, 3(7):503–509, 2006

Book Chapters

- Jin Liang, **Gabriel Renaud**, and Shawn M Burgess. Sequencing-based expression profiling in Zebrafish. *Methods in Cell Biology*, 104:379, 2011

Theses

- **Gabriel Renaud**. *Protein Secondary Structure Prediction using inter-residue contacts*. Master's thesis, University of Waterloo, 2005

Invited Talks

2015	McMaster University, Hamilton, ON, Canada *Computational endogenous consensus calling and contamination estimate for ancient hominin samples.*
2014	Leipzig University, Leipzig, Germany *Processing of Next Generation Sequencing*
2013	Leipzig University, Leipzig, Germany *The MPI-EVA sequencing pipeline for ancient and modern DNA*
2009	5th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, Angra dos Reis, Brazil *Identical sequence patterns in the ends of exons and introns of human genes*

Teaching Activities

2015	Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany *Ancient DNA from a computational perspective* 21 students
2013	Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany *Introduction to UNIX for Next Generation Sequencing processing* 16 students
2003 to 2005	University of Waterloo, Waterloo, ON, Canada Teaching assistant for Java programming and algorithms >100 students in entire class