Indels and large scale variation in archaic hominins compared to present day humans

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet
Informatik

Vorgelegt

von Master of Technology **Manjusha Chintalapati**
geboren am 14. August 1988 in Machilipatnam, India

Die Annahme der Dissertation wurde empfohlen von:
1. Prof. Dr. Peter F. Stadler, Universität Leipzig
2. Dr. Kay Prüfer, Max Planck Institute for Evolutionary Anthropology

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 06.12.2018 mit dem Gesamtprädikat *magna cum laude*.

“And, when you want something, all the universe conspires in helping you to achieve it.”

— Paulo Coelho

Chintalapati Manjusha
Indels and large scale variation in archaics and present day humans
Max-Planck Institute for Evolutionary Genetics, Dissertation 2018
134 pages, 125 references, 26 Figures, 17 tables

Acknowledgements

I am in deeply grateful to my PhD supervisor Dr. Kay Prüfer, who provided me the wonderful opportunity to work at Max-Planck Institute, for his support, patience, time and guidance which made this dissertation possible. I am eternally thankful to him. I would like to thank my co supervisor Janet Kelso, who is a constant source of inspiration and Svante Pääbo for providing me the opportunity to carry out my PhD at MPI. I would like to thank my buddy and my colleague in Kay's group, Stephane who helped me in many possible ways and made my life easier in times of difficulty. I would like to thank my entire gang at MPI: Mike, Martin, Fabrizio, Mateja, Alex, Viviane, Amin, Homa, Petra, Elena, and many more for making MPI a pleasant place to work and making it my second home. It has been a pleasure to have you all in my life. I would like to thank the Bioinformatics group at MPI for their constant input and mentoring in every Monday meeting.

I would like to express gratitude to my mom Venkata Lakshmi and my grandmother Vijaya Lakshmi who are my role models. I am grateful to my family- my mother in law Sujatha, my brothers Srikanth, Ravi Shankar, Ravindra Reddy, my only god given sister Divya, my family in Europe Shruti, Sahi and my lovely nephews Ishan, Vihan and Kathyayan for making my life happy and pleasant.

Last but not the least I would like to thank my advisor Peter Stadler for his input in my PhD projects and wonderful retreat trips. During my study I knew many nice people at Bioinf- Marible, Steve, Katja, Heni, Bia, Irena, thank you all for making me part of your life.

In India we say "Behind every successful man there is a woman" but in my case I would rather say "Behind a successful woman there are many men". My heroes in my life are my father Markandeya Sastry Chintalapati, my Father in law Suresh Babu Vattakandiyil and the love of my life Sree Rohit Raj Kolora without whom this work would have never happened. As a token of gratitude I dedicate my work to my heroes.

Summary

The study of ancient DNA provided us with the genomes of our closest hominin relatives: Neandertals and Denisovans. One of the major findings from the analysis of these archaic genomes was that 2% of the genomes of present-day non-Africans originate from admixture with Neandertals. Further investigation suggested that some of the introgressed regions underlie both advantageous and disadvantageous traits.

Since single nucleotide polymorphisms (SNPs) constitute the largest class of variation in the genomes of humans, most previous studies focused on using SNPs for genetic analysis. However, other mutations such as insertion/deletion variants (indels) or structural variation have a larger potential to affect phenotype or cause disease. These non-SNP classes of mutations are often excluded from evolutionary studies since they are more difficult to detect. In this dissertation, I aimed to study indels and rearrangements in archaic genomes compared to present day humans.

In the first part of my thesis I focus on small indels (1-5 base pairs in length) on the human lineage. The archaic genome of the Altai Neandertal allows mutations to be classified into those occurring before the split of humans and Neandertals, those that occur after the split, i.e. those specific to modern humans, and those introgressed into modern humans from Neandertals. Using these three datasets, I studied the evolutionary forces acting on deletion and insertion events. I found that deletions are, on average, more deleterious than insertions. Furthermore, introgressed variants appear to be less deleterious than modern human specific variants, suggesting that negative selection removed a larger proportion of deleterious variants either before or after introgression. Despite this evidence for stronger selection, some introgressed variants may still contribute to modern human phenotypic diversity, and I discovered one such introgressed indels which is associated with the time to menarche in humans.

In the second part of my thesis, I studied large scale genomic structural variation in archaics compared to humans. Existing methods to identify rearrangements in ancient genomes use read alignments to a reference genome. I took an alternative approach: *de novo* assembling the archaic genomes to re-construct pieces of contiguous genomic sequence (contigs) and inferring rearrangements from discontinuous alignments to the human reference genome. I identified four different types of rearrangements from these alignments: deletions, insertion, duplication and inversions. The identified rearrangements were further classified into human derived, Neandertal derived or ancestral events. The rearrangements overlapping exons are catalogued which could be a resource for functional testing. This study also yielded contigs that are on average more than 10 times longer than reads and allow for more of the Neandertal genome to be reconstructed confidently.

In summary, I analyzed non-SNP mutations, encompassing small indels and large genomic structural variation, in archaic and human genomes. My study resulted in a collection of mutation events that may underlie some of the phenotypic differences observed between archaic and modern humans and provide a starting point for further investigation.

Contents

List of Figures

List of Tables

List of Abbreviations

DNA	Deoxyribonucleic acid
SNP	Single Nucleotide polymorphism
Indel	Insertion or deletion
PCR	Polymerase chain reaction
BWA	Burrows wheeler alignment
GATK	Genome Analysis ToolKit
k-mer	strings of read of length “k”
SVs	Structural Variations
ORF	Open Reading Frame
NAHR	Non-Allelic Homologous recombination
NHEJ	Non homologous end joining
FoSTeS	Fork stalling and template switching
UCSC genome browser	University of California Santa Cruz genome browser
MK Test	McDonald-Kreitman test
AFS	Allele frequency spectrum
VEP	Variant effect predictor
CADD	Combined Annotation Dependent Depletion tool
GWAS	Genome-Wide Association Studies
rDI	ratio of Deletions to Insertions
FDR	False Discovery Rate
UDG	uracil DNA glycosylase
DBG	De Bruijin Graph
OLC	Overlap layout consensus
SOAPdenovo	Short Oligonucleotide Analysis Package *de novo*
NCBI	National center for Biotechnology information
SAMTOOLS	Sequence Alignment/Map tools
LINE	Long Interspersed Nuclear Elements
SINE	Short Interspersed Nuclear Elements
IGV	Integrative Genomics Viewer

Chapter 1 Introduction

The central question in evolutionary biology over the past decades has been "what makes us human". Advancements in evolutionary genomics through the advent of high throughput sequencing facilitated the generation of whole genome sequences of many species including the human genome. Furthermore, the in-depth study of human evolution took a leap forward through advances in the field of ancient DNA genomics, which availed us with archaic genomes of Neandertals and Denisovan genomes. Previous studies on human evolution were based on identifying changes in the human genome in comparison to the closest primate genomes, which yields a list of mutations accumulated over past 6 million years. The archaic (Neandertal and Denisovan) genomes allow these mutations to be further categorized into those that are unique to present-day humans and those that are older.

Most genetic analysis comparing archaic genomes and human genomes are performed using single nucleotide polymorphisms since these types of changes are the most abundant class of mutations. Here, I extend these genetic analyses to other classes of mutations such as small insertions or small deletions also termed "indels" and genomic rearrangements which are large scale changes in the genome encompassing insertions, deletions, duplications, inversions or translocations.

Although indels are less abundant in the genome, they can have significant functional impact and may be over represented among variants that are associated with disease risk. In spite of their functional potential, most evolutionary studies use single nucleotide polymorphisms (SNPs) and avoid indels, in part due to the difficulty in genotyping and identifying indels. In my first study, I identified small indels arising on the human lineage and used the Neandertal genome to classify indels into those that are shared between humans and Neandertals and those indels that are specific to present-day humans. Using this dataset, I gained insight

into the selection pressures that affect indels and studied indels which are potentially introgressed from Neandertals.

In the second part of my study, I studied large scale structural variation in archaic genomes by comparing these genomes to the human reference genome. The available read based approaches for the identification of rearrangements rely on using paired end sequencing data or split read alignments to a reference. However, these methods are limited since ancient sequences are often too short and damaged over time. Therefore, I *de novo* assembled the archaic genomes to construct longer sequences called "contigs" after error correction of ancient DNA damage and used these contigs to identify rearrangements between archaics and humans.

In summary, this dissertation encompasses a study of genomic structural variation ranging from small indels to large scale rearrangements between humans and archaics.

Chapter 2 Background

2.1 DNA Sequencing

2.1.1 From Sanger to Next Generation Sequencing Technologies

DNA (DeoxyriboNucleic Acid) is composed of two chains of nucleotides carrying the bases Adenine (A), Guanine (G), Cytosine (C) and Tyrosine (T), called strands, that are connected by hydrogen bonds between bases (A with T; and C with G). The order in which the bases occur on a strand of DNA can code for information that is required for the survival, functioning and reproduction of any living cell. A part of this information is required to synthesize proteins, whereas other information is required to regulate the abundance of synthesized proteins and for other regulatory purposes. The process of determining the exact order of these nucleotides is called DNA sequencing.

DNA sequencing uses a type of enzyme called "DNA polymerase" which synthesizes a new strand of DNA using deoxyribonucleotides and a reference DNA template. This reference DNA template consists of a double stranded part from which synthesis starts and a single stranded part for which the second strand is synthesized in the reaction. The DNA polymerases used *in vitro* are often extracted from different organisms (Miura et al. 2013). One of the first attempts to sequence DNA without any prior knowledge of the region to be sequenced was using radioactive labelled nucleotides called "Maxam-Gilbert sequencing" (Maxam and Gilbert 1977) developed by Allan Maxam and Walter Gilbert, which was soon replaced by "chain termination method" developed by Sanger and colleagues since it involved less toxic chemicals (Sanger et al. 1977). With the advent of other techniques such as capillary electrophoresis and fluorescent labelling this method was used to develop a first generation of sequencing machines that could be loaded with specifically prepared DNA and yielded read-outs of sequences as long as ~1000bp, often for hundreds of sequences in parallel.

It was this technology which allowed for the first full human genome to be sequenced in 2001 (Lander et al. 2001).

In the early 1980's, the polymerase chain reaction (PCR) was developed, that allows a region of interest on a DNA strand to be copied several times. This reaction proceeds by heat-denaturing double stranded DNA into single strands, attaching primers (small synthesized single stranded DNA molecules) to the DNA strands and then filling in the remaining strand by using a DNA polymerase (Mullis 1990). This technique helps targeted amplification and sequencing.

DNA sequencing took a leap forward in the mid-2000's when a new generation of sequencers became available. These "next generation" sequencers produced shorter reads of sequences at a much higher throughput. These high throughput sequencing technologies were referred to as "Next-generation sequencing" (NGS) technologies (Reuter et al. 2015). Some of the technologies which sequenced short reads are ion-torrent, 454 pyrosequencing (Margulies et al. 2005), Ion torrent (Rothberg et al. 2011), SOLiD sequencing (Shendure et al. 2005) and Illumina (Solexa) sequencing (Bentley et al. 2008). As these short read sequencing technologies operated at much lower cost per sequenced base pair, it was possible to correct errors in the reads by sequencing regions or fragments multiple times.

2.1.2 Illumina Sequencing

Illumina Sequencing technology uses a method called "reversible terminated chemistry" invented by Bruno Canard and Simon Sarfati at the Pasteur Institute in Paris and was implemented by Shankar Balasubramanian and David Klenerman of Cambridge University to develop into a sequencing technology. The DNA is first fragmented into smaller pieces. Each DNA fragment is ligated with a pair of synthesized oligonucleotides, called adaptors, on either side of the molecule along with primers, forming a DNA sequencing library. The DNA library molecules are attached to a sequencing plate or "flow cell" coated with primers and these DNA molecules are copied several times by running the polymerase chain reaction, which amplifies the DNA molecules to form "DNA

clusters" of identical molecules in a small area around the original library molecule. The amplified DNA is sequenced with DNA polymerase using fluorescently labelled reversible terminator bases (RT-bases). The RT bases when incorporated during sequencing release fluorescence and inhibit the addition of other bases. The fluorescence emitted by the binding of the RT bases is captured by a camera and the resulting images are processed to determine the sequence of DNA. The flow cell is then washed for the next cycle of RT-bases. This method is also referred to as "sequencing by synthesis".

2.1.3 Paired End Sequencing

Sequences generated with second generation technologies, but also those generated by older technologies, are much shorter than the length of a human chromosome (>50Mb). To gain information over longer distances, the next generation sequencing technologies implemented an approach whereby the ends of a larger DNA molecules are sequenced (Figure 2.1). The DNA sequence which is attached to adaptors is sequenced by hybridizing primers to adaptors which initiate DNA sequencing and thus sequence the whole fragment. In paired end sequencing, this process takes place from both ends of the DNA molecule resulting in two read pairs. The paired reads are separated by a known length of sequence known as insert size.

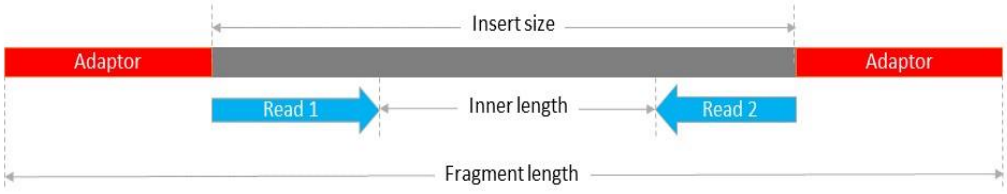


Figure 2.1 Terminology in paired end sequencing

2.2 Ancient DNA

DNA that is preserved in the remains of a long dead individuals is called Ancient DNA. The extraction of DNA from fossils or dead tissue is a complex process as the DNA present in the fossil is degraded over time. Over the last decade continuous development of library preparation protocols and extraction methods have maximized the DNA content that can be recovered from ancient remains while minimizing contamination by modern DNA (Rohland and Hofreiter 2007; Kuch and Poinar 2012; Gansauge and Meyer 2013).

2.2.1 Sequencing Library Preparation Protocol for Ancient DNA

A double stranded library preparation method involves attaching adapters to the ends of a double stranded molecule after single stranded overhangs, which naturally occur when the DNA deteriorates over time, are repaired to form a double stranded molecule with the same number of bases on each strand (also referred to as "blunt end"). The ancient DNA we work with is damaged and degraded over time resulting in depletion of endogenous DNA (DNA coming from the fossil) and the DNA often acquires single stranded overhangs due to double stranded breaks. In a standard double stranded library preparation protocol the single-stranded overhangs of an ancient DNA molecule are repaired and the DNA molecule is sequenced. However, in a single stranded protocol, the DNA is denatured in the first step producing single stranded molecules which are then library prepared for sequencing (Gansauge and Meyer 2013). The single stranded molecules are attached to adaptors with biotin, a substance which attaches to a substrate called streptavidin, thereby immobilizing DNA molecules. A primer is added to hybridize to the adaptor and the DNA molecule is copied. The resulting copied double stranded molecules are eluted out and sequenced using Illumina technology. This immobilization of DNA on to streptavidin reduces loss of information.

Treating ancient DNA with a single stranded library preparation protocol instead of a double stranded protocol has the advantage that a larger proportion

of molecules yields information, since a sequencing library molecule can start from each of the two single strands of an ancient DNA molecule. In addition, this protocol retains the orientation of nucleotide substitutions originating from ancient DNA damage (see next section).

Independent of the library preparation protocol, additional synthetic oligomers are attached to DNA molecules called "indices" which help to differentiate between different samples when sequenced simultaneously and to detect contaminating molecules from other sequencing libraries (Figure 2.2).

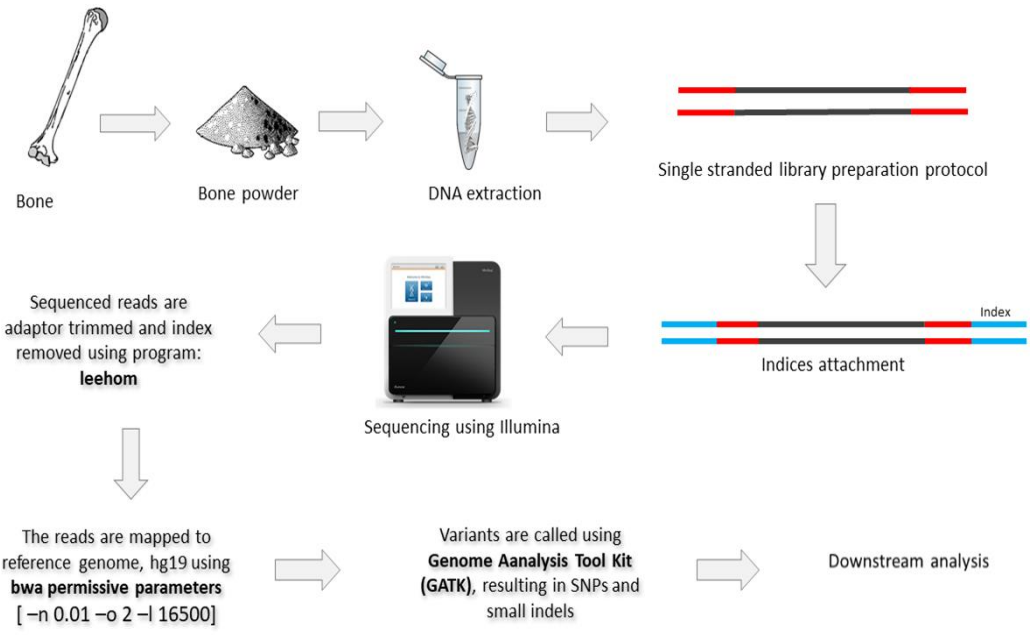


Figure 2.2 Flowchart of data processing of ancient DNA in-house

2.2.2 Characteristics of Ancient DNA

Ancient DNA differs in its features from present day DNA as a result of the degradation of the DNA over long periods of time. The three main features of ancient DNA are a short fragment size, base damage and contamination from external sources such as bacteria.

2.2.2.1 Fragmented DNA

DNA extracted from an ancient sample is fragmented due to a lack of repair mechanisms that prevent the degradation of DNA in living organisms. Breaks in DNA strands are often affecting one strand, leaving single stranded overhangs in the resulting DNA fragments. The average length of ancient DNA varies between samples but tends to be shorter for older samples (Allentoft et al. 2012). Environmental factors such as temperature and humidity also play a role in DNA preservation (Schwarz et al. 2009) and this varies between different fossil types (Hansen et al. 2017).

2.2.2.2 Damaged Bases

The bases in the DNA of any organism are susceptible to oxidation, hydrolysis or deamination, leading to damaged bases. While these bases are reverted back by DNA repair mechanisms in living cells, these damages start to accumulate in ancient DNA after the death of the organism. One of the most common types of base damage observed in ancient DNA is deamination of cytosine leading to uracil which accumulates approximately two orders faster in single-stranded compared to double stranded DNA, so that the error is most prevalent at the ends of ancient DNA fragments and leads to C to T exchanges after sequencing of libraries prepared with single-stranded protocols and C to T as well as G to A exchanges for libraries prepared with double-stranded protocols (Briggs et al. 2007; Dabney et al. 2013).

2.2.2.3 Contamination

Two types of contamination occur for ancient samples and can influence downstream analysis: contamination by modern human DNA, caused for instance by the handling of the bones by excavators and researchers, and contamination from environmental sequences, mostly from microbes. The contamination from bacteria and other fauna can be excluded from further analyses by mapping DNA sequences to a close reference genome. However, modern human contamination cannot easily be distinguished from the original DNA of ancient modern and archaic humans. Hence methods have been developed to minimize the chance of

contamination during excavation and laboratory procedures, and to quantify the presence of contamination. True ancient DNA fragments are expected to show damage at the ends of molecules while present-day contaminating sequences are not expected to show such changes at high frequency. Consequently, the presence of uracils at the ends of DNA fragments have been used to exclude present-day contamination through laboratory procedures. Ancient DNA sequences can also be authenticated *in silico* by selecting for those sequences that contain C to T exchanges compared to a close reference genome.

2.3 Sequencing and Processing of Ancient DNA Sequences

2.3.1 Archaic Genomes

Neandertals are the closest extinct relatives of humans and inhabited Eurasia from before 400kya until 40kya when they disappear from the fossil record (Pinhasi et al. 2011; Galvan et al. 2014; Higham et al. 2014). Neandertals are morphologically different from anatomical modern humans (present day individuals) in various features such as skull shape, occipital bun, brow ridge, rib cage and the whole body stature (Helmuth 1998; Sawyer and Maley 2005; De Groote 2011). Fossil evidence shows overlap of the Neandertal existence in Europe and the arrival of modern humans to Eurasia (Higham et al. 2014), leading to speculation that admixture between Neandertals and modern humans may have occurred (Trinkaus et al. 2003).

A study of draft Neandertal nuclear sequences from three different individual fossils revealed that Neandertal genome shares more genetic variants with non-Africans than Africans whereas the sharing remains the same when comparing non-African individuals to each other or African individuals to each other. This result suggests gene flow from Neandertals into the ancestors of non-Africans (Green et al. 2010).

In 2010, a proximal toe phalanx was found in the eastern gallery of Denisovan cave in the Altai Mountains. It was identified to be from a Neandertal, which was

named the “Altai Neandertal”, after the mountains at the site where the fossil was excavated. The genome sequences from the Altai Neandertal was sequenced to a coverage of ~50X and compared to different present day human populations, confirming earlier results that indicated admixture between Neandertals and non-Africans (Prüfer et al. 2013). The proportion of Neandertal ancestry in all present day non-Africans was estimated to be ~2-5%. This admixture was estimated to have occurred 50,000-60,000 years ago (Fu et al. 2014), i.e. around the time when anatomically modern humans moved out of Africa. The location of the regions with Neandertal ancestry in the genomes of present day humans have been inferred from comparisons of this Neandertal genome to the genomes of over 1000 human genomes (Sankararaman et al. 2014; Vernot and Akey 2014; The Genomes Project 2015).

In 2008, a hominin finger phalanx was discovered in the eastern gallery of Denisova cave and the DNA from this sample was extracted and sequenced to a coverage of 30X (Meyer et al. 2012). DNA sequences from this hominin were closer to Neandertals than humans, but more distant to Neandertals than any two Neandertals sequenced till date. This indicated the individual to be a sister group of Neandertals, named “Denisovans”, after the cave where the fossil was found. Nuclear genome analysis of the Denisovan genome and present day humans revealed that Oceanians (south east region comprising of Melanesia, Micronesia, Polynesia, and Australasia) share more Denisovan alleles than Eurasians or Africans. The Denisovan component in Oceanians was estimated to be ~4-6% (Gittelman et al. 2016; Sankararaman et al. 2016).

2.3.2 Processing of Ancient DNA Sequencing Data

The library-prepared ancient DNA molecules could be sequenced using any short read sequencing technology given their fragment length, however the data generated in house are sequenced using Illumina technology. These reads are further processed to remove adaptors and indices using leehom program (Renaud et al. 2014). The processed reads after adaptor trimming and indices removal are called DNA sequences.

2.3.2.1 Mapping

The DNA sequenced from a Neandertal fossil are a mix of sequences from the Neandertal's genome, microbial contamination and potentially contamination from other fauna. One way to extract hominin sequences and to eliminate microbial contamination is by aligning all DNA sequences to a close reference genome, the human genome. Genome alignment is the process of finding the best match for a short query string (e.g. "reads") in a larger, pre-formatted string database (e.g. "reference genome"). All sequences presented here were aligned to the human reference (hg19 reference) using the program `BWA` (Burrows wheeler alignment tool) which implements a search for reads in an indexed burrows wheeler transformed reference genome database. BWT is a text compressing algorithm which sorts all possible rotations of a string including spaces in lexical order and constructs an index by taking the last column of the sorting output. This helps in data compressing and allows for string matching for larger genomes such as the human genomes. The algorithm implemented in `BWA` aligns the reads to the reference index allowing for mismatches and gaps which can be caused by sequencing errors or which can represent polymorphisms in the sequenced genome. To accommodate a larger fraction of mismatches along the damaged ancient sequences, `BWA` was run with parameters –n 0.01 –o 2 –l 16500.

2.3.2.2 Genotyping

After mapping, sequences covering each position are compared to infer what bases the two genomes this individual inherited from his parents carried (called genotyping). In addition to bases, genotyping can also infer insertion/deletion differences between the two parental copies of the genome. For the data presented here, the Genome Analysis ToolKit (`GATK`) was used with parameters --output_mode EMIT_ALL_SITES --genotype_likelihoods_model BOTH (McKenna et al. 2010). The output variant calls are stored in a variant call format (VCF) file.

2.3.3 Contamination Estimates

Methods for the quantification of modern human contamination can be based on modern human specific sites or on the ratio of alignments to X and autosomal chromosomes. For the samples used here, the contamination rates were estimated to lie consistently below 1%.

2.3.3.1 Human Mitochondrial Contamination Estimate

Modern human mitochondrial contamination was estimated on sequences longer than 35bp in length. This method uses a set of diagnostic sites which are defined based on differences between Neandertal consensus mitochondrial DNA and 311 human mitochondrial DNA. In addition, deamination on the ends of the sequences (first and last two bases) was used as pre-requisite to identify ancient reads.

A read is classified as human contaminant and is filtered out if a read overlap any of the derived human diagnostic sites or if the diagnostic site base is a transition compared to reference or if the base in the read overlapping a diagnostic site overlaps the bases A or T which could be a result of deamination or if a read has low/no levels of deamination (Prüfer et al. 2013).

2.3.3.2 Human Autosomal Contamination Estimate

Human nuclear contamination was estimated using the human genome excluding sex chromosomes, on sequences which are a minimum length of 35bp and a minimum mapping quality and base quality of 30. These sequences were further filtered to be uniquely mapping to human reference with mapability filter (map35_100%).

A list of derived fixed changes in present day humans compared to great ape outgroups are catalogued and the frequency of the derived variant in overlapping reads is tabulated. The method uses the fact that modern human contamination will introduce reads that carry the derived variant at positions where the archaic individual carries a homozygous or heterozygous ancestral genotype. A

maximum likelihood method was used to estimate contamination level based on the frequency of reads carrying derived alleles in a sample (Prüfer et al. 2013).

2.4 Whole Genome *de novo* Assembly

One of the challenges in genomics is reconstructing the genome sequence of an organism from the comparatively short sequences produced by sequencing technology. The length of sequences varies between types of sequencers; for instance, Sanger produces ~1000bp reads and next generation sequencers such as Illumina produce ~150bp reads. For Sanger sequencing, the predominant algorithm used for *de novo* assembly is often called a "overlap, layout, consensus" (OLC) assembler. However, with the advent of next generation sequencers, a different type of algorithm, based on a *de Bruijn* graph, is often used (Li et al. 2012).. The details of these algorithms are discussed below.

Figure 2.3 General steps involved in *de novo* assembly of a genome

A *de novo* assembly for constructing a genome from sequencing data can involve an initial step of error correction of reads to remove sequencing errors. Reads are then used to re-construct contiguous sequence (contigs), and paired-end reads are added to further join these contigs into scaffolds. Each of these steps are discussed in detail below (Figure 2.3).

1) Error Correction of Reads Before Assembly

Errors in read data can be problematic for assemblies, since similarity between sequences from the same location in the genome are harder to detect. The read correction is generally carried out using two methods: read alignment and k-mer frequency spectrum.

The read alignment method is a probability based approach which identifies and corrects reads with errors by using multiple alignments between all reads. This method is computationally intensive as it involves computing all possible pairwise alignments. Some of the software which implement this method are `coral` (Salmela and Schröder 2011) and `ECHO` (Kao et al. 2011).

The second method uses k-mers which are possible sub strings of a read, of length "k". This method uses a k-mer frequency, which is a frequency distribution of all possible k-mers from a given sequencing data which results into a normal distribution. The normal distribution is then divided up into trusted k-mers and untrusted k-mers based on a hard cut-off. K-mers with k-mer frequency less than the given cut-off are classified as untrusted and k-mers with frequency greater than a given cut-off are classified trusted. The bases in an untrusted k-mer are modified to match a known trusted k-mer thereby correcting errors in the reads. In order to avoid correcting real unique k-mers, some of the softwares have a limit on the number of changes per length of the read thereby finding the minimal change path from untrusted to trusted k-mers e.g. `Musket` (Liu et al. 2013), `SOAPdenovo` (Li et al. 2010) (Figure 2.4).

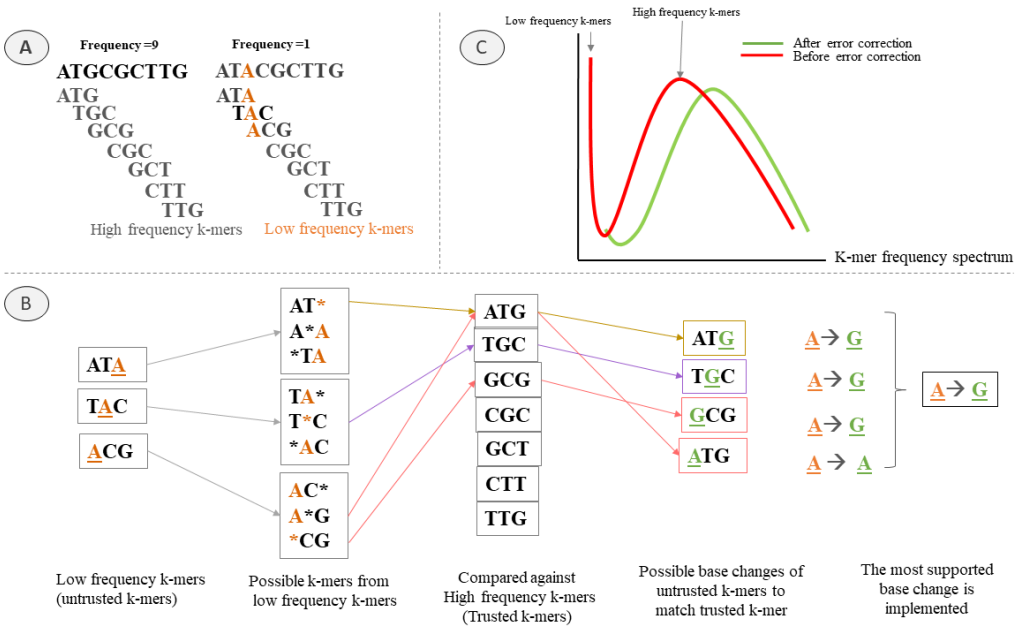


Figure 2.4 Schematic showing error correction using k-mer frequency spectrum The k-mer frequency spectrum works by classifying k-mers into trusted and un-trusted (A) Error correction of un-trusted k-mer to trusted k-mer (B) thereby observing a shift in k-mer frequency spectrum (C).

2) Contig Construction

This step involves construction of continuous sequences called "contigs" using sequencing reads. The approach varies depending on the read length as long reads from Sanger sequencing used overlap, layout, consensus (OLC) approach, whereas later next generation sequencing technologies which produce short reads use *de Bruijin* graph approach. These algorithms are discussed in detail below.

Overlap Layout Consensus (OLC)

In this algorithm the overlaps between reads are first detected by computing pairwise alignment between reads. This is a computational intensive step as the number of pairwise comparisons grows roughly quadratic with the number of sequences.

By identifying overlap between reads, the algorithm constructs a graph with each read as a node and the overlap between reads as edges. Since each node in this graph represents a read the number of nodes will increase linearly with increase

in the number of reads (sequencing depth), while the edges which represent the overlap between reads will increase at logarithmic scale. The continuous sequences also called "contigs" are called from the constructed graph called "layout" by following a Hamiltonian path i.e. walking in the graph covering each node only once (Peltola et al. 1984) (Figure 2.5a). The Hamiltonian path problem is a NP hard problem. Some of the softwares implementing OLC algorithm are `Celera` Assembler (Denisov et al. 2008), `Phusion` (Mullikin and Ning 2003), `Newbler` (Margulies et al. 2005) and `PCAP` (Huang et al. 2003).

De Bruijin Graph (DBG)

Due to the steep increase in the number of comparisons with the number of sequences the OLC method is often avoided when dealing with short read data. Instead, algorithms based on *de Bruijn* graphs are often used. A *de Bruijin* graph is a directed graph with the number of incoming edges equal to outgoing edges. It is constructed by first chopping reads into smaller chunks of a given length called "k-mers". These k-mers are then overlapped to form continuous sequences with an overlap of length k-1. This process of chopping and storing the information of the k-mer overlap is done simultaneously. The algorithm constructs a k-mer graph with k-mers as the nodes and the edges in the graph are given by the overlap between k-mers. This graph is more space-efficient than the graph produced by the OLC approach; for a genome of size G, a DBG graph is constructed with (G-K+1) nodes and G-K edges. The contigs are read from the constructed graph by following Euler path i.e. the shortest path where each edge is covered once (Pevzner et al. 2001) (Figure 2.5b).

The k-mer abundance of a given k-mer can be calculated using genomic coverage. For sequencing data of N reads of length L, the number of bases sequenced is given by N_b (N_b=N*L). Given the coverage of bases for this data as (d_b) the number of k-mers possible of length k is given by: N_k=N_b*(L-K+1) and the coverage of these k-mers is given by: d_k=d_b*((L-k+1)/L). Using the information of k-mer coverage and base coverage, we can get an estimate of genome size of an unknown genome G=N_k/d_k.

Some of the softwares implementing DBG approach are `VELVET` (Zerbino 2010), `SOAPdenovo` (Luo et al. 2012), `IDBA` (Peng et al. 2010), `Minia` (Chikhi and Rizk 2013).

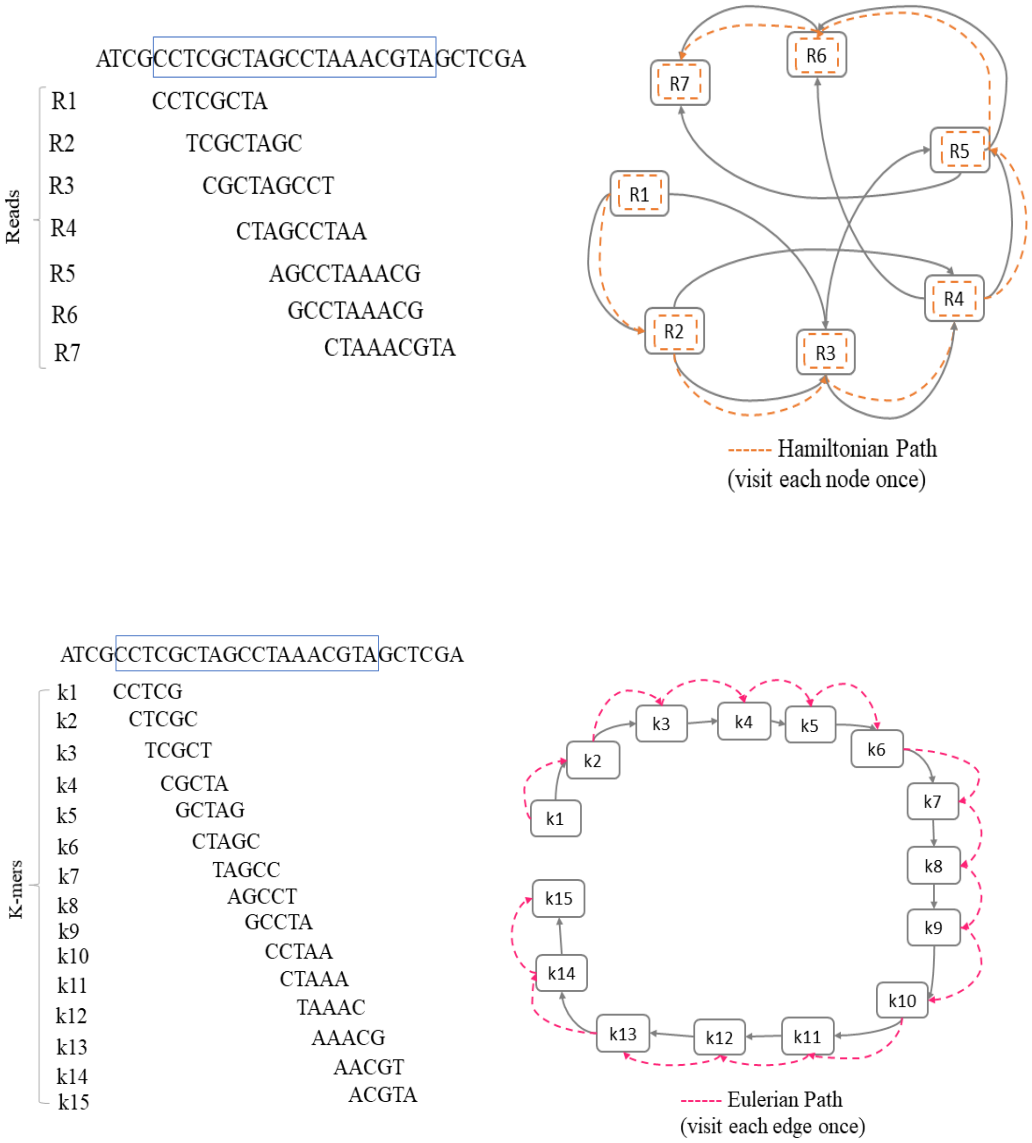


Figure 2.5 Graph reading approaches used in OLC and DBG *de novo* assembly

Role of Heterozygosity and Repeats in Contig Construction

Heterozygosity in a genome can affect contig construction. In the OLC approach this is overcome by allowing for few mis-matches. However, in the DBG approach, will form small "bubbles" in the graph that need to be eliminated. Repeats in a genome are a major obstacle for constructing contigs. A repeat which is shorter than the read length can be resolved, so in the case of OLC method is comparatively easily handled. But repeats increase computational time in OLC method as they align with many other reads. In DBG repeats are often collapsed into few nodes with many connections to other nodes. This structure is hard to resolve and various heuristics are used to either resolve repeats or separate them from neighboring contigs.

3) Scaffolding of Contigs

Scaffolding is the process of resolving the repeats and constructing contiguous genomic sequences from contigs by using the information from paired-end sequences. If one end of a paired read aligns to a contig and the other end to another contig, then these contigs can either be merged or oriented in respect to each other.

4) Gap Closure in Scaffolds

In order to get a chromosomal level assembled genome sequence, the gaps between and within the scaffolds needs to be filled and closed. The gaps within a scaffold are called in-gap and the gaps between scaffolds are called out-gaps. The in-gap mostly result from a repeat region which was not resolved and could be closed by paired end sequencing information. The out-gaps are much harder to fill since these are repeat region which are longer than the paired end insert size, so they cannot resolve these gaps.

2.5 Indels and Genomic Rearrangements

Short indels (insertion or deletion), i.e. those less than or equal to 50bp, are the second most abundant type of mutations after SNPs. However, variation in the genomes of modern human exists also in the form of genomic rearrangements or structural variations (SVs) which include large insertions, deletions, inversions and translocations. Although the rate of formation of short indels and genomic rearrangements is much lower than SNPs, they have a stronger functional potential as they can alter one or more genes at a time. The SVs account for ~1% of variation among humans and a previous comparative study of an individuals from different populations to the reference genome revealed that SVs account for up to 1.2% variation in the genome, much higher than SNPS which account for only 0.1% (Pang, MacDonald et al. 2010). Given that structural variants have strong functional potential their characterization is important. Some small Indels have also been linked to various genomic disorders.

The length of an indel can determine its effect on gene function. For instance, indels in coding regions of length that is not divisible by 3 are depleted since they lead to a shift in the open reading frame (ORF) resulting in amino acid sequence differences for the protein product. In contrast, indels in coding sequences with a length divisible by 3 would only result in the loss or addition of a single amino acid while subsequent amino-acids remain unaltered.

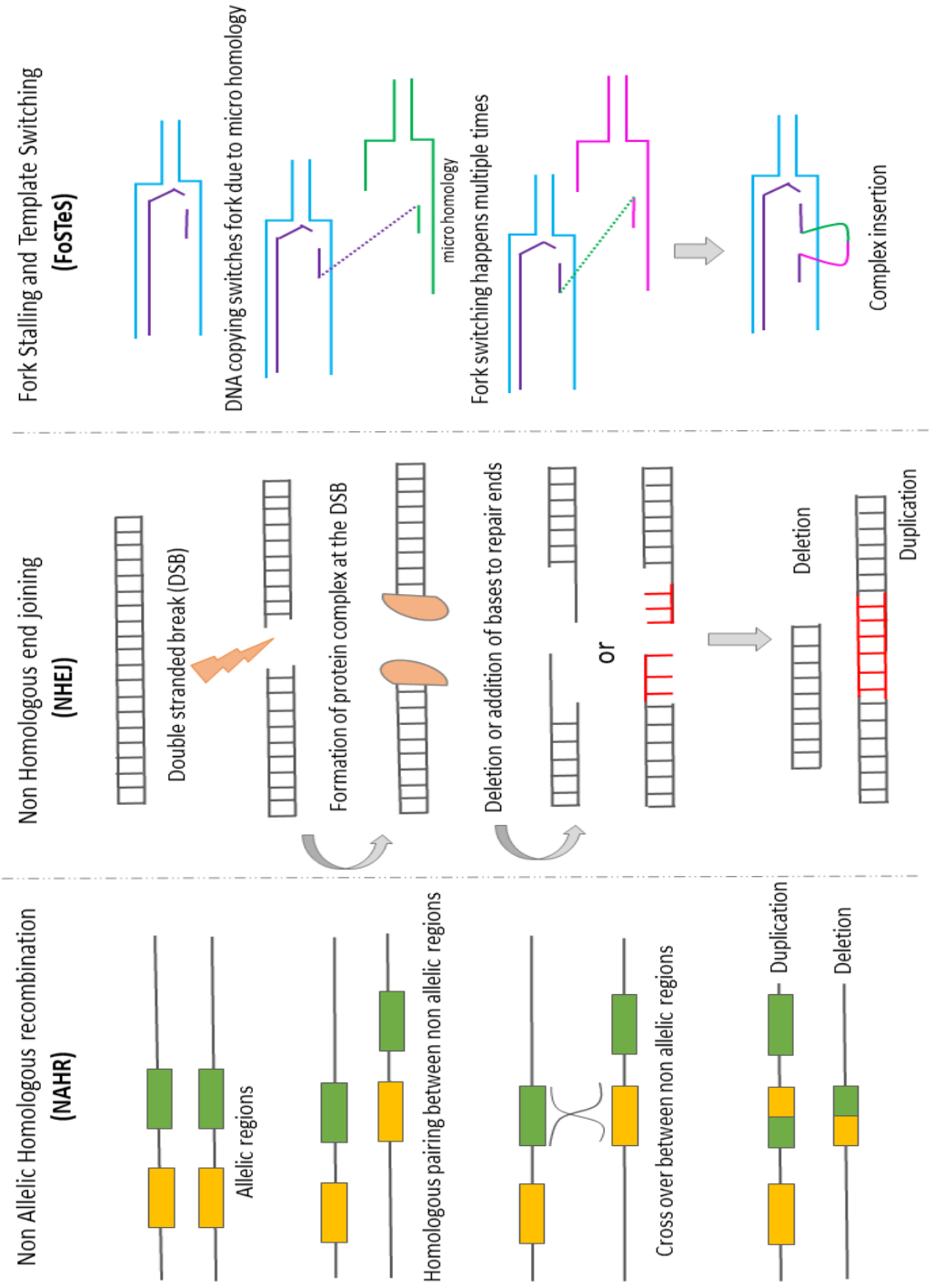


Figure 2.6 Three major mechanisms involved in the formation of small indels and genomic rearrangements in human genome

These mechanisms in a cell lead to indels (1-10bp) and large scale rearrangements. Figures adapted from (Lieber 2010; Chen et al. 2014; Ottaviani et al. 2014)

2.5.1 Biology Behind the Formation of Indels and Genomic Rearrangements

One of the well-studied mechanisms for the indels formation is replication slippage (Levinson and Gutman 1987; Taylor et al. 2004). During the replication of DNA, the polymerase skips a few bases creating a small deletion or adds new nucleotides creating an insertion. The slippage often occurs in tandem repeat rich regions as these regions are unstable during replication. One of the well-known genomic disorder caused by tandem duplication is Huntington's disease caused by expansion of trinucleotides.

Over the years, mechanisms leading to the formation of genomic rearrangements altering gene order or gene orientation by deletion, duplication, inversion or translocation have been well studied. There are three major mechanisms proposed explaining the formation of rearrangements: Non Allelic Homologous Recombination (NAHR), Non-Homologous End-Joining (NHEJ) and Fork Stalling and Template Switching (FoSTeS) models (Gu et al. 2008). Each of these mechanisms lead to formation of different kinds of rearrangements. NAHR leads to formation of deletions, duplication, inversion and translocation, NHEJ have been proposed to create deletions and at times duplications and FoSTeS leads to complex rearrangements (Figure 2.6). Most rearrangements formed are in the size of 1kb to 1Mb and sometimes larger than 3MB which can be visualized microscopically.

2.5.1.1 Non Allelic Homologous Recombination (NAHR)

An allele is a DNA region present at the same genetic locus on both the chromosomes in a diploid organism (e.g. Humans). Allelic regions on both chromosomes are homologous, meaning they have high sequence similarity. In a regular recombination event, the cross over occurs between two alleles after their homologous pairing during meiosis.

However, there are regions in the genome such as low copy repeats or segmental duplications which are 10 to 300bp with 95-97% sequence similarity which cause

glitches during homologous pairing. These low copy repeats often undergo non allelic homologous pairing due to their sequence similarity and crossover of these non-allelic region causes unequal products. This mis alignment and unequal recombination could occur both in meiosis or mitosis. When the low copy repeat regions are on the same chromosome and in same direction, recombination between them results in a deletion or duplication. When the low copy repeat regions are on the same chromosome and in opposite direction, they cause inversion and if the low copy repeat regions are on different chromosomes, recombination between them results in a translocation event.

NAHR is one of the major mechanism leading to the formation of genomic rearrangements. Since this mechanism is guided by low copy repeats or segmental duplication, they have known hotspots in human genome (Reiter et al. 1996; Lopez-Correa et al. 2001). These events have been mostly found in the regions prone to double stranded breaks, hence there are studies suggesting a correlation between NAHR and double stranded breaks in DNA.

Genomic rearrangements caused by NAHR are known to be associated with genomic disorders. Some of the known genomic disorders involve Charcot-Marie-Tooth disease type 1A (CMT1A) which is heritable, suggesting that duplications/deletions could have been a
result of NAHR in the gametes (Raeymaekers et al. 1991). An example of a sporadic non-heritable genetic disorder is Potocki-Lupski syndrome (PTLS) which could be due to deletion/duplication through NAHR in somatic cells (Potocki et al. 2007). It has been suggested that, NAHR is involved in many cancers creating a mosaic of cells with and without rearrangements. In addition, there are studies suggesting significant difference of NAHR between mitosis and meiosis and as well as between males and females (Steinmann et al. 2007).

2.5.1.2 Non Homologous End Joining (NHEJ)

Non homologous end joining is second major mechanisms used to repair a double stranded break in DNA. The mechanism first identifies the double stranded break, then a molecular bridge is formed at the break which repairs the broken ends by

adding or deleting bases making it compatible for ligation in the next step, then the repaired ends are ligated using ligase. During the repair of broken ends in a double stranded break, the possibility for addition or deletion of bases leads to the formation of duplication or deletion. A deletion/duplication by NHEJ at the Duchenne Muscular Dystrophy (DMD) gene locus is a known cause of muscular dystrophy (Toffolatti et al. 2002).

2.5.1.3 Fork Stalling and Template Switching (FoSTeS)

Apart from the above two mechanisms, fork stalling and template switching is a mechanism which explains complex rearrangements. This mechanism is different from NAHR and NHEJ which involve double stranded breaks, since FoSTeS occurs during replication. During replication of DNA, the double stranded DNA is opened into single strands and this site is called replication fork. Each single strand is used as a template and DNA copying occurs through DNA polymerase. The strand which has synthesized from 3' to 5' is called the leading strand and the other strand is called lagging strand where DNA synthesis occurs in small fragments. In the FoSTeS model, the replication fork stalls and the lagging strand switches to another replication fork which has sequence similarity with the lagging strand of the first fork. Later the lagging strand switches back to the original fork and synthesizes the remaining DNA. This switching between forks can happen from one to multiple times causing complex genomic rearrangements. Depending on the strand that was invaded and the

location of invading fork upstream or downstream it can lead to a deletion or duplication (Lee et al. 2007).

2.5.2 Computational Approaches to Detect Genomic Rearrangements

One of the widely used methods to detect genomic rearrangements was microarrays (Pinkel, Segraves et al. 1998, Iafrate, Feuk et al. 2004) which was lately replaced by NGS technologies and novel computational algorithms. Some

of the large databases which report genomic variants in humans are 1000 genomes (Conrad, Pinto et al. 2010, The Genomes Project 2015); other databases which report SVs and their disease associations are dbVar, DGVa and OMIM (Amberger, Bocchini et al. 2009, Lappalainen, Lopez et al. 2013).

There are four major computational strategies to identify structural variants which all require that the reads of the query genome (the sequenced genome) are aligned to a reference genome (Figure 2.7). Each of these strategies is discussed in detail below.

1) Read Coverage Approach

This method uses the coverage of sequenced data in a given genomic region of the reference genome. Assuming the genome is sequenced uniformly, the coverage follows a Poisson distribution (Lander and Waterman 1988) and any deviation from this expectation will reflect a genomic variation which could be either deletion or duplication (Magi, Tattini et al. 2012). For instance, in a diploid genome (e.g. humans) a genomic region with a novel duplication with respect to reference would show an increased coverage (by a factor of 1.5) and a deletion would result in lower coverage (factor 0.5) at the respective locus. This method was further improved by normalizing reads based on their length or by correcting for GC content, which often biases coverage due to different efficiencies in extraction, library preparation or sequencing for molecules of different GC content (Iakovishina et al. 2016). Some of the softwares which implement this method are PSCC and ReadDepth (Miller, Hampton et al. 2011, Li, Chen et al. 2014).

1) Paired End Mapping Approach

Paired end sequencing is advantageous over normal single end sequencing since it carries the information of the distance between the read pairs also called “insert size”. Both the insert size and the orientation of the paired end mapping can be used to infer rearranged regions. The discordant mapping of read pairs with a given insert size can indicate rearranged regions with respect to the reference

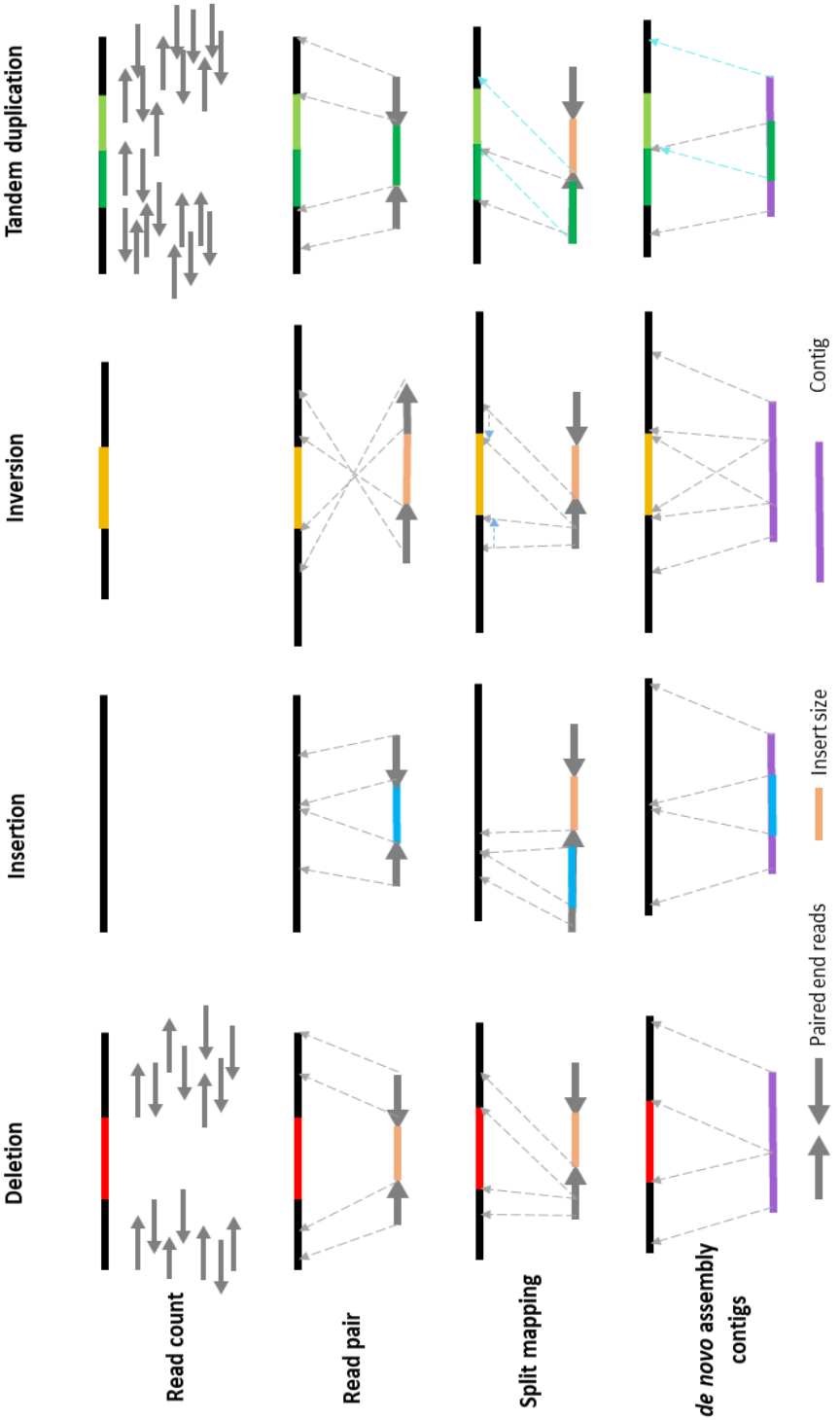


Figure 2.7 Computational techniques for identifying four different kinds of rearrangements.
As shown here, there are four distinct ways: using single/paired end abundance using coverage, using paired end reads, split mapping of read and using contigs generated from *de novo* assembly. Figure adapted from (Tattini et al. 2015)

genome (Korbel, Urban et al. 2007). A gap in the alignment not concurrent to the insert size indicates an insertion/deletion; a flip in the direction of the mapping read in a read pairs indicates an inversion and mapping of read pairs to two different chromosomes indicates a translocation. This approach is powerful as it can detect several classes of rearrangements compared to the read coverage method, which is limited to copy number variation. The implementation of this method requires detection of clusters of read pairs with a minimum read support threshold and a standard deviation from the insert size supporting the discordant mapping to call a rearranged region. Some of the softwares which

implement this approach are BreakDancer (Chen, Wallis et al. 2009) and CLEVER (Marschall, Costa et al. 2012)

2) Split Read Mapping Approach

Split read alignment is a mapping technique exclusively developed to detect genomic rearranged regions, however it uses long reads for reliably detection of breakpoints. This technique was implemented by several softwares which use read pairs and insert information to identify rearranged regions. The presence of a rearranged region is indicated by a signature of breakpoint, region in the genome where the query read breaks and the rest of the read maps elsewhere due to an underlying rearranged region. This approach also detects almost all classes of rearrangements (Zhang, Du et al. 2011). One of the caveat of this method is that its detection efficiency for deletions is higher than insertions and the efficiency for detection of insertions is dependent on insert size of the paired end sequencing.

3) *De novo* Assembly Approach

This approach is based on *de novo* assembly of short reads using *de Bruijin* approach or long reads using OLC approach to construct contigs and then using these contigs to infer rearranged regions by mapping them to a reference. This approach is ideal in detecting most types of rearrangements. One notable exception are rearrangements due to mobile elements or overlapping repeat regions which are hard to detect since assembly methods using short reads tend

to collapse regions with high sequence similarity into a single copy (Liu, Huang et al. 2015). One of the disadvantages of this approach is that it identifies homozygous variants as assembly constructs a haploid genome.

Though all the above mentioned methods have their pros and cons in identifying rearranged regions, these methods are often used in combination to get full spectrum of results. For instance, SVDetect (Zeitouni, Boeva et al. 2010) uses both read count and paired end information to detect rearranged regions, GenomeSTRIP infers population scale genomic rearrangements using read count, paired end and split read mapping approaches. Tools such as HYDRA (Quinlan, Clark et al. 2010) and NovelSeq (Hajirasouliha, Hormozdiari et al. 2010) use paired end reads and local assembly approaches to infer rearrangements.

Chapter 3 Open Questions to be Addressed

In this dissertation I address two open questions in the field of ancient DNA. Nearly all studies in the field of ancient DNA focus on single nucleotide polymorphisms (SNPs) and carry out genetic analysis. In my graduate study, I carried out two studies which focus on variation contributed by mutations other than SNPs, such as indels and rearrangements which are described in detail in chapter 4 and chapter 5.

3.1 Analysis of Small Indels on the Human Lineage Using the Neandertal Genome

Small insertions and deletions (also known as indels) are a class of mutations that are formed by the loss/gain of a small number of bases (here I focus on indels with a length 1-5bp). Small indels are roughly 10-fold less abundant than SNPs, but they may contribute over-proportionally to functional variants and disease. Besides, some of the indels segregating within the human population are also expected to originate from the admixture with Neandertals. To gain insight into the evolution of small indels, I focused on comparing indels that were formed at different time-frames (before and after the split from Neandertals) and those that originate from admixture. I tested whether the action of selection can be observed as differences between different time-frames and to introgressed indels. In addition, this study looked for variants for further study: modern human specific indels that went to fixation may underlie modern human specific traits, whereas some introgressed indels may contribute to phenotypic variation among present-day people.

3.2 Large Scale Genomic Variation in Archaics Hominins Compared to Modern Humans by *De Novo* Assembly of Archaic Genomes

Genomic variation due to rearrangements are often overlooked while studying divergence between populations or species. These mutations are not as common as SNPs or small indels, however given their potential to influence multiple genes and regulatory networks, they can have wider functional consequences and are often associated with disease risk. Large scale rearrangements in humans were in the past identified using primate genome sequences which yielded a list of changes that occurred during the past 6 million years or longer ago. Now due to the availability of archaics genomes, we can identify the more recent changes on the human lineage after the split of humans from archaics. As an extension of my first work on small indels, I explored large scale variation differences between humans and archaics. This study aims at establishing whether *de novo* assembly, in contrast to read-based approaches that have been used before, is a viable option to detect structural rearrangements from ancient DNA reads from archaic humans and to provide a list of variants detected in the archaic human genomes. However, *de novo* assembly of short damaged ancient DNA reads is a complex task. In this part of study, I address the impediments posed by ancient DNA and obtain fragmented assemblies for rearrangement detection. The aim of the study is to identify rearranged regions which are derived on the human and the Neandertal lineage respectively which could give us hints about their phenotypic impact.

Chapter 4 Evolution of Small Insertions and Deletions in Modern Humans

4.1 Introduction and Motivation

Mutations are mainly composed of single nucleotide changes which effect one base at a time and indels (insertion or deletion) which add or delete one or more bases at a time. While most of the sequence variation among human individuals is due to single nucleotide changes, Indels contribute around 10% to the total variation. The rate of occurrence of small indels is approximately one order of magnitude less abundant than SNPs but have a higher probability to affect function than nucleotide substitutions (Montgomery et al. 2013).

Given their functional importance and their substantial influence on diversity in a population, indels are often excluded in evolutionary studies. This is likely due to the particular challenges of indel genotyping (Mullaney et al. 2010; Neuman et al. 2013; Hasan et al. 2015) and the heterogeneous processes generating indels that lead to a large variation in mutation rates along the genome (Belinky et al. 2010; Kvikstad and Duret 2014). Several studies were performed which found contradictory results, for example; deletions were found to evolve, on average, under stronger negative selection on the human lineage than insertions by one study that compared fixed to polymorphic indels (Sjödin et al. 2010), while a later study found the opposite signal using the allele frequency spectrum between populations (Huang et al. 2013). The cause for this discrepancy may lie in homoplasy, i.e. the independent occurrence of identical changes on several lineages, which can lead to the mis-assignment of the ancestral state and type of the mutation (insertion or deletion) (Kvikstad and Duret 2014).

In this chapter I use the Neandertal genome (Fu et al. 2014) together with data of present-day humans from the 1000 Genomes data (Genomes Project et al. 2015) to identify indels and divide the set of indels further into those that likely occurred

after the split from Neandertals, those that arose before the split from Neandertals and likely introgressed indels. I test for different patterns of selection between these sets and compile a list of introgressed and modern-human-fixed indels that may contribute to modern human phenotype.

4.2 Methods

4.2.1. Primate Multiple Sequence Alignment

Pairwise alignments between the human reference genome (Lander, Linton et al. 2001) (GhRch37/hg19) and six primates (chimpanzee (The Chimpanzee Sequencing and Analysis Consortium 2005) (panTro4), gorilla (Scally et al. 2012) (gorGor3), orangutan (Locke et al. 2011) (ponAbe2), gibbon (Carbone et al. 2014) (nomLeu1), rhesus macaque (Gibbs et al. 2007) (rheMac3) and marmoset (2014) (calJac3)) were downloaded from the UCSC genome browser (Speir et al. 2016) and converted into MAF format. In addition, the bonobo (Prufer et al. 2012) (panpan1.1) pairwise whole genome alignment to hg19 was prepared in house following the processing applied to genomes for inclusion in the UCSC genome browser. All seven pairwise alignments were joined into one multiple sequence alignment using the reference guided alignment program multiz (Version: roast.v3; Command-line: “roast + E=hg19 '(((((hg19(panTro4,panpan1.1) gorGor3)ponAbe2)nomLeu1)rheMac3)calJac3)' <input_files.sing.maf> <output_file.maf> ”, (Blanchette et al. 2004)). The resulting file was filtered to retain only those alignment blocks that include sequence from the genomes of all eight species.

4.2.2. Inferring Fixed Derived and Polymorphic Indels on the Human Lineage

Human polymorphic indels were extracted from the 1000 Genomes phase 3 dataset (The 1000 Genomes Project Consortium 2015). The indels were further filtered by requiring overlap with the eight species whole genome alignment and requiring all seven non-human reference sequences in this alignment to agree. The ancestral state of polymorphic indels was then called as the non-human state

and the alternative labeled as a derived human-specific indel. Further filtering was carried out to remove sites with more than one derived variant and long variants marked as variable in copy number (denoted as <CN> for the derived state in the 1000 Genomes data).

Human-specific derived indels were called fixed if all non-human species showed an identical insertion or deletion difference compared to the human reference sequence and if the position was not listed as polymorphic in the 1000 Genomes data.

4.2.3. Inferring Modern Human Specific Indels and Putatively Introgressed Indels Using the Neandertal Genome

I used the genotype calls of a Neandertal from the Altai Mountains (Fu et al. 2014) to divide derived human-specific indels into those that are shared with Neandertals and those that are specific to modern humans.

Two percent of the genomes of present day non-Africans show high similarity to the Neandertal genome due to a recent admixture event with Neandertals (Fu et al. 2014). To infer putatively introgressed indels I used our set of human polymorphic indels and filtered for variants that are fixed in individuals from sub-Saharan African populations (Luhya, Yoruba, Gambian, Mende and Esan) and show an alternate allele in the Europeans (Utah, Finland, British and Scotland, Iberian, Toscani) or East-Asians (Chinese Dai, Han Chinese, Southern Han Chinese, Japanese, Kinh) that is shared with the Neandertal. I used the same process to infer introgressed SNPs (Figure 4.1).

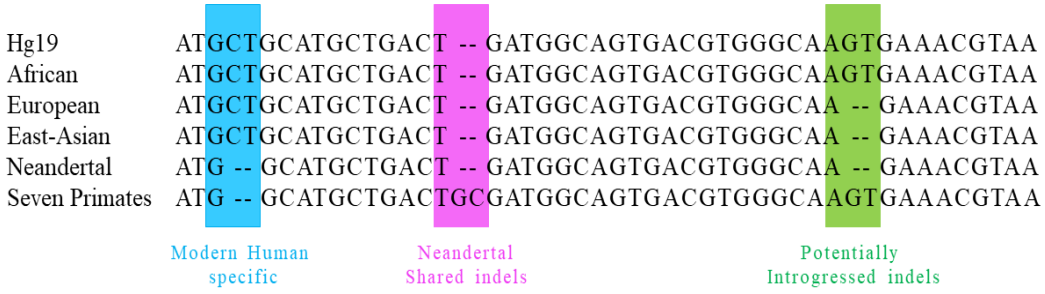


Figure 4.1 schematic showing different categories of indels on the human lineage

4.2.4. Contrasting Fixed and Polymorphic Insertions and Deletions

The McDonald–Kreitman test (McDonald and Kreitman 1991) compares the number of polymorphic changes within one species to the number of fixed changes when comparing to another species between two types of sites, neutral and non-neutral. Under neutrality the ratio of non-neutral to neutral changes is expected to be equal when comparing fixed to polymorphic changes. Negative selection is expected to reduce the number of non-neutral changes that reach fixation, while repeated positive selection is expected to increase the number of non-neutral changes due to the rapid fixation of advantageous alleles. Following the approach of Sjödin et al. and Kvikstad and Duret (Sjödin et al. 2010; Kvikstad and Duret 2014), I applied the concept of the McDonald-Kreitman test to indels by comparing the number of insertions and deletions that are polymorphic to those that are fixed-derived on the human-lineage. P-values were calculated using Fisher's exact test as implemented in R (R Core Team 2014).

4.2.5. Derived Site Frequency Spectra of Polymorphic Indels

I used the average allele-frequency for different populations from the 1000 Genomes phase 3 data to tabulate the site frequency spectra. Site frequency spectra were compared by applying a two-sided Wilcoxon rank sum test with continuity correction to the distribution of indel frequencies.

The minor allele frequencies for potentially introgressed indels in the European populations and the East Asian populations from the 1000 Genomes Project phase 3 were tabulated to arrive at an AFS of introgressed indels.

4.2.6. Annotation of Indels

Indels were annotated using the variant effect predictor (`VEP`) (McLaren et al. 2010) version 78 using the option "–most_severe" to limit the output to one annotation per indel. For each annotated region and for each pair of classes of indels, I determined the significance by calculating Fisher's exact test on a 2x2 contingency table contrasting the two classes and the counts inside and outside of the annotated region. The combined list of p-values from all variance effect predictor tests was FDR adjusted using the p.adjust() function implemented in R.

In addition the Combined Annotation Dependent Depletion (`CADD` v1.3) tool (Kircher et al. 2014) was used to score the tentative phenotypic impact of indels. `CADD` annotates each indel with a phred-scaled C-score. A cutoff of 20 on the C-score was applied to generate lists of indels with an increased chance of affecting phenotype.

4.2.7. Genome wide Association Studies (GWAS)

I used a collection of genome-wide association studies (GWASdb, version: 2015 August, hg19 dbSNP142, (Allentoft et al. 2012)) to find potential phenotype associations for introgressed indels. Since indels are typically excluded in the process of GWAS, I sought to detect SNP that are in perfect LD (linkage disequilibrium) with introgressed indels in the 1,000 Genomes. Indels that showed an identical combination of reference/non-reference genotypes as the GWAS associated SNP in all individuals were considered completely linked. I report phenotype associations for each indel that is in perfect LD with a SNP that has been associated with the corresponding phenotype with a p-value of at least 1e-6.

4.2.8. Gene Ontology Enrichment

Enrichment of indels in specific gene categories was tested using the software package FUNC version 0.4.7 (Briggs et al. 2007). For this, I selected indels that were assigned to genes based on the `VEP` annotation and further annotated these indels to gene categories used the Gene Ontology [version: Ensembl Genes 75 (GRCh37)]. To account for all the plausible effects, for instance when an indel overlaps more than one gene, I allowed multiple annotations of each indel. Genes were assigned corresponding GO categories using the Ensembl database (Cunningham et al. 2015).

In addition to explanations involving selection, the number of indels in a gene category can vary due to differences in mutation rates or due to a difference in gene-length between categories. In order to avoid these issues, I compared the number of two types of indels per category using the FUNC implementation of the binomial test. The following types of indels were compared:

1. Indels shared with Neandertals to those that are modern human specific
2. Indels that are shared with Neandertals to those that have come by introgression from Neandertals.

I chose a p-value cutoff of less than or equal to 0.05 for the family wise error rate (FWER) to filter for significantly enriched categories.

4.3 Results

4.3.1. Indels on the Human Lineage

To identify insertion and deletion events on the modern human lineage and to alleviate the problem of mis-assignment of the ancestral state, I aligned the human reference genome with seven primate genomes and inferred the derived state in the human lineage by requiring an identical ancestral allele in all seven primate genomes. An insertion on the human lineage is called only when all non-human primates show a deletion compared to the human state, and a human-specific deletion when all primates show an insertion. Our method detected 315,513

indels of 1-5bp in length in the human reference genome. Of these, most indels (315,412) were covered in the Altai Neandertal genome.

I used data from the 1000 Genomes project phase 3 (Genomes Project et al. 2015) to further increase the set of variable indels. Variants marked as copy number variants ("<CN>") exceeded the length of variants considered here and were excluded. A total of 2,982,740 were inferred from 1000 Genomes data after filtering out sites with more than one derived variant. These indels were assigned an ancestral and derived state by comparison to seven non-human primate genomes, and overlapped with Altai Neandertal genotypes, resulting in 989,138 indels of length 1-5bp. Combining indels identified using the human reference and those identified using the 1000 Genomes data, yielded 1,232,285 indels of size 1-5bps on the human lineage (245,520 appear fixed and 986,765 were segregating in present day populations) (Figure 4.2).

I computed the ratio of deletions to insertions for fixed (1.449) and polymorphic indels (2.06) and found ratios higher than 1, consistent with deletions accumulating approximately twice as fast as insertions (Ophir and Graur 1997; Fan et al. 2007; Matthee et al. 2007; Sjödin et al. 2010).

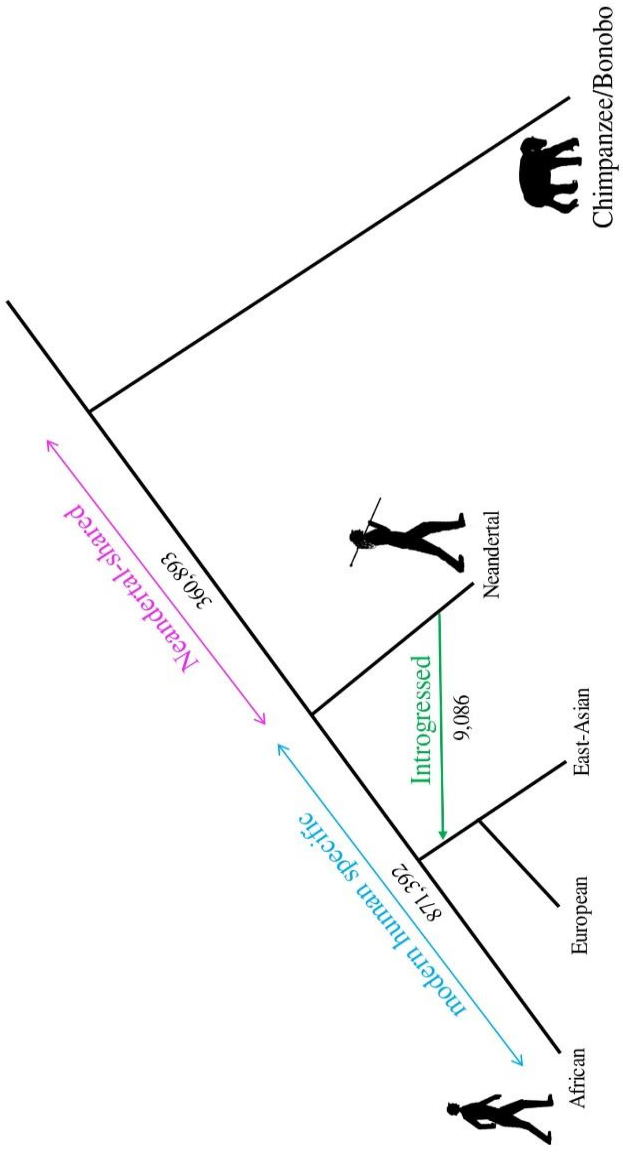


Figure 4.2 Indels analyzed in this study.
Indels on the human lineage divided into three categories: a) Indels which likely arose on the human lineage after the split from Neandertals and are specific to modern humans (blue) b) Indels which occurred before humans split from Neandertals and are shared with Neandertals (pink) c) Indels introduced into non-Africans due to introgression from Neandertals (green).

4.3.2. Modified McDonald–Kreitman Test on the Human Lineage Indels

Previous studies have used a modified version of the McDonald-Kreitman test (McDonald and Kreitman 1991; Sjödin et al. 2010; Kvikstad and Duret 2014) -- comparing the ratio of fixed deletions to fixed insertions to the ratio of polymorphic deletions to polymorphic insertions -- to test whether insertions and deletions are affected differently by selection. Under neutrality both the fixed and polymorphic ratios are solely dependent on the rate at which insertions and deletions are generated, i.e. at a roughly 2-fold higher rate for deletions than for insertions. Under this assumption, the ratios of deletions to insertions are not expected to differ significantly from each other when comparing fixed to polymorphic sites. However, a departure from this expectation can emerge if one type of change is selectively favored over the other, and is thus biased towards fixation. Note that such a signal requires only the average selection pressures on insertions and deletions to differ; the majority of both types of changes can still be selectively neutral.

I first applied the modified McDonald Kreitman test to all 1-5 base pair long indels described in the previous section and found a significant difference between the ratio of fixed to the ratio of polymorphic indels (p<2.2e-16). In order to test whether this signal is driven by a certain length of indels, I repeated the test for each length, separately, and found that the signal persists in all comparisons (Table 4.1). This result is consistent with the results of Kivkstat and Duret (Kvikstad and Duret 2014) and study by Sjödin et al. (Sjödin et al. 2010) suggesting that deletions are under stronger negative selection than insertions.

It is interesting to note, that the ratio of polymorphic insertions and polymorphic deletions also differs significantly between all lengths (pairwise comparisons between lengths 1-5bps: p-values< 0.05).

Indel length category (in bp)	1	2	3	4	5	Sum: 1-5
Fixed deletions	86791	26860	14802	12161	4689	145303
Fixed insertions	66333	13589	8022	9406	2867	100217
Fixed rDI	1.30	1.97	1.845	1.29	1.635	1.449
Polymorphic deletions	344533	121548	82114	84393	31607	664195
Polymorphic insertions	226712	38545	21147	27180	8986	322570
Polymorphic rDI	1.519	3.15	3.88	3.10	3.52	2.06

Table 4.1 Fixed and polymorphic indels on the human lineage by length.
Rate of deletions to insertions (rDI) is given for polymorphic and fixed indels of different lengths in the human lineage. Fisher's exact tests were applied to the counts of fixed and polymorphic insertions and deletions in each column and yielded p-values<2.2e-16 in all comparisons.

4.3.3. Derived Allele Frequency of the Human Lineage Indels

The derived allele frequency spectra (AFS) of polymorphic insertions and deletions can be used as an alternative to test for differences in selection pressure affecting both types of changes (Gibbs et al. 2007). The test is based on the idea that a favorable allele will on average segregate at higher frequency compared to neutral alleles, and neutral alleles will in turn segregate at higher frequencies compared to deleterious alleles (Fay et al. 2001). I found that the AFS for deletions differs significantly from the AFS for insertions (two-sided Wilcoxon rank sum test; $p < 2.2e\text{-}16$; Figure 4.3), with deletions showing an excess of low-frequency alleles compared to insertions. This signal is detected consistently in all 1000 Genomes populations and for all sizes of indels (1-5bp).

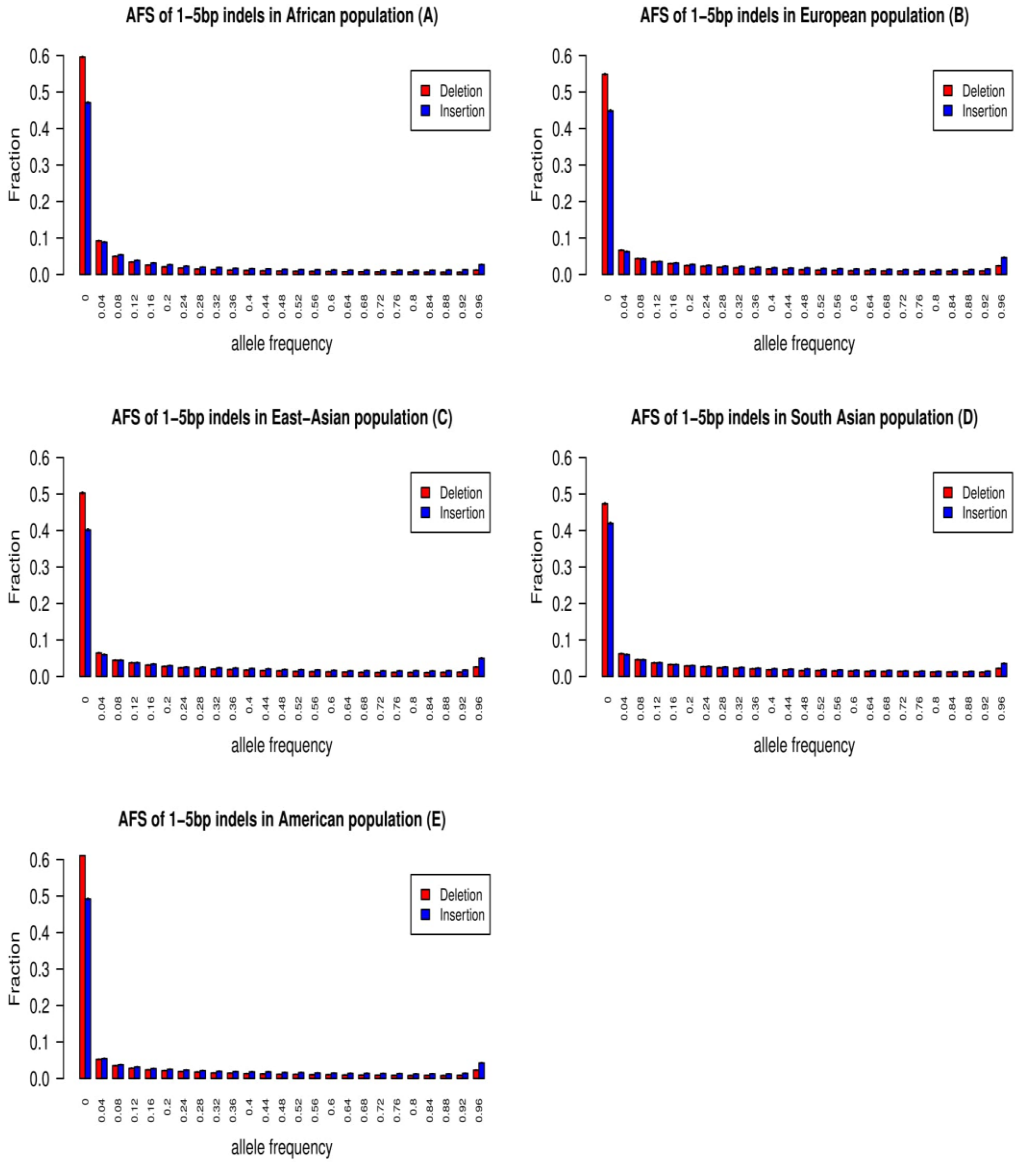


Figure 4.3 Derived allele frequency spectra (AFS) of indels in different populations the 1000 Genomes dataset.
The AFS for populations in 1000 Genomes dataset. Wilcoxon rank sum tests (two-sided) show that the frequency distributions of insertions and deletions differ significantly for all populations ($p < 2.2e\text{-}16$).

4.3.4. Genomic Distribution of the Human Lineage Indels

The previous two tests examined the difference in selection pressure between insertions and deletions by comparing allele frequencies. However, if one type of change is more often deleterious, a difference may also be visible in the fraction of insertions and deletions residing in regions that are more likely functional as compared to regions that are more likely neutral. I tested this hypothesis by annotating indels by their genomic location using the Variant Effect Predictor (McLaren et al. 2010). As expected, a major fraction of indels fall in intronic and intergenic regions while a much smaller fraction fall in coding regions. In addition, intergenic regions show a statistically significant higher fraction of deletions than insertions (binomial test; p=7.3e-119; FDR adjusted p=7.8e-117) while the opposite is true for intronic regions (p-value = 3.6e-59; FDR adjusted p=1.3e-57; Figure 4.4(A)). This observation is compatible with the notion that deletions are more constraint than insertions. However, these results should be interpreted cautiously as these differences in insertion and deletion frequencies may also be influenced by other factors, such as sequence context (Kondrashov and Rogozin 2004; Kvikstad et al. 2007; Kvikstad et al. 2009) leading to unequal insertion and deletion mutation rates between classes of genomic regions.

4.3.5. Modern Human Specific and Neandertal Shared Indels

I divided indels into those that were identified in the modern human reference and the Altai Neandertal, and those that were only detected in the human reference. A total of 37,443 indels were modern human specific and 265,975 were shared. The frequency of modern human specific indels can be used to calculate a relative divergence of the human reference to the Neandertal genome. I calculate a divergence of 12.3% relative to the divergence to the common ancestor with chimpanzee, close to the range of values calculated using nucleotide differences (11.2-11.8%, see SI6a in (Fu et al. 2014)).

I classified polymorphic indels from the 1000 Genomes Project (Genomes Project et al. 2015) into those for which the derived variant is shared with the Neandertal and those where the derived variant is only observed in modern humans, and

pooled the dataset with human-reference specific indels. As expected by the difference in age, the majority of the 360,893 shared indels were fixed (243,060 fixed and 117,833 polymorphic) while the majority of the 871,392 modern human specific indels were polymorphic (2,460 are fixed and 868,932 are polymorphic).

Neandertal-shared indels are expected to be on average older than indels that are specific to modern humans. I use this expectation to test again for differences between the ratios of deletions to insertions of both age-classes, similar to the McDonald-Kreitman test. The ratio of deletions to insertions is significantly lower for shared compared to modern human specific indels (Table 4.2,4.3) consistent with earlier comparisons between fixed and polymorphic indels.

Table 4.2 Contingency table contrasting modern human specific and shared indels.

Category	Shared	Modern Human specific
Deletions	205075	604423
Insertions	155818	266969
Ratio(Deletions/Insertions)	1.316	2.26

The ratios of deletions to insertions are significantly different between the shared and modern human specific classes (Fisher's exact test; p<2.2e-16, odds ratio=0.58).

Table 4.3 : Counts of insertions to deletions compared between modern human specific and Neandertal shared indels.

Category	Shared	Modern Human specific
Deletions	199041	604423
Insertions	152840	266969
Ratio(Deletions/Insertions)	1.30	2.26

Introgressed indels were removed from the counts of Neandertal-shared indels. The ratios differ significantly (Fisher's exact test p<2.2e-16, odds ratio=0.58)

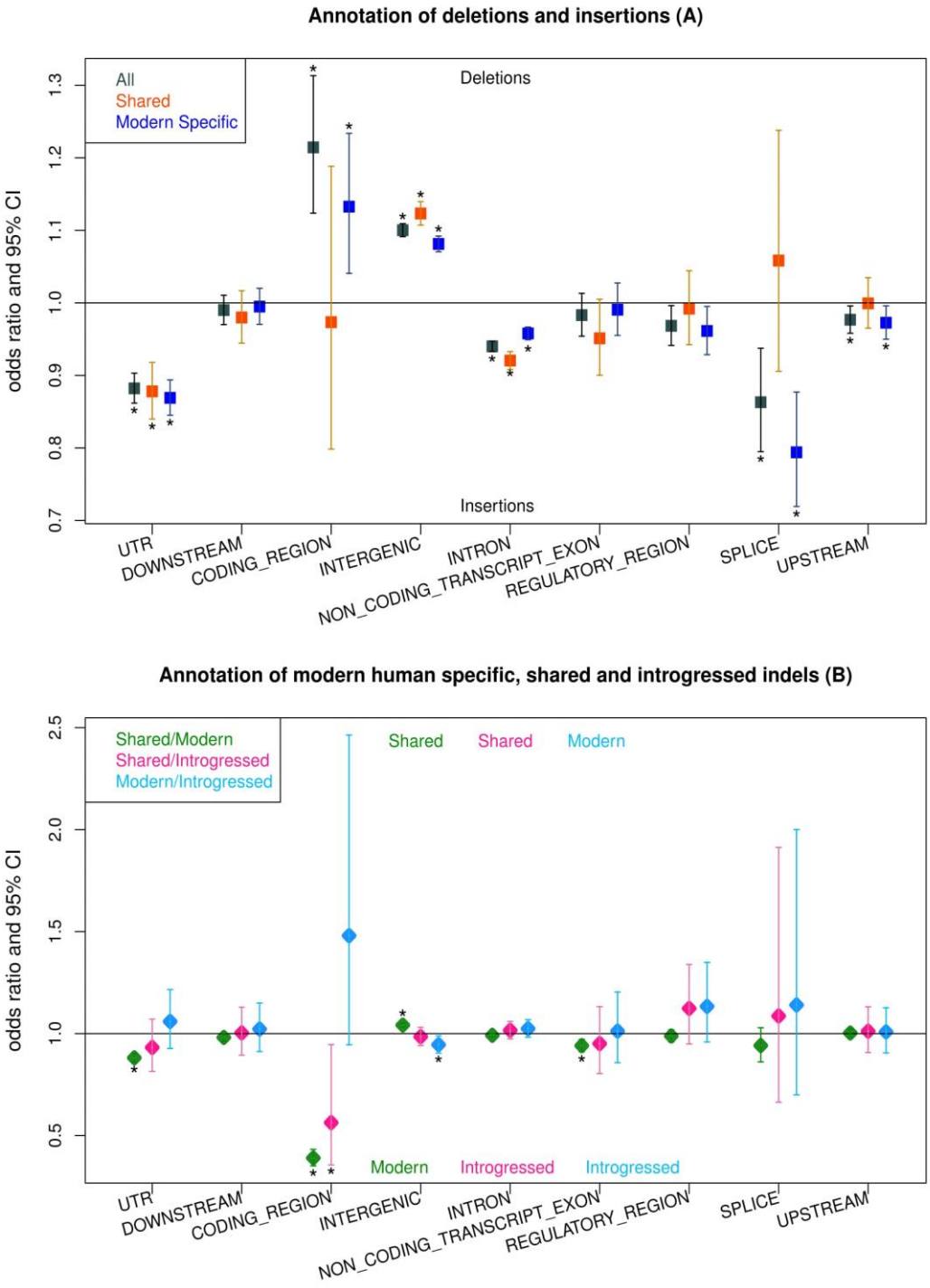


Figure 4.4 Proportion of different types of indels in classes of genomic regions. Odds ratios with 95% confidence intervals comparing A) insertions to deletions, and B) modern human specific, Neandertal shared and introgressed indels. Categories with FDR adjusted p<0.05 are marked with (*).

When annotating indels with the class of genomic regions that is most likely to influence phenotype, I find that a significantly higher fraction of Neandertal-shared indels fall in intergenic regions compared to modern human specific indels (Fisher's exact test; p =1.77e-21; False Discovery Rate (FDR) adjusted p=9.57e-21; odds ratio: 0.9599) while modern human specific indels fall more often in intronic regions compared to shared indels, although this difference is not significant after multiple testing correction (Fisher's exact test; p=0.0369, FDR adjusted p=0.083; odds ratio: 1.0087). These signals are consistent with a longer exposure to selection for Neandertal-shared indels as compared to modern human specific indels (Figure 4.4(B)). For both classes, a higher fraction of insertions resides in coding regions compared to deletions and the opposite pattern is observed for intergenic regions (Figure 4.4(A)).

4.3.6. Putatively Introgressed Indels

A subset of the indel variants segregating in non-African populations trace their ancestry back to Neandertals, through an admixture event between non-Africans and Neandertals 50-60 thousand years ago (Sankararaman et al. 2012; Carbone et al. 2014). By conditioning on the absence of the derived variant in Africans and the presence of the derived variant in Neandertals and either the East-Asian or European population, I identified 9,086 putatively introgressed indels. Of these 6,070 are deletions and 3,016 insertions with an average allele frequency of 0.027 in Europeans and 0.048 in the East-Asian population (Wilcoxon rank test for European frequencies smaller less than East-Asian frequencies: p=1.8e-35). The difference in allele frequencies between both populations is similar to the one observed for introgressed SNPs (Europeans: 0.026; East-Asians: 0.046; Figure 4.5).

Following the patterns observed for all indels, I found that a higher fraction of introgressed deletions fall in intergenic regions compared to introgressed insertions (Figure 4.6). Our previous results, comparing modern-human specific to Neandertal shared indels, remain significant when putatively introgressed indels are removed (Table 4.4).

Table 4.4 : Genic and Intergenic variants in Shared and modern human specific indels.

	Before filtering introgressed indels from Neandertal-shared			After Filtering introgressed indels from Neandertal-shared		
Category	**p-value**	**Odds ratio**	**FDR**	**P-value**	**Odds ratio**	**FDR**
Intergenic	1.77e-21	0.9599	9.57e-21	3.01e-21	0.960	3.22e-20
Intronic	0.0369	1.0087	0.083	0.0233	1.0080	0.1008

Proportion of Neandertal-shared vs modern human specific indels in intergenic and intronic regions before and after filtering introgressed indels.

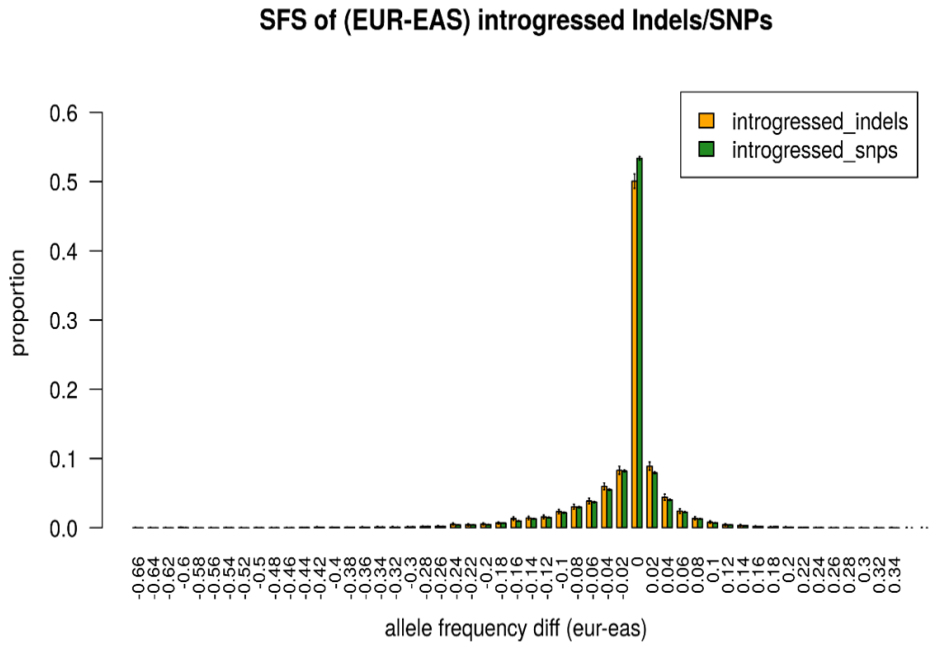


Figure 4.5 Histogram comparing the European to East-Asian allele frequency differences between indels and SNPs.

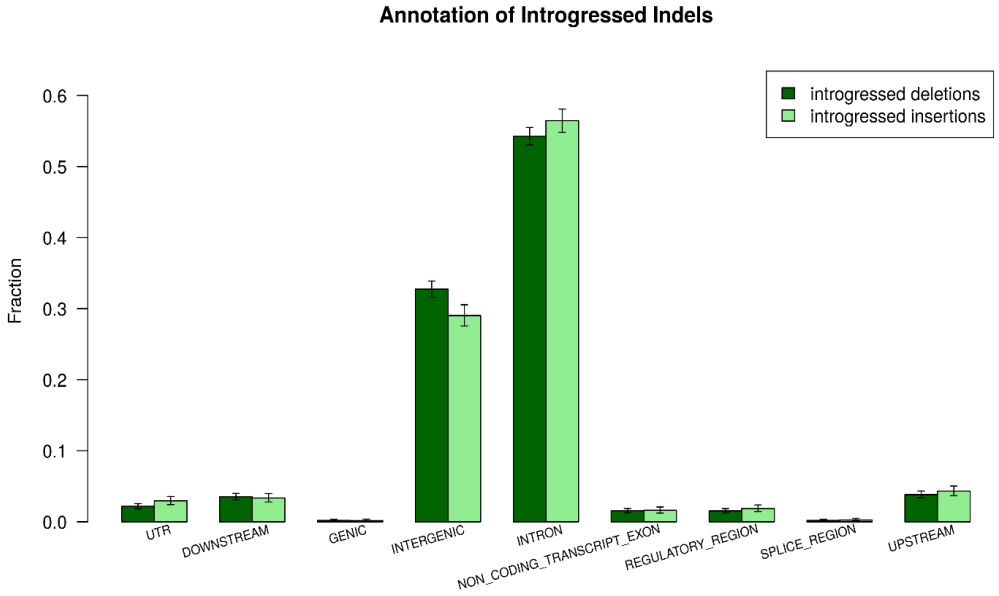


Figure 4.6 Annotation of introgressed indels
Relative frequency of variant effect predictor annotation of introgressed deletions (dark green) and introgressed insertions (light green).

To gain insight into the selection pressures that acted on introgressed indels, I compared their distribution over classes of genomic regions with those of Neandertal-shared (but without introgressed) and modern human specific indels (Figure 4.4(B)). Interestingly, I find that a slightly smaller proportion of introgressed indels fall in intron regions compared with the other two classes of indels (55.3% versus 55.7% and 55.9% for Neandertal-shared and human specific, respectively), and a slightly larger proportion of introgressed indels fall into intergenic regions (31.5% versus 31.2% and 30.3%) (Table 4.5). For Neandertal-shared variants this difference to introgressed indels is not statistically significant (Fisher's exact test, one-sided, p=0.229, odds ratio: 1.016 and p=0.258, odds ratio: 0.985 for intron and intergenic regions, respectively), while modern human specific variants show a significant difference to introgressed variants for intergenic (p=0.0074; FDR adjusted p=0.022; odds ratio: 0.945) but not intron regions (p=0.130, odds ratio: 1.024). Coding regions,

however, contain a significantly lower proportion of Neandertal-shared variants than introgressed variants (1.2% versus 2.1%, p=0.0153; FDR adjusted p=0.044) while the comparison to modern human specific indels shows a non-significant trend in the opposite direction (3.0% versus 2.0%, p=0.046; FDR adjusted p=0.101). These results raise the possibility that introgressed indels have been subjected to stronger negative selection, either before or after the introgression event, compared to modern human specific indels.

Table 4.5 Annotation of modern human and shared indels

Class	Utr	Downstream	Coding	Intergenic	Intron	NonCoding Transcript	Regulatory	Splice	Upstream
Shared	2.3	3.4	0.118	31.19	55.68	1.50	1.739	0.191	3.86
Modern	2.61	3.5	0.309	30.32	55.89	1.59	1.75	0.201	3.849
Introgressed	2.46	3.4	0.209	31.51	55.3	1.57	1.55	0.176	3.818

Percentage of indels annotated using `VEP` for Neandertal-shared indels, modern human specific indels and introgressed indels.

4.3.7. Comparison of Shared, Modern and Putatively Introgressed Indels

In order to understand the evolution of introgressed indels which are segregating in non-Africans, I use the information of indels which are recent (modern specific indels) and old indels (shared indels). The ratio of deletions to insertions (rDI) for shared indels is 1.3 compared to modern humans 2.26. However, the rDI for introgressed indels 2.01 falls somewhere in between the value of rDI for shared and modern human specific indels (Table 4.6). The ratio of introgressed indels is thus close to that of modern human specific but significantly different from shared indels. The slightly higher ratio of modern human specific compared to introgressed indels suggests that introgressed indels underwent selection either

before or after introgression from Neandertals to humans. However, the class of introgressed indels may still contain on average more deleterious alleles than the older class of shared indels.

Table 4.6 Ratio of deletion to insertions in all three categories of indels

Category	Shared	Modern Human specific	Introgressed indels
Deletions	199041	604423	6,070
Insertions	152840	266969	3016
Ratio(Deletions/Insertions)	1.30	2.26	2.01

4.3.8. Genome Wide Association Studies of Introgressed Indels

To find further evidence for a potential impact of introgressed indels on human phenotypes, I searched for introgressed indels that are in perfect linkage to SNPs that are linked to specific traits by genome wide association studies (Table 4.7). I found 9 traits ($p<1e-6$) related to neurological, immunological, developmental and metabolic phenotypes, among others. Interestingly, one SNP at chromosome 2: 157,096,776 (in perfect LD with an indel in chromosome 2: 157,099,707) is associated with menarche (Elks et al. 2010). Human carriers of the Neandertal allele showed an earlier menarche compared to non-carriers and the Neandertal allele has a higher prevalence in Europeans (allele frequency = 0.06) compared to Asians (allele frequency = 0.01).

To further corroborate that the menarche associated indel is introgressed, I plotted putatively introgressed variants in the individuals from the 1000 genomes surrounding the location of the indel (Figure 4.7). In concordance with the low frequency in present-day Europeans and East-Asians, few individuals showed the homozygous derived state for introgressed variants in the vicinity of the indel. I observe haplotypes of different lengths, two of which encompass an additional introgressed indel upstream. Regions overlapping the indel have also been found

to be introgressed in two independent maps of introgressed segments in non-Africans (Sankararaman et al. 2014; Vernot and Akey 2014).

Considering introgressed variants shared between non-African individuals, I estimate a minimum length of 180,900 bp for the introgressed segment. The recombination rate in this region is 0.23 cM/Mb, which is lower than the genome wide average of ca. 1cM/Mb (Hinch et al. 2011). I calculated the probability of a region to retain a length of at least ~180kb if it was generated by incomplete lineage sorting (see (Huerta-Sanchez et al. 2014; Dannemann et al. 2016)) and found that this scenario is unlikely (p=0.003).

Table 4.7 Introgressed indels linked to genome-wide association studies candidates.

Chr	Indel position	Snp position	Snp rsid	P-value	Gwas trait	EAS_AF	EUR_AF	Gene	PMID
1	196365712	196376474	rs16839886	7.26E-06	Age-related macular degeneration	0.013	0.075	KCNT2	(Fritsche et al. 2013)
1	209987712	209988047	rs10863790	1.00E-14	Cleft lip	0.429	0.014	NA	(Beaty et al. 2010)
1	210174981	210174417	rs11119388	4.57E-09	Cleft lip	0.445	0.009	SYT14	(Beaty et al. 2010)
14	55769446	55808151	rs17673930	1.89E-40	Protein biomarker	0.006	0.081	CHMP4BP1	(de Boer et al. 2012)
2	157099707	157096776	rs17188434	1.00E-09	Menarche (age at onset)	0.001	0.061	NA	(Elks et al. 2010)
3	23386162	23385942	rs17013049	2.78E-06	Type 2 diabetes	0.113	0.024	UBE2E2	(Cho et al. 2011)
3	100671648	100647927	rs13060137	8.96E-08	Suicide attempts in bipolar disorder	0.002	0.153	RNU6-865P	(Perlis et al. 2010)
8	20253488	20263408	rs1016646	9.45E-06	Preeclampsia	0.092	0.064	NA	(Fritsche et al. 2013)
9	87171753	87177586	rs35640669	5.17E-08	Insulin-related traits	0.055	0.035	NA	(Chen et al. 2012)

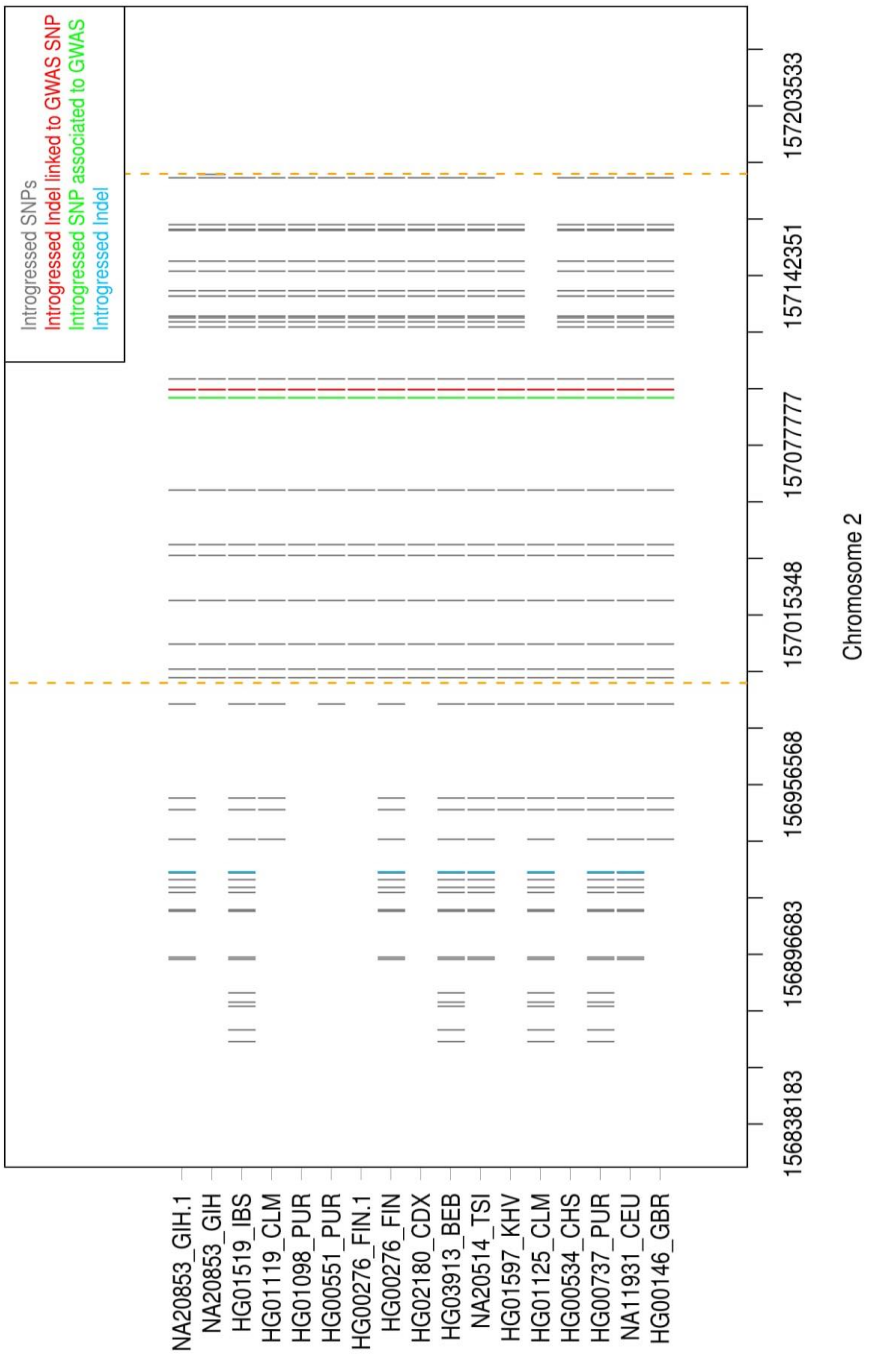


Figure 4.7 Introgressed region around an introgressed indel linked to menarche.

Introgressed haplotypes carrying introgressed indels (red) linked to an introgressed SNP associated with menarche GWAS (green) in individuals from 1000 Genomes phase 3. The borders of the shared region over all introgressed haplotypes are indicated by the dashed yellow lines.

4.3.9. Gene Ontology Enrichment

To test whether any group of functionally related genes experienced a shift in constraint from before the split to after the split from Neandertals, I used the Gene Ontology to group and compare the number of shared and modern human specific indels annotated to genes. Two Gene Ontology categories, ion channel complex and transmembrane complex, showed significant enrichment for modern human specific indels compared to shared indels (Table 4.8). This result could be explained by a relaxation of constraint for these genes in modern humans since the split from Neandertals. No significant enrichment was found in the opposite direction, or when comparing introgressed indels to shared indels.

Table 4.8 Gene ontology categories with enrichment for modern human specific changes.

Component	**Category**	**Gene ontology**	**Genes in Modern human specific**	**Genes in Shared with Neandertals**	**FWER Modern human**	**FWER Shared**
Cellular component	ion channel complex	GO:0034702	15748	7377	0.005	1
Cellular component	transmembrane transporter complex	GO:1902495	16123	7559	0.005	1

4.3.10. List of Potentially Disruptive Indels

Identifying the molecular basis for modern human specific traits remains a challenge for the study of human evolution. Here I provide a list of candidates that have been fixed in modern humans since the split from Neandertals and that are annotated as a top 1% disruptive change according to the `CADD` package (Table 4.9). Further study is needed to test whether some of these changes play a role in modern human specific traits.

In addition, I provide a list of putatively introgressed indels which have been classified as likely disruptive (Table 4.10). Variants with the highest allele frequency differences (measured by F_{ST}) between Europeans and East Asians that also show some evidence for disruptiveness are listed in Table 4.11.

Table 4.9 Top 1% c-score fixed modern human indels.

Chr	Position	Ref	alt	Type	Consequence	Gene	c-score
7	115542344	TAGAG	T	Del	Intergenic	NA	22.1
3	25739437	C	CA	Ins	Intergenic	NA	22.1
2	221679644	TAATC	T	Del	Intergenic	NA	21.9
7	156283580	CA	C	Del	Intronic	LINC01006	21.4
2	160083677	AGAGT	A	Del	Intronic	TANC1	21.3
9	119310385	CTGTT	C	Del	Intronic	RP11-264C15.2	21.1
9	119310385	CTGTT	C	Del	Intronic	ASTN2	21.1
9	37265129	C	CT	Ins	Intronic	ZCCHC7	21.0
8	65910625	ATAGT	A	Del	Intergenic	NA	20.7
12	122590799	TTC	T	Del	Intronic	MLXIP	20.6
2	168891430	C	CA	Ins	Intronic	STK39	20.2
2	144225349	CTT	C	Del	Intronic	RP11-570L15.2	20.2

2	144225349	CTT	C	Del	Intronic	ARHGAP15	20.2
2	144225349	CTT	C	Del	Intronic	AC096558.1	20.2
20	40295358	CA	C	Del	Intergenic	NA	20.2
20	38267544	GC	G	Del	Intergenic	NA	20.2
11	117229118	AAT	A	Del	Intronic	CEP164	20.1
1	108038937	G	GC	Ins	Intergenic	NA	20.1

Table 4.10 Top 1% c-score introgressed indels

This is the original table

Chr	Positions	Ref	Alt	Type	Consequence	Gene	C-score	EAS AF	EUR AF
14	74060511	T	TTCAA	Ins	Frame_shift	ACOT4	34	0	0.006
22	23011159	AG	A	Del	Frame_shift	IGLV3-27	24.8	0.014	0
6	146185477	T	TA	Ins	Frame_shift	SHPRH	24.3	0.052	0
2	236693080	CTAAT	C	Del	Upstream	AC064874.1	22.9	0.015	0
10	27687534	C	CT	Ins	Frame_shift	PTCHD3	22.9	0.224	0.012
7	21068784	T	TG	Ins	Intergenic	NA	22.8	0	0.001
22	24313530	GGA	G	Del	Frame_shift	DDTL	22.7	0.028	0.002
4	151508852	CA	C	Del	Downstream	MAB21L2	22.6	0	0.004
2	179301055	CAG	C	Del	Intronic	PRKRA	22.6	0.154	0.025
2	177503917	C	CT	Ins	Upstream	LINC01116	22.6	0.004	0
2	177503917	C	CT	Ins	Intronic	LINC01117	22.6	0.004	0
16	28915046	C	CTT	Ins	Downstream	RABEP2	22.3	0.011	0
4	117649011	CT	C	Del	Intergenic	NA	22.2	0	0.039
13	72876666	CA	C	Del	Intergenic	NA	22.2	0	0.016

1	209738399	TG	T	Del	Intronic	RP1-272L16.1	22.2	0.245	0.012
9	98096123	TAA	T	Del	Intergenic	NA	22.1	0.045	0
3	184071131	C	CCGG	Ins	Inframe	CLCN2	22.1	0	0.023
14	99742823	TTA	T	Del	Upstream	BCL11B	22.1	0.014	0
1	205293177	TAAAC	T	Del	Upstream	NUAK2	22.0	0.002	0.008
12	102125452	TATAAA	T	Del	Downstream	CHPT1	21.9	0.292	0.008
11	44026804	AC	A	Del	Downstream	RP11-613D13.4	21.9	0.029	0.001
6	69910260	TA	T	Del	Intronic	BAI3	21.8	0	0.001
14	99240729	GT	G	Del	Intergenic	NA	21.8	0.172	0.039
14	66722909	ATAAT	A	Del	Intronic	RP11-72M17.1	21.8	0	0.019
8	4762649	TA	T	Del	Intronic	CSMD1	21.7	0.044	0
14	65936081	ATAG	A	Del	Upstream	RPL21P8	21.7	0.087	0.003
8	107927163	AAC	A	Del	Intergenic	NA	21.6	0	0.003
5	175215413	TG	T	Del	Intergenic	NA	21.6	0	0.008
5	58519627	CCAAT	C	Del	Intronic	PDE4D	21.4	0.054	0.004
5	117827264	TTTAA	T	Del	Intronic	CTD-2281M20.1	21.4	0.088	0.003
18	75697997	CA	C	Del	Upstream	LINC01029	21.4	0.063	0.025
1	46966402	TA	T	Del	Intergenic	NA	21.4	0.054	0
1	14020411	C	CAG	Ins	Downstream	SCARNA11	21.4	0.022	0.021
11	94667080	CAAG	C	Del	Intergenic	NA	21.3	0	0.015
5	52608156	C	CT	Ins	Intergenic	NA	21.2	0.065	0
1	83216710	TTAAG	T	Del	Intergenic	NA	21.2	0.031	0
17	37815323	TGAA	T	Del	Inframe	STARD3	21.2	0	0.003

1	218868860	AGTTT	A	Del	Intergenic	NA	21.2	0.005	0.067
2	223154225	T	TG	Ins	Intronic	PAX3	21.1	0.001	0.001
16	79076425	TACTC	T	Del	Intronic	WWOX	21.1	0.001	0.050
5	154878448	CAAT	C	Del	Intergenic	NA	21.0	0.047	0
3	169381200	C	CT	Ins	Regulatory	NA	21.0	0.077	0.002

Table 4.11 Introgressed indels with F_{st} between Europeans and East Asians above 0.15 and c-score above 10.

Chr	Position	Ancestral	Derived	Type	Annotation	Gene	C-score	F_{st}	EAS AF	EUR AF
11	120175419	TAGAAA	T	Del	Regulatory	NA	17.71	0.600	0.603	0.002
12	102374341	T	TAG	Ins	Intronic	DRAM1	11.54	0.424	0.455	0.015
1	209986054	GTGAC	G	Del	Intergenic	NA	10.82	0.399	0.429	0.014
1	215924632	CAGT	C	Del	Intronic	USH2A	11.57	0.379	0.397	0.008
14	58320402	TA	T	Del	Intronic	SLC35F4	16.76	0.361	0.392	0.014
7	13620958	CT	C	Del	Intronic	AC011288.2	12.06	0.345	0.001	0.348
12	114745134	TA	T	Del	Intergenic	NA	12.74	0.301	0.348	0.021
3	169313681	GA	G	Del	Intronic	MECOM	10.53	0.286	0.289	0.001
2	169866296	ACT	A	Del	Intronic	ABCB11	15.80	0.278	0.279	0
12	102125452	TATAAA	T	Del	Downstream	CHPT1	21.9	0.272	0.292	0.008
12	102622272	CT	C	Del	Downstream	RP11-18O15.1	10.18	0.268	0.301	0.014
16	72814910	G	GT	Ins	Downstream	ZFHX3	12.19	0.255	0.268	0.005
6	131324263	TA	T	Del	Intronic	EPB41L2	13.38	0.253	0.261	0.003
10	27845466	CAG	C	Del	Intergenic	NA	10.66	0.238	0.263	0.001

6	131069041	AAAG	A	Del	Intergenic	NA	14.48	0.233	0.269	0.015
5	36203135	AAGAG	A	Del	Regulatory	NA	15.35	0.224	0.225	0
5	36194785	TTCTC	T	Del	3prime_utr	NADK2	11.07	0.224	0.225	0
5	36193223	CAG	C	Del	Downstream	NADK2	20.3	0.224	0.225	0
1	209738399	TG	T	Del	Intronic	RP1-272L16.1	22.2	0.216	0.245	0.012
12	125147092	ATGGCC	A	Del	Intergenic	NA	10.11	0.201	0.315	0.055

4.4 Discussion

Small indels are a common type of sequence variation among present-day humans (Mills et al. 2011). Here I used several outgroups to divide indels into derived insertions and derived deletions. Each class was further categorized using the Neandertal genome into those derived variants that are shared with Neandertals and those that are only observed in modern humans.

Previous studies have compared allele frequencies and the proportion of fixed to polymorphic insertions and deletions to gain insight into differences in selection pressures affecting each type of change. Some of these studies found that deletions appear to be more deleterious than insertions (Sjödin et al. 2010) while others found the opposite (Huang et al. 2013), a discrepancy that may in parts be explained by homoplasy, i.e. the independent formation of identical indels on several lineages (Kvikstad and Duret 2014). Here I used seven primate outgroups to reduce the effect of homoplasy and to confidently call the ancestral state. Comparing allele frequencies, fixed to polymorphic indels, and Neandertal-shared indels to modern human specific, I found that the proportion of deletions is consistently smaller for older time-frames and higher frequencies, suggesting that deletions are on average more deleterious than insertions. Interestingly, this signal is further corroborated by the genomic distribution of insertions and deletions, where I found a higher fraction of insertions in coding regions

compared to deletions, which show a higher fraction that fall in intergenic regions. Despite these consistent results, I caution that our strong requirement of several primate outgroups selects for sites that remain stable over millions of years of evolution, and that our results only holds for this subset of indels, which will be biased towards conserved and against repetitive genomic regions. I also caution that insertions and deletions are influenced by other factors than selection (Kondrashov and Rogozin 2004; Kvikstad et al. 2007; Kvikstad et al. 2009), and that they may form at unequal rates in different functional classes of the genome.

In principle, a Neandertal-shared derived variant could originate through two processes: either the variant came into existence before the Neandertal and modern human populations split, or the variant was contributed to modern humans after the split, through admixture. I make use of previous results that found Neandertal admixture in out-of-African populations to select indels that likely entered through admixture by selecting those Neandertal-shared variants that are only observed in out-of-African populations. Putatively introgressed indels showed similar differences in the genome-wide distribution of insertions and deletions, with a higher fraction of insertions residing in coding regions and a higher fraction of deletions in intergenic regions. This suggests that introgressed deletions are more deleterious than introgressed insertions.

At least 40% of the introgressing Neandertal genomes can be reconstructed from Neandertal segments segregating in out-of-African populations (Sankararaman et al. 2014) (Vernot and Akey 2014). However, the distribution of these segments has been found to be non-uniform, with genes and conserved regions of the genome showing an underrepresentation of Neandertal introgression. The patterns of depletion of Neandertal-ancestry near genes have been used to estimate the strength of selection against introgressed segments (Juric et al. 2015) and simulations suggest that Neandertals may have had a reduction in fitness compared to modern humans (Harris and Nielsen 2016). Comparing Neandertal-shared indels, which represent older events and which are mostly fixed, to putatively introgressed indels, I find no evidence for stronger negative selection acting on introgressed variants. However, compared to derived indels on the

modern human lineage, Neandertal introgressed variants show some signals that are compatible with more selective constraint, suggesting that selection acted on these variants either before or after introgression.

Some introgressed indels may also convey an advantage to the carrier and there are several examples of variants that have been positively selected after introgression (Dannemann et al. 2016; Deschamps et al. 2016; Gittelman et al. 2016; Racimo et al. 2017). Among the introgressed indels that were present in both Europeans and East-Asians and that scored highest for affecting phenotype I found a frame shift insertion in PTCHD3 (patched domain-containing protein-3), a gene which has a role in sperm development or sperm function (Fan et al. 2007) and that has been found to contain a risk-allele for asthma (White et al. 2016). However, due to the high-frequency in which null-mutations are encountered in present-day humans, the gene has also been suggested to be non-essential in humans (Ghahramani Seno et al. 2011). Some introgressed indels were also in perfect linkage with SNPs associated with different traits and diseases in genome-wide association studies. One such indel was linked to a variant associated with a decrease in the time to menarche in humans. The direction of effect for this variant is in line with research suggesting that Neandertals may have reached adulthood earlier than present-day humans (Ramirez Rozzi and Bermudez De Castro 2004; Elks et al. 2010).

4.5 Outcome

Indels in modern humans contribute not only to genetic variation, but also appear to be subject to stronger selective forces than nucleotide substitutions. Here, I studied the differences between insertions and deletions using the Neandertal genome as an additional outgroup and found signals that suggest that deletions are more often deleterious than insertions. Among the indels segregating in modern humans are those that entered out-of-African populations by admixture with Neandertals. While these introgressed indels show weak signals of negative selection compared to other variants that segregate in modern humans, I find some variants that may contribute to functional variation in present-day humans.

Arguably the most interesting variant with phenotype association is an introgressed indel variant associated with a decreased time to menarche, raising the possibility that some of the introgressing Neandertals' life history traits now form part of the modern human variation.

Chapter 5 Study of Large Scale Variation in Archaic Genomes by *de novo* Assembly

5.1 Introduction and Motivation

The field of paleo genomics is growing vastly since the last decade and the advancements in this field provided us with high coverage genomes from two archaic hominin groups: Neandertals and Denisovans. Most of the analyses using these genomes have focused on comparing single nucleotide variants. However, there are some existing studies on large scale variation in archaics. They used read-coverage based approaches to study copy number variants (Prüfer et al. 2013), polymorphic deletions in modern humans shared with archaics (Lin et al. 2015), or paired-end sequences to detect translocations in archaic genomes which may have acted as barriers to genetic exchange between Neandertals and humans (Rogers 2015).

The data from the high-coverage Altai Neandertal and Denisovan genomes constitute the currently best available data to investigate large scale variation such as rearrangements and larger insertion deletion differences to modern human genomes, in that both samples have low estimates of human contamination, are rich in endogenous DNA fraction (>70%) and yielded sequences that were enzyme treated to remove damage due to deamination. However, not all approaches applicable to modern DNA are equally useful in the ancient DNA context. For instance, split alignment analysis, which is often used to investigate large genomic variation with next generation sequencing data, is of limited use in for ancient DNA analysis since only a small fraction of ancient sequences are sufficiently long to prevent ambiguous alignments. Similarly, since only a small fraction of the ancient genome data is present in the form of paired end reads, analysis of inconsistent read pair orientation is limited in power and may also be

biased due to contamination, which can be enriched among molecules longer than the average length although such a difference is not always present (Briggs et al. 2007; Green et al. 2009).

An alternative approach which has been used also for the study modern DNA would be to *de novo* assemble the ancient genome to construct contigs and use these to detect rearrangements compared to present-day genomes. This has two advantages: first, the approach yields longer contigs that can be aligned more easily even when affected by rearrangements than shorter sequences; secondly, as the contigs are *de novo* assembled they are not affected by alignment bias that may lead to the loss of sequences that are too different, and thus aid the detection of insertions and inversion specific to the ancient genome. In this chapter, I explore this approach to ancient genome analysis using the high-coverage Neandertal and Denisovan genomes from the Denisova cave.

5.2 Methods

5.2.1 Data

The genomic data of the Altai Neandertal genome, which was sequenced to ~52x coverage (Fu et al. 2014), and the data of the Denisovan genome, sequenced to a ~30x coverage (Meyer et al. 2012), were used for assembly. The DNA from both samples was pre-treated with uracil DNA glycosylase (UDG) to remove uracils generated by deamination of cytosines. The raw sequence data was adapter-trimmed and paired-end reads were merged. Merged sequences were filtered for a minimum length of 35bp for all downstream analysis.

5.2.2 Read Correction

Read correction was performed on reads using a tool called `Musket` (Liu et al. 2013) with default parameters and a k-mer length of 29. Merged reads and unmerged paired end reads were corrected separately. To test the effectiveness of error correction, I compared the k-mer frequency spectrum before and after

correction using the program `Jellyfish` (Marcais and Kingsford 2011) and measured the frequency of ancient DNA damage associated cytosine to thymine substitutions along sequences before and after error correction.

5.2.3 Assembly

Two *de Bruijin* graph assemblers were used to assemble the error corrected reads: `SOAPdenovo` (Luo et al. 2012) and `Minia` (Chikhi and Rizk 2013). The parameters used for `SOAPdenovo`2 were –R –L 200 –M 3 -d 2 -D 3, and `Minia` which is a memory efficient implementation of DBG with parameter -abundance-min 3 in addition to the default settings. The k-mer used for assembly was estimated using kmer-genie (Chikhi and Medvedev 2014) which takes error corrected reads as an input to estimate the k-mer value at which the data has the highest fraction of unique k-mers. Although kmer-genie gives a point estimate of the best k-mer for the data (k), I processed the data with both assemblers using two additional k-mers around the best point estimate (k-4, k-2, k, k+2 and k+4) to explore the quality of the contigs produced at different k-mer lengths. The assembly with the highest N50 and longest contigs was further used for downstream analysis.

5.2.4 Contig Filtering

The assembled contigs were mapped to the human genome (hg19/GRCh37) using `BWA-MEM` with default parameters. Only mapped contigs were regarded hominin contigs and retained for downstream analysis. Unmapped contigs were further classified by mapping to a database of bacterial genomes from NCBI genbank (Benson et al. 2005). The hominin contigs were separated into contiguous alignments and split alignments. Contigs with split alignments were used to infer rearranged regions.

5.2.5 Rearrangement Calls

Contigs with split alignment yield a primary alignment (`SAMTools` flag: 0 for forward and 16 for reverse) and one or more supplementary alignments (`SAMTools` flag: 2048 for forward and 2064 for reverse).

In split alignments that occurred within the same chromosome (intra chromosomal), a gap between primary and supplementary alignments indicates a deletion in the query, whereas a split alignment with no gap between primary and supplementary alignments but with additional unaligned sequence between both parts indicates an insertion in the query. Split alignment with no gap between primary and supplementary alignments but with a contig length that is less than the mapped length must contain some parts that are mapped twice to the reference, and the region in the reference genome thus contains a duplication that is not presented in the contig. Inversions are indicated by contigs that produce three adjacent alignments in forward/reverse/forward or reverse/forward/reverse orientation (Figure 5.1).

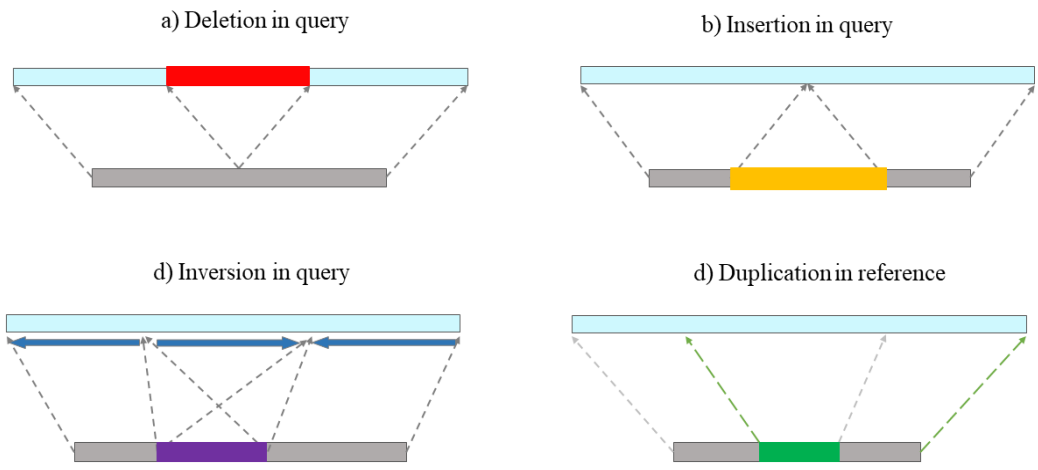


Figure 5.1 Schematic showing rearrangement calls using split mapping of contigs to a reference genome.

In addition to the configuration of the split alignment, sequence coverage with a mapping quality of 25 at split junctions was used as an additional source of information to infer rearrangements. I used the ratio of inferred intra- to inter-chromosomal as an indicator of the quality of the inference and tested the effect

of a range of different cutoffs on minimal alignment span, mapping quality and coverage at junctions on this measure.

5.3 Results

5.3.1 Error Correction of Ancient DNA Damage

A typical feature of ancient DNA is the presence of deamination at the ends of the DNA fragments, which is also referred to as ancient DNA damage (Briggs et al. 2007; Dabney et al. 2013). Although the DNA of both ancient genomes was pre-treated with UDG, an enzyme to remove uracil's created by deamination of cytosine's, some cytosine to thymine substitutions remain at the ends of the sequences due to inefficiency of the enzyme at the ends of molecules (Figure 5.2). These substitutions can cause incorrect overlap between sequences or, more likely, to overlaps between sequences to be missed.

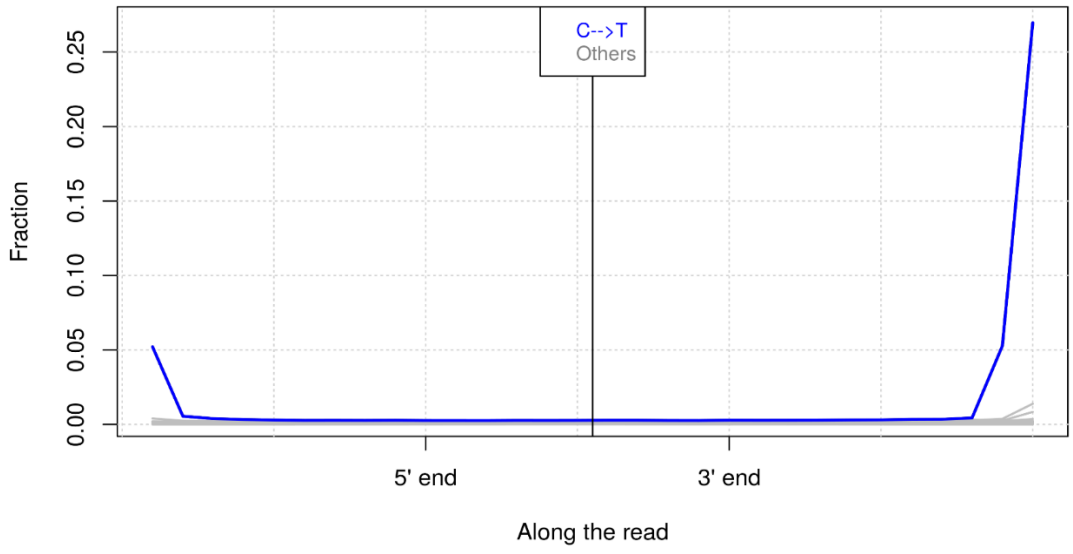
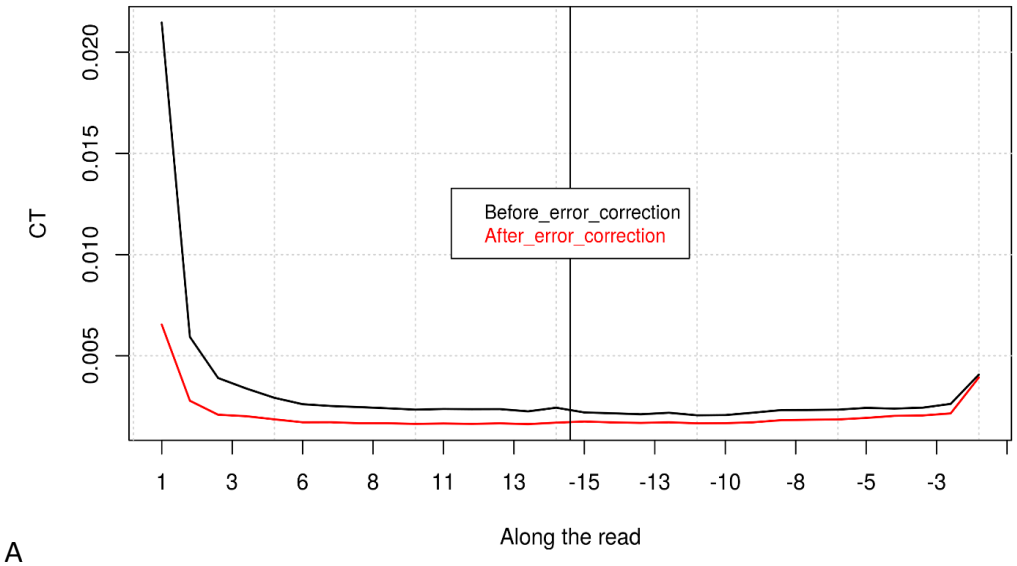
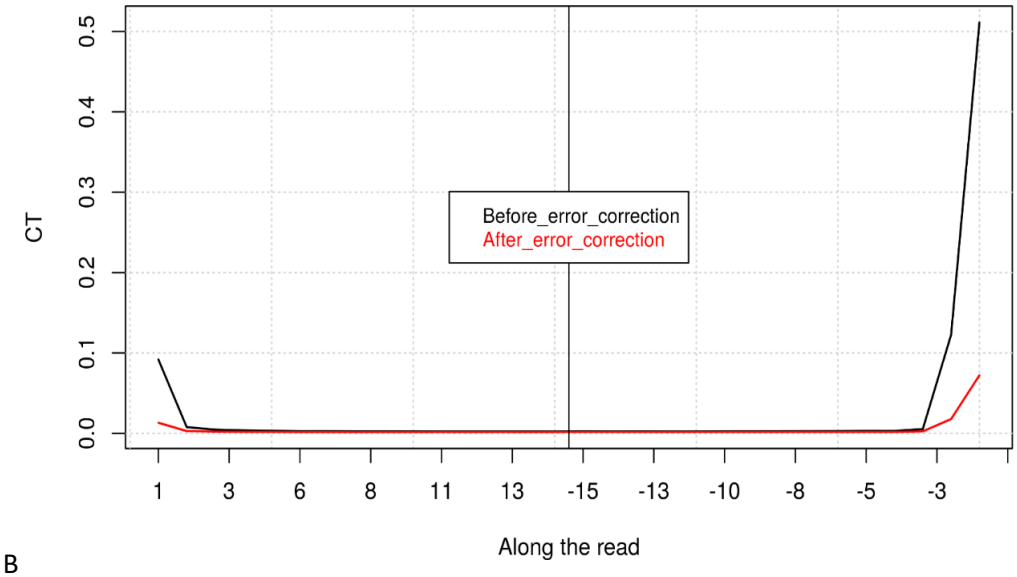


Figure 5.2 Different kinds of ancient DNA damage in Altai Neandertal genomic reads.

The plot shows all kinds of base changes accumulated over time, especially Cytosine deamination (showed in blue) to be much higher than other base changes (grey) at the ends of the reads.



A



B

Figure 5.3 Deamination patterns at the ends of the reads before and after error correction in Altai Neandertal (A) and Denisovan genome (B).

I used a k-mer-based error correction tool named `Musket` (Liu et al. 2013) to correct damaged bases in the sequenced DNA fragments. The algorithm constructs a kmer-frequency spectrum from the input sequences, i.e. the distribution of frequencies of short sequence motifs of k base pair length, where the k-mers at high frequency are considered correct and the k-mers at low frequency are considered to be the result of errors in sequences. By identifying the most likely high-frequency k-mer that gave rise to a low-frequency k-mer through an error, sequences can be corrected. Since I filter our reads to be at least 35bp long, I choose a k-mer size of less than 35 to maximize the k-mer abundance in order to error correct damage. Hence I used a k-mer value of 29 to error correct the deamination pattern at the ends of the sequences.

After correction, I observe a lower fraction of low-frequency k-mers and a higher fraction of higher-frequency k-mers. The frequency of k-mers in the error corrected sequences is larger than those before error correction for both archaic genomes (Figure 5.3).

If `Musket` is effective in removing damage-associated substitutions, I would expect that cytosine to thymine substitutions at the ends of corrected sequences are less frequent than in uncorrected sequences. I observed a significant decrease in the amount of cytosine to thymine substitutions on both 3' and 5' ends of sequences in both the archaic genomes (*p-value*= 3.55e-09 for Altai Neandertal genome and *p-value*= 5.96e-08for Denisovan genome) (Figure 5.4).

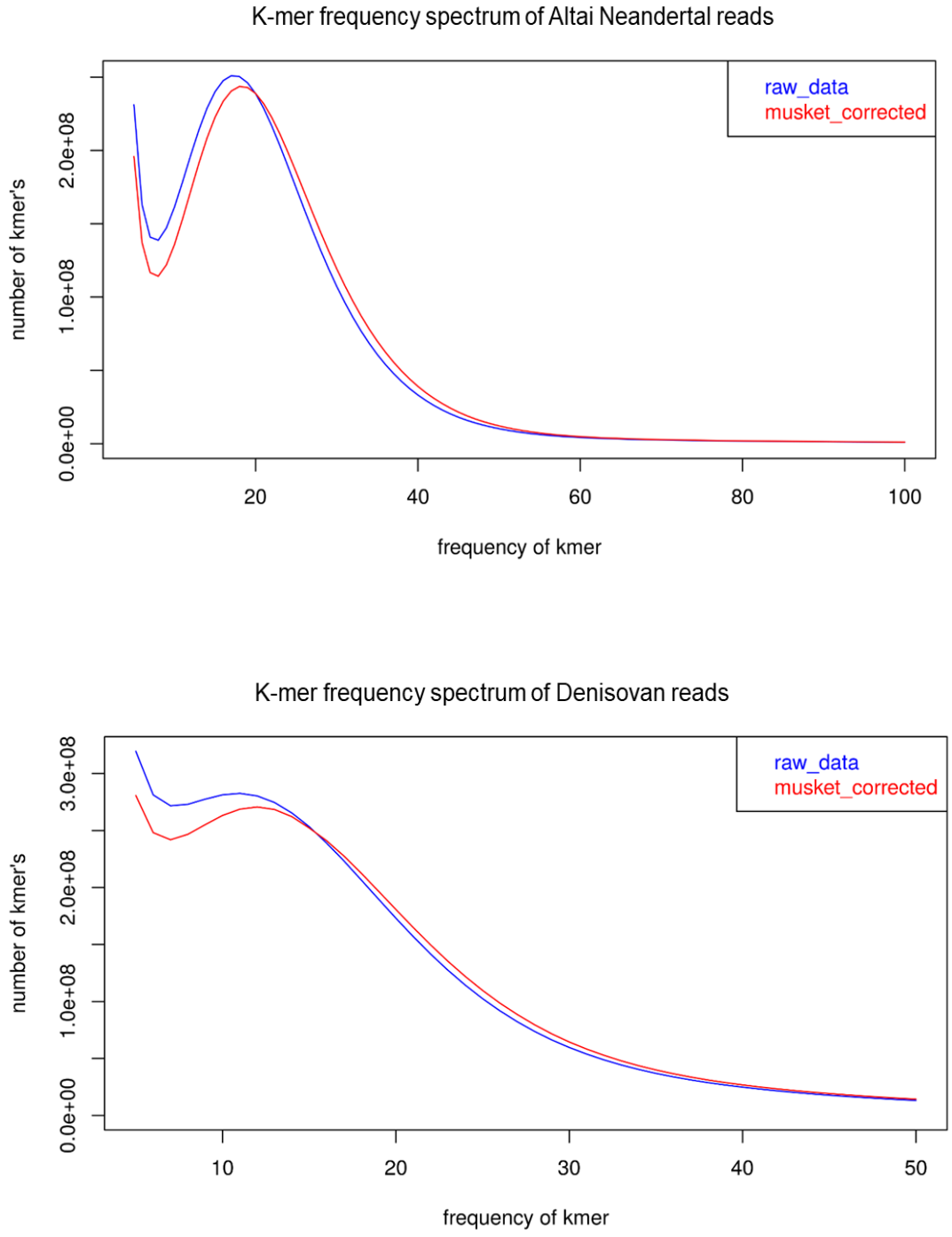


Figure 5.4 K-mer frequency spectrum before and after error correction for reads of Altai Neandertal genome (above) and Denisovan genome (below).

5.3.2 *De novo* Assembly of the Altai Neandertal Genome and the Denisovan Genome

I used two implementations of a *de Bruijn* graph approach, `SOAPdenovo2` (Luo et al. 2012) and `Minia` (Chikhi and Rizk 2013), to assemble the error corrected sequences of the Altai Neandertal and the Denisovan. The k-mers used for assemblies were estimated using `KmerGenie` (Chikhi and Medvedev 2014), yielding values of 55 and 47 for Altai Neandertal and Denisovan data, respectively. To test whether contiguity could be increased by changing the k-mer parameter, I assembled the genomes with two additional k-mers around the point estimate (k=51, 53, 55, 57, 59 for Altai and k=43, 45, 47, 49, 51 for Denisovan data).

A common measure of contiguity of assemblies is N50, the length of contigs at which the cumulative length of all contigs arranged in descending order exceeds half of the estimated length of the genome. The largest N50 over all tested k-mers was 2011bp for the Altai and 808bp for the Denisovan assembly Figure 5.5. The longest contig produced in Altai assembly was ~197kb, larger than the longest contig produced in Denisovan genome assembly ~45kb (Table 5.1). Although these numbers are promising, I caution that contigs may originate partly from microbial contaminants instead of the archaic hominin or may constitute mis-assemblies. These two issues are discussed in detail in the next sections.

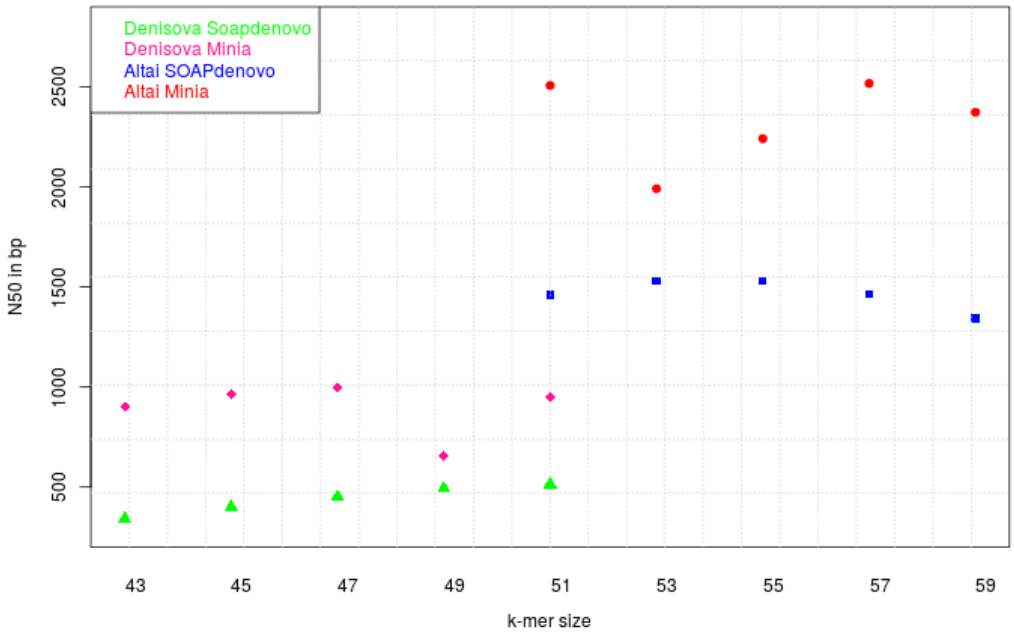


Figure 5.5 N50 of hominin contigs from Altai Neandertal and Denisovan genome assemblies using two different DBG assemblers `SOAPdenovo` and `Minia`.

Table 5.1 N50 and longest hominin contig from two different assemblers `Minia` and `SOAPdenovo` for Altai Neandertal Denisovan assembly.

Altai	**Minia assembler**				**SOAPdenovo assembler**			
	N50 from raw contigs	**N50 from hominin contigs**	**Longest raw contig**	**Longest hominin contig**	**N50 from raw contigs**	**N50 from hominin contigs**	**Longest raw contig**	**Longest hominin contig from soap**
51	1950	2507	117903	40953	1345	1459	134382	35958
53	1842	1991	191468	27638	1426	1529	197914	37590
55	2013	2236	171083	33135	1446	1531	197916	31018
57	2102	2517	111223	36871	1401	1466	197918	32914
59	2028	2374	111227	32246	1300	1343	197920	28306

Denisovan	N50 from raw contigs	N50 from hominin contigs	Longest raw contig	Longest hominin contig	N50 from raw contigs	N50 from hominin contigs	Longest raw contig	Longest hominin contig from soap
43	675	901	43243	13607	270	397	36445	13623
45	747	968	33645	14367	328	449	42612	13625
47	797	1000	35544	16476	390	494	35829	14856
49	609	656	45158	10497	437	528	35831	15435
51	808	947	29475	14865	464	541	52545	13393

5.3.3 Filtering Contigs from *de novo* Assembly

Contamination is a major problem in ancient DNA studies as fossils from which DNA is extracted are colonized by bacteria after the death of the organism and additional human contamination may be introduced during handling of the fossil. Although the human contamination estimate for both archaic genomes is less than 1% (0.8% Altai Neandertal; 0.22 % Denisovan Genome) and the fraction of archaic DNA is around 70% for both samples, a substantial fraction of sequences remains that likely represent bacterial contamination. I used the human genome (hg19), the closest available reference genome to the archaic hominins, to map the assembled contigs. This approach helped us filter out non-hominin contigs, if present. Subsequently I re-estimated N50 on hominin contigs (Table 5.1). The N50 for both assemblies increase marginally, ~2.5kb for Altai assembly and 1kb for Denisovan genome assembly. The longest assembled contig mapping to human reference genome for Altai Neandertal is ~40kb and for Denisovan assembly is ~16kb. I used the assembly with the highest estimated N50 after mapping contigs to human reference for both the archaic genomes for all downstream analysis (assembly with k-mer 51 for Altai genome, assembly with k-mer 47 for Denisovan genome).

5.3.4 *De novo* Assembly Coverage

Since the *de novo* assembly of the archaic genomes generates larger contiguous sequences compared to single sequences, it is, in principle, possible that more of the archaic genome can be mapped uniquely and compared to the human reference genome. Previous analysis of the Altai Neandertal sequences used `BWA` with more permissive parameters to align sequences to the human genome. Around 2.8Gb of the human genome are covered by aligning sequences, even when a mapping quality of 25 is applied (2.80Gb before MQ25, 2.77Gb after). However, since many regions in the genome are not unique, previous analysis employed a mapability track of 35mers to ensure correct mapping of the short ancient sequences. Applying this filter left 2.03Gb of the human genome covered. In contrast, using our assembly of Altai Neandertal and by mapping contigs to human reference using `BWA-MEM` with default parameters I cover 2.61Gb of human genome, with additional criteria of minimum mapping quality of 25 and a contig length cut-off of greater than equal to 100bp I cover 2.58Gb of human genome (Table 5.2). These estimates are similar for the Denisovan genome assembly. The assembly thus yields contigs of sufficient length, so that a larger fraction of the human genome can be covered with reliable alignments. By using this approach, I am able to cover ~582 Mb more of human genome (more than a 25% improvement) that were previously excluded due to uncertainty in the mapping of short sequences.

Table 5.2 Coverage of contigs to human reference for different k-mers and under different filtering criteria.

Altai Neandertal genome		
Altai reads	**Mq>=25**	**Mq>=25 && mapable regions**
2,834,010,540	2,765,803,904	2,027,696,381

K-mer	**Coverage of mapped contigs mq>=25;cl>=100 from minia assembly**	**Coverage of mapped contigs mq>25;cl>=100 from SOAPdenovo assembly**
51	2,583,148,543	2,662,764,654
53	2,589,507,874	2,666,532,448
55	2,604,288,811	2,667,857,496
57	2,620,771,090	2,665,868,542
59	2,630,517,936	2,659,935,363

Denisovan genome		
Denisovan reads	**Mq>=25**	**Mq>=25 && mapable regions**
2,815,409,056	2,732,881,365	2,119,839,809

K-mer	**Coverage of mapped contigs mq>=25;cl>=100 from minia assembly**	**Coverage of mapped contigs mq>=25;cl>=100 from SOAPdenovo assembly**
43	2,494,106,853	2,591,896,181
45	2,515,352,006	2,592,402,681
47	2,533,097,858	2,588,810,659
49	2,513,703,335	2,578,643,363
51	2,557,836,092	2,560,070,567

5.3.5 Split Alignment of Contigs

I use the archaic contig alignments to the human reference to identify regions which may be rearranged between humans and archaic (Neandertal/Denisovan) genomes. The majority of all aligned contigs showed one contiguous alignment (98.8%) whereas 1.2% showed a split alignment indicative of a rearrangement.

To select a subset of alignments with higher confidence for rearrangements, I require the split alignments to have a minimum contig length with a minimum mapping quality of 25 for both primary and supplementary alignments. To select appropriate cutoffs for these filters I made use of the fact that rearrangements are

expected to occur more frequently within one chromosome (intra chromosomal) than between two or more chromosomes (inter chromosomal). For this, I measure the ratio of intra/inter-chromosomal split alignments and test whether filtering increases this ratio. Before filtering, the Altai Neandertal and the Denisovan genome assembly yield more inter-chromosomal than intra-chromosomal split alignments (intra/inter ratio=0.3457 and 0.3156 for Altai and Denisova respectively). Requiring a minimum mapping quality of 25 increases the ratios to 1.33 and 0.6799. Hence, a contig length threshold along with mapping quality is improving the quality of predicted rearrangements.

Sequence coverage of the mapped original archaic sequences can be used as an additional source of information to distinguish between true and false split alignments, since true archaic sequences are not expected to span regions that are rearranged in the archaic genome compared to the human reference. When I compute the sequence coverage at all splits junctions with no filtering on contig length and mapping quality in both Altai Neandertal and Denisovan genomes, I observe a bimodal distribution (Figure 5.6). The first peak of this distribution is at coverage 2-3 in both archaic genomes whereas the second peak is close to the average genomic coverage of each genome (Altai data ~50 and Denisovan data ~30). The lowest point between these peaks is at a coverage of around 15 and 10 for Altai and Denisovan genomes respectively. Note that 15 or 10 does not constitute a strict cutoff, since contaminating modern human sequences are expected to be much too rare to yield sequence coverage as high as 15 or 10.

Using the Mapping quality and the split coverage observed from the coverage at split junctions, I aim to find an optimum contig length which yields a ratio of 50 for intra/inter chromosomal splits (Turner et al. 2008; Hansen et al. 2017). I observed that at a contig length of 60 with split coverage of 15 for

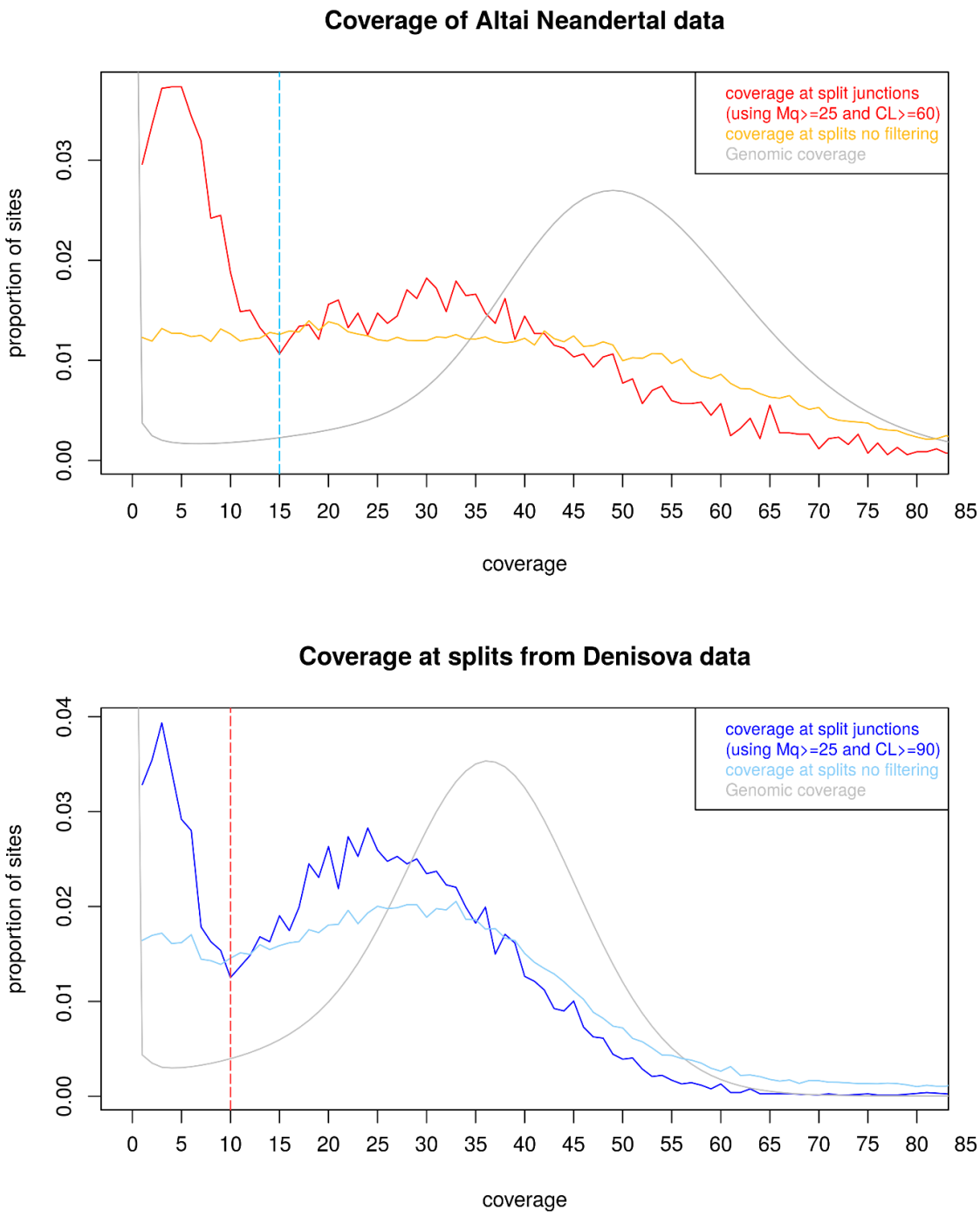


Figure 5.6 Coverage at split junctions of rearrangement calls for Altai Neandertal (red) and Denisovan (blue) genomes for different filtering criteria.

Altai Neandertal and a contig length of 90 with split coverage of 10 for the Denisovan resulted in ratios of just above 50 (Figure 5.7). I restricted all future analyses to contigs that passed all three filters with these cutoffs.

5.3.6 Rearrangements in Archaic Genome Assemblies

I used filtered split alignments to infer the type of rearrangement based on the structure of the split alignment (see Figure 5.1). I define two types of insertion/deletion differences between the modern human genome and Neandertals: 1) indels where the modern human reference genome carries additional sequence compared to Archaics (N- for Neandertals and D- for Denisovans) and 2) indels where the Archaics carry additional sequence not observed in the human reference genome (N+ for Neandertals and D+ for Denisovans). Using the split alignment of archaic contigs to human reference, I observed 2050 N- and 194 N+ sequences and 1413 D- and 105 D+ sequences in Neandertal and Denisovan assemblies, respectively, compared to human reference. To gain further insight into the excess of N- sequences compared to N+ sequences, I overlapped our detected rearranged regions with repeat regions in human genome. I observe that 1582 of N- and 1088 D- sequences overlap SINE/LINE transposable elements whereas only 25 N+ and 16 D- sequences overlap the repeat regions. This suggests that this discrepancy is at least partly driven by the fact that the human genome is of high quality allowing for most repeat insertions to be correctly identified, whereas the archaic assemblies are of lesser quality and often do not resolve repeat regions correctly.

Split alignments with no gaps between the alignments but where the orientation of the alignments is in order reverse/forward/reverse or forward/reverse/forward are inferred to be inversions. I observe 5 inversions in Neandertal assembly and 2 in Denisovan assembly, much fewer events than insertions and deletions but these events are evolutionarily more important so these are the regions of divergence accumulation and at times causation of genetic homogenization barrier (Pang et al. 2010).

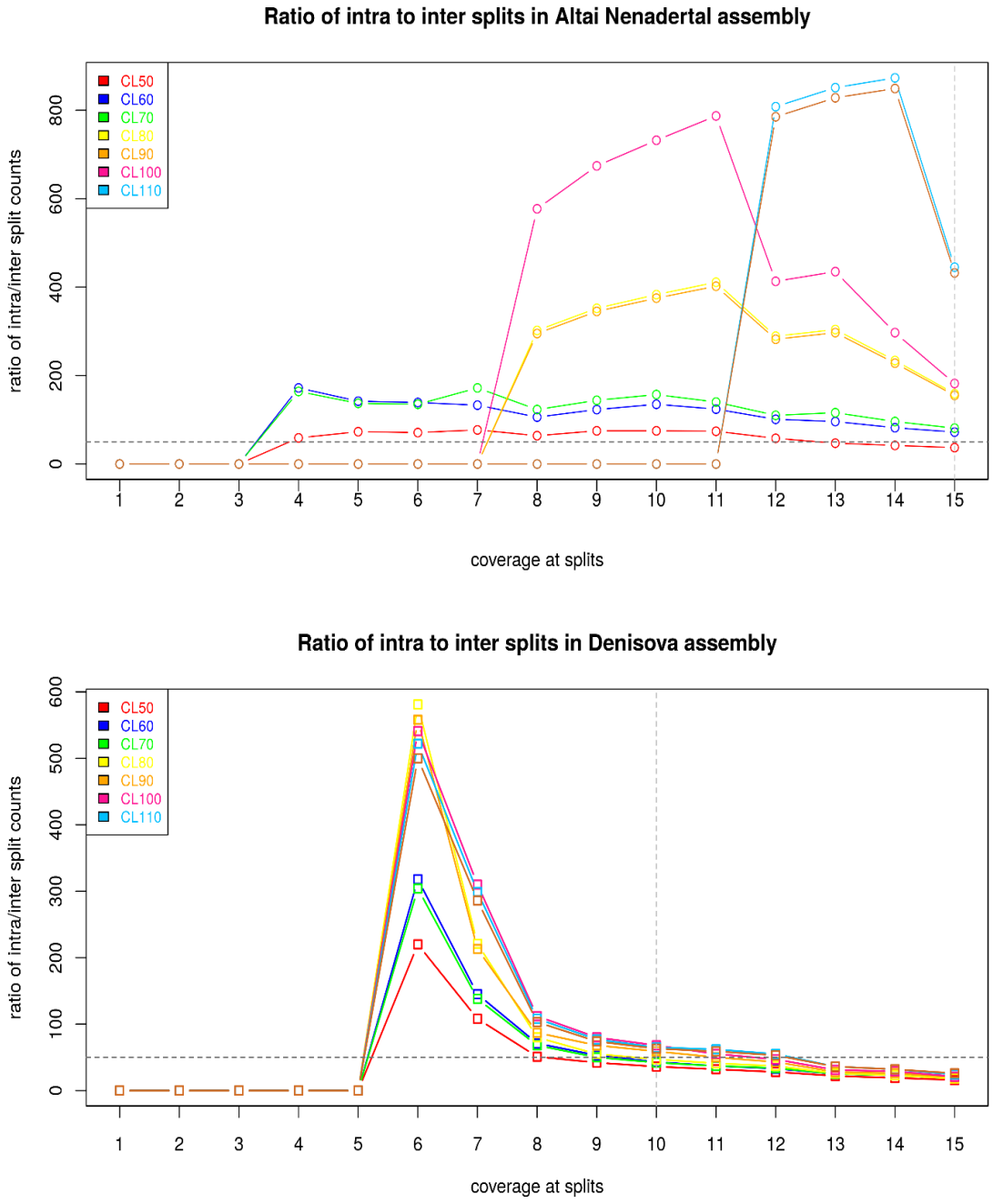


Figure 5.7 Ratio of intra to inter chromosomal splits in Altai Neandertal assembly and Denisovan assembly for different contig length filtering.

Ratio of intra/inter chromosomal splits for different contig length and split read coverage. Grey lines indicate a ratio of 50 for intra/inter with the split read coverage found from coverage plots.

Split alignments with no gaps between the alignments but where the orientation of the alignments is in order reverse/forward/reverse or forward/reverse/forward are inferred to be inversions. I observe 5 inversions in Neandertal assembly and 2 in Denisovan assembly, much fewer events than insertions and deletions but these events are evolutionarily more important so these are the regions of divergence accumulation and at times causation of genetic homogenization barrier (Pang et al. 2010).

A duplicated or deleted duplicated sequence is indicated by split alignments with no gap between the alignments where one part of the contig maps twice to the reference, i.e. the human genome carries one copy whereas the Neandertal genome carries multiple. I observe 55 rearrangements of this kind in Neandertal assembly and 44 in Denisovan assembly.

The overall rearrangement detection rate is lower for the Denisovan genome compared to the Altai Neandertal genome, in line with a lower assembly quality of the Denisovan genome than that of the Altai Neandertal (Table 5.3).

Table 5.3 Number of rearranged regions in Altai Neandertal and Denisovan assemblies.

Rearrangement category	Count in Altai Assembly	Count in Denisovan assembly
Deletions	2050	1413
Duplication	55	56
Insertion	194	105
Inversion	5	2

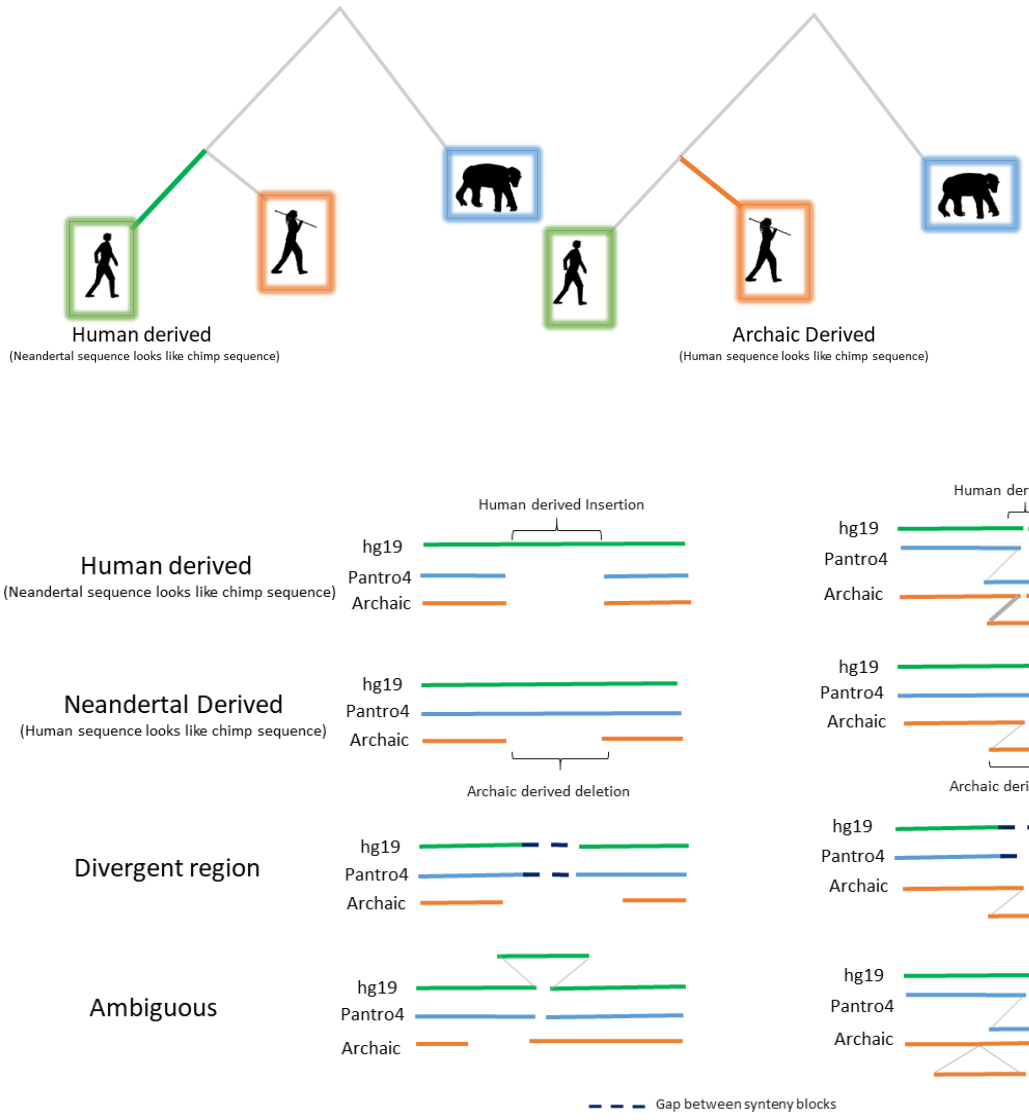


Figure 5.8 Schematic explaining derived and ancestral assignment of rearranged regions identified.

The ancestral and derived state assignment for indels detected using archaic genome assemblies mapped to human reference and hg19-chimp pairwise alignment data.

5.3.7 Ancestral and Derived State Assignment

Neandertal and Denisovan rearrangements were so far described as differences compared to the human reference genome. To further classify these rearrangements as human derived or Neandertal derived, I used the chimpanzee genome (Pantro4) to assign the ancestral and derived state. If a N-/D- sequence corresponds to a deletion compared to hg19, these sequences were inferred as ancestral and called as an insertion in human genome else these were classified as Neandertal derived deletions. Similarly, N+/D+ sequences that match the chimpanzee state in pairwise alignment of chimpanzee were inferred as human derived deletion, and as Neandertal/Denisovan derived when they did not match Table 5.4, Table 5.5. Calls were labelled as divergent if chimpanzee did not match either state, but differed by only few base pairs from the called insertion or deletion, and as ambiguous if the chimpanzee genome mismatches substantially from both options. (Figure 5.8).

Table 5.4 Counts of derived and Ancestral indels identified using Denisovan assembly

Indels identified using Denisovan assembly				
Category	**Deletions**	**Insertions**	**Deletions without repeats**	**Insertions without repeats**
Human derived	35	762	29	42
Denisovan derived	314	56	169	51
Divergent region	10	0	5	0
Ambiguous	282	13	76	7

Table 5.5 Counts of ancestral and derived rearrangements identified using Altai Neandertal genome

Indels identified using Altai Neandertal assembly				
Category	**Deletions**	**Insertions**	**Deletions without repeats**	**Insertions without repeats**
Human derived	62	1127	55	79
Neandertal derived	420	94	200	81
Divergent region	21	0	11	0
Ambiguous	339	52	138	41

5.3.8 List of Rearranged Regions Between Archaics (Neandertal/Denisovan) and Humans

By overlapping the rearranged regions with EnsEMBL genes version 78, I identified rearrangements overlapping exonic regions. These exonic regions were further filtered for those that also show exonic annotation in `IGV`, which uses the EnsEMBL gene annotation. Table 5.6 catalogues all identified indels with further classification into those detected only in this study and those previously identified.

I next analyzed in detail one deletion which was found to be present in archaic humans while it is polymorphic in present-day people (Lin et al. 2015). Figure 5.9 shows an Altai Neandertal contig which spans this deletion, indicating a deletion. In total, I observe 25 exonic rearrangements (deletions, insertion, duplications) which overlap genes and pseudogenes in all three archaic genomes (coverage support in Altai Neandertal, Vindija Neandertal genome, Denisovan genome). Eight of these rearrangements were previously found to be shared with present day humans (Lin et al. 2015), while I were unable to detect 8 previously described deletions. This indicates that the conservative filtering used here leads to false negatives and that this and previous approaches are complementary.

Among the newly detected candidates is a derived modern human specific duplication in gene ANKRD30A, an ankyrin repeat domain coding for a transcription factor that is uniquely expressed in mammary and testis epithelium. The absence of the duplication in archaics is further corroborated by coverage and the presence of nucleotide variants that are specific to one of the two copies in modern humans (Figure 5.10).

Further candidate deletions, insertions, inversions and duplications overlapping exons are shown in Figure 5.11 for the Altai Neandertal and Figure 5.12 for the Denisovan. In addition to the split mapping, these plots show the `IGV` (Integrated Genome Viewer) plots of read coverage supporting the presence of the rearrangement.

Among the newly detected candidates is a derived modern human specific duplication in gene ANKRD30A, an ankyrin repeat domain coding for a transcription factor that is uniquely expressed in mammary and testis epithelium. The absence of the duplication in archaics is further corroborated by coverage and the presence of nucleotide variants that are specific to one of the two copies in modern humans (Figure 5.10).

Further candidate deletions, insertions, inversions and duplications overlapping exons are shown in Figure 5.11 for the Altai Neandertal and Figure 5.12 for the Denisovan. In addition to the split mapping, these plots show the `IGV` (Integrated Genome Viewer) plots of read coverage supporting the presence of the rearrangement.

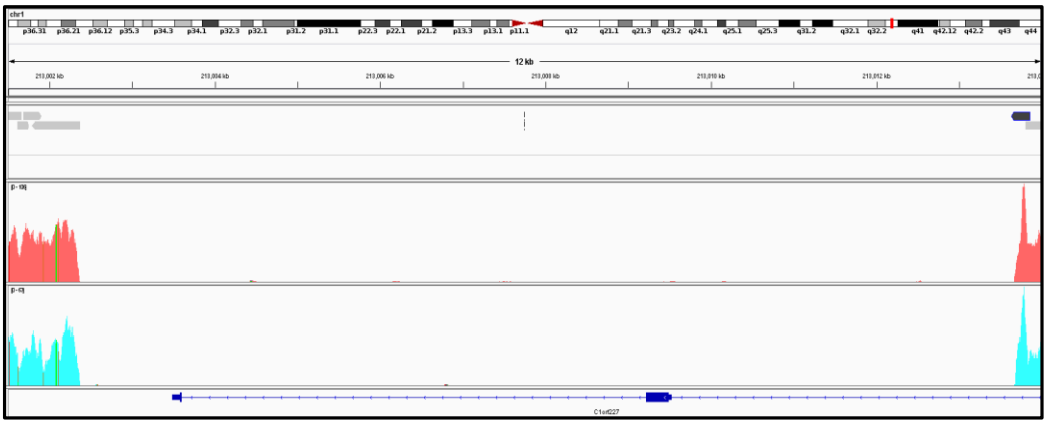


Figure 5.9 IGV of human polymorphic deletion previously identified with split mapping of contig from Altai Neandertal assembly.

This IGV figure shows the presence of previously studied deletion in (Lin et al. 2015) supported by split mapping of contig from Altai genome assembly (grey lines indicate primary alignment and black indicate supplementary alignment) along with coverage support at the breakpoints from Altai Neandertal reads (coral red) and Vindija Neandertal genome (blue).

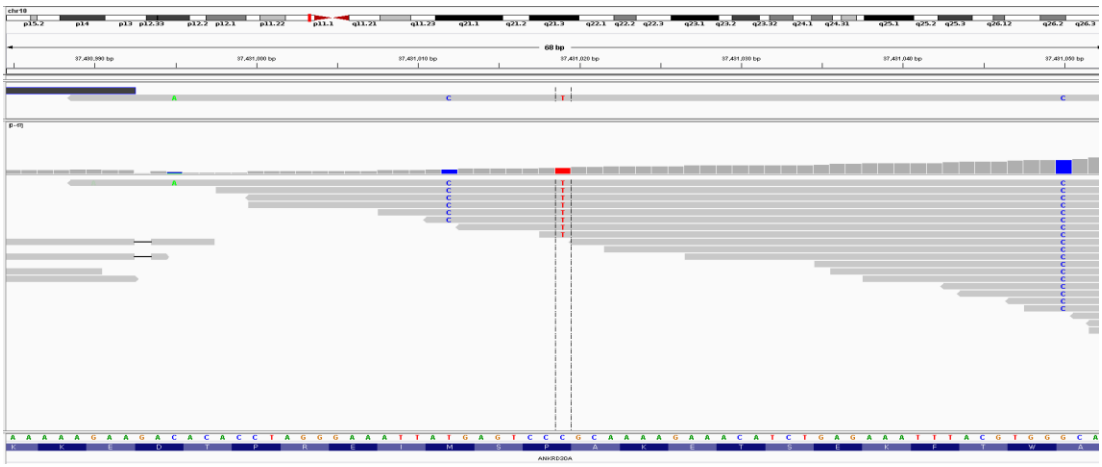


Figure 5.10 IGV of human derived duplication in ANKRD30A gene inferred using Altai contigs

The SNPs shown in the figure support deletion of a duplicated copy in Altai Neandertal genome.

Table 5.6 Rearrangements overlapping exons identified in both Altai Neandertal and Denisovan genomes using split contig mapping to human reference genome.

Chr	Start	End	Type	gene	Function	Altai Split support	Vindija coverage support	Denisovan split support	Previous study
X	152105544	152107450	deletion	ZNF185	Zinc finger	Y	N	N	N
1	32373737	32373779	deletion	PTP4A2	Protein tyrosine phosphatase	Y	Y	Y	N
1	111031022	111032229	deletion	RP11-470L19.2 /CYMP	Chymosin Pseudogene	Y	N	N	N
1	213002372	213013666	deletion	C1orf227	Spermatogenesis Associated 45	Y	Y	N	Y
5	42628554	42630991	deletion	GHR	Growth hormone reporter	Y	Y	Y	Y
7	99461394	99463563	deletion	CYP3A43	cytochrome P450	Y	Y	Y	Y
8	1733551	1733821	deletion	CLN8	CLN8, Transmembrane ER And ERGIC Protein	Y	Y	N	N
8	27662521	27662831	deletion	ESCO2	Establishment Of Sister Chromatid Cohesion N-Acetyltransferase 2	Y	Y	Y	N
8	144634068	144636240	deletion	GSDMD	Gasdermin D	Y	Y	N	Y
10	7793832	7794039	deletion	KIN	Kin17 DNA And RNA Binding Protein	Y	Y	Y	N
10	37430989	37430993	duplication	ANKRD30A	Ankyrin Repeat Domain 30A	Y	Y	N	N
11	1269344	1272625	deletion	MUC5B	Mucin 5B, Oligomeric Mucus/Gel-Forming	Y	Y	N	N
11	3238739	3244087	deletion	MRGPRG-AS1	MRGPRG Antisense RNA gene	Y	Y	Y	Y
11	60228166	60229387	deletion	MS4A1	Membrane Spanning 4-Domains A1	Y	Y	N	Y
11	128682715	128683411	deletion	FLI1	Fli-1 Proto-Oncogene, ETS Transcription Factor	Y	N	N	Y

12	9555876	9722104	deletion	DDX12P	DEAD/H-Box Helicase 12, Pseudogene	Y	Y	N	N
12	9555876	9722104	deletion	RP11-726G1.1	Psudogene	Y	Y	N	N
12	27648144	27655164	deletion	RP13-200J3.2	arginyl-tRNA synthetase (RARS) pseudogene	Y	Y	N	N
12	27648144	27655164	deletion	SMCO2	Single-Pass Membrane Protein With Coiled-Coil Domains 2	Y	Y	N	N
12	66527652	66529877	deletion	RP11-745O10.2	Protein coding	Y	Y	Y	N
12	66527652	66529877	deletion	TMBIM4	Transmembrane BAX Inhibitor Motif Containing 4	Y	Y	Y	N
14	24408476	24408497	insertion	DHRS4-AS1	DHRS4 Antisense RNA 1	Y	Y	Y	N
14	65660358	65728072	deletion	CTD-2509G16.2	Long Intergenic Non-Protein Coding RNA 2324	Y	N	N	N
17	18280750	18283242	deletion	RP1-37N7.1 /EVPLL	Envoplakin-Like Protein	Y	Y	N	N
17	79285361	79286613	deletion	TMEM105	Transmembrane protein 105	Y	Y	N	Y

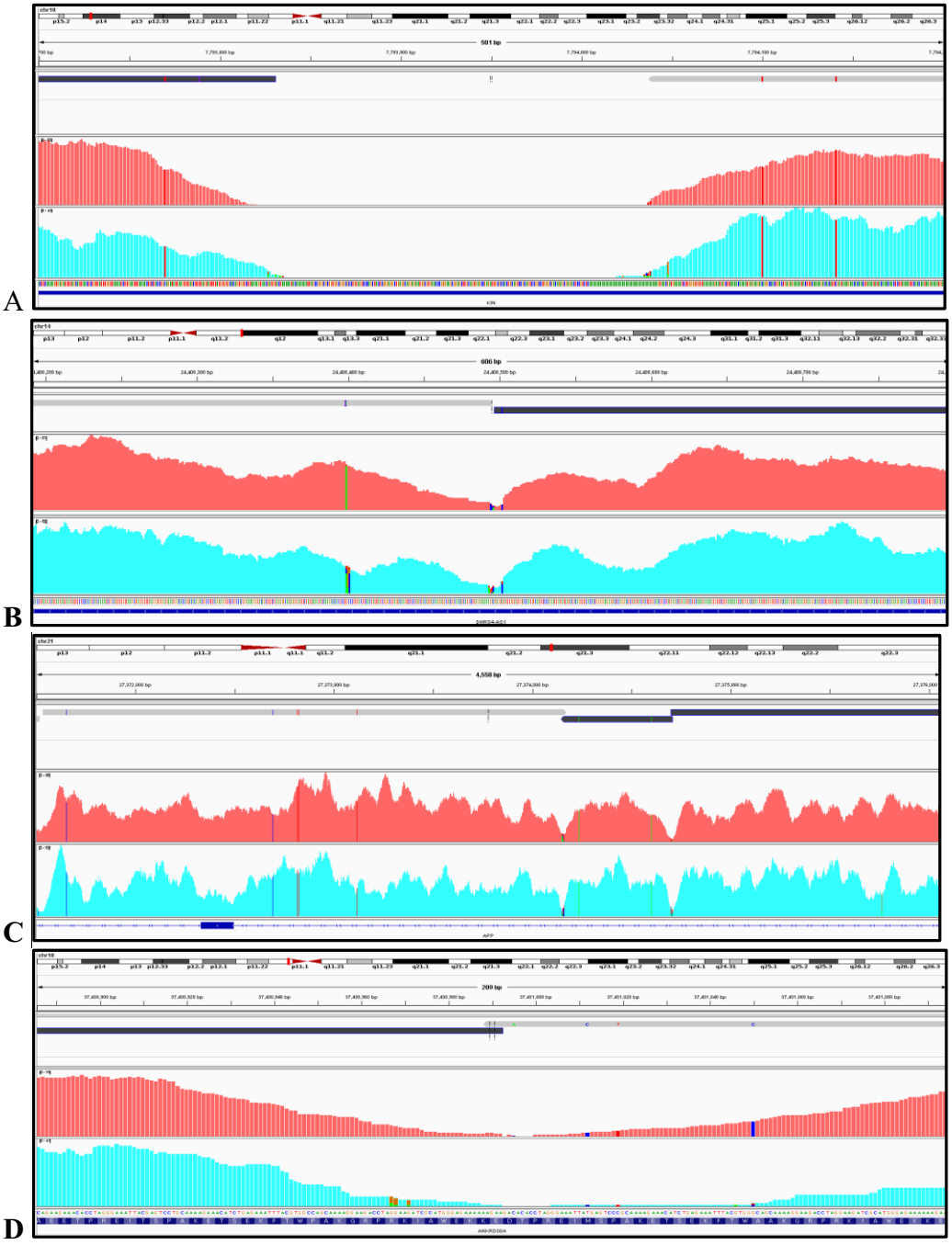


Figure 5.11 IGV images of four different kinds of rearrangements using Altai Neandertal assembly using split mapping of contigs.

The split mapping of Altai Neandertal assembly contigs (grey if primary alignment and black for supplementary alignment). Read coverage at breakpoints shown for Altai Neandertal genome (coral red) and Vindija genome (cyan blue). (A) represents a missing sequence in Neandertal with respect to human genome (B) is additional sequence in Neandertal with respect to human genome (C) is inversion in Neandertal genome and (D) is tandem duplication in human reference genome corresponding to a deletion in Neandertal genome.

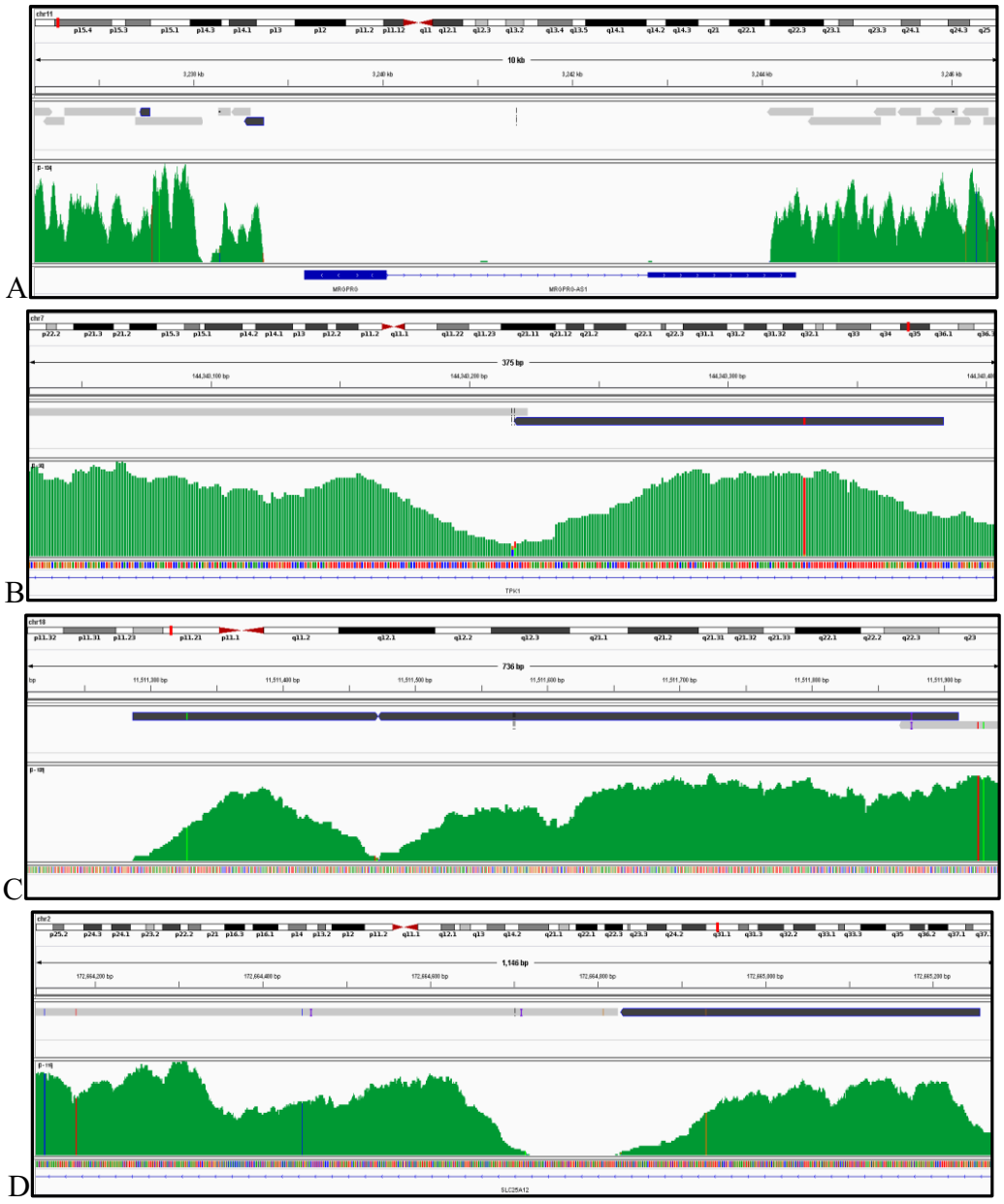


Figure 5.12 IGV images of four kinds of rearrangements in Denisovan genome assembly using split mapping of contigs.

The contigs mapped to human genome by split mapping (grey if primary alignment and black for supplementary alignment. Read coverage at breakpoints shown for Denisovan genome reads shown in green (A) represents a missing sequence in Denisovan genome compared to human reference genome (B) Additional sequence in Denisovan compared to reference human genome (C) is inversion in Denisovan (D) is tandem duplication in human reference genome corresponds to a deletion in Denisovan genome.

5.4 Discussion

Large-scale genomic rearrangements affect a larger fraction of an individuals' genome compared to single nucleotide SNPs (Pang et al. 2010). Continuous efforts are being made to characterize genomic rearrangements in humans and study their potential functional impact.

The availability of archaic hominin genomes (Meyer et al. 2012; Prüfer et al. 2013) allows in principle to further characterize human chromosomal rearrangements and two recent studies have aimed specifically at this goal. The first study used known variation within present-day humans and tested for the absence or presence of an insertion/deletion variant in archaic genomes by studying sequence coverage (Lin et al. 2015). This approach yielded a list of candidate variants that are shared with archaic humans, some of which appear to affect genes. While effective, this approach is limited to variants that are already known. A second study published in the same year looked at indels and translocations in archaic human genomes to identify regions that may act as barriers for genetic exchange (Rogers 2015). This study was carried out using paired end data from archaics and the rearranged regions were called based on discordant mapping of read pairs. This approach, in comparison to the former, allows for archaic variation to be characterized without prior knowledge. However, due to scarcity of paired end sequences in archaic genomes, the power to detect variants may be limited. Here, I used a *de novo* assembly approach that aims at overcoming the limitations of these both approaches, in that new variation can be detected while using all available sequence data.

The *de novo* assembly of an archaic genome is complicated by the properties of ancient DNA (Seitz and Nieselt 2017). Ancient DNA reads are short compared to modern DNA reads, they contain base changes that accumulate over time due to degradation and some of the sequences originate from other sources such as bacteria that invaded the sequenced material after the death of the organism (Briggs et al. 2007). These limitations are partly overcome by the fact that the Altai Neandertal and the Denisovan genomes are high coverage and low in

contaminating DNA from other organisms. An issue that remains, however, are errors due to ancient DNA damage that are expected to mask true similarity between sequences but could also cause mis-assemblies. Here, I was able to show that error correction methods that are typically used to decrease the effect of sequencing error are also effective in reducing errors due to ancient DNA damage (Liu et al. 2013).

Assembling the Altai Neandertal and Denisovan genomes with a *de Bruijn* graph approach yielded a large fraction of contigs with a length of over 1000 base pairs. I found that many of these contigs are indeed of human origin, indicating that assembly is a viable option to study rearrangements as long as ancient DNA damage is error corrected and the sample is sufficiently well preserved. Using the assembled contigs, I could cover half a mega base more of the human genome compared to short read mapping as they require stringent filtering.

In identify chromosomal rearrangements using split mapping of contigs to the human reference. To further improve the accuracy of the identified chromosomal rearrangements, I employed additional filtering that uses the ratio of called inter-chromosomal to intra-chromosomal variants as a measure of quality. This choice is motivated by the fact that intra-chromosomal variants are expected to be overrepresented at an approximate ratio of 50:1, while erroneous variant calls are expected to randomly sample chromosomes and appear to be mostly inter-chromosomal events (Turner et al. 2008). Without filtering, I detected 2050 and 1413 rearrangements in Altai Neandertal and Denisovan genome respectively with an intra- to inter-chromosomal ratio of 0.3457 and 0.3156. By filtering on a minimum mapping quality of 25 and a contig length of 60 and 90 with coverage at the split junctions of at least 15 and 10 for Altai and Denisova, respectively, the ratio of intra to inter chromosomal splits increases to 72 and 59, respectively, close to estimates of the ratio in humans (Turner et al. 2008; Hansen et al. 2017).

Out of the different categories of rearrangements, I observe a higher proportion of deletions compared to other events, similar to previous studies. To further test the accuracy of calls, I compared our detected deletions to those found in the study based on known present-day human variation. The test showed that a substantial proportion of variants were not called (50%). Among the false calls is one instance in which the variant (a deletion) is longer than the average contig length, which likely led to this variant being missed. The remaining cases show support in the contig alignment but fail the strict subsequent filters.

The comparison of these variants with the chimpanzee genome enabled us to identify 186 rearrangements for which some or all present-day humans carry the derived variant, and 281 Neandertal derived and 220 Denisovan derived variants. Among the newly identified rearrangements are several that overlap exons. One example is an insertion in Neandertals in the RNA gene DHRS4-AS1. Another is an inversion followed by a Neandertal deletion overlapping an exon of the gene SPINK14, a serine peptidase inhibitor. The analysis also yielded rearrangements that occurred on the human lineage and potentially affects genes. One of these is a duplication, present in human reference, in an exon of ANKRD30A, a ankyrin repeat domain gene which codes for a transcription factor expressed in mammary glands and testis. This duplication has been further validated by the presence of three variants in the archaics that occur in only one of the two copies.

5.5 Outcome

A detailed study of genomic rearrangements between archaics and present-day humans could help our understanding of phenotypic difference between both groups of humans. In this chapter I used *de novo* assembly to compile a list of likely rearrangements between archaic and present-day humans. The approach yielded 501 previously unknown derived variants detected in the archaic humans and 136 variants that appear derived in present-day humans. Among these variants I also detected some that overlap exons. These variants can serve as starting point for further functional testing.

Chapter 6 Conclusions

In this thesis I used genomic sequences from archaic hominins and present day humans to study two classes of mutations: indels and rearrangements, which are often excluded from evolutionary studies. But it is a well-known fact that the indels and large scale rearrangements have larger functional impact than SNPs. Nearly all previous genetic analysis involving archaic genomes were carried out using single nucleotide polymorphisms. Hence I explore the role of indels and rearrangements in archaic genomes compared to human reference.

In the first part of this thesis I presented the results of analyzing small indels on the human lineage using the Neandertal genome. The study had the aim to understand the evolutionary forces acting on deletions and insertions events that occurred at different time-frames and those that were introgressed from Neandertals. The ratio of deletions to insertions decreased with increase in age, allele-frequency and functional potential of the region, consistent with deletions being, on average, more deleterious than deletions. This result is consistent with most earlier studies. Using the ratio of deletions to insertions I was also able to infer that introgressed indels appear to be less deleterious than other variants in present-day humans. However, among the introgressed indels are also those that are associated with phenotypes in genome-wide association study. I discuss one indel that is associated with a shorter time to menarche, that represents a candidate for further study to understand the contribution of introgressed Neandertal variants to present-day human phenotypic variation.

In the second part of my thesis I identified large genomic rearrangements. I used an assembly-based approach on archaic genomes that differs from previously applied read-based approaches. However, *de novo* assembly using short reads ancient reads is complicated due to the typical properties of ancient DNA. I overcome the ancient DNA damage i.e. deamination at the ends of reads using a k-mer based error correction method. The assembly of the error corrected

reads using DBG based assemblers resulted in contigs which are at least 10 times longer than the input reads. Mapping these contigs to human reference genome and identifying split mapping contigs aided us identify rearranged regions between archaic and human genomes. This approach has the potential to identify new variants and variants of different types than these previous approaches. By using the ratio inter- to intra-chromosomal events, I show that detected rearrangements are low in error as long as sufficiently stringent filters are used. In total, I detect 2304 rearrangements in Altai Neandertal and 1576 in Denisovan genomes, some of which overlap genes and constitute candidates for further study.

In summary, my thesis provides a comprehensive analysis of non-SNP variants, ranging from small indel to larger rearrangement variants. Together with previous studies on single nucleotide polymorphisms between humans and archaics, my study provides a complete comprehensive understanding of the genomic differences between archaics and humans for a better understanding of the human specific genomic changes.

It is my hope that some of the specific classes of events that I detected, such as human-specific fixed variants that overlap genes or Neandertal-introgressed variants with phenotype-association, may prove useful for functional testing.

Bibliography

Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, Campos PF, Samaniego JA, Gilbert MTP, Willerslev E et al. 2012. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences* doi:10.1098/rspb.2012.1745.

Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, Liang KY, Wu T, Murray T, Fallin MD et al. 2010. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature genetics* **42**: 525-529.

Belinky F, Cohen O, Huchon D. 2010. Large-scale parsimony analysis of metazoan indels in protein-coding genes. *Molecular biology and evolution* **27**: 441-451.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2005. GenBank. *Nucleic Acids Research* **33**: D34-D38.

Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research* **14**: 708-715.

Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prufer K, Meyer M, Krause J, Ronan MT, Lachmann M et al. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 14616-14621.

Carbone L, Harris RA, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**: 195-201.

Chen G, Bentley A, Adeyemo A, Shriner D, Zhou J, Doumatey A, Huang H, Ramos E, Erdos M, Gerry N et al. 2012. Genome-wide association study identifies novel loci association with fasting insulin and insulin resistance in African Americans. *Hum Mol Genet* **21**: 4530-4536.

Chen L, Zhou W, Zhang L, Zhang F. 2014. Genome Architecture and Its Roles in Human Copy Number Variation. *Genomics & Informatics* **12**: 136-144.

Chikhi R, Medvedev P. 2014. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**: 31-37.

Chikhi R, Rizk G. 2013. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms for Molecular Biology* **8**: 22.

Chintalapati M, Dannemann M, Prüfer K. 2017. Using the Neandertal genome to study the evolution of small insertions and deletions in modern humans. *BMC Evolutionary Biology* **17**: 179.

Cho YS, Chen CH, Hu C, Long J, Ong RT, Sim X, Takeuchi F, Wu Y, Go MJ, Yamauchi T et al. 2011. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nature genetics* **44**: 67-72.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al. 2015. Ensembl 2015. *Nucleic Acids Research* **43**: D662-D669.

Dabney J, Meyer M, Pääbo S. 2013. Ancient DNA Damage. *Cold Spring Harbor Perspectives in Biology* **5**: a012567.

Dannemann M, Andres AM, Kelso J. 2016. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *American journal of human genetics* **98**: 22-33.

de Boer RA, Verweij N, van Veldhuisen DJ, Westra H-J, Bakker SJL, Gansevoort RT, Muller Kobold AC, van Gilst WH, Franke L, Leach IM et al. 2012. A Genome-Wide Association Study of Circulating Galectin-3. *PloS one* **7**: e47385.

De Groote I. 2011. The Neanderthal lower arm. *Journal of human evolution* **61**: 396-410.

Denisov G, Walenz B, Halpern AL, Miller J, Axelrod N, Levy S, Sutton G. 2008. Consensus generation and variant detection by Celera Assembler. *Bioinformatics* **24**: 1035-1040.

Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova JL, Patin E, Quintana-Murci L. 2016. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *American journal of human genetics* **98**: 5-21.

Elks CE Perry JR Sulem P Chasman DI Franceschini N He C Lunetta KL Visser JA Byrne EM Cousminer DL et al. 2010. Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nature genetics* **42**: 1077-1085.

Fan J, Akabane H, Zheng X, Zhou X, Zhang L, Liu Q, Zhang YL, Yang J, Zhu GZ. 2007. Male germ cell-specific expression of a novel Patched-domain containing gene Ptchd3. *Biochemical and biophysical research communications* **363**: 757-761.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* **158**: 1227-1234.

Fritsche LG Chen W Schu M Yaspan BL Yu Y Thorleifsson G Zack DJ Arakawa S Cipriani V Ripke S et al. 2013. Seven new loci associated with age-related macular degeneration. *Nature genetics* **45**: 433-439, 439e431-432.

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, de Filippo C et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**: 445.

Galvan B, Hernandez CM, Mallol C, Mercier N, Sistiaga A, Soler V. 2014. New evidence of early Neanderthal disappearance in the Iberian Peninsula. *Journal of human evolution* **75**: 16-27.

Gansauge M-T, Meyer M. 2013. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols* **8**: 737.

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.

Ghahramani Seno MM, Kwan BY, Lee-Ng KK, Moessner R, Lionel AC, Marshall CR, Scherer SW. 2011. Human PTCHD3 nulls: rare copy number and sequence variants suggest a non-essential gene. *BMC Med Genet* **12**: 45.

Gibbs RA Rogers J Katze MG Bumgarner R Weinstock GM Mardis ER Remington KA Strausberg RL Venter JC Wilson RK et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science (New York, NY)* **316**: 222-234.

Gittelman RM, Schraiber JG, Vernot B, Mikacenic C, Wurfel MM, Akey JM. 2016. Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. *Current biology : CB* **26**: 3375-3382.

Green RE, Briggs AW, Krause J, Prüfer K, Burbano HA, Siebauer M, Lachmann M, Pääbo S. 2009. The Neandertal genome and ancient DNA authenticity. *The EMBO Journal* **28**: 2494-2502.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y et al. 2010. A Draft Sequence of the Neandertal Genome. *Science (New York, NY)* **328**: 710-722.

Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. *PathoGenetics* **1**: 4.

Hansen HB, Damgaard PB, Margaryan A, Stenderup J, Lynnerup N, Willerslev E, Allentoft ME. 2017. Comparing Ancient DNA Preservation in Petrous Bone and Tooth Cementum. *PloS one* **12**: e0170940.

Harris K, Nielsen R. 2016. The Genetic Cost of Neanderthal Introgression. *Genetics* **203**: 881-891.

Hasan MS, Wu X, Zhang L. 2015. Performance evaluation of indel calling tools using real short-read data. *Human genomics* **9**: 20.

Helmuth H. 1998. Body height, body mass and surface area of the Neanderthals. *Zeitschrift fur Morphologie und Anthropologie* **82**: 1-12.

Higham T, Douka K, Wood R, Ramsey CB, Brock F, Basell L, Camps M, Arrizabalaga A, Baena J, Barroso-Ruíz C et al. 2014. The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature* **512**: 306.

Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL et al. 2011. The landscape of recombination in African Americans. *Nature* **476**: 170-175.

Huang S, Li J, Xu A, Huang G, You L. 2013. Small insertions are more deleterious than small deletions in human genomes. *Human mutation* **34**: 1642-1649.

Huang X, Wang J, Aluru S, Yang SP, Hillier L. 2003. PCAP: a whole-genome assembly program. *Genome research* **13**: 2164-2170.

Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**: 194-197.

Iakovishina D, Janoueix-Lerosey I, Barillot E, Regnier M, Boeva V. 2016. SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read mappability. *Bioinformatics* **32**: 984-992.

Juric I, Aeschbacher S, Coop G. 2015. The Strength of Selection Against Neanderthal Introgression. *bioRxiv* doi:10.1101/030148.

Kao WC, Chan AH, Song YS. 2011. ECHO: a reference-free short-read error correction algorithm. *Genome research* **21**: 1181-1192.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**: 310-315.

Kondrashov AS, Rogozin IB. 2004. Context of deletions and insertions in human coding sequences. *Human mutation* **23**: 177-185.

Kuch M, Poinar H. 2012. Extraction of DNA from paleofeces. *Methods in molecular biology (Clifton, NJ)* **840**: 37-42.

Kvikstad EM, Chiaromonte F, Makova KD. 2009. Ride the wavelet: A multiscale analysis of genomic contexts flanking small insertions and deletions. *Genome research* **19**: 1153-1164.

Kvikstad EM, Duret L. 2014. Strong heterogeneity in mutation rate causes misleading hallmarks of natural selection on indel mutations in the human genome. *Molecular biology and evolution* **31**: 23-36.

Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD. 2007. A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS Comput Biol* **3**: 1772-1782.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**: 231-239.

Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**: 1235-1247.

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular biology and evolution* **4**: 203-221.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**: 265-272.

Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B et al. 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in functional genomics* **11**: 25-37.

Lieber MR. 2010. NHEJ and its backup pathways in chromosomal translocations. *Nature Structural &Amp; Molecular Biology* **17**: 393.

Lin YL, Pavlidis P, Karakoc E, Ajay J, Gokcumen O. 2015. The evolution and functional impact of human deletion variants shared with archaic hominin genomes. *Molecular biology and evolution* **32**: 1008-1019.

Liu Y, Schroder J, Schmidt B. 2013. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* **29**: 308-315.

Locke DP Hillier LW Warren WC Worley KC Nazareth LV Muzny DM Yang SP Wang Z Chinwalla AT Minx P et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**: 529-533.

Lopez-Correa C, Dorschner M, Brems H, Lazaro C, Clementi M, Upadhyaya M, Dooijes D, Moog U, Kehrer-Sawatzki H, Rutkowski JL et al. 2001. Recombination hotspot in NF1 microdeletion patients. *Hum Mol Genet* **10**: 1387-1392.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**: 18.

Marcais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764-770.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376.

Matthee CA, Eick G, Willows-Munro S, Montgelard C, Pardini AT, Robinson TJ. 2007. Indel evolution of mammalian introns and the utility of non-coding nuclear markers in eutherian phylogenetics. *Molecular Phylogenetics and Evolution* **42**: 827-837.

Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**: 560-564.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**: 652-654.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**: 1297-1303.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**: 2069-2070.

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science (New York, NY)* **338**: 222-226.

Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, Kemeza DM, Strassler DS, Ponting CP, Webber C et al. 2011. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome research* **21**: 830-839.

Miura M, Tanigawa C, Fujii Y, Kaneko S. 2013. Comparison of six commercially-available DNA polymerases for direct PCR. *Revista do Instituto de Medicina Tropical de Sao Paulo* **55**: 401-406.

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS et al. 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome research* **23**: 749-761.

Mullaney JM, Mills RE, Pittard WS, Devine SE. 2010. Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics* **19**: R131-R136.

Mullikin JC, Ning Z. 2003. The phusion assembler. *Genome research* **13**: 81-90.

Mullis KB. 1990. The unusual origin of the polymerase chain reaction. *Scientific American* **262**: 56-61, 64-55.

Neuman JA, Isakov O, Shomron N. 2013. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Briefings in bioinformatics* **14**: 46-55.

Ophir R, Graur D. 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* **205**: 191-202.

Ottaviani D, LeCain M, Sheer D. 2014. The role of microhomology in genomic structural variation. *Trends in genetics : TIG* **30**: 85-94.

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC et al. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome biology* **11**: R52.

Peltola H, Soderlund H, Ukkonen E. 1984. SEQAID: a DNA sequence assembling program based on a mathematical model. *Nucleic Acids Research* **12**: 307-321.

Peng Y, Leung HCM, Yiu SM, Chin FYL. 2010. IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. In *Research in Computational Molecular Biology*, (ed. B Berger), pp. 426-440. Springer Berlin Heidelberg, Berlin, Heidelberg.

Perlis RH, Huang J, Purcell S, Fava M, Rush AJ, Sullivan PF, Hamilton SP, McMahon FJ, Schulze TG, Potash JB et al. 2010. Genome-wide association study of suicide attempts in mood disorder patients. *The American journal of psychiatry* **167**: 1499-1507.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 9748-9753.

Pinhasi R, Higham TFG, Golovanova LV, Doronichev VB. 2011. Revised age of late Neanderthal occupation and the end of the Middle Paleolithic in the northern Caucasus. *Proceedings of the National Academy of Sciences* **108**: 8611-8616.

Potocki L, Bi W, Treadwell-Deering D, Carvalho CMB, Eifert A, Friedman EM, Glaze D, Krull K, Lee JA, Lewis RA et al. 2007. Characterization of Potocki-Lupski Syndrome (dup(17)(p11.2p11.2)) and Delineation of a Dosage-Sensitive Critical Interval That Can Convey an Autism Phenotype. *American journal of human genetics* **80**: 633-649.

Prufer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**: 527-531.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C et al. 2013. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**: 43.

R Core Team. 2014. R: A Language and Environment for Statistical Computing.

Racimo F, Gokhman D, Fumagalli M, Ko A, Hansen T, Moltke I, Albrechtsen A, Carmel L, Huerta-Sanchez E, Nielsen R. 2017. Archaic Adaptive Introgression in TBX15/WARS2. *Molecular biology and evolution* **34**: 509-524.

Raeymaekers P, Timmerman V, Nelis E, De Jonghe P, Hoogendijk JE, Baas F, Barker DF, Martin JJ, De Visser M, Bolhuis PA et al. 1991. Duplication in chromosome 17p11.2 in Charcot-Marie-Tooth neuropathy type 1a (CMT 1a). The HMSN Collaborative Research Group. *Neuromuscular disorders : NMD* **1**: 93-97.

Ramirez Rozzi FV, Bermudez De Castro JM. 2004. Surprisingly rapid growth in Neanderthals. *Nature* **428**: 936-939.

Reiter LT, Murakami T, Koeuth T, Pentao L, Muzny DM, Gibbs RA, Lupski JR. 1996. A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nature genetics* **12**: 288-297.

Renaud G, Stenzel U, Kelso J. 2014. leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Research* **42**: e141.

Reuter JA, Spacek D, Snyder MP. 2015. High-Throughput Sequencing Technologies. *Molecular cell* **58**: 586-597.

Rogers RL. 2015. Chromosomal Rearrangements as Barriers to Genetic Homogenization between Archaic and Modern Humans. *Molecular biology and evolution* **32**: 3064-3078.

Rohland N, Hofreiter M. 2007. Ancient DNA extraction from bones and teeth. *Nat Protoc* **2**: 1756-1762.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**: 348.

Salmela L, Schröder J. 2011. Correcting errors in short reads by multiple alignments. *Bioinformatics* **27**: 1455-1461.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**: 5463-5467.

Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**: 354-357.

Sankararaman S, Mallick S, Patterson N, Reich D. 2016. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current biology : CB* **26**: 1241-1247.

Sankararaman S, Patterson N, Li H, Paabo S, Reich D. 2012. The date of interbreeding between Neandertals and modern humans. *PLoS Genet* **8**: e1002947.

Sawyer GJ, Maley B. 2005. Neanderthal reconstructed. *Anatomical record Part B, New anatomist* **283**: 23-31.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169-175.

Schwarz C, Debruyne R, Kuch M, McNally E, Schwarcz H, Aubrey AD, Bada J, Poinar H. 2009. New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. *Nucleic Acids Research* **37**: 3215-3229.

Seitz A, Nieselt K. 2017. Improving ancient DNA genome assembly. *PeerJ* **5**: e3126.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, NY)* **309**: 1728-1732.

Sjödin P, Bataillon T, Schierup MH. 2010. Insertion and deletion processes in recent human history. *PloS one* **5**: e8650.

Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS et al. 2016. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Research* **44**: D717-725.

Steinmann K, Cooper David N, Kluwe L, Chuzhanova Nadia A, Senger C, Serra E, Lazaro C, Gilaberte M, Wimmer K, Mautner V-F et al. 2007. Type 2 NF1 Deletions Are Highly Unusual by Virtue of the Absence of Nonallelic Homologous Recombination Hotspots and an Apparent Preference for Female Mitotic Recombination. *American journal of human genetics* **81**: 1201-1220.

Tattini L, D'Aurizio R, Magi A. 2015. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology* **3**: 92.

Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome research* **14**: 555-566.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69-87.

The Genomes Project C. 2015. A global reference for human genetic variation. *Nature* **526**: 68.

Toffolatti L, Cardazzo B, Nobile C, Danieli GA, Gualandi F, Muntoni F, Abbs S, Zanetti P, Angelini C, Ferlini A et al. 2002. Investigating the mechanism of chromosomal deletion: characterization of 39 deletion breakpoints in introns 47 and 48 of the human dystrophin gene. *Genomics* **80**: 523-530.

Trinkaus E, Moldovan O, Milota S, Bilgar A, Sarcina L, Athreya S, Bailey SE, Rodrigo R, Mircea G, Higham T et al. 2003. An early modern human from the Pestera cu Oase, Romania. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 11231-11236.

Turner DJ, Miretti M, Rajan D, Fiegler H, Carter NP, Blayney ML, Beck S, Hurles ME. 2008. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature genetics* **40**: 90-95.

Vernot B, Akey JM. 2014. Resurrecting surviving Neandertal lineages from modern human genomes. *Science (New York, NY)* **343**: 1017-1021.

White MJ, Risse-Adams O, Goddard P, Contreras MG, Adams J, Hu D, Eng C, Oh SS, Davis A, Meade K et al. 2016. Novel genetic risk factors for asthma in African American children: Precision Medicine and the SAGE II Study. *Immunogenetics* **68**: 391-400.

Zerbino DR. 2010. Using the Velvet de novo assembler for short-read sequencing technologies. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* **CHAPTER**: Unit-11.15.

Curriculum Vitae

Manjusha Chintalapati

Summary

Masters of technology in Bioinformatics and a PhD student in bioinformatics at Max Planck Institute for Evolutionary Anthropology. Experienced in handling NGS genomic data, ancient DNA and developing algorithms for biological systems.

Education and occupation

PhD Student *2014-2018*
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

PhD in Computer Science
Faculty of Mathematics and Computer Science
PhD Dissertation: Indels and large scale variation in archaic hominins compared to present day humans
Supervisor: Dr. Kay Prüfer (Genomes group, Department of Genetics, MPI-EVA)
Advisor: Prof Dr. Peter Stadler (Universität Leipzig)

Software programmer *2012-2014*
Tata Consultancy Services
Gachibowli, Hyderabad
India

Masters of Technology in Bioinformatics *2010-2012*
University of Hyderabad, Hyderabad, Telegana, India
Masters in Bioinformatics
Department of Life sciences
Master Thesis: Disease Causing Silent Mutations: Development of Database and Prediction algorithm
Advisor: Dr. H. A. Nagarajaram (Laboratory of Computational Biology, Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad)

Bachelors of Technology in Biotechnology *2006-2010*
Anil Neerukonda Institute of Technology, Andhra University
Bachelors of biotechnology
Bachelor Thesis: Purification of and Chemical modification studies on Snake Gourd (*Trichosanthes Anguina*) Phloem Lectin
Advisor: Prof. Musti J Swamy (School of Chemistry, University of Hyderabad)
Co-advisor: Dr. Sridevi (Head of Department, Department of Biotechnology, ANITS college of engineering, Visakhapatnam)

Talks and Posters at conferences

Next Generation Sequencing conference (NGS) organized by International society for computation biologist, Barcelona, Spain. (2014)

Talk at 13th Herbstseminar der Bioinformatik, Doubice (Czech Republic) organized by Prof. Peter Stadler, University of Leipzig. (2015)

Poster presentation at International Society for molecular bio systems (ISMB), Orlando, Florida, USA. (2016)

Talk at the Genome Informatics CSHL, Cold Spring Harbor Laboratories, New York, USA. (2017)

Technical skills

Programming languages	:	Perl, C, C++, Bash, R
Operating systems	:	Linux, Mac OS, Windows

Fellowships and achievements

Summer research fellowship from Indian Academy of Sciences for 2009-2010.

Gold medalist for the Best performance in academics during 2006-2010, in the Department of Biotechnology, ANITS, Visakhapatnam, India.

Topper of the batch in Bachelors, Department of Biotechnology for the Batches 2006-2010 consequently.

GATE (Graduate Aptitude Test in Engineering) Fellowship awardee for the year 2010-12.

Topper of the batch in Masters, Bioinformatics for the period of 2010-2012 at University of Hyderabad, Hyderabad.

Selected in **T**ata **C**onsultancy **S**ervices (TCS), Hyderabad trough campus placements 2012.

Publications

Chintalapati, Manjusha, Michael Dannemann, and Kay Prüfer. 2017. "Using the Neandertal Genome to Study the Evolution of Small Insertions and Deletions in Modern Humans." ***BMC Evolutionary Biology***. DOI: 10.1186/s12862-017-1018-8

Kay Prüfer, Cesare De Filippo, Steffi Grote, Fabrizio Mafessoni, Petra Korlević, Mateja Hajdinjak, Benjamin Vernot, Laurits Skov, Pinghsun Hsieh, Stéphane Peyrégne, David Reher, Charlotte Hopfe, Sarah Nagel, Tomislav Maricic, Qiaomei Fu, Christoph Theunert, Rebekah Rogers, Pontus Skoglund, **Manjusha Chintalapati**, Michael Dannemann, Bradley J. Nelson, Felix M. Key, Pavao Rudan, Željko Kućan, Ivan Gušić, Liubov V. Golovanova, Vladimir B. Doronichev, Nick Patterson, David Reich, Evan E. Eichler, Montgomery Slatkin, Mikkel H. Schierup, Aida Andrés, Janet Kelso, Matthias Meyer, Svante Pääbo. 2017. "A high-coverage Neandertal genome from Vindija Cave in Croatia". ***Science***. DOI: 10.1126/Science.Aao1887

Sree Rohit Raj Kolora; Anne Weigert; Amin Saffari; Stephanie Kehr; Maria Beatriz Walter Costa; Cathrin Spröer; Henrike Indrischek; Gero Doose; **Manjusha Chintalapati**; Konrad Lohse; Jörg Overmann; Boyke Bunk; Christoph Bleidorn; Klaus Henle; Katja Nowick; Rui Faria; Peter F Stadler; Martin Schlegel. 2018. Divergent evolution in the genomes of the closely related lacertids, *Lacerta viridis* and *L. bilineata* and implications for speciation. ***Giga science*** (***manuscript under revision***).

Chintalapati Manjusha, Sree Rohit Raj Kolora, Kay Prüfer. "*de novo* assembly of archaic genomes to study large scale genomic variation compared to humans" (***manuscript under preparation***).

Declaration of Authorship

Hereby I declare to have prepared the present dissertation independently and without undue foreign help. I have not used any sources or resources other than those listed, and any textually or literally taken from published or unpublished texts, and all statements based on oral information, have been identified as such. Likewise, all materials or services provided by other persons are marked as such.

I hereby accept the doctoral regulations of the Faculty of Mathematics and Computer Science of the University of Leipzig from 6, September, 2018. The submitted work was not submitted to another examination authority for the purpose of a doctoral or other examination procedure. Figures, tables and texts of this work have already been presented in parts in publications of which I am one of the main authors (see references (Chintalapati et al. 2017)).