**DISSERTATION**

# Addressing Challenges of Ancient DNA Sequence Data Obtained with Next Generation Methods

submitted in fulfillment of the requirements for the degree

Doctorate of natural science

*doctor rerum naturalium*

at the Faculty of Biology

Johannes Gutenberg University Mainz

by

Christian Sell

born 12.12.1980

in Merzig, Germany

Mainz, 26.04.2017

# Abstract

This thesis addresses challenges in the bioinformatic analysis of palaeogenomes that were generated by Next Generation Sequencing of highly degrade ancient DNA from archaeological skeletal remains. It establishes a pipeline that incorporates a correction for postmortem damage as well as sequencing errors, to facilitate the comparison with sequence data from modern specimen. By applying the pipeline to published ancient genomes from the Aegean Neolithic and by comparing the results to data from the 1000 Genomes project, it could be shown that an excess of Cytosine to Thymine transitions linked to deaminations during the postmortem degradation of the DNA, can be reverted by bioinformatic processing.

In another attempt to address the complexity and scarcity of DNA from prehistorical specimen, an in-solution hybridization enrichment was designed. This method can counteract the relatively low endogenous DNA content in samples from prehistoric human skeletal remains by selectively enriching specifically designed regions. The developed capture array was analyzed in 21 skeletal human remains from a Bronze Age battlefield, resulting in an average read depth of 1.71x over the whole genome. The statistical analysis of data produced by this approach enables genomic inferences similar to those based on full genomes.

Third the thesis addresses the false assignments of individual bar-code-indices to sequence samples. In a data set comprising 38 capture enriched mitochondrial genomes from prehistoric human remains, it could be shown that this sequencing error can mimic a cross contamination event during lab work. By identifying and removing affected reads, false positive variants could be reduced from ∼38% to 0%.

# Contents

# 1 Introduction

New technologies like Next Generation Sequencing (NGS) and the enrichment of DNA using hybridization capture approaches have drastically increased the amount of sequence data, that can be produced in a given time period. This is even more so in the field of ancient DNA. While ancient DNA (aDNA) research started with a single 229bp long sequence from the quagga in 1984 [39], today there are more than 30 complete mammalian genomes from ancient individuals available. Most of them from anatomically modern humans [15, 29, 40, 49, 54, 64, 74, 97] but also including archaic hominids like a Neanderthal [92] or a Denisovan genome [80]. While these new techniques generate enormous amounts of information, insights in human evolutionary changes and the reconstruction of population history "they require new methods of data processing and analysis, as well as conceptual changes in interpreting the results" [41]. In this thesis, a pipeline for the bioinformatic analysis of aDNA sequence data will be presented. The pipeline consists of a chain of operations, that each generate the input for the following process. The work flow will include pre-processing of the raw sequence data, aligning the sequences to a reference genome and the thereafter refinement of the alignment. The refinement will specially account for DNA damage and sequencing errors. This pipeline will be applied and tested on sequences from a shotgun sequencing experiment, as well as sequences from a specially developed in solution hybridization enrichment approach, hereafter called "capture". Included in the capture are two kinds of new supposed selective neutral regions. The selection of those and the applied categories will be explained. DNA sequences obtained using the capture, from 21 prehistoric humans will be compared to modern and ancient data using current population genetic methods. Further an approach to remove cross contamination during multiplex sequencing with single indices will be presented. Results of 33 prehistoric human mtDNA, sequenced in parallel will be described, that show how samples with high read depth, can influence other samples sequenced in parallel.

## 1.1 aDNA characteristics

aDNA or DNA from prehistorical samples, in this case human skeletal remains, is exposed to many different environmental conditions and reactions leading to its degradation [102]. As a result of hydrolytic degradation, strand breaks, interstrand cross-links, abasic sites and several atypic nucleotide bases can be formed [83]. Often only trace amounts of degraded endogenous DNA is available in the bone, while large fractions of a DNA extract from an ancient sample will consist of exogenous contamination, mainly microorganisms from the surrounding soil [88]. Especially while working with human samples the risk oft contamination by the researcher is given and has to be prevented with special care during lab work [13]. Two aDNA characteristics are of particular interest during this thesis. One being the low amount of endogenous DNA available in a sample, and the resulting risk of contamination, and the other is nucleotide mis-incorporation mostly driven by deaminated forms of Cytosin (Uracil). The low amount of endogenous DNA can to some extend be compensated by a hybridization capture approach (see below in Chapter 1.2.2), while possible contamination has to be monitored carefully during the analysis of aDNA sequencing data. Since non human contaminants are easily removed during the mapping of the sequences to the human reference genome, it is human contamination that has to be specially accounted for. Such human contamination can result from modern DNA of the researchers, due to cross contamination events during wet lab work or sequencing (for further details see Chapter **5 On lane contamination**). The nucleotide misincorporations resulting from the deaminated Cytosin will result in a typical damage pattern. This damage pattern can be visualized in the sequenced reads after the alignment to a reference. It will manifest as increased frequency of Cytosin (C) to Thymin (T) transitions towards the 5'-end of a read and an increased frequency of Guanin (G) to Adenin (A) mutations towards the 3'-end [14, 101]. This pattern, now referred to as post *mortem* damage (PMD), can

influence variant detection and the resulting summary statistics, like heterozygous rate, if not specially cared for (see Chapter **2 Pipeline** for details). Therefore, C to T and G to A changes have often been removed from modern data sets for comparison with ancient data [29, 106]. PMD patterns, like the elevated frequencies of C to T at the 3'-end or G to A changes at the 5'-end, seem to be dependent on the sequencing library protocol used [79] (see Chapter 1.2.1 below) and can vary if e.g. a single strand protocol is applied [80]. It has to be noted that with low coverage genomes, often no genotypes are called [5, 27, 29, 36, 63, 75, 106] allowing only to compare the most likely allele.

## 1.2 Methods introduction

### 1.2.1 Sequencing library preparation

To apply NGS technologies, DNA molecules need to be transferred into sequencing libraries. The library preparation used here is specially suited for multiplexed sequencing on an Illumina Hiseq sequencer [57, 79]. To amplify and sequence a certain region from both directions using PCR, the classic Sanger sequencing needed two known primer sequences , on opposite sites of a molecule. A NGS library will allow for unspecific amplification and sequencing of all unknown sequences present in an DNA extract. While modern DNA needs to be randomly fragmented in the lab, this can be ignored for aDNA because of its fragmentation due to PMD. To build a library for NGS sequencing, two artificial DNA molecules, called adapters, are attached to both ends of every DNA molecule. A fully realized adapter sequences contains several specialized sequence parts:

- So called anchor sequences, that will hybridize with special oligo nucleotides bound to a surface during the actual sequencing.
- A primer sequence to sequence the read and the optional index.
- The optional index sequences, that are used to separate multiplexed samples after sequencing, bioinformaticaly.
- An optional second primer sequence for a possible second read and second index.


**Figure 1.1:** *Schematic overview of fully realized library molecule indicating the sequencing direction for reads 1 and 2 as well as for the index reads 1 and 2; P5 and P7 = Anchor sequences for the class slide during read 1 and read 2 respectively*

The library protocol and the adapters are designed, to allow only for the ligation of correctly oriented adapter sequences at both ends of the DNA molecule, with different adapter sequences for each end. A full library molecule will consist of the appropriate adapter being bound to both ends of both strands in the correct 5' to 3'-end orientation. Only fully formed library molecules can later be sequenced (see Figure 1.1 for a schematic of a complete library molecule). A completed DNA library involves at least one amplification step of all molecules. This can be done by using the adapters as primers and thus amplifying the whole library unspecific to the included DNA molecules.
While the index sequence is theoretically optional on both ends of the library molecule, besides the cost efficient multiplex sequencing strategy, indices are also used to exclude cross contamination due to wet lab

work. Although using only a single index at one end of the library molecule would be enough to identify unique experiments, it can result in mis-identification or on lane cross contamination events [57]. Therefor a different index at both ends of the library molecule is often used in aDNA libraries (see Chapter **5 On lane contamination** for further details).

### 1.2.2 Capture enrichment approach

Capture means the selective enrichment of selected loci prior to sequencing. Simultaneously, this reduces the exogenous molecules in aDNA sequencing libraries, compared to a so called shotgun sequencing approach, that uses the full library. The method used here is an in solution hybridization capture [32]. It can very easily be described as bait fishing, with certain loci of interest being the fish and a single stranded DNA probe of defined length, as bait. To perform such a capture, certain regions of interest in the studied organisms genome, have to be selected a priori. The regions used here are specific nuclear loci in the human genome as well the whole human mitochondrial genome (see Chapters **3 Nuclear capture enrichment approach** and **5 On lane contamination**). From those regions artificial biotinylated oligo nucleotides, here 80bp long, will be generated, based on a specified tiling. The tiling defines how often each base in the target regions will be covered by the baits. The baits will then be mixed with the DNA library. After denaturation, the baits can hybridize with the library molecules, that match the sequence of the baits. With the use of streptavidin coated magnetic beads, the biotinylated baits binding to the target library molecule can be extracted from the library (see Figure 1.2). By denaturation, the library molecules can be separated from the baits, amplified and treated as a standard DNA library for sequencing.

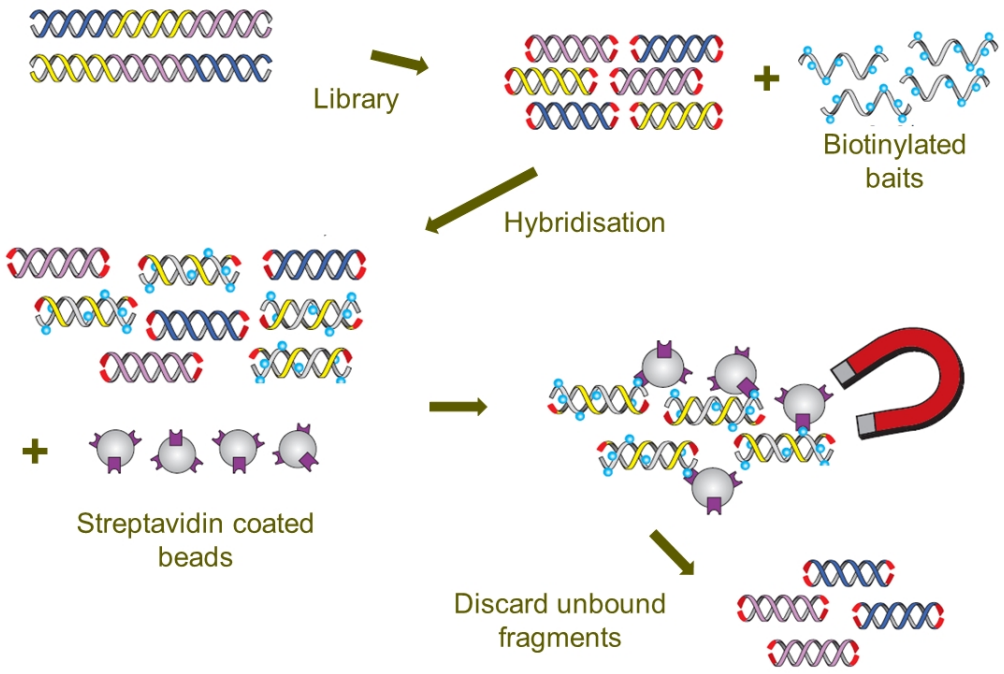



**Figure 1.2:** *Schematic overview of an in solution hybridization capture, based on the Agilent Sure Select*

### 1.2.3 Illumina HiSeq sequencing

To sequence a DNA library on an Illumina Hiseq sequencer, the library molecules have to be clonally amplified to increase the fluorescent signal in the latter sequencing reaction. This amplification is done in a solid phase amplification or bridge PCR, on a glass slide. Two different oligo nucleotides called anchors, one with a sequence matching the anchor sequences in the one library adapter and the other matching the reverse complement of the anchor sequence in the other adapter, are covalently bound to the class slide. When the

DNA library gets flushed over this class slide and denaturated, one of the anchors can hybridize with the appropriate anchor sequence in one strand of the library molecule while the other strand gets washed away. Using a single cycle of PCR, the anchor can then be used as a primer and the reverse complement of the hybridized library molecule will be elongated out of the anchor, thus creating a covalently bound library molecule on the glass slide. Then the hybridized strand of the library molecules is removed from the glass slide. Now the 3'-end of the synthesized strand can be hybridized with the other anchor present on the glass slide and can again be used for a single elongation, building a bridge between the two anchors. This procedure is repeated several times, generating dense sequence clusters on the glass slide, with each cluster consisting of copies of a single library molecule (compare Figure 1.3). After enzymatic removal of one of the strands, sequencing by syntheses can be performed, using the signal of one whole sequence cluster for one read.
Sequencing by synthesis is performed by using nucleotides that are color labeled with a dye and are reversibly blocked at the 3'-end. A so called blocker prevents any further elongation of the molecule as soon as one nucleotide is build in, while the dye labels the base, that is included in the nucleotide. The attached dye will be excited with a laser and the signal is monitored. In a following step the blocking group at the 3'-end of the nucleotide is removed, allowing a new nucleotide to be build in and the dye of the new nucleotide to be read again. In that way a whole sequence can be synthesized and simultaneously read. A read is generated by sequentially listing all monitored bases at a certain sequence cluster. To start the elongation a sequence matching the 3'-end of the adapter at the 5'-end of the library molecule is used. Simultaneously the 3'-end of the library molecule is attached to the glass slide (compare Figure 1.4).
To allow for paired end sequencing the reverse complement of the already sequenced strand is elongated using the above described bridge amplification. After elongation the already sequenced strand is enzymatically cleaved from the anchor and washed from the glass slide. Using a matching primer to start the elongation for the sequencing by synthesis on the other strand of the library molecule, as the second read. Index sequences are read with different primers separated from each read for each of the strands. A paired end sequencing using a multiplex approach with two indices, one on each side of the library molecule, will result in four different reads per sequence cluster. The affiliation of the reads will then be determined by the location they were detected, on the glass slide (see [10, 57, 78] for further details on the sequencing).

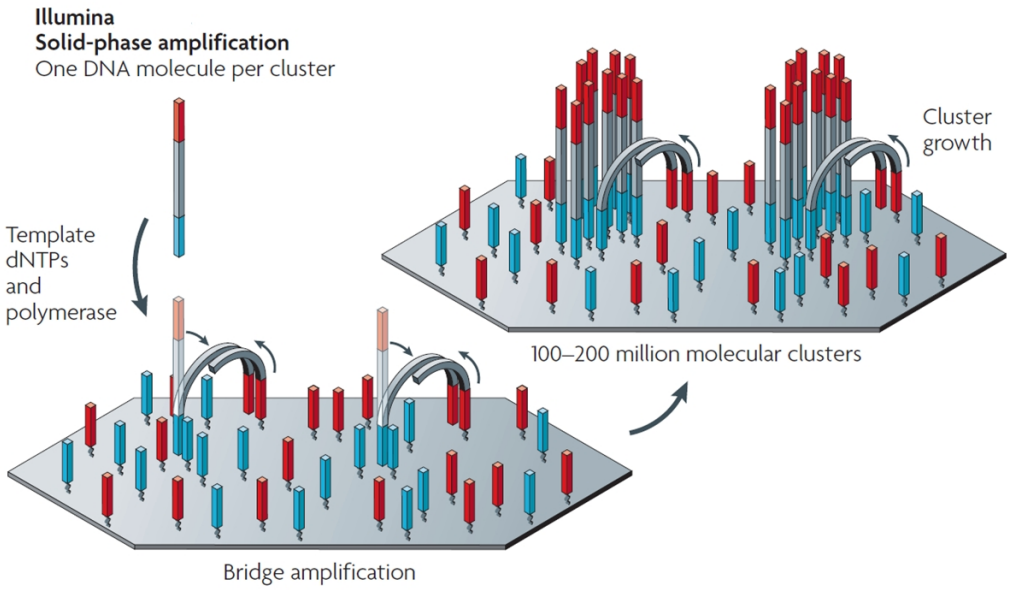



**Figure 1.3:** *Schematic overview of the bridge PCR used prior to the Illumnia HiSeq sequencing; edited from Metzker 2010 [78]*

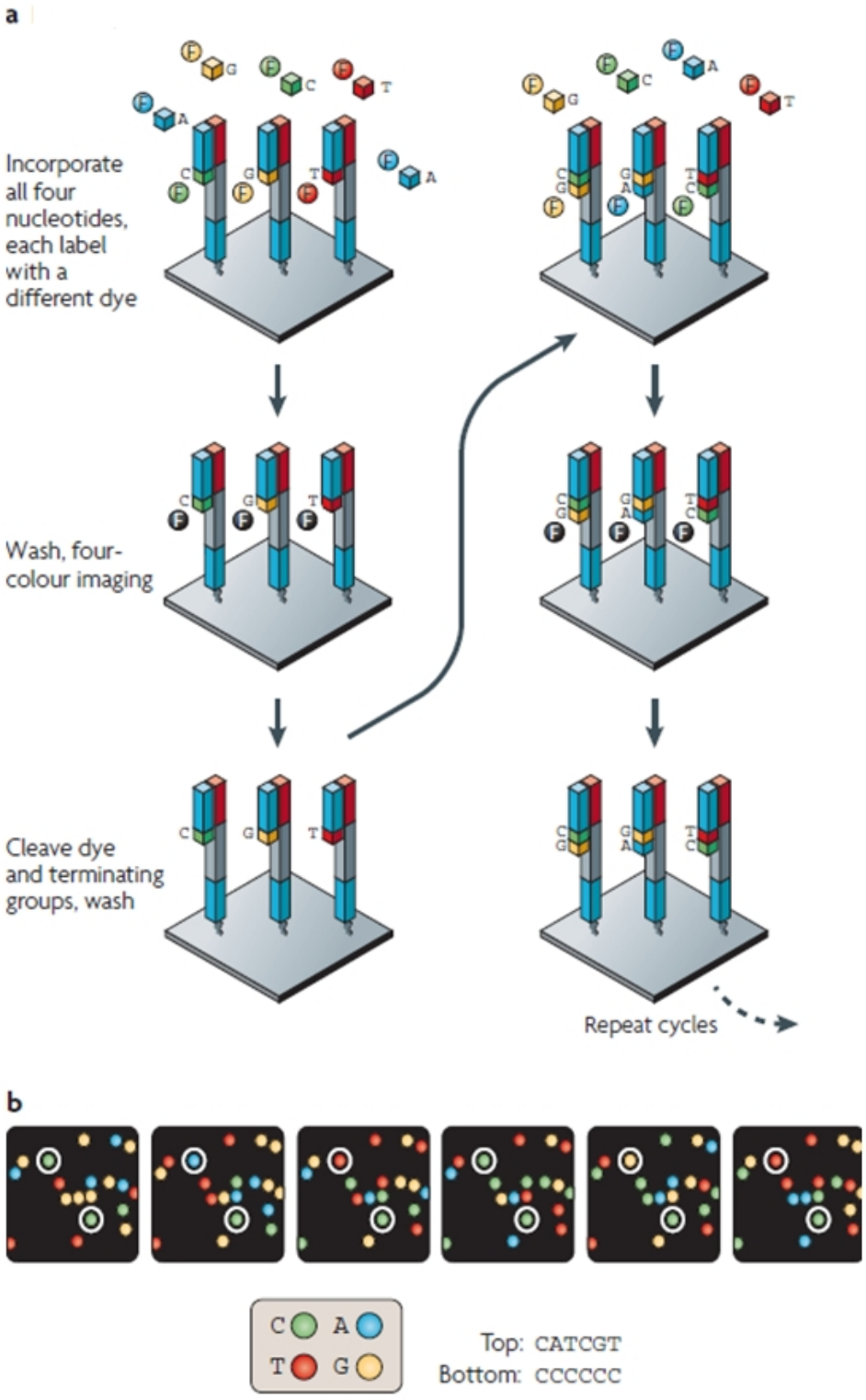



**Figure 1.4:** *Schematic overview the sequencing procedure during the Illumina HiSeq sequencing; edited from Metzker 2010 [78]; a = showing two cycles of sequencing by synthesis in detail, with the incorporation of the labeled nucleotide, the washing and imaging during the excitement of the dye, and the cleaving of the terminator; b = showing the imaging process of six consecutive nucleotides in two different sequences on the same flow cell; circles indicate the sequence clusters top and bottom, with the colour codes used and the resulting sequence shown below.*

## 1.3 Motivation

The aim of this thesis is to combine the use of NGS technologies with aDNA on a population wide scale. To achieve this, a pipeline will be presented and tested for comparability with modern methods and data, while simultaneously reacting on the needs of ancient DNA. The pipeline counteracts the challenges intertwined with aDNA on a bioinformatics level and a developed capture will react on the molecular level. The main questions this thesis tries to answer are:

- Can sequence analysis tools for modern NGS data be used for sequences generated from aDNA?
- Can such aDNA sequence data then be used for comparison with modern data and allow for genotype calls in low coverage data?
- How has aDNA sequence data to be processed informatically to achieve the above?
- Can the influence of post mortem damage (PMD) on the sequence data be reduced through bioinformatics?
- How does a capture approach influence NGS data from aDNA and can it be comparable to a shotgun approach?

To test the above several filter criteria, a correction for PMD as well sequencing error calibration will be applied on published human ancient genomes [40]. The data will then be compared to the unmodified ancient genomes and modern data, sub set from 1000 Genomes [20]. Additionally unpublished capture experiments of 21 prehistoric human remains, related to a battle during the bronze age, will be used to asses the capture, answer archaeological questions using current population genetic methods and compare it with results from published ancient human genomes.

# 2 Pipeline

An analysis pipeline can be understood as a chain of operations that each needs the output of the former to generate a result for the following process. The data analysis pipeline described in the following Chapters (see Figure 2.5) uses a combination of programs and scripts especially suited to handle large ancient DNA (aDNA) data sets generated with Next Generation Sequencing (NGS) Methods. Although the amount of relative high quality genomes with ∼10x read depth from ancient individuals is constantly increasing [15, 27, 29, 49, 64, 92], the majority of ancient human individuals is only available at poor quality, often below a 1x read depth. For the latter variant calls have often been restricted to haploid calls [29, 36, 64, 75]. Those haploid calls are generated by randomly picking one of the present alleles at each position or picking the allele that is the most frequent at a position. The following pipeline is established to generate a reliable alignment from raw sequencing data to allow for further analysis including a diploid variant and genotype call. Its main purpose is to generate data sets that will be comparable to modern sequencing data and therefore tries to include topics such as DNA damage patterns and human contamination whilst using software that is widely used in modern human genomics, like the Genome Analysis Toolkit (GATK) [76] that has been used in the 1000 Genomes [20] project. The goal is to create alignments and variant calls based on those, that can be used for reliable population genetic statistics not biased by possible DNA damage or sequencing artifacts. Since the influence of sequencing error or PMD, can result in the removal of sequence information due to bad quality, this pipeline also focuses on retaining as as much information as possible. In the following subsections the step by step explanation of the work flow including detailed results for the refinement of alignments and their discussion will be described.

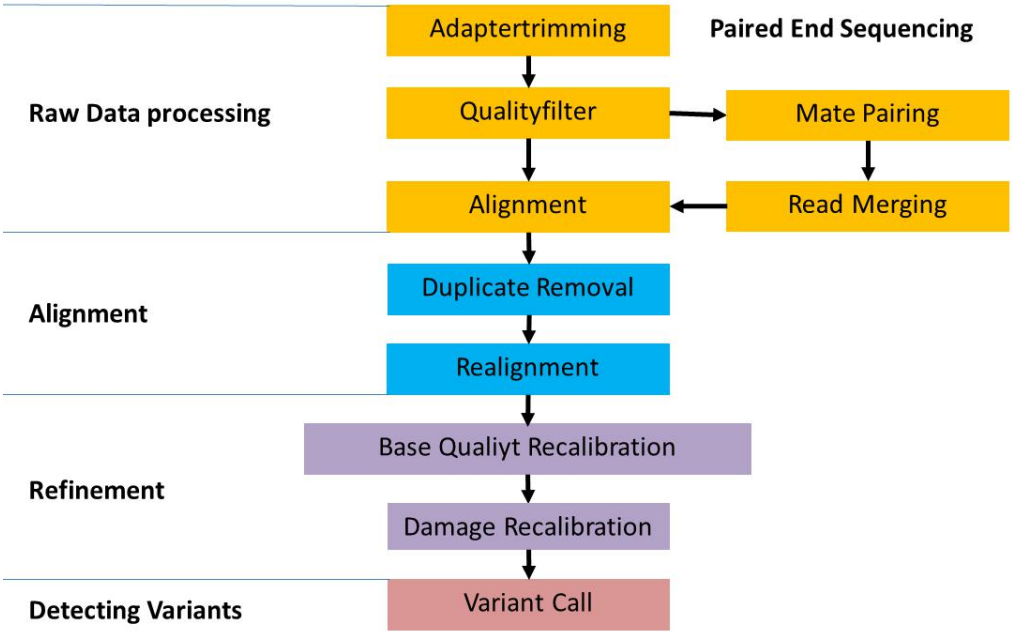



**Figure 2.5:** *Schematic view of the developed pipeline.*

## 2.1 Methods

In the following, each step in the analysis pipeline will be described. This description includes file formats, programs, their functions and the use of parameters.

### 2.1.1 Raw data processing

In general raw data from NGS experiments is available in the fastq format [19]. This is the current default format for raw sequencing data and combines the actual sequence information with a quality value. This quality represents a PHRED scaled probability for an incorrect base being called and thus giving a measurement for the certainty of a base call. For a more detailed explanation of the PHRED scores see below.
The raw data processing involves operations performed directly on the sequences in the fastq format and is independent from the organism or sample sequenced. It is intertwined with the lab methods used to prepare a library and the applied sequencing method. Both, the sample preparation and the sequencing method can have an influence on the base quality, possible contaminating sequences introduced in the data set and sequencing artifacts. One example, the index mis-identification as a special sequencing artifact, is discussed in detail in the Chapter **5 On lane contamination**. The raw data processing reduces the influence of both factors and increase the alignment quality for any single sequence and thereby for the alignment as a whole.

#### *Adapter trimming*

During the sequencing process it can occur that parts of an adapter sequence, used during wet lab library preparation (see Chapter 1.2.1), are read at the 3'-end of a read. This happens when the read length exceeds the length of the actual DNA molecule sequenced. Full or partial adapter sequences need to be removed prior to the alignment. Otherwise, they will decrease the alignment quality of a read or completely prevent the mapping of the read to a reference. The outcome depends on the alignment algorithm and the hereby allowed mismatches as well as the length of the adapter. To remove adapter sequences from the sequence reads, the raw data is processed with the script *KeyAdapterTrimmFastQ_cc.py* [56]. This program tries to overlap a given adapter sequence with every read, starting at the last base of the read and the first base of the adapter. Default parameters are used with the exception that chimeras and key sequences are deactivated. The default parameters allow an overall sequence identity of 90% and a length of one base pair overlap between the adapter and the trimmed sequence. For a paired end sequencing, this program is executed separately on the forward and the reverse read. For each of the reads, the appropriate adapter sequence is given. If a read is trimmed, a "T_" is placed in front of the sequencing name.

#### *Quality filtering*

To ensure the sequencing machine called the correct base at any position, it reports a quality score for each base. This quality score (Q) is PHRED based, meaning it is logarithmically related to the base-calling error probabilities (P).

$$Q = -10\log_{10}P$$

Therefore a high quality represents a low error probability. Thus a base having a high quality is more likely to be correct and vice versa. Base qualities are considered in most steps of the downstream analysis. They influence mapping/alignment quality as well as the later SNP call. In this step, the quality will be checked for every base in each read. If a read contains a certain number of bases below a quality threshold, it will be removed. This is done by the script *QualityFilterFastQ.py* [56]. A minimum quality value of 15 is allowed in 5% of the bases in the read.

***Mate pairing***

In case of paired end sequencing, separate fastq files will be provided for the forward and the reverse reads. Both will be processed separately for the steps prior to mate pairing. Most of the programs used downstream will assume that a pair consisting of the forward and reverse read are found in the same line in their respective files. Therefore the order in which the reads are listed will be relevant for pairing. Since adapter trimming as well as quality filtering can remove reads from each of the fastq files or change the name of a modified sequence without considering the pair information, some reads loose their partner. This will result in a different order of the sequences in the files. In this step, a python script is used to find and remove reads that are only available as either forward or reverse read and sort the files with complete pairs for the forward and reverse reads accordingly (script available as supplementary file remove_reads_no_mate.py; see table A.43). The unpaired reads are called orphans and will be stored in a separate file. In order to allow trimmed reads to be paired, sequence names have to be modified by removing the "T_" at the beginning of each trimmed read.

***Read merging***

After mate pairing, a resulting read pair can be merged into one single read. A merged read resulting from two reads of a paired end read, can increase the mapping quality due to its greater length and higher base qualities compared to any of the two single reads it consists of. This merging is only possible if the actual molecule sequenced was shorter than the length of the forward and the reverse read combined. If this occurs, bases at the 3'-end of the forward read will be reverse complement to parts of the 5'-end in the reverse read. By reverse complementing the reverse read, adding the non-overlapping bases to the 3'-end of the forward read and keeping the base with higher quality at overlapping bases, a read pair can be merged into one read. The merging described in this section is likely to work with the majority of reads in ancient DNA samples due to their small fragment size. To perform the read merging *fastq-join* from *ea-utils* is used with default parameters [8].

### 2.1.2 Alignment generation and processing

After the raw reads have been processed, each remaining sequence will be aligned to a reference genome. All samples analyzed were aligned either against the human genome hg19 (build GRch37) or, for experiments regarding only the human mitochondrial DNA (mtDNA) against the revised Cambridge reference sequence (rCRS) [6]. Both genomes were obtained via NCBI [82]. Mapping is performed to the indexed references using the Burrows-Wheeler Aligner *bwa* version 0.7.5 [65]. Prior to mapping, genomes were indexed with *bwa* using the *is* algorithm for the rCRS and the *bwtsw* algorithm for the hg19. All sequences are mapped after raw data processing with *bwa aln* using default parameters. Alignments are then printed in SAM/BAM format [66] using *bwa samse* for single end sequencing runs, orphan reads and merged reads. For printing the alignments of the non-merged reads *bwa sampe* is used. To save disk space, the alignments are not stored but redirected (piped) into *samtools view* [66], combining printing and filtering in a single step. The filtering removes unmapped reads, reads with mapping quality below 25 and not properly paired reads if the alignment contains non-merged reads from a paired end sequencing run. A read pair is considered a proper pair if both reads are mapped in a reasonable distance given the expected maximum molecule length and are correctly oriented, meaning one is mapped to the reverse strand and the other against the forward strand. The reasonable distance between two reads in a pair, is generally inferred by *bwa* itself during the printing of the alignment, but can be manually set. In the latter case the default setting is 500bp as reasonable distance. Mapping quality is again a PHRED score that represents the error rate of a sequence being falsely aligned to a certain position, which means the lower the quality the higher the error rate.
The before mentioned SAM/BAM (Sequence/Binary Alignment Map) format, is the current standard to save

alignment information. It consists of a header section and an alignment section. The former can include meta information about the alignment, like the reference genome, programs used with the alignment and a read group which includes specific information about the sample and lab protocols used. The alignment section includes all the information about how a read aligns to a given reference in columns separated by tabs. Multiple reads are then separated by lines. The single columns include read name, position in the reference, the sequence, mapping quality, information about mis-matches and indels and more. The difference between SAM and BAM is that the first is a text format that is human readable without further processing and the latter is binary coded, which needs less disk space but cannot be read without decoding it first. For further details see [66].
Duplicate sequences in the alignment, resulting from PCR, are removed from the alignment to ensure that biased amplification of certain molecules during the laboratory procedures will have no influence on any downstream analysis like the SNP call. To remove duplicates *MarkDuplicates* from *Picardtools* [2] is used.

### 2.1.3 Alignment refinement

***The Genome Analysis Toolkit***

The Genome Analysis Toolkit (GATK) [76] is a software package for analyzing next-generation sequencing data. This toolkit is developed and maintained by the Broad Institute and offers multiple tools focusing on variant discovery and genotyping as well as assuring the quality of a generated data set. GATK is used in this pipeline for realigning sequences around indels, base quality recalibration and variant calling.

Indel realignment

The indel realignment process improves the original alignment of the reads, by applying a multiple sequence alignment to positions with possible indels [24, 109] and maximizing the quality of the whole alignment including all contained sequences. In the first alignment with *bwa* a local alignment is performed [65] maximizing the quality of the one read that is currently aligned. Programs used for mapping can falsely map reads, allowing multiple mismatches instead of an actual indel or vice versa. This mis-alignment can result in errors during the following base quality recalibration, variant detection and further downstream analysis. In a first step, positions are located by walking through the alignment and checking each position, if a realignment could increase the global quality. Reads covering those positions will be realigned in a second step with and without possible indels and the alignment with the highest overall mapping quality is then kept at this position.

Base quality recalibration

It is known that sequencing machines are biased in the base quality they report [76]. Those biases can for example be dependent on the prior or the following nucleotide, the di nucleotide context, or the position of the current base in the read. Base quality scores are crucial for all downstream analysis, like the variant call. To remove possible biases, the GATK Base Recalibrator (BQSR) analyzes the co-variation of several features of a base, for example the reported quality score, the position within the read and the preceding nucleotide. This is done by passing through every position that is present in the alignment and not given in a mask. Masks are typically sites with known variation and are needed to differentiate between true variation and possible sequencing errors. It is assumed that all variation is given in a file and therefore every mismatch to the reference is the result of an error. Therefore the data set used for known variation is crucial and should contain as much variation as possible. The data sets used as a mask for recalibration are the ones suggested and included in the GATK resource bundle if not stated otherwise. The work flow suggested by GATK will be referred to as **"default recalibration"** [24, 109].

If no reference population is available or one wants to discover new variants, not given in the known variation of an organism, a mask can be generated from the actual sample. To generate such a data set of known variation, an un-calibrated variant call has to be performed. Resulting variants have to be filtered in a way to secure that no false variation is introduced, for example using only homozygote sites with a genotype quality of >30 (see below in Chapter 2.1.4). The detected and assumed true variants will then be used in the base recalibration step as given variation. This recalibration on the given variation of a data set itself will be referred to as **"self-recalibration"** [24, 44, 109].

***Damage Recalibration***

Next to using the GATK toolkit, a python script provided by Krishna Veeramah was used to recalibrate the base quality according to the known damage patterns generated with *mapDamage* [50] for each read (python script aDNAbamRecal.py is available as supplementary file; see table A.43) . During this process, base qualities of Cytosin and Thymin reads are reduced according to the transition frequencies in relation to the base position in the read relative to the appropriate end (see Chapter 2.1.5 for further details on *mapDamage* and Chapter 1.1 for damage patterns in aDNA).

***Merging of multiple Sequencing runs***

If a given sample was sequenced multiple times in separately lanes, each lane of each sequencing run of the sample was processed separately until after base recalibration. To combine the information of multiple experiments, the recalibrated bam files can be merged into one single alignment using *picardtools MergeSam*.

### 2.1.4 Detecting variants

There are two variant callers available in the GATK toolkit, the *UnifiedGenotyper* and the *HaplotypeCaller*. Although the latter is the newer and more sophisticated tool, it did not allow a variant call on a haploid genome prior to version GATK 3.3 thus still keeping the first more basal tool alive. Here both callers are used. The *UnifiedGenotyper* is used with the *ploidy* set to 1 for the mtDNA as well as the Y and X chromosomes in case of male individuals. The *HaplotypeCaller* is used for all autosomes and the X chromosome in female individuals. Versions after GATK 3.3 can be used for all haploid chromosomes using the same *ploidy* stetting as before with the *UnifiedGenotyper*. The *ploidy* parameter allows to specify the *ploidy* of the organism investigated, here the default of two is used for all diploid autosomes and the setting of one is used for all haploid chromosomes like the mtDNA and X and Y in male individuals. The *HaplotypeCaller* is able to not only emit the sites that vary in an alignment compared to the reference but also emit every non varying site in the alignment. For a population wide comparison, the latter is especially useful because it allows to differentiate between not covered sites and sites that actually carry the reference allele. If only variant sites are emitted, the state of positions not shown in the call would be unknown. For all analysis only callable regions were used from the ancient samples. A callable region is here defined as any locus in the alignment that is at least covered twice, with one of the two bases covering the locus having a base quality of $\geq$15 and a mapping quality of the corresponding read $\geq$20. To identify such callable regions GATK's *CallableLoci* was used.

A SNP call with either of the above described tools will result in a so called **V**ariant **C**all **F**ormat (VCF)[23] file. While having two sections, a header section and a section that includes variants, the variants are displayed for each position line by line with the information in each line separated by tabulators. The first nine columns in each line will show: chromosome, 1 based position of current variant, reference allele, ID of the position, possible alternate alleles, the general quality of the variant, if and which filter was applied, varying information that will be explained in the header of each file and the format in which all samples will be shown. The vcf format also allows to hold multiple samples. Those samples will be displayed as shown in the format field, one

sample per column after the ninth column. In this project, filtering was only applied on parts of the format field of each sample, namely depth (DP), genotype quality (GQ) and genotype (GT). DP states the number of high quality reads that cover the given position. Meaning DP will show a 1 if one of two reads covering this position will have a mapping quality $\leq$20 or the base in question has a base quality $\leq$15. The genotype quality shows the quality of the genotype in relation to the likelihood of the second most likely genotype as PHRED scaled value. By showing the likelihood of the second most likely genotype, GQ states the error probability of the called genotype. The GT states the genotype at the position by showing the state of the two alleles like allele1/alllele2. There are two possible states for each allele, 0 means reference allele and 1 states the alternative allele. Possible genotypes are:

- 0/0 for positions that are homozygous with the reference allele.
- 0/1 for positions that are heterozygous.
- 1/1 for positions that are homozygous with the alternative allele.

While allowing to show non variant sites in a variant call the use of the -ERC BP_RESOLUTION option in the *HaplotypeCaller* will introduce "[...] a symbolic allele pair called <NON_REF> to indicate that this site is not homozygous reference, [...]" [47]. This symbolic allele pair is introduced to keep the allele specific fields in the vcf format without having an actual alternative allele and thus calculate genotype likelihoods for positions having a reference base. Those genotype likelihoods are calculated by "Estimating the confidence that no SNP exists at the site by contrasting all reads with the reference base vs all reads with any non-reference base" and "Estimate the confidence that no indel of size <X (determined by command line parameter) could exist at this site by calculating the number of reads that provide evidence against such an indel, and from this value estimate the chance that we would not have seen the allele confidently"[47] (for further details see [23, 47]).
For filtering the called variants, a self-written python script is used that could be better integrated in the following steps and had a faster performance than the GATK counter part (see Chapter 2.1.6 for further details; see supplementary file filter_snps_bgz.py in Table A.43 for the script).

### 2.1.5 Contamination assessment and/or authenticity of aDNA

To quantify possible contamination and check the authenticity of a sample, five methods were used during the course of this thesis. Two methods, MIA [33, 80] and *ContaMix* [28], allow to measure a rate of contamination on the mitochondrial genome. One, *FastQ Screen* [11] screens the data for a given set of sequences. Here a continuously used artificial sequence to assess possible cross contamination during wet lab work was given(see Chapter 5.2.1). Another method, *mapDamage* [50] visualizes typical aDNA damage patterns in nucleotide sequences. Whereas the fifth application, ANGSD [59], allows to estimate a rate of contamination on the basis of X chromosomal reads in male individuals.

#### *Mapping iterative assembler (MIA)*

MIA is a tool for creating short read assemblies. It assembles fragmented reads to a single consensus sequence assisted by a given reference. The program was designed to specially fit the needs for aDNA in the context of sequencing the first Neanderthal genome and the Denisovan genome [33, 80]. Despite assembling short reads, *ccheck* is integrated in MIA that allows to quantify possible contaminants. *Ccheck* needs a consensus sequence of the sample generated by MIA and an additional data set consisting of 311 human mitochondrial sequences, available with MIA, to estimate the contamination rate. The program looks for differences between the consensus and the putative contaminants. Those differences are then used to classify whether a read belongs to the sample or the possible contaminant. To use MIA, the sequences that should be assembled, in this

case sequences aligning to the human mtDNA, have to be present in the fastq format. Fastq files can be extracted from the previously generated SAM/BAM alignments in several ways. The two possibilities used here are either a simple bash script or the *bamToFastq* command present in the *bedtools* [96] package. MIA will then iteratively assemble the reads from the fastq file alongside a given reference. The assembler will output multiple assemblies numbered by their iteration. Each iteration uses the information of the prior assembly to refine the following until there is no increase in quality. The assembly from the highest iteration is considered the most accurate one and used for further analysis. MIA was used with default parameters and the rCRS [82] as reference. The manual page of *ccheck* explicitly warns not to use this program to assess contamination rates in anatomical modern humans. It is explained above but was not used to authenticate a sample after *ContaMix* (see next paragraph) was published. MIA is used to assemble a consensus sequence from mitochondrial genomes for usage with *ContaMix*.

#### *ContaMix*

*ContaMix* [28] estimates the proportion of contamination in sequences aligning to the human mtDNA with a likelihood-based method. Additionally, it predicts the sequencing error rate present in the data set as a uniform per base error rate. Estimates are made under the assumption that contamination is less than 50%. Since the number of distinct contaminating individuals as well as the frequency of mtDNA haplotypes present in the contaminating population is unknown, this approach uses a Markov chain Monte Carlo probabilistic model to estimate the contamination rate.
As input data, the program needs all mtDNA reads of a sample aligned to its own consensus sequence as well as a multiple sequence alignment of all possible sequences of contaminating mtDNA haplotypes and the samples consensus sequence. During this work, the consensus of a sample was built using MIA as described above. The dataset of 311 mitochondrial genomes included in MIA was used as a reference for the contaminating population. The multiple sequence alignment of the 311 human mtDNA sequences and the samples consensus sequence was carried out with *mafft* [52]. *Bwa* was used with default parameters to realign all mitochondrial reads to their own consensus sequence.
*ContaMix*'s output is a single line with the following tab-delimeted fields:
Estimated error rate, estimated maximum posterior proportion authentic, 2.5% credible quantile for proportion authentic, 97.5% credible quantile for proportion authentic, Gelman and Rubin diagnostic point estimate, Gelman and Rubin diagnostic upper confidence limit.
The proportion authentic displays the relative amount of reads that can be assumed as authentic and are mapping best to the sample in question. For relative contamination rates the proportion authentic has to be subtracted from 100%.

#### *Fastq screen*

This application is written in perl and allows the user to search a sequence data set against a panel of different databases containing nucleotide sequence data. *FastQ Screen* [11] utilizes *bowtie2* [62] for screening of raw sequence data for given sequences. It is used to search the fastq data for artificial sequences present during sample preparation. For further details see Chapter 5.2.2

#### *mapDamage*

*MapDamage* 2 [50] is a software written in Python and R that can track and quantify damage patterns in ancient DNA sequencing reads. In the context of this work it was mainly used to verify that the sequenced sample, could be seen as ancient by visualizing typical damage patterns and to supply the damage recalibration with the necessary observed transition frequencies. The typical damage pattern for aDNA shows elevated

frequencies of G → A and C → T transitions. Those frequencies are particularly elevated for C → T at the 3'-end and G → A at the 5'-end of a sequence.

#### *ANGSD*

ANGSD [59] is a software package that can perform several tasks in downstream analysis like estimating site frequency spectrum, testing for admixture and estimating contamination on chromosomes with one genecopy. It is used to estimate contamination on the basis of X chromosomal reads in male individuals. ANGSD expects only a single allele at each position and thus if a second allele is observed at a given site it implies this arose due to contamination or sequencing error [97]. In short, ANGSD calculates the rate of heterozygosity at SNP sites and compares that to the mismatch rate at adjacent supposedly monomorphic sites. In order to do this, the software needs a list of polymorphic sites with their respective allele frequency. For the X chromosome in humans, files are available in the ANGSD source package. To perform this analysis, a binary count file has to be generated from the bam file in a first step which will then be used to perform Fisher's exact test to find a p-value and jackknifing to estimate the contamination [1] in the following step.

### 2.1.6 Method testing

#### *Filtering*

To determine the influence of filtering for read depth and genotype likelihood on the variants used in downstream analysis, several differently filtered data was produced and compared to average values generated on modern data (see below in Chapter 2.1.7). The filter parameters used on the callable regions, for read depth at a given position are 0x, 1x, 5x and 10x and for genotype quality values of 0, 15 and 30. Of those nine possible filter combinations one set of filters is chosen and used to test the influence of the recalibration methods.

#### *Recalibration*

In an experiment the GATK base recalibration itself was explored by comparing the two methods described above for base quality recalibration, the default GATK recalibration and the self-recalibration. The latter was performed iteratively by recalibrating the previously recalibrated alignment with variants called on itself, to test the influence of the known variation on the data and see if this recalibration can be performed on organisms with unknown variation
Self-recalibration is done by first emitting all callable sites in the sample and selecting only homozygous sites with a genotype quality of $> 30$ (Filter1). For this variant call the GATK *HaplotypeCaller* is used as described in Chapter 2.1.4. Homozygous sites include also sites that do not differ from the references. The resulting sites are then used as known variants during the base recalibration of the alignment. A variant call as described before is again performed on this recalibrated alignment. The variant calls of the recalibrated and the uncalibrated data set are then compared using a different filter. A filter that allows both homozygous and heterozygous sites with a genotype quality of 15 (Filter2). The recalibrated and uncalibrated data sets filtered with Filter2 will then be compared with regards to the total number of variants for each of the three possible genotypes, homozygous variant calls, homozygous reference calls and heterozygous calls. If a difference of 0.01% occurs between any of the data sets the Filter1 is applied to the new variant call. The resulting variants are then used in the next iteration as known variants, to recalibrate the base qualities of the alignment resulting from the previous recalibration. This is repeated till no more than 0.01% differences between two consecutive SNP calls occur. The final alignment is then considered to be recalibrated (see supplementary file recal_on_recal.sh in Table A.43 for the bash script used to perform the self recalibration). Damage recalibration was tested alone and used either before or after the GATK default recalibration was performed (see Chapter 2.2.2).

### 2.1.7 Sample description

***Generating a modern average***

To compare the ancient data sets with modern data, several statistics were calculated over the variant calls of each of the tested methods using self-written python script (see supplementary file array_stats.py in Table A.43 for the script). To reduce RAM usage and disk space all variant calls were randomly sub sampled to 3 million positions.
Calculated statistics:

- Read depth as the average number of reads covering one position in the genome.
- Frequencies for homozygous and heterozygous sites in all variants. Those will be referred to as *hom freq* for the first and *het freq* for the latter in tables shown in the following chapters.
- The distribution of allele changes in contrast to the human reference genome. Resulting from this, the difference between C→T and T→C changes as well as between G→A and A→G changes are calculated. The two differences will be referred to as *Diff CT-TC* and *Diff GA-AG* in later tables.
- The total number of variants in the given subset of position of that sample.
- The average base quality per different base and position in each read. As well as the average standard deviation for each base over all positions and reads.

The above statistics were calculated on 3 million randomly selected positions of 7 individuals in the 1000G project phase 3[20] to give expected values for the calculated statistics that will be referred to as *modern average*. To reduce the amount of data only chromosomes 1, 5, 10, 11 and 22 were used.
1000 Genomes Individuals (average read depth): HG02583 (9.81x), NA20513 (6.66x), HG02107 (40.28x), HG00732 (30.95x), HG01873 (24.04x), HG01784 (18.35x), NA12342 (16.9x), HG02973 (10.95x), NA19311 (5.2x), NA18534 (3.16x).
To show the little variation in *het freq*, *hom freq*, *Diff CT-TC* and *Diff GA-AG* calculated from the modern data, the modern sample NA18534 was selected as an example. The average read depth is the closest of all modern samples to the ancient samples shown later but far from the modern average. With a read depth of 3.16x it is even below the average including the standard deviation. The frequencies of heterozygous and homozygous variants in NA18534 is in the range of the calculated average ± the standard deviation (see Table 2.1). It should also be pointed out that there is no standard deviation measurable in the modern average for *Diff CT-TC* and *Diff GA-AG*. Hence the value in NA18534 is the same as in modern average. Filtering for read depth had no influence on the modern average and taking GQ into account was not possible since the GQ value was not reported in the available vcf files. It also has to be pointed out that only variant sites are available in the used variant calls from the 1000 Genomes phase 3 calls [20].

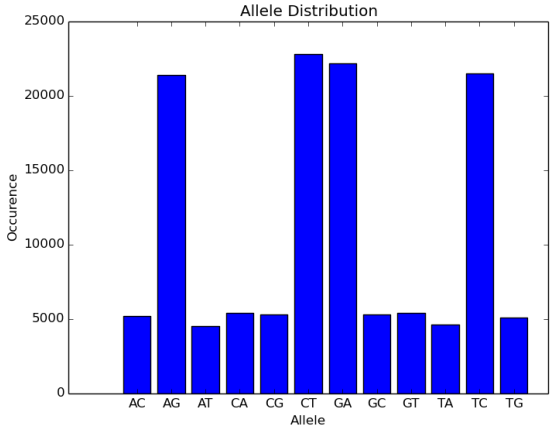



**Figure 2.6:** *Showing the allele change distribution for NA18534.*

| Sample | NA18534 | modern average | STD |
|---|---|---|---|
| Read depth | 3.16 | 17.90 | 12.08 |
| het freq | 0.58 | 0.62 | 0.04 |
| hom freq | 0.42 | 0.38 | 0.04 |
| Diff CT-TC | 0.01 | 0.01 | 0.00 |
| Diff GA-AG | 0.01 | 0.01 | 0.00 |
| Total variants | 101882 | | |

**Table 2.1:** *Summary statistics of allele change distribution NA18534, the average of the 7 modern samples (modern average) and the standard deviation(STD) for the modern average.*

***Prehistoric samples***

Both prehistoric samples used here, were sequenced as 100bp paired end runs on the Illumina Hiseq. The alignments and variant calls were produced as described above. All methods are tested on two Neolithic samples from Greece, Klei10 and Pal7. The shotgun sequencing experiments of the two samples were produced within another project [40]. The samples used are considered low coverage data sets and will be described in detail below. All statistics described here were calculated on the here generated bam files available from Hoffmanova & Kreutzer *et. al* 2016 [40] to establish a base line on which to evaluate the influence of the previously described methods.

Sample Klei10:

Sample Klei10 dates to the Neolithic in the Aegeis (cal. BCE = 4230-3995, 95.4% calibration range) and consists of two separate libraries sequenced on one lane. This data set has an average read depth of 2.01 (± 2.2) and is published in Hoffmanova & Kreutzer *et al.* 2016 [40]. Mitochondrial contamination is estimated between 0-0.50% using ContaMix. With the sex of the sample being male, contamination on the X chromosome using ANGSD could be estimated to 1.49% (SE = 0.13; p-value=2.20E-16) using Method one and to 1.54% (SE=0.21;p-value=2.83E-12) using Method two (see Chapter 2.1.5). The damage patterns of this sample show a C → T change with a frequency of 0.28 at the 5'-end that rapidly declines towards the middle of the read, having reached a frequency of 0.05 already at base 5. This pattern is almost symmetrical at the 3'-end for the G → A changes (see Table 2.2 and Figure 2.7).

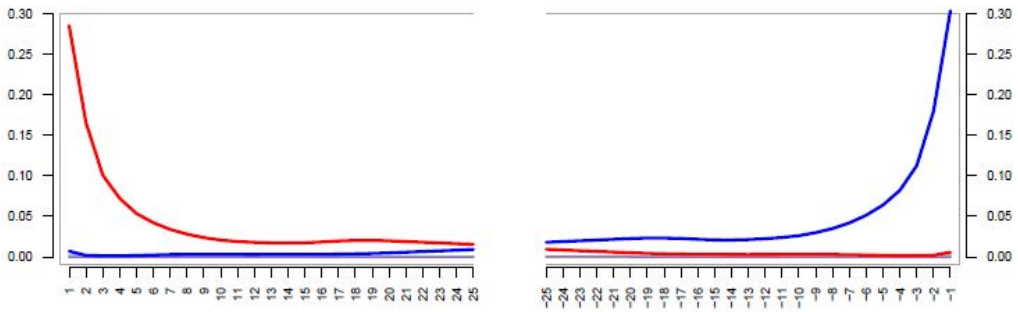



**Figure 2.7:** *Klei10;Damage pattern generated with mapDamage; C to T changes relative to the 5'-end in red on the left and of G to A changes in blue on the left relative to the 3'-end.*

**Table 2.2:** *Klei10; Frequencies of C to T changes relative to the 5'-end and of G to A changes relative to the 3'-end.*

| relative position | 5pC → T | 3pG → A |
|---|---|---|
| 1 | 0.28 | 0.30 |
| 2 | 0.17 | 0.18 |
| 3 | 0.10 | 0.11 |
| 4 | 0.07 | 0.08 |
| 5 | 0.05 | 0.06 |
| 6 | 0.04 | 0.05 |
| 7 | 0.03 | 0.04 |

Looking at the base change distribution in Figure 2.8 and Table 2.3, it can be seen that in Klei10 more C → T and G → A changes occurred than T → C and A → G. This results in a *Diff CT-TC* of 0.31 and a *Diff GA-AG* of 0.27. Both differences are at least 27x higher than in modern average (see Table 2.8). Additionally one can see that the frequency of heterozygous genotypes is raised by 0.29 to 0.89 in Klei10 compared to the *modern average*. The per base quality distribution of the unrefined alignment shows the base quality for all four bases rising from 36 in the first few bases to 39-40 and decaying towards the end of the read to 31-32 (Table 2.4

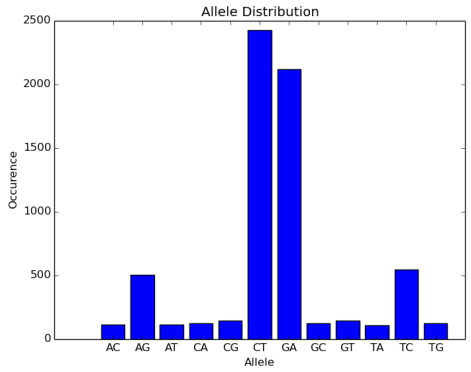
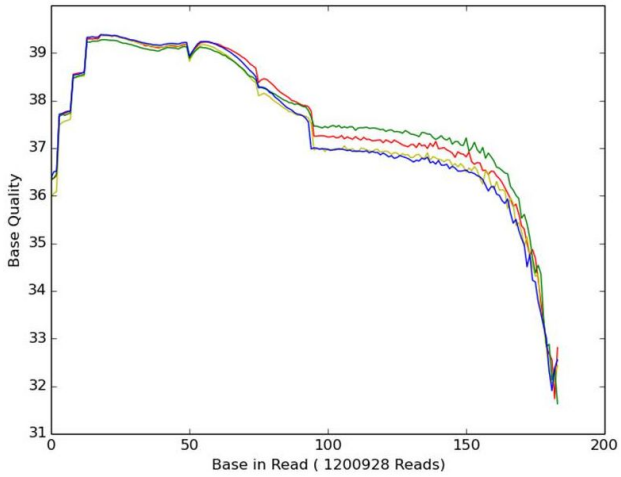



**Figure 2.8:** *Occurrences of base changes for Klei10.*

**Table 2.3:** *Klei10;Frequencies and total counts for above allele change distribution.*

| Change | AC | AG | AT | CA | CG | CT | GA | GC | GT | TA | TC | TG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Count | 164 | 700 | 159 | 185 | 187 | 3906 | 3466 | 186 | 182 | 143 | 734 | 195 |
| Frequency | 0.02 | 0.07 | 0.02 | 0.02 | 0.02 | 0.42 | 0.37 | 0.02 | 0.02 | 0.02 | 0.08 | 0.02 |

and Figure 2.9).




| Base | Average | Min | Max |
|---|---|---|---|
| T (red) | 37.63 | 31.76 | 39.37 |
| STD | 3.14 | 1.87 | 4.08 |
| A (blue) | 37.47 | 31.90 | 39.39 |
| STD | 3.21 | 2.03 | 4.13 |
| C (yellow) | 37.46 | 31.23 | 39.38 |
| STD | 3.30 | 2.15 | 4.63 |
| G (green) | 37.70 | 32.07 | 39.28 |
| STD | 3.14 | 1.94 | 3.89 |

**Figure 2.9:** *Average base qualities for Klei10 unprocessed (y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

**Table 2.4:** *Average base qualities in Klei10 unprocessed for each base over all positions in all reads; Min and Max are chosen from the averages per position and therefore the std shows variance at that position.*

Sample Pal7:

The second sample is individual Pal7, published in Hoffmanova & Kreutzer *et al.* 2016 which also dates to the Aegean Neolithic (cal. BCE = 4452-4350, 95.4% calibration range). Pal7 has an average read depth of 1.29 (±1.53). Mitochondrial contamination is estimated between 0.006-0.77% using ContaMix (see Chapter 2.1.5). Since Pal7 is a female individual, an X chromosomal contamination could not be estimated. Although the sample is only ∼200 years older than Klei10, the damage patterns with a frequency of 0,42 C → T at 5'-end and 0,41 G → A at the 3'-end are 1.5x elevated compared to Klei10 ( see Table 2.5 and Figure 2.10 ).

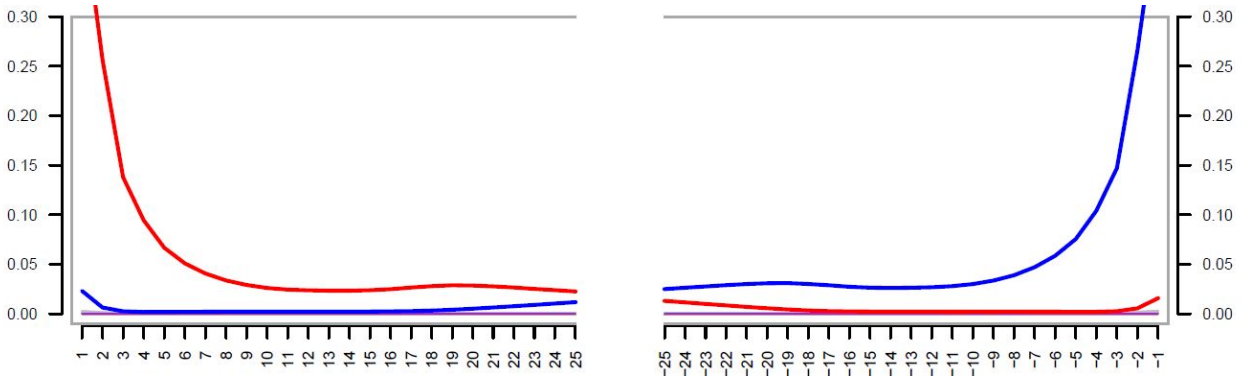



**Figure 2.10:** *Pal7; Damage pattern generated with mapDamage; C to T changes relative to the 5'-end in red on the left and of G to A changes in blue on the left relative to the 3'-end.*

**Table 2.5:** *Frequencies of C to T changes relative to the 5'-end and of G to A changes relative to the 3'-end.*

| relative position | 5pC → T | 3pG → A |
|---|---|---|
| 1 | 0.42 | 0.41 |
| 2 | 0.26 | 0.27 |
| 3 | 0.14 | 0.15 |
| 4 | 0.09 | 0.10 |
| 5 | 0.07 | 0.08 |
| 6 | 0.05 | 0.06 |
| 7 | 0.04 | 0.05 |

The *Diff CT-TC* of 0.36 and *Diff GA-AG* 0.31 are even more elevated in Pal7 compared to Klei10 and are 36x or 31x higher than in the modern average. Additionally the heterozygous frequency in Pal7 is also raised to 0.89 (see Table 2.8). The average quality scores reported by the same sequencer as in Klei10 are very similar and show almost the same pattern as in the sample before. Starting at around 37 and raising in the first few bases to 39-40 and decrease towards the end of the read to a quality of 31-32 (see Table 2.7 and Figure 2.12).

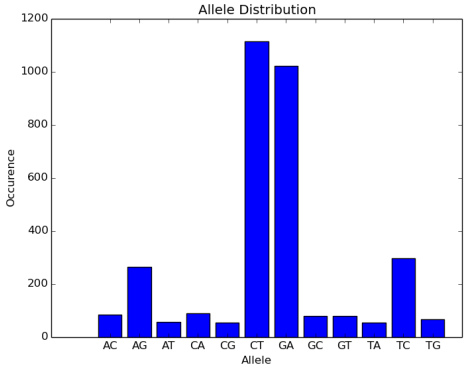



**Figure 2.11:** *Occurrences of base changes for Pal7.*

**Table 2.6:** *Frequencies and total counts for above allele change distribution.*

| Change | AC | AG | AT | CA | CG | CT | GA | GC | GT | TA | TC | TG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Count | 68 | 185 | 54 | 63 | 58 | 1677 | 1457 | 59 | 73 | 68 | 232 | 61 |
| Frequency | 0.02 | 0.05 | 0.01 | 0.02 | 0.01 | 0.41 | 0.36 | 0.01 | 0.02 | 0.02 | 0.06 | 0.02 |

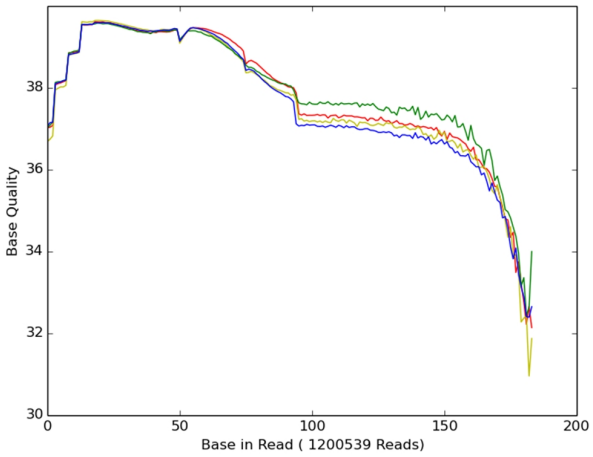



**Figure 2.12:** *Average base qualities for Pal7 unprocessed (y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

| Base | Average | Min | Max |
|---|---|---|---|
| T (red) | 37.79 | 32.14 | 39.62 |
| STD | 3.09 | 1.8 | 4.18 |
| A (blue) | 37.64 | 32.4 | 39.59 |
| STD | 3.12 | 2.11 | 4.2 |
| C (yellow) | 37.69 | 30.96 | 39.65 |
| STD | 3.18 | 1.17 | 4.83 |
| G (green) | 37.93 | 32.43 | 39.58 |
| STD | 3.01 | 0 | 4.12 |

**Table 2.7:** *Shows average base qualities in Pal7 unprocessed for each base over all positions in all reads; Min and Max are chosen from the averages per position and therefore the std shows variance at that position.*

**Table 2.8:** *Summary statistics of allele change distributions from Klei10 and Pal7.*

| Sample | Klei10 (filter 5x) | Modern average | Pal7 (filter 5x) |
|---|---|---|---|
| Read depth | 2.02 | 17.90 | 1.29 |
| het freq | 0.83 | 0.62 | 0.89 |
| hom freq | 0.17 | 0.38 | 0.11 |
| Diff CT-TC | 0.31 | 0.01 | 0.36 |
| Diff GA-AG | 0.24 | 0.01 | 0.31 |
| Total variants | 3901 | | 4055 |

## 2.2 Results

### 2.2.1 Filtering results

Before filtering, both samples show a high difference between the average genotype qualities in homozygous positions and heterozygous postions (see Tables A.30 & A.31). In Klei10, the average genotype qualities for positions that are called as homozygous reference are equal to positions called as homozygous for alternate alleles with a quality of seven. The average quality of heterozygous positions in contrast is seven times higher with 49. The same pattern can be found Pal7, with an average quality of six for any homozygous positions and a 7.67x higher quality for the heterozygous positions with 46. Any filter criteria used on the raw data decreases the difference in the average quality between heterozygous and homozygous positions and generally increases genotype qualities. Additionally all filters increase the frequency of heterozygous positions. This increase is stronger using higher genotype qualities. In combinations of filter criteria including 10x read depth, the genotype quality has little influence on the heterozygous frequency. The lowest value in heterozygous frequency after applying any filter, is reached with read depth 5x (see upper graph in Figure 2.13) and a genotype quality of <15, additionally filtering for 5x read depth and genotype quality zero has the lowest data loss. The *Diff CT-TC* and *Diff GA-AG* reach their lowest values at read depth 10x and genotype quality 30 (see lower graph in Figure 2.13). The relative amount of variants relative to all called positions is increased no matter what filtering step is used. In both samples the data loss is enormous, loosing 84.22 % of all positions in Klei10 and 94.27% in Pal7 by just applying the filter for 5x read depth (see Tables A.30 & A.31).

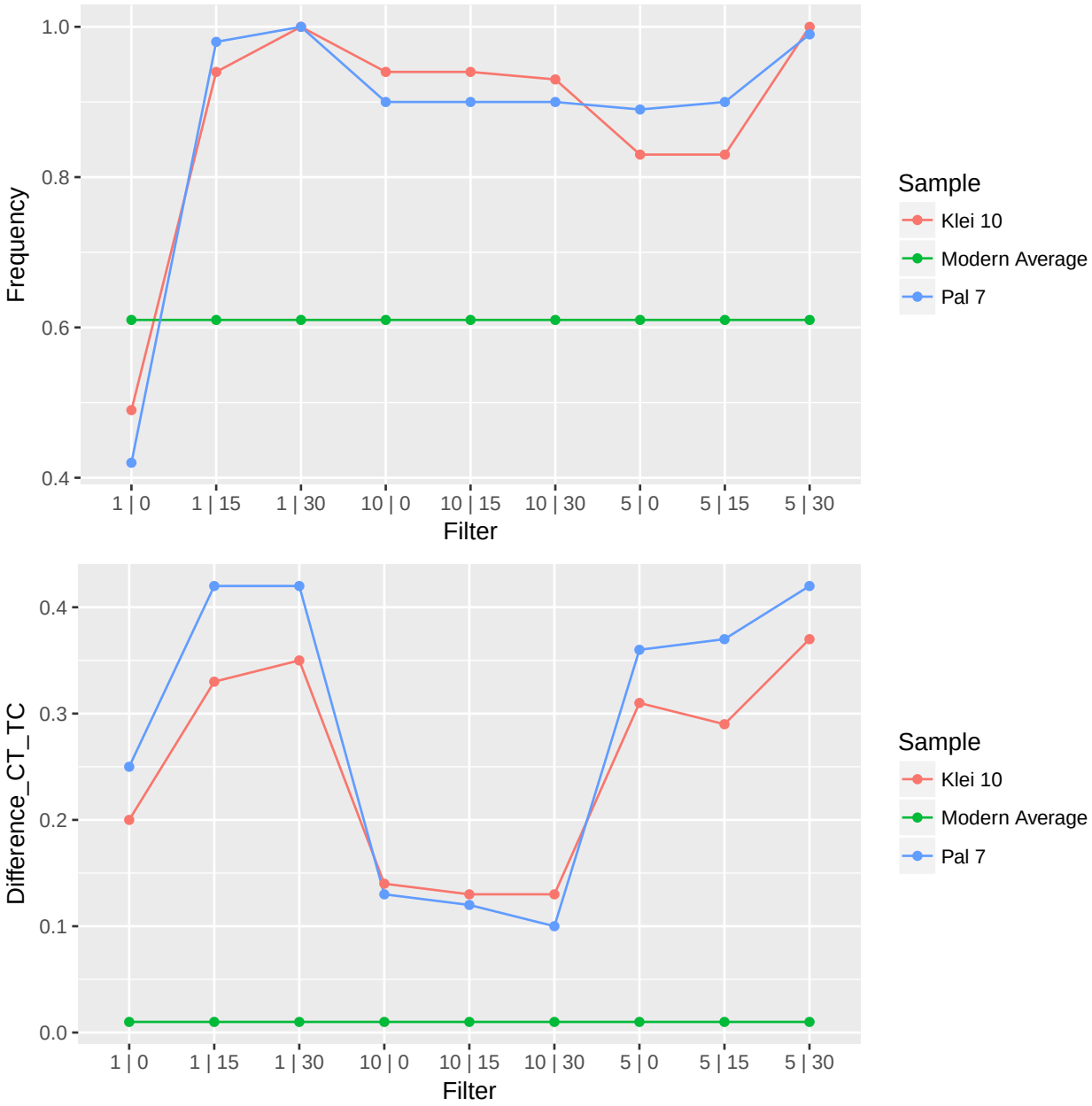



**Figure 2.13:** *Top:Heterozygous genotype frequencies Klei10 (red), Pal7(blue) and Modern Average (green) over multiple filter steps; filter is given with read depth | genotype quality.*
*Bottom: Diff CT-TC for Klei10 (red), Pal7 (blue) and Modern Average (green) over multiple filter steps; filter is given with read depth | genotype quality.*

### 2.2.2 Results of recalibration methods

***Default GATK base recalibration***

For both samples, Klei10 and Pal7, a generally lower base quality can be observed after applying the GATK recalibration.(see Figures 2.14 & A.37 and Table 2.9). Additionally the average base quality in Thymine (T) and Adenine (A) reads is overall lower than the qualities for Cytosine (C) and Guanine (G) calls. For C and G calls, the average base quality has dropped by ∼14. Whereas the average base quality for T and A dropped by ∼18 after base recalibration. For all base calls, the average standard deviation fell by ∼0.1-0.4, indicating less variation in the base qualities then prior to the GATK base recalibration.

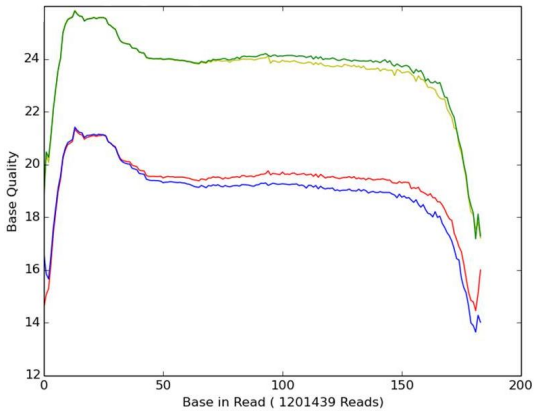



**Figure 2.14:** *Average base quality distribution for Klei10 recalibrated with GATK (y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

**Table 2.9:** *Shows average base qualities for each base over all positions in all reads for Klei10 and Pal7 recalibrated with GATK; Min and Max are chosen from the averages per position and therefore the std shows variance at that position.*

| Sample | Klei10 | | | Pal7 | | |
|---|---|---|---|---|---|---|
| Base | Average | Min | Max | Average | Min | Max |
| T (red) | 19.33 | 14.65 | 21.40 | 17.71 | 12.91 | 20.17 |
| STD | 2.82 | 0.99 | 3.50 | 2.97 | 1.75 | 3.95 |
| A (blue) | 19.17 | 14.39 | 21.44 | 17.51 | 12.26 | 20.18 |
| STD | 2.91 | 1.69 | 3.63 | 3.09 | 1.83 | 3.97 |
| C (yellow) | 23.59 | 17.45 | 25.85 | 21.72 | 15.53 | 24.27 |
| STD | 2.81 | 1.45 | 4.02 | 2.92 | 1.46 | 4.15 |
| G (green) | 23.71 | 17.60 | 25.81 | 21.83 | 14.29 | 24.25 |
| STD | 2.71 | 1.51 | 3.89 | 2.83 | 1.51 | 4.07 |

The summary statistics of the allele change distribution in Table 2.10 shows in the columns marked as "Default", that base recalibration brought all values closer to the modern average in comparison to the uncalibrated alignment (compare Table 2.8). The frequency of heterozygous genotypes as well as both differences between allele changes, *Diff CT-TC* and *Diff GA-AG*, are lowered by ∼0.1 compared to the values in the not recalibrated data set. The difference between the self recalibration step 1 (see Chapter 2.1.6) and default GATK recalibration is 0.01-0.02 for all values except read depth (compare columns "Self recal step 1" & "Default" in Table 2.10). It should also be noted that there is a reduction in actual variants found after any recalibration in both samples. For Klei10, there is a loss of ∼16.7% of variants in both recalibration methods. For Pal7 a ∼25% loss during the default recalibration and a loss of ∼32.4% after applying the first self recalibration step can be observed. Recalibration reduces the genotype quality of heterozygous variants in both samples to ∼40-50% of the average quality present after filtering. For Klei10, the heterozygous genotype quality is reduced to 37 and for Pal7 to 33 (see Tables A.30 & A.31).

**Table 2.10:** *Summary statistics of allele change distribution. Default GATK recalibration and GATK self recalibration step 1*

| Sample | Klei10 | | modern average | Pal7 | |
|---|---|---|---|---|---|
| Method | Default(filter x5) | Self recal step 1 (filter 5x) | modern average | Default(filter x5) | Self recal step 1 (filter 5x) |
| Read depth | 2.02 | 2.02 | 17.90 | 1.29 | 1.29 |
| het freq | 0.73 | 0.72 | 0.62 | 0.78 | 0.8 |
| hom freq | 0.27 | 0.28 | 0.38 | 0.22 | 0.2 |
| Diff CT-TC | 0.17 | 0.16 | 0.01 | 0.21 | 0.22 |
| Diff GA-AG | 0.16 | 0.18 | 0.01 | 0.19 | 0.22 |
| Total variants | 3252 | 3243 | | 3042 | 2742 |

***Self recalibration***

Figure 2.15 shows the base quality distribution after the first step in the described iterative self recalibration from Chapter 2.1.3. As the comparison between Figure 2.14 and Figure 2.15 shows, using the samples own variants as known variation results in a base quality distribution comparable to the default recalibration method. Both recalibration methods show the general reduction in base quality and a reduced quality for A and T compared to G and C (for average base qualities see Table A.29). Base qualities after self recalibration in Pal7 behave similar to Klei10 (compare Figure A.38 & 2.12 for the base qualities in Pal7 )

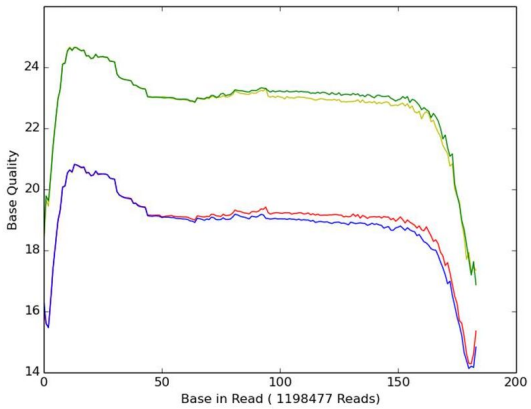



**Figure 2.15:** *Klei10 selfrecalibration step 1; Average base quality (y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

Displayed in Figure 2.16 is the average base quality after iteration 6 in the self recalibration for Klei10. The sixth iteration was the last of the recalibration steps. No differences greater than 0.1% could be measured between the number variant calls from step 5 and step 6 using any of the filters. The graph shows that the quality scores of A and T bases follow the measured damage pattern (see Figure 2.7) to some extend. As $G \rightarrow A$ mismatches rise towards the 3'-end of a read, so drops the quality for Adenosine base calls. And the other way around for Thymine base qualities. With a reduction in $C \rightarrow T$ towards the 3'-end, the base quality rises (for average base qualities see Table A.32). For Pal7 similar patterns arise after iteration 5, which was the last step in self recalibration for Pal7 (see Figure A.38).

Although the first step in self recalibration has the highest impact on the calculated statistics in terms of shifting them towards the modern average, following steps further decrease the frequency of heterozygous sites, *Diff CT-TC* and *Diff GA-AG*. With the last step of self recalibration a difference in the frequency of heterozygous genotypes of 0.07 for Klei10 and 0.12 for Pal7 towards the modern average is reached. The *Diff CT-TC* and *Diff GA-AG* are still at least 10x higher than the modern average although they have been reduced by >0.09.

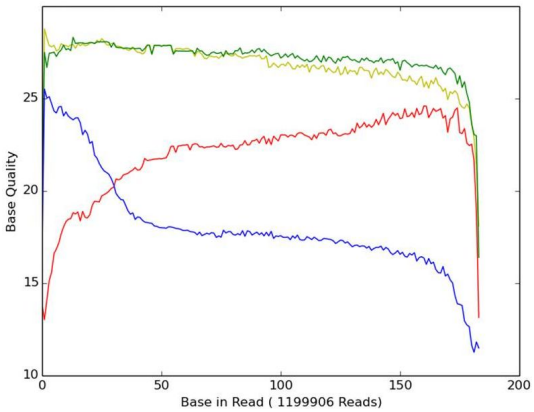



**Figure 2.16:** *Klei10 selfrecalibration step 6; Average base qualities (y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

**Table 2.11:** *Summary statistics of allele change distribution. Self recalibation step 2 and 6 for Klei 10 and step 2 and 5 for Pal7*

| Sample | Klei10 | | modern average | Pal7 | |
|---|---|---|---|---|---|
| MethodS | 2nd Step (filter 5x) | 6th Step (filter 5x) | modern average | 2nd Step (filter 5x) | 5th Step (filter 5x) |
| Read depth | 2.02 | 2.02 | 17.90 | 1.29 | 1.29 |
| het freq | 0.70 | 0.69 | 0.62 | 0.75 | 0.74 |
| hom freq | 0.30 | 0.31 | 0.38 | 0.25 | 0.26 |
| Diff CT-TC | 0.12 | 0.13 | 0.01 | 0.15 | 0.16 |
| Diff GA-AG | 0.10 | 0.11 | 0.01 | 0.15 | 0.15 |
| Total variants | 2871 | 2968 | | 2107 | 2457 |

### *Damage recalibration*

Applying only damage recalibration, the *Diff CT-TC* and *Diff GA-AG* are reduced to only $<$ 4x of the difference present in the modern average for Pal7. The values in Klei10 are even $>$ 5x lower than the modern average in both differences, showing that the both CT-TC and GA-AG are even closer to being equal than in the modern average (see Table 2.12). The frequencies for heterozygous genotypes are reduced drastically, almost equal to the frequency of homozygous positions in the modern average for Pal7 and equal for Klei10. The base quality distribution is similar to the raw data (data not shown).

**Table 2.12:** *Summary statistics of allele change distribution. After damage recalibration*

| Sample | Klei10 | modern average | Pal7 |
|---|---|---|---|
| Method | Damage recal | modern average | Damage recal |
| Read depth | 2.02 | 17.90 | 1.29 |
| het freq | 0.62 | 0.62 | 0.69 |
| hom freq | 0.38 | 0.38 | 0.31 |
| Diff CT-TC | -0.002 | 0.01 | 0.04 |
| Diff GA-AG | 0.001 | 0.01 | 0.04 |
| Total variants | 3020 | | 3042 |

***Damage recalibration followed by GATK base recalibration***

Applying the GATK default base recalibration after the damage recalibration results, as with the damage correction alone, in the frequencies of homozygous and heterozygous genotypes being very close to the genotype frequencies in the modern average. The values for *Diff CT-TC* and *Diff GA-AG* on the other site are negative, showing that $T \rightarrow C$ and $A \rightarrow G$ mutations are higher in frequency than $C \rightarrow T$ or $G \rightarrow A$ (see Table 2.13). This combination of GATK and damage recalibration completely removed $C \rightarrow T$ and $G \rightarrow A$ changes from the data set. As Figure 2.17 displays, the allele changes that are associated with post *mortem* damage have completely vanished from the data sets (see Figure A.40 and Tables A.33 & A.34 for results of Pal7 and mutation frequencies). The base quality distribution, showing a similar pattern to the raw data with the exception of a drop in all base qualities between positions ∼25-50 and T having the highest base qualities in general. Compared to the default GATK recalibration alone, the average base qualities are higher and a higher standard deviation in A and T base qualities can be seen (see Figures A.41 & A.42 and Table A.35 for base quality distributions).

**Table 2.13:** *Summary statistics of allele change distribution. After damage recalibration followed by GATK default recalibration*

| Sample | Klei10 | modern average | Pal7 |
|---|---|---|---|
| Method | Damage + GATK | modern average | Damage + GATK |
| Read depth | 2.02 | 17.90 | 1.29 |
| het freq | 0.62 | 0.62 | 0.65 |
| hom freq | 0.38 | 0.38 | 0.35 |
| Diff CT-TC | -0.25 | 0.01 | -0.24 |
| Diff GA-AG | -0.25 | 0.01 | -0.24 |
| Total variants | 2004 | | 1863 |

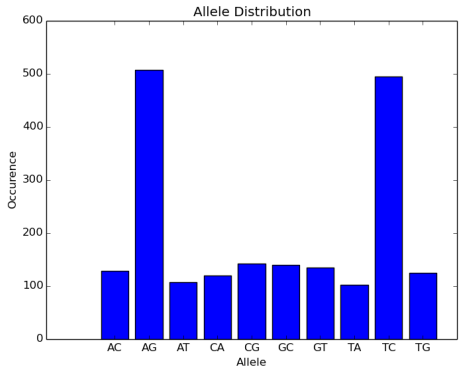



**Figure 2.17:** *Occurrences of base changes for Klei10 after damage recalibration followed by GATK recalibration.*

***GATK recalibration followed by damage recalibration***

Starting with the GATK recalibration and applying the damage recalibration in a second step does not remove the two types of transitions completely from the data set and brings *Diff CT-TC* and *Diff GA-AG* almost to zero. As the negative values for both allele change differences in Table 2.14 show, $C \rightarrow T$ as well as $G \rightarrow A$ mutations occur less than the corresponding opposite transitions (see Figure 2.18 and A.43 for Pal7). The average base qualities look similar to the ones seen after only applying GATK recalibration (see Figures A.45 & A.44). The frequency of heterozygous variants decrease in Pal7 to 0.68 and in Klei10 to 0.58.

**Table 2.14:** *Summary statistics of allele change distribution. After GATK default recalibration followed by damage recalibration*

| Sample | Klei10 | modern average | Pal7 |
|---|---|---|---|
| Method | GATK + damage | modern average | GATK + damage |
| Read depth | 2.02 | 17.90 | 1.29 |
| het freq | 0.58 | 0.62 | 0.68 |
| hom freq | 0.41 | 0.38 | 0.32 |
| Diff CT-TC | -0.01 | 0.01 | -0.02 |
| Diff GA-AG | -0.04 | 0.01 | -0.03 |
| Total variants | 2679 | | 1856 |

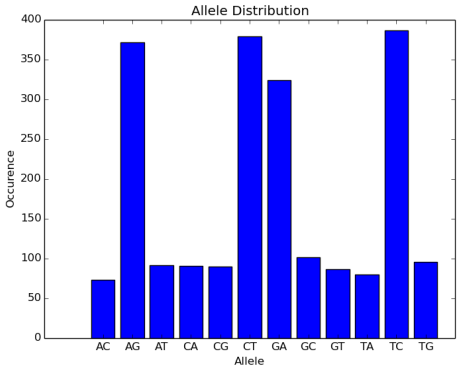



**Figure 2.18:** *Occurrences of base changes for Klei10 after GATK followed by damage recalibration.*

## 2.3 Discussion

### 2.3.1 Modern data set

Several factors have to be taken into consideration while using this modern average as a proxy in the calculated statistics for the ancient samples. First, it should be noted that an ancient population or individual has had a different population history than the modern ones and therefore a different frequency of heterozygous variants could be possible in a prehistoric individual. To solve this, coalescent theory could be used to simulate statistics for a population at the time of interest and use those as a proxy.
Another factor is the difference in calling methods used. Although the same software, the GATK *HaplotypeCaller* [76] was used, a cohort calling was performed in the 1000 Genomes project [20] instead of the single individual call performed on the ancient samples. This so called cohort calling utilizes the information present in a population to increase the certainty of a call by using a specific model [46]. This can influence the frequency of heterozygous variants in a sample after filtering and makes the genotype frequencies dependent on the cohort it is called in. While this publicly unavailable model [46] influences the genotype quality, the impact of the cohort on the general quality of a variant is even stronger, due to a genotype refinement process [45]. The influence of the samples, present in a cohort call, on the general quality of a variant, facilitated the choice of the genotype quality over the general quality of variant as a filter criteria. Although a genotype quality was not available for the 1000 Genomes data. This cohort calling is often not practical for ancient samples, since such cohort call of multiple samples only are suitable if all individuals can be seen as belonging to the same population.
While the average read depth varies in between individuals chosen for the modern average, it seems to have not as much influence on calculated statistics as in the ancient samples. This is supported, by the facts, that a read depth filter on modern average had no influence on the statistics and that there is a read depth difference of ∼12x between the worst and the best modern sample chosen here but they only vary by 0.08 between each other (see Table 2.1) in the heterozygous frequency. This difference in frequencies can be reached by applying a filter of 5x read depth to one of the ancient samples. While damage and low read depth could likely explain this difference, it is unclear how much the individuals with high read depth influence individual with low read depth in a cohort calling.
Another difference in the modern data set is that only sites that are called as a variant in at least one sample, are present in the available vcf files from 1000 Genomes[20]. Thus the calculation of the frequency of heterozygous sites was done only on the variable sites in the ancient samples.

### 2.3.2 Filtering and recalibration

To understand the evolutionary processes shaping genetic variation, quantifying and comparing genetic diversity is essential. While obtaining parameters like heterozygosity from high quality next generation sequencing data is done regularly, inferring such parameters reliably from low coverage data, like the two samples presented here is challenging. For one, this is due to a high probability of observing only one of the possible alleles at a given locus with a low read depth, which can only be solved by obtaining more data. Other reasons why this task will prove challenging, are sequencing errors that can occur up to an order of magnitude higher [60] than heterozygous positions and especially with ancient DNA, post *mortem* damage. The main objective of this pipeline is to account for the latter two, with GATK base recalibration and damage recalibration.
The filtering steps should to some extend reduce the effects of all three previously mentioned effects that can introduce biases into the measured genetic variation. Simultaneously it will reduce the amount of informative positions than can be included in the analysis, with a data loss of up to 94.27% by applying 5x read depth as the only filter criterion. Any filter criterion used on the raw data will increase the relative amount of heterozygous positions in the data. In addition, the used filters for genotype quality >15 in positions with

read depth <10, all show an increase in the frequency of heterozygous variants compared to the same read depth (see Figure 2.13 and Tables A.30 & A.31). In combination with the average qualities for different genotypes, this shows that any filter applied will favor heterozygous genotypes. This has several reasons including sequencing error, post *mortem* damage and that the *HaploytpeCaller* calls heterozygous genotypes with a higher likelihood than homozygous positions of the same read depth, which can be seen in the low genotype quality in homozygous variants , with an average value of 7 in the unfiltered Klei10, in comparison to heterozygous variants that have an average quality of 49. This does not mean that the heterozygous positions are wrong, just that they are more likely to be called before a certain read depth, because of the high probability of observing only one of the possible alleles with few reads covering a position. It has to be pointed out that because of the lower qualities in the positions called as homozygous references, every filtering step involving genotype qualities increases not only the frequency of heterozygous variants compared to all variants but also increases the relative amount of called variants in the sample itself. The results of filtering for genotype quality with a read depth >10 suggest that this bias towards heterozygous variants can only be overcome by higher read depth, since the genotype qualities in homozygote variants and heterozygote become closer, with values of 64 for the first and 66 for the second, in Pal7 with applying just a 10x read depth filter. At a read depth of 10 filtering for genotype quality does not make any difference in the genotype frequency of the variants(see Figure 2.13 and Tables A.30 & A.31).
The observed values for the differences between C to T and T to C or G to A and A to G transitions get constantly lower, the stricter the filtering criteria are chosen, from 0.42 *Diff CT-TC* in Pal7 with read depth 1x and genotype quality 15 to 0.10 using read depth 10x and genotype qualiyt of 30. This decrease can likely be attributed to the post *mortem* damage. Since PMD introduces an actual different base in the molecule, the sequencer can read a falsely integrated base with a high quality. In combination with low read depth and low genotype qualities, C to T and G to A transitions are observed more often in SNP calls of ancient DNA data. Since a PMD associated transition happens most of the time at the beginning and end of a read, and different molecules break somewhat randomly, base changes due to PMD do not accumulate at certain positions in a genome. Thus its influence will be reduced by higher read depth. The 5x read depth as only filter criterion for all following recalibration steps is chosen due to the lowest values in heterozygous frequency in combination with the lowest amount of data loss in all filtered data sets.

After filtering, the base recalibration with GATK and the damage recalibration are applied to increase the reliability of the a variant call. The average base qualities in Figures 2.14 & A.37 and Table 2.9 show that the GATK default recalibration detects some pattern in the ancient DNA that can be related to PMD and as a result lowers the base qualities in A and T throughout the read. After the last self recalibration step with GATK, the base qualities even resembles the PMD patterns. After the last recalibration step, an increase in the base qualities for A towards the 5'-end and an increase for T towards the 3'-end compared to the first recalibration step could be measured in both samples (see Figures 2.16 & A.39). GATK needs several steps of recalibrating the base qualities to integrate the information of the PMD observed in the base changes into the base quality. While self-recalibration is ideal for non model organisms or organisms with unknown variation and seems to achieve the better results, the computational time needed is enormous. Using a 12 core CPU with 32 GB RAM, a single iteration with recalibration, calling and filtering took ∼1 day with Klei10. The self recalibration with 6 iterations took ∼6 days. For further testing of the recalibration methods, the default recalibration was therefor chosen, although the final step in self recalibration shows ∼0.4 reduced heterozygous frequency and a reduction of ∼0.05 in both *Diff CT-TC* and *Diff GA-AG* (compare Tables 2.10 & 2.11).
Applying the damage recalibration alone on the filtered data brings Klei10, the sample with the higher read depth, to a heterozygous frequency of the modern standard and reduces both *Diff CT-TC* and *Diff GA-AG* even beyond the average for the modern samples (see Table 2.12). Pal7 also shows a drastic reduction in all three statistics. In comparison with the results form the default and self recalibration, the damage recalibration

has more impact on the calculated statistics and brings the values closer to the modern average. This points towards PMD having a higher influence on the observed positions than sequencing error at positions with a read depth >5 in the samples used here. Although the summary statistics for Klei10 might indicate that the damage recalibration alone is enough to remove a bias introduced due to PMD , in Klei10 with *Diff CT-TC* being -0.002 and the heterozygous frequency 0.62, the same as the *modern average*, it still has to be assumed that the the sequencing error is present after damage recalibration alone. Especially since the GATK default recalibration is often used in modern data sets, like the 1000 Genomes project [20] and should therefore be applied just for the sake of comparability.
Integrating the PMD patterns into base quality with the damage recalibration prior to the GATK default recalibration results in an over correction of the PMD patterns by completely removing any mutation that is related to PMD from the data. It seems that GATK turned the effect of PMD upside down. The damage recalibration only reduces the base quality at positions that could possibly be influenced by PMD, C to T or G to A mutations, according to the observed damage pattern. Therefore the process generates bases with lower qualities for T and A, at mismatch positions, closer to the appropriate end of the read. Looking into a bam file that was first damage recalibrated and than GATK recalibrated at known transitions associated with PMD, one can see that GATK reduces the base quality of every mismatch at such position to 1. Apparently the introduced lower base quality associated with C to T and G to A mismatches in combination with base position in the read generates a pattern that makes GATK account for all PMD associated transitions in a read as sequencing errors. Thus excluding such bases from the SNP call by setting their base quality to 1.
Using damage recalibration after the default GATK recalibration does not have this drastic impact on the data. The *Diff CT-TC* and *Diff GA-AG* do shift to negative values indicating that the non PMD mutations are favored. This also can be interpreted as an over correction, since the modern data has slightly more C to T and A to G changes. In general, the combination of the GATK default recalibration followed by damage recalibration brings the summary statistics closest to the modern average for both samples. Although the frequency of Klei10 heterozygous variants drops below the modern average it is still in the range of the standard deviation of 0.04 in the modern average.
All methods tested here have a different impact on each of the samples. This might be due to the different quality and preservation of the two samples tested but could also point towards the impossibility of generating reliable data from extremely low coverage genomes.

### 2.3.3 General pipeline

Although this pipeline has been used in three publications [12, 15, 40] several programs used here can and should be replaced. In general this pipeline was based on a collection of tools and parameters by Martin Kirchner 2012 [56], who developed this in the context of the Neanderthal project [33, 34]. Several ancient genomes have been published since then [15, 27, 29, 49, 64, 92] and different and improved pipelines were used.
Currently there are at least two other analysis pipelines specially suited for ancient DNA and NGS sequencing published, PALEOMIX [103] and EAGER [87]. PALEOMIX being the older one includes additional phylogenetic analysis and microbial taxonomical profiling compared to the one described here. Although the steps are very similar in the processing of the raw data, different software is used in several steps. The first difference is the use of the program *AdapterRemoval* [69] to trim, filter and merge reads prior to the alignment. Despite performing all tasks in one step, the program does not remove low quality reads, it allows to cut low quality ends instead and could thus increase the amount of reads present in the final alignment. The next difference is that *mapDamage* 2.0 [50] is used to recalibrate/rescale the quality scores according to the damage patterns generated. The rescaling of mapDamage works similar as the damage recalibration used her but does not include sequencing errors which is done in the base recalibration with GATK. Finally PALEOMIX uses *sam-*

*tools* for genotyping with the argument that GATK needs high quality data sets of known variation which is impossible to obtain for a lot of organisms. While PALEOMIX can be installed as a whole and is used via terminal on a linux environment, single parts could be separately integrated into, different pipelines, since most of the steps make use of different free software packages.
The second pipeline, EAGER is the more sophisticate one, allowing for several options at certain steps and having an integrated Graphical User Interface. There are four new tools added that either can be used optional or replace certain programs used in similar steps here. The program *Clip&Merge* combines all pre alignment read processing steps in one program and the authors claim that it performs better in terms of computational time and resulting mapped reads compared to several other tools including *AdapterRemoval* that is used in PALEOMIX. The program called *DeDup* is used to remove duplicates and would replace *markDuplicates* from *Picartools* [2] which seems to incorrectly identify duplicates in single-end reads or merged paired reads. It accepts reads that have the same start point as an other read, but are shorter, as a duplicate of the larger read. This way of duplicate removal will lead to a defined maximum number of sequences in a genome which is equal to two times the length of a genome or contig. This again is equal to the amount of unique strand breaks a molecule can have considering reverse and forward strand. In the human mtDNA genome for example, this will result in a maximum of 16569 unique reads from each direction. Additionaly EAGER contains a special mapping software for circular genomes called *CircularMapper* and a tool called *VCF2Genome* that can generate a consensus sequence from a vcf file. At the state of this thesis it is recommended to exchange the software used in the steps prior to the alignment with *Clip&Merge* and *MarkDuplikates* with *DeDup*.
Both of the pipelines described before, use advantageous programs in several steps but have no alternative for the recalibration processes of sequencing errors and post *mortem* damage included. PALEOMIX uses *mapDamge* 2.0 for damage which uses the same model as the method used in this pipeline. It also has to be pointed out that there are parameters suggested for the mapping of ancient DNA to a reference genome that increase the gain of mapped reads, without reducing the quality of the alignment [104]. In early tests those parameters increased computational time enormously with only gaining $\sim$1% of mapped sequences. Recent communication with colleagues seems to point towards different reasons for both effects with one being related to how *markDuplicates* identifies duplicates. With increasing computational power and continous improvement of the algorithms and tools used, the suggested parameters from [104] should be applied instead of the default parameters.

## 2.4 Conclusion

In terms of the methods used, the question remains, whether ancient samples can be comparable to a modern data set at all. With the pipeline shown here, an ancient sample sequenced with the Illumina Hiseq sequencer to low read depth can reach values close to the modern average for heterozygous frequencies and differences between C→T and T→C or G→A and A→G transitions. The combination of a base quality recalibration followed by a calibration for post *mortem* damage is the most promising combination of refinements to reach summary statistics similar to the modern standard while filtering for at least a read depth of 5x. Although the options for calling only a single allele instead of full genotype will in many cases not suffice to estimate parameters useful to population genetics, like heterozygosity [15], it remains an option for comparing positions with read depth between 2x and 5x. The single allele calls either as a random pick or picking the most likely allele, will at least allow for comparison of populations using f-statistics or other methods that do not need genotypic data [36, 75].

In general there are five possible influences that can introduce differences between individuals while comparing ancient samples with modern samples, like the ones from the 1000 Genomes project[20]. Sequencing error and PMD which can be integrated in the pipeline, read depth which can be overcome by additional sequencing runs, different SNP call methods and an actual difference in population history.

On the basis of the obtained results a pipeline fully integrating sequencing error and post *mortem* damage was developed, called ATLAS [60, 71]. The major advantages of this pipeline are the integration of the post *mortem* damage in the base quality recalibration step and during the actual variant call and allowing a reference free recalibration step as long as there is a haploid region of a certain length in the genome, like the X chromosome in mamalian male individuals. ATLAS will most likely replace the alignment refinement steps as well as the variant calling in this pipeline.

# 3 Nuclear capture enrichment approach

## 3.1 The motivation for the nuclear capture

An alternative to the the random nature of a shotgun sequencing experiment is a capture approach. This enrichment of selected loci circumvents the problem of too little coverage in the regions of interest, allows for a set of comparable loci over all processed samples, and is cost effective due to the possibility of parallel sequencing of numerous samples. While the human mtDNA capture is well established [12], its informative value for population genetics is limited. This limited power is mainly due to the fact that mitochondrial DNA is haploid and is only inherited maternally. Therefore it does not recombine, represents one single maker and can only account for the female population history. While limiting the explanatory power of population genetic inference, those factors simultaneously allow for "easier/ more precise" interpretation of such inference. To improve the informative value of a capture approach and allow for the reconstruction of population history using population genetic methods, a nuclear capture was designed. This capture approach is a combination of multiple genomic regions that are either of interest for reconstructing certain phenotypes or fulfill, to some extent, the criteria of neutrality and free recombination. The regions that contain phenotypic informative markers were developed in a different project and described in detail elsewhere (currently unpublished). Those will be referred to as *phenotypic SNPs*. Next to the advantages for population genetics the capture should above all increase the quality of the data, by increasing the coverage of the captured regions and ensure that there will be comparable loci between samples. The capture should add to the improvement in an overall quality of the alignments and a more secure variant call established within the chapter **2 Pipeline**.
The non phenotypic regions were selected so that they can be considered as being inherited independently, non coding and thus selectively neutral. Those will be referred to as *neutral regions*. For those neutral regions two different sets exist, one being more conservative the *conservative neutral regions* and the other being more relaxed the *relaxed neutral regions*. The reasoning behind the selection of neutral regions was influenced by coalescent theory and its possible application on nuclear data. Underlying the coalescent theory as a basic concept is the idea, that in the absence of selection sampled lineages/individuals in a population can be seen as randomly picking their parents going back in time. If two lineages/individuals pick the same parent, they coalesce into a single lineage. Ultimately all lineages will coalesce into one single Most Recent Common Ancestor. In that way, a genealogy is built backwards in time on which selectively neutral mutations can be randomly superimposed [99]. Under the absence of selection, any deviation from genetic drift can be explained by possible population events (e.g. size increase or decrease, gene flow with other population) or population structure and thus allow reconstructing of population history.

The next chapters will describe the selection of the neutral regions and differences between the relaxed and conservative neutral regions, the influence of the capture on the summary statistics described in chapter 2.1.7 and on a Principal Component Analysis (PCA). Although an inference based on PCA is not driven by coalescent theory and is not considered as a true population genetics method it will be used to show possibilities and limitations of the capture approach.
A detailed Sample description can be found in the following chapter **4 Case study Welzin**. The samples are used here as a case study and proof of concept for the developed capture. All samples are processed using the Pipeline described in chapter **2 Pipeline** on page 7. The captured regions are extracted as an additional step after the refinement. To allow for contamination assessment each samples sex was determined by comparing the read depth and coverage of all regions included in the capture, falling on the Y chromosome, to all other autosomal regions per eye. If more than 90% of the Y chromosomal regions were not covered the sample is assumed to have two X chromosomes.

## 3.2 Methods

### 3.2.1 Selection of conservative neutral regions

The prior description of the coalescent process explains why the neutral regions have to fulfill the criterion of selective neutrality to be used in coalescent simulations. To identify possible selectively neutral regions with a distance of at least 0.5cM to genes, the background selection coefficient (BSC) was used. BSC states the reduction in diversity as a result of linkage to sites under negative selection and explains the expected fraction of neutral diversity that is present. Values range between 0 and 1.0. 0 indicates the presence of background selection and 1 its absence [77].
The criterion of independent inheritance was applied by choosing only those loci that have at least 0.5cM distance between each other, ensuring that the regions are in linkage equilibrium. This principle was used to ensure that theoretically each of the chosen regions can be used as a single linage/marker using a coalescence approach.
In a first step, the Neutral Region Explorer (NRE) [7, 53] was used to find all regions that do not overlap any gene. NRE is an online tool to find putatively neutral sequences in the human genome. Since NRE was not working as intended, several parameters were disabled and in a second step manually filtered using bash. The disabled Parameters are: length of a region, the regions minimum distance to a gene and its recombination rate. Additionally all repeats were subtracted from the found regions. For this, bedtools [96] was used in combination with the databases from *repeatmasker* [108] and simple repeats obtained with UCSC genome table browser [51]. The regions obtained for each chromosome were then manually filtered for a minimal length of 500 bp, a background selection coefficient of at least 0.95 and a distance of at least 0.5 cM to a neighboring gene. The remaining regions were sorted and the genetic distance between each neighboring region on the same chromosome was calculated using a genetic map. Regions were only kept, if they could reach a distance of at least 0.5 cM to each of their neighboring regions. For the genetic distance calculated, the genetic map from HapMap [26] was used. The remaining regions were checked against all known genes and the EST database obtained from USC Genome Browser. Two regions overlapped with a known gene and were removed. 63 regions overlapped with ESTs and are currently kept.

### 3.2.2 Workflow relaxed neutral regions

The work for those regions was done by Krishna Veeramah slightly based on the work flow described above, while simultaneously using regions that were already identified as not being under selection. A set of 37,574 1kb regions that were previously identified by Gronau *et al.* 2011 [35] were lifted over from hg18 to hg19 and used as an inclusion mask within NRE. Only regions identified by NRE and completely overlapping the regions from Gronau *et al.* 2011 were selected and subjected to additional filtering:

- BSC >0.85.
- Recombination rate of >0.01 and <10cM/MB.
- A distance of >0.01cM to neighboring genes.
- No overlap with the conservative neutral regions, extreme phastCons [105] and simple repeats.
- Distance of 50kb to each other and to the conservative neutral regions.

#### *Testing probe specificity for relaxed neutral regions*

To make sure the selection of loci being partially present in the *repeatmasker* database will not influence unambiguous sequence alignments, the relaxed regions were subjected to additional procedures.

Blasting of the regions (Analysis done by Krishna Veeramah):
For all of the regions, 100bp sequences with 20bp sliding window were generated and blasted [17] against the human genome (hg19). Alignments scores for the first two best matches were recorded. If multiple matches were found the 100bp long sequences are extracted with additional 1-19bp upstream and downstream of the matching sequence. Those were blasted again and the two best alignment scores were kept. This procedure was repeated for 50bp fragments with a 10bp sliding window.
Mapping of the regions:
For each of the 1kb loci, fastq sequences were generated by randomly picking 40-180bp from anywhere in the region and assigning a random base quality between 40 and 60. Those reads were aligned to hg19 using *bwa* [65] default parameters and filtered for mapping quality of 25.

### 3.2.3 Bait design

The actual bait design for all selected regions was performed by the company MYcroarray©. For the bait design, all nucleotide sequences of the regions were consolidated into contiguous sequences in case fragments overlap. Each sequence was cut into 80mers that overlap by 60bp to generate the baits in a 5x tiling. To ensure unique mapping, the resulting bait candidates were then blasted against the human genome while masking the input sequence. This masking should prevent blast hits on the original selected capture regions and allow for finding ambiguously mapping baits. For each obtained blast hit, a melting temperature (Tm) was predicted under the hybridization conditions from the lab protocol Mybaits user manual v1 [81]. For each bait, the number of blast hits were counted according to six bins based on their Tm: 40-60°C, 60-62.5°C, 62.5-65°C, 65-67.5°C, 67.5-70°C and above 70°C. MYcroarray© applied three different selection filters to these baits. In this case the most stringent filter was chosen and bait candidates were accepted if they satisfied one of the following conditions:

1. No blast hit with a Tm above 60 °C.
2. $\leq$2 blast hits at 62.5-65 °C, or 10 hits in the same 80bp interval and at least one neighbor candidate being rejected.
3. $\leq$2 blast hits at 65-67.5°C, and 10 hits at 62.5-65 °C, and two neighbor candidates on at least one side being rejected.
4. $\leq$1 at or above 70°C, and $\leq$ 1 hit at 65-67.5°C, and $\leq$2 hits at 62.5-65°C, and two neighbor candidates on at least one side being rejected.

Phenotypic SNPs were designed with a 5x tiling in a region 60bp in both directions of the selected marker. For selected markers that could not be covered with 5 baits after applying the above filter, the region surrounding the SNP was widened to 300bp instead of 120bp, and more relaxed filter criteria were used to reach a 5x tiling: In addition to the prior mentioned criteria, candidates are accepted if they do not have more than 10 hits at 62.5 - 65°C and more than 2 hits above 65°C and are not surrounded by two consecutive selected baits on each side.

### 3.2.4 PCA

To see if it is possible to use the sequencing data from captured samples for Principal Component Analysis (PCA) and how the limited data set will influence this analysis, several PCAs were generated using LASER [113]. The influence of the alignment refinement described in Chapter 2.4, as well as the influence of reads aligning outside of the capture region was tested with four PCAs. One for each of the refined and unrefined

alignments of both, the reads aligning to the full genome as well as only to the capture regions. To generate the PCA plots, two separate processes need to be performed.
In a first step, a reference space is generated, by performing a standard PCA on the genotype data of modern individuals, with imputation of missing entries thru averaging over all individual genotypes at that position. As a modern reference data set, the combined data from Hellenthal *et al.* [37] and Busby *et al.* [16] based on Illumina's genotyping platforms (Illumina 550, 610, 660W arrays) was used as processed in [40] with a total of 1051 individuals and 510,811 autosomal polymorphic loci.
In the following step, the captured samples are mapped on the reference space. By using *pileup* files generated with *samtools* [66], LASER allows to directly use the read information present in the alignment without the need of a prior genotype call. These pileup files show detailed information for each of the positions in the alignment. Pileup files are generated with *samtools mpileup* using the filter criteria of minimum mapping quality 30 and minimum base quality of 20 suggested [111]. After a second required re-formatting, the PCA can be generated. This is done by first simulating data for all reference individuals similar to the read depth and coverage of the given ancient individuals, which will be used together with the ancient samples to produce a first PCA. This PCA will then be projected into the before generated reference space by Procrustes Transformation [112] to generate the final result seen in the plots. Additionally a similarity score will be reported that allows to qualify the samples placement in the reference PCA after Procrustes transformation [114]. In the PCA the two principal components that account for the highest variance are plotted against each other. In general the vectors composed of the first to two principal components of SNP data for each sample should correlate with the broad geographic distribution of the samples. For clarity reasons the majority of populations is removed from the PCA plots. This includes all non European populations as well as some European populations. This reduction of displayed samples, zooms into the PCA and allows a more detailed view of the clustering of European samples. Although not shown in the plots, the whole reference data set including all individuals is used for calculation of the principal components.

## 3.3 Results

### 3.3.1 Capture development

Blasting the 100bp fragments of the relaxed neutral regions against the human genome resulted in 0.75% of the regions containing at least one fragment with >90% identity to another position in the human genome than the originating fragment. This in total sums up to 6522bp or 0.14% of all bases covered in the relaxed neutral regions, spread on multiple 100bp fragments, while allowing single positions being covered by more than one fragment. Reducing the fragment size to 50bp increases the fraction of ambiguous mapping fragments to 5.36% including 1.22% of all bases covered in those regions (compare Table 3.15).

**Table 3.15:** *Summary of BLAST results for the 100 & 50bp fragments. Total loci = Showing the number of loci that include at least one fragment mapping with the minimum identity given in columns ; Total bp = Showing the actual number of base pairs mapping with the minimum identity given in columns*

| Indentity | 100bp Fragments | | | 50bp Fragments | | |
|---|---|---|---|---|---|---|
| | >70% | >80% | >90% | >70% | >80% | >90% |
| Total loci | 194 | 100 | 35 | 895 | 581 | 251 |
| % of all | 4,14% | 2,13% | 0,75% | 19,10% | 12,40% | 5,36% |
| Total bp | 83.809 | 40.854 | 6.522 | 186.036 | 122.883 | 57.138 |
| % of all | 1,79% | 0,87% | 0,14% | 3,97% | 2,62% | 1,22% |

After bait design and the above described quality control done by MYcroarray© the relaxed neutral regions have 133 regions not being fully covered by baits. Leaving a total of 12928bp or 0.28% of the former selected regions uncovered. The conservative neutral regions have no base not covered by 5 baits.

**Table 3.16:** *Summary of BLAST results for the 50bp fragments; Showing the Numbers of prior selected regions (# of selected), selected base pairs (# select bp), the number of baits(# of baits), the number of base pairs covered with the baits (# bp in baits), the number of regions that were split due to fragments having multiple hits (# split regions), the base pairs covered by the split regions (# bp in split) and the number of base pairs not covered from the previously selected (# bp not covered).*

| Category | # of selected | # selected bp | # of baits | # bp in baits | # split regions | # bp in split | # bp not covered |
|---|---|---|---|---|---|---|---|
| Relaxed neutral | 4653 | 4653000 | 207796 | 4644592 | 133 | 120072 | 12928 |
| Conservative Neutral | 429 | 214500 | 9006 | 214929 | 0 | 0 | 0 |

### 3.3.2 Captured samples

In the 21 samples analyzed here, the amount of reads mapping to the human genome including the captured regions is on average 76.93% (±9.22) of the total reads in all samples. Only ∼10% less map to the capture alone, with 66.53% (±9.37) on average, compared to the total reads. This is reduced after removing PCR duplicates from the alignments, to 8.71%(±3.5) of the total reads mapping to whole human genome and to 1.77% (±1.54) mapping to the capture.

Overall the 21 samples reach an average read depth of 18.13x (± 13.69) in the captured regions, with no significant difference found between the above described neutral regions. Seven samples are below a read depth of 10x in those regions. Excluding those samples, the average read depth is increased to 24.96x (± 12.27) (see Tables 3.17 & A.37 and Table 3.17 column 4). Both values are higher than the modern average and have a lower standard deviation (see Chapter 2.1.7 Table 2.1). The read depth in the full genome including the capture regions is on average 1.71x (± 0.44) with the averaged standard deviation for all samples being 3.4 (± 2.24). Average values for frequency of heterozygous variants and the differences between CT-TA and GA-AG in the unrefined alignments are very similar between all samples and only those with a read depth >10x. Two distinctions have to be mentioned in the unrefined samples, for one the values for the differences between

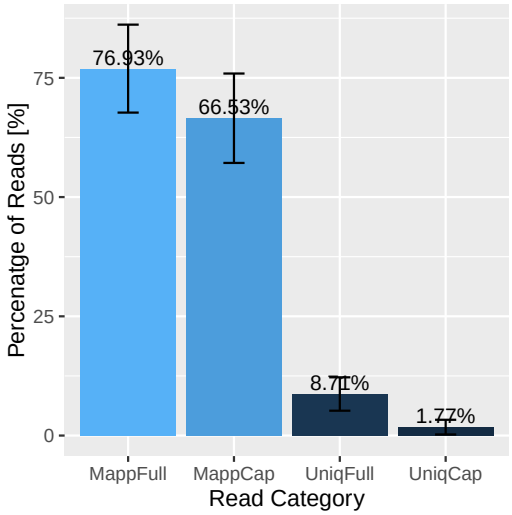



**Figure 3.19:** *Showing the Percentage of all reads mapping to the human genome for different categories; MappFull = percentage of all reads mapping to the whole human genome(including the captured regions); MappCap = percentage of all reads mapping only to the captured regions; UniqFull = percentage of all unique reads mapping to the whole human genome; UniqCap = percentage of all unique reads mapping to only to the captured regions.*

CT-TA and GA-AG, which are ∼0.03 higher and have up to ∼2x higher standard deviation in the samples with >10x coverage(see Table 3.17 columns 2 & 3). As for the second, the total amount of variants includes on average 1816.88 more variants and lower standard deviation, if the samples with low read depth are excluded. For all unrefined alignments it has to be stated that both genotype frequencies, 0.88 for heterozygous and 0.12 for homozygous genotypes, as well as the differences between CT-TA and GA-AG, with 0.30 and 0.31 for the latter, are higher than the modern average (compare Table 3.17 column 4).

The refined alignments are in general closer to the modern average than the unrefined ones, in addition they have >2x less variants than the unrefined alignments. Removing the samples with read depth below 10x shifts all values even closer to the expected frequencies and the muattaion differences calculated on the modern samples (see Table 3.17 columns 5 & 6), with genotype frequencies of 0.66 and 0.34 for the heterozygous and homozygous genotypes as well as 0.03 for *Diff CT-TC* and 0.06 *Diff GA-AG*.

An average of 96757.43 (±52371.86) reads on the X chromosome and of 2345.05 (±1342.67) from the mtDNA were extracted for contamination analysis from the alignments against the full genome. Three samples, WEZ16, 61 and 74, could be typed as female individuals thus not allowing to test for contamination using ANGSD [59]. All other samples are considered male individuals based on inspection of the read depth and coverage in Y chromosomal regions. Sex determination using a regression between region size and sequenced nucleotides or the method described in Skoglund *et al.* 2013 [107] were all inconclusive (results not shown).

Two samples, WEZ63 and 77, showed high contamination on both chromosomes tested. Mitochondrial contamination tested with *ContaMix* [28] (see Chapter 2.1.5 for details) was 9.77% for WEZ63 and 11.04% for WEZ77. X chromosomal contamination ranges from 20.15% (p-value<0.5; Method2) to 23.28% (p-value<0.5; Method1) for WEZ63, and from 33.84% (p-value<0.5; Method2) to 35.43% (p-value<0.5; Method1) for WEZ77. Similar values for X contamination could be obtained in the three female individuals (see Table A.37). The general average for contamination on the mitochondrial genome excluding the two individuals WEZ 63 and 77 is 1% (±0.02) with an estimated sequencing error of <0.01 (±0.001). Averages for contamination estimated on the X chromosome, excluding the contaminated samples and the female individuals, are 1% (±1.4) with p-values of 0.5 (±0.41) for method 1 and 1% (±1.4) with p-values of 0.6 (±0.36) for method 2 (for details see Table A.42).

**Table 3.17:** *Summary statistics for the samples captured.*

| Sample | Unrefined | | Modern Average | Refined | |
|---|---|---|---|---|---|
| | Capture Average | Capture Average >x10 | | Capture Average | Capture Average >x10 |
| Read depth | 18.43 ± 13.69 | 24.96 ± 12.27 | 17.9 ± 12.08 | 18.43 ± 13.69 | 24.96 ± 12.27 |
| het freq | 0.88 ± 0.03 | 0.88 ± 0.04 | 0.62 ± 0.04 | 0.70 ± 0.07 | 0.66 ± 0.05 |
| hom freq | 0.12 ± 0.03 | 0.12 ± 0.04 | 0.38 ± 0.04 | 0.30 ± 0.07 | 0.34 ± 0.05 |
| Diff CT-TC | 0.30 ± 0.04 | 0.34 ± 0.10 | 0.01 ± 0 | 0.06 ± 0.06 | 0.03 ± 0.04 |
| Diff GA-AG | 0.31 ± 0.05 | 0.34 ± 0.07 | 0.01 ± 0 | 0.10 ± 0.09 | 0.06 ± 0.04 |
| Total variants | 10381.05 ± 5017.01 | 12197.93 ± 4611.18 | | 3813.33 ± 1015.41 | 4025.93 ± 379.65 |

### 3.3.3 PCA

The first two principal components from the resulting PCA using the un-refined whole genome data of the her captured WEZ samples, are displayed in Figure 3.20. The overlap with the positions in the reference data set is on average 8.13 (±4.53)% for each capture sample. All of the WEZ samples (red squares) fall into the European variation displayed in the figure. Those can be divided into two groups and some outliers. One smaller wider spread group clustering with “middle and eastern” European individuals consisting of WEZ15, WEZ39, WEZ40, WEZ48, WEZ58 and WEZ83, and a second tighter and larger group clustering closer to the “southern” Europeans including all other individuals except the three outliers WEZ16, WEZ54 and the contaminated WEZ77.

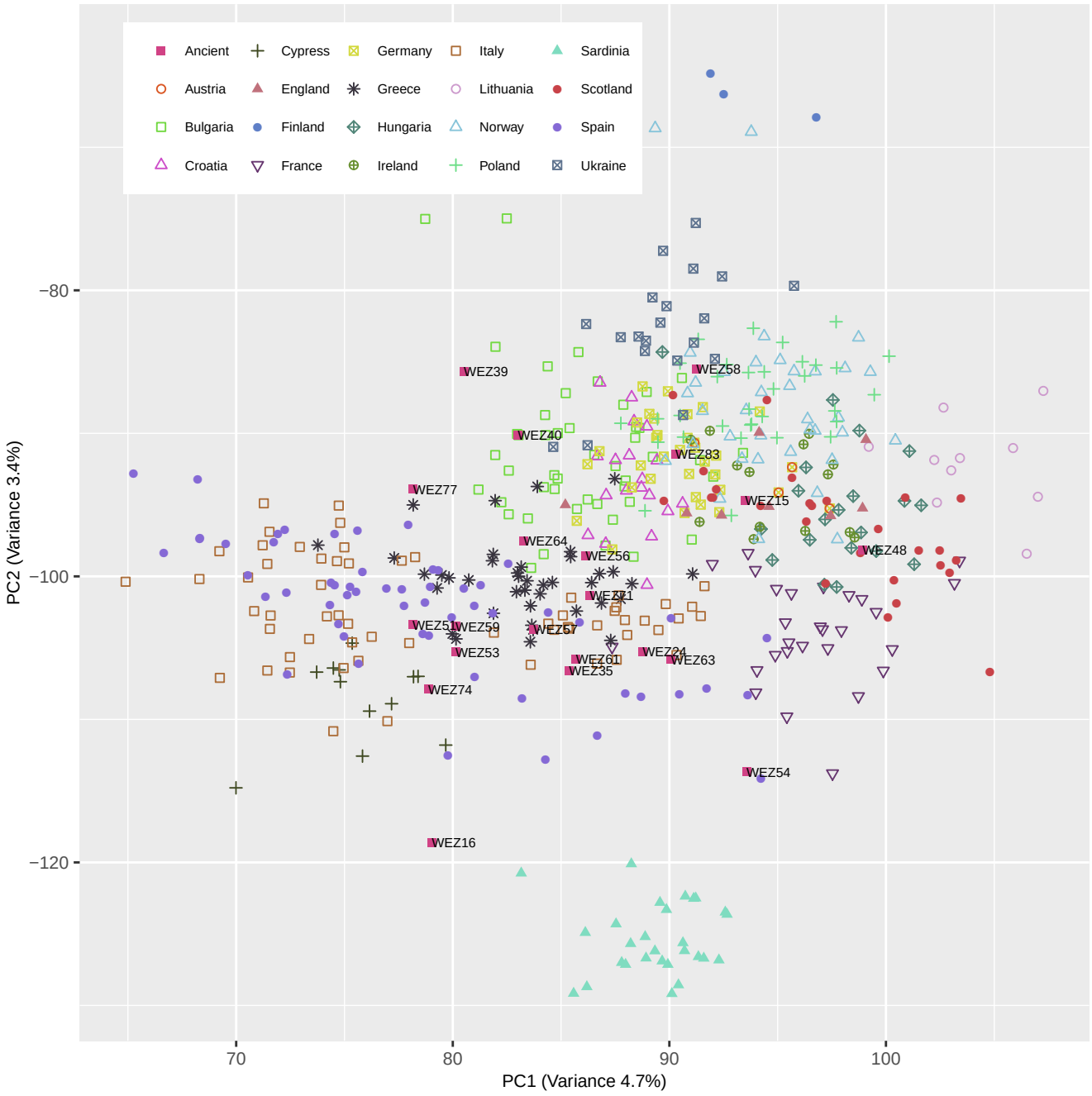



**Figure 3.20:** *Projection of the complete WEZ samples using Procrustes on a PCA generated with a worldwide data set; WEZ samples are shown in red squares and are named; for visualization the majority of populations have been removed and only selected European populations are displayed; the first two principal components are plotted with x-axis = PC1 and; y-axis = PC2.*

The PCA plot with the same reference PCA as used before was performed three more times. Once more on the refined whole genome alignment of the captured samples and twice using only the reads mapping to the Capture regions, with one using the refined alignments and the other with the unrefined ones of the WEZ samples. As displayed in Table 3.18, the refinement does reduce the markers overlapping with the reference positions in all samples on average by 17.5 (±0.02)%. This reduces the average read depth in all reference positions to 0.12 (± 0.06) for the full genome data, which was 0.15 (± 0.08) reads per locus prior to refinement. Using only the capture regions, reduces the average read depth in position from the WEZ samples covering the reference to 0.05 (± 0.04) leaving only 3.9% of the overlap with the reference positions from the full genome alignments of the captured samples. The loss in the overlap between the WEZ samples and the reference data due to the refinement of the capture regions with 7.45%, of the unrefined capture regions, is lower than with in the full genome alignments. It should be pointed out that the Procrustes similarity score is always >0.99. As displayed in Table A.36, the intersection of the designed capture regions with the reference data used for the PCA is 1488 markers.

**Table 3.18:** *Overlap with the reference data used for PCA and average Procrustes similarity; shown in average number of loci and average read depth of the whole reference data set;*

| Alignment Regions | Average Loci# | Average Ref Depth[%] | Average Procrustes Similarity |
|---|---|---|---|
| Capture | 1625.33 ± 200.23 | 0.05 ± 0.04 | 0.99 ± 0.00 |
| Full Genome | 41541.24 ± 23153.87 | 0.15 ± 0.08 | 1 ± 0.00 |
| Refined Capture | 1499.43 ± 207.47 | 0.04 ± 0.03 | 0.99 ± 0.00 |
| Refined Full Genome | 34304.4 ± 19228.83 | 0.12 ± 0.06 | 1 ± 0.00 |

After applying the recalibration methods described in chapter 2.4, all samples are positioned in general further left along PC1 in the PCA than their unrefined alignments against the full human genome (see Figure 3.21). On average the samples positions on PC1 change by 15.22 (±5.72) whereas the change in PC2 is smaller with an average of -4.05 (±3.62). This includes two samples WEZ39 and WEZ48 shifting down along PC2, the opposite direction than the shift for every other sample. Two of the three greatest differences in PC1 between the refined and unrefined alignments can be seen in the two contaminated samples WEZ63 with a difference of 24.71 and WEZZ77 with 22.66 (compare Table A.38). The last of those three samples is WEZ71 with a difference of 22.68.

Using only the capture regions from the refined alignment, the samples fall in general further left and downwards compared to the refined full genome alignment (see Figure 3.22 and Table A.38). The difference in PC1 to the unrefined alignment is on average larger with 33.31 (±12.92) compared to the shift between the refined and unrefined full genome alignments. Although two samples, WEZ71 with a difference in PC1 of 14.99 and WEZ77 with 22.1 shift slightly back to the right along PC1. WEZ63 with a difference of 60.19 further increases the distance towards the unrefined alignment if only the refined capture regions are used in the PCA. The loss in overlap with the reference data due to the recalibration process in the full genome data is 17.55 (±1.85)% and due to the reduction to the capture regions 94.09 (±3.82)%.

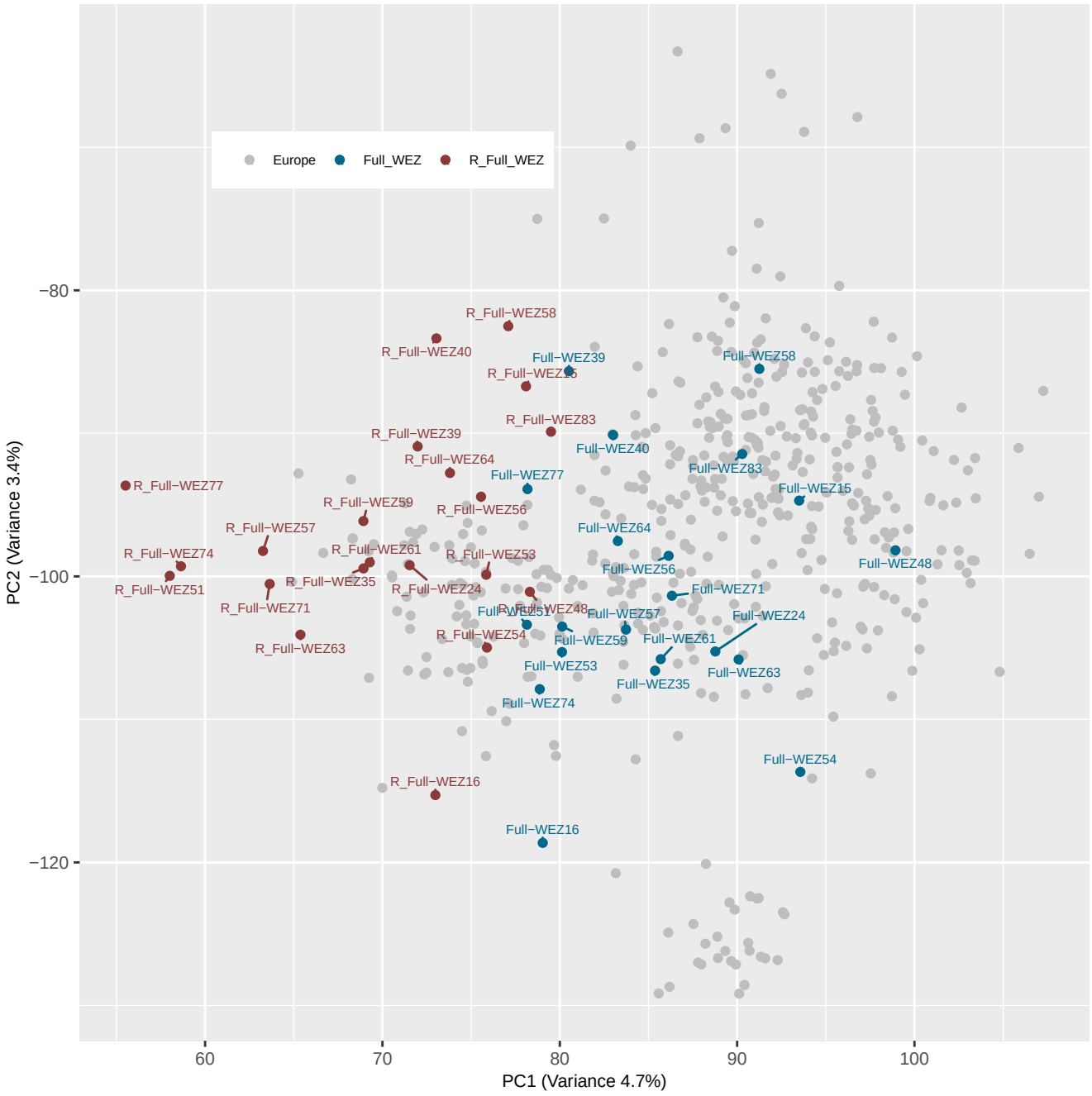



**Figure 3.21:** *Combined plot of two separately performed Procrustes transformations of the refined and unrefined full genome data of the WEZ samples; red = refined samples (R_Full); blue = unrefined samples (Full), the same individuals and positions as in Figure 3.20; grey = European individuals as seen in Figure 3.20.*


**Figure 3.22:** *Combined plot of two separately performed Procrustes transformations of the refined data for the capture regions and the full genome WEZ samples; blue = capture samples (Full); red = full genome samples (R_Full), the same individuals and positions as in Figure 3.20; grey = European individuals as seen in Figure 3.20.*

## 3.4 Discussion

While the conservative and relaxed regions were developed differently, there was no significant difference found in terms of read depth or coverage. After Sequencing the capture, the differences remain in size and thus amount of variable positions (see Table A.36). Whether the stricter choice of parameters in the background selection coefficient, distance from genes as well as between each other has an impact on later analysis remains to be tested. After all if a stricter parameter set is needed for any analysis it can be applied on the generated data afterwards.

In general the here designed capture outperforms a shotgun experiment in information quality. Through the increased coverage in the targeted regions, a higher quality of the included positions can be reached, which can be seen in the summary statistics obtained after the refinement of the alignment (see Table 3.17). After removing the seven samples with less than 10x coverage, the summary statistics improve towards the modern standard. The average heterozygous frequency with 0.66 of the ancient samples falls into the range of the modern average 0.62(± 004). This points towards the increase in coverage as an counter measure to PMD and sequencing error as suggested before (see Chapter 2.4). Although the estimated sequencing error rates obtained with *ContaMix* [28] are very similar throughout the samples, the differences between the ancient samples with >10x is still higher than in the modern average, while the average coverage in the ancient samples is also higher. This can point towards several things including remaining effects of PMD and sequencing error that could not be countered with recalibration, or the difference in the variant calling strategy. The modern samples have been called in a cohort allowing the alleles present in the cohort to influence the genotype call in single samples (see Chapter 2.3.1). While this is useful, to prevent over-interpretation of errors or damage it could also suppress rare variants and thus homogenize a cohort and reduce the contained variation. Although this is speculative, it should be kept in mind while comparing ancient data sets to modern samples, where especially the suppressed rare variants might be of interest.

While the majority of reads fall in the capture regions, the general distribution of the reads sequenced after a capture experiment shows (see Figure 3.19), that a large amount of unique information is available throughout the rest of the genome, ∼7% of all reads are unique reads mapping anywhere in the human genome except the capture regions. This corresponds to ∼79% of all unique reads available in a captured sample. While this information is only available in low coverage it allows for additional analysis like PCA and contamination estimates. The amount of reads mapping to the mtDNA is, even for the sample with lowest coverage, sufficient to use *ContaMix*. Using ANGSD to determine the X chromosomal contamination in males on the other hand is problematic, as the average for p-values in both methods with ∼0.55 shows, without taking females and possibly contaminated individuals into account. Saying that for most of those individuals a contamination estimate is not possible. It seems promising that in female individuals and actually contaminated samples the p-value for both methods is <0.001. As this suggests that contamination will easily be detected and only the proof of absence of contamination is difficult. In combination with the sex determination, that needs to be done visually, those results seem to advocate the addition of loci to the capture. Those should allow for better contamination assessment and sex determination like loci from the X chromosome and the mitochondrial DNA.

Although with the data available here the best positioning in the PCA has to be speculative, the placing of the un-refined alignments using the full genome data seems to represent the best positioning, since this data set has the largest overlap with the reference data set used. With an average amount of 41541.24 (±23153.87) positions this is 1.2x of the overlap in the refined full genome alignments and 27.7x the overlapping positions in the refined capture (compare Table A.38). Further the information we have about some samples as well as the general background add to favoring the PCA with the unrefined full genome data (see Figure 3.20). The Neolithic WEZ16 sample falls close to the Sardinian population, which show a Neolithic like signal in

PCA [29, 54]. Although a similar structure is somewhat kept in all of the different PCAs between the WEZ individuals, the first PCA including the unrefined full genome data, seems to be the only one where all WEZ individuals fall in the European variation shown here.

Although the suggested refinement method and the capture developed here show higher similarity in the selected summary statistics compared to a modern average, the here obtained results seem to suggest that a PCA utilizing Procrustes analysis with LASER [113] could be robust to sequencing error and DNA damage. Simultaneously it seems to be affected more by loss of information. It has to be pointed out that even with the full genome data there was only an overlap 8.13 ($\pm$4.53)%. As already stated, qualifying the positioning obtained during any of those analysis is difficult. Although LASER reports a Procrustes similarity score that allows to qualify the positioning of a sample, in relation to the actual PCA obtained with the down sampled reference, it does not help to qualify the positioning in the PCA used in the Procrustes Transformation.

# 4 Case study Welzin

To further test the potential of the developed capture, it is used here to answer archaeological questions with current population genetic methods formerly applied on shotgun sequencing and SNP array data. To see if this is possible, the sample set from the archaeological site Welzin, used in the previous chapter will be analyzed with those population genetic methods. F3-outgroup, D-statistics and ADMIXTURE will be applied on the samples and the results will be interpreted in the light of the known population structure during the Bronze Age.

## 4.1 Introduction

### 4.1.1 Sample background

The 21 samples available to this study stem from skeletal remains found in the Tollense valley in north eastern Germany and date to the bronze age (ca. 3200 BP), except for sample WEZ16, which dates to the neolithic (ca. 5000 BP) and was found in a burial context. Although several samples from the Welzin site have been dated using the $C^{14}$ method, from the samples used for this study only the neolithic WEZ16 (2960BC ±66) and the Bronze Age sample WEZ15 (1007BC ±102) were radiocarbon dated. All individuals except WEZ16 were found in a non burial context, widely dispersed and dis-articulated [48] along the river bank of the Tollense river. Recent excavations and ongoing research suggest a large scale battle in the Tollense valley with an estimated number of up to 4000, mostly male participants [21]. The artifacts that have been found, include weapons such as wooden clubs, bronze spear and axe heads as well as bronze and flint arrow heads, but also bronze arm and finger rings. So far those findings allow no correlation between skeletal remains and separate cultures or the identification of different parties involved in the conflict.

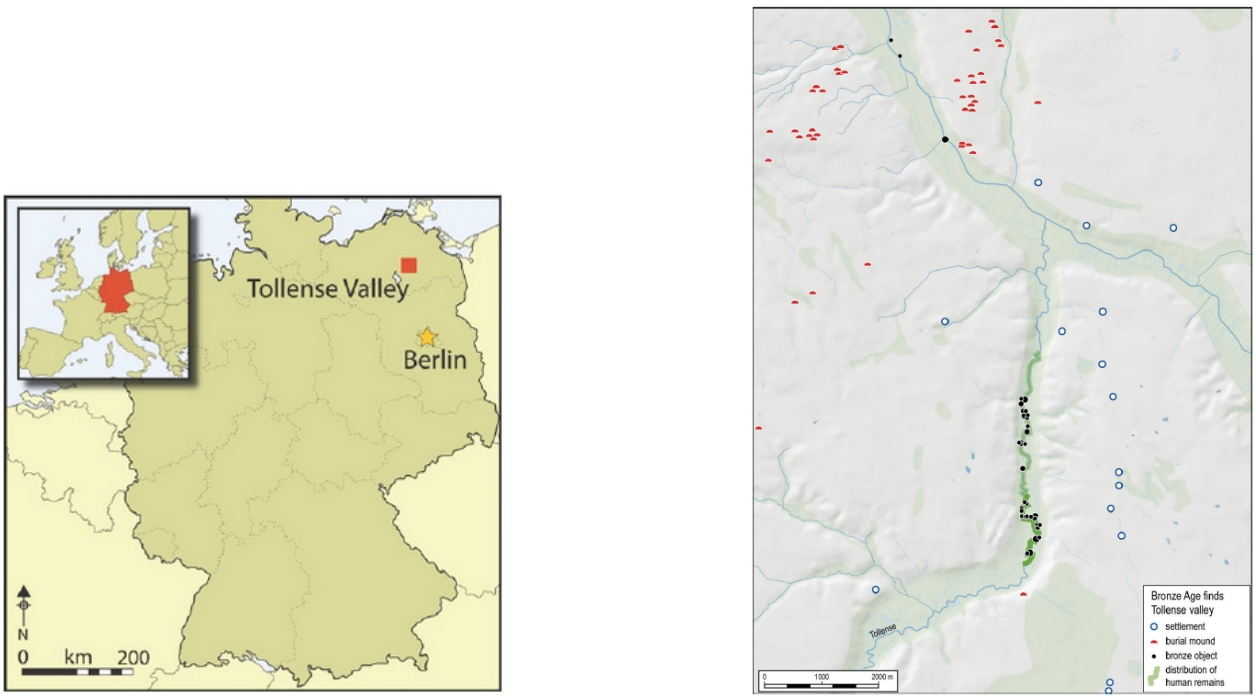



**Figure 4.23:** *On the left: A map of Germany showing the loctaion of the Tollense Valley; on the right: A detailed map of the Tollense valley including the Bronze Age finds; from Jantzen et al. 2010 [48]*

### 4.1.2 Genetic history of Bronze Age Germany

Around ∼8000-7000 years ago, early farmers arrived in Germany, introducing domesticated animals and plants, new pottery and a sedentary lifestyle, known as the Neolithic culture. Individuals associated with Neolithic cultures, like the Linearbandkeramik (LBK), found in today's Germany, show a close genetic affinity to ancient populations from the Aegaeis and western Anatolia, that are linked with the onset of the Neolithic [36, 40, 64]. During the following millenia, European farmers eventually admixed with local hunter-gatherer populations, that were already present prior to the arrival of farmers [12, 36]. About 4000 years ago during the late Neolithic and early Bronze Age period, a second migration event associated with the Yamnaya and Corded Ware culture can be detected. The Yamnaya are a late Neolithic/early Bronze Age culture present in the Eurasian Steppe between ∼ 5,500-4000 BP that shows ancestry components of eastern hunter gatherers and Near Eastern Neolithic populations [36]. The Corded Ware culture is closer to the Yamnaya than any other late Neolithic/early Bronze Age culture in central Europe. Other cultures during the late Neolithic and Bronze Age, like the Bell Beaker or the later Unetice culture are genetically very similar to each other, but show differences in the amount of their Yamnaya ancestry, although still showing similar amounts of Yamnaya ancestry as present day European populations. It has to be pointed out that, all modern European populations can be modelled as a three-way mixture of Western hunter-gatherer, Early Neolithic and Yamnaya populations [36]. Individuals associated with late Bronze Age cultures dating to ∼3,500 BP mostly fall inside modern European genetic variation [5, 36, 75] (see Figures A.46 & A.47 for a graphical explanation of the prehistoric population movements).

### 4.1.3 Archaeological background of the Bronze Age

In the late Neolithic period, ∼5000-4000 years ago, two main cultures can be distinguished in central Europe, the Bell Beaker Culture (BB) and The Corded Ware Culture (CW). The BB arose in western Europe from the Iberian peninsula and was spread from North Africa to the British Isles [90, 115]. The lifestyle of the BB people can be described as traveling artisans skilled in metal work [38]. The CW was spread over an area between the Rhine and Volga and from the Danube up to Scandinavia. They were mostly living of pastoralism with some farming elements. Their burial rites as well as their lifestyle is shared with the Yamnaya culture from the Eurasian Steppe. Archaeological evidence suggests that the BB and CW cultures coexisted for ∼300 years in Central Europe [22, 110]. Both cultures are considered to have brought drastic changes into the existing mainly farming lifestyle. For the BB it is assumed that they introduced metal work into already existing cultures. The CW can be linked to burning down forests creating large free space for their cattle to graze [61, 84]. While metal work with copper and gold was already known by the BB, the actual Bronze Age starts with the use of the alloy bronze ∼4000 years ago and the appearance of the Unetice culture in Central Europe [89]. The dispersal of the Unetice culture widely overlaps with the BB and CW, making it likely that this new culture of metalworkers and miners arose from the substrate of the two previous cultures. Next to technological changes archaeological records indicate social and economic changes, like a commodity based trade for the relatively rare metals needed for bronze [25].
Regarding the limited archaeological context of the Tollense site, one questions seems to be immanent:
Can different parties of a possible conflict be separated ?
Including the genetic prehistory of the Bronze age one can try to associate the Welzin individuals to cultures or find population structure and thus maybe identify the possible parties involved.

## 4.2 Methods

21 samples with an endogenous DNA content >10% were selected according to low depth sequencing or screening results from a total of 46 Samples. The screening and the NGS sequencing libraries were processed

as described in Hofmanova & Kreutzer et al. 2016 [40]. The capture was applied twice, first on the NGS library and a second time on the first capture product. Both captures were performed according to the Mybaits user manual v1 [81]. Sequencing data of the 21 samples was processed using the Pipeline described in Chapter 2. The pipeline was stopped after the realignment step during the alignment refinement (see Chapter 2.1.3) for all sequencing data. The refining of the alignment was done with ATLAS as described below (see Chapter 4.2.1). The whole genome was used during all following methods. Two possibly contaminated individuals, WEZ63 and WEZ77, were removed from the data set (see Tables A.37 & A.40 for further details on the samples).

### 4.2.1 ATLAS

As already mentioned in Chapter 2.4, a tool set called ATLAS was developed by Vivian Link *et al.* [71] on the basis of the here generated results and built on a genotyping model accounting for post *mortem* damage (PMD) [40, 60]. ATLAS can infer PMD, recalibrate base quality scores including sequencing errors and PMD in the qualities, call accurate genotypes and estimate genetic diversity based on genotype likelihoods. The main differences between ATLAS and the pipeline shown in Chapter 2, are the separate estimation of PMD and sequencing errors and the possibility of a reference free base recalibration, which does not introduce a bias as present in GATK [43] and allows non model organisms to be processed as well.
For each of the Welzin samples three files where generated, the recalibrated bam file and two variant calls in the vcf format [23] according to the paired end pipeline available at [70] by Vivian Link. As input for ATLAS, realigned bam files were generated according to Chapter 2.1.3. With the exception, that duplicates were only marked instead of being removed, as suggested in the original pipeline. The duplicated reads were kept to ensure that the needed thresholds for the recalibration using *recal* [70] could be reached. The *recal* method utilizes regions in a genome with no polymorphisms or only haploid genotypes, like the X chromosome in male individuals, to estimate sequencing errors. To use this reference free recalibration method, samples were grouped according to sequencing runs and lanes if possible, thus allowing to use sample specific PMD detection and lane or run specific sequencing error estimation. Because of this grouping the sequencing error in female individuals (WEZ16, WEZ61 and WEZ74) could be recalibrated using *recal* with the information obtained from the male samples, sequenced on the same run/lane.
The two variant calls available, are the *allelePresence* call and a maximum likelihood estimate of the diploid genotype (MLE). Whereas the latter produces diploid genotypes the *allelePresence* calls only the most likely single allele at a position and thus generates a haploid genotype. To be able to use such a haploid call with a diploid reference data set it has become practice [36, 40, 63, 64, 75] to double the called allele at all positions and generate a pseudodiploid genotype. While previous studies were limited to picking the most frequent base at a given position as the most likely allele, ATLAS assigns a quality to each of the alleles based on the likelihood, including PMD, sequencing error, base and mapping quality of all bases covering the position. The variant calls need to be filtered and all samples merged into one vcf file for further use. Filtering for genotype quality and coverage as well as generating pseudodiploid genotypes out of the *allelePresence* calls was done by a bespoke python script (see additional file pseudoDIP_filter.py; see Table A.43). Merging all samples for each of the filtered data sets and generating subsets including only positions of the used reference data was done using *bcftools* [18]. To remove possible duplicated sites because of different alleles at a position the option -m was set to all, to merge different alleles of a given position into one entry and including indels as well as SNPs in one position. The X and Y chromosomal positions as well as all positions on the mitochondrial genome have been excluded for all further analysis.

### 4.2.2 Relatedness

For the later applied population genetic methods only unrelated individuals can be used, therefore the samples needed to be tested for possible relatedness between each individual. While the Plink software [94] is able to test for relatedness between two individuals, the use of the here generated pseudodiploid calls would have resulted in an overestimation of the identity by descent (IBD). Instead an approach by Lipatov *et al.* 2015 [72] for the maximum likelihood estimation of relatedness is used. It includes genotype likelihoods and thus the uncertainty involved in variant calls with low coverage, which is also not possible using Plink. The software lcMLkin [73] includes a python script, SNPbam2vcf.py, which allows to generate those likelihoods directly from a bam file. To make use of this approach the bam files recalibrated with ATLAS were used. Due to differences in the vcf format generated by ATLAS compared to the format needed by lcMLkin it was not possible to use the vcf files from ATLAS directly. To reduce time and memory usage all possible variant sites in the Welzin samples were extracted from the MLE calls and thinned to achieve a distance $>$ 100k bp between all variable positions using *vcftools* [23]. The resulting positions were used to generate genotype likelihoods for all positions in all samples using SNPbam2vcf.py. Since no relatedness could be found and the the number of positions was always below the suggested threshold results are only shown in supplementary files (see file WEZ.relate in Table A.43).

### 4.2.3 Plink and the reference data set

To compare the Welzin individuals with other ancient and modern populations two data sets were used: Lazaridis *et al.* 2016 [63] published a large collection of ancient samples, collected from several publications, which is available in the EIGENSTRAT format and can therefore be used for the later described outgroup f3, D-statistics and ADMIXTURE [3] analysis. For the use with LASER [113] and the projection of the ancient samples in a PCA, selected bam files were separately downloaded and processed as described in Chapter 3.2.4. Independent of the analysis, all samples were kept in the groups and detailed labels as defined in Lazaridis *et al.* 2016 [63]. For several groups the detailed labels are used to achieve a more precise grouping in the geographical region and time span of the Welzin individuals. See Table A.39 for a detailed information about the publication, the grouping, the age and location of the samples. For the comparison with modern populations the combined data from Hellenthal *et al.* [37] and Busby *et al.* [16] was used as processed in [40]. Additionally, to allow for better visual comparison, the reference space from [40] was used for both PCAs in this Chapter.

To combine the calls of the Welzin individuals generated in the vcf format with data from the references [16, 37, 63] the Welzin data had to be converted to Plink ped/map format [95] using Plink 1.9 [94] and the merged data into the EIGENSTRAT format using the program *convertf* [91], as part of the later described Admixtools [85, 86]. The Plink ped/map format can store variants of multiple individuals and consists of two files the ped and the map file: The ped file holds individual genotypes separated by white spaces with individuals separated by new lines, whereas the map file holds the information for each of the positions separated by new lines. Thus a column in the ped file represents a line in the map file. To ensure the Welzin data and the reference data are all referring to the same alleles, the reference alleles of the EIGENSTRAT data from [63] were used with the option –reference-allele in Plink ped/map files prior to merging any of the data sets, to force a definition of reference an alternate allele according to the reference data. Additionally the rs number or SNP identifier in the map file was changed to "chr:position" for all SNPs in both, the reference and the Welzin data set, to allow merging in Plink. Sites that were reported as having more than two alleles in the combined data set, including the Welzin samples and all reference individuals, were removed, since Plink 1.9 is not able to handle more than two alleles at a given position. The resulting ped/map files consisted only of biallelic positions from all samples available in Lazaridis *et al.* 2016 [63] and overlapping positions from

selected modern populations [16, 37] and all Welzin samples.

### 4.2.4 Admixtools

*Admixtools* [85, 86] is a collection of programs that allow to calculate several population genetic statistics. In this case it was used to calculate outgroup f3 and D-statistics. Further it was used to convert between the Plink ped/map and the EIGENSTRAT formats. The first format is mainly used during merging and filtering of the data and the latter is needed in *Admixtools* to calculate f3 or D-statistics. The programs used are *qp3Pop* for f3, *qpDstat* for D-statistics and *convertf* [91] for the format conversion. The latter is available with *Admixtools* but was developed separately.
*A Note on convertf:*
While theoretically *convertf* should be able to use the binary Plink bed format as input, there is an error in *convertf* that produces false EIGENTSRAT files from the binary Plink bed files. The information for having two or none of the reference alleles is switched in the EIGENSTRAT geno file, while using the binary Plink format. After conversion to EIGENSTRAT, positions showing two reference alleles in the geno file showed no reference alleles in both Plink formats, the vcf or the bam format and positions with no reference allele at a given position in the geno file showed two reference alleles in all other used file formats. It is therefore necessary to only use the human readable Plink ped format with *convertf*.

#### *Outgroup f3*

Outgroup f3 statistic is a special case of the f3 statistics. f3 is generally described as $F_3$(C:A,B), with C being the target population that is tested for admixture with populations A and B. Setting C to a population that has not shared any gene flow with populations A or B since they diverged from C, results in positive values for the f3 statistics proportional to the length of the shared drift between population A and B. If one of the populations, let it be A, is fixed and only population B is exchanged, a relative measure of genetic similarity for all populations B to A can be inferred, based on the value of the f3 statistic. Since population C can be understood as an outgroup for populations A and B this special case is called outgroup f3 and will here be used in the form $F_3$(outgroup:test,sample). The Mbuti population from the reference data will serve as an outgroup, as has been done before [63]. Test will be substituted with all populations/individuals for which the relative similarity to the sample is of interest. This will be repeated with sample being set to each of the Welzin individuals and possible groupings of those ( see Chapters 4.3.3 & 4.4.1. To account for the pseudodiploid genotypes generated from the *allelePresence* calls, the inbreed option was set to *YES* according to the recommendations in Admixtools [85].

#### *D-statistics*

D-statistics can be depicted as D(W,X;Y,Z) [86]. Setting Z as outgroup like described before in Outgroup f3, the violation of the tree (((W,X)Y)Z) is tested by estimating relative derived allele sharing between W and Y against X and Y. If the D-statistic significantly deviates from 0, it suggest inconsistency of the tested tree and some form of unique shared ancestry in the populations ancestral to W, X and Y. This shared ancestry can be the result of either population structure or gene flow. A positive value will indicate shared ancestry between W and Y, and a negative D-statistic will indicate shared ancestry between X and Y.
To test for population structure D-statistics of the form D(WEZ,test,WEZ,Outgroup) were performed. Substituting WEZ for each Welzin sample and test with all available ancient populations form the reference data set [63]. Any significant negative result will indicate that there is no shared ancestry between two WEZ individuals but rather between the Test population and the second Welzin sample in the tree. Additionally it was used to test if the results of the outgroup f3 statistics could be verified with D-statistics by choosing X and Y populations so that population X is followed by population Y in a given outgroup f3 ranking.

### 4.2.5 Admixture

ADMIXTURE is a program for estimating ancestry proportions of individual samples in a model-based manner from large autsomal SNP genotype data sets [3]. Prior to the ADMIXTURE analysis, Plink [94] was used to remove SNPs that showed evidence for linkage disequilibrium by setting $r^2$ to 0.5 and sliding window size to 200 SNPs with a 25 SNPs sliding step to prune the data set. Additionally the data set had to be filtered for positions that are not covered in any sample. Only a missingness <95% was allowed for a given locus. Admixture was used in two ways. First, to estimate the number of ancestral populations (K) unsupervised, without using the information of known ancestry. Second the analysis is run supervised, by setting populations that are supposed to be ancestral as fixed and using a known K.
According to previous studies [5, 36, 75] at least three cultural groups could be expected as ancestral populations for the Welzin individuals, the Yamnaya, any early Neolithic culture and hunter-gatherer cultures. Therefore, all individuals associated with the Yamnaya culture available in the reference data set [63] were used (Yamnaya_Kalmykia & Yamnaya_Samara; according to detailed labels see Table A.39). As representatives of neolithic populations all individuals associated with the LBK in Germany (LBK_EN) were chosen and all available hunter-gatherer individuals were included. Because of similar dating and geographic proximity, individuals from the Unetice, the Bell Beaker and the Corded Ware Culture found in Germany were also included. This will be referred to as the large data set. The unsupervised analysis was run for 100 iterations with cross validation for each K2-K6 to estimate the best number of ancestral populations for the data set. The supervised analysis was run once with K2 as the best K (see Chapter 4.3.4) after cross validation, setting all HG populations combined to one and LBK_EN as ancestral.
A second ADMIXTURE analysis was performed, unsupervised and supervised, with a reduced data set only including individuals from LBK_EN, Western hunter-gatherer (WHG) including the Switzerland_HG as a WHG individual, Swedish hunter-gatherer (SHG) and the Welzin individuals. This will be referred to as the small data set. The unsupervised analysis with the cross validation was done for K2-3 and 100 iterations. A supervised analysis was done again with K2 as the most likely K after cross validation (see Chapter 4.3.4), with this small data set, setting again all hunter-gatherers, WHG and SHG combined as one and LBK as the other ancestral component.
To visualize the estimated ancestry from the ADMIXTURE outputs PONG [9] is used. It can compare and combine several runs on the same data set and allows a user friendly visualization of the results. It combines given ADMIXTURE runs according to a pairwise similarity (default value 0.97) and displays all possible modes from the distribution of ADMXTIURE runs, while also choosing a representative run from the most likely mode. Which is defined as the mode with the highest support given by the number of ADMIXTURE runs that can be combined under the given pairwise similarity. From here on I will use mode to revere to a combination of ADMIXTURE runs under a given pairwise similarity.

## 4.3 Results

### 4.3.1 Reference overlap

In Table 4.19 the missingness and overlap with the reference positions of a subset of Welzin samples and reference individuals is shown, to show the heterogeneity of the data with regards to the number of positions available. The missingness is defined as the fraction of missing genotypes in the data set and is given as a frequency. What needs to be pointed out is the general heterogeneity of all ancient populations and individuals in comparison to the modern populations. Both groups have an average of 61% of missing positions, whereas the standard deviation of the ancient samples is with $\sim$ 30%, 15x higher than in the modern samples. All samples from the Tollense valley have a higher missingness than the ancient reference including the standard

deviation, with an average of 95% ± 3%. A similar missingness is reached by the Nordic_LBA individual and the Nordic_BA populations (see Table A.41 for details on all Welzin samples).

**Table 4.19:** *Showing missingness per individual/population in the used reference [63]; #Pos = the number of actual overlapping positions; Individuals = states the number of individuals available for a given label; in case of > 1 individuals, average values ± standard deviation are shown*

| Population | Missingness | #Pos | Individuals |
|---|---|---|---|
| WEZ16 | 0.99 | 15671 | 1 |
| WEZ54 | 0.93 | 83331 | 1 |
| WEZ56 | 0.93 | 79734 | 1 |
| WEZ57 | 0.92 | 98550 | 1 |
| WEZ | 0.95 ± 0.03 | 61891.79 ±34125.33 | 19 |
| Nordic_LBA | 0.93 | 92836 | 1 |
| Nordic_MN_B | 0.76 | 293134 | 1 |
| Nordic_BA | 0.93 ±0.02 | 89050 ± 27091.16 | 3 |
| Nordic_LN | 0.60 ± 0.34 | 489971.5 ± 414461 | 4 |
| SHG | 0.44 ± 0.16 | 681212.83 ± 195070.34 | 6 |
| Unetice | 0.57 ± 0.25 | 512774.73 ± 295403.17 | 11 |
| All Modern | 0.61 ± 0.02 | 480632.318 ± 20375.49 | 349 |
| All Ancient | 0.61 ± 0.30 | 479105.38 ± 367424.38 | 309 |
| Ancient Reference | 0.59 ± 0.29 | 509472.17 ± 361995.13 | 290 |

### 4.3.2 PCA

The PCA in Figure 4.24 shows modern Eurasian individuals in grey and ancient individuals in colour according to their assigned population (for details on the modern populations see Figure A.48). The majority of Welzin individuals fall within the variation of modern populations from the northern central part of Europe (compare Figure A.48), with hunter gatherers, the Yamnaya and the LBK populations appearing on the outer range of PC1 and PC2. The Nordic_LN, the Unetice_EBA as well as the Nordic_BA population fall in a similar range as the Welzin individuals. With the first two stretching more along PC2 towards the Corded Ware and Yamnaya clusters. The Bell Beaker individuals are more wide spread and expand more along PC1 compared to the Welzin cluster. Outliers from the Welzin cluster are: WEZ16, which falls closer to the Sardinians and neolithic LBK along PC2, WEZ54, which clusters with the Basques and also fall closer to LBK individuals along PC2, WEZ57, which falls in between the former individual and the Welzin cluster, and WEZ56, which separates from the main cluster of Welzin individuals along PC2 in the opposite direction as the former three, towards the Corded Ware or Yamnaya.

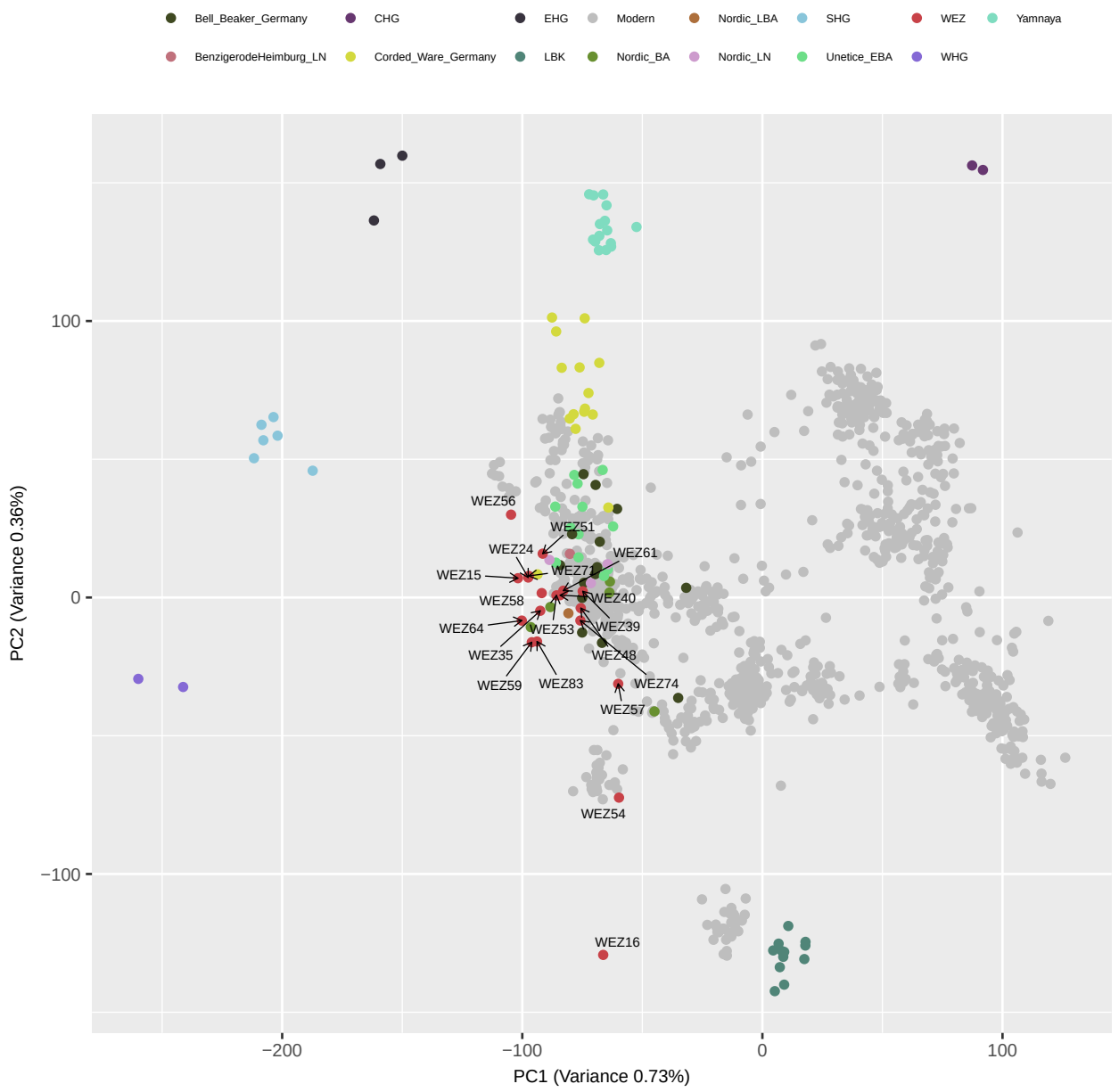



**Figure 4.24:** *PCA of Weltzin samples and selected ancient samples [63] projected on the reference space used in Hoffmanova & Kreutzer 2016 [40]; for a more detailed view of the modern populations see Figure A.48*

### 4.3.3 F3 and D-statistics

The Welzin individuals except the outliers WEZ16, WEZ54, WEZ56 and WEZ57 based on the PCA results were grouped in one population. The results of the outgroup f3 analysis as described in Chapter 4.2.4 with the grouped Welzin individuals substituted for WEZ can be seen in Figure 4.25. The left plot shows the top 15 results in hierarchical order, using ancient populations as test population and the right plot shows all results for the compared 16 modern populations. The ancient population that share the most drift with the Welzin group are WHG and the SHG population followed by the Unetice, the Bell Beaker and the Corded Wear. Starting with the Unetice the following f3 values fall in the range of the standard error of each other. The average difference between two consecutive f3 values is 0.0021 ± 0.0024 and the average standard error in each f3 value is 0.0037 ± 0.0007. The most similar modern populations are the Polish, Austrians and the Scottish. The differences between f3 values are as small as in the ancient populations, with an average difference between two consecutive f3 values for the modern populations of 0.0025 ± 0.0048 and an average standard error of 0.0035 ± 0.0008 (precise f3 values can be found in supplementary files for all compared populations; see Table A.43).

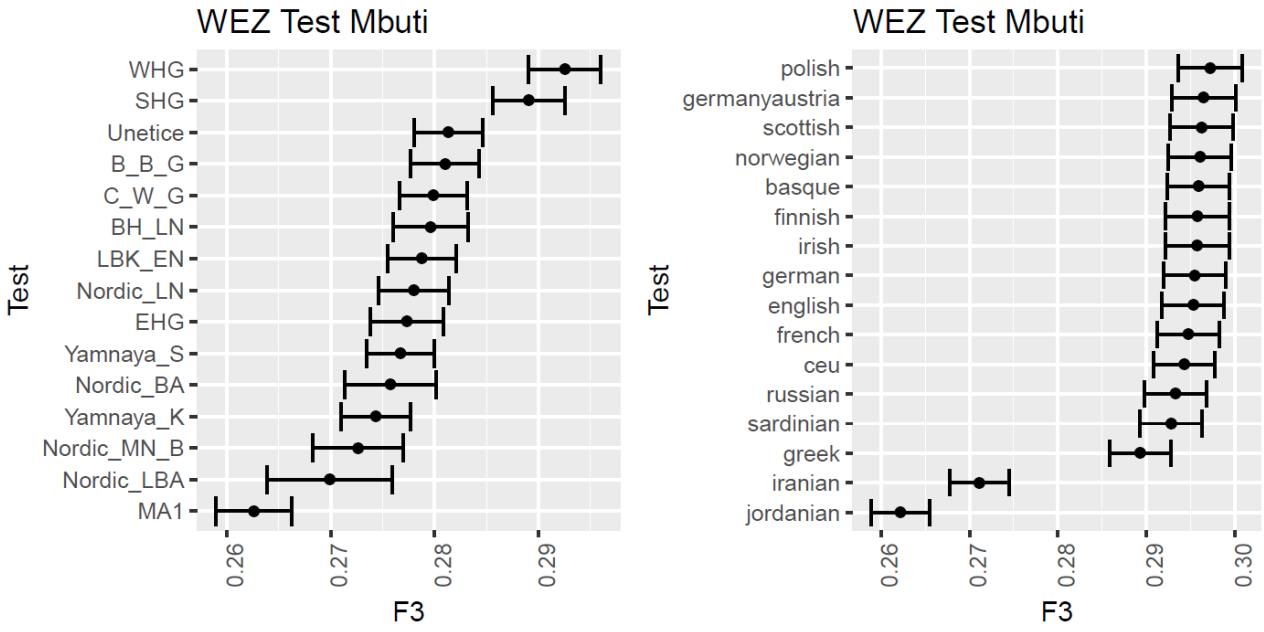



**Figure 4.25:** *F3 outgroup statistics for the Welzin group; on the left,the ancient populations using detailed labels from Table A.39; B_B_G = Bell Beaker Germany; C_W_G = Corded Ware Germany; BH_LN = BenzigerodeHeimburg_LN; on the right with selected populations from the modern reference data set described in Chapter 4.2.3; the tree tested, is shown on top left of each graph; the f3 values are depicted along the x axis with the standard error as error bars.*

It has also been tested which ancient and modern populations shares the most genetic drift with each of the Welzin individuals itself. Shown in Figures 4.26 and 4.27 are the outgroup f3 statistics for the ancient and modern populations of the PCA outliers WEZ16, WEZ54, WEZ56 and WEZ 57. While consecutive f3 values again fall in the error range of each other, the amount of shared drift to other populations is different for each of the samples, also compared to the results from the Welzin group shown above. A striking difference can be found in the results for individual WEZ54, which shows Nordic_LBA as the populations with the highest amount of shared drift. While the other three individuals keep the first two populations in the same order as the results from the Welzin group. The amount of shared drift with the LBK_EN population is in WEZ16, WEZ54 and WEZ57 higher than in the analysis using the Welzin group. In the comparison of the modern populations WEZ16, and WEZ54 share more drift with the Sardinians and the basque population than the Welzin group. In all plots, it can be seen that the error is higher if the compared reference samples have a high missingnes. This is the case in the Nordic_BA and the Nordic_LBA, both with a missingnes of ∼ 93%

(see Table 4.19) and the largest error bars for each plot (for specific values of all Welzin individuals as well as plots for the outgroup f3 of the remaining individuals see supplementary files; Table A.43) .

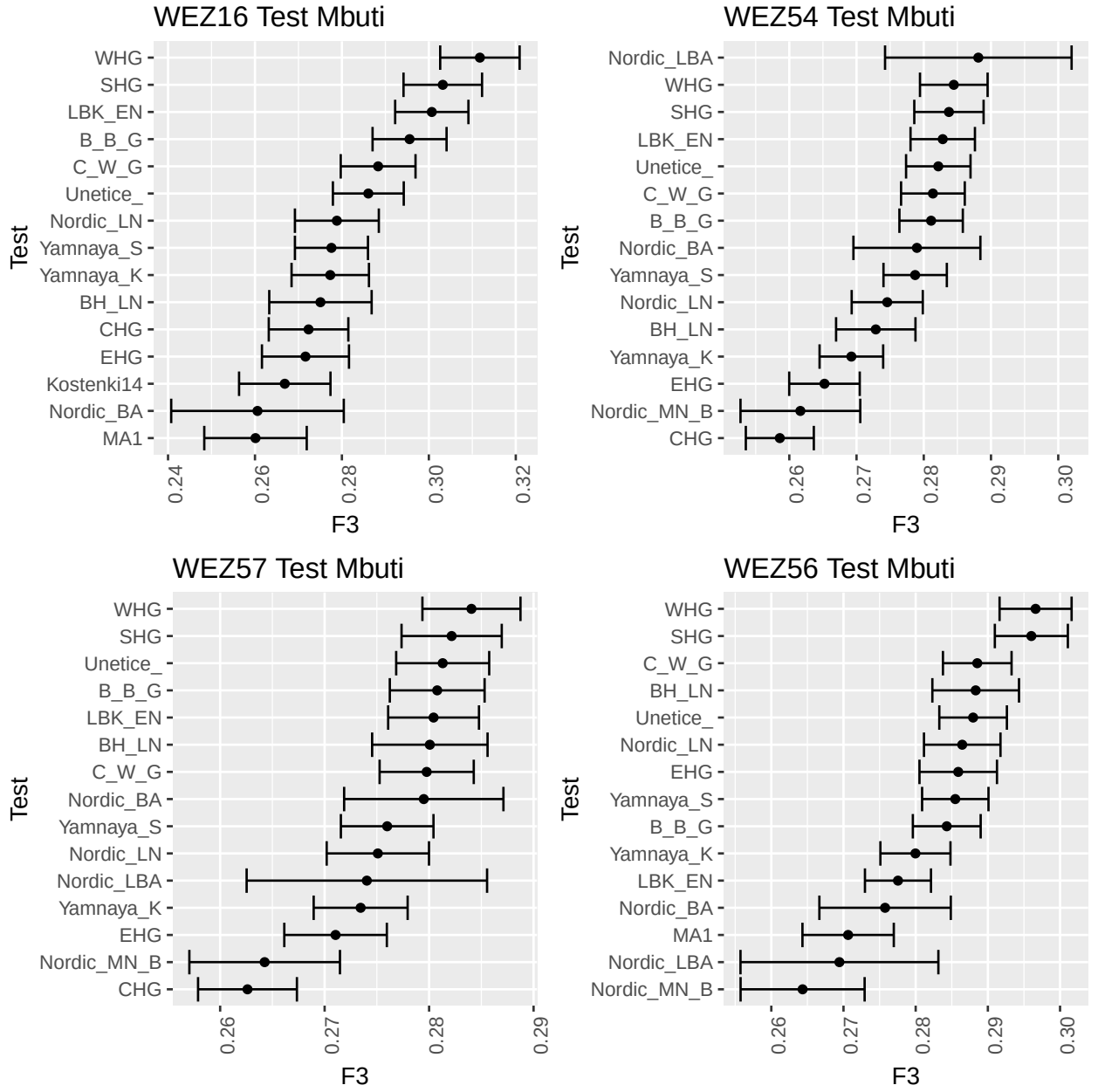



**Figure 4.26:** *Top 15 outgroup f3 statistics for WEZ16, WEZ54, WEZ56 and WEZ57 with the ancient populations using detailed labels from Table A.39; B_B_G = Bell Beaker Germany; C_W_G = Corded Ware Germany; BH_LN = BenzigerodeHeimburg_LN; the tree tested is shown on top left of each graph; the f3 values are depicted along the x-axis with the standard error as error bars.*

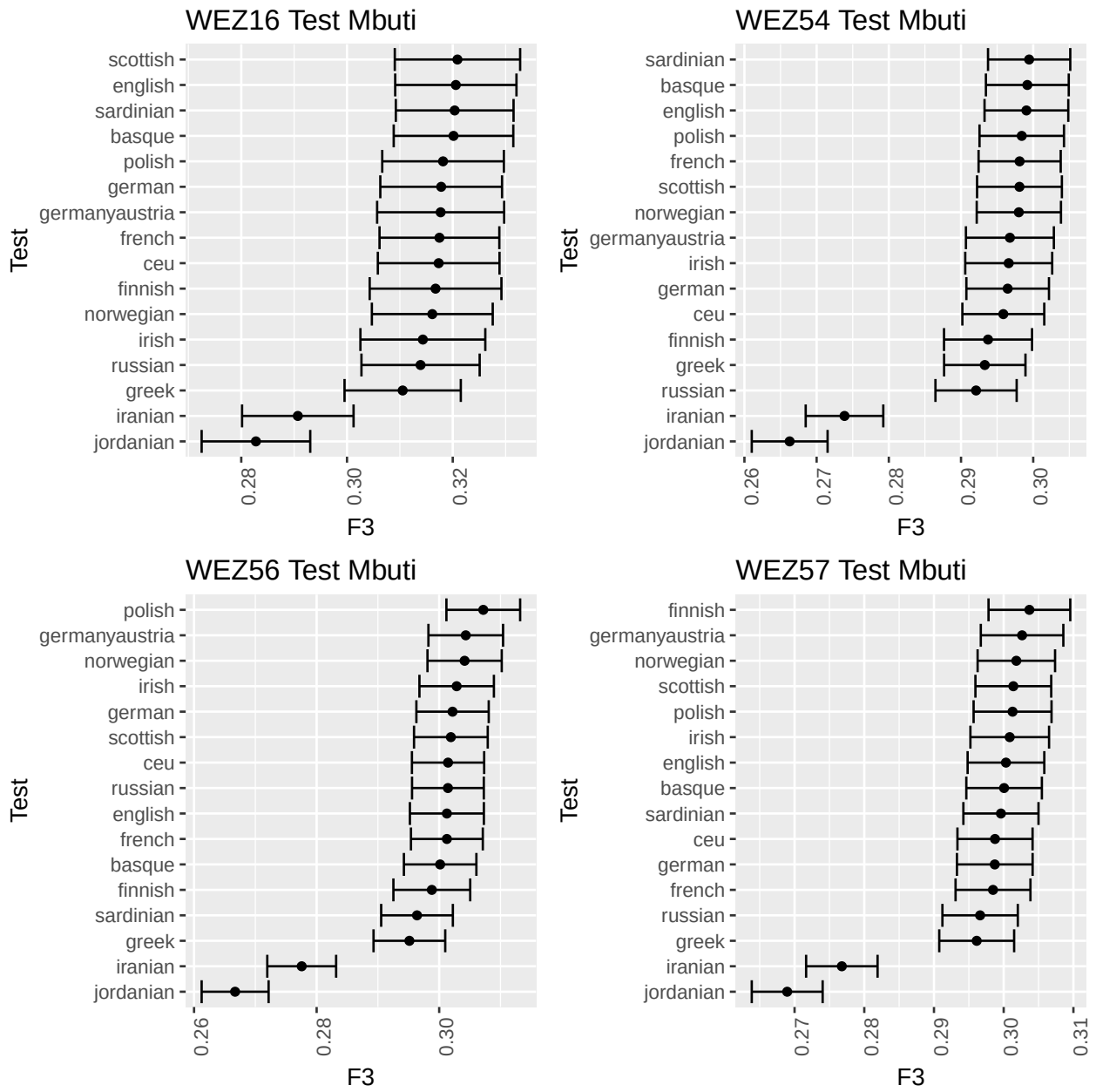



**Figure 4.27:** *F3 outgroup statistics for WEZ16, WEZ54, WEZ56 and WEZ57 with selected populations from the modern reference data set described in Chapter 4.2.3; the tree tested is shown on top left of each graph; the f3 values are depicted along the x axis with the standard error as error bars.*

**Table 4.20:** *Shown are significant D-statistics for the WEZ population; The firts four columns show the populations set in the tree (((W,X)Y)Z), followed by the values for the D-statistik and the Z-score. The last line shows the only negative value optained by using single Welzin individuals for W and Y.*

| W | X | Y | Z | D-statistik | \|Z-*score*\| |
|---|---|---|---|---|---|
| Unetice | WHG | WEZ | Mbuti | -0.002576 | 9.112 |
| Unetice | SHG | WEZ | Mbuti.DG | -0.002118 | 7.769 |
| SHG | Bell_Beaker_Germany | WEZ | Mbuti | 0.002211 | 7.883 |
| WHG | Bell_Beaker_Germany | WEZ | Mbuti | 0.002683 | 9.379 |
| polish | french | WEZ | Mbuti | 0.000684 | 4.979 |
| norwegian | french | WEZ | Mbuti | 0.000376 | 3.058 |

The D-statistics to test for a population structure among the Welzin samples showed significant negative values (|Z-*score*| >3) only if D(W,X;Y,Z), with both W and Y assuming any Welzin individual and Z being set to an outgroup, if X was chosen as a Welzin individual as well. Showing for all Welzin individuals, that they are closer to each other than any other ancient population. Each combination of Welzin individuals was tested, including the contaminated ones, with each of the available ancient populations from the reference data set, resulting in 25200 trees of which 7688 gave significant results (the details for each tested tree that gave any result are available from the file F4_WEz_test_Wez_Out.log in the supplementary files; see Table A.43). Additional D-statistic were run on all combinations of the five closest ancient and ten closest modern populations to the Welzin group according to outgroup f3 statistics. Only significant values are shown in Table 4.20. The significant D-statistic results support the outgroup f3 analysis, by pointing towards a shared ancestry between the WEZ group and the population that has more shared drift according to outgroup f3, of the two other populations, except the outgroup, included in D(W,X;Y,Z). For example in the Test D(Unetice,WHG;WEZ,Mbuti) the result is negative, thus supporting the tree (((WHG,WEZ)Unetice)Mbuti) and pointing towards more shared ancestry between Welzin and WHG than between Welzin and the Unetice.

### 4.3.4 Admixture

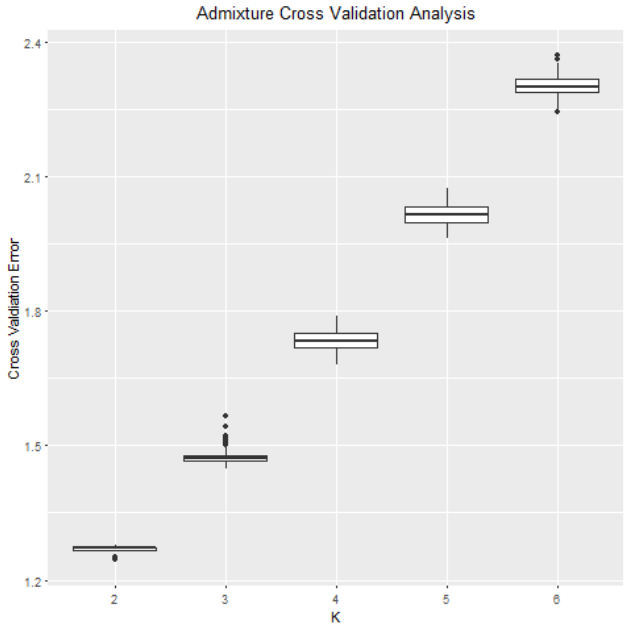



**Figure 4.28:** *Showing the cross validation errors for K2-K6 from the First admixture analysis*

For the ADMIXTURE runs with the large data set, the lowest cross validation error can be reached with K2 (see Figure 4.28). The analysis of the large data set using K2 is shown in Figures 4.29, A.50 & A.49. Any K >2 resulted in 100 modes from the distribution of 100 runs using a pairwise similarity of 0.90. The most likely

mode for K2 in the larger data set includes 35 ADMIXTURE runs with an average pairwise similarity of 0.923 and displays the LBK population (red) as one ancestry component and a combination of Yamnaya, EHG, Switzerland_HG and most parts of the WHG as the other (blue) (see Figure 4.29). Looking at the 8 modes available (compare Figure A.49) the next likely one, combining 24 ADMIXTURE runs with K2 and an average pairwise similarity of 0.932, changes the WHG population and the Switzerland_HG to the same component (red) as the LBK population, removing most of the the blue component from the Welzin individuals and redistributing both components in the Corded Ware and the Bell Beaker populations. The main difference of the other 6 modes, consisting of 41 ADMIXTURE runs are mostly changes in single Welzin individuals. Despite the upside down change of the orientation in the bars, the supervised run with the hunter-gatherers and LBK set to ancestral, resulted in a plot similar to the most likely mode, with changes in four of the Welzin individuals and one Europe_LNBA individual, as well the removal of the LBK_EN (red) component from the WHG ( see Figure A.50). It has to be pointed out that the individuals WEZ16,WEZ54 and WEZ57 show a higher proportion of LBK_EN ancestry (red) than combined hunter-gatherer ancestry (blue) in all plots. Especially WEZ16 shows in all modes only LBK_EN ancestry, WEZ54 in all but one mode consisting of 13 ADMIXTURE runs. WEZ56 on the other hand shows the most hunter-gatherer ancestry of all samples in all plots. The ancestry proportions of WEZ56 only vary in the second most likely mode combining 24 ADMIXTURE plots, but still showing more hunter-gatherer than LBK ancestry.

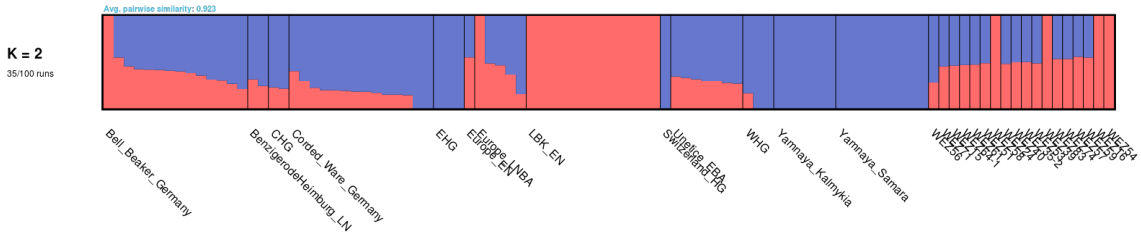



**Figure 4.29:** *Showing the most likely ADMIXTURE result for K2 for the large data set; 35 of 100 runs with a pairwise similarity of 0.923 show similar ancestry proportions.*

As can be seen in Table A.41 the large ADMIXTURE analysis had less variants available than the following one with the smaller subset of individuals. The difference between the two used data sets except the individuals included, is the order of filtering and sub setting the data set. While the large set, was first sub set in individuals and then filtered for LD and missing genotypes, the following small data set was first filtered and than individuals were sub set. Additionally sample WEZ48 is missing form the above analysis due to an error. The small subset consisted only of the LBK_EN, the SHG, the WHG including the Switzerland_HG as WHG, and the Welzin population. This small set was only run with cross validation for K2 and K3. The cross validations errors were 1.88 $\pm$ 0.04 and 2.68 $\pm$ 0.7 respectively. The most likely mode for K2 chosen by PONG, is shown in Figure A.51 with the support of two runs and an average pairwise similarity of 0.902. The supplementary file pong_k2_small_all_modes.pdf shows all 90 modes available, allowing an average pairwise similarity of 0.90 (see Table A.43). Five modes each consisting of a single run show a difference of ancestry proportions in any other individual than Welzin. All 85 other modes and thus 95 ADMIXTURE runs show different ancestry proportions only in the Welzin individuals. Shown in Figure 4.30 is the supervised analysis of the small subset with K2, all hunter-gatherers combined as one known ancestry component and LBK_EN as the other. Again WEZ56 is the individual with the highest amount of hunter-gather ancestry and WEZ16, WEZ54 and WEZ57 have the highest amount of LBK_EN ancestry of all Welzin individuals.

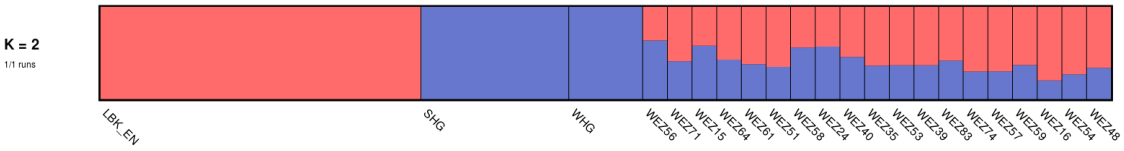



**Figure 4.30:** *The plot of a supervised ADMIXTURE analysis including WHG, SHG, LBK and Welzin individuals with K 2; WHG and SHG were combined to one fixed ancestry component, LBK was set as the other*

## 4.4 Discussion

### 4.4.1 Interpretation of the results in context of population history and Archeology

According to existing studies, that include populations in a similar time frame, Bronze Age populations in Central Europe are genetically very close to the modern populations of respective regions [5, 36, 75]. This is also true for the Welzin individuals, as can be seen in the PCA (see Figure 4.24). With the exception of four individuals the Welzin samples fall close to modern central European populations. Their distribution is similar to the Nordic_BA and Unetice population with a slightly higher affinity towards the WHG and SHG individuals along the first principal component. This affinity is also reflected in the outgroup f3 results for the Welzin group. They show the WHG as the closest population and the SHG as second to the WEZ individuals, followed by the Unetice, Bell Beaker and Corded Ware culture, which fits the assumption that the Welzin population as a whole belongs to the Bronze Age populations in Central Europe (see Figure 4.25). Under the assumption that WHG ancestry can still be traced in all modern European populations [36], and that a early Bronze Age population is close to modern population, both the PCA and outgroup f3 results are plausible. One could argue against the outgroup f3 results, that the Unetice, Corded Ware and Bell Beaker populations are not included by the statement that all modern European populations can be modeled as combination of Yamnaya, early Neolithic and WHG populations, but since the Unetice, Bell Beaker and Cord Ware were shown to be a mixture of those three [5, 36, 75], the f3 results still hold. Since f3 measures shared genetic drift between two individuals/populations, these results suggest that the Tollense population has close affinities to hunter-gatherers but is otherwise similar to contemporary populations, like the Unetice. The close affinity to the hunter-gatherers is further emphasised by the D-statistics with the grouped Welzin individuals, that only showed significant values for shared ancestry with SHG or WHG even if tested with any of the late Neolithic or Bronze Age populations. Thus not allowing to separate the Unetice, Bell Beaker or Corded Ware populations, according to their shared ancestry with the Welzin group.
Because of the PCA results and the fact that the D-statistics of the form D(WEZ,*test*;WEZ,Mbuti) only resulted in significant positive D values if *test* was an ancient population, the Welzin individuals were grouped together with the exceptions of the samples WEZ16, WEZ54, WEZ56 and WEZ57 and the contaminated samples WEZ63 and WEZ77 that were excluded from all analysis. This grouping was mainly based on the PCA results, since the D-statistics results were no different for the four outliers, but can additionally be supported by the individual f3 and ADMIXTURE results and in case of WEZ16 also based on archaeological evidence. Individual WEZ16 is the only one that was found in a burial context and is directly dated to the Neolithic, about $\sim$ 2000 years earlier than the other Welzin individuals used in this study. In all ADMIXTURE analysis WEZ 56 is the individual with the most hunter-gatherer ancestry and WEZ16 the individual with the most Neolithic ancestry followed by WEZ54 and WEZ57. The individual outgroup f3 results have all different rankings and especially WEZ16 and WEZ54 rank the LBK in the ancient and the Sardinians and Basque in the modern populations higher than the grouped f3. Also indicating a higher affinity towards a Neolithic population as suggested by their PCA positioning.

### 4.4.2 Data quality

The pseudodiploid variant calls with the allele PHRED scaled likelihood of >15 are used here because of the PCA from the previous chapter 3.3.3, with the general idea to start with as much information as possible. Additionally this way of calling a genotype by just doubling the most likely allele is an often used practice [5, 15, 36, 40, 63, 75] in ancient DNA, which is motivated by the limited amount of ancient individuals, bad preservation and thus a low coverage of the available genomes.
As can be seen in Table A.41 the filtering for linkage between the markers only left eight of the 21 Welzin samples with > 50,000 positions, which is the suggested minimum for dividing populations on a continental scale according to the ADMIXTURE manual [4]. It has been shown that ADMIXTURE on ancient DNA can be used with ∼ 10,000-20,000 positions [5], which is reached by all Welzin samples after LD filtering is applied on the full data set including all available ancient and selected modern populations used in the small ADMIXTURE analysis. The multi-modality of the unsupervised ADMIXTURE analysis or the large standard errors in the outgroup f3 can be interpreted in favor or against the low filter criteria and can have several reasons.
There is little support for the most likely modes in ADMIXTURE, 35 runs of 100 with an average pairwise similarity of 0.923 in the larger first analysis and 2 of 100 runs with a similarity of 0.902 in the smaller second data set. While the larger data set has less variants, ∼ 50% of the variants from the small second ADMIXITURE analysis (compare Table A.41), it still has the better support for the most likely mode. While this indicates that less positions as a result of higher filter criteria might be sufficient it also means that the WEZ individuals introduce uncertainty to the ADMIXTURE analysis. This uncertainty can be due to bad data quality, limited data or an actual highly admixed population that is represented by the Welzin individuals, and is visual in the multi-modality of the unsupervised ADMIXTURE analysis. That the samples analyzed here are the main reason for the multi-modality, can be seen in the variable ancestry components in Welzin individuals between all modes (see Figure A.49 and supplementary file pong_k2_small_all_modes.pdf) and the fact that the small data set consists of only 22 other, non WEZ individuals, whereas the big data set includes 80 non WEZ individuals. In the small data set the 19 Welzin individuals account for ∼46% of the data set and thus could reduce the pairwise similarity by ∼46%, if they had no ancestry with any available group in the data set and the ADMIXTURE plot would generate random patterns. The combination of the two previous arguments, first, the Welzin individuals are variable in the ADMIXTURE results and second the WEZ individuals are a much larger proportion of the less certain small data set than the large data set, allows the following assumption: A higher number of individuals with a secure ancestry reduces the number of modes in the unsupervised ADMIXTURE analysis as a whole, but not in the Welzin individuals, thus leaving their ancestry components unclear. Since the variant calls from the ancient individuals used as reference [5, 75] are called with similar criteria by just doubling the single most likely allele at a position to generate a pseudo diploid genotype, I tend to argue towards a limited amount of data or even an admixed population as reasons for the results rather than bad data quality of the positions. Especially under consideration of Table 4.19, which shows that while somehow reaching the minimum requirements suggested for ADMIXTURE [4, 5] the Welzin individuals are missing 0.95% ±0.03 of the reference positions. It has to be pointed out that both ADMIXTURE runs indicate the four outliers. Thus supporting the PCA results, by showing that WEZ56 has more hunter-gatherers ancestry than any other individual form the Tollense valley, and WEZ16, WEZ54 and WEZ57 have more Neolithic ancestry than any other.
For the outgroup f3 analysis all results showed significant values (Z-score >3), but the standard error of the f3 was on average higher than the difference between to consecutive values with an average of 0.0021 ±0.0024 and an average standard error of 0.0037 ± 0.0007 in the grouped analysis with the modern populations and similar values for the comparison to the ancient populations. By Comparing the outgroup f3 results of the Welzin group with the PCA (see Figure A.48) it becomes visible that even modern populations might be difficult to

separate using outgroup f3 with this amount of data. Even more so for the widespread WEZ individuals in comparison to a modern population like the Polish on the PCA. Although the ranking according to shared drift in the outgroup f3 makes sense, the high errors make it impossible to distinguish individuals based on the ranking of there f3 values. The outliers WEZ16, WEZ54, WEZ57 selected via PCA statistics could not be shown to have more shared ancestry with other populations than Welzin using D-statistics.
Although the analyses are based on very few loci and should therefore interpreted with caution they are consistent and seem to support each other. PCA, f3 and ADMIXTURE identifies outliers, while ADMIXTURE, D-statistics and PCA support the grouping of the individuals.

## 4.5 Conclusion

As mentioned in the previous Chapter 3.1 this capture was designed to be used with coalescence based methods. The results generated for this study were used to see how and if this data can be used comparable to shotgun sequencing data without the use of coalescence based methods. It has to be pointed out that the f3 and D-statistics [86] as well as ADMIXTURE [3], were developed for the use with shotgun sequencing data or large SNP arrays and that both references used here are based on such arrays [16, 37, 63].
With regards to the quality of the results and the amount of data available in the capture, that is comparable to the ancient reference [63], the capture seems to be close to the lower limit of data that can be used. Considering the archaeological question if different parties were involved in the Tollense battle or if a population structure in the Welzin samples can be uncovered, the results suggest that the data is not sufficient for the methods and references used here to uncover such fine scale differences. Nevertheless the data generated with this capture could show that the Welzin individuals are similar to other contemporary populations, like individuals from the Unetice culture, with a stronger affinity to hunter-gatherer populations than to the contemporary populations. No population structure could be found between the Welzin individuals. Four outliers were identified and the sum of the results points in the same direction: Three of the four outliers WEZ16, WEZ54 and WEZ57 show closer affinities to neolithic populations than the grouped Welzin individuals and WEZ56 shows closer affinities to hunter-gatherers. Any interpretation regarding possible parties that might have been involved in the conflict in the Tollense valley $\sim$ 3200 ago can only be speculative with regards to the here shown data.
With the resolution given here, an educated guess for different involved parties could be, that both parties were relatively local and more closely related than any ancient DNA study was able to separate so far. Maybe similar to people from Hessen versus people from Rhineland-Palatinate in modern Germany.

# 5 On lane contamination

Here Agilent's SureSelect in solution capture was used with different protocols, on 33 human archaeological samples of different age and state of preservation. To make economical use of this approach, samples can be provided with specific indices and sequenced in parallel. To allow for simultaneous sequencing of all samples using six lanes on the Illumina Hiseq, each sample was tagged with one specific index. To overcome known difficulties with false index assignments of reads on the same lane double indexing of a library is theoretically available [57]. Despite the general possibility, not all sequencing centers had their machines adjusted at the time of this work. Further Li & Stoneking 2012 [67] showed that although double indexing was used, the problem of false index assignment persists for low level mutations and that cross contamination during simultaneous handling of multiple samples is also an issue. Results obtained for the here presented study show that it is possible to identify and remove mis-assigned sequences as well as possible cross contamination and still be able to perform reliable downstream analysis. The method presented below allows a variant call with no falsely called variants and only a loss of 2.47% of correct called SNPs whereas alignments unprocessed with the shown method, display a false negative rate of 35.8% of all known variants and a false positive rate of 38.64% of all called SNPs.The following chapter will explain lab protocols resulting in this "On lane Contamination" and the following bio-informatic processing that can overcome this issue.

## 5.1 Samples

33 different samples were used in this study whereof five were sequenced twice on separate lanes. Sample age varies between 8800 cal. BC and 475 cal. AD; sampling sites are in Iran, Greece, Central Germany and Russia. Samples were either found in caves or at open air sites. For 28 of the 33 samples processed in this analysis, the sequence of a partial HVR 1 (16013-16409) was available through other projects done in this Institute. All samples are unpublished, therefore no connection between specific samples and their sites, age and precise sequence will be made. Altogether 81 SNPs are known in 32 samples including those sequenced twice which could be used for further analysis.

## 5.2 Methods

### 5.2.1 Lab methods

***DNA extraction***

DNA extractions and amplifications were performed in a specific ancient DNA laboratory and followed the guidelines described previously by Bramanti et *al.* 2009 [13].

***Library preparation***

Libraries for all samples were built in specific ancient DNA laboratory from 50$\mu$l of unshared DNA extract mainly following the protocol developed Meyer *et al.* 2010 [79] with given modifications. All index primer sequences were designed and ordered without the additional base T at 3'-ends (Biospring GmbH) to overcome index primer dimer structures during amplification. Extracts are first purified after adapter ligation to avoid loss of DNA molecules due to small fragment size in combination with cut-off values of the Qiagen MinElute Purification Kit and purification procedure itself. End-repair reaction mix was made up as follows:

- 7.05$\mu$l T4 Ligase Buffer (10x)
- 0.7$\mu$l dNTPs (10mM each)

- 0.35$\mu$l BSA (20mg/ml)
- 3.5$\mu$l T4 PNK (10U/$\mu$l)
- 1.4$\mu$l T4 DNA Pol (5U/$\mu$l)
- 7.05$\mu$l UV-HPLC-$H_2O$

After incubation for 15 min at 25 °C and 5 min at 12 °C, T4 DNA polymerase was inactivated at 75 °C for 10 minutes. The reaction was cooled down to room temperature to allow renaturation of short DNA molecules. Subsequent reaction volume for adapter ligation was increased to 100$\mu$l adding:

- 5$\mu$l ATP (10mM)
- 10$\mu$l PEG-4000 (50%)
- 1$\mu$l Adapter-Mix (each 100$\mu$M)
- 2.5$\mu$l T4 DNA Ligase (5U/$\mu$l)
- 11.5$\mu$l, UV-HPLC-H2O

The reaction was incubated at 22 °C for at least 30 min followed by a clean-up step with Qiagen MinElute Purification Kit. Adapter sequence fill-in reaction was performed as described by Meyer *et al.* 2010 [79]. Libraries were amplified by performing 20 cycles of PCR. PCR setup comprised:

- 0.5$\mu$l AmpliTaq Gold DNA polymerase
- 5$\mu$l GeneAmp 10 X PCR Gold Buffer
- 5$\mu$l $MgCl_2$ solution (25mM)
- 1$\mu$l dNTP Mix (10mM each)
- 1$\mu$l BSA (20mg/ml)
- Is4 Primer (10$\mu$M)
- Indexing Primer (10$\mu$M)
- 25.5$\mu$l Uv-HPLC-H2O

For the amplification of libraries without adding full adapter-sequence before capture enrichment reaction Primer Is5 (10$\mu$M) and Is6 (10$\mu$M) were used instead of Is4 and Indexing Primer. Indexed or only amplified libraries were purified and eluted in 50$\mu$l elution buffer. Blank controls were carried out for every library preparation and DNA amplification to control the reagents in every step and check for cross-contamination during hands-on time. We created an artificial Library-Control (LC) to guarantee full efficiency of all used enzymes during preparation. The LC sequence is assembled from a 40bp nonsense sequence DNA region [100] and Roche Primer Sequences A and B on both ends of the nucleotide (5' CCATCTCATCCCTGCGTGTCAACT-GACTAAACTAGGTGCCACGTCGTGAAAGTCTGACAACCTATCCCCTGTGTGCCTTG 3'). Based on the extension of fragments by ligation of adapters, a shift in length could act as a positive control. Amplified Libraries, all blank controls and diluted Library-Control (LC) were visualized and screened using an Agilent Bioanalyzer DNA High Sensitivity chip.

***mtDNA Capture enrichment***

The enrichment of the complete human mitochondrial genome was performed by in-solution hybridization with the Sure Select Target enrichment XT kit (Agilent). One reaction contains an oligo bait library consisting of approximately 55k biotinylated RNA baits matching a specific target region [32]. The oligo bait library was custom designed by Agilent's web-based software eArray with 4x tiling and the Cambridge reference sequence [6] as target sequence. Since hybridization sensitivity for ancient DNA molecules to RNA baits is unknown, due to short fragment size and deamination, the most common mtDNA mutations were introduced into the bait library. Furthermore regions with lower GC content were added separately to counteract a GC bias during hybridization reaction. Due to the fact that the mitochondrial control region carries the most genetic variation, complementary baits in this region (position: 16021-450) were increased to gain a higher coverage after sequencing. In sum 87% of the oligo bait library cover the whole mtDNA genome with a theoretical 368-fold sequencing depth of each base, 10% contain additional baits with low GC content and 3% of the baits are covering the HVR 1.

***Hybridization***

All post-library-amplification experiments and hybridization reactions were carried out in a separate laboratory that is exclusively used for NGS experiments. Working steps were performed in UV radiated boxes. To avoid cross-contamination between indexed or non-indexed library molecules all articles of daily use were incubated with bleach and UV irradiated after usage. SureSelect In-solution target capture enrichment was performed according to manufacturer's protocol except following modifications:
Due to our tiling design of mitochondrial bait library the recommended volume could be 4.8x diluted to achieve a theoretical read depth of nearly 76 fold. For libraries which were already indexed the Agilent SureSelectIndexing Block #3 was replaced by our index sequence specific self-designed Blocking-Oligo-Mix containing BO1.P5.F, BO2.P5.R, Ind*Block_F, Ind*Block_R (100$\mu$M each), whereby asterisk symbol acts as a placeholder for individual index number. The capture products were amplified in three to four parallels with AmpliTaq Gold DNA Polymerase as described above by using primer pair Is5 and Is6. To add an individual index sequence for samples captured without an index, different primer systems (Is4 and Index-Primer) had to be chosen. Reamplification of some of the capture products was performed using Herculase DNA Polymerase, 5x Herculase Buffer, DMSO, Primer Is5 and Is6 (100$\mu$M) after initial denaturation for 30 s at 98 °C with following cycling conditions for 8-10 cycles:

- 98 °C 10 s
- 60 °C 30 s
- 72 °C 30 s

Followed by a final elongation at 72 °C for 5 min. DNA was purified using MSB Spin PCRapace from STRATEC Molecular GmbH and eluted in 40-60$\mu$l Elution buffer. Libraries and all blank controls were visualized and screened using an Agilent Bioanalyzer DNA High Sensitivity chip.

***Sequencing***

All samples were sequenced on one flow cell of the Illumina HiSeq2000 at the IMSB of J.-G.-University in Mainz according to the manufacturer's protocol. A 50 bp paired-end (PE) run was sequenced using TruSeq PE cluster Kit v3, flow cell type v3 and TruSeq SBS Kit v3 (200)(Illumina, Inc). Corresponding to measurements of dsDNA HS Assay on a Qubit® 2.0 and on Agilent 2100 BioanalyzerTM with the High Sensitivity DNA Assay 5-7 samples were pooled in equimolar ratio and sequenced in parallel on one single lane. Indices for pooled samples were selected that they differ in at least three positions.

### 5.2.2 Bioinformatics

For sequence analysis the pipeline described in chapter 1.3 was used. Except variations noted in the following section.

#### *Correction for index mis-identification*

After alignment sequences were clustered by 98% sequence identity in base composition and equal sequence length using cd-hit-454 4.5.4 [68]. For each cluster one representing sequence and the number of reads contained in the cluster was obtained. Sequences that stem from a cluster of size one (singletons) were removed. The theoretical amount of sequences that could have resulted in an index mis-identification event from each cluster, was calculated assuming a probability of 0.003 per sequence that was suggested in Kircher *et al.* 2012 [57]. E.g a cluster with 1000 reads would have 3 possible index mis-identifations (1000 x 0.003). All sequence clusters with a probability of 0.1 for such an event were now considered as possible contaminants. For each sample only sequences that defined a cluster were extracted from the first alignment using a perl script and realigned using *bwa* [65] and *samtools* [66] as stated above. For each read in every alignment, the start position in the alignment (POS according to SAM format specifications [66]) and the string for mismatching positions according to the reference (MD tag according to SAM format specifications) were noted. This information was compared between every aligned read and each sequence considered as a possible contaminant, that was sequenced on the same lane and not resulted from the same sample. If a match between sequence and possible contaminant was found and the sequence in the alignment resulted from a cluster that was smaller than 100 times the number of possible index miss identifications, it was removed from the alignment. This method will be referred to as *decon* throughout this work.

#### *Variant detection*

Both alignments, before and after *decon*, of each sample were processed further for SNP calling. Duplicates were removed using *picardtools MarkDuplicates* version 1.56 [2] followed by indel realignment and base recalibration with GATK [76] using a variant data set obtained from 1000Genomes [20]. Read depth and coverage was calculated on the hereby obtained alignments using *samtools* [66]. Variant calls were performed simultaneously for all samples but separately for the two different alignments using the *UnifiedGenotyper* [24] setting sample ploidy to 1 and allowing a minimum base quality of 10. Only variants with at least 4x read depth and a quality of 50 were considered as called. Variants called for each sample located in the partial HVR 1 (16013-16409) were then compared to the known polymorphisms. To quantify the impact of both methods used for generating the alignment on the SNP call, following categories were used:

- Known SNP: A variant that is known through multiple Sanger sequences and considered as correct.
- Missing Information: Positions, which possessed neither the required read depth nor the quality to be considered as called.
- Error: False negative variants that were not called despite a 4x read depth and high quality or false positive variants.

#### *Blast*

Local *blast* [17] searches were carried out for four samples on the high performance computing cluster at the Johannes Gutenberg University Mainz. The data used for the *blast* resulted from libraries and capture experiments, that were conducted temporarily independent from each other. For each sample, all reads not aligning to the target region were identified, duplicates of those were removed using cd-hit-454 [68] with the settings stated above, for the identification of contaminants; the defining sequence for each of the generated

clusters was extracted and information of the copy number kept. Two different searches were executed using *megablast* from *blast* version 2.2.25, setting identity percentage cut-off to 80. For the first search all sequences off each sample that did not align to the rCRS were *blasted* against the NCBI nucleotide database and for the second, a data base for *blast* was generated from all non-matched reads of each sample. The LC, IS1 and IS2 were blasted against the latter databases using *megablast*. Sequences were considered as a hit if the blast identity threshold (bit) score was 10% below the maximum score achievable. The maximum of the achievable score was set to two times the smaller sequence length of either database sequence hit or the blasted sequence. Results from the search against the NCBI nucleotide database [93](accessed 20.09.2012) were analyzed using MEGAN [42] to detect different taxa present. LCA-Parameters were set to MinScore = 80, TopPercent = 10, WinScore = 100 and MinSupport = 5 to assign taxa to reads. All reads belonging to the Hominidae family were extracted and aligned to the human reference genome using *bwa* with default settings. After alignment duplicate removal was done once more using *picartools*.

***DNA sequence authenticity***

To check for typical damage patterns in aDNA, *mapdamage* version 0.3 [31] was used separately on all samples for both alignments. Since fragmentation pattern alone cannot be used as authenticity criteria for endogenous ancient DNA [30, 58] and cannot be fully determined in a 50bp read, a known partial HVR 1 sequence for each sample was used for analysis. Those fragments were independently generated with at least three classical PCRs and Sanger sequencing from different extracts. All known sequences were reproduced in our ancient DNA laboratory following guidelines for aDNA [13].

## 5.3 Results

### 5.3.1 Capture efficiency

In total ∼43Gb (Gigabases) were generated for this dataset. ∼5Gb (11.6%) were removed during quality filter and adapter trimming. A total of ∼9Gb (20.8%) could be aligned to the target region. After duplicates were removed 12Mb (Megabases) (0.13%) of all aligned reads remain in the non *decon* alignments and ∼4.5Mb (0.05%) in the *decon* alignments. On target percentage varied between 80.24% and 0.05% of all reads per sample after adapter trimming and quality filter with an average of 22.81% (±26.40). For all samples an average genome coverage of 92.47% (±14.53) was reached before decontamination, with the minimum being at 33.19% and maximum 100%. Applying *decon* reduces the average for genome coverage to 70.98% (±31.13) with the minimum being at 7.79% and the maximum 100%. Seven samples of 33 resulted in whole mitochondrial genomes, whether *decon* was applied or not. Another six samples resulted in a coverage >98% for the targeted region with at least 5x per base read depth independent from the method used. (For details of all samples see Table A.42). Influences of different protocols used during lab work are shown in Table 5.21. Significant differences occur only between capturing with index and without index. Adding the index after capture increased the on target ratio in average by ∼21% (p-value=0.021) and the amount of genome that was covered after decontamination by ∼20% (p-value=0.0465).

**Table 5.21:** *Average Values for the data of the whole Flow Cell and groups within all samples according to different lab protocols used; Samples occur in two protocols either in Index or No index and in Reamplified or Not reamplified; Index = capture of indexed library; No index = capture with non indexed library; Reamplified =additional PCR of capture products; Not reamplified = only one round of PCR after capture; On Target = Percentage of reads from total aligned to the mtDNA; Genome covered no decon = Percentage of Genome covered before decontamination; Genome covered decon = Percentage of mtDNA covered after performing decontamination.*

| Protocol | Nr. of samples | On target [%] | Genome covered no decon [%] | Genome covered decon [%] |
|---|---|---|---|---|
| Flow Cell | 38 | 22.81 ± 26.40 | 92.37 ±14.53 | 70.98 ± 31.13 |
| Index | 22 | 13.86 ± 22.61 | 88.56 ± 17.73 | 62.47 ± 33.46 |
| No Index | 16 | 35.12 ± 26.86 | 97.60 ± 5.46 | 82.68 ± 23.89 |
| Reamplified | 19 | 17.01 ± 22.14 | 89.93 ± 19.23 | 70.82 ± 31.04 |
| Not reamplified | 19 | 28.6 ± 29.49 | 94.79 ± 7.18 | 71.13 ± 32.05 |

The variance in the average on target ratio 27.92% (±26.40) is 872.35. Without *decon* being applied to the alignments the variance for the percentage of genome covered is 211.11 at an average of 92.37% (±14.53). If *decon* is applied to the alignments, variance in the fraction of genome covered between samples is 969.22 at an average of 70.98% (±31.13). As also displayed in Figure 5.31 variance in samples is smaller during data processing without *decon* being applied. Figure 5.31 shows that samples from lane 5 seem to have a major impact on the variance of the average fraction of genome covered per sample. Removing lane 5 from this data-set results in an average on target percentage of 30.5% (±30.29) and a variance of 917.7. The corresponding values for the percentage of genome covered are 96.34% (±6.02) with a variance of 36.2 for the alignments before *decon* and 74.97% (±27.33) with a variance of 751.14 for the ones *decon* was applied. The only significant difference between lanes is found between fraction of genome covered in the non *decon* alignments between lane 5 and lane 6 (p-value = 0.03). Correlating fractions of genome covered after processing (y/dependent variable) and the percentage of on target sequences (x/independent variable) for each sample shows that there is a higher correlation when the *decon* alignments are chosen. Regression analysis including the *decon* alignments shows that the on target ratio in each sample can explain 37.2% (corrected R2 = 35.55%) of the variation observed in genome coverage (standard error = 24.99; slope = 0.63; intercept = 53.01; F-Value = 21.43; critical-f = 4.6310-5; Figure A.52). Whereas on target ratio can only explain 9.66% (corrected R2 = 7.15) using the alignments without *decon* (standard error = 14; slope = 0.153; intercept = 88.09; F-Value = 3.84; critical-f = 0.06; Figure A.53).

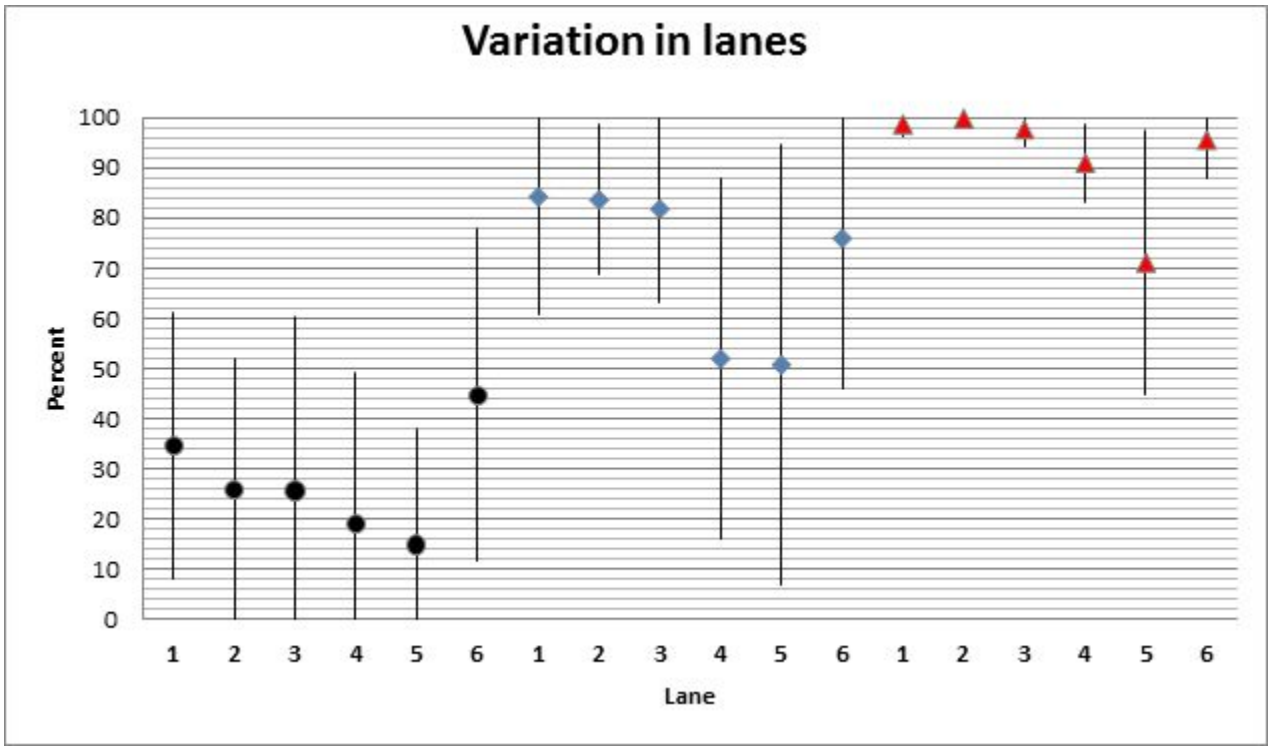



**Figure 5.31:** *Showing average variation per lane including ± standard deviation (black lines); Results are sorted by category; black circle = On target as a fraction of all reads after quality filtering and adapter trimming ; blue diamond = Fraction of genome covered by decon alignments; red triangle = Fraction of genome covered by non decon alignments.*

**Table 5.22:** *Showing corresponding values for Figure 5.31.*

| Lane | Nr. of samples | On target[%] | Genome covered non decon [%] | Genome covered decon[%] |
|---|---|---|---|---|
| Flow Cell | 38 | 27,92 ± 29,54 | 92.37 ± 14.53 | 70.98 ± 31.13 |
| 1 | 5 | 34.57 ± 26.46 | 98.66 ± 2.23 | 84.21 ± 23.45 |
| 2 | 7 | 25.94 ± 25.99 | 99.79 ± 0.29 | 83.61 ± 14.96 |
| 3 | 6 | 27.21 ± 36.46 | 97.61 ± 3.22 | 81.78 ± 18.83 |
| 4 | 7 | 20.17 ± 31.45 | 90.98 ± 7.69 | 51.98 ± 35.83 |
| 5 | 6 | 14.94 ± 22.93 | 71.16 ± 26.39 | 50.78 ± 43.96 |
| 6 | 7 | 44.64 ± 33.1 | 95.54 ± 7.84 | 75.96 ± 29.82 |

### 5.3.2 Correction for index mis-identification

For aligned reads, including duplicates of all samples, a total of ∼442Mb (4.95%) was removed during *decon*. 3.61% of all aligned reads were removed because of being singletons and 1.33% had a copy number higher than one and were removed due to sequence similarities in other samples. Sample averages are 14.93% (±21.11) for total removed reads, 5.78% (±6.78) for removed singletons and 8.86% (±16.1) removed reads due to sequence similarities with possible contaminants. No significant differences exist between all samples captured with index and without index for any of the three ratios. This also holds true for comparison of lane 5 and 6. Looking at the aligned reads after duplicate removal, *decon* removed ∼7.5 Mb (64.1%) from the not decontaminated alignments. ∼4.5 Mb (60.5%) of all hereby removed reads were singletons.

### 5.3.3 SNP calling

Variant calling in all 32 samples on the known partial HVR 1 for the *decon* data sets shows no error and allows 25 called SNPs. Meaning 30.85% of all 81 known variants in all samples, after applying the filter for the *decon* alignments were called. If compared to all SNPs called this means that 100% of called SNPs are correct. Whereas SNP calling in the non *decon* alignments for all samples generates 46 errors and 27 correct variant calls corresponding to 33.32% of the known variants and only accounting for 61.35% of all SNPs called. This means that 38.65% of all called SNPs in the non *decon* alignments are false positives.
The amount of known positions containing not sufficient information for a SNP call is 2,24x times higher in the *decon* alignments with 56 or 69.15% compared to unprocessed ones with 25 or 30.85% of 81 known SNPs (see Figure 5.24). Changing filter parameters to higher values reduces the power to detect real variants in both methods but also reduces errors in the non *decon* alignments (see Table 5.23). Allowing a minimum read depth of ten and a minimum SNP quality of 80 removes 16% off the correct called SNPs in the *decon* alignments. In the unprocessed alignments 11.11% of the right calls are removed and also 23% of the errors. Discrepancies occur because erroneous calls had a quality below 80 but read depth above ten.

**Table 5.23:** *Showing results of SNP calling for different filter criteria summarized in categories (as described in methods) Alignments used: filter criteria; read depth = either 4x or 10x read depth of position to allow SNP call; qual = PHRED scaled error rate for SNP either 50 or 80; Categories are described in methods*

| Alignment used | decon (32) | | non decon (32) | |
|---|---|---|---|---|
| Filter criteria | read depth 4 qual 50 | read depth 10 qual 80 | read depth 4 qual 50 | read depth 10 qual 80 |
| Known | 81 | 81 | 81 | 81 |
| Right calls | 25 | 21 | 27 | 24 |
| Missing Information | 56 | 60 | 25 | 35 |
| Error | 0 | 0 | 46 | 36 |

**Table 5.24:** *Showing Results of SNP calling for 4x read depth and quality 80 with errors divided into false negatives and false positives*

| Alignment used | decon(32) | non decon(32) |
|---|---|---|
| Known | 81 | 81 |
| Right calls | 25 | 27 |
| Missing Information | 56 | 25 |
| False Negative | 0 | 29 |
| False Positive | 0 | 17 |

Reducing both data sets to samples with at least a 4x read depth removes 19 samples from the *decon* data set and four from the unprocessed ones. Figure 5.32 shows that in the *decon* data set 96% of all called variants belong to the 13 samples with at least 4x read depth. Applying the criteria of 4x read depth also increases the missing information to 25% of all known variants. The removal of four samples in the non *decon* alignments reduces the error rate to 34.16% of all called SNPs and the amount of missing information at known variant

sites to 19.12%. None of the accurately called variants in all samples belonged to one of the removed samples in the unprocessed data. All four removed samples belonged to lane 5. Comparing only the alignments of the 13 samples that have at least fourfold read depth without regards to the method shows an error rate of zero for both, the *decon* and the non *decon* alignments. The comparison further shows that the *decon* alignments have a slightly higher rate of variant detection in relation to the unprocessed ones. With 78.13% of the 32 known SNPs in the 13 samples with >4x compared to 75% in the *decon* alignments (see Figure A.54).

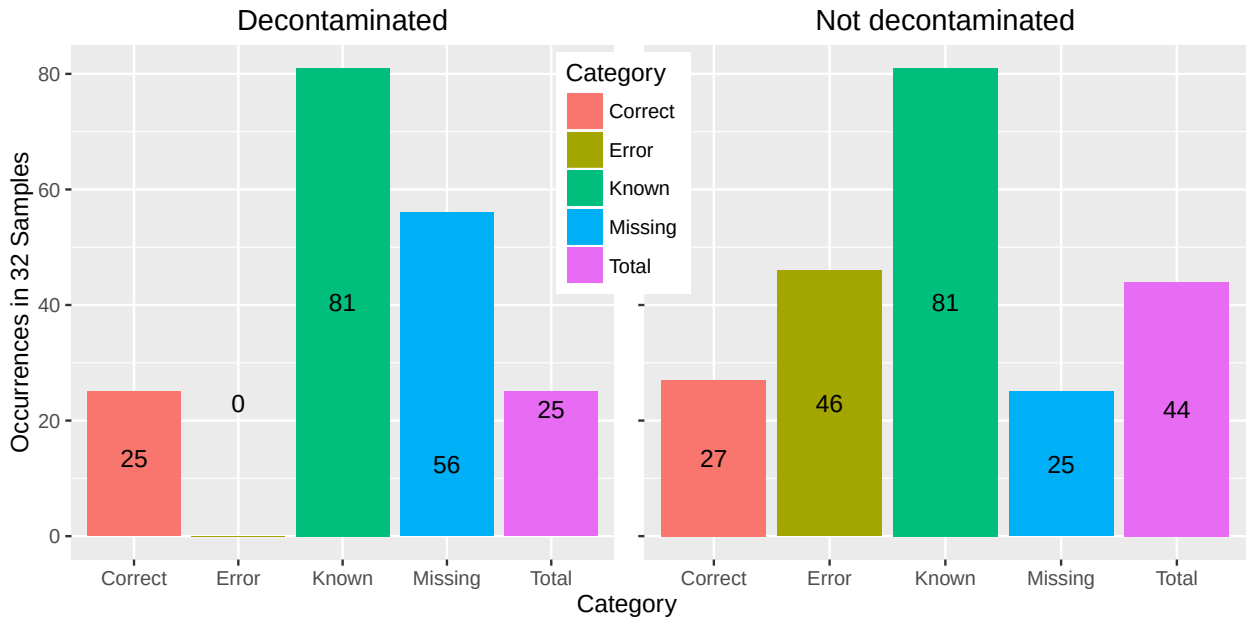



**Figure 5.32:** *Showing results of SNP calling for the two different alignments, summarized in categories as described in methods for 32 Samples; number of occurrences is shown inside the bars.*

**Table 5.25:** *General values for samples shown in Figure 5.33 and 5.35; On Target = Percentage of reads from total aligned to the mtDNA; nodecon/decon reads = Total amount of reads aligned before/after decontamination; Genome covered no decon = Percentage of Genome covered before decontamination; Genome covered decon = Percentage of mtDNA covered after performing decontamination;*

| Sample | on target [%] | no decon reads | non decon genome covered [%] | decon reads | decon genome covered [%] |
|---|---|---|---|---|---|
| BLA8 | 0.21 | 3694 | 99.84 | 427 | 65.46 |
| Krk4 | 0.24 | 5288 | 99.90 | 590 | 79.68 |
| HV1 | 10.35 | 4119 | 99.99 | 1138 | 95.20 |
| BLA17 | 21.48 | 2319 | 99.21 | 398 | 66.26 |
| ZV317b_XXII | 27.21 | 2884 | 99.67 | 593 | 78.65 |
| MIN4_XXVI | 60.50 | 18035 | 100 | 12620 | 100 |
| WC1 | 61.68 | 19183 | 100 | 13263 | 100 |

Illustrated in Figure 5.33 - 5.35 is the influence of samples with relative high amounts of reads at a certain position, or stronger samples, on weaker samples with a small amount of reads at the same position. All samples shown were sequenced on one lane and except ZV317b_XXII captured with an index. True or known variable positions from one sample (WC1) are the only ones called in five other samples (see Figure 5.33). In addition all known variants of those five samples are falsely typed as the reference allele in the capture experiment. After decontamination nearly all of the sequence information in the five samples is removed (see Figure 5.35). Table 5.25 shows, that the five weaker samples have less than 50% of the on target ratio than Min4 and WC1, and only have between 12.1-29.32% of the reads after duplicate removal aligned to the mitochondrial genome compared to the two samples.

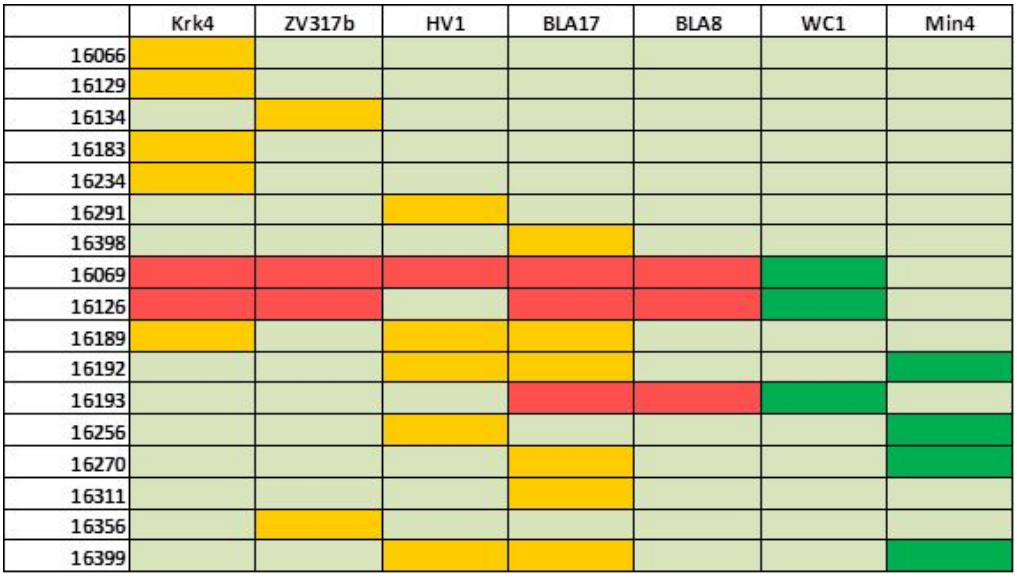



**Figure 5.33:** *Showing samples on Lane 2 with all variable positions known and called; Lines = Positions; Columns = Samples*

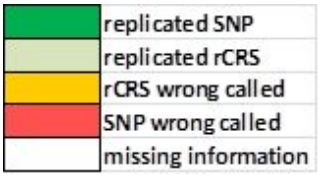



**Figure 5.34:** *Showing colour distribution for Figure 5.33 and Figure 5.35*


**Figure 5.35:** *Showing samples on Lane 2 with all variable positions known and called; Lines = Positions; Columns = Samples*

### 5.3.4 Blast

The *blast* search for the four samples, BLA14, MIN4, ZV122BXXXV and ZV122BXXVIII, against the NCBI nucleotide database (20.09.2012) with a total of 18275734 sequences, resulted in 5.99% of sequences with a hit in the database. Of those, 2.59% could not be assigned to any taxa using MEGAN. Of the 598610 reads that could be assigned, 19.95% were allocated to bacteria, 0.06% to archea, 79.3% to eukaryota with 66.2% belonging to hominidae and 0.69% were viruses, which almost exclusively represented phiX, which is added as a standard during sequencing. A total of 311711 reads corresponding to 78.65% of all reads being assigned to the hominidae could be aligned to the human reference genome (hg19). These ratios are in concordance with results from *blast* analysis of shotgun samples from literature [54, 55, 97, 98] (see Table 5.26).

**Table 5.26:** *Blast results obtained from human aDNA samples through shotgun sequencing from other studies and BLAST results obtained from not aligned sequences after capture of the four samples analyzed here (highlighted in grey); values are given in percent / assigned reads if not stated otherwise; data supplemented from Khaira et al. 2013*

| Sample | Eukaryota | Bacteria | Archaea | Viruses | Assigned of total | Reference |
|---|---|---|---|---|---|---|
| S2000 | 5,2 | 93,24 | 0,41 | 0,03 | 2,3 | Khairat et al. 2013 |
| Mummy 4 | 6,86 | 92,7 | 0,01 | 0,05 | 12,5 | Khairat et al. 2013 |
| Mummy 1b | 9,84 | 89,81 | 0 | 0,06 | 11,8 | Khairat et al. 2013 |
| Mummy 2b | 15,16 | 84,39 | 0 | 0,05 | 10,8 | Khairat et al. 2013 |
| BLA14 | 17,21 | 81,33 | 0,06 | 0,03 | 5,39 | This Study |
| Mummy 5 | 19,19 | 72,45 | 0,03 | 0,08 | 13 | Khairat et al. 2013 |
| S1000 | 38,13 | 60 | 0,4 | 0,1 | 1,76 | Khairat et al. 2013 |
| ZV122BXXXV | 63,72 | 35,25 | 0,15 | 0 | 3,61 | This Study |
| ZV122BXXVIII | 66,09 | 30,51 | 0,27 | 0,04 | 4,13 | This Study |
| Mummy 3 | 83,16 | 16,2 | 0 | 0,29 | 0,33 | Khairat et al. 2013 |
| Mummy 2a | 88,27 | 10,61 | 0,05 | 0,41 | 0,19 | Khairat et al. 2013 |
| Denisova SL3003 | 89,6 | 0,69 | 0,02 | 9,57 | 26,3 | Reich et al. 2010 |
| Mummy 1a | 96,03 | 3,6 | 0 | 0,16 | 2,55 | Khairat et al. 2013 |
| Iceman | 98,79 | 0,79 | 0 | 0 | 10,84 | Keller et al. 2012 |
| Min4 | 99,14 | 0,26 | 0 | 0 | 10,74 | This Study |
| Saqqaq | 99,7 | 0,07 | 0 | 0 | 74,72 | Rasmussen et al. 2010 |

Using the same reads that were *blasted* to build a *blast* databases, and *blasting* the LC as well as IS 1 and IS 2 against those self build databases, resulted in a total of 135 hits for the LC corresponding to $1.52 * 10^{-4}$% of all reads in all databases of the four samples analyzed (see Table 5.27). For IS 1, 89 hits were obtained and 21 reads matched IS 2. Representing a fraction of $4.87 * 10^{-4}$% of all reads in the databases for the first and $1.15 * 10^{-}4$%for the latter. As the Library Control (LC) consists of multiple copies of a single molecule, it cannot be differentiated as being amplified and stemming from a single molecule or stemming from multiple different molecules, the total copy number of each LC hit in the databases was summed for the total number of hits. Therefore it must be related to all reads generated for a sample after adapter trimming and quality filtering to quantify cross contamination during library creation.

**Table 5.27:** *Showing per sample results of the blast search against the databases generated from all not aligned reads for each sample; Protocol = shows protocol differences used with the sample; LC = reads that matched LC; IS 1 or 2 = reads that matched IS 1 or 2; Of Database = showing percentage of sequence found in Database of sample; Size = Average copy number of each sequence in Blasted or in aligned reads of not decontaminated alignments; Time Lib/Cap = shows time difference between libraries/captures*

| Sample | Protocoll | On target [%] | LC | LC Database [%] | IS 1 | IS1 Database [%] | IS 2 | IS2 Database [%] | Size blasted | Size aligned | Time Lib. | Time Cap. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MIn4 | Index | 60.50 | 0 | 0.00E+00 | 2 | 4.27E-05 | 0 | 0.00E+00 | 1.39 | 304.85 | 0 | 0 |
| ZV122 BXXVIII | no Index & reamplified | 39.23 | 88 | 2.13E-04 | 61 | 1.95E-03 | 1 | 3.19E-05 | 4.83 | 91.76 | 1 M | 0 |
| ZV122 BXXXV | no Index | 4.36 | 38 | 4.31E-04 | 25 | 3.85E-04 | 16 | 2.47E-04 | 1.39 | 1033.25 | 1 M | 1 M |
| BLA14 | Index | 21.34 | 9 | 4.36E-05 | 1 | 2.52E-05 | 4 | 1.01E-04 | 3.82 | 546.50 | 3 M | 1 M 1D |

Detailed results for each sample show that both ZV122 samples, both captured without index, contain higher amounts for all artificial sequences LC, IS 1 and IS 2. Whereas the reamplified sample ZV122BXXVIII has the highest amount of hits for the LC and IS 2. Copy numbers for aligned reads are at least 20.47 times higher than for all blasted reads (see Table 5.27).
Figure 5.36 shows damage patterns in Min4 as a representative for all samples. An elevated frequency of C to T transitions (red line) can be found at 5'-end of reads. In general this transition is elevated over the whole read length.

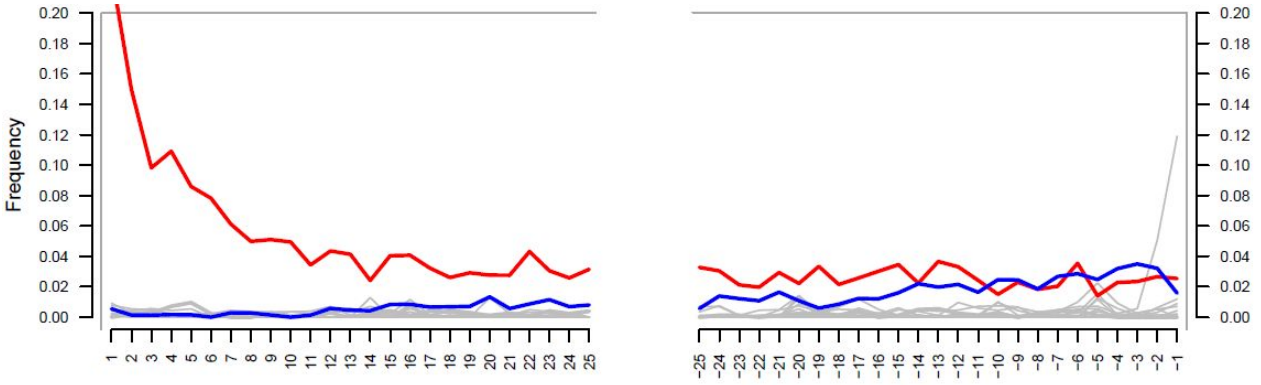



**Figure 5.36:** *Showing Mapdamge results of Min4; x axys = position in read; y axys = transition frequency; red line = C/T transitions; blue line = G/A transitions.*

### 5.3.5 ContaMix

To additionally show the influence of index mis-identification, ContaMix (see Chapter 2.1.5) was applied on both alignments of each sample from lane 2 prior and after *decon*. As can be seen in Table 5.28, the application of the decontamination reduces the sequencing error as well as the estimated amount of contaminating sequences in each sample. Only for HV1 and Krk4 the sequencing error estimate does not change.

**Table 5.28:** *Results of ContaMix for each of the samples before and after decon is applied; Alignment = shows if Decon or No Decon was applied; Error = shows estimated sequencing error probability by ContaMix, lower = shows 1 quartile of reads uniquely mapped to that sample; upper = shows 3rd quartile of reads uniquely mapped to that sample*

| Sample | Alignment | Error | Lower | Upper |
|---|---|---|---|---|
| BLA17 | Decon | 0.004 | 0.94 | 1.00 |
| BLA17 | No Decon | 0.007 | 0.74 | 0.94 |
| BLA8 | Decon | 0.006 | 0.95 | 1.00 |
| BLA8 | No Decon | 0.007 | 0.72 | 0.89 |
| HV1 | Decon | 0.008 | 0.76 | 0.93 |
| HV1 | No Decon | 0.008 | 0.69 | 0.90 |
| Krk4 | Decon | 0.007 | 0.87 | 0.98 |
| Krk4 | No Decon | 0.007 | 0.74 | 0.89 |
| MIN4 | Decon | 0.013 | 0.94 | 0.98 |
| MIN4 | No Decon | 0.015 | 0.93 | 0.97 |
| WC1 | Decon | 0.007 | 0.99 | 1.00 |
| WC1 | No Decon | 0.012 | 0.95 | 0.97 |
| ZV317b | Decon | 0.006 | 0.81 | 0.96 |
| ZV317b | No Decon | 0.007 | 0.72 | 0.90 |

## 5.4 Discussion

### 5.4.1 Capture efficiency & correction for index mis-identification

The variation in the average on target percentage suggests, that the amount of the captured target region covered should also span a high variance (Figure 5.31 & Table 5.21). A linear regression analysis to correlate those two values for the two different alignments, shows that the correlation of decontaminated alignments with $r^2$=0.373 is 3.85x higher (4.96x for corrected $r^2$) than for the unprocessed with $r^2$=0.116. Assuming the non *decon* alignments are true, this would suggest that only 9.66% of the genome fraction, that is covered, is explained by the capture efficiency. This correlation raises to 37.2% taking the *decon* alignments instead. Even for a small genome like the mtDNA one would expect that whether a sample has 60% or 2% on target ratio, the difference in percentage of genome covered should be more than 10% between the two samples. Those results combined with differences in variance between both alignments, in contrast to on target ratio, suggest that samples homogenize during lab processing or sequencing. Performing all post library-PCR labwork with indexed libraries should rule out cross contamination between simultaneously processed samples during capture. Which in reverse means a risk of cross contamination for samples processed without an index after library PCR. The fact that capture reactions performed with an index have a significant lower on target ratio than the ones performed without an index, and a significant lower amount of genome covered after *decon*, suggests that cross contamination happens and/or that the efficiency of a capture reaction without index is increased. Observing no significant differences in the ratio of removed reads during *decon*, points in direction of higher capture efficiency without using an index. One reason for the influence of lane five on the variance in on average target percentage and genome coverage might be the fact that all samples have been captured with an index. This phenomenon could point to cross contamination during the capture reaction and is emphasized by the actuality that the only significant difference between lane five and other lanes can be observed for lane 6 in the percentage of mtDNA covered before decontamination. In lane 6, all samples have been captured without an index. Another reason for lane five having a high variance in the amount of genome covered before decontamination might be the generally poor quality of all samples on that lane. This quality is represented in the worst average for all analyzed lanes in on target ratio and percentage of genome covered for both alignments, *decon* and non *decon* (see Table 5.22). Poor sample quality will result in less reads which again will result in less miss-identification or cross contamination events arising from that sample. Pooling with a high quality sample for which the contrary is true will greatly increase the ratio of reads identified as being falsely assigned or cross contaminants compared to all aligned reads in the low quality sample.

The total percentage of reads removed by *decon* is 4.95% with, or 1.33% without the singeltones, which is 16,5x or 4.43x higher than the suggested ∼0.3% by Kircher 2011 [57] but at least 3.03x lower than the 10-15% found by Li & Stoneking 2012 [67]. Several reasons could account for that discrepancy. One being the amount of samples pooled which is at least eight times higher in Li & Stoneking 2012, where the result was obtained from 40 to 90 samples pooled with double indices. In this study the authors refer to “chimeric reads” meaning that 10-15% of all reads had index pairs that were not assigned to any of the studied libraries. The general false assignment of indices will be higher than the actual amount of sequences that were falsely assigned to the wrong sample, because two specific indices will have to be falsely paired to actually assign a sequence to another sample. Another factor is that during the application of *decon*, not all reads were compared between samples. To reduce computational time, singletons were removed and only sequences with a copy number high enough to result in at least one mis-identification event per 100 copies were compared. This could result in the overestimation of the influence of index mis-identification or the underestimation of cross contamination between samples with no index. Further Garcia-Garcera *et al.* 2011 indicated that fragmentation patterns in aDNA are not random [30]. This would increase the probability that reads in different samples are identical by chance. Again this would overestimate the amount of sequences originating from simultaneously processed

samples pointing towards the *decon* approach being rather conservative.

### 5.4.2 SNP calling

Whether index mis-identification during sequencing or due to the influence of cross contamination during simultaneously capturing, the comparison with the known variants allowed us to uncover the combined effect of both events on downstream variant analysis. The presented method for decontamination eliminated all errors that occurred in variant calling before complete data processing by a cost of only two correct variants. It could be shown that all wrong calls were actually positions that lacked true information from the sample and just showed carried over sequence information from the simultaneously sequenced samples. Whether a SNP is called at an unknown position or false positive variant can be called, depends on the sequence composition and read depth of the other samples on the same lane (see Figures 5.33 - 5.35). Seven complete mitochondrial genomes in this data set, no matter whether sequences are removed during *decon* or not, suggest that some samples do not contain enough DNA to generate the necessary data needed for a correct SNP call if the here used capture method is applied. This is further fortified if the criterion for considering a SNP call is an overall read depth of 4x for decontaminated samples. Comparing the remaining 13 samples in the *decon* and non *decon* alignments, shows that without errors the vast majority of actual present SNPs can be found in those 13 (96% *decon*; 92.59% non *decon*), suggesting that read depth alone could overcome the index mis-identification.

### 5.4.3 Blast

The Blast results for all reads that could not be identified as being on target against the NCBI nt database are comparable to shotgun sequencing runs from literature that were analyzed in a similar way. Data shown here implies that reads, that are not on target, reflect a random sample of the library that was captured, similar to a shotgun experiment. Average copy numbers of all sequences that could be aligned are at least 20 times higher than the ones that were not on target. This indicates that molecules from the whole library of a sample are "co-captured". Different reasons might be: baits capture unspecific molecules, secondary structures between adapter sequences of two different library molecules or streptavidin beads used for removing not captured molecules from the extract bind unspecific to DNA.

### 5.4.4 DNA sequence authenticity

Damage patterns generated with Mapdamage do not show characteristic damage patterns, expected for aDNA. The rate of G to A transitions should also be elevated towards the last base. Most likely this happens because during adapter trimming a match with adapter sequence of one base pair was sufficient to be cut. Also with a read length of 50bp in a single end sequencing run, the actual end of a DNA fragment will not often be sequenced, since most of the fragments will be longer than 50bp.

Precautions against possible contamination during labwork before capture were strictly taken according to standards known for aDNA. Because of the new methods used, samples with a previously replicated HVR 1 were used to evaluate the confidence of these methods that need to be conducted outside the dedicated aDNA lab. Although no library blank control showed amplifiable molecules the results of the non *decon* alignments and the blast results from the LC against all reads not matching the mtDNA suggest that there is cross contamination. Between an actual cross contamination event during lab work prior to an index ligation and an index mis-identification cannot be distinguished. Assuming cross contamination happens sporadically and foreign molecules introduced into a sample will have a very low quantity, the outcome will appear in the same way for both events. Molecules from mis-identification and cross contamination events result in a low copy number of that molecule in the contaminated sample and should only influence loci with low read depth (see Figures 5.33 - 5.35). The amount of copies should be influenced by the point in time a

cross contamination event happens during the used protocols. If a contamination is introduced during library preparation before any amplification the contaminating sequence could rise to high copy numbers. At this point a high read depth with endogenous molecules after duplicate removal at an affected position can be the only way to counteract or detect a cross contamination. Using the method described here will not only remove seqeunces from index mis-identification but also cross contamination events posterior to library PCR if samples that were processed simultaneously during labwork are also analyzed simultaneously. The amount of cross contamination in library preparation can be quantified to $1.52 * 10^{-4}\%$ according to the LC blast results for all four samples, assuming all reads not matching the target region are a random sample of the captured DNA extract. Taking into account that two of the samples used for a blast database had no index and thus contamination with LC could have happened during capture, the cross contamination for indexed samples goes down to $2.33 * 10^{-5}\%$. The fact that no errors occurred after *decon*, suggests that any cross contamination event that happened during library preparation was removed during data processing or is negligible. The combined effect of barcode mis-identification and cross contamination can be quantified to an average of 14.93% ($\pm$21.11) of all mapped reads, by taking the reads into account that were removed during *decon*. Singletons also result from base calling errors which cannot be distinguished from any other error source with this method. Therefore not taking those into account leaves a lower boundary for the combined effect at 8.86% ($\pm$16.1).

The copy number of each molecule needs to be taken into account during any capture or other experiment that involves more amplification steps than those involved in library preparation. The amount of copies present for any given molecule can give substantial information on its reliability. Further all singletons should be removed from any analysis in such cases. Looking in detail at positions suggests that removing sequencing reads up to a copy number of three is advisable. Positions now classified as missing information still show reads with a copy number below four that would result in an erroneous SNP call if considered.

## 5.5 Conclusion

The aim was to assess the possibility of capturing a broad range of samples in different states of preservation and to evaluate the influence of different protocols used. Sample ages vary between $\sim$ 1600-8900 BC and different burial conditions like permafrost and cave sites from sites in Iran, Greece, Central Germany and Russia. From all different protocols applied, capturing with or without an index or applying a amplification or not. Results suggest that not all samples are appropriate to be enriched from a single library with percentages of reads aligning to the target region from $\sim$ 0.14 - 78.66% in different capture experiments. In general a large variance in on target percentage or capture efficiency between samples can be observed. This surely has various reasons like, different sample quality or state of preservation and lab protocols used during sample processing. To investigate the influence of preservation, samples were grouped according to age, locality, archaeological context or lab protocol and compared in capture efficiency. Only capturing with an index had significant negative influence on the capture efficiency in decreasing it by $\sim$21% in comparison to capturing with an index. Since capture efficiency has an impact on the total amount of reads, the magnitude of read depth and the fraction of genome covered, quality assessment of a sample before sequencing is a crucial step. It is shown that pooling of samples with variable quality can result in erroneous variant calls. A total of 38 samples resulted in seven complete mitochondrial genomes weather a decontamination was applied or not. With minimum read depth of 4x after the application of the *decon* process 13 samples remain that contain 96% of all found non erroneous variants, additionally for those 13 samples there are no false positive variants called with or without *decon*. This implies that coverage alone can overcome the problem of index miss-identification. It is therefore imperative to process such sequencing data with great caution. Variant calling in simultaneously sequenced libraries can be crucially biased towards SNPs that are present in samples with

higher quality. This study suggests that a 4x coverage increase towards the low quality samples is enough to produce erroneous variants from index mis-identification. To find appropriate samples for pooling, quality assessment that will only consider few loci or just the general amount of complete library molecules is not sufficient. Although the general sample quality might be comparable, high discrepancies in one locus between two different simultaneously sequenced samples might occur. This can result in the underrepresented sample having the same variant than the dominant one at such a locus, or if diploid samples are analyzed might change the genotype of the weaker sample from homozygote to heterozygote or vice versa. Before pooling different aDNA samples it is recommended that for each of the samples shotgun runs and the measure of the endogenous content are available as a reliable source for comparison of sample quality. Further it is recommended to take copy numbers of molecules per library generated by qPCR into account. A combination of both methods, the qPCR and the low depth shotgun sequencing as screening, are used in current ancient DNA studies [15, 40].

# Conclusion

This thesis was started to allow access to NGS technologies for a broad range of archaeological samples and generate data that would be ideal for population genetics and be comparable to available modern reference data. While the three chapters **2 Pipeline**, **3 Nuclear capture enrichment approach** and **4 Case study Welzin** were developed from specific questions, Chapter **5 On lane contamination** is based on a special phenomenon that arose in an early stage during the development of the others. In general, all Chapters show that there can be a solution for every challenge that NGS data from ancient samples might provide, with the exception of to little data.

The results from the capture experiments in both chapters **3 Nuclear capture enrichment approach** and **5 On lane contamination** support read depth as the ultimate solution to counteract any bias or error introduced through sequencing or PMD. For one, summary statistics calculated on the nuclear capture with samples having a >10x read depth are the closest to a modern standard. Second, the samples that were most resistant to index mis-identification or cross contamination are those with >4x read depth. The latter is solvable prior to data analysis by carefully selecting and screening the samples of interest as well as utilizing index pairs instead of single indices and has become common practice [15, 40] now. This thesis shows that even with a limited amount of target DNA, as in skeletal remains, there are solutions to counteract the factors of sequencing errors or PMD bioinformatically. With careful choice of method, shotgun or capture, the filtering and recalibration of the sequencing data, methods used on modern NGS sequences as well as modern reference data can be utilized to analyze aDNA sequencing data.

A lot of progress has been made in the fields of aDNA and palaeogenomics, especially in combination with NGS since the start of this pipeline. Which results in alternative pipelines being developed or existing ones being improved. While the described refinement of the alignments from chapter **2 Pipeline** was at the time the most complex available, the improvement and further developed ATLAS [60, 71] was used in Chapter **4 Case study Welzin**. Although the general idea and work flow of the here developed pipeline can be kept, programs and scripts should be reviewed regularly and if necessary exchanged with newer versions or different and better performing ones (see chapter 2.3.3).

While the capture results look promising in terms of summary statistics, a capture reduces the availability of comparable data sets. It could be shown, in the nuclear and the mitochondrial capture, that the sequences that are “co captured” resemble a shotgun experiment (see Chapters 3.3.2 & 5.3.4), and can be used to circumvent the comparability issues that might arise with such a theoretical limit of possible loci (see Chapter 3.3.3). While population genetic analysis can be run on the here generated capture data, Chapter **4 Case study Welzin** shows that the power is limited to large scale questions similar to a continental population structure or separating hunter-gatherer individuals from Neolithic ones. To actually detect finer population structure, like different parties that might have been involved in a possible battle field, or differentiating people from the same country, the amount of loci in the capture should be increased 10 fold with regards to a given reference population. Whereas it still remains to be seen if the used methods f3, D-statistics and ADMIXTURE are suited to solve such fine population structure with low coverage data. Because of the low power, the challenging estimation of X chromosomal contamination and the sex determination, it should be kept in mind to redesign or at least add markers to the existing capture. Ideally a reference data set using the capture could be created. Both would remove the dependency on additional loci that are sequenced by chance. Finally it still remains to be tested how the capture performs and can be utilized with coalescent based methods for inferring population history.

Although index mis-identification can now be solved through double indexing of the DNA library, and several suggestions from the **Chapter 5 On lane contamination** are now common practice, they were not at the time of the performed experiments. This chapter emphasizes the need to carefully control sequence data

for contamination or other artifacts introduced through any wet lab work or sequencing, that need special treatment, during bioinformatical processing.

# Appendix

## A.1 Pipeline

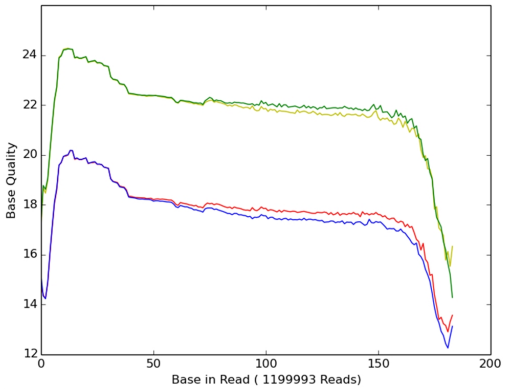



**Figure A.37:** *Average base quality distribution for Pal7 recalibrated with GATK(y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

**Table A.29:** *Shows average base qualities for each base over all positions in all reads for Klei10 and Pal7 after self recalibration step 1; Min and Max are chosen from the averages per position and therefore the std shows variance at that position.*

| Sample | Klei10 | | | Pal7 | | |
|---|---|---|---|---|---|---|
| Base | Average | Min | Max | Average | Min | Max |
| T (red) | 18.97 | 14.35 | 20.80 | 17.87 | 12.67 | 20.16 |
| STD | 2.59 | 0.88 | 3.39 | 2.82 | 0.42 | 3.66 |
| A (blue) | 18.82 | 13.90 | 20.83 | 17.67 | 12.66 | 20.15 |
| STD | 2.68 | 1.67 | 3.49 | 2.96 | 1.80 | 3.78 |
| C (yellow) | 22.76 | 17.23 | 24.66 | 21.68 | 13.33 | 24.13 |
| STD | 2.59 | 1.30 | 3.70 | 2.82 | 1.35 | 4.05 |
| G (green) | 22.86 | 16.44 | 24.65 | 21.84 | 15.00 | 24.11 |
| STD | 2.50 | 1.34 | 3.56 | 2.72 | 1.43 | 3.85 |

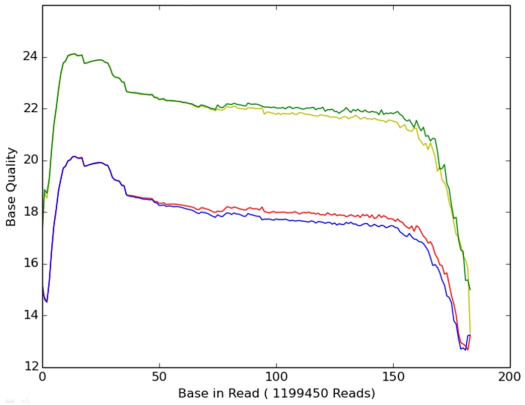



**Figure A.38:** *Pal7 self recalibration step 1; Average base quality distribution (y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

**Table A.30:** *Statistics for Klei10 after several filter steps; filter parameters are given with coverage genotype-quality.*

| Klei10 | Total Variants | het freq | het qual | hom freq | hom qual | ref qual | CT-TC | GA-AG | Total Pos | Data los % | Sampled |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Raw | 3401 | 0.49 | 49 | 0.51 | 7 | 7 | 0.20 | 0.18 | 1731952382 | | 3000000 |
| Filter 1 15 | 18843 | 0.94 | 49 | 0.06 | 17 | 16 | 0.33 | 0.27 | 159761844 | 90.78 | 3000000 |
| Filter 1 30 | 841540 | 1.00 | 52 | 0.00 | 59 | 42 | 0.35 | 0.29 | 1877103 | 99.89 | 1877103 |
| Filter 5 0 | 3901 | 0.83 | 60 | 0.17 | 17 | 13 | 0.31 | 0.24 | 273336618 | 84.22 | 3000000 |
| Filter 5 15 | 6787 | 0.83 | 61 | 0.17 | 17 | 16 | 0.29 | 0.25 | 159111798 | 90.81 | 3000000 |
| Filter 5 30 | 259303 | 1.00 | 66 | 0.00 | 59 | 42 | 0.37 | 0.31 | 1285820 | 99.93 | 1285820 |
| Filter 10 0 | 14398 | 0.94 | 68 | 0.06 | 55 | 29 | 0.14 | 0.12 | 2390709 | 99.86 | 2390709 |
| Filter 10 15 | 13237 | 0.94 | 74 | 0.06 | 58 | 33 | 0.13 | 0.12 | 2138223 | 99.88 | 2138223 |
| Filter 10 30 | 12510 | 0.93 | 77 | 0.07 | 60 | 42 | 0.13 | 0.11 | 1034733 | 99.94 | 1034733 |

**Table A.31:** *Statistics for Pal7 after several filter steps; filter parameters are given with coverage genotype-quality.*

| Pal7 | Total Variants | het freq | het qual | hom freq | hom qual | ref qual | CT-TC | GA-AG | Total Pos | Data los % | Sampled |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Raw | 3272 | 0.42 | 46 | 0.58 | 6 | 6 | 0.25 | 0.23 | 1150174070 | | 3000000 |
| Filter 1 15 | 46606 | 0.98 | 46 | 0.02 | 18 | 16 | 0.42 | 0.34 | 33473433 | 97.09 | 3000000 |
| Filter 1 30 | 472186 | 1.00 | 48 | 0.00 | 72 | 55 | 0.42 | 0.35 | 744530 | 99.94 | 744530 |
| Filter 5 0 | 4055 | 0.89 | 62 | 0.11 | 18 | 12 | 0.36 | 0.31 | 65849896 | 94.27 | 3000000 |
| Filter 5 15 | 7851 | 0.90 | 62 | 0.10 | 18 | 16 | 0.37 | 0.31 | 33033771 | 97.13 | 3000000 |
| Filter 5 30 | 70755 | 0.99 | 66 | 0.01 | 72 | 55 | 0.42 | 0.35 | 339560 | 99.97 | 339560 |
| Filter 10 0 | 5067 | 0.90 | 66 | 0.10 | 64 | 37 | 0.13 | 0.12 | 502977 | 99.96 | 502977 |
| Filter 10 15 | 4547 | 0.90 | 73 | 0.10 | 69 | 44 | 0.12 | 0.11 | 411059 | 99.96 | 411059 |
| Filter 10 30 | 4136 | 0.90 | 78 | 0.10 | 73 | 55 | 0.10 | 0.10 | 272133 | 99.98 | 272133 |

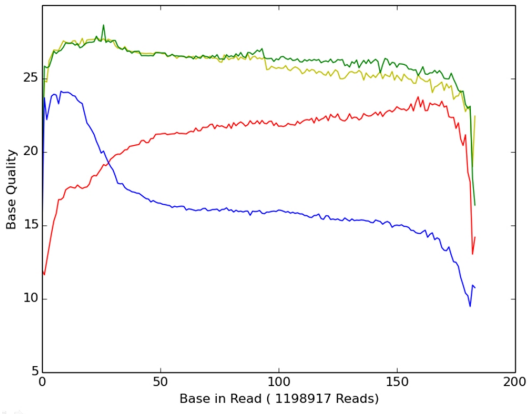



**Figure A.39:** *Pal7 sel frecalibration step 5; Average base quality distribution (y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

**Table A.32:** *Shows average base qualities for each base over all positions in all reads for Klei10 and Pal7 after the last self recalibration step ; Min and Max are chosen from the averages per position and therefore the std shows variance at that position.*

| Sample | Klei10 Step 6 | | | Pal7 Step 5 | | |
|---|---|---|---|---|---|---|
| Base | Average | Min | Max | Average | Min | Max |
| T (red) | 21.99 | 13.02 | 24.68 | 20.99 | 11.63 | 23.76 |
| STD | 3.68 | 2.62 | 7.07 | 4.16 | 2.68 | 6.52 |
| A (blue) | 18.07 | 11.36 | 25.57 | 16.55 | 9.48 | 24.13 |
| STD | 3.52 | 2.17 | 5.18 | 3.84 | 2.18 | 5.43 |
| C (yellow) | 26.82 | 18.88 | 28.81 | 25.85 | 18.25 | 27.74 |
| STD | 2.97 | 1.51 | 7.71 | 3.15 | 1.52 | 8.15 |
| G (green) | 27.15 | 18.44 | 28.15 | 26.18 | 16.38 | 28.65 |
| STD | 2.77 | 1.70 | 7.38 | 2.95 | 1.49 | 7.84 |

**Table A.33:** *Frequencies and total counts for above allele change distribution after applying damage and GATK recalibration to Klei10.*

| Change | AC | AG | AT | CA | CG | CT | GA | GC | GT | TA | TC | TG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Count | 129 | 508 | 108 | 120 | 142 | 0 | 0 | 140 | 135 | 102 | 495 | 125 |
| Frequncy | 0.06 | 0.25 | 0.05 | 0.06 | 0.07 | 0.00 | 0.00 | 0.07 | 0.07 | 0.05 | 0.25 | 0.06 |

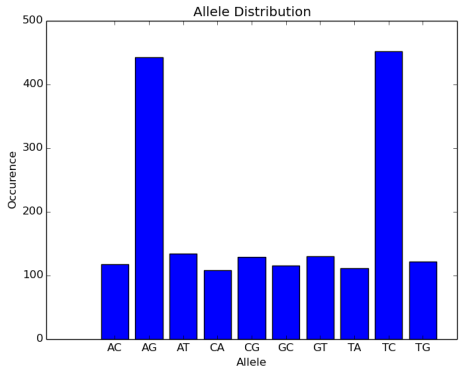



**Figure A.40:** *Occurrences of base changes for Pal7 after damage recalibration and GATK recalibration.*

**Table A.34:** *Frequencies and total counts for above allele change distribution after applying damage and GATK recalibration to Pal7.*

| Change | AC | AG | AT | CA | CG | CT | GA | GC | GT | TA | TC | TG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Count | 118 | 443 | 134 | 108 | 129 | 0 | 0 | 116 | 130 | 111 | 452 | 122 |
| Frequncy | 0.06 | 0.24 | 0.07 | 0.06 | 0.07 | 0.00 | 0.00 | 0.06 | 0.07 | 0.06 | 0.24 | 0.07 |

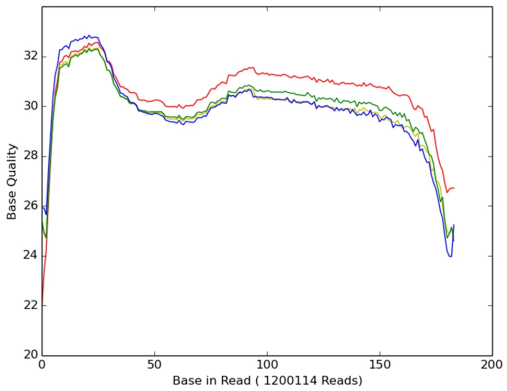



**Figure A.41:** *Klei10 damage recalibration followed by GATK; Average base quality distribution (y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

**Table A.35:** *Shows average base qualities for each base over all positions in all reads for Klei10 and Pal7 after applying damge and GATK recalibration ; Min and Max are chosen from the averages per position and therefore the std shows variance at that position.*

| Sample | Klei10 | | | Pal7 | | |
|---|---|---|---|---|---|---|
| Base | Average | Min | Max | Average | Min | Max |
| T (red) | 30.54 | 21.79 | 32.56 | 29.64 | 20.38 | 31.89 |
| STD | 4.11 | 2.19 | 9.61 | 4.19 | 2.16 | 9.01 |
| A (blue) | 29.83 | 23.97 | 32.84 | 28.54 | 20.56 | 32.24 |
| STD | 5.60 | 2.36 | 7.65 | 6.31 | 2.18 | 9.35 |
| C (yellow) | 29.81 | 24.72 | 32.34 | 29.17 | 24.00 | 32.12 |
| STD | 3.37 | 1.42 | 4.63 | 3.48 | 0.82 | 4.49 |
| G (green) | 29.96 | 24.59 | 32.29 | 29.34 | 24.08 | 32.09 |
| STD | 3.23 | 1.03 | 4.20 | 3.41 | 1.41 | 4.32 |

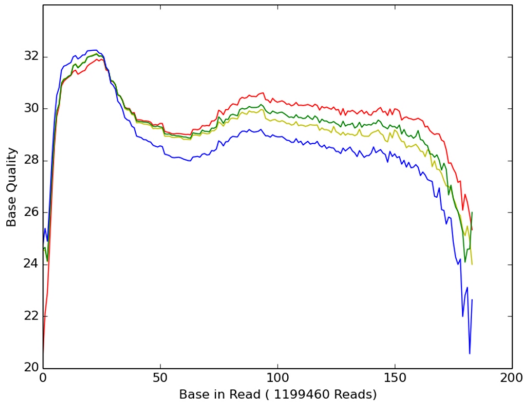



**Figure A.42:** *Pal7 damage recalibration followed by GATK; Average base quality distribution (y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

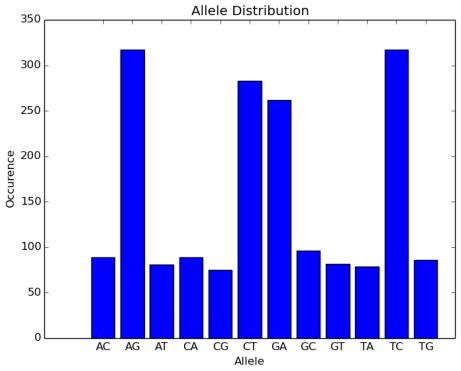



**Figure A.43:** *Occurrences of base changes for Pal7 after GATK followed by damage recalibration.*

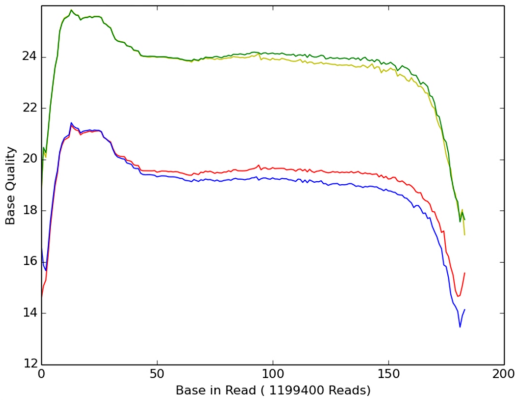



**Figure A.44:** *Klei10 damage recalibration followed by GATK; Average base quality distribution (y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

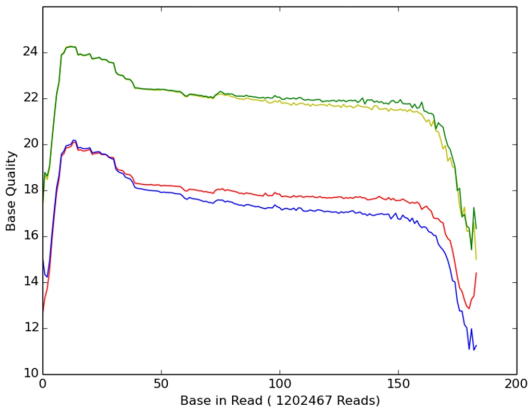



**Figure A.45:** *Klei10 damage recalibration followed by GATK; Average base quality distribution (y-axis) per position in read (x-axis); red =Thymine; blue = Adenine; yellow = Cytosine; green = Guanine*

## A.2 Capture

**Table A.36:** *Showing the overlap of the capture regions with the reference data[16, 37, 40]*

| Regions | Overlapping positions |
|---|---|
| Full capture | 1488 |
| Neutral | 74 |
| Relaxed neutral | 1260 |
| Phenotypic markers | 154 |

**Table A.37:** *Summary of sequencing results for each sample and contamination results from ContaMix and ANGSD*

| Sample | Capture Cov | Genome Cov | Endogen [%] | On target [%] | Uniqfull [%] | UniqCap[%] | Sex | mtDNA | | Chrom X Method1 | | | Chrom X Method2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Seq Error | con [%] | con [%] | SE | p-value | con [%] | SE | p-value |
| WEZ15 | 5.02 ± 2.98 | 1.25 ± 1.35 | 86.99 | 66.41 | 15.35 | 0.33 | XY | 0.01 | 0.61 | 0.01 | 0.00 | 0.58 | 0.11 | 0.00 | 1.00 |
| WEZ16 | 1.80 ± 5.80 | 1.08 ± 1.22 | 81.14 | 74.30 | 4.22 | 0.17 | XX | 0.01 | 0.10 | 40.73 | 0.00 | 0.00 | 68.29 | 0.00 | 0.00 |
| WEZ24 | 9.61 ± 4.51 | 1.50 ± 1.90 | 78.57 | 69.41 | 7.59 | 1.23 | XY | 0.01 | 0.11 | 0.01 | 0.00 | 1.00 | 0.01 | 0.00 | 1.00 |
| WEZ35 | 18.64 ± 8.15 | 1.64 ± 2.92 | 76.43 | 64.73 | 9.74 | 1.81 | XY | 0.01 | 0.08 | 1.74 | 0.01 | 0.02 | 0.45 | 0.01 | 0.14 |
| WEZ39 | 6.33 ± 3.50 | 1.40 ± 1.44 | 76.40 | 68.87 | 4.45 | 0.59 | XY | 0.01 | 0.74 | 1.89 | 0.02 | 0.30 | 0.00 | 0.05 | 0.30 |
| WEZ40 | 14.37 ± 6.89 | 1.52 ± 2.50 | 83.15 | 71.42 | 11.26 | 1.93 | XY | 0.01 | 2.36 | 1.02 | 0.01 | 0.48 | 0.87 | 0.01 | 0.56 |
| WEZ48 | 6.59 ± 10.71 | 1.54 ± 3.31 | 92.56 | 87.55 | 1.40 | 0.32 | XY | 0.01 | 6.05 | 0.01 | 0.00 | 1.00 | 0.01 | 0.00 | 1.00 |
| WEZ51 | 30.17 ± 15.12 | 2.09 ± 5.15 | 76.29 | 65.79 | 9.29 | 2.51 | XY | 0.01 | 0.14 | 0.22 | 0.00 | 1.00 | 0.01 | 0.00 | 0.61 |
| WEZ53 | 4.64 ± 4.23 | 1.16 ± 1.30 | 85.06 | 74.49 | 9.87 | 0.70 | XY | 0.01 | 0.15 | 0.01 | 0.00 | 1.00 | 0.01 | 0.00 | 1.00 |
| WEZ54 | 19.48 ± 7.77 | 1.68 ± 3.18 | 78.28 | 66.97 | 11.87 | 2.44 | XY | 0.01 | 3.88 | 0.74 | 0.01 | 0.66 | 3.31 | 0.04 | 0.18 |
| WEZ56 | 41.89 ± 20.34 | 2.03 ± 5.87 | 79.41 | 65.46 | 6.74 | 1.74 | XY | 0.01 | 0.15 | 0.01 | 0.00 | 1.00 | 0.56 | 0.01 | 0.38 |
| WEZ57 | 28.19 ± 14.29 | 1.97 ± 3.97 | 69.15 | 49.85 | 13.95 | 2.08 | XY | 0.01 | 0.06 | 0.10 | 0.00 | 0.70 | 0.01 | 0.00 | 0.72 |
| WEZ58 | 13.35 ± 6.71 | 1.62 ± 2.33 | 87.32 | 79.20 | 5.38 | 0.74 | XY | 0.01 | 0.12 | 2.80 | 0.02 | 0.03 | 4.55 | 0.04 | 0.07 |
| WEZ59 | 10.72 ± 4.69 | 1.38 ± 1.75 | 80.54 | 67.00 | 10.34 | 0.97 | XY | 0.01 | 0.19 | 5.17 | 0.02 | 0.00 | 2.20 | 0.03 | 0.09 |
| WEZ61 | 26.40 ± 10.03 | 1.77 ± 3.73 | 82.20 | 73.96 | 7.74 | 1.40 | XX | 0.01 | 0.77 | 28.07 | 0.00 | 0.00 | 26.78 | 0.00 | 0.00 |
| WEZ63 | 53.40 ± 30.46 | 2.89 ± 10.29 | 81.60 | 71.33 | 14.25 | 6.74 | XY | 0.01 | 9.70 | 23.28 | 0.00 | 0.00 | 20.15 | 0.00 | 0.00 |
| WEZ64 | 3.56 ± 2.30 | 1.17 ± 0.95 | 78.77 | 64.18 | 8.39 | 0.24 | XY | 0.01 | 0.08 | 1.81 | 0.01 | 0.10 | 0.01 | 0.00 | 1.00 |
| WEZ71 | 28.44 ± 16.19 | 2.27 ± 5.31 | 52.71 | 44.20 | 6.72 | 1.77 | XY | 0.01 | 0.80 | 1.18 | 0.01 | 0.05 | 0.00 | 0.00 | 0.63 |
| WEZ74 | 13.75 ± 6.03 | 2.07 ± 3.45 | 62.26 | 57.40 | 4.28 | 1.47 | XX | 0.01 | 1.44 | 27.90 | 0.00 | 0.00 | 21.05 | 0.00 | 0.00 |
| WEZ77 | 12.79 ± 5.62 | 1.59 ± 2.41 | 76.36 | 68.62 | 7.22 | 1.39 | XY | 0.01 | 11.04 | 35.43 | 0.00 | 0.00 | 33.84 | 0.00 | 0.00 |
| WEZ83 | 37.86 ± 19.25 | 2.29 ± 7.04 | 63.15 | 55.72 | 9.83 | 3.31 | XY | 0.01 | 0.18 | 0.01 | 0.00 | 1.00 | 0.01 | 0.00 | 1.00 |

**Table A.38:** *PCA resuts for each PCA and sample*

| Sample | Full Genome | | | | | Refined Full Genome | | | | | Refined Capture | | | | | Capture | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Loci | Cov | t | PC1 | PC2 | Loci | Cov | t | PC1 | PC2 | Loci | Cov | t | PC1 | PC2 | Loci | Cov | t | PC1 | PC2 |
| WEZ15 | 113487 | 0.27 | 1.00 | 93.50 | -94.70 | 93965 | 0.22 | 1.00 | 78.09 | -86.71 | 1436 | 0.01 | 0.99 | 55.25 | -102.71 | 1612 | 0.01 | 0.99 | 82.52 | -104.31 |
| WEZ16 | 11239 | 0.02 | 1.00 | 79.04 | -118.63 | 8874 | 0.02 | 1.00 | 72.98 | -115.29 | 642 | 0.00 | 0.97 | 42.79 | -166.51 | 769 | 0.00 | 0.98 | 73.55 | -154.43 |
| WEZ24 | 23856 | 0.07 | 1.00 | 88.77 | -105.25 | 19878 | 0.06 | 1.00 | 71.53 | -99.22 | 1532 | 0.02 | 0.99 | 61.17 | -117.08 | 1655 | 0.03 | 0.99 | 81.36 | -121.17 |
| WEZ35 | 46545 | 0.16 | 1.00 | 85.37 | -106.59 | 39083 | 0.13 | 1.00 | 68.93 | -99.44 | 1629 | 0.04 | 0.99 | 50.10 | -126.89 | 1720 | 0.05 | 0.99 | 61.26 | -120.18 |
| WEZ39 | 19268 | 0.05 | 1.00 | 80.52 | -85.65 | 15903 | 0.04 | 1.00 | 71.98 | -90.92 | 1451 | 0.01 | 0.99 | 64.46 | -103.21 | 1600 | 0.02 | 0.99 | 73.18 | -93.29 |
| WEZ40 | 42754 | 0.13 | 1.00 | 83.00 | -90.11 | 34070 | 0.10 | 1.00 | 73.04 | -83.36 | 1526 | 0.03 | 0.99 | 36.61 | -93.13 | 1673 | 0.04 | 0.99 | 52.31 | -98.84 |
| WEZ48 | 9999 | 0.03 | 1.00 | 98.93 | -98.18 | 8378 | 0.03 | 1.00 | 78.31 | -101.08 | 1458 | 0.01 | 0.99 | 54.07 | -119.90 | 1600 | 0.02 | 0.99 | 77.47 | -123.99 |
| WEZ51 | 40445 | 0.19 | 1.00 | 78.14 | -103.38 | 34524 | 0.16 | 1.00 | 58.01 | -99.96 | 1638 | 0.08 | 0.99 | 41.59 | -109.12 | 1717 | 0.09 | 0.99 | 61.09 | -117.71 |
| WEZ53 | 34904 | 0.08 | 1.00 | 80.13 | -105.29 | 28299 | 0.07 | 1.00 | 75.85 | -99.89 | 1374 | 0.01 | 0.99 | 61.15 | -129.77 | 1536 | 0.01 | 0.99 | 74.77 | -121.31 |
| WEZ54 | 44811 | 0.16 | 1.00 | 93.56 | -113.67 | 37253 | 0.13 | 1.00 | 75.89 | -104.98 | 1581 | 0.05 | 0.99 | 61.83 | -99.47 | 1708 | 0.05 | 0.99 | 95.41 | -90.21 |
| WEZ56 | 56346 | 0.25 | 1.00 | 86.14 | -98.56 | 47450 | 0.21 | 1.00 | 75.56 | -94.43 | 1636 | 0.10 | 0.99 | 50.85 | -98.80 | 1725 | 0.11 | 0.99 | 64.23 | -97.99 |
| WEZ57 | 76654 | 0.31 | 1.00 | 83.74 | -103.72 | 62775 | 0.25 | 1.00 | 63.26 | -98.22 | 1553 | 0.07 | 0.99 | 17.76 | -118.55 | 1710 | 0.08 | 0.99 | 54.39 | -113.38 |
| WEZ58 | 34915 | 0.11 | 1.00 | 91.25 | -85.49 | 28086 | 0.09 | 1.00 | 77.10 | -82.51 | 1541 | 0.03 | 0.99 | 57.25 | -107.43 | 1692 | 0.04 | 0.99 | 68.77 | -111.20 |
| WEZ59 | 55832 | 0.15 | 1.00 | 80.14 | -103.51 | 44446 | 0.12 | 1.00 | 68.93 | -96.13 | 1553 | 0.02 | 0.99 | 61.61 | -123.38 | 1696 | 0.03 | 0.99 | 85.68 | -123.97 |
| WEZ61 | 56972 | 0.21 | 1.00 | 85.69 | -105.79 | 48691 | 0.18 | 1.00 | 69.29 | -99.02 | 1646 | 0.06 | 0.99 | 56.35 | -100.33 | 1729 | 0.07 | 0.99 | 75.89 | -109.06 |
| WEZ63 | 30302 | 0.22 | 1.00 | 90.08 | -105.81 | 24479 | 0.18 | 1.00 | 65.38 | -104.08 | 1630 | 0.13 | 0.99 | 29.89 | -109.87 | 1711 | 0.15 | 0.99 | 66.03 | -108.34 |
| WEZ64 | 58532 | 0.13 | 1.00 | 83.27 | -97.53 | 48963 | 0.11 | 1.00 | 73.81 | -92.77 | 1378 | 0.01 | 0.99 | 47.48 | -108.56 | 1522 | 0.01 | 0.99 | 60.73 | -100.36 |
| WEZ71 | 38049 | 0.18 | 1.00 | 86.32 | -101.35 | 31588 | 0.15 | 1.00 | 63.64 | -100.53 | 1587 | 0.07 | 0.99 | 71.33 | -90.03 | 1683 | 0.08 | 0.99 | 102.41 | -87.14 |
| WEZ74 | 15191 | 0.07 | 1.00 | 78.87 | -107.89 | 12793 | 0.06 | 1.00 | 58.64 | -99.29 | 1570 | 0.03 | 0.99 | 51.59 | -104.81 | 1690 | 0.04 | 0.99 | 75.89 | -98.77 |
| WEZ77 | 30447 | 0.10 | 1.00 | 78.18 | -93.90 | 24523 | 0.08 | 1.00 | 55.52 | -93.65 | 1574 | 0.03 | 0.99 | 55.97 | -111.59 | 1685 | 0.03 | 0.99 | 79.97 | -104.25 |
| WEZ83 | 31818 | 0.17 | 1.00 | 90.28 | -91.44 | 26372 | 0.14 | 1.00 | 79.50 | -89.88 | 1553 | 0.09 | 0.99 | 66.31 | -84.90 | 1699 | 0.10 | 0.99 | 82.92 | -84.56 |

## A.3 Case study Welzin

**Table A.39:** *Details of ancient individuals and population grouping; [75] refers to Mathieson et al. 2015; [5] refers to Allentoft et al. 2015;[49] referes to Jones et al. 2015; [64] refers to Lazaridis et al. 2014; this table is an edited subset of the* ***Supplementary Table 1*** *from Lazaridis et al. 2016 [63].*

| Genetic ID | Analysis Label | Detailed Label | Date | Location | Country |
|---|---|---|---|---|---|
| RISE559 [5] | Europe_LNBA | Bell_Beaker_Germany | na | Augsburg | Germany |
| RISE560 [5] | Europe_LNBA | Bell_Beaker_Germany | na | Augsburg | Germany |
| RISE562 [5] | Europe_LNBA | Bell_Beaker_Germany | na | Landau an der Isar | Germany |
| RISE563 [5] | Europe_LNBA | Bell_Beaker_Germany | na | Osterhofen-Altenmarkt | Germany |
| RISE564 [5] | Europe_LNBA | Bell_Beaker_Germany | na | Osterhofen-Altenmarkt | Germany |
| I0805 [75] | Europe_LNBA | Bell_Beaker_Germany | 2467-2142 calBCE | Quedlinburg VII | Germany |
| I0806 [75] | Europe_LNBA | Bell_Beaker_Germany | 2431-2150 calBCE | Quedlinburg VII | Germany |
| I0112 [75] | Europe_LNBA | Bell_Beaker_Germany | 2457-2142 calBCE | Quedlinburg XII | Germany |
| I0111 [75] | Europe_LNBA | Bell_Beaker_Germany | 2475-2204 calBCE | Rothenschirmbach | Germany |
| I0108 [75] | Europe_LNBA | Bell_Beaker_Germany | 2575-2299 calBCE | Rothenschirmbach | Germany |
| I0113 [75] | Europe_LNBA | Bell_Beaker_Germany | 2346-2033 calBCE | Quedlinburg XII | Germany |
| I0060 [75] | Europe_LNBA | Bell_Beaker_Germany | 2428-2149 calBCE | Rothenschirmbach | Germany |
| I1546 [75] | Europe_LNBA | Bell_Beaker_Germany | 2500-2050 BCE | Benzingerode-Heimburg | Germany |
| I1549 [75] | Europe_LNBA | Bell_Beaker_Germany | 2500-2050 BCE | Benzingerode-Heimburg | Germany |
| I0059 [75] | Europe_LNBA | BenzigerodeHeimburg_LN | 2337-2138 calBCE | Benzingerode-Heimburg | Germany |
| I0171 [75] | Europe_LNBA | BenzigerodeHeimburg_LN | 2287-2041 calBCE | Benzingerode-Heimburg | Germany |
| KK1 [49] | CHG | CHG | 7940-7600 calBCE | Kotias Klde | Georgia |
| SATP [49] | CHG | CHG | 11430-11180 calBCE | Satsurblia | Georgia |
| RISE435 [5] | Europe_LNBA | Corded_Ware_Germany | 2863-2498 calBCE | Tiefbrunn | Germany |
| RISE434 [5] | Europe_LNBA | Corded_Ware_Germany | 2880-2630 calBCE | Tiefbrunn | Germany |
| RISE436 [5] | Europe_LNBA | Corded_Ware_Germany | 2868-2580 calBCE | Tiefbrunn | Germany |
| RISE446 [5] | Europe_LNBA | Corded_Ware_Germany | 2829-2465 calBCE | Bergrheinfeld | Germany |
| I0104 [75] | Europe_LNBA | Corded_Ware_Germany | 2559-2296 calBCE | Esperstedt | Germany |
| I0103 [75] | Europe_LNBA | Corded_Ware_Germany | 2578-2468 calBCE | Esperstedt | Germany |
| I0049 [75] | Europe_LNBA | Corded_Ware_Germany | 2464-2210 calBCE | Esperstedt | Germany |
| I0106 [75] | Europe_LNBA | Corded_Ware_Germany | 2464-2210 calBCE | Esperstedt | Germany |
| I1539 [75] | Europe_LNBA | Corded_Ware_Germany | 2625-2291 calBCE | Esperstedt | Germany |
| I1532 [75] | Europe_LNBA | Corded_Ware_Germany | 2500-2050 BCE | Esperstedt | Germany |
| I1538 [75] | Europe_LNBA | Corded_Ware_Germany | 2500-2050 BCE | Esperstedt | Germany |
| I1540 [75] | Europe_LNBA | Corded_Ware_Germany | 2500-2050 BCE | Esperstedt | Germany |
| I1542 [75] | Europe_LNBA | Corded_Ware_Germany | 2500-2050 BCE | Esperstedt | Germany |
| I1534 [75] | Europe_LNBA | Corded_Ware_Germany | 2500-2050 BCE | Esperstedt | Germany |
| I1544 [75] | Europe_LNBA | Corded_Ware_Germany | 2500-2050 BCE | Esperstedt | Germany |
| I1536 [75] | Europe_LNBA | Corded_Ware_Germany | 2500-2050 BCE | Esperstedt | Germany |
| I1504 [75] | Europe_LNBA | Late_Bronze_Age | 1270-1110 calBCE | Ludas-Varju-Dulo | Hungary |
| I0054 [75] | Europe_EN | LBK_EN | 5222-5022 calBCE | Unterwiederstedt | Germany |
| I0048 [75] | Europe_EN | LBK_EN | 5211-5009 calBCE | Halberstadt-Sonntagsfeld | Germany |
| I0057 [75] | Europe_EN | LBK_EN | 5218-5019 calBCE | Halberstadt-Sonntagsfeld | Germany |
| I0046 [75] | Europe_EN | LBK_EN | 5212-4989 calBCE | Halberstadt-Sonntagsfeld | Germany |
| I0795 [75] | Europe_EN | LBK_EN | 5216-5036 calBCE | Karsdorf | Germany |
| I0797 [75] | Europe_EN | LBK_EN | 5500-4775 BCE | Karsdorf | Germany |
| I1550 [75] | Europe_EN | LBK_EN | 5500-4775 BCE | Halberstadt-Sonntagsfeld | Germany |
| I0100 [75] | Europe_EN | LBK_EN | 5202-4852 calBCE | Halberstadt-Sonntagsfeld | Germany |
| I0659 [75] | Europe_EN | LBK_EN | 5211-4963 calBCE | Halberstadt-Sonntagsfeld | Germany |
| I0025 [75] | Europe_EN | LBK_EN | 5500-4800 BCE | Stuttgart-Muehlhausen | Germany |
| I0026 [75] | Europe_EN | LBK_EN | 5500-4800 BCE | Stuttgart-Muehlhausen | Germany |
| I0022 [75] | Europe_EN | LBK_EN | 5500-4800 BCE | Stuttgart-Muehlhausen | Germany |
| I0821 [75] | Europe_EN | LBK_EN | 5201-4850 calBCE | Halberstadt-Sonntagsfeld | Germany |
| Loschbour [64] | WHG | Loschbour_HG | 6210-5990 calBCE | Echternach | Luxembourg |
| I0012 [75] | SHG | Motala_HG | 5714-5575 calBCE | Motala, Kanaljorden | Sweden |
| I0017 [75] | SHG | Motala_HG | 5721-5631 calBCE | Motala, Kanaljorden | Sweden |
| I0011 [75] | SHG | Motala_HG | 5721-5516 calBCE | Motala, Kanaljorden | Sweden |
| I0013 [75] | SHG | Motala_HG | 5964-5638 calBCE | Motala, Kanaljorden | Sweden |
| I0014 [75] | SHG | Motala_HG | 5877-5629 calBCE | Motala, Kanaljorden | Sweden |
| I0015 [75] | SHG | Motala_HG | 5964-5629 calBCE | Motala, Kanaljorden | Sweden |
| RISE47 [5] | Europe_LNBA | Nordic_BA | 1499-1324 calBCE | Sebber skole | Denmark |
| RISE175 [5] | Europe_LNBA | Nordic_BA | 1395-1132 calBCE | Abekås I | Sweden |
| RISE210 [5] | Europe_LNBA | Nordic_BA | 1432-1292 calBCE | Ängamöllan | Sweden |

| | | | | | |
|---|---|---|---|---|---|
| RISE71 [5] | Europe_LNBA | Nordic_LN | 2196-2023 calBCE | Falshöj | Denmark |
| RISE179 [5] | Europe_LNBA | Nordic_LN | 2010-1776 calBCE | Abekås I | Sweden |
| RISE98 [5] | Europe_LNBA | Nordic_LN | 2275-2032 calBCE | L Beddinge 56 | Sweden |
| RISE97 [5] | Europe_LNBA | Nordic_LN | 2025-1885 calBCE | Fredriksberg | Sweden |
| RISE577 [5] | Europe_LNBA | Unetice_EBA | na | Velke Prilepy | Czech Republic |
| RISE586 [5] | Europe_LNBA | Unetice_EBA | na | Moravska Nova Ves | Czech Republic |
| RISE154 [5] | Europe_LNBA | Unetice_EBA | 1925-1765 calBCE | Szczepankowice | Poland |
| RISE109 [5] | Europe_LNBA | Unetice_EBA | 1954-1772 calBCE | Wojkowice | Poland |
| RISE150 [5] | Europe_LNBA | Unetice_EBA | 1885-1693 calBCE | Przeclawice | Poland |
| I0117 [75] | Europe_LNBA | Unetice_EBA | 2272-2039 calBCE | Esperstedt | Germany |
| I0115 [75] | Europe_LNBA | Unetice_EBA | 1954-1760 calBCE | Esperstedt | Germany |
| I0116 [75] | Europe_LNBA | Unetice_EBA | 2134-1939 calBCE | Esperstedt | Germany |
| I0803 [75] | Europe_LNBA | Unetice_EBA | 2132-1942 calBCE | Eulau | Germany |
| I0804 [75] | Europe_LNBA | Unetice_EBA | 2137-1965 calBCE | Eulau | Germany |
| I0047 [75] | Europe_LNBA | Unetice_EBA | 2111-1891 calBCE | Halberstadt-Sonntagsfeld | Germany |
| I0164 [75] | Europe_LNBA | Unetice_EBA | 2023-1894 calBCE | Quedlinburg VIII | Germany |
| Bichon [49] | Switzerland_HG | Bichon | 11820-11610 calBCE | Grotte du Bichon | Switzerland |
| Stuttgart [64] | Europe_EN | LBK_EN | 5310-5070 calBCE | Stuttgart-Muehlhausen | Germany |
| RISE546 [5] | Steppe_EMBA | Yamnaya_Kalmykia | na | Temrta IV | Russia |
| RISE548 [5] | Steppe_EMBA | Yamnaya_Kalmykia | na | Temrta IV | Russia |
| RISE547 [5] | Steppe_EMBA | Yamnaya_Kalmykia | 2887-2634 calBCE | Temrta IV | Russia |
| RISE552 [5] | Steppe_EMBA | Yamnaya_Kalmykia | 2849-2143 calBCE | Ulan IV | Russia |
| RISE240 [5] | Steppe_EMBA | Yamnaya_Kalmykia | 2880-2632 calBCE | Sukhaya Termista I | Russia |
| RISE550 [5] | Steppe_EMBA | Yamnaya_Kalmykia | 3334-2635 calBCE | Peshany V | Russia |
| I0370 [75] | Steppe_EMBA | Yamnaya_Samara | 3300-2700 BCE | Samara | Russia |
| I0441 [75] | Steppe_EMBA | Yamnaya_Samara | 3010-2622 calBCE | Samara | Russia |
| I0444 [75] | Steppe_EMBA | Yamnaya_Samara | 3335-2882 calBCE | Samara | Russia |
| I0429 [75] | Steppe_EMBA | Yamnaya_Samara | 3339-2918 calBCE | Samara | Russia |
| I0231 [75] | Steppe_EMBA | Yamnaya_Samara | 2921-2762 calBCE | Samara | Russia |
| I0438 [75] | Steppe_EMBA | Yamnaya_Samara | 3021-2635 calBCE | Samara | Russia |
| I0439 [75] | Steppe_EMBA | Yamnaya_Samara | 3321-2921 calBCE | Samara | Russia |
| I0443 [75] | Steppe_EMBA | Yamnaya_Samara | 3300-2700 BCE | Samara | Russia |
| I0357 [75] | Steppe_EMBA | Yamnaya_Samara | 3090-2913 calBCE | Samara | Russia |
| I0061 [75] | EHG | Karelia_HG | 6850-6000 BCE | Karelia | Russia |
| I0211 [75] | EHG | Karelia_HG | 5500-5000 BCE | Karelia | Russia |
| I0124 [75] | EHG | Samara_HG | 5657-5541 calBCE | Samara | Russia |

**Table A.40:** *Lab code and archaeological Inventory number*

| Lab Code | Inv.-Nr. |
|---|---|
| WEZ15 | ALM 2000/1382-0001 |
| WEZ16 | ALM 2008/0459-0001 |
| WEZ24 | ALM 2011/1145-0428 |
| WEZ35 | ALM 2008/0460-0223 |
| WEZ39 | ALM 1996/0855-0055 |
| WEZ40 | ALM 1996/0855-0141 |
| WEZ48 | ALM 2010/1093-1022 |
| WEZ51 | ALM 2013/0463-1946 |
| WEZ53 | ALM 2013/0463-2000 |
| WEZ54 | ALM 2013/0463-1056 |
| WEZ56 | ALM 2013/0463-1009 |
| WEZ57 | ALM 2013/0463-0967 |
| WEZ58 | ALM 2013/0463-0277 |
| WEZ59 | ALM 2013/0463-1980 |
| WEZ61 | ALM 2013/0463-1584 |
| WEZ63 | ALM 2013/0463-1534 |
| WEZ64 | ALM 2013/0463-1729 |
| WEZ71 | ALM 2013/0463-0408 |
| WEZ74 | ALM 2013/0463-1588 |
| WEZ77 | ALM 2013/0463-0658 |
| WEZ83 | ALM 2013/0463-0932 |

**Table A.41:** *Showing Missingness per Welzin individual in the used reference [63]; Missingness is defined as the fraction of missing genotypes in the data set; #Pos = shows the number of overlapping positions with reference; #Pos LD small = shows the number of overlapping positions after LD filtering on the small data set; #Pos LD large = shows the number of overlapping positions after LD filtering on the large subset*

| Individual | Mssingness | #Pos full | #Pos LD small | #Pos LD large |
|---|---|---|---|---|
| WEZ15 | 0.8676 | 159055 | 119338 | 63983 |
| WEZ16 | 0.987 | 15671 | 11902 | 6382 |
| WEZ24 | 0.9676 | 38925 | 29494 | 16210 |
| WEZ35 | 0.9325 | 81094 | 62091 | 34269 |
| WEZ39 | 0.9778 | 26639 | 20389 | 11439 |
| WEZ40 | 0.949 | 61224 | 46382 | 25230 |
| WEZ48 | 0.9847 | 18363 | 14063 | na |
| WEZ51 | 0.9632 | 44190 | 34099 | 19380 |
| WEZ53 | 0.9659 | 40987 | 30334 | 16119 |
| WEZ54 | 0.9306 | 83331 | 63271 | 34589 |
| WEZ56 | 0.9336 | 79734 | 60859 | 33598 |
| WEZ57 | 0.9179 | 98550 | 76506 | 43451 |
| WEZ58 | 0.9491 | 61177 | 46758 | 25811 |
| WEZ59 | 0.9342 | 79049 | 59498 | 32207 |
| WEZ61 | 0.9176 | 98995 | 75393 | 42015 |
| WEZ64 | 0.9383 | 74072 | 56607 | 31388 |
| WEZ71 | 0.9625 | 45012 | 35170 | 20220 |
| WEZ74 | 0.9763 | 28418 | 21909 | 12143 |
| WEZ83 | 0.9655 | 41458 | 31039 | 16609 |

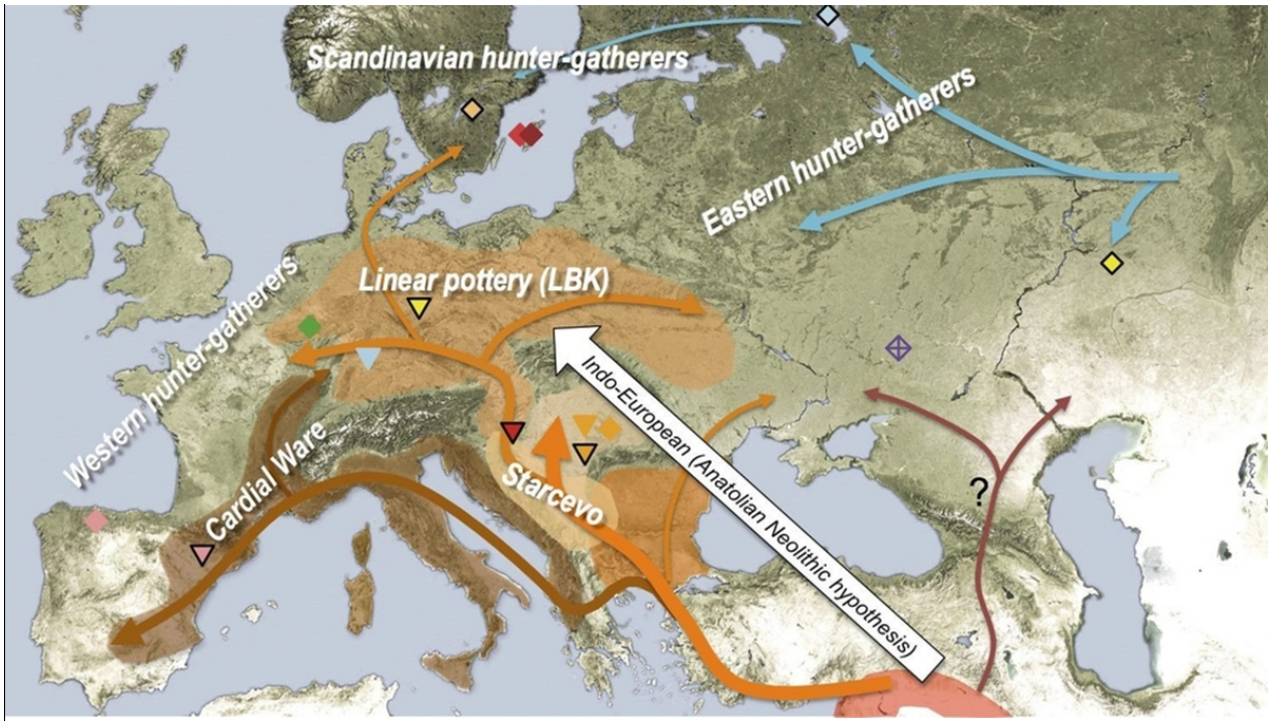



**Figure A.46:** *Showing cultural distribution and proposed population movement during the Neolthic period; modified from Haak et al. 2015 [36]*


**Figure A.47:** *Showing cultural distribution and proposed population movement during the early Bronze Age period; modified from Haak et al. 2015 [36]*

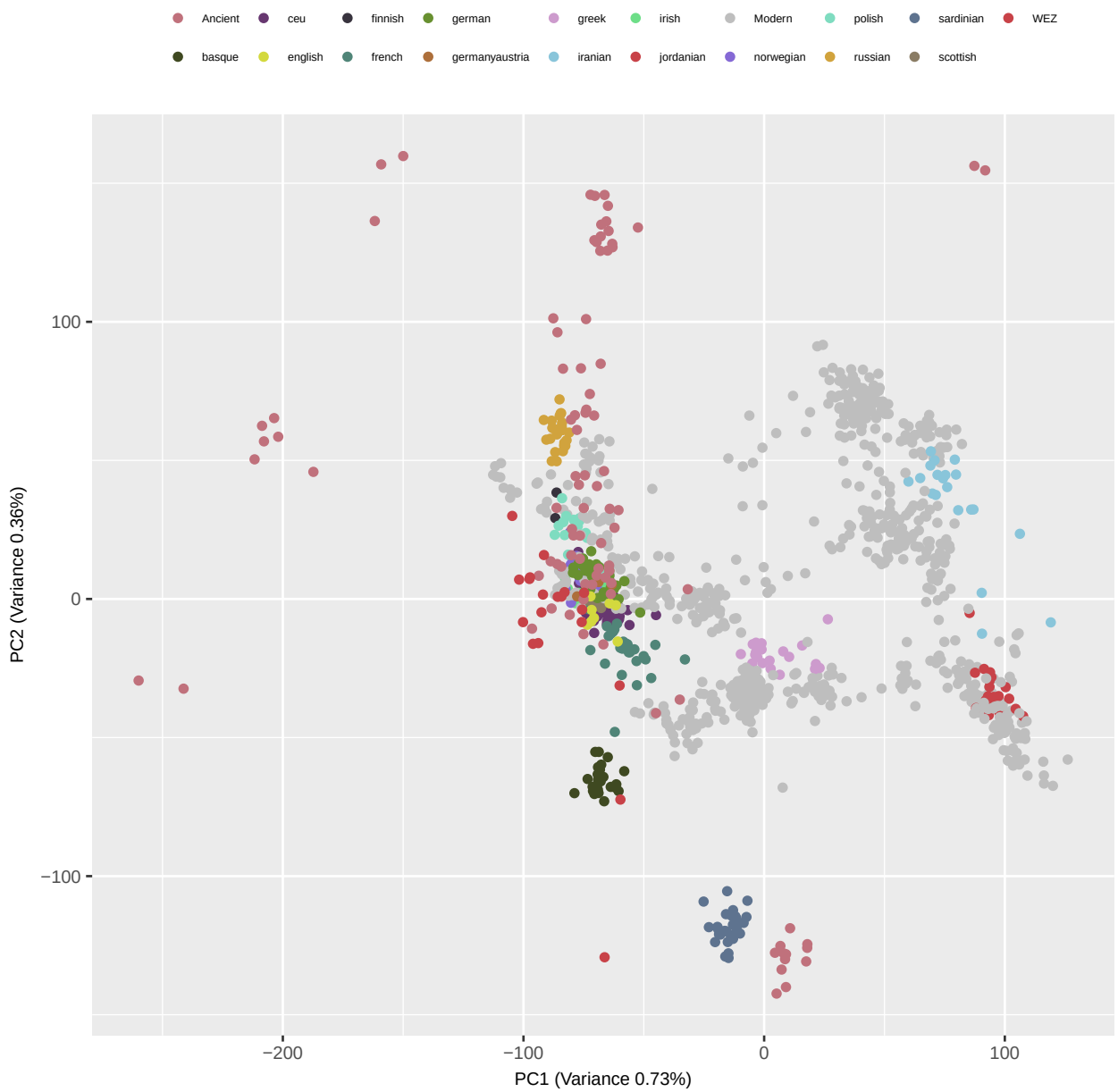



**Figure A.48:** *PCA of Weltzin samples and selected ancient samples[63] projected on the reference space used in Hoffmanova & Kreutzer 2016 [40]*

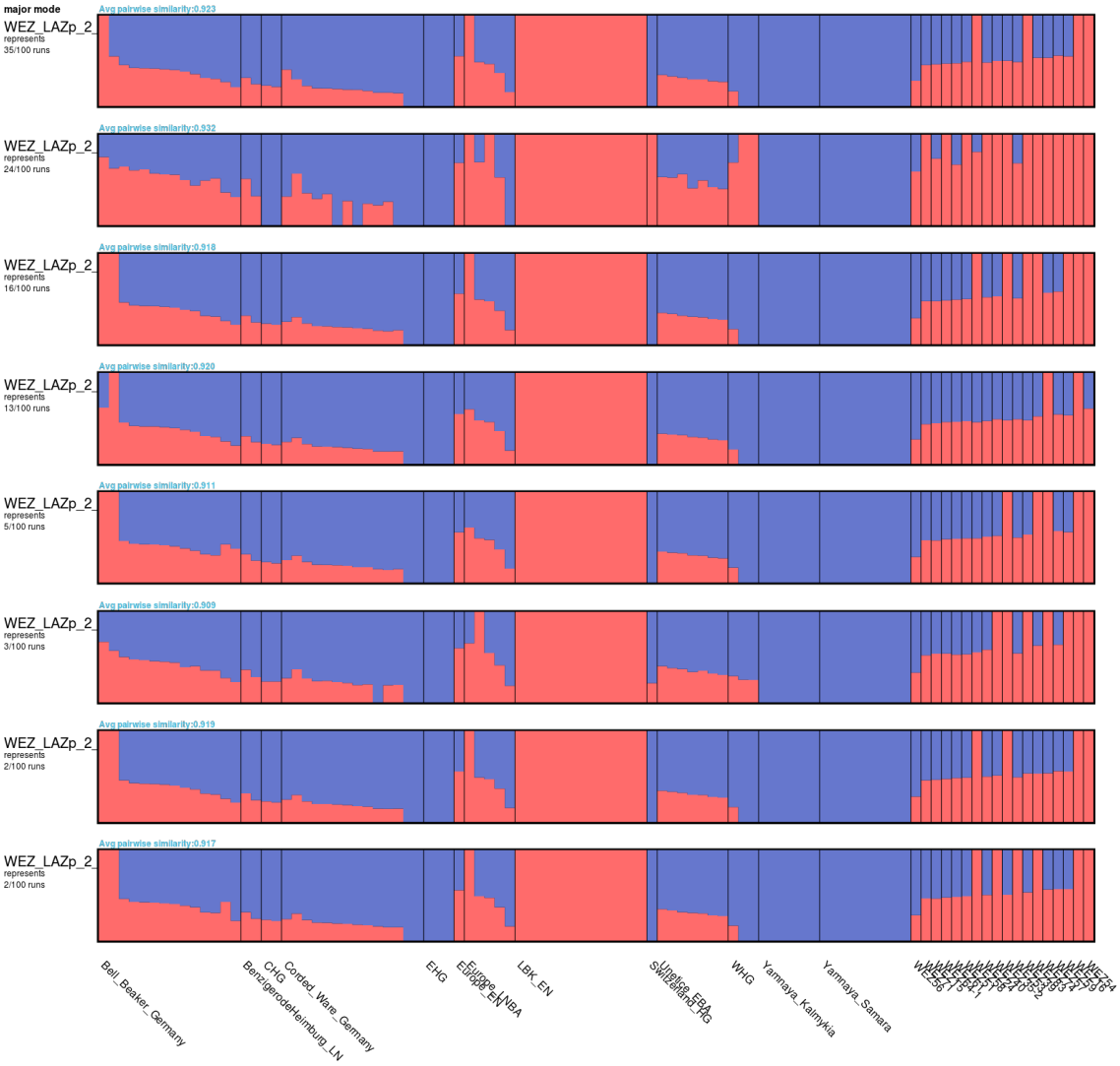



**Figure A.49:** *Showing all modes off the ADMIXTURE result for K 2 for the first analysis done; pairwise similarity is given in the upper left corner; the number of ADMIXTURE runs is given on the left.*

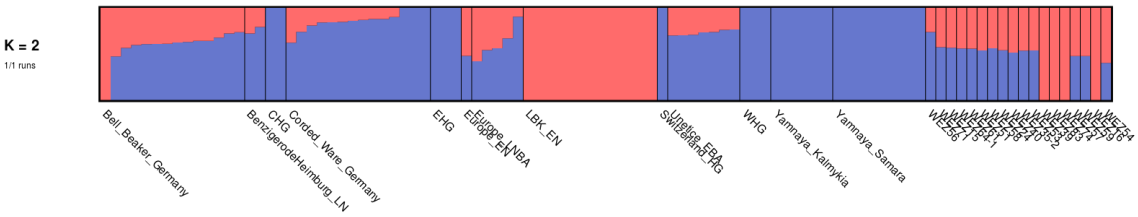



**Figure A.50:** *Showing the supervised ADMIXTURE result for K 2 from the first analysis done; ancestry components have the same colours but are switched from top to bottom;*

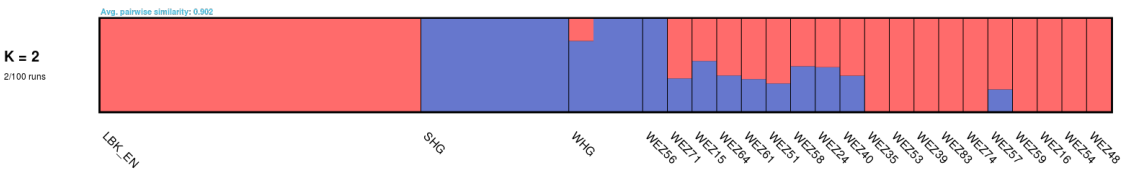



**Figure A.51:** *Showing the most likely plot of admixture analysis including WHG, SHG, LBK and Welzin individuals; consists of 21 out of 100 admixture runs with a pairwise similarity of 0.845*

## A.4 On lane contamination

**Table A.42:** *Summary of Sequencing results for each Sample; including known HVRI Positions*

| Lane | Lab code | Age | Site Situation | Region | Index in Cap (y/n) | On target [%] | Genome covered no decon [%] | Genome covered decon [%] | Known HVR1 Mutations |
|---|---|---|---|---|---|---|---|---|---|
| 1 | BLA14 | 3603± 38 bc | cave | Blätterhöhle, Hagen | y | 21,34 | 1,00 | 0,99 | 16270_T |
| 1 | BLA7 | 3666±20 bc | cave | Blätterhöhle, Hagen | n | 78,66 | 1,00 | 0,99 | 16304_C |
| 1 | QR8.2 | ca. 6400 bc | open air site | Qaleh Rostam, Iran | n | 13,07 | 0,95 | 0,46 | 16239_T |
| 1 | QR7 | ca. 6400 bc | open air site | Qaleh Rostam, Iran | n | 20,52 | 0,99 | 0,78 | no replikated HVR1 available |
| 1 | ZV122b_XXVIII | 5383±68 bc | open air site | Zvejnieki | n | 39,23 | 1,00 | 1,00 | 16192_T,16256_T,16270_T |
| 2 | BLA17 | 3681±19 bc | cave | Blätterhöhle, Hagen | y | 21,48 | 0,99 | 0,66 | 16189_C,16192_T,16270_T, 16311_C,16398_A |
| 2 | BLA8 | 3726±38 bc | cave | Blätterhöhle, Hagen | y | 0,21 | 1,00 | 0,65 | no Known SNPs |
| 2 | HV1 | ca. 1600 bc | open air site | Haftavan Tepe, Iran | y | 10,35 | 1,00 | 0,95 | 16189_C,16192_T,16256_T, 16291_T,16399_G |
| 2 | Krk4 | ca. 5000 bc | open air site | Kremasti Koilada, Northern Greece | y | 0,24 | 1,00 | 0,80 | 16066_G,16129_A,16183_C, 16189_C,16234_T |
| 2 | MIN4_XXVI | 8740-8420 bc | open air site | Minino | y | 60,50 | 1,00 | 1,00 | 16192_T,16256_T,16270_T, 16399_G |
| 2 | WC1 | ca. 1600 bc | open air site | Wezmeh Cave, Iran | y | 61,68 | 1,00 | 1,00 | 16069_T,16126_C,16193_T |
| 2 | ZV317b_XXIII | 3890±67 bc | open air site | Zvejnieki | n | 27,21 | 1,00 | 0,79 | 16134_T,16356_C |
| 3 | BLA16 | 3429±60 bc | cave | Blätterhöhle, Hagen | y | 1,22 | 0,97 | 0,53 | 16186_T,16293_G,16311_C, 16398_A |
| 3 | BLA20 | 8652±58 bc | cave | Blätterhöhle, Hagen | y | 67,40 | 1,00 | 1,00 | 16192_T,16256_T,16270_T |
| 3 | BLA9 | 3681±19 bc | cave | Blätterhöhle, Hagen | y | 6,06 | 0,92 | 0,68 | 16270_T |
| 3 | Min1_XXXV | 8850-8550 bc | open air site | Minino | n | 80,24 | 1,00 | 1,00 | 16356_C |
| 3 | ZV317b_XXXV | 3890±67 bc | open air site | Zvejnieki | n | 8,18 | 1,00 | 0,91 | 16134_T,16356_C |
| 3 | BLA6 | 8676±44 bc | cave | Blätterhöhle, Hagen | y | 0,14 | 0,96 | 0,78 | 16172_C,16189_C,16270_T, 16398_A |
| 4 | Kru3a | Neolihic | open air site | Krukow | y | 0,17 | 0,95 | 0,37 | 16342_C |
| 4 | Kru3b | Neolihic | open air site | Krukow | y | 4,25 | 0,81 | 0,17 | 16342_C |
| 4 | Min1_XXIII | 8850-8550 bc | open air site | Minino | n | 67,10 | 1,00 | 1,00 | 16356_C |
| 4 | BLA5 | 3335±136 bc | cave | Blätterhöhle, Hagen | y | 65,13 | 0,88 | 0,63 | no Known SNPs |
| 4 | Lie4a | Neolihic | open air site | Liepen | y | 0,10 | 0,90 | 0,28 | no replikated HVR1 available |
| 4 | Lie4b | Neolihic | open air site | Liepen | y | 0,06 | 0,83 | 0,20 | no replikated HVR1 available |
| 4 | ZV122b_XXXV | 5383±68 bc | open air site | Zvejnieki | n | 4,36 | 1,00 | 0,99 | 16192_T,16256_T,16270_T |
| 5 | Fra1 | ca. 8000 bc | cave | Franchthi cave, Southern Greece | y | 0,17 | 0,33 | 0,08 | 16189_C,16223_T,16274_A, 16278_T |
| 5 | Klei1 | ca. 5000 bc | open air site | Kleitos, Northern Greece | y | 60,20 | 0,90 | 0,87 | 16224_C,16311_C,16327_C |
| 5 | Mau3 | ca. 450 ad | open air site | Mauropigi, Northern Greece | y | 3,30 | 1,00 | 1,00 | 16126_C,16163_G,16169_T, 16186_T,16189_C,16263_C, 16294_T,16311_C |
| 5 | Theo1 | ca. 7050 bc | cave | Theopetra cave | y | 16,68 | 0,71 | 0,17 | 16224_C,16311_C |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5 | Theo5 | ca. 7050 bc | cave | Theopetra cave | y | 2,28 | 0,46 | 0,08 | 16126_C,16291_T,16294_T, 16296_T |
| 5 | Xir2 | ca. 6200 bc | open air site | Xirolimni, Northern Greece | y | 7,00 | 0,87 | 0,85 | 16126_C,16163_G,16186_T, 16189_C,16294_T |
| 6 | BLA10 | 3418±63 | cave | Blätterhöhle, Hagen | n | 75,00 | 1,00 | 1,00 | no Known SNPs |
| 6 | BLA11 | 3922±60 | cave | Blätterhöhle, Hagen | n | 60,72 | 1,00 | 0,94 | 16270_T |
| 6 | BLA13 | 3513±102 | cave | Blätterhöhle, Hagen | n | 72,05 | 1,00 | 1,00 | 16304_C |
| 6 | QR6 | ca. 6400 bc | open air site | Qaleh Rostam, Iran | n | 0,56 | 0,79 | 0,28 | no replikated HVR1 available |
| 6 | BLA19 | 8638±56 | cave | Blätterhöhle, Hagen | n | 0,25 | 0,98 | 0,48 | no replikated HVR1 available |
| 6 | BLA2 | 8748±67 | cave | Blätterhöhle, Hagen | n | 34,32 | 0,92 | 0,62 | no replikated HVR1 available |
| 6 | HV8 | ca. 1600 bc | open air site | Haftavan Tepe, Iran | n | 69,56 | 1,00 | 1,00 | 16126_C,16163_G,16294_T |

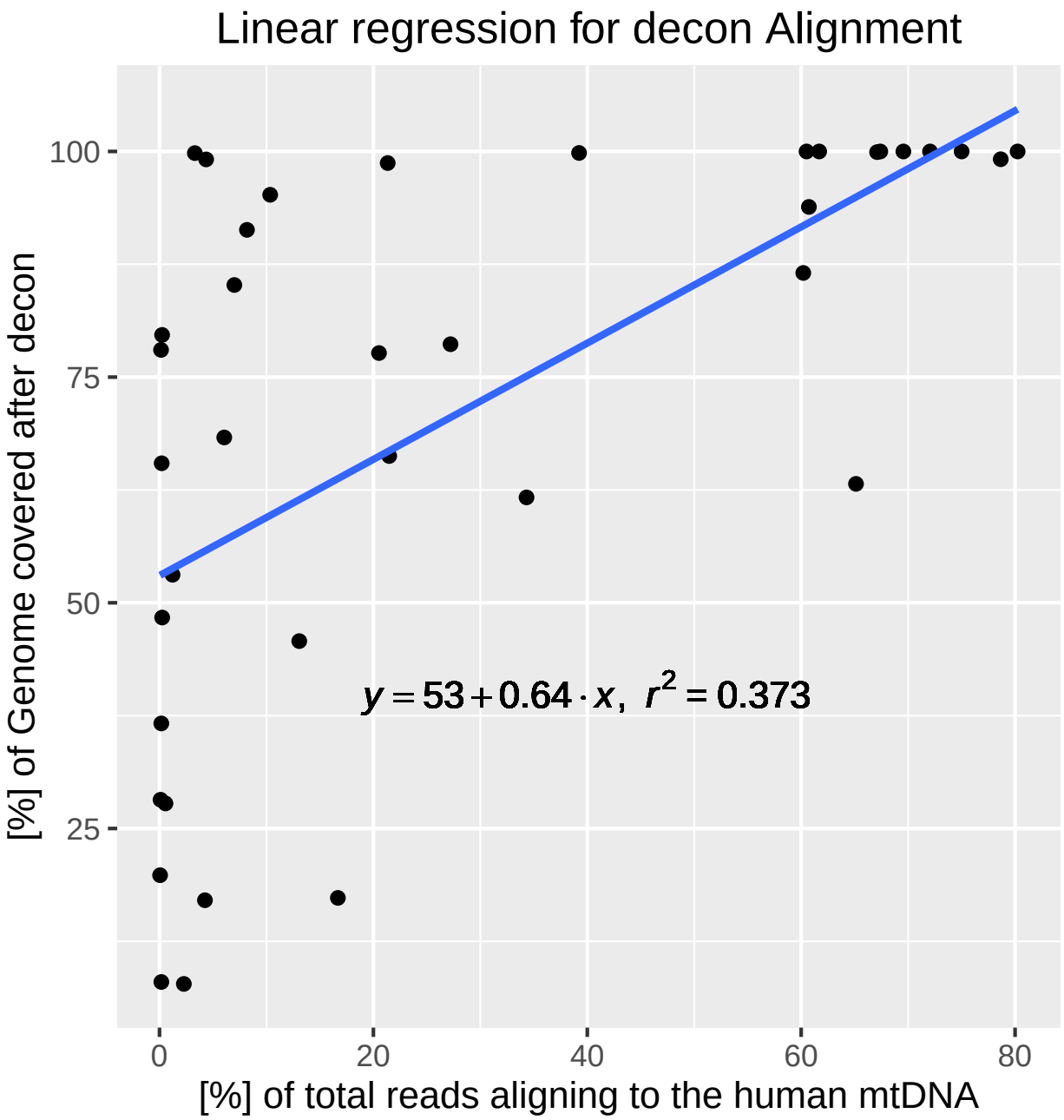



**Figure A.52:** *Showing linear regression for all 33 Samples seqeunced ; x axys = Percenatge of all reads including duplicates aligning to the human mtDNA(rCRS) ; y axys = Percentage of the human mtDNA(rCRS) that is covered in the decontaminated alignment; Residual standard error: 24.99 on 36 degrees of freedom; Multiple R-squared: 0.3731, Adjusted R-squared: 0.3557; F-statistic: 21.43 on 1 and 36 DF, p-value: 4.644e-05*

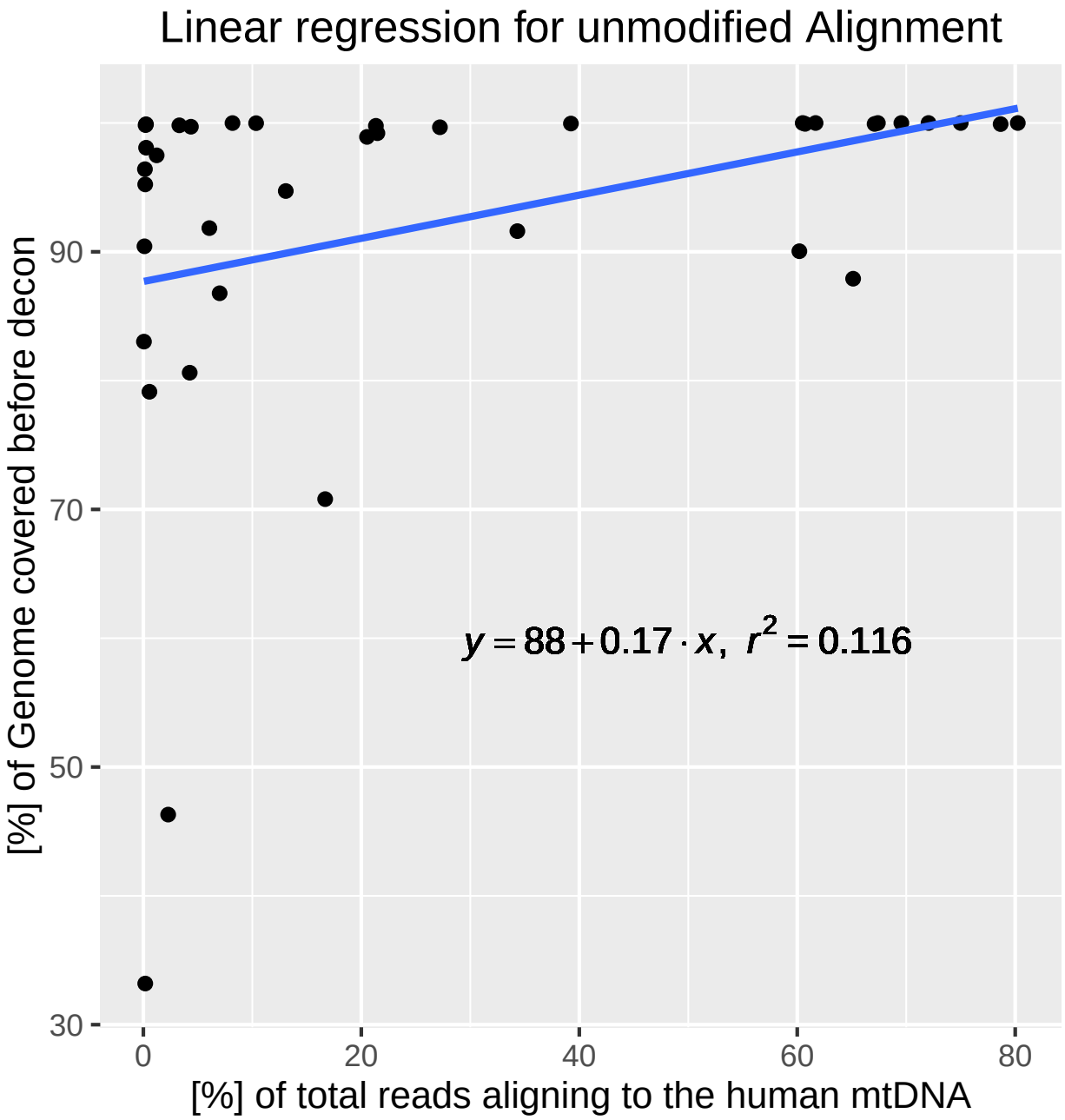



**Figure A.53:** *Showing linear regression for all 33 Samples seqeunced ; x axys = Percenatge of all reads including duplicates aligning to the human mtDNA(rCRS) ; y axys = Percentage of the human mtDNA(rCRS) that is covered in the NOT decontaminated alignment; Multiple R-squared: 0.1162, Adjusted R-squared: 0.09162, F-statistic: 4.732 on 1 and 36 DF, p-value: 0.03626*

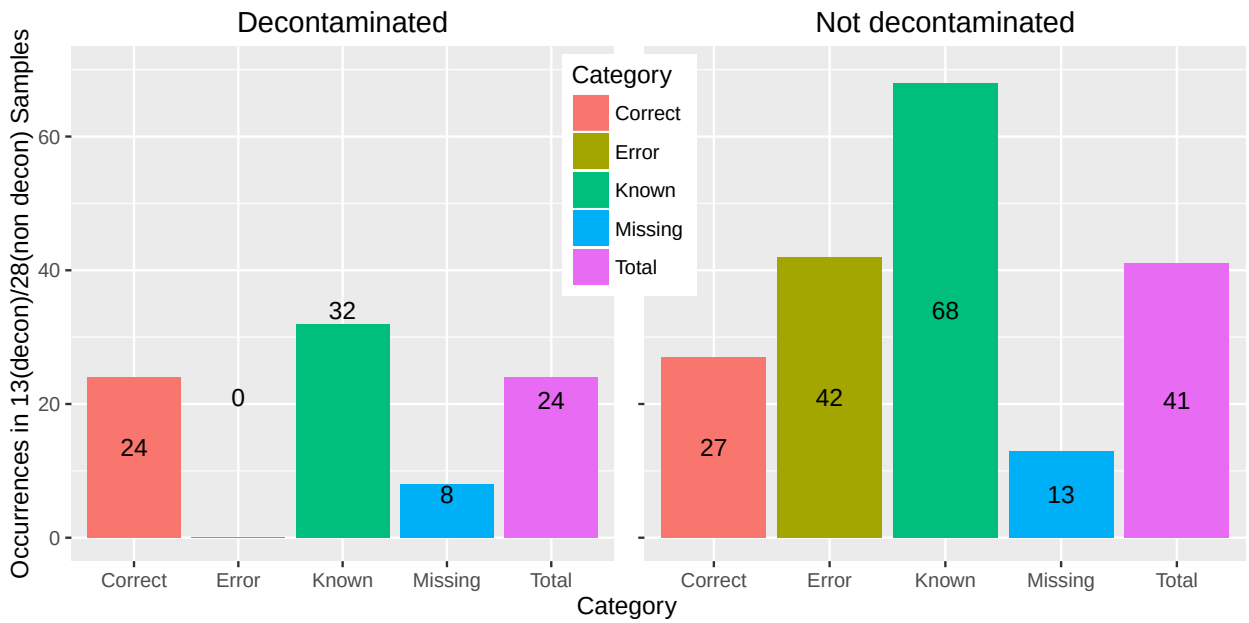



**Figure A.54:** *Showing results of SNP calling for the two different alignments of 13 Samples with 4x coverage after decontamination and 28 samples before decontamination; summarized in categories as described in methods*

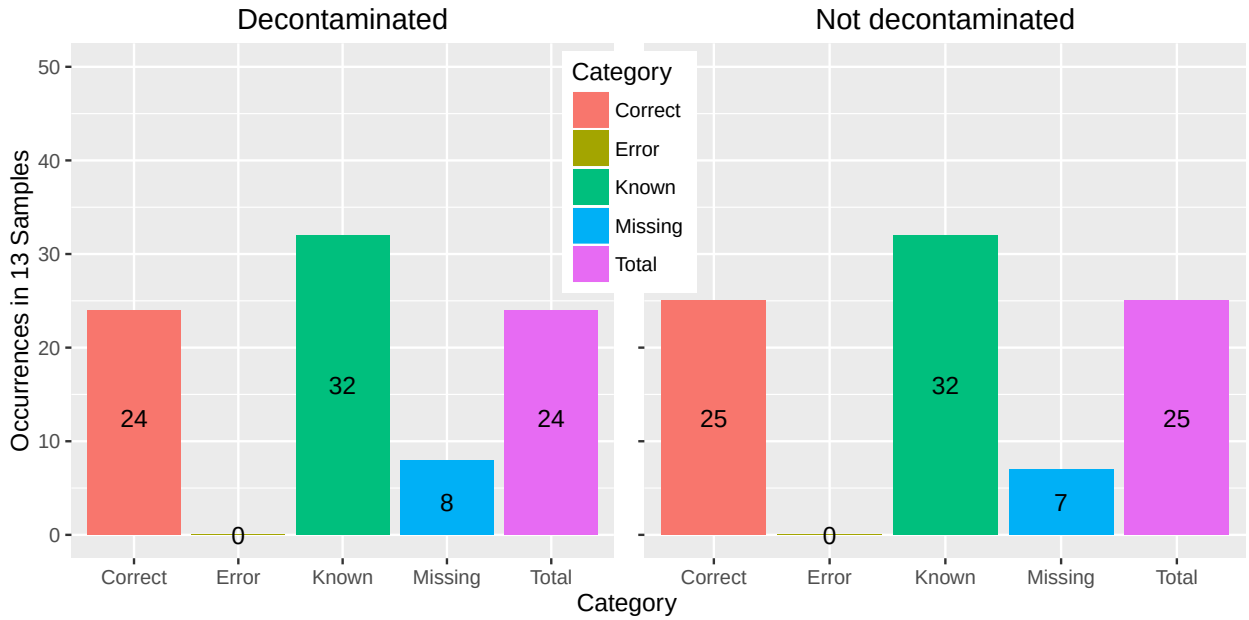



**Figure A.55:** *Showing results of SNP calling for the two different alignments of 13 Samples with 4x coverage after decontamination summarized in categories as described in methods*

## A.5 Supplementary files

Raw fastq files for Welzin are available on the archive drive of the Palaeogenetics group. Please refer to following path for the raw fastq files of the Welzin samples:
/uni-mainz.de/archiv/10Palaeogenetik/RAW_DATA/Sell/Weltzin

Final Alignments and Calls from the Welzin individuals, recalibrated alignments from the Aegean individuals and decon and nodecon alignments stored on an external hard drive deposited at the Palaeogenetics group. Please refer to the external hard drive submitted with this thesis.

Requests can be send to: Palaeogenetics Mainz <anthro2@uni-mainz.de>

**Table A.43:** *Supplementary data files available for download under https://seafile.rlp.net/d/2ac8729845/*

| File | Description |
|---|---|
| Ancient_wez15_24_35_39.pdf | Plot of top 15 F3 values for the given samples and Ancient populations |
| Ancient_wez40_48_51_53.pdf | Plot of top 15 F3 values for the given samples and Ancient populations |
| Ancient_wez58_59_61_64.pdf | Plot of top 15 F3 values for the given samples and Ancient populations |
| Ancient_wez71_74_83.pdf | Plot of top 15 F3 values for the given samples and Ancient populations |
| F4_WEZ_test_WEZ_Out.log | Log file with the results for all f4 trees tested with the each single Welzin individual |
| Modern_wez15_24_35_39.pdf | Plot of top 15 F3 values for the given samples and Modern populations |
| Modern_wez40_48_51_53.pdf | Plot of top 15 F3 values for the given samples and Modern populations |
| Modern_wez58_59_61_64.pdf | Plot of top 15 F3 values for the given samples and Modern populations |
| Modern_wez71_74_83.pdf | Plot of top 15 F3 values for the given samples and Modern populations |
| WEZ#.Ancient.F3.log | Log file for F3 statistics of a given Weltzin individual with Ancient populations |
| WEZ#.Modern.F3.log | Log file for F3 statistics of a given Weltzin individual with Modern populations |
| WEZ_Ancient_Modern.bed | Binary Plink file including all here used Individuals |
| WEZ_Ancient_Modern.bim | Binary Plink file including all here used Individuals |
| WEZ_Ancient_Modern.fam | Binary Plink file including all here used Individuals |
| pong_k2_small_all_modes.pdf | All modes for unsupervised ADMIXTURE analysis for the smaller subset |
| WEZ.relate | Output of lcMLkin |
| aDNAbamRecal.py | Python script to include PMD into base quality in bam files |
| remove_reads_no_mate.py | Python script to sort paired end sequences according to read pairs in fastq files |
| pseudoDIP_filter.py | Python script to filter and generate pseudodiploid genotypes from Allelepresence calls |
| filter_snps_bgz.py | Python script to filter GATK genrated VCF files including the <NON_REF> allele |
| recal_on_recal.sh | bash script to iterate the self recalibration with GATK |
| array_stats.py | python script to generate statistics from bam files |

# Bibliography


[1] Angsd contamination estimates. `http://popgen.dk/angsd/index.php/Contamination`. Accessed:2016-07-22.

[2] A set of command line tools (in java) for manipulating high-throughput sequencing (hts) data and formats such as sam/bam/cram and vcf. `http://broadinstitute.github.io/picard/`. Accessed:2016-07-22.

[3] D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664, 2009.

[4] D.H. Alexander, J. Novembre, K. Lange, and S. Shringarpure. `https://www.genetics.ucla.edu/software/admixture/admixture-manual.pdf`, 2015. Accessed:2017-02-8.

[5] Morten E Allentoft, Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B Damgaard, Hannes Schroeder, Torbjörn Ahlström, Lasse Vinner, et al. Population genomics of bronze age eurasia. *Nature*, 522(7555):167–172, 2015.

[6] Richard M Andrews, Iwona Kubacka, Patrick F Chinnery, Robert N Lightowlers, Douglass M Turnbull, and Neil Howell. Reanalysis and revision of the cambridge reference sequence for human mitochondrial dna. *Nature genetics*, 23(2):147–147, 1999.

[7] Leonardo Arbiza, Elaine Zhong, and Alon Keinan. Nre: a tool for exploring neutral loci in the human genome. *BMC bioinformatics*, 13(1):1, 2012.

[8] E Aronesty. Command-line tools for processing biological sequencing data, ea-utils. *Expression Analysis. Durham, NC: Available online at: http://code. google. com/p/ea-utils*, 2011.

[9] Aaron A. Behr, Katherine Z. Liu, Gracie Liu-Fang, Priyanka Nakka, and Sohini Ramachandran. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, 32(18):2817, 2016.

[10] 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.

[11] Babraham Bioinformatics. Fastq screen. *Babraham Bioinformatics*, 2013.

[12] Ruth Bollongino, Olaf Nehlich, Michael P Richards, Jörg Orschiedt, Mark G Thomas, Christian Sell, Zuzana Fajkošová, Adam Powell, and Joachim Burger. 2000 years of parallel societies in stone age central europe. *Science*, 342(6157):479–481, 2013.

[13] Barbara Bramanti, MG Thomas, Wolfgang Haak, M Unterlaender, P Jores, K Tambets, I Antanaitis-Jacobs, MN Haidle, R Jankauskas, C-J Kind, et al. Genetic discontinuity between local hunter-gatherers and central europess first farmers. *science*, 326(5949):137–140, 2009.

[14] Adrian W Briggs, Udo Stenzel, Philip LF 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.

[15] Farnaz Broushaki, Mark G Thomas, Vivian Link, Saioa López, Lucy van Dorp, Karola Kirsanow, Zuzana Hofmanová, Yoan Diekmann, Lara M. Cassidy, David Díez-del Molino, Athanasios Kousathanas, Christian Sell, Harry K. Robson, Rui Martiniano, Jens Blöcher, Amelie Scheu, Susanne Kreutzer, Ruth Bollongino, Dean Bobo, Hossein Davudi, Olivia Munoz, Mathias Currat, Kamyar Abdi, Fereidoun Biglari, Oliver E. Craig, Daniel G Bradley, Stephen Shennan, Krishna R Veeramah, Marjan Mashkour, Daniel Wegmann, Garrett Hellenthal, and Joachim Burger. Early neolithic genomes from the eastern fertile crescent. *Science*, 2016.

[16] George BJ Busby, Garrett Hellenthal, Francesco Montinaro, Sergio Tofanelli, Kazima Bulayeva, Igor Rudan, Tatijana Zemunik, Caroline Hayward, Draga Toncheva, Sena Karachanak-Yankova, et al. The role of recent admixture in forming the contemporary west eurasian genomic landscape. *Current Biology*, 25(19):2518–2526, 2015.

[17] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1):1, 2009.

[18] Nicholas Clarke, Travis Collier, Petr Danecek, Javier Herrero, Warren Kretzschmar, Heng Li, Shane McCarthy, John Marshall, Joel Martin, and Stephan Schiffels. Bcftools. `https://github.com/samtools/bcftools`. Accessed:2017-02-7.

[19] Peter JA 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.

[20] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[21] Andrew Curry. Slaughter at the bridge: Uncovering a colossal bronze age battle. `http://www.sciencemag.org/news/2016/03/slaughter-bridge-uncovering-colossal-bronze-age-battle`, 2016. Accessed:2017-02-17.

[22] Janusz Czebreszuk and Marzena Szmyt. The northeast frontier of bell beakers. In *Proceedings of the symposium held at the Adam Mickiewicz University, Poznań (Poland), May 26-29 2002 (= British Archaeological Reports, International Series*, volume 1155, 2003.

[23] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[24] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[25] Timothy K Earle. *Bronze Age economics: the beginnings of political economies*. Westview Pr, 2002.

[26] Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, 2007.

[27] Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M Slepchenko, Aleksei A Bondarev, Philip LF 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.

[28] Qiaomei Fu, Alissa Mittnik, Philip LF 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.

[29] Cristina Gamba, Eppie R Jones, Matthew D Teasdale, Russell L McLaughlin, Gloria Gonzalez-Fortes, Valeria Mattiangeli, László Domboróczki, Ivett Kővári, Ildikó Pap, Alexandra Anders, et al. Genome flux and stasis in a five millennium transect of european prehistory. *Nature communications*, 5, 2014.

[30] Marc García-Garcerà, 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.

[31] Aurelien Ginolhac, Morten Rasmussen, M Thomas P Gilbert, Eske Willerslev, and Ludovic Orlando. mapdamage: testing for damage patterns in ancient dna sequences. *Bioinformatics*, 27(15):2153–2155, 2011.

[32] Andreas Gnirke, Alexandre Melnikov, Jared Maguire, Peter Rogov, Emily M LeProust, William Brockman, Timothy Fennell, Georgia Giannoukos, Sheila Fisher, Carsten Russ, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature biotechnology*, 27(2):182–189, 2009.

[33] Richard E Green, Johannes Krause, Adrian W Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, et al. A draft sequence of the neandertal genome. *science*, 328(5979):710–722, 2010.

[34] Richard E Green, Anna-Sapfo Malaspinas, Johannes Krause, Adrian W Briggs, Philip LF 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.

[35] Ilan Gronau, Melissa J Hubisz, Brad Gulko, Charles G Danko, and Adam Siepel. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics*, 43(10):1031–1034, 2011.

[36] Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson, et al. Massive migration from the steppe was a source for indo-european languages in europe. *Nature*, 522(7555):207–211, 2015.

[37] Garrett Hellenthal, George BJ Busby, Gavin Band, James F Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, 2014.

[38] Volker Heyd. Families, prestige goods, warriors & complex societies: Beaker groups of the 3rd millennium cal bc along the upper & middle danube. In *Proceedings of the Prehistoric Society*, volume 73, pages 327–379. Cambridge Univ Press, 2007.

[39] Russell Higuchi, Barbara Bowman, Mary Freiberger, Oliver A Ryder, and Allan C Wilson. Dna sequences from the quagga, an extinct member of the horse family. 1984.

[40] Zuzana Hofmanová, Susanne Kreutzer, Garrett Hellenthal, Christian Sell, Yoan Diekmann, David Díez-del Molino, Lucy van Dorp, Saioa López, Athanasios Kousathanas, Vivian Link, et al. Early farmers from across europe directly descended from neolithic aegeans. *Proceedings of the National Academy of Sciences*, page 201523951, 2016.

[41] Michael Hofreiter, Johanna LA Paijmans, Helen Goodchild, Camilla F Speller, Axel Barlow, Gloria G Fortes, Jessica A Thomas, Arne Ludwig, and Matthew J Collins. The future of ancient dna: Technical advances and conceptual shifts. *BioEssays*, 37(3):284–293, 2015.

[42] Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. Megan analysis of metagenomic data. *Genome research*, 17(3):377–386, 2007.

[43] Sohyun Hwang, Eiru Kim, Insuk Lee, and Edward M Marcotte. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific reports*, 5:17875, 2015.

[44] Broad Institute. Gatk base quality score recalibration without known snps. `http://gatkforums.broadinstitute.org/gatk/discussion/44/base-quality-score-recalibration-bqsr`. Accessed:2016-07-22.

[45] Broad Institute. Gatk methods and algorithms; genotype refinement workflow:mathematical details. `https://software.broadinstitute.org/gatk/documentation/topic?name=method`. Accessed:2016-07-22.

[46] Broad Institute. Gatk methods and algorithms; hc step 4: Assigning per-sample genotypes. `https://software.broadinstitute.org/gatk/documentation/topic?name=method`. Accessed:2016-07-22.

[47] Broad Institute. Gatk methods and algorithms; how the haplotypecaller's reference confidence model works. `https://software.broadinstitute.org/gatk/documentation/topic?name=method`. Accessed:2016-07-22.

[48] Detlef Jantzen, Ute Brinker, Jörg Orschiedt, Jan Heinemeier, Jürgen Piek, Karlheinz Hauenstein, Joachim Krüger, Gundula Lidke, Harald Lübke, Reinhard Lampe, et al. A bronze age battlefield? weapons and trauma in the tollense valley, north-eastern germany. *Antiquity*, 85(328):417–433, 2011.

[49] Eppie R Jones, Gloria Gonzalez-Fortes, Sarah Connell, Veronika Siska, Anders Eriksson, Rui Martiniano, Russell L McLaughlin, Marcos Gallego Llorente, Lara M Cassidy, Cristina Gamba, et al. Upper palaeolithic genomes reveal deep roots of modern eurasians. *Nature communications*, 6, 2015.

[50] Hákon Jónsson, Aurélien Ginolhac, Mikkel Schubert, Philip LF Johnson, and Ludovic Orlando. mapdamage2. 0: fast approximate bayesian estimates of ancient dna damage parameters. *Bioinformatics*, page btt193, 2013.

[51] Donna Karolchik, Angela S Hinrichs, Terrence S Furey, Krishna M Roskin, Charles W Sugnet, David Haussler, and W James Kent. The ucsc table browser data retrieval tool. *Nucleic acids research*, 32(suppl 1):D493–D496, 2004.

[52] 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.

[53] Keinanlab. Nre: Neutral region explore; webtool. `http://nre.cb.bscb.cornell.edu/nre/index.html`. Accessed:2016-08-10.

[54] Andreas Keller, Angela Graefen, Markus Ball, Mark Matzas, Valesca Boisguerin, Frank Maixner, Petra Leidinger, Christina Backes, Rabab Khairat, Michael Forster, et al. New insights into the tyrolean iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature communications*, 3:698, 2012.

[55] Rabab Khairat, Markus Ball, Chun-Chi Hsieh Chang, Raffaella Bianucci, Andreas G Nerlich, Martin Trautmann, Somaia Ismail, Gamila ML Shanab, Amr M Karim, Yehia Z Gad, et al. First insights into the metagenome of egyptian mummies using next-generation sequencing. *Journal of applied genetics*, 54(3):309–325, 2013.

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

[57] Martin Kircher, Susanna Sawyer, and Matthias Meyer. Double indexing overcomes inaccuracies in multiplex sequencing on the illumina platform. *Nucleic acids research*, page gkr771, 2011.

[58] Karola Kirsanow and Joachim Burger. Ancient human dna. *Annals of Anatomy-Anatomischer Anzeiger*, 194(1):121–132, 2012.

[59] Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. Angsd: analysis of next generation sequencing data. *BMC bioinformatics*, 15(1):1, 2014.

[60] Athanasios Kousathanas, Christoph Leuenberger, Vivian Link, Christian Sell, Joachim Burger, and Daniel Wegmann. Inferring heterozygosity from ancient and low coverage genomes. *Genetics*, 205(1):317–332, 2017.

[61] KV Kremenetski. Steppe and forest-steppe belt of eurasia: Holocene environmental history. *Prehistoric steppe adaptation and the horse*, pages 11–28, 2003.

[62] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[63] Iosif Lazaridis, Dani Nadel, Gary Rollefson, Deborah C Merrett, Nadin Rohland, Swapan Mallick, Daniel Fernandes, Mario Novak, Beatriz Gamarra, Kendra Sirak, et al. Genomic insights into the origin of farming in the ancient near east. *Nature*, 536(7617):419–424, 2016.

[64] 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, 2014.

[65] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[66] 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.

[67] Mingkun Li and Mark Stoneking. A new approach for detecting low-level mutations in next-generation sequence data. *Genome biology*, 13(5):1, 2012.

[68] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[69] Stinus Lindgreen. Adapterremoval: easy cleaning of next-generation sequencing reads. *BMC research notes*, 5(1):1, 2012.

[70] Vivian Link. Atlas: Analysis tools for low-depth and ancient samples; homepage. `https://bitbucket.org/phaentu/atlas/wiki/Home`. Accessed:2017-06-02.

[71] Vivian Link, Athanasios Kousathanas, Krishna Veeramah, Christian Sell, Amelie Scheu, and Daniel Wegmann. Atlas: Analysis tools for low-depth and ancient samples. *bioRxiv*, 2017.

[72] Mikhail Lipatov, Komal Sanjeev, Rob Patro, and Krishna Veeramah. Maximum likelihood estimation of biological relatedness from low coverage sequencing data. *bioRxiv*, 2015.

[73] Mikhail Lipatov, Komal Sanjeev, Rob Patro, and Krishna Veeramah. Maximum likelihood estimation of biological relatedness; program page. `https://github.com/COMBINE-lab/maximum-likelihood-relatedness-estimation`, 2015. Accessed:2017-02-16.

[74] M Gallego Llorente, ER Jones, A Eriksson, V Siska, KW Arthur, JW Arthur, MC Curtis, JT Stock, M Coltorti, P Pieruccini, et al. Ancient ethiopian genome reveals extensive eurasian admixture in eastern africa. *Science*, 350(6262):820–822, 2015.

[75] Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernandes, Mario Novak, et al. Genome-wide patterns of selection in 230 ancient eurasians. *Nature*, 528(7583):499–503, 2015.

[76] 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.

[77] Graham McVicker, David Gordon, Colleen Davis, and Phil Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet*, 5(5):e1000471, 2009.

[78] Michael L Metzker. Sequencing technologies-the next generation. *Nature reviews genetics*, 11(1):31–46, 2010.

[79] Matthias Meyer and Martin Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*, 2010(6):t5448, 2010.

[80] 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.

[81] MYcroarray. Mycorarray: Mybaits user manual v1. `http://nre.cb.bscb.cornell.edu/nre/index.html`. used on baits generated before 2014.

[82] Resource Coordinators NCBI. Database resources of the national center for biotechnology information. *Nucleic acids research*, 41(Database issue):D8, 2013.

[83] 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. *Annu. Rev. Genet.*, 38:645–679, 2004.

[84] Galyna Pashkevych. Environment and economic activities of neolithic and bronze age populations of the northern pontic area. *Quaternary International*, 261:176–182, 2012.

[85] Nick Patterson. Admixtools 4.1. `https://github.com/DReichLab/AdmixTools`. Accessed:2017-02-7.

[86] Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

[87] Alexander Peltzer, Günter Jäger, Alexander Herbig, Alexander Seitz, Christian Kniep, Johannes Krause, and Kay Nieselt. Eager: efficient ancient genome reconstruction. *Genome biology*, 17(1):1, 2016.

[88] Hendrik N Poinar, Carsten Schwarz, Ji Qi, Beth Shapiro, Ross DE 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.

[89] D Pokutta. Population dynamics, diet and migrations of the únětice culture in poland. *Gothenburg: Gothenburg University*, 2013.

[90] Joachim Preuß. Das neolithikum in mitteleuropa. *Kulturen–Wirtschaft–Umwelt vom*, 6:1–3, 1998.

[91] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[92] 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.

[93] Kim D Pruitt, Tatiana Tatusova, Garth R Brown, and Donna R Maglott. Ncbi reference sequences (refseq): current status, new features and genome annotation policy. *Nucleic acids research*, 40(D1):D130–D135, 2012.

[94] Shaun Purcell. Plink 1.9. `https://www.cog-genomics.org/plink2/`. Accessed:2017-02-7.

[95] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[96] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[97] Morten Rasmussen, Yingrui Li, Stinus Lindgreen, Jakob Skou Pedersen, Anders Albrechtsen, Ida Moltke, Mait Metspalu, Ene Metspalu, Toomas Kivisild, Ramneek Gupta, et al. Ancient human genome sequence of an extinct palaeo-eskimo. *Nature*, 463(7282):757–762, 2010.

[98] David Reich, Richard E Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y Durand, Bence Viola, Adrian W Briggs, Udo Stenzel, Philip LF Johnson, et al. Genetic history of an archaic hominin group from denisova cave in siberia. *Nature*, 468(7327):1053–1060, 2010.

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

[100] Juan J Sanchez, Claus Børsting, Charlotte Hallenberg, Anders Buchard, Alexis Hernandez, and Niels Morling. Multiplex pcr and minisequencing of snps a model with 35 y chromosome snps. *Forensic Science International*, 137(1):74–84, 2003.

[101] Akira Sassa, Yuki Kanemaru, Nagisa Kamoshita, Masamitsu Honma, and Manabu Yasui. Mutagenic consequences of cytosine alterations site-specifically embedded in the human genome. *Genes and Environment*, 38(1):17, 2016.

[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, Luca Ermini, Clio Der Sarkissian, Hákon Jónsson, Aurélien Ginolhac, Robert Schaefer, Michael D Martin, Ruth Fernández, Martin Kircher, Molly McCue, et al. Characterization of ancient and modern genomes by snp detection and phylogenomic and metagenomic analysis using paleomix. *Nat Protoc*, 9(5):1056–1082, 2014.

[104] Mikkel Schubert, Aurelien Ginolhac, Stinus Lindgreen, John F Thompson, Khaled AS Al-Rasheid, Eske Willerslev, Anders Krogh, and Ludovic Orlando. Improving ancient dna read mapping against modern reference genomes. *BMC genomics*, 13(1):1, 2012.

[105] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

[106] Pontus Skoglund, Helena Malmström, Maanasa Raghavan, Jan Storå, Per Hall, Eske Willerslev, M Thomas P Gilbert, Anders Götherström, and Mattias Jakobsson. Origins and genetic legacy of neolithic farmers and hunter-gatherers in europe. *Science*, 336(6080):466–469, 2012.

[107] Pontus Skoglund, Jan Storå, Anders Götherström, and Mattias Jakobsson. Accurate sex identification of ancient human remains using dna shotgun sequencing. *Journal of Archaeological Science*, 40(12):4477–4482, 2013.

[108] AFA Smit, R Hubley, and P Green. Repeatmasker open-4.0. 2013–2015. *URL http://www. repeatmasker. org*, 2014.

[109] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. *From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline*. John Wiley and Sons, Inc., 2013.

[110] Marc Vander Linden. For equalities are plural: reassessing the social in europe during the third millennium bc. *World archaeology*, 39(2):177–193, 2007.

[111] Chaolong Wang and Xiaowei Zhan. Laser. `http://genome.sph.umich.edu/wiki/LASER`. Accessed:2016-08-06.

[112] Chaolong Wang and Xiaowei Zhan. Laser: Locating ancestry from seequence reads, 2.03 edition. `http://csg.sph.umich.edu//chaolong/LASER/LASER_Manual.pdf`. Accessed:2016-08-06.

[113] Chaolong Wang, Xiaowei Zhan, Jennifer Bragg-Gresham, Hyun Min Kang, Dwight Stambolian, Emily Y Chew, Kari E Branham, John Heckenlively, The FUSION Study, Robert Fulton, et al. Ancestry estimation and control of population stratification for sequence-based association studies. *Nature genetics*, 46(4):409–415, 2014.

[114] Chaolong Wang, Xiaowei Zhan, Jennifer Bragg-Gresham, Hyun Min Kang, Dwight Stambolian, Emily Y Chew, Kari E Branham, John Heckenlively, The FUSION Study, Robert Fulton, et al. Ancestry estimation and control of population stratification for sequence-based association studies. *Nature genetics*, 46(4):409–415, 2014.

[115] Jean Susorney Wellington. Ancient europe 8000 bc-ad 1000: Encyclopedia of the barbarian world. *Reference & User Services Quarterly*, 44(1):81–83, 2004.