COMPUTATIONAL METHODS OF VARIANT CALLING AND THEIR APPLICATIONS

by

JOSEPH KENNETH KAWASH

A dissertation submitted to the

Graduate School-Camden

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computational and Integrative Biology

written under the direction of

Dr. Andrey Grigoriev

and approved by

____
Andrey Grigoriev

____
Eric Klein

____
Debashish Bhattacharya

____
Ilya Serebriiskii

____
Jongmin Nam

Camden, New Jersey

January 2018

ABSTRACT OF THE DISSERTATION

Computational Methods of Variant Calling and Their Applications

By JOSEPH KENNETH KAWASH

Dissertation Director:
Andrey Grigoriev

Genome sequencing is becoming an indispensable part of biological research. Mutations identified in genomic sequence contribute to explanations of disease, phenotypic variation, and evolutionary adaptation. Increasing reliance on next generation sequencing (NGS) data necessitates efficient and accurate means of genome analysis.

We developed two algorithms, GROM-RD and GROM, to address current issues of mutation calling in NGS data. GROM-RD analyzes multiple biases in read coverage to improve copy number variation (CNV) detection in NGS data. GROM-RD takes a two-tiered approach to complex and repetitive segments, while incorporating excessive coverage masking, GC weighting, GS bias normalization, dinucleotide repeat bias normalization, and a sliding-window break-point calculator. Current NGS projects produce massive amounts of data, often on multiple samples; with several approaches designed specifically for each variant, use of multiple algorithms is necessary. GROM provides comprehensive genome analysis into a single algorithm, identifying single nucleotide polymorphisms (SNPs), indels, CNVs, and structural variants (SV), with superior sensitivity and precision while reducing the time cost up to 72 fold.

Comparative genomics studies typically limit their focus to SNVs, such as in previous comparisons of woolly mammoth and another comparison of eastern gorilla. We extended these analyses to identify SVs and indels. Our analysis found mammoth-specific variants suggesting adaptations to Arctic conditions, including variants associated with metabolism, immunity, circadian rhythms, and structural features. In gorilla populations, variants were identified that associate with physical features used to distinguish between the two subspecies. Within the gorilla subspecies was also found unique genetic evidence related to disease and abnormality, evidence of dwindling populations.

Untested and ad hoc methods of mutation calling are often used in ancient DNA (aDNA) studies. While aDNA NGS analysis is highly susceptible to aDNA degradation, many studies utilize standard mutation calling algorithms, not taking into account unique aDNA challenges of excessive contamination, degradation, or environmental damage. We present ARIADNA, a novel approach based on machine learning techniques, using specific aDNA characteristics as features, to yield improved mutation calls. In our comparisons of variant callers across several ancient genomes, ARIADNA consistently detected higher-quality variants, while reducing the false positive rate compared to other approaches.

DEDICATION

To my wife, Katarina

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Andrey Grigoriev, for all the guidance and support he provided me through our many years together. I would like to extend thanks to the members of my committee, Dr. Eric Klein, Dr. Jongmin Nam, Dr. Ilya Serebriiskii, and Dr. Debashish Bhattacharya, whose knowledge and input improved my research.

My work and my life have been enriched from the companionship I have had at Rutgers. I want to thank my lab mates, Sean and Spyros, for all of the help and teamwork we shared over the years. I owe a great deal of gratitude to Sean, whose collaboration has been invaluable. I would like to thank my closest friends, Matt, Shirley, and Steve, who have shared in this experience with me; all the struggle, encouragement, commiseration, and celebration. I want to express a special thanks to Dr. Bill Saidel, for his many years of thought-provoking conversation and perspective.

None of this would have been possible without the love and support I have received from my family. For that I want to thank my parents, and my wife, Katarina, who has (for a time, literally) been at my side throughout this journey.

Table of Contents

CHAPTER 1: INTRODUCTION

1.1 Next Generation Sequencing and Copy Number Variation

Rapid advances in DNA sequencing technologies are producing unprecedented amounts of data at a continually falling cost. This new found accessibility to whole genome sequencing combined with the development of computational tools are transforming the landscape of biological research. Next generation sequencing (NGS) is a powerful tool for the discovery of genetic causes in phenotypic variation, disease, and evolutionary adaptation (Xue et al. 2015, Prüfer et al. 2014, Lynch et al. 2015, Palkopoulou et al. 2015, Lazaridis et al. 2014, Abecasis et al. 2010, Berger et al., 2011, Campbell et al., 2010, Stephens et al., 2009, Stefansson et al., 2009, Marshall et al., 2008). A torrent of new sequencing projects not limited to thousands of human genomes, plant and animal species, has produced enormous amounts of data requiring efficient and accurate analysis.

The most ubiquitous variant in NGS data is the single nucleotide polymorphism (SNP). SNPs are prevalent in every genome, potentially the simplest to detect, and are responsible for a multitude of variation and disease. Yet genome structural variants (SVs), including; copy number variants (CNVs), insertions, inversions, and translocations, account for more differences between human genomes (Baker, 2012) in terms of the number of nucleotides and potentially have a greater impact on phenotypic variation compared to SNPs (Korbel et al., 2007).

A multitude of methods have been developed to detect SVs in NGS data. Paired-read methods identify clusters of discordant (aberrant insert size or orientation) read pairs. Split-read methods remap otherwise unplaced reads through division of the reads. Read depth (RD) methods identify CNVs by measuring variance in read coverage. De-novo methods assemble reads into contigs, accounting for differences against the reference.

Yet detection is complicated by GC bias inherent in NGS technologies, as read coverage varying with regional GC content. This problem is especially persistent in RD analysis methods. Existing RD methods reduce GC bias by GC bin mean normalization (Abyzov et al. 2011, Yoon et al. 2009), polynomial fitting (Boeva et al. 2011), and LOESS regression (Miller et al. 2011). However, these methods do not consider the differences in read depth variance that coincides with GC content, even after GC bias correction. Complex and repetitive regions are also challenging for CNV detection methods. Regions near telomeres and centromeres are known to be SV hotspots (Mills et al., 2011) and sequencing bias has been frequently observed in repeat regions (Ross et al., 2013). To overcome these challenges, we developed GROM-RD, an algorithm that analyzes multiple biases in read coverage to find CNVs in NGS data.

1.2 Comprehensive Mutation Calling in Next Generation Sequence Data

With the rapid decline of NGS cost and increasing throughput, large scale genomics projects are becoming more feasible and prevalent. One of the first such projects, the 1000 Genomes Project (Abecasis et al. 2010) was launched in 2008 with the goal of producing and analyzing whole genome sequencing for 1,000 genomes. This is now being followed by Human Longevity, Inc.’s 10,000 publicly available WGS genomes (Telenti et al 2016), the United Kingdom’s 100,000 Genomes Project (Genomics England 2017), and even larger sequencing projects involving 1,000,000 participants in the United States with the Precision Medicine Initiative (NIH All of Us Reseach Program 2017) and Million Veteran Program (USVA Million Veteran Program 2017), and China’s own million genomes project (Cyranoski et al 2017). Such projects produce massive amounts of data, straining computational resources and requiring much faster analytical methods than currently available (Stephens et al 2015).

Comprehensive analysis of genomic differences in these studies requires the detection of a wide range of variants, including SNVs, indels, CNVs and SVs; including deletions, duplications, insertions, inversions, and translocations. While methods have been developed for each type of variant, a typical WGS analysis work flow requires running multiple algorithms for complete analysis. This is wasteful of computational resources and drastically increases the time required for WGS analysis. We therefore developed, Genome Rearrangement Omni-Mapper (GROM), a novel comprehensive method of variant detection, combining mismatch, split-read, read pair, and read depth WGS evidence to detect all variant types in a single process.

We applied both GROM-RD and GROM to perform a comparative analysis of variants in several published genomes, including those of woolly mammoths, elephants and two gorilla populations.

1.3 Woolly Mammoth and Asian Elephant

The woolly mammoth (*Mammuthus primigenius*) is an ancient species of megafauna that inhabited the upper Arctic regions of the globe until the demise of the most recent population about 3500 years ago. The last known population was a small group located in the far north of the globe on Wrangel Island (Palkopoulou et al. 2015, Vartanyan et al. 2008). Though the woolly mammoth did not persist, its sister species, the Asian elephant, has successfully carried on to the present date. Woolly mammoths lived in a cold, dry steppe-tundra where average winter temperatures ranged from -30°C to -50°C, much different from the tropical and subtropical environments of modern African and Asian elephants (MacDonald et al. 2012). Mammoths had many anatomical adaptations for success in its harsh environment, such as thick fur, small ears, and small tails (compared with modern elephants), and a thick layer of fat under the skin to reduce heat loss and possibly serve as a heat source or fat reservoir for the winter (Fisher et al. 2012, Hill et al. 2009, Repin et al. 2004). Despite being studied extensively anatomically, little is known about the genetic divergence and evolution between woolly mammoth and Asian elephant.

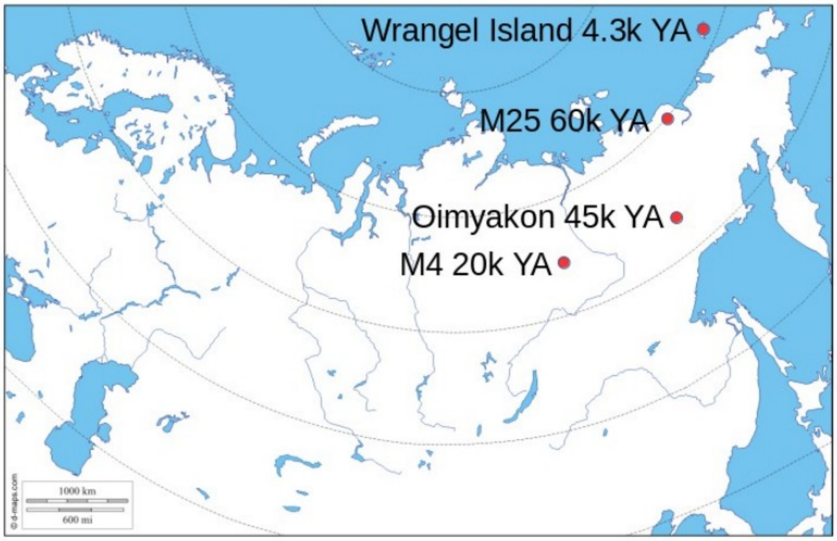


Figure 1. The locations of woolly mammoth samples studied.

Until recently, whole genome sequence availability of woolly mammoth specimens has been sparse and typically of low quality, a common challenge for ancient genomes; therefore, few comparative genetic studies have been performed. These studies have been limited in sample size as well as breadth of mutation type. In 2008, the first whole genome sequencing (<1×) of a woolly mammoth was published (Miller et al 2008). However, only recently have high coverage WGS datasets become available from two studies that identified SNVs unique to woolly mammoths to infer the genetic basis of adaptations to the Arctic (Lynch et al. 2015) and to analyze species diversity prior to extinction (Palkopoulou et al 2015). We analyzed the patterns of variation across the genomes of these four mammoths and compared them to the available elephant genomes. Using GROM-RD and GROM, we systematically identified variants, ranging from single nucleotide changes and short indels to deletions and amplifications of regions encompassing gene fragments and complete genes.

1.4 Eastern Gorilla

Gorilla gorilla are a great ape species that populate a wide range of the central African continent. The eastern populations consist of the closely related subspecies, eastern lowland gorilla, *Gorilla graueri*, and the mountain gorilla, *Gorilla beringei*. Though all gorillas are currently threatened, the eastern species are of specific interest due to their low numbers. Mountain gorillas are estimated to have a total population of approximately 800 individuals (Gray et al. 2013) while eastern lowland are estimated to have a population size of about 3,800 (Plumptre et al. 2016). Conservation efforts have prevented further decline of gorilla populations, however already diminished population sizes may reduce genetic diversity and affect long term survival of the species (Xue et al. 2015).

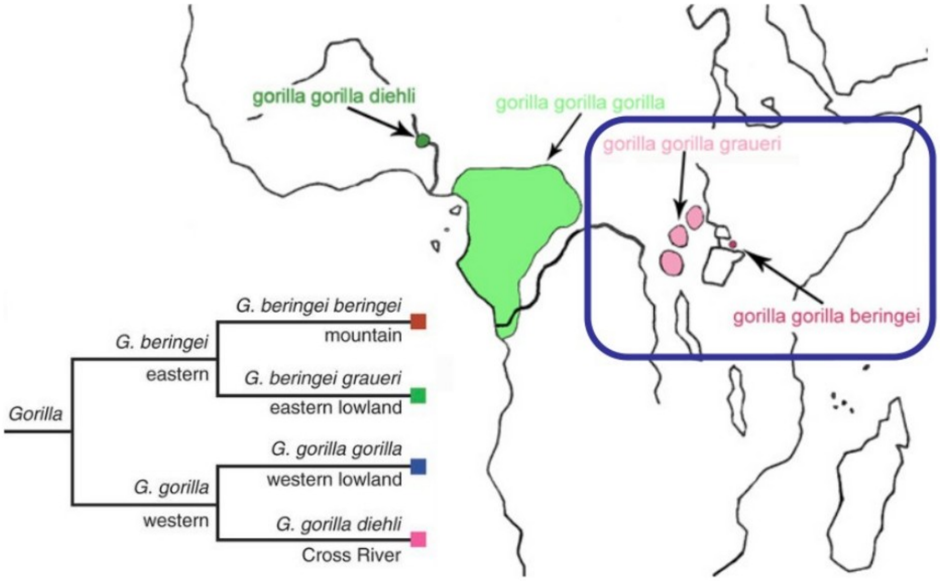


Figure 2. Locations of gorilla subspecies populations used in this study. Figure modified from Xue et al. 2015.

Though the eastern lowland gorilla and the mountain gorilla are closely related and geographically separated by a relatively small distance, they have distinct habitats and the populations are not reported to intermix. As well as being geographically distinct, both groups have distinct physical characteristics that are used to discern between the subspecies: The eastern lowland gorilla is reported to have shorter hair, a larger nose, and a greater incidence of age-related balding. The mountain gorilla is larger in size and heftier, though with shorter arms than the eastern lowland relatives. The mountain gorilla also has a larger and more prominent jaw and teeth (Fossey 1983, Pilbrow 2010, Schultz 1934, Stanford 2001). Previous work by Xue et al. focused on the analysis of SNVs. We analyzed whole genome sequence for 12 gorillas from these two subspecies, seven mountain gorillas and five eastern lowland gorillas, expanding the comparative analysis of these two gorilla populations to include such structural variation.

1.5 Ancient Genome Sequence Analysis

Prior to the development of next generation sequencing (NGS) methods for aDNA sequencing, comparative analysis relied on the physical analysis of remains. The combination of increased quality from NGS technology as well as new methodology for reliable extraction and library preparation of ancient DNA samples has led to an influx of genomic studies of extinct species (Prüfer 2014, Parks 2015, Rizzi 2012, Gansauge 2014, Prüfer 2010). Early studies adopted mitochondrial genome sequence analysis representing a leap forward in evolutionary research. Until now, many such aDNA studies have been mostly limited to mitochondrial and small genome regions (Krings 1997, Krings 1999, Noro 1998, Krause 2006, Green 2006), given the problems with extraction of usable DNA in sufficient quantity (Rasmussen 2010, Rizzi 2012, Green 2009, Prüfer 2010). Recent advances have enabled amplifying enough nuclear DNA to allow for complete genome sequencing of ancient samples, although also leading to potentially compounding effects on coverage, quality, and contamination (Rasmussen 2010, Malmstrom 2005, Rizzi 2012, Gansauge 2014). Accuracy in these studies is especially important for comparative and evolutionary analysis against living species (Prüfer 2014, Lynch 2015, Palkopoulou 2015, Lazaridis 2014, Rasmussen 2010, Schuenemann 2017). Experimental and computational methods were developed to mitigate complications of aDNA sequencing including contamination from microbes and human handling, fragmentation, depurination, and deamination (Rasmussen 2010, Sawyer 2012, Noonan 2005, Hoss 1996, Briggs 2007, Brotherton 2007, Paabo 1989, Paabo 2004, Hofreiter 2001, Malmstrom 2005, Rizzi 2012, Green 2009).

Such methods often rely on detecting degradation of aDNA due to extensive exposure to the environment and the physical handling of samples over time (Rasmussen 2010, Noonan 2005, Hoss 1996), which is used to differentiate between sample and noise (Sawyer 2012, Briggs 2007, Brotherton 2007, Paabo 1989, Gansauge 2014). Filtering out contamination of contemporary DNA from aDNA samples uses short read lengths, as aDNA is often highly fractured; thus long read lengths are comparatively rare and likely a contemporary contaminant (Parks 2015, Sawyer 2012, Paabo 2004, Green 2009, Paabo 1989). Other features can be used for reducing error in aDNA studies, such as substitutions arising from depurination events frequently occurring before strand breaks (Briggs 2007), or deamination events often found at the ends of fragments (Briggs 2007, Brotherton 2007). Compensation for these nucleotide change events is frequently made by masking such substitutions if they occur towards the end of reads (Lynch 2015, Parks 2015).

Although these methods have been shown to decrease the bias caused by aDNA damage and contamination (Sawyer 2012, Rizzi 2012, Gansauge 2014), there is yet to be a consensus method to address the issue of mutation calling. Typical approaches are often ad hoc extensions of existing algorithms, such as GATK (McKenna et al. 2010). However, there is little similarity among these extensions in the filtering of read depth, quality, masking locations, or mapping characteristics (Prüfer 2014, Lynch 2015, Palkopoulou 2015, Lazaridis 2014, Rogers 2016, Parks 2015). For example, recent publications on the sequencing of various woolly mammoth and ancient human whole genomes have all utilized differing methods in the quality control of sequencing reads despite working with similar datasets (Rogers 2016). Additionally, a study in neandertal genomes has demonstrated that the use of GATK on even highly processed sequence data potentially yields inaccurate results (Prüfer 2017).

Due to the variation of quality and quantity of usable DNA in ancient samples, and divergent methods to extract the maximum amount of information, there can be large discrepancies in aDNA findings and interpretations (Lynch 2015, Palkopoulou 2015, Lazaridis 2014, Rogers 2016, Rasmussen 2010, Parks 2015). Many of the currently employed variant calling algorithms are utilized with limited validation of results or proof of efficacy due to the constraints of aDNA sample availability.

Here, we introduce ARIADNA, a novel approach using machine learning for detecting SNV variants in ancient DNA samples. In essence, it uses our fast GROM genome scanning engine (Smith 2017) to find all potential SNVs (PSNVs) found as deviations between sample and reference genomes and then utilizes a boosted regression tree algorithm for training and classification of potential mutation sites. The unique features of the corresponding sites are used by our algorithm to determine the difference between bona fide mutations in aDNA and noise due to aDNA degradation or contamination. We compared ARIADNA results on (i) woolly mammoth genomes with the most commonly employed mutation caller, GATK, and (ii) the Altai neandertal genome with output from Prüfer in 2014 and 2017. Our comparisons demonstrate that ARIADNA provides the most accurate and comprehensive mutation call sets in these ancient genomes.

Chapter 2: GROM-RD: resolving genomic biases to improve read depth detection of copy number variants

Smith, Sean D., Joseph K. Kawash, and Andrey Grigoriev. "GROM-RD: Resolving genomic biases to improve read depth detection of copy number variants." PeerJ 3 (2015): e836.

We developed GROM-RD, an algorithm for identifying copy number variation in next generation sequencing data. GROM-RD has several unique features that are not incorporated in other algorithms such as; excessive coverage mapping, quantile normalization of GC bias, dinucleotide repeat bias normalization, and a variable sized sliding window for the identification of break points. Utilizing these unique approaches to NGS analysis of read depth, GROM-RD outperformed CNVnator and RDXplorer with improved CNV detection and breakpoint accuracy.

My contributions to GROM-RD include the development of the excessive coverage masking module, which improved deletion and duplication sensitivity in low coverage datasets by 6% and 7% respectively, and in high coverage datasets by 4% and 15% respectively. In addition I extensively tested the algorithm in its entirety, and verified its performance on the datasets used in publication. I wrote the section of the paper on excessive coverage masking and assisted with the writing of the results section.

GROM-RD: resolving genomic biases to improve read depth detection of copy number variants

Introduction

Copy number variants (CNVs) have been linked to several diseases including cancer (Berger et al., 2011; Campbell et al., 2010; Stephens et al., 2009), schizophrenia (Stefansson et al., 2009), and autism (Marshall et al., 2008). Compared to single nucleotide polymorphisms (SNPs), structural variants (or SVs, which include CNVs, insertions, inversions, and translocations) account for more differences between human genomes (Baker, 2012) in terms of the number of nucleotides and potentially have a greater impact on phenotypic variation (Korbel et al., 2007). Modern sequencing technologies, often identified as next-generation sequencing (NGS), have enabled higher resolution of CNVs compared to older methods such as array comparative genome hybridization (aCGH) and fosmid paired-end sequencing (Korbel et al., 2007). NGS produces sequenced reads, either single- or paired-end, that are mapped to a reference genome. Several strategies have been developed to detect SVs. Paired-read (PR) methods search for clusters of discordant (aberrant insert size or orientation) read pairs. Split-read methods map previously unmapped reads by splitting the reads. Read depth (RD) methods identify CNVs by detecting regions of low or high read coverage. De novo methods assemble reads into contigs, particularly useful for detecting insertions. Each detection strategy has advantages and disadvantages, and they complement each other by detecting SVs not found or not detectable using the other strategies. For example, RD does not depend on paired reads for finding SVs and is able to detect CNVs with mutated or rough breakpoints that may not be detectable with paired or split reads, but RD is unable to detect insertions, translocations, and inversions.

Several whole genome sequencing (WGS) RD methods, CNV-seq (Xie & Tammi, 2009), SegSeq (Chiang et al., 2009), rSW-seq (Kim et al., 2010), CNAseg (Ivakhno et al., 2010), and CNAnorm (Gusnanto et al., 2012), require a control sample. Other WGS RD methods, such as JointSLM (Magi et al., 2011) and cn.MOPS (Klambauer et al., 2012), require multiple samples. Often multiple samples or a suitable control are not available. Whole exome sequencing (WES) RD methods, including ExomeCNV (Sathirapongsasuti et al., 2011), CONTRA (Li et al., 2012), EXCAVATOR (Magi et al., 2013), CoNIFER (Krumm et al., 2012), and XHMM (Fromer et al., 2012) are limited to detection in coding regions of the genome (Sims et al., 2014). WGS RD methods that do not require a control include FREEC (Boeva et al., 2011), ReadDepth (Miller et al., 2011), CNVnator (Abyzov et al., 2011), and RDXplorer (Yoon et al., 2009).

Detecting CNVs is complicated by GC bias of NGS technologies, whereby read coverage varies depending on the GC content of the genome region. Existing RD methods reduce GC bias by GC bin mean normalization (CNVnator and RDXplorer), polynomial fitting (FREEC), and LOESS regression (ReadDepth). However, these methods do not consider differences in read depth variance with GC content, which may exist after GC bias correction. Complex and repetitive regions are challenging for all CNV detection methods including RD. Complex regions near telomeres and centromeres are known to be SV hotspots (Mills et al., 2011) and sequencing bias has been observed in repeat regions (Ross et al., 2013). However, RD methods have not been tailored for the difficulties of complex and repetitive regions. Additionally, RD methods suffer from low breakpoint resolution.

We have developed GROM-RD, a control-free WGS RD algorithm with several improvements and novel features compared to existing RD algorithms, such as excessive coverage masking, GC bias mean and variance normalization, GC weighting, dinucleotide repeat bias detection and adjustment, and a size-varying sliding window CNV search. These features address weaknesses in existing RD methods and biases in genomic sequencing that limit CNV sensitivity, specificity, and breakpoint accuracy, as evidenced by comparison of our algorithm to two most commonly used control-free WGS RD tools, RDXplorer (Yoon et al., 2009) and CNVnator (Abyzov et al., 2011). GROM-RD showed improved predictive capabilities and breakpoint resolution for CNVs, as well as excellent scalability for different NGS datasets, both simulated and real.

Methods

GROM-RD outputs a union set from two pipelines that differ based on the inclusion or exclusion of a pre-filtering step, excessive coverage masking (Fig. 3). Each step from Fig. 3 will be described in the following subsections.

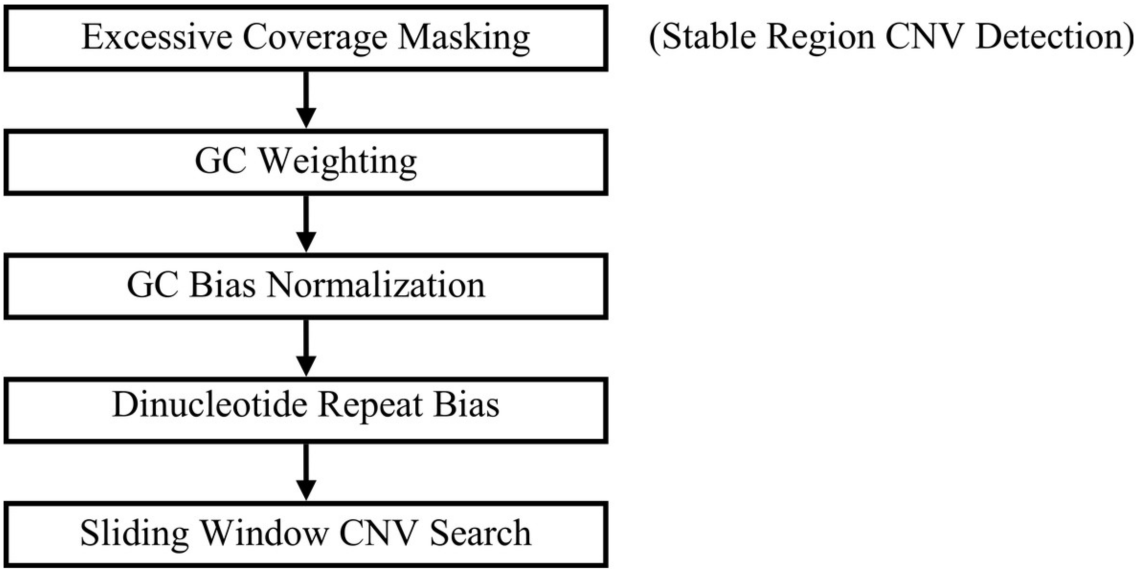

Excessive Coverage Masking
(Stable Region CNV Detection)
GC Weighting
GC Bias Normalization
Dinucleotide Repeat Bias
Sliding Window CNV Search

Figure 3: GROM-RD pipeline summary.
Two iterations of the pipeline are combined into a union set of CNV predictions. For the first iteration (step 1 included), CNV detection in stable regions is improved by masking regions of excessive coverage. Without masking (step 1 excluded), CNVs are detected in complex and repetitive regions that are characterized by excessive coverage.

Excessive coverage masking

Abnormal read coverage has been reported in centromere and telomere regions (Rausch et al., 2012). Similarly, we observed excessive read coverage in certain regions, particularly near centromeres (data not shown). This might be due to complex and repetitive segments, which are common in the human genome and can complicate CNV detection. Such high read coverage may result in false positives and also reduce CNV sensitivity in less complex regions. GROM-RD uses a two-pipeline approach to detect CNVs in complex and repetitive segments and improve sensitivity in less complicated regions. In the first pipeline, we mask clusters of blocks (10,000 base segments) with high read coverage (default: >2× chromosome average) and run GROM-RD on the masked genome. A cluster is defined as a section of the genome where >25% of the blocks have high read coverage and a minimum of four blocks have high read coverage. High coverage regions have been shown to have a high concentration of SVs (Mills et al., 2011). Thus, in the second pipeline, we run GROM-RD on the unmasked genome. GROM-RD outputs a union set of predicted CNVs from the two pipelines. Many false positives may be produced from spikes in read coverage, particularly for the unmasked genome. Thus during later steps in the pipeline, read coverage greater than twice the chromosome average is adjusted (described in ‘GC bias normalization’).

GC weighting

Variation in the GC content of genome regions affects read coverage produced by NGS platforms. A post-sequencing approach used by many RD algorithms, such as CNVnator and RDXplorer, is to bin genome regions by GC content and adjust the average read depth of each bin to the average read depth of the genome, referred to as GC bias normalization. Here we discuss the first step of this approach, calculating GC content of genome regions. RD algorithms often divide a chromosome into regions, referred to as windows, of a fixed size and estimate read depth in each window by counting reads within the window. GC content for a window is calculated from the proportion of reference sequence G and C bases within the window. Previous studies (Aird et al., 2011; Benjamini & Speed, 2012; Bentley et al., 2008) have identified PCR bias as the main contributor to GC bias in NGS. Thus, reference bases outside a window may affect read coverage within a window, especially for long reads and paired-end reads. Benjamini & Speed (2012) showed a higher correlation between GC content and read depth when considering the GC content of the entire PCR-replicated DNA fragment rather than the sequenced segment. Based on these observations, we developed a novel GC weighting method to consider all bases within an average insert size. To maximize sensitivity, we do not calculate GC weighting for a window of bases; instead, GC weighting is calculated for each base i as $h_i = \sum w_j a_j / \sum w_j$, where j is a base that may affect read depth for base i, w_j is the weight of base j and is equivalent to the sum of average inserts with unique starting locations and that overlap base j and base i, and a_j is 1 if base j is a G or C and 0 otherwise. For single-end reads, the insert size is equivalent to read length.

GC bias normalization

As referred to previously, "GC bias" in this context denotes variation in read coverage produced by NGS platforms as a result of variation in the GC content of genome regions. Many RD algorithms, such as CNVnator and RDXplorer, bin genome regions (windows) by GC content and adjust the average read depth of each bin to the average read depth of the genome:

(1) $r_{i,norm} = ri\ m/m_{GC}$

where $r_{i,\mathrm{norm}}$ is the read coverage of a window after normalization, ri is the read coverage of window i prior to normalization, m is the global mean read coverage of all windows in the genome, and m_{GC} is the mean read coverage of all windows with similar GC content (Yoon et al., 2009). Although this method normalizes the read depth means across the GC bins, we found differences in variance after GC bias correction (Fig. 4). From this observation, we expect methods using this approach to over-predict CNVs when a GC region has high variance and under-predict CNVs when a GC region has low variance.

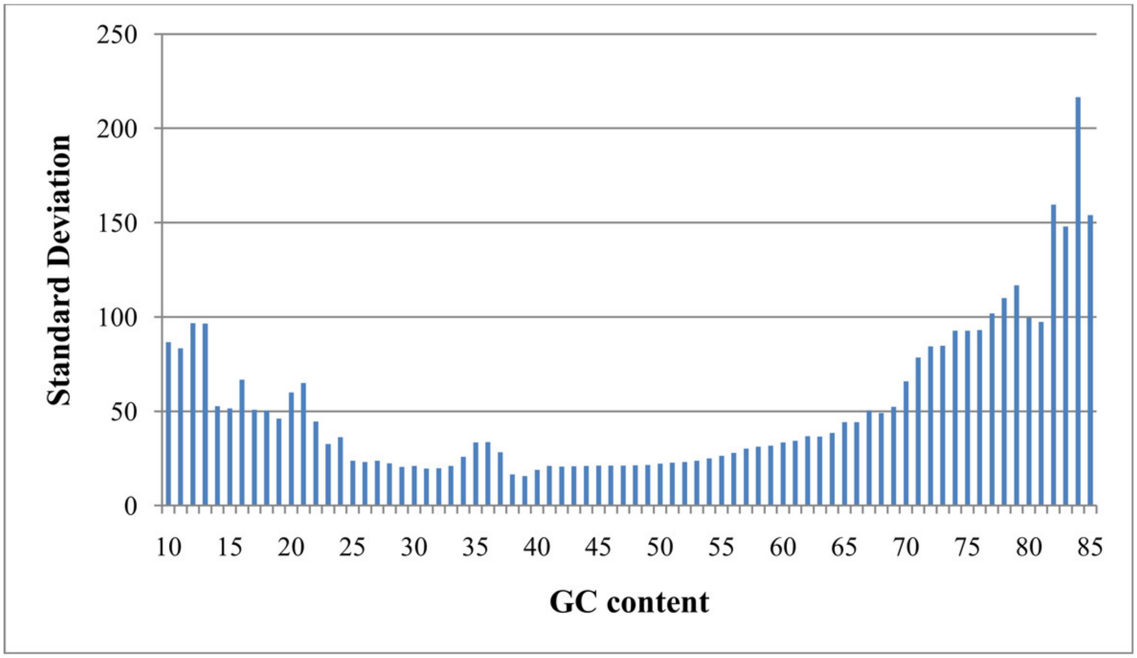


Figure 4: Standard deviation after GC bias normalization.
Data produced from chromosome 19 of NA12878 (Illumina high coverage paired-end read dataset aligned with BWA to human reference hg18) (DePristo et al., 2011) using 100-base non-overlapping windows. Reads were assigned to a window if the read center was within the window. After correcting for GC bias using a common approach, the standard deviation varies with GC content. This negatively impacts further analysis by CNV detection algorithms.

We use a quantile normalization approach to correct for variance across bins of GC weighted bases (Lin et al., 2004). For this approach, we rank bases in each bin based on read depth and calculate a rank proportion pi for each base i using:

$$\begin{aligned} p_i &= R_i/n && \text{if } 2R_i \le n \\ p_i &= (n - R_i)/n && \text{if } 2R_i > n \end{aligned}$$

where R_i is the read depth rank for base i and n is a count of bases with a particular GC weighting. When R_i is 0 (for $2R_i \le n$) or n − Ri is 0 (for $2R_i > n$), the numerator in Eq. (2)

is set to 0.5. Subsequently, p_i is converted to standard deviation units, x_i, using a pre-computed normal distribution table. Note when n is identical for all GC bins and there are no read depth ties within a GC bin, each bin distribution will have identical statistical properties, including mean and variance, after quantile normalization. Statistical properties of quantile normalized distributions may vary across GC bins when n varies, however this effect is negligible when n is large. GROM-RD requires a GC bin to have at least 100 bases. GROM-RD does not produce a normalized read depth as in Eq. (1) because it is not necessary for further analysis. Instead, read depth in standard deviation units is used. As mentioned previously in ‘Excessive coverage masking,’ to reduce false positives, read coverage greater than twice the chromosome average is adjusted by averaging the rank of the observed read coverage and the rank of read coverage equivalent to twice the chromosome average read coverage. CNVs may occur in low mapping quality regions; however, read coverage distributions tend to differ between low mapping quality and high mapping quality regions. To compensate for variation of read coverage distributions with mapping quality, GROM-RD calculates the average mapping quality for each window and creates separate distributions for low mapping quality (default: <5) and high mapping quality windows. The nature of the read depth distribution for NGS data has not been clearly defined. A rank-based approach does not assume a specific distribution and is less affected by outliers when compared to parametric methods.

Dinucleotide repeat bias normalization

Repeat bias has been observed with NGS technologies (Ross et al., 2013). We found similar repeat biases in our investigations. Additionally, these biases may vary with sequencing technology and genomes. For instance, we observed decreased coverage for AT repeats in human (Fig. 5) but not for other genomes (data not shown). We found that dinucleotide repeats as short as 20 bases affected coverage. GROM-RD detects dinucleotide repeat biases and uses a quantile normalization method in the respective genomic regions. Dinucleotide repeats with average read coverage that is more than 1.5 standard deviations above or below the genome average read coverage, are considered biased. For a biased dinucleotide repeat, we use a quantile normalization approach similar to our GC bias normalization, except R_i is the read depth rank of occurrence i of a particular dinucleotide repeat. From this we obtain read depth in standard deviation units for each biased dinucleotide repeat occurrence. As we move further from a repeat, GROM-RD creates separate sample distributions in 10 base increments to adjust for the decreasing influence of repeat bias. Thus, we bin bases by distance from the repeat, in contrast to binning by GC weighting as described in ‘GC weighting.’ Repeat bias normalization is applied within a distance of half-insert size from biased dinucleotide repeats. For genomic regions with dinucleote repeat bias, dinucleotide repeat bias normalization replaces GC bias normalization. To our knowledge, GROM-RD is the first RD method to specifically adjust for repeat bias.

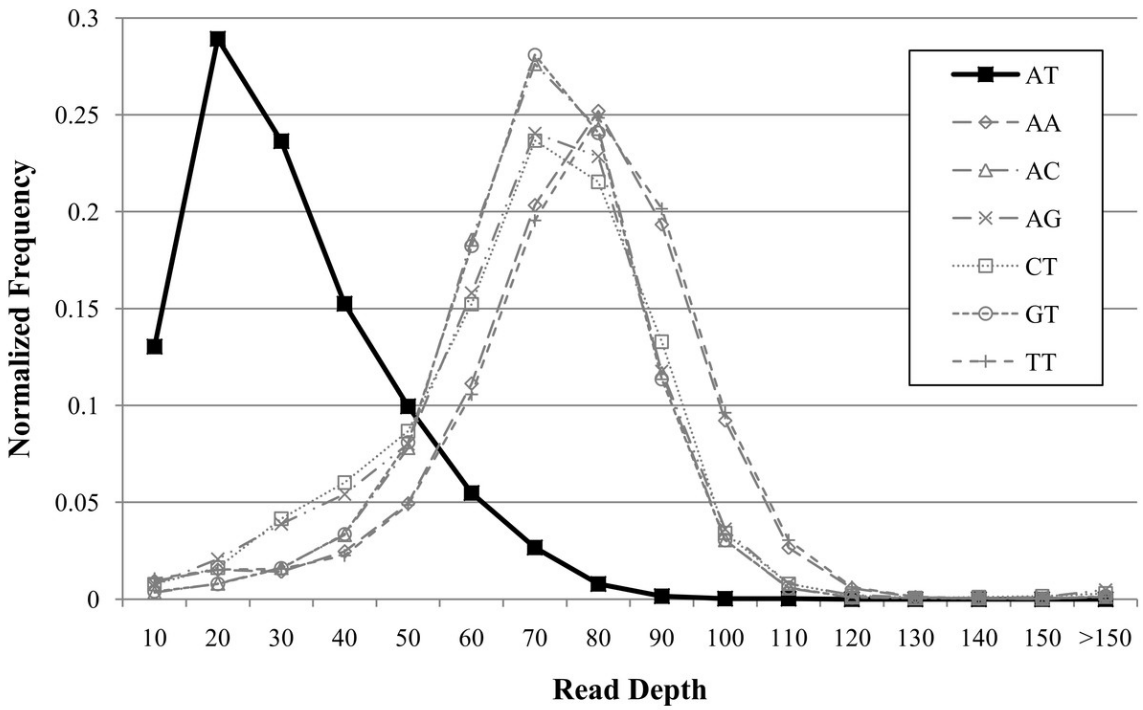


Figure 5: Example of dinucleotide repeat bias in a human genome.
AT repeats had lower coverage compared to other dinucleotide repeats for human genome NA12878 (Illumina high-coverage paired-end read dataset aligned with BWA to human reference hg18) (DePristo et al., 2011). Dinucleotide repeats less than 20 bases were filtered. Dinucleotide combinations with less than 50 occurrences in the genome are not shown.

Sliding window CNV search

RD methods typically suffer from reduced breakpoint resolution compared to other methods, such as split-read. One reason for low resolution is fixed-size, non-overlapping windows. We employ sliding windows that sequentially increase in one-base increments to improve breakpoint resolution. Fixed-size, non-overlapping windows also reduce sensitivity when CNVs start or end near the center of a non-overlapping window. Using sliding windows, GROM-RD is equally sensitive to CNVs regardless of start or end points. Additionally, by creating distributions for incremental window sizes, GROM-RD improves sensitivity on a range of CNV sizes.

As described in the previous sections, GROM-RD normalizes GC bias or, if necessary, dinucleotide repeat bias for each base. However, we do not expect to find one base deletions or duplications; instead, GROM-RD combines normalized bases into windows by averaging standard deviation units of all bases in a window. Since the means and variances of the bases have been normalized with respect to GC bias or dinucleotide repeat bias, GC and dinucleotide bias are not associated with the windows.

For each window size, we sample a set of windows from the dataset and obtain a read depth mean and standard deviation. Then, we identify base positions with abnormal read coverage $\geq 1.3_{rave,h}$ for duplications or $\leq 0.70_{rave,h}$ for deletions (for diploids) as potential breakpoints, where $r_{ave,h}$ is the average read depth for bases with h weighted GC content. If at least half of the bases have abnormal coverage for a minimum window size, $w_{l,min}$ (default = 100) beginning at a potential breakpoint j, we calculate a z-score, z, based on a sample distribution of read depths for wl,min and the read depth of a window i having size $w_{l,min}$ and beginning at j.

Several parameters affect calling CNVs as outlined below (and they can potentially be modified by a user). A CNV is called if z < α, (default: α = 1 × 10−6). We increase the window size in one-base increments and recalculate z to either extend or detect a CNV until a maximum window size $w_{l,max}$ (default = 10,000) is reached. If no CNV has been detected, we move to the next potential breakpoint and repeat our statistical testing. Attempts to extend or detect a CNV will end before reaching $w_{l,max}$ if less than half the bases have abnormal read coverage (≥ 1.3 or $\leq 0.70_{rave,h}$ for diploids). If a CNV was found and $w_{l,max}$ has been reached, we try to extend the CNV by sliding a window of size $w_{l,max}$ and recalculating z. Attempts to extend a CNV continue until thresholds related to read coverage and distance from the CNV end breakpoint have been reached. A flowchart for the sliding window CNV search is provided in Fig. 6.

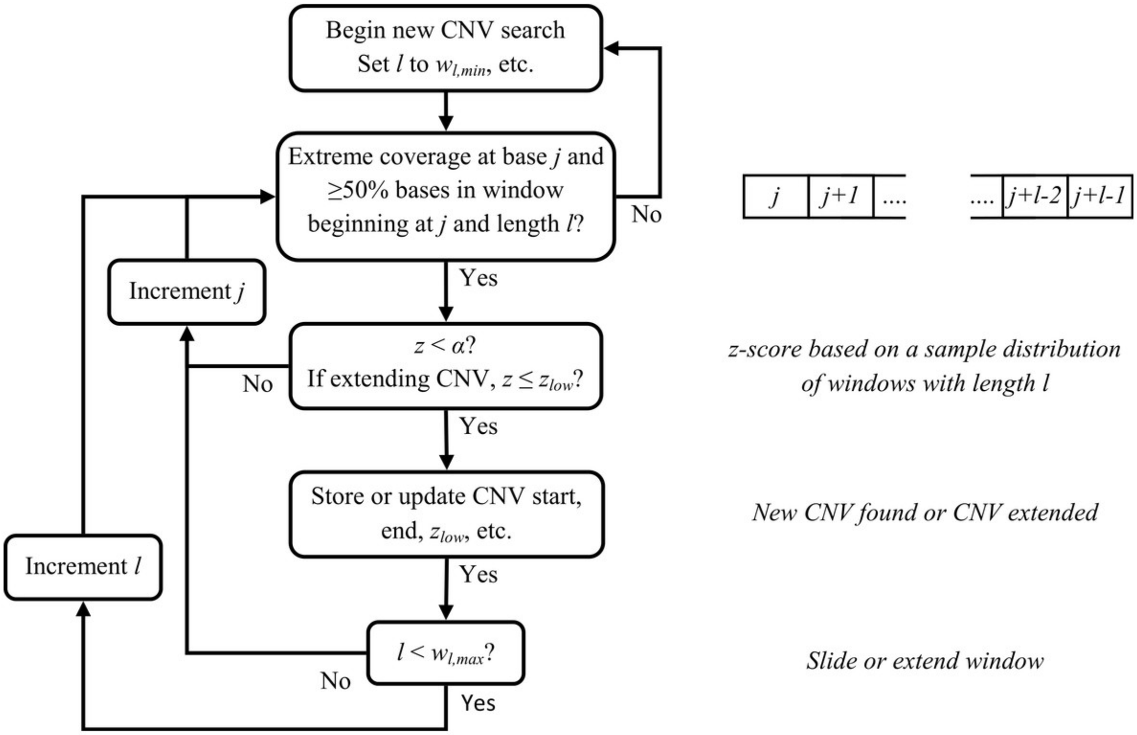
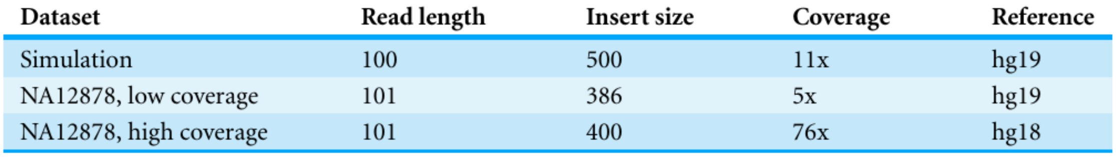

Begin new CNV search
Set l to $w_{l,min}$, etc.
Extreme coverage at base j and ≥50% bases in window beginning at j and length l?
No
Yes
Increment j
$z < \alpha$?
If extending CNV, $z \leq z_{low}$?
No
Yes
Store or update CNV start, end, z_{low}, etc.
Yes
Increment l
$l < w_{l,max}$?
No
Yes
j
$j+1$
....
....
$j+l$-2
$j+l$-1
z-score based on a sample distribution of windows with length l
New CNV found or CNV extended
Slide or extend window

Figure 6: Flowchart for sliding window CNV search.
For clarity, some conditions for ending a CNV search have been omitted.

Results

Datasets

To test GROM-RD's performance, we used both simulated (with known SVs) and experimental (with a large number of validated SVs) datasets for a human genome (Table 1). We first compared our approach with two commonly used RD algorithms, CNVnator and RDXplorer, on a simulated dataset. We used RSVSim (Bartenhagen & Dugas, 2013) to simulate 10,000 deletions and duplications ranging from 500 to 10,000 bases using the most recent human reference genome (hg19). RSVSim assumed a beta distribution to create a distribution of CNV sizes based on SVs from the Database of Genomic Variants with lengths between 500 and 10,000 bases, resulting in a decreasing frequency of CNVs with increasing size. Deletions were heterozygous (1 copy number) and duplications ranged from 3 to 10 copy numbers. RSVSim biased SVs to certain types of repeat regions and corresponding mechanisms of formation, such as non-allelic homologous recombination, based on several studies (Chen et al., 2008; Lam et al., 2010; Mills et al., 2011; Ou et al., 2011; Pang et al., 2013). We then used pIRS (Hu et al., 2012) to simulate 100-base Illumina paired-end reads with 500 base inserts and read coverage above ten. pIRS is designed to simulate Illumina base-calling error profiles and GC bias. The simulated reads were mapped to human reference genome hg19 using BWA (Li & Durbin, 2009).



Dataset	Read length	Insert size	Coverage	Reference
Simulation	100	500	11x	hg19
NA12878, low coverage	101	386	5x	hg19
NA12878, high coverage	101	400	76x	hg18

Table 1:
Summary of simulated and gold standard datasets.

We also compared CNVnator, RDXplorer, and GROM-RD on two human datasets (both from NA12878). To better assess algorithm performance with current sequencing technologies (longer reads, lower error rates, etc.), we used the more recent sequence datasets of low coverage NA12878 produced as part of the main project alignments for the 1000 Genomes Project (Abecasis et al., 2012) and high coverage

NA12878 produced at the Broad Institute and released to the 1000 Genomes Project (DePristo et al., 2011). Both datasets contain Illumina paired-end reads. We tested algorithm performance using a large set of experimentally validated and high confidence SVs produced during the pilot phase of the 1000 Genome Project and commonly referred to as the "gold standard" (Mills et al., 2011). We will use the term "gold standard" to refer to the above set of validated SVs and the sequence datasets.

Simulation results

CNVnator, RDXplorer, and GROM-RD prediction results for the simulated dataset are shown in Fig. 7. At least 10% reciprocal overlap between a predicted CNV and a simulated CNV was required for a true positive. Default parameters were used for all algorithms, except for the window (bin) size for CNVnator. We estimated the optimal window size for CNVnator (230 bases) by curve fitting the window size and read coverage combinations (resulting in bin size = $2205x^{-0.941}$, where x is the read depth) recommended by the program's authors (Abyzov et al., 2011). The default window size for RDXplorer and GROM-RD is 100 bases. For GROM-RD, we found a 100 base-window to be suitable for all datasets tested.

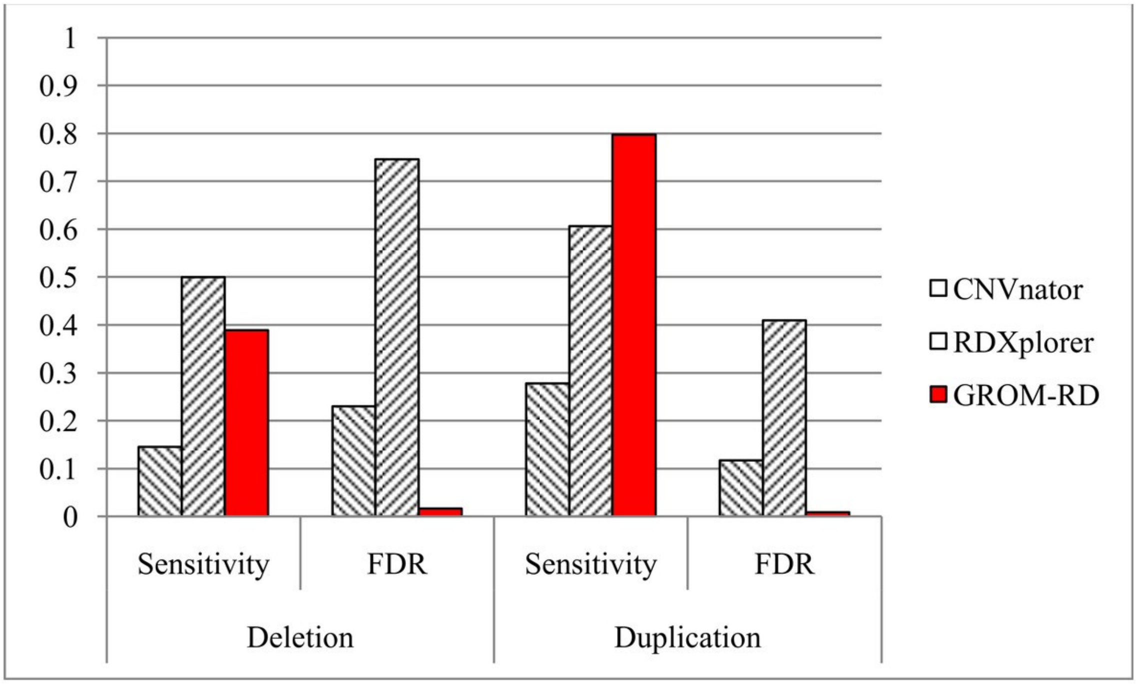
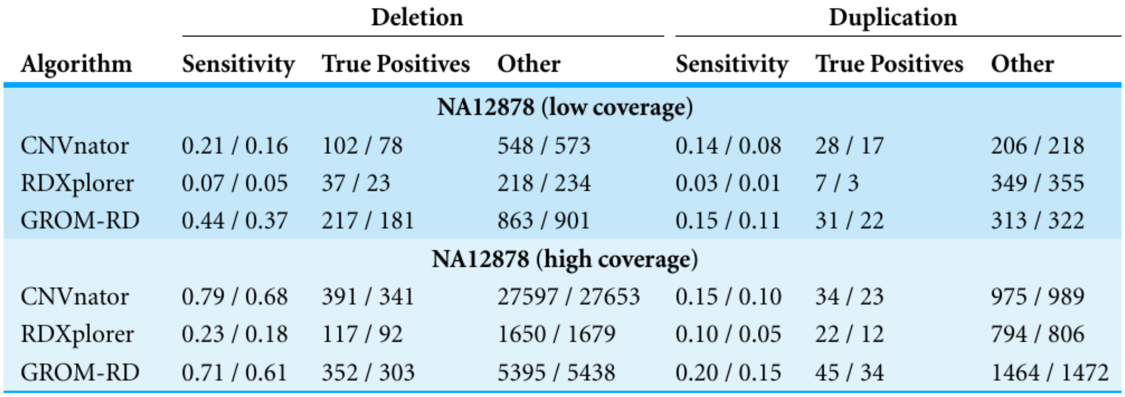


Figure 7: Sensitivity and FDR for simulated dataset.
GROM-RD had the highest sensitivity and lowest FDR for duplications. GROM-RD's sensitivity was lower than RDXplorer's sensitivity for deletions, but GROM-RD had a much lower FDR. Ten thousand deletions and duplications were simulated from human reference hg19 using RSVSim. CNVs were biased to repeat regions. One hundred-base paired-end Illumina reads with 500 base inserts were simulated at 11x coverage using pIRS and mapped to hg19 using BWA.

For the simulated dataset, GROM-RD had the highest sensitivity and lowest false discovery rate (FDR, or the proportion of predictions that were false positives) for duplications. For deletions, our method also had the lowest FDR and second-best sensitivity after RDXplorer, which showed a very high FDR (0.75) when compared to GROM-RD (0.02). When the FDR is very high, it may be more informative to consider the false positive counts. RDXplorer had 13,457 false positives compared to only 61 false positives for GROM-RD. All methods had lower sensitivity and a higher FDR for deletions than duplications, which may be due to the fact that 3 to 10 copy number changes for duplications should be easier to detect than halved RD deletions.

Gold standard results

Prediction results for the gold standard datasets are shown in Table 2. True positives indicate at least 10 or 50% reciprocal overlap between a predicted CNV and the gold standard. CNV predictions not overlapping the gold standard were labeled “Other.” Default parameters were used for all algorithms, except for the window size for CNVnator. Using the previously described curve fitting for CNVnator, we estimated 450 and 100 base windows for the low and high coverage (NA12878) datasets, respectively.



	Deletion			Duplication		
Algorithm	Sensitivity	True Positives	Other	Sensitivity	True Positives	Other
NA12878 (low coverage)						
CNVnator	0.21 / 0.16	102 / 78	548 / 573	0.14 / 0.08	28 / 17	206 / 218
RDXplorer	0.07 / 0.05	37 / 23	218 / 234	0.03 / 0.01	7 / 3	349 / 355
GROM-RD	0.44 / 0.37	217 / 181	863 / 901	0.15 / 0.11	31 / 22	313 / 322
NA12878 (high coverage)						
CNVnator	0.79 / 0.68	391 / 341	27597 / 27653	0.15 / 0.10	34 / 23	975 / 989
RDXplorer	0.23 / 0.18	117 / 92	1650 / 1679	0.10 / 0.05	22 / 12	794 / 806
GROM-RD	0.71 / 0.61	352 / 303	5395 / 5438	0.20 / 0.15	45 / 34	1464 / 1472

Table 2:
CNV prediction results for gold standard datasets.
Results indicate 10%/50% reciprocal overlap between predicted CNV and gold standard. CNV predictions not meeting overlap criteria were classified as “Other.”

Again, GROM-RD had the highest sensitivity for deletions and duplications in the low coverage dataset and duplications in the high coverage dataset. However, CNVnator found 39 more true deletions than GROM-RD in the high coverage dataset with 10% reciprocal overlap or 38 with 50% overlap.

In Table 3, we compared algorithm performance with CNV size (500–10,000 and >10,000 bases) for the gold standard datasets. True positives indicate 10% reciprocal overlap. GROM-RD had the highest sensitivity for all comparisons except for short (500–10k) high coverage NA12878 CNVs. The paucity of supporting evidence makes detecting deletions in low coverage datasets difficult for any method (Fig. 8). However, GROM-RD excelled at detecting deletions in the low coverage dataset, correctly calling more than twice and five times as many deletions as CNVnator and RDXplorer, respectively. Regarding the contribution of individual steps of our pipeline, we note that implementation of the dinucleotide repeat bias adjustment reduced GROM-RD's deletion predictions in low and high coverage NA12878 by 4 and 48%, respectively, while losing only one true positive prediction. Using quantile normalization for GC bias improved deletion and duplication sensitivity by 768 and 933%, respectively, for low coverage NA12878 and 15 and 73% for high coverage NA12878. Additionally, when employing the two-pipeline approach for excessive coverage masking, deletion and duplication sensitivity increased 6 and 7%, respectively, for the low coverage gold standard dataset and 4 and 15% for the high coverage gold standard dataset.

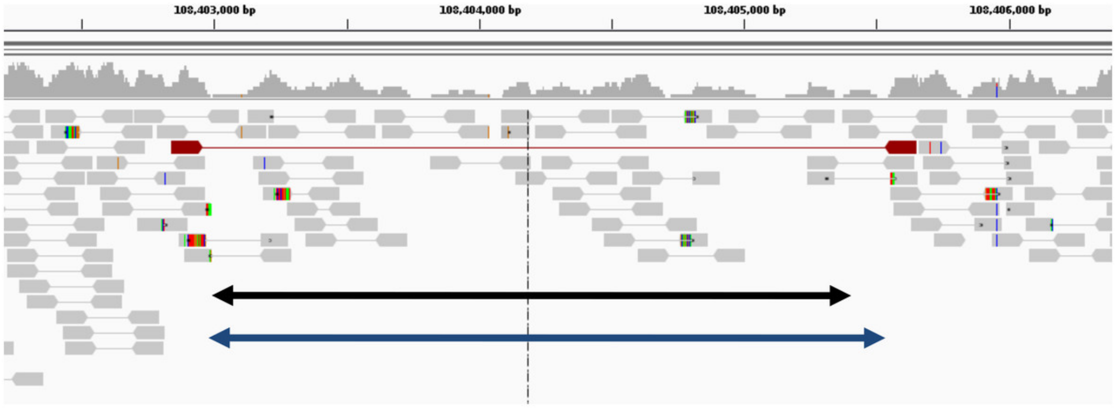
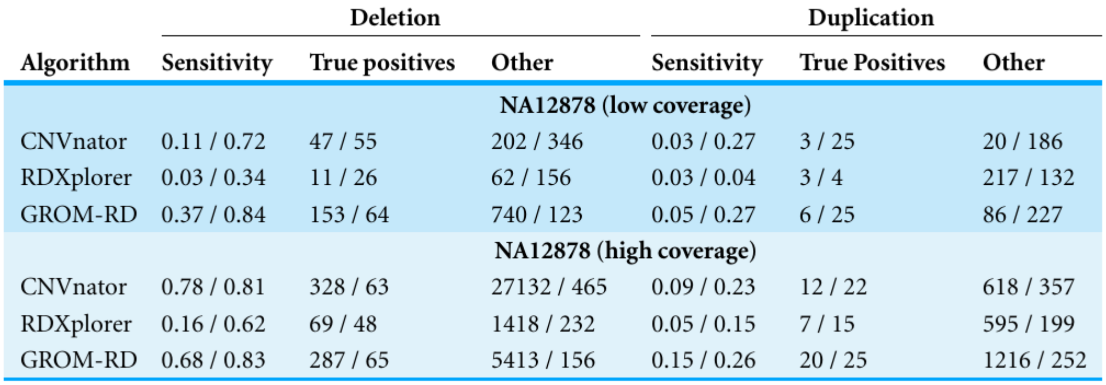


Figure 8: Example of deletion in chromosome 1 of NA12878 (detected only by GROM-RD in the low coverage dataset).
Histogram at the top reflects read coverage across the region. Grey pointed rectangles connected by lines represent paired reads. Gold standard validation (108402984–108405403) and GROM-RD's prediction (108402966–108405569) are represented by the black and blue double-arrowed lines, respectively. CNVnator and RDXplorer did not predict the deletion. We note that deletions in low coverage datasets are difficult for any method to detect as evidenced by only one discordant read pair (red pointed rectangles connected by red line) supporting the deletion making detection unlikely by a PR method. Example is shown using human reference hg19 in IGV viewer (Robinson et al., 2011).

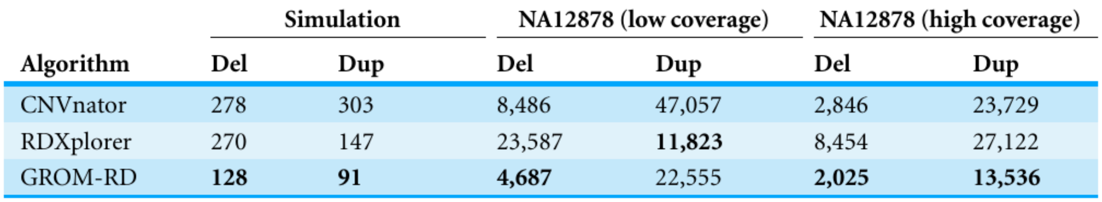


	Deletion			Duplication		
Algorithm	Sensitivity	True positives	Other	Sensitivity	True Positives	Other
			NA12878 (low coverage)			
CNVnator	0.11 / 0.72	47 / 55	202 / 346	0.03 / 0.27	3 / 25	20 / 186
RDXplorer	0.03 / 0.34	11 / 26	62 / 156	0.03 / 0.04	3 / 4	217 / 132
GROM-RD	0.37 / 0.84	153 / 64	740 / 123	0.05 / 0.27	6 / 25	86 / 227
			NA12878 (high coverage)			
CNVnator	0.78 / 0.81	328 / 63	27132 / 465	0.09 / 0.23	12 / 22	618 / 357
RDXplorer	0.16 / 0.62	69 / 48	1418 / 232	0.05 / 0.15	7 / 15	595 / 199
GROM-RD	0.68 / 0.83	287 / 65	5413 / 156	0.15 / 0.26	20 / 25	1216 / 252

Table 3:
Comparison of algorithm performance for different CNV sizes.
Results shown for short (500–10,000 bases) / long (>10,000 bases) CNVs. True positives indicate 10% reciprocal overlap. CNV predictions not meeting overlap criteria were classified as "Other."

Breakpoint accuracy

Breakpoint accuracy is one of the traditional weaknesses of the RD methods and improvements in this area can help in narrowing down CNV borders and facilitate subsequent validation experiments. CNVnator, RDXplorer, and GROM-RD breakpoint accuracy on the simulated and NA12878 gold standard datasets is summarized in Table 4. GROM-RD had the lowest deletion and duplication breakpoint error for all datasets, except duplications for low coverage NA12878 where RDXplorer had lower breakpoint error (11823 bases) compared to GROM-RD (22555). We note that RDXplorer had only seven true positive duplication calls for low coverage NA12878, limiting the reliability of the breakpoint error estimation. We observed larger breakpoint error for the NA12878 gold standard datasets relative to the simulation dataset. This was partly due to the simulation study not having large CNVs (>10k) which had larger breakpoint error compared to short (500–10k) CNVs in the gold standard datasets. Additionally, breakpoints have complexities (such as microhomology of sequence around breakpoints, repeat sequences, etc.) that are not well understood and simulated.

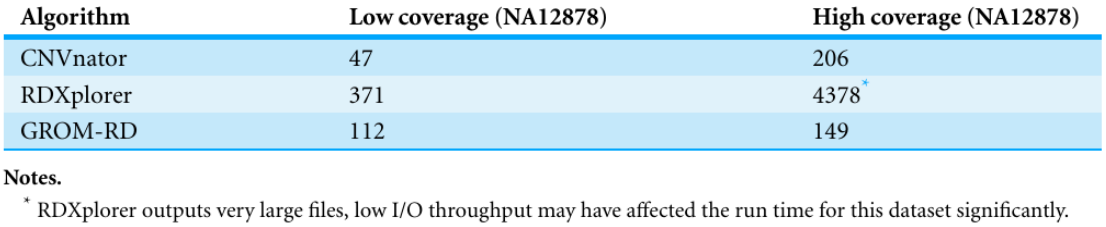


	Simulation		NA12878 (low coverage)		NA12878 (high coverage)	
Algorithm	**Del**	**Dup**	**Del**	**Dup**	**Del**	**Dup**
CNVnator	278	303	8,486	47,057	2,846	23,729
RDXplorer	270	147	23,587	**11,823**	8,454	27,122
GROM-RD	**128**	**91**	**4,687**	22,555	**2,025**	**13,536**

Table 4:
Mean breakpoint error for simulated and gold standard datasets.
Lowest error for each measurement is bolded. GROM-RD had the lowest deletion (Del) and duplication (Dup) breakpoint error for all datasets.

Algorithm metrics

Run times for the algorithms on the gold standard datasets are provided in Table 5. We tested all three programs on a single CPU (Intel Xeon E31270, 3.4 GHz) on a Linux workstation with 16 GB RAM memory. Standard BAM files were used as input. In contrast to other tools, GROM-RD's run time is relatively insensitive to read coverage with a 15-fold increase in coverage resulting in only a 33% increase in run time. GROM-RD is written in C, uses standard BAM files as input, is able to utilize paired or single reads, and is available at http://grigoriev.rutgers.edu/software/.



Algorithm	Low coverage (NA12878)	High coverage (NA12878)
CNVnator	47	206
RDXplorer	371	4378*
GROM-RD	112	149

Notes.

* RDXplorer outputs very large files, low I/O throughput may have affected the run time for this dataset significantly.

Table 5:
Run times (in minutes) on gold standard datasets.

Discussion

We developed a novel RD approach for detecting CNVs in NGS data. Many RD algorithms, such as CNVnator and RDXplorer, correct GC bias by binning genome regions based on GC content and normalizing the read depth mean of each bin to the global average. However, read depth variance tends to vary with GC content after normalizing the means (Fig. 4). GROM-RD normalizes variance by using a quantile normalization approach to convert read depth to standard deviation units. As a result, our method produces fewer false positives overall. GROM-RD, CNVnator, and RDXplorer were tested on a simulated and two gold standard datasets. GROM-RD performed well on the simulated data having the highest sensitivity and lowest FDR. Although RDXplorer had a somewhat higher sensitivity for deletions compared to GROM-RD, it came at the expense of extreme overprediction: RDXplorer had a very high FDR resulting in 13,457 false positives compared to only 61 false positives for GROM-RD. GROM-RD had the highest sensitivity for deletions and duplications on the low coverage gold standard dataset and for duplications on the high coverage gold standard dataset. For deletions in the high coverage dataset, GROM-RD had comparable sensitivity (0.71) to CNVnator (0.79). GROM-RD excelled at detecting deletions in the low coverage NA12878 dataset, correctly calling more than twice and five times as many deletions as CNVnator and RDXplorer, respectively. When comparing performance by CNV size, GROM-RD had the highest sensitivity for all comparisons except for short (500–10k) high coverage NA12878 CNVs, where GROM-RD had comparable sensitivity (0.68) to CNVnator (0.78). GROM-RD’s dinucleotide repeat bias normalization reduced GROM-RD’s deletion predictions by 4 and 48% on the low and high coverage datasets, respectively, while losing only one true positive, suggesting an improvement in specificity. As expected, duplication predictions were not affected by dinucleotide repeat bias normalization. Using quantile normalization for GC bias normalization improved deletion and duplication sensitivity by 768 and 933%, respectively, for low coverage and 15 and 73%, respectively, for high coverage NA12878. Compared to one pipeline with no excessive coverage masking, our two pipeline approach with excessive coverage masking increased deletion and duplication sensitivity 6 and 7%, respectively, for the low coverage gold standard dataset and 4 and 15% for the high coverage gold standard dataset.

Often RD algorithms analyze read depth in non-overlapping windows with a fixed size. A read is placed in a window if the read's center (CNVnator) or start (RDXplorer) occurs in the window. Fixed-size, non-overlapping windows result in low breakpoint resolution. GROM-RD utilizes sliding windows with sizes varying in one-base increments to improve breakpoint accuracy. For all datasets except duplications for low coverage NA12878, GROM-RD had the lowest deletion and duplication breakpoint error, thus improving this common weakness of RD methods.\

RD algorithms are complementary to and have some advantages compared to other CNV detection methods. For instance, RD algorithms may be able to detect CNVs with rough breakpoints and duplications with few uniquely mapped reads that paired- and split-read methods may have difficulty detecting. We observed a number of such cases for validated CNVs in the low coverage NA12878 dataset, with just one discordant read pair spanning a deletion (Fig. 8) or even with no support from discordant paired reads at all. However, RD methods frequently have low breakpoint resolution. Our results suggested that GROM-RD was able to improve RD sensitivity, specificity, and breakpoint accuracy compared to CNVnator and RDXplorer, the two most frequently used RD algorithms. Additionally, GROM-RD had a short run time that was relatively insensitive to read coverage indicating excellent scalability of the method for different datasets.

Chapter 3: Lightning-fast genome variant detection with GROM

Smith, Sean D., Joseph K. Kawash, and Andrey Grigoriev. "Lightning-fast genome variant detection with GROM." GigaScience 6, no. 10 (2017): 1-7.

GROM was developed as a comprehensive and extremely efficient variant detection algorithm for next generation sequencing data. This need had become increasingly apparent to us with the abundance of large scale NGS studies coming to fruition. GROM utilizes multiple forms of evidence for variant detection including mismatch, split read, read pair and read depth genome sequence information in a single algorithm through a single pass of the BAM file. GROM identifies all variants, including SNVs, CNVs, SV, and indels. GROM outperformed other methods on several benchmarking data sets in sensitivity, specificity, as well as speed; reducing computational time for variant detection from 41% of a typical pipeline to < 1%.

My contribution to GROM includes the extensive testing and setting of optimal values for all parameters. I also verified all call types made by GROM in several testing datasets, writing code to correct issues and improve performance (such as identifying regions with multiple types of overlapping variants, and adjusting call parameters in these regions) that were later incorporated into the algorithm. I also wrote the code converting GROM output to various formats such as vcf, as well as code that is used for comparative genomics, to identify shared and unique mutations in populations. In addition I wrote the section of results on variant discovery in NA12878, including Fig. 13, and assisted in writing the results section of the paper.

Lightning-fast genome variant detection with GROM

Introduction

The 1000 Genomes Project (Genomes Project Consortium 2010) was launched in 2008 with the goal of producing and analyzing whole genome sequencing (WGS) for 1000 genomes. By 2016 decreasing costs and increasing sequencing throughput had led to an exponential increase in the size and scope of WGS projects from Human Longevity, Inc.'s 10 000 publicly available WGS genomes (Telenti et al. 2016) to the United Kingdom's 100 000 Genomes Project (Genomics England 2017) to even larger, though less-clearly defined, sequencing projects involving 1 000 000 participants proposed in the United States (Precision Medicine Initiative (National Institutes of Health 2017) and Million Veteran Program (U.S. Department of Veterans Affairs 2017)) and China (Cyranoski 2016). Such projects produce massive amounts of data, straining computational resources and requiring much faster methods than current capabilities (Stephens et al. 2015).

Comprehensive analysis of genomic differences requires detection of a wide range of variants, including single nucleotide variations (SNVs), indels (insertions and deletions <50 bases), and larger copy number variants (CNVs) and structural variants (SVs), which include deletions, duplications, insertions, inversions, and translocations. Methods have been developed for each type of variant; subsequently, a typical WGS analysis workflow requires running multiple algorithms. A recent pipeline, SpeedSeq (Chiang et al. 2015), focused on reducing the computational resources needed for WGS analysis, though it still employed 4 variant detection algorithms. This can be wasteful of computational resources due to repetitive input/output and analysis of the same read sequences by several algorithms.

We present our method, Genome Rearrangement Omni-Mapper (GROM), a novel comprehensive method of variant detection, combining mismatch, split-read, read pair, and read depth WGS evidence. GROM boasts lightning-speed runtimes an order of magnitude faster than state-of-the-art variant detection pipelines. While drastically reducing computational time, GROM detects SNVs, indels, SVs, and CNVs in a single algorithm and provides superior overall variant detection compared with commonly employed algorithms.

Algorithm

Differences in variant types (Fig. 9) have resulted in separate algorithms designed for a limited range of variants. GROM achieves fast, comprehensive variant analysis via a compact workflow (Fig. 10), efficiently analyzing and gathering information at each reference base in 1 pass through a BAM file. Base information includes average mapping and base qualities, overlapping discordant pairs, unmapped mate reads, and split-reads, and read depth. Discordant pairs are identified based on abnormal read orientation or abnormal insert size. GROM determines abnormal insert size based on a sample of 10 million paired reads. Since insert size distributions tend to have right skewness, GROM calculates the median insert size and uses a rank-based method to determine abnormal insert size thresholds corresponding to 3 standard deviations from the median under a normal distribution (after outliers more than 5× the median insert size have been filtered). Each read with a split mapping, indel, discordant mate, or unmapped mate contributes breakpoint evidence to each potential reference base breakpoint. For simple cases such as a 2-base deletion within a read, there is 1 potential reference base start breakpoint and 1 potential reference base end breakpoint. Other cases may have less precise breakpoints, such as a read from a discordant deletion pair (abnormally large insert size). In this case, the exact breakpoint is unknown and a potential breakpoint is recorded for each reference base consistent with forming a concordant pair in the sample, where a concordant pair corresponds to insert sizes $\geq i_{min}$ and $\leq i_{max}$, where i_{min} and i_{max} represent the minimum and maximum insert size thresholds, respectively (Fig. 11). Using the deletion example in Fig. 11, a breakpoint distant from both reads would necessitate an insert size that is too large to be consistent with a concordant pair (and the source DNA fragment), and thus would not be a potential breakpoint. When soft-clipping (≥5 bases) or a split-read (each mapped split ≥20 bases) occurs in the potential breakpoint region, the reference base immediately adjacent to the soft-clipping or split-read is recorded as a potential breakpoint and other potential breakpoints are recorded with half-weighting. This enables base resolution of breakpoints while limiting a single aberrant read mapping from misidentifying the true breakpoint.

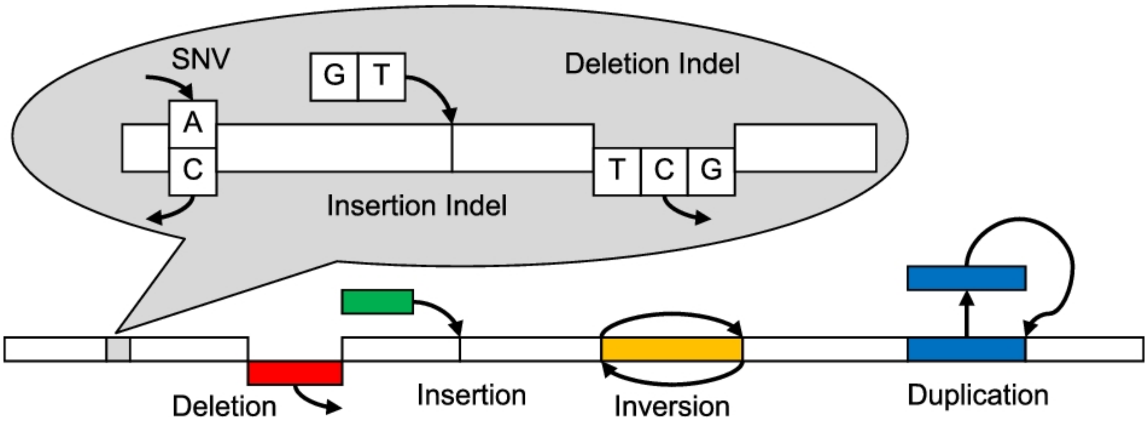


Figure 9.

Examples of variants detected by GROM. GROM detects a comprehensive range of variants (SNVs, indels, deletions, insertions, inversions, and duplications). GROM also detects translocations spanning more than 1 chromosome (not shown).

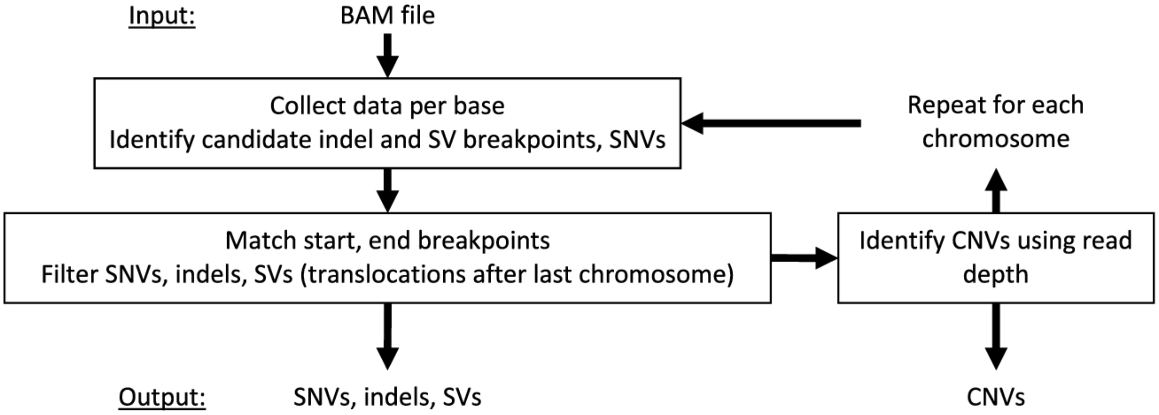

Input:
BAM file
Collect data per base
Identify candidate indel and SV breakpoints, SNVs
Repeat for each chromosome
Match start, end breakpoints
Filter SNVs, indels, SVs (translocations after last chromosome)
Identify CNVs using read depth
Output:
SNVs, indels, SVs
CNVs

Figure 10.

GROM workflow. GROM simultaneously collects data for each reference base and identifies candidate breakpoints and SNVs in 1 pass through a BAM file. After each chromosome, SNVs are filtered; start and end breakpoints are matched and filtered for each indel and SV type (excluding translocations), and CNVs are identified (using read depth).

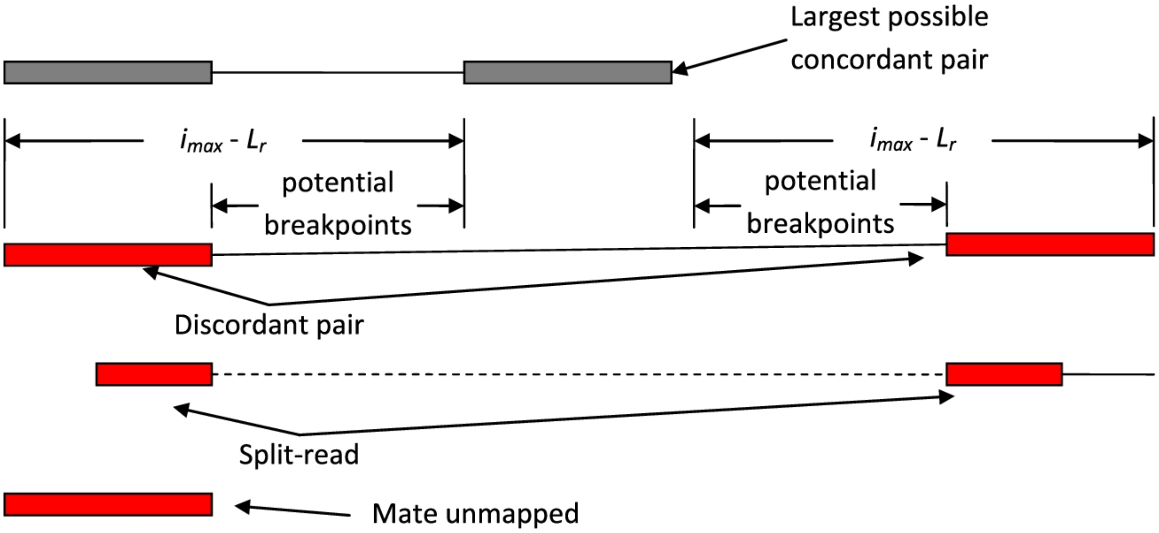


Figure 11.
Example of SV evidence and potential breakpoints. GROM considers multiple input features at each reference base position to statistically determine the likelihood of an SNV, indel, SV, or CNV. Inputs in this example (discordant pairs, split-reads, and unmapped mate reads) are primarily used for SV detection. Discordant deletion pairs identified by insert size exceeding imax. For discordant pairs, potential start and end breakpoints are recorded for each reference base capable of forming a concordant pair in the sample. Lr indicates read length.

Base by base of the reference, breakpoint evidence is stored for each distinct indel or SV. In some cases, it is difficult to distinguish variants. For instance, 2 heterozygous deletions may overlap and have similar start and end breakpoints and similar lengths. Thus, for each potential breakpoint, we cluster read evidence by variant type and length. Such clustering can be a computationally intensive task. We use the following efficient method.

We define a cluster or breakpoint cluster as a specific reference base location with a set of reads supporting a breakpoint at that location for a specific indel or SV type (deletion, duplication, etc.) of a certain length. A read from a discordant pair provides imprecise breakpoints and thus may be a member of multiple clusters, 1 cluster per reference location. A read is placed into an existing breakpoint cluster if the read and cluster support the same indel or SV type and the variant lengths are close, i.e.,

$|L_{bc} - L_{disc}| \leq (i_{max} - i_{min} + i_{median} - 2L_r)(1 + 1/x_{bc}),$

(1)

where L_{bc} is the mean indel or SV length for the breakpoint cluster, L_{disc} is the length of the indel or SV pertaining to the candidate read, Lr is the read length, x_{bc} is the number of previously recorded reads supporting the breakpoint cluster, and imax, imin, and imedian are the maximum, minimum, and median concordant pair lengths, respectively. If a candidate read does not fit in any existing breakpoint clusters, a new cluster is created. If a candidate read fits in more than 1 breakpoint cluster at the same reference position, the breakpoint cluster with the most reads is chosen. This method is efficient and has the benefit of a read being considered in multiple clusters.

Additionally, the number of previously recorded reads influences whether a read is added to a breakpoint cluster because we expect our estimated (averaged) variant length to be closer to the true SV length as supporting reads are incorporated into the SV length average. For example, in Equation (1), let insert size statistics be such that $i_{max} - i_{min} + i_{median} - 2L_r = 500$, let an SV be a deletion of 1200 bases, and let our first discordant pair indicate an SV of length $L_{disc} = 1700$. One read is a poor estimate of the true SV length. Thus, in our example, the second read's SV length may differ from the first read's SV length by 1000 bases, $|1700 - L_{disc}| \leq 1000$. However, as the number of supporting reads increase, we expect the average SV length (L_{bc}) to converge to the true

SV length of 1200, at which point we will not add the read as evidence unless its estimated SV length (L_{disc}) is within 500 bases of the true SV length, $|1200 – L_{disc}| \leq 500*(1 + \varepsilon)$, where $\varepsilon \ll 1$.

For each reference base, a mismapping probability, pbc, is calculated for each possible SNV, indel, and SV. pbc is the binomial probability of at least xbc reads supporting the breakpoint cluster given nbc read depth and a mapping quality threshold m. Thus, pbc indicates the likelihood that all of the supporting reads are mismappings. Read depth includes all mapped reads, unsequenced segments between concordant pairs, and potential breakpoints, and thus is an estimate of physical coverage. Physical coverage provides a more comprehensive representation of genome coverage than read coverage. It also helps GROM define deletion and duplication breakpoints when soft-clipping is unavailable as a decrease in coverage will affect breakpoint probability estimates. The mapping quality threshold m indicates the probability of a read mismapping, p = $10^{-m/10}$. Thus, pbc is given as

p_{bc}= Pr (X≥x)= 1 − ∑k = 0x−1(nk)$p^{k}q^{n-k}$,

(2)

where q = 1- p. To reduce computational time, binomial probability tables are precomputed and stored as data files. GROM will compute additional probability data files if the default mapping quality threshold (m = 20) is adjusted.

Potential indel and SV breakpoints are retained for further analysis. After processing reads for a chromosome (or the whole genome for translocations), GROM identifies indels and SVs with matching start and end breakpoints. Matching SV breakpoints must meet the following criteria:

$|B_S+L_S-B_e|\leq c\times(i_{max}-i_{min})$,

(3)

$|B_e-L_e-B_{s|}\leq c\times(i_{max}-i_{min})$,

(4)

where c = 3/8, B_s and B_e are the start and end breakpoints, respectively, and L_s and L_e are the average variant length of reads supporting the start or end breakpoints, respectively. Matching translocation breakpoints follow the same concept modified due to the start and end breakpoints occurring on different chromosomes,

$|M_S-B_e|\leq c\times(i_{max}-i_{min})$,

(5)

$|M_e-B_s|\leq c\times(i_{max}-i_{min})$,

(6)

where c = 3/8, B_s and B_e are the start and end breakpoints, respectively, and M_s and M_e are the average mate read reference locations of reads supporting the start or end breakpoints, respectively.

Mixed libraries/BAM files, e.g., with insert size distributions appreciably different as to affect Equations (3–6) for matching breakpoints, or libraries containing paired-end with mate-pair data, require separate runs of GROM. Also, GROM can analyze exome or RNA sequencing reads with detection limited to SNVs and indels.

GROM will also work for libraries of non-paired reads using (in addition to finding SNVs and SVs within reads) our earlier method for finding copy number variants (CNVs), GROM-RD (Smith et al. 2015). GROM-RD also performs well compared with the standard tools such as CNVnator (Abyzov et al 2011). GROM and GROM-RD have the same foundation of collecting information for each reference base, but GROM-RD detects CNVs based on read depth, where low or high coverage is evidence of a deletion or duplication, respectively. This method is complementary to the core GROM approach described above.

GROM is able to simultaneously perform duplicate filtering; its duplicate filter is conceptually similar to Picard's MarkDuplicates (Broad Institute Picard Releases 2017) and SAMtools rmdup (Li et al. 2009), which have been shown to have similar performance. Duplicate filtering may improve predictive accuracy relative to no filtering (Ebbert et al. 2016). GROM provides an option to include such filtering, if necessary. GROM filters read pairs with identical orientation and external mapping coordinates, retaining the pair with highest mapping quality. Unlike SAMtools, GROM and Picard's MarkDuplicates are able to filter duplicates with reads mapping to different chromosomes and adjust external coordinates based on soft-clipping (Ebbert et al. 2016). For the sake of speed optimization and 1-pass analysis, soft-clipping is not considered for a read's mate.

Results

We compared GROM's performance to 4 commonly used algorithms, GATK HaplotypeCaller (GATK-HC) (Despristo et al. 2011), SAMtools (Li et al. 2009), LUMPY (Layer et al. 2014), and Manta (Chen et al. 2016) using 2 extensively validated human WGS data sets, 51× NA12878 "platinum" genome (Eberle et al. 2017) and 68× HX1, a recent Chinese genome (Shi et al. 2016). GATK-HC, considered a gold standard in SNV/indel detection, has been shown to outperform state-of-the-art algorithms (Yi et al. 2014), and SAMtools is present in most pipelines. Because GROM integrates multiple lines of evidence, we also specifically compared it with a similar SV tool in the SpeedSeq pipeline (SpeedSeq, RRID:SCR_000469), LUMPY, shown to outperform other algorithms (Layer et al 2014), such as DELLY (DELLY, RRID:SCR_004603) (Rausch et al. 2012), Pindel (Pindel, RRID:SCR_000560) (Ye et al. 2009), and GASVPro (GASVPro, RRID:SCR_005259) (Sindi et al. 2012). As part of a 10 000 genome sequencing study, presently the largest human WGS variant study, a comparison of 7 SV detection algorithms (BreakDancer (Chen et al. 2009), DELLY (Rausch et al 2012), GenomeSTRiP (Handsaker et al. 2015), LUMPY (Layer et al. 2014), Manta (Chen et al. 2016), MatchClip2 (Wu et al. 2013), and Pindel (Ye et al. 2009), showed that Manta performed the best for SV detection (Telenti et al. 2016). We evaluated SNV and indel detection with the Illumina Platinum pedigree-validated benchmark sets (Eberle et al. 2017). GROM exhibited the highest SNV and insertion indel sensitivity and precision and the highest deletion indel sensitivity when compared with GATK-HC and SAMtools for the NA12878 genome (Supplementary Table S1). SVs are notoriously difficult to reliably detect (Telenti et al. 2016). Thus, we extensively analyzed GROM's performance using 4 benchmark sets for NA12878: Database of Genomic Variants Gold Standard (DGV-GS, deletions and duplications) (Macdonald et al 2014), Mills Gold Standard (Mills-GS; deletions, duplications, and insertions) (Mills et al. 2011), Genome in a Bottle (GIAB, deletions and insertions) (Zook et al. 2014); and Pendleton PacBio (deletions and inversions) (Pendleton et al. 2015). And we utilized 3 deletion and duplication benchmark sets for HX1: DGV-GS, Shi PacBio (Shi et al. 2016), and Shi IrysChip (Shi et al. 2016) (see the Methods section for a more complete description of benchmark/validation sets). A summary of the deletion and duplication comparison with LUMPY and Manta indicated superior deletion and duplication detection (Supplementary Table S2), with GROM being the highest in 10 of 14 deletion (Supplementary Table S3) and 7 of 10 duplication (Supplemental Table S4) metrics (sensitivity and precision) across the benchmark data sets. Additionally, GROM was highest in all inversion (Supplemental Table S5) and insertion (Supplemental Table S6) metrics. GROM also detected 545 and 472 translocation events in NA12878 and HX1, respectively. However, these events were not included in the benchmarking due to the lack of validated translocation data sets for either genome.

With dropping sequencing costs and growing data throughput, it is imperative to reduce the computational costs of big data analysis. GROM was 1.7× (NA12878) and 2.1× (HX1) faster than the next fastest algorithm, Manta (Supplementary Table S7). Since typical analyses involve running separate algorithms for SNV/indel and SV detection, we compared a simple 24-thread parallelized GROM version (allocating a thread per 1/24 of the genome) with the fastest and best-performing 2-algorithm workflow (GATK-HC/Manta). Strikingly, GROM ranged from 24× (HX1, no duplicate filtering) to 72× (NA12878 with duplicate filtering) faster than a combination of 22-thread GATK-HC/2-thread Manta (Supplementary Table S8), drastically reducing variant detection and duplicate filtering from 41% to <1% of a typical WGS analysis pipeline (Fig. 12). For 1000 genomes on a 24-thread server, it may literally save years of computation.

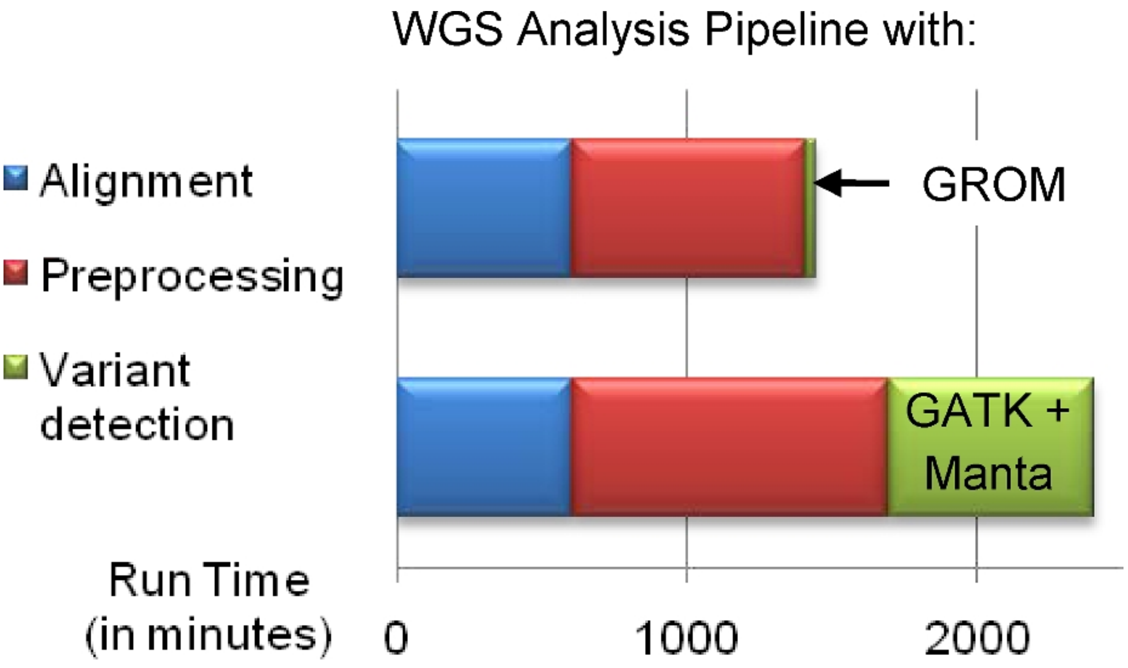


Figure 12.
Total WGS pipeline timing on NA12878. GROM reduces WGS analysis time by drastically cutting run time for variant detection (green). It enables further speedup in preprocessing (red) by simultaneously performing an optional step, duplicate filtering. For visibility in the bar chart, GROM's variant detection run time was artificially increased 3-fold.

Comparing the variants predicted by different tools, we identified 33 validated NA12878 SVs detected by GROM (but unreported by LUMPY and Manta) that overlapped genes and ranked them using the number of independent validations (Supplementary Table S9). A variant was considered validated if it occurred in at least 1 of the NA12878 benchmarks corresponding to the SV type (DGV-GS, Mills-GS, GIAB, Pendleton PacBio for deletions; DGV-SV, Mills-GS for duplications; Mills-GS, GIAB for insertions; and Pendleton PacBio for inversions).

Among these variants, we noted 4 deletions with significant health-related impact for NA12878: RHD, GSTM1, IFI16, and UGT2B17 (Fig. 13). GROM predicted a deletion spanning the entire RHD gene, 1 of 2 genes responsible for Rh blood group antigens (Wagner et al. 2009). Decreased copy numbers or null genotype of GSTM1 have been associated with hepatotoxicity (Singh et al. 2017) and higher risk of many cancers including lung cancer (Yang et al. 2015), gastric cancer (Lao et al. 2014), and bladder cancer (Norskov et al. 2011). UGT2B17 copy number variation has been associated with changes in bone mineral density and risk of osteoporosis (Yang et al. 2008). IFI16 is involved in viral defense (Orzalli et al. 2013) and p53-mediated apoptosis (Aglipay et al. 2003, Johnstone et al. 200).

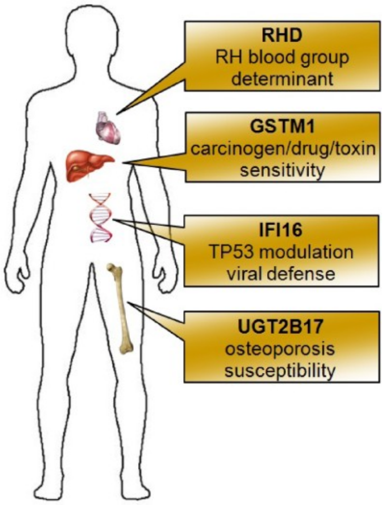
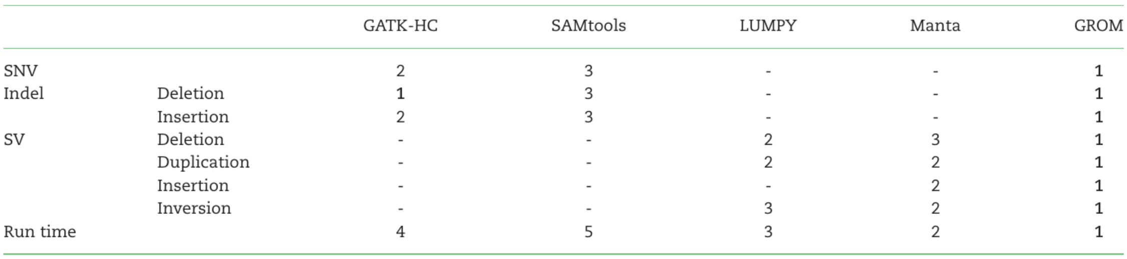

RHD
RH blood group determinant
GSTM1
carcinogen/drug/toxin sensitivity
IFI16
TP53 modulation viral defense
UGT2B17
osteoporosis susceptibility

Figure 13.
Example of genes overlapped by validated GROM-specific SVs. In the example are 4 of 33 genes overlapped by validated SVs that were identified by GROM and unreported by LUMPY and Manta. Biological significance listed below gene.

Additionally, GROM provides an option to include duplicate filtering. This leads to minor accuracy gains in a number of cases (see example in Supplementary Table S10) and achieves additional speedup (Supplementary Table S8). Lastly, we have summarized GROM's relative performance in Table 6.



		GATK-HC	SAMtools	LUMPY	Manta	GROM
SNV		2	3	-	-	**1**
Indel	Deletion	**1**	3	-	-	**1**
	Insertion	2	3	-	-	**1**
SV	Deletion	-	-	2	3	**1**
	Duplication	-	-	2	2	**1**
	Insertion	-	-	-	2	**1**
	Inversion	-	-	3	2	**1**
Run time		4	5	3	2	**1**

Table 6.
Performance based on sensitivity and precision rankings (1 = highest, 3 = lowest) averaged across benchmarks for NA12878 and HX1. Bold text indicates the best-performing algorithm in each category. A dash sign indicates that an algorithm does not detect variant type.

Methods

All timings were performed on an Intel Xeon E5–2690 v. 3 processor, 2.60 GHz, with 24 threads and 128 GB RAM.

Rankings in Table 6 and Supplementary Table S2 were based on average ranking across benchmarks (1-highest to 3-lowest). Ranking for each benchmark was based on sensitivity and precision values in Supplementary Tables S3–S6. For instance, GROM had the highest value for 10, second highest for 2, and lowest for 2 of the 14 deletion sensitivity and precision benchmarks (average benchmark rank, 1.4) Subsequently, the algorithms were ranked after sorting by their average benchmark ranking, resulting in deletion rankings of GROM, 1; LUMPY, 2; and Manta, 3 (as shown in Table 6).

Unlike most SV variant callers, GROM is able to analyze data sets with single or paired reads. However, all SV tests included only paired reads since most of the other callers operate on those.

While state-of-the-art detection methods for SNVs and indels have been deemed adequate for the clinical setting, SV detection is notably more difficult (Telenti et al. 2016). Additionally, synthetic data sets have suffered from oversimplifications and misleading conclusions (Telenti et al. 2016). Thus, we extensively analyzed GROM's SV detection performance using 4 validation benchmark sets for NA12878:

Database of Genomic Variants Gold Standard (deletions and duplications) (Macdonald et al. 2014) in Supplementary Tables S2–S4;

Mills Gold Standard (deletions, duplications, and insertions) (Mills et al. 2011) in Supplementary Tables S2–S5;

Genome in a Bottle (deletions and insertions) (Zook et al. 2014) in Supplementary Tables S2, S3, S5; and

Pendleton PacBio (deletions and inversions) (Pendleton et al. 2015) in Supplementary Tables S2, S3, S6.

Additionally, we utilized 3 deletion and duplication benchmark sets for HX1: DGV-GS (as above), Shi PacBio, and Shi IrysChip (Shi et al. 2016) in Supplementary Tables S2–S4. For NA12878 DGV-GS benchmarks, all deletions and duplications with the "NA12878" tag were extracted from the DGV-GS. The HX1 DGV-GS benchmarks were created by extracting deletions and duplications with the "Asian" tag. To obtain a benchmark set of common Asian variants, deletions and duplications with fewer than 200 "Asian"-tagged samples were filtered.

To limit potential biases, we selected benchmarks covering a range of technologies, including Illumina, PacBio, and IrysChip, and inclusive of multiple variant detection algorithms (Illumina platinum pedigree-validated, DGV-GS, Mills-GS, and GIAB). Indels were defined as deletions and insertions <50 bases, whereas SVs were ≥50 bases. To identify true positives, indel benchmarking required variant call breakpoints within 2 bases of the benchmark. Insertion SV calls within 10 bases of the benchmark were considered true. All other SV benchmarking required a 50% (10% for IrysChip due to low resolution) reciprocal overlap of a variant call and the benchmark. Some false positives may potentially be true positives not represented in the benchmark. To limit false positives due to unrepresented calls, for each SV type (excluding insertions where the length is often unknown), we ignored SV calls smaller or larger than a particular benchmark's shortest or longest SV, respectively.

NA12878 and HX1 Illumina platinum fasta files were mapped to human references hg19 and GRCh38, respectively, using BWA mem (Li et al. 2009), version 0.7.15, with the -M parameter to mark shorter read splits as secondary. Duplicate filtering comparisons were performed using default parameters for SAMtools (Li et al. 2009), version 1.3.1, and Sambamba (Tarasov et al. 2015), version 0.6.4. GATK version 3.6.0 HaplotypeCaller (GATK, RRID:SCR_001876) (Depristo et al. 2011), SAMtools (SAMTOOLS, RRID:SCR_002105) (Li etl al. 2009), LUMPY (LUMPY, RRID:SCR_003253; version 0.2.11) (Layer et al. 2014), and Manta (version 1.0.1) (Chen et al. 2016) were run with default parameters.

Conclusion

Our extensive performance analysis indicates that GROM achieves superior variant detection and is significantly faster than current state-of-the-art methods by incorporating comprehensive variant detection (SNV, indel, SV, CNV), duplicate filtering, and multi-threading in 1 algorithm. GROM’s superior variant detection makes it valuable for WGS analysis projects of all sizes, and its “lightning”-fast speed is especially critical for keeping pace with increasingly higher sequencing throughput and larger data projects.

Availability of data and materials

NA12878 raw short-read Illumina platinum WGS data, as well as pedigree-validated SNVs and indels, supporting the results in this study are available from the Database of Genotypes and Phenotypes under accession number phs001224.v1.p1 (National Center for Biotechnology Information, 2017). HX1 raw short-read Illumina WGS data supporting the results in this study are available from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), study PRJNA301527 (National Center for Biotechnology Information, 2017). DGV-GS-validated SVs supporting the results in this study are available from the Database of Genomic Variants website (Database of Genomic Variants, 2017) . Mills-GS-validated SVs supporting the results in this study are available as Supplementary Table S5 in the associated paper (Mills et al. 2011). GIAB validation data supporting the results in this study are available from NCBI at separate locations for deletions (v. 3.3.1) (Zook et al. 2016) and insertions (Wang et al. 2017). Pendleton PacBio–validated deletions and inversions supporting the results in this study are available as Supplementary Tables S5 and S6, respectively, in the associated paper (Pendleton et al. 2015). Shi PacBio– and Shi IrysChip–validated SVs supporting the results in this study are available from the corresponding author's website (Wang et al. 2017). Human reference genomes hg19 and GRCh38 are available from the Broad Institute (Broad Institute, 2017) and UCSC (UCSC, 2017), respectively. Snapshots of the GROM project code are available via the Open Science Framework (Smith et al. 2017) and the GigaScience database, GigaDB (Smith et al. 2017).

Chapter 4: Evolutionary adaptation revealed by comparative genome analysis of woolly mammoths and elephants

Smith, Sean D., Joseph K. Kawash, Spyros Karaiskos, Ian Biluck, and Andrey Grigoriev. "Evolutionary adaptation revealed by comparative genome analysis of woolly mammoths and elephants." DNA Research (2017): dsx007.

Comparative genomics studies typically limit their focus to single nucleotide variants (SNVs) and that was the case for previous comparisons of woolly mammoth genomes. However, structural variants (SVs) account for more differences between human genomes (Baker, 2012) in terms of the number of nucleotides and potentially have a greater impact on phenotypic variation (Korbel et al., 2007). Using GROM, we extended the analysis identifying SVs and indels, as well as SNVs. This led to the discovery of multiple mammoth-specific deletions and duplications affecting exons or even complete genes that were not previously reported. We found many unique genetic changes in woolly mammoths that suggest adaptations to life in harsh Arctic conditions, including variants in cold adaptation, circadian rhythms, and anatomy and physiology.

My contributions to this paper includes the code to perform comparative analysis between woolly mammoth and Asian elephant genomes as well as code to identify variants that are present in the exon regions of genes. In addition, I identified variants in genes for several sections of the paper, including cold adaptation (APOB, TRPM8, and GPR83), circadian rhythms (ARNTL, CRTC1, KDM5A, and KMT2A), and physical differences (ZDHHC23 and CCDC94). I also wrote the sections of the paper related to these topics, as well as assisted in the writing of the paper in its entirety.

Evolutionary adaptation revealed by comparative genome analysis of woolly mammoths and elephants

Introduction

The woolly mammoth (Mammuthus primigenius) was the last surviving species of the mammuthus genus with the last known population on Wrangel Island about 4,000 years ago (Palkopoulou et al. 2015, Vartanyan et al. 2008). Woolly mammoth is one of the most studied extinct species, although much is still unknown, e.g. why they became extinct, how they evolved, and how they differ from elephants, their closest living relatives. Perhaps the most common theories for the cause of their extinction are a warming climate, hunting by humans, or both. Woolly mammoths lived in a cold, dry steppe-tundra where average winter temperatures ranged from -30°C to -50°C, much different from the tropical and subtropical environments of modern African and Asian elephants (MacDonald et al. 2012). Mammoths had many anatomical adaptations minimizing heat loss in its harsh environment, such as thick fur, small ears, and small tails (compared with modern elephants), and a thick layer of fat under the skin to reduce heat loss and possibly serve as a heat source or fat reservoir for the winter(Fisher et al. 2012, Hill et al. 2007, Repin et al. 2004). Mitochondrial analysis has suggested there were three clades (I–III) with clade I surviving ~30,000 years after the extinction of clades II and III despite sharing overlapping territory with clade II in Northeastern Siberia and possibly clade III in Europe (Palkopoulou et al. 2013).

In 2008, the first whole genome sequencing (WGS) (<1×) of a woolly mammoth was published (Miller et al. 2008). However, only recently have high coverage WGS datasets become available from two studies that identified SNVs unique to woolly mammoths to infer the genetic basis of adaptations to the Arctic (Lynch et al. 2015) and to analyse species diversity prior to extinction (Palkopoulou et al. 2015). Our combined dataset included mammoths M4 and M25 from the Lynch study, and mammoths Wrangel and Oimyakon from the Palkopoulou study. We used the combined dataset, as well as four Asian elephant WGS datasets, (Lynch et al. 2015, Dastjerdi et al 2014) to analyse structural variants (SVs), copy number variants (CNVs) and indels, as well as SNVs, to further investigate genetic adaptations and diversity. Combining datasets also enabled a comparison of clades, as each study had one clade I (M4, Wrangel) and one clade II (M25, Oimyakon) mammoth. All mammoths remains originated from northern Siberia (M4, ~20,000 years ago; M25, ~60,000 years ago; Oimyakon, ~44,800 years ago) or Wrangel Island, off the coast of northern Siberia (Wrangel, ~4,300 years ago), although the exact location is unknown for M4 (Palkopoulou et al. 2017, Lynch et al. 2015, Gilbert et al. 2008).

In this paper, we analysed the patterns of variation across the genomes of these four mammoths and compared them to the available elephant genomes. Using algorithms from our GROM (Genome Rearrangement Omni-Mapper) suite; GROM-RD (Smith et al. 2015) and GROM (Smith et al. 2017), we systematically identified variants, ranging from single nucleotide changes and short indels to deletions and amplifications of regions encompassing gene fragments and complete genes. These variants reveal the signs of evolutionary adaptation of mammoths to the harsh and cold environment. We identified changes in many parts of the genome, including genes associated with metabolism, circadian rhythms, immunity and skeletal/body shape. Taken together, they describe a rich evolutionary history of the woolly mammoth species and may shed light on causes of their extinction.

Materials and methods

WGS data

WGS fasta files for woolly mammoths M4 and M25 and Asian Elephants Asha, Parvathy, and Uno were downloaded from the Sequence Read Archive (SRA), http://www.ncbi.nlm.nih.gov/sra (9 February 2017, date last accessed) (project accession number: PRJNA281811). WGS fasta files for the Wrangel and Oimyakon mammoths were downloaded from the European Nucleotide Archive (ENA), http://www.ebi.ac.uk/ena (9 February 2017, date last accessed) (accession number: ERP008929). WGS fasta files for the Asian elephant Emelia were downloaded from ENA (accession: ERP004241). WGS fasta files were mapped to the African reference genome loxAfr3, downloaded from UCSC (https://genome.ucsc.edu, http://hgdownload.soe.ucsc.edu/goldenPath/loxAfr3/bigZips/ (9 February 2017, date last accessed)), using BWA MEM, (Li et al. 2009) version 0.7.4, with default parameters. Duplicates were removed using SAMtools, (Li et al. 2009) version 0.1.19.

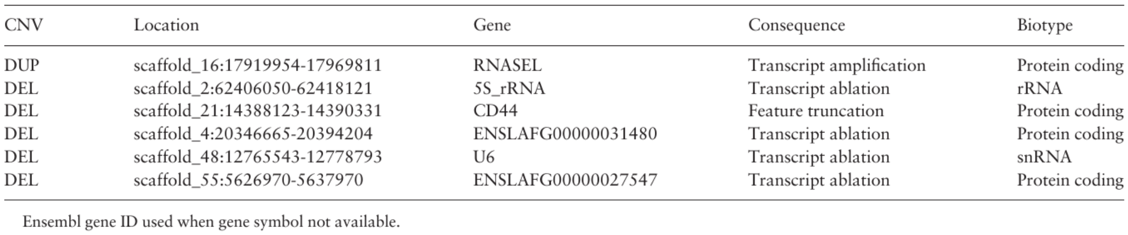
Variant detection and analysis

We limited analysis to supercontigs/scaffolds ≥1,000,000 bases. Visual inspection of the mapped files using IGV (Robinson et al. 2011) indicated an increase in low mapping quality reads and highly variable coverages for smaller (<1,000,000) supercontigs. For GROM-RD, (Smith et al. 2015) this threshold was set to 5,000,000 bases to provide adequate sampling and increase specificity when utilizing read depth methods. GROM-RD identifies regions of abnormal coverage that are unlikely to occur by chance based on analysis of read depth variance with nucleotide composition and mapping quality across the genome. This reduces false positives in potentially biased, highly variable ancient DNA datasets. GROM-RD measures read-depth variance for various window sizes (100–10,000 bases). A benefit of this approach is limiting spikes in false positives with decreasing CNV length. CNVs were detected using GROM-RD with default parameters. Homozygous deletions (GROM-RD copy number estimate <0.5) and duplications (copy number estimate >3.5) found in all woolly mammoths (80% reciprocal CNV overlap) were analysed using GROM-RD and filtered if the GROM-RD copy number estimate was <1.5 (deletions) or >2.5 (duplications) in any of the Asian elephants. Indels and SNVs were detected using our variant calling framework GROM (Smith et al. 2017). Additionally, GROM-RD calls were filtered if CNVnator (run with default parameters) detected the CNV in any Asian elephant (20% reciprocal overlap) or did not detect the CNV in all mammoths (50% reciprocal overlap). A simplified SNV/indel caller module in GROM was implemented to replicate, with a few differences, the SNV detection method used in a previous mammoth study (Lynch et al. 2015). Indels and SNVs predicted as being heterozygous (<80%, indels, or <90%, SNVs, of overlapping reads contained the variant) or supported by <4 reads with mapping quality ≥ 20 in any of the mammoths, or with at least one variant-supporting read (no mapping quality threshold) in any of the Asian elephants were filtered. We also required each Asian elephant to have at least 4× coverage at the indel or SNV site predicted in mammoth. These requirements were reciprocated when calling elephant-specific variants. Fixed, derived CNVs, indels, and SNVs were then uploaded to VEP (Variant Effect Predictor, www.ensembl.org) to identify variants potentially affecting genes. GO term and KO phenotypes obtained from the Gene Ontology Consortium (geneontology.org), and MGI (www.informatics.jax.org). GO term, KEGG pathway, and KO phenotype enrichment for woolly mammoth fixed, derived amino acid substitutions was analysed using PANTHER (http://pantherdb.org/ (9 February 2017, date last accessed)) (Thomas et al. 2003), WebGestalt (http://www.webgestalt.org/ (9 February 2017, date last accessed)) (Zhang et al. 2005), and Vlad (http://proto.informatics.jax.org/prototypes/vlad/ (9 February 2017, date last accessed)), respectively. Using a set of SNVs (four variant-supporting reads, mapping quality ≥20, allele frequency ≥0.4) found in at least one mammoth or Asian elephant, FST estimates were calculated, as described in a prior publication (Weir et al. 1984), with VCFtools (Danecek et al. 2011), version 0.1.12, using the –weir-fst-pop function with a 100,000 base window size (–fst-window-size) and 10,000 base sliding window (–fst-window-step). Protein multiple sequence alignments produced using T-Coffee (Notredame et al. 2000) for protein sequences shorter than the recommend limit of 2,500 amino acids or MUSCLE (Edgar 2004) for longer protein sequences.

CNVs

As noted in Lynch et al. (2015) mammoths and Asian elephants diverged after branching from African elephants (reference genome), thus fixed CNVs in mammoths are considered derived. Using an intersection of GROM-RD (Smith et al. 2015) and

CNVnator (Abyzov et al. 2011) calls, we found 56 fixed, derived mammoth CNVs, including 55 deletions and one amplification. According to VEP (ensembl.org), three deletions with putative loss of function and one amplification with potential gain of function occurred in the exons of protein-coding genes, and two deletions affected RNA genes (Table 7).



CNV	Location	Gene	Consequence	Biotype
DUP	scaffold_16:17919954-17969811	RNASEL	Transcript amplification	Protein coding
DEL	scaffold_2:62406050-62418121	5S_rRNA	Transcript ablation	rRNA
DEL	scaffold_21:14388123-14390331	CD44	Feature truncation	Protein coding
DEL	scaffold_4:20346665-20394204	ENSLAFG00000031480	Transcript ablation	Protein coding
DEL	scaffold_48:12765543-12778793	U6	Transcript ablation	snRNA
DEL	scaffold_55:5626970-5637970	ENSLAFG00000027547	Transcript ablation	Protein coding

Ensembl gene ID used when gene symbol not available.

Table 7
Fixed, derived woolly mammoth CNVs

Perhaps the most famous CNV in the elephant genome is the TP53 gene (Abegglen et al. 2015). However, analysis with GROM-RD and visual inspection with IGV did not reveal any copy number change relative to elephants in the corresponding genome regions of woolly mammoths.

One deletion occurred in an exon of CD44. The CD44 gene is expressed in multiple tissues including the central nervous system, lung, epidermis, liver, and pancreas (Sneath et al. 1998). Its product is involved in multiple functions, including cellular adhesion, hyaluronate degradation, lymphocyte activation, lymph node homing, angiogenesis, and the release of cytokines (Sneath et al. 1998). CD44 contributes to the maintenance of stem cell features, such as apoptosis resistance (Zoller 2015). Intriguingly, the deleted exon is the first of the ten so-called variable exons of CD44, whose splicing and histone mark deposition has been shown to be modulated by Argonaute proteins and strongly affected in Ago2-/- mouse embryonic fibroblasts (Ameyar-Zazoua et al. 2012), thus its loss may affect the functionality and tissue distribution of CD44 isoforms. Among a multitude of CD44 GO terms and knockout (KO) mammalian phenotypes listed in Supplementary Table S1, there are two phenotypes of special interest to woolly mammoths, increased diameter of the tibia and short tibia. Mammoth limb bones were much greater in diameter than the limb bones of modern day elephants (Haynes 1993). Additionally, the hind leg to fore leg ratio was smaller in mammoths compared with elephants (Haynes 1993), with their body size and stature decreasing towards the end of the last glacial advance in the Pleistocene (Haynes 1993). On the basis of a small sample, the tibia to femur ratio of mammoths (0.57 ± 0.04, n = 5) was less than the tibia to femur ratio of African elephants (0.60 ± 0.02, n = 5) (Brennan-Laun et al 2014).

One may argue that deletions detected with an alignment of reads against an evolutionary distant reference may represent false positives arising due to rapid evolution and low sequence similarity in the respective regions. Even if that was true, this argument does not change the relevance of our logic on the possible effects on the exon of CD44, since any significant sequence alteration would also modify the function of the region in woolly mammoths, compared with African and Asian elephants. Further, the large sizes of the detected deletions (Table 7) suggest extreme changes, not seen in other genes. Other deletions were observed outside of the coding regions and are thus also outside the scope of the study and do not affect our conclusions.

We detected a fixed, derived woolly mammoth amplification encompassing the RNase L gene (scaffold_16:17939571-17966869, forward strand), including ∼20,000 bases upstream. GROM-RD predicted five (Wrangel, Oimyakon), six (M25), or nine (M4) copies of RNase L (6.6 copies, stdev 0.8) in the mammoths. Figure 14 shows normalized coverages and the location of RNase L within the CNV. RNase L has several critical roles including antiviral response, adipocyte differentiation, tumorigenesis, cell proliferation, innate immune response, and apoptosis (Brennan-Laun et al. 2014). Antiviral response involves endonucleolytic cleavage of single-stranded foreign RNAs, ribosomal RNAs, and mRNAs by activated RNase L (Brennan-Laun et al. 2014, Fabre et al 2012, Salehzada et al. 2009).

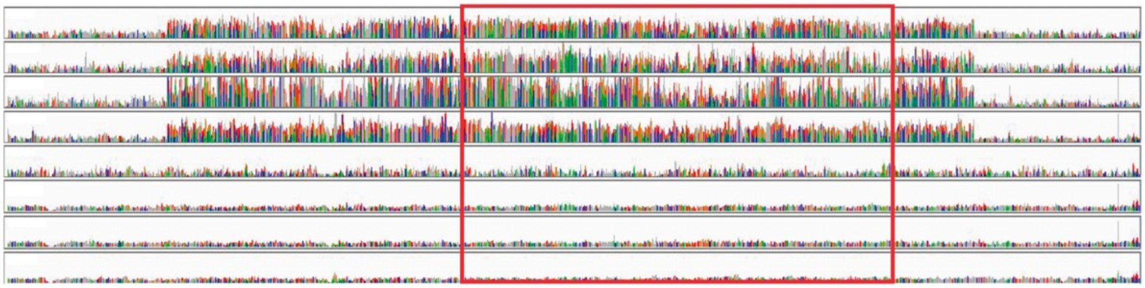
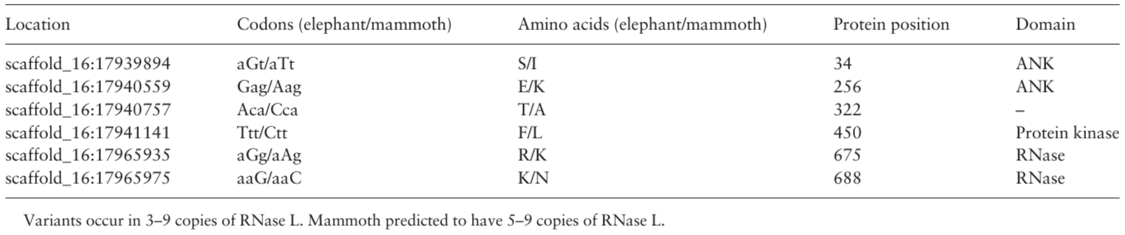


Figure 14.
RNase L amplification unique to woolly mammoths. Additional 10,000 bases shown upstream and downstream of duplication. Read coverage normalized (y axis maximum = 6× average genome read depth). Top four tracks show mammoths (Wrangel, Oimyakon, M4, M25). Bottom four tracks show Asian elephants (Emelia, Asha, Parvathy, Uno). Box indicates region containing RNase L exons.

Additionally, we identified nine derived SNVs predicted to occur in 3–9 of the RNase L copies and with no evidence in any of the Asian elephant samples. Six of these were non-synonymous (Table 8) and may reflect adaptation in woolly mammoths. We used T-Coffee (Notredame et al. 2000) to align the RNase L protein sequence for mammoth with the protein sequence of 16 other species, including human, mouse, and cold-adapted species polar bear, alpine marmot, and walrus (Supplementary Fig. S1). A more concise alignment of human and African elephant RNase L protein sequence is shown in Fig. 15. Substitution S34I is adjacent to residue G35 (Supplementary Fig. S1), which is involved in 2–5A interaction (Huang et al. 2014), needed for dimerization of RNase L and activation of its antiviral activity. V34 was most prevalent with S34 occurring in African and Asian elephants, T34 in dolphin, and I34 in mammoth, Alpine marmot, and platypus. T322A is close to several residues involved in self-domain dimerization (Y310, S312, R316, and L319)(Huang et al. 2014) and 2-5A sensing (Y308, Y310) (Han et al. 2014). V322 was predominant with I322 occurring in Anole lizard, T322 in African and Asian elephants, and A322 in mammoth, cow, sheep, armadillo, dolphin, and platypus. Of interest, woolly mammoth substitution R675K was identified as a ribonuclease active site in pigs by Huang et al. (2014), who has shown that mutant pK672A (mammoth 675) results in a defective RNase L. In this position, K occurs in all of the aligned species except African and Asian elephants, which have R, and in a previous alignment by Huang, R occurs in chicken (Huang et al. 2014). Both R and K occur in the mammoth. Together with the significant amplification of the locus in the mammoth genomes, these variants likely reflect adaptation processes related to RNase L functionality and specificity.

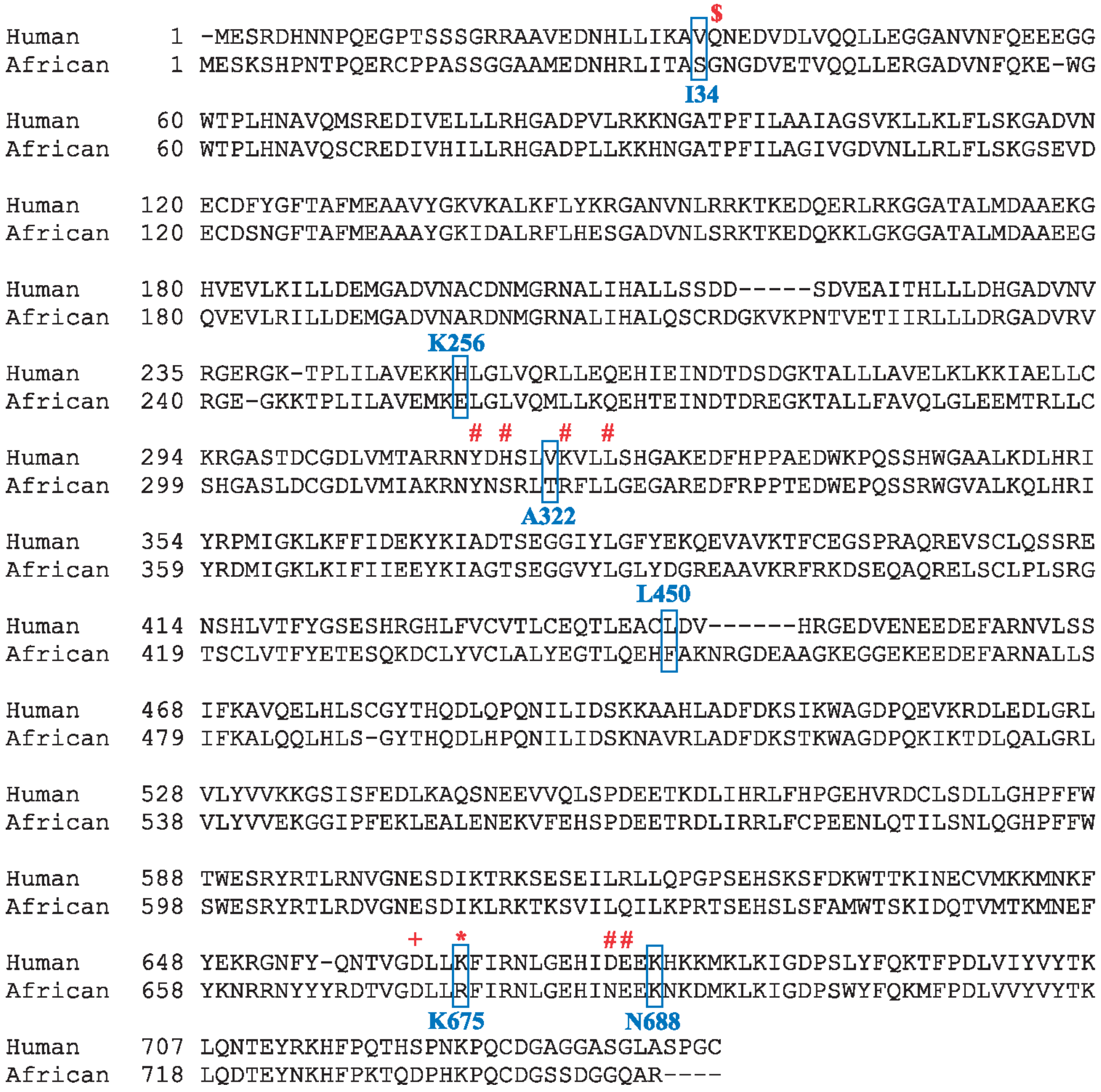


Location	Codons (elephant/mammoth)	Amino acids (elephant/mammoth)	Protein position	Domain
scaffold_16:17939894	aGt/aTt	S/I	34	ANK
scaffold_16:17940559	Gag/Aag	E/K	256	ANK
scaffold_16:17940757	Aca/Cca	T/A	322	–
scaffold_16:17941141	Ttt/Ctt	F/L	450	Protein kinase
scaffold_16:17965935	aGg/aAg	R/K	675	RNase
scaffold_16:17965975	aaG/aaC	K/N	688	RNase

Variants occur in 3–9 copies of RNase L. Mammoth predicted to have 5–9 copies of RNase L.

Table 8.

Amino acid variants in mammoth RNase L



```
                                               $
Human      1 -MESRDHNNPQEGPTSSSGRRAAVEDNHLLIKAVQNEDVDLVQQLLEGGANVNFQEEEGG
African    1 MESKSHPNTPQERCPPASSGGAAMEDNHRLITASGNGDVETVQQLLERGADVNFQKE-WG
                                              I34

Human     60 WTPLHNAVQMSREDIVELLLRHGADPVLRKKNGATPFILAAIAGSVKLLKLFLSKGADVN
African   60 WTPLHNAVQSCREDIVHILLRHGADPLLKKHNGATPFILAGIVGDVNLLRLFLSKGSEVD

Human    120 ECDFYGFTAFMEAAVYGKVKALKFLYKRGANVNLRRKTKEDQERLRKGGATALMDAAEKG
African  120 ECDSNGFTAFMEAAAYGKIDALRFLHESGADVNLSRKTKEDQKKLGKGGATALMDAAEEG

Human    180 HVEVLKILLDEMGADVNACDNMGRNALIHALLSSDD-----SDVEAITHLLLDHGADVNV
African  180 QVEVLRILLDEMGADVNARDNMGRNALIHALQSCRDGKVKPNTVETIIRLLLDRGADVRV
                             K256
Human    235 RGERGK-TPLILAVEKKHLGLVQRLLEQEHIEINDTDSDGKTALLLAVELKLKKIAELLC
African  240 RGE-GKKTPLILAVEMKELGLVQMLLKQEHTEINDTDREGKTALLFAVQLGLEEMTRLLC
                               # #   #  #
Human    294 KRGASTDCGDLVMTARRNYDHSLVKVLLSHGAKEDFHPPAEDWKPQSSHWGAALKDLHRI
African  299 SHGASLDCGDLVMIAKRNYNSRLTRFLLGEGAREDFRPPTEDWEPQSSRWGVALKQLHRI
                                   A322
Human    354 YRPMIGKLKFFIDEKYKIADTSEGGIYLGFYEKQEVAVKTFCEGSPRAQREVSCLQSSRE
African  359 YRDMIGKLKIFIIEEYKIAGTSEGGVYLGLYDGREAAVKRFRKDSEQAQRELSCLPLSRG
                                           L450
Human    414 NSHLVTFYGSESHRGHLFVCVTLCEQTLEACLDV------HRGEDVENEEDEFARNVLSS
African  419 TSCLVTFYETESQKDCLYVCLALYEGTLQEHFAKNRGDEAAGKEGGEKEEDEFARNALLS

Human    468 IFKAVQELHLSCGYTHQDLQPQNILIDSKKAAHLADFDKSIKWAGDPQEVKRDLEDLGRL
African  479 IFKALQQLHLS-GYTHQDLHPQNILIDSKNAVRLADFDKSTKWAGDPQKIKTDLQALGRL

Human    528 VLYVVKKGSISFEDLKAQSNEEVVQLSPDEETKDLIHRLFHPGEHVRDCLSDLLGHPFFW
African  538 VLYVVEKGGIPFEKLEALENEKVFEHSPDEETRDLIRRLFCPEENLQTILSNLQGHPFFW

Human    588 TWESRYRTLRNVGNESDIKTRKSESEILRLLQPGPSEHSKSFDKWTTKINECVMKKMNKF
African  598 SWESRYRTLRDVGNESDIKLRKTKSVILQILKPRTSEHSLSFAMWTSKIDQTVMTKMNEF
                           +  *         ##
Human    648 YEKRGNFY-QNTVGDLLKFIRNLGEHIDEEKHKKMKLKIGDPSLYFQKTFPDLVIYVYTK
African  658 YKNRRNYYYRDTVGDLLRFIRNLGEHINEEKNKDMKLKIGDPSWYFQKMFPDLVVYVYTK
                             K675         N688
Human    707 LQNTEYRKHFPQTHSPNKPQCDGAGGASGLASPGC
African  718 LQDTEYNKHFPKTQDPHKPQCDGSSDGGQAR----
```

Figure 15.
African elephant RNase L alignment to human RNase L. Wolly mammoth residue in bold text are shown next to the corresponding elephant and human residues (boxed). Residues of interest near or coinciding with woolly mammoth amino acid substitutions marked above with the following signs: $ (2-5A interaction site), # (self-domain dimerization), or * (ribonuclease active site). RNases L cleavage site H683 (hH672) is marked with a plus sign (+).

RNase L is able to selectively target specific cellular RNA (Brennan-Laun et al. 2014) and is involved in antiviral activity against numerous virus families (Ezelle et al.

2016). RNase L activation in virus-infected cells was shown to trigger the NLRP3 inflammasome (Chakrabarti et al. 2015), the latter being implicated in the host response to many different types of RNA and DNA viruses, including herpesviridae (Nour et al. 2011). Notably, elephant endotheliotropic herpesviruses (EEHV) can cause a highly fatal hemorrhagic disease when transmitted to young Asian elephants: two available genomes of Asian elephants were obtained postmortem from the animals affected by this disease (Dastjerdi et al. 2014). The EEHV genotypes found in African elephants appear to be generally less virulent (Richman et al. 2014).

Arctic conditions may be of additional significance with regard to the antiviral action of this protein. Temperature-dependent transmission of rotavirus (family Reoviridae) has been shown in humans, with a 13 percent decrease in infections per 1°C increase in temperature above 5°C (Atchison et al. 2010). In river water in Japan, the peak reovirus level was found in winter during the cold weather months (Tani et al. 1995). Influenza virus was found to favor cold and dry conditions for transmission in guinea pigs (Lowen et al. 2007). Macaques infected with SRV-4, family Retroviridae, had less SRV-4 antibodies in cold weather (Zao et al. 2011). Mammoths lived in cold dry steppe tundra, and likely in close matriarch-led groups similar to modern elephants, possibly increasing pressure for the species to evolve defenses against viruses adapted to their environment. Not surprisingly, RNase L KO mammalian phenotypes included increased susceptibility to viral infection, as well as abnormal thymus morphology, enlarged thymus, and thymus hyperplasia (Supplementary Table S1).

Given the close arrangement of the SNVs near the ribonuclease active site, we further analysed their co-occurrence in the individual reads (Fig. 16). For all mammoths, non-synonymous A allele (R675K, aGg/aAg, scaffold 16:17,965,935) always co-occurred with synonymous A allele (H683, scaffold_16:17,965,951) for reads containing both nucleotide locations (Fig. 16). The patterns of co-occurrence of the non-synonymous C allele (K688N, scaffold_16:17,965,975) with the A alleles at 17,965,935 and 17,965,951 appeared to be linked with the clade structure. The nine RNase L copies of M4 showed almost equal occurrence of GGG (same as the African elephant reference) or AAC (all three nucleotides substituted compared with the reference) haplotypes at these three positions. Another clade I member, Wrangel, showed prevalence of AAC and GGC in the reads containing the three SNV locations, while GGC haplotype was strongly preferred in the clade II genomes.

Wrangel (I)	M4 (I)	Oimyakon (II)	M25 (II)
2	15	2	7
16	39	10	4
6	2	0	2
1	0	2	0
1	3	9	5
2	6	0	0
5 (11)	22 (24)	7 (7)	3 (5)
2 (3)	0 (0)	1 (3)	4 (4)
7 (8)	3 (3)	22 (31)	17 (22)
2 (4)	24 (30)	0 (0)	4 (4)

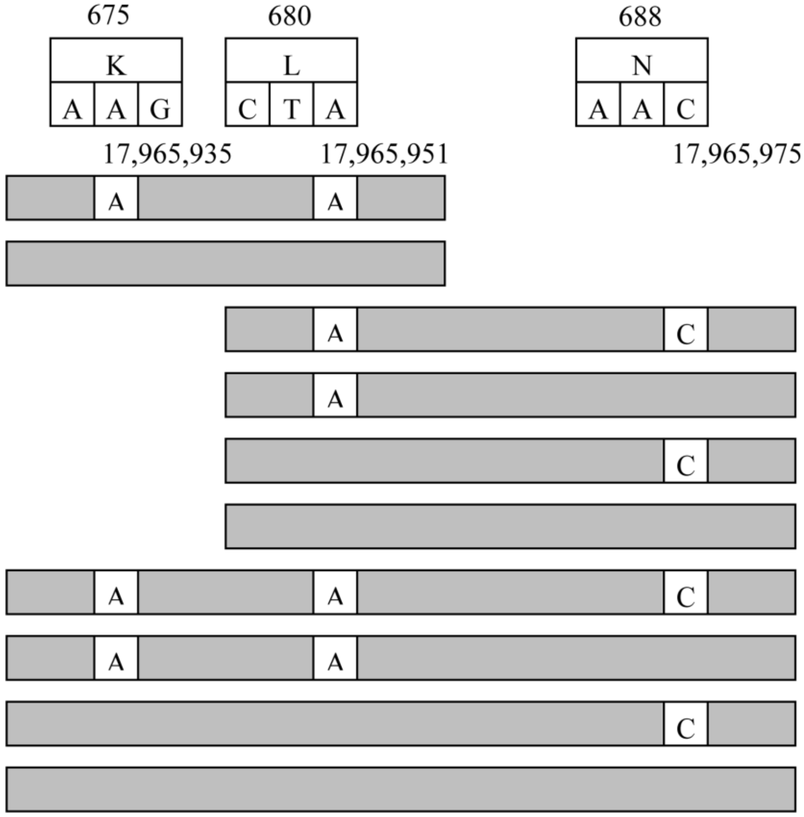
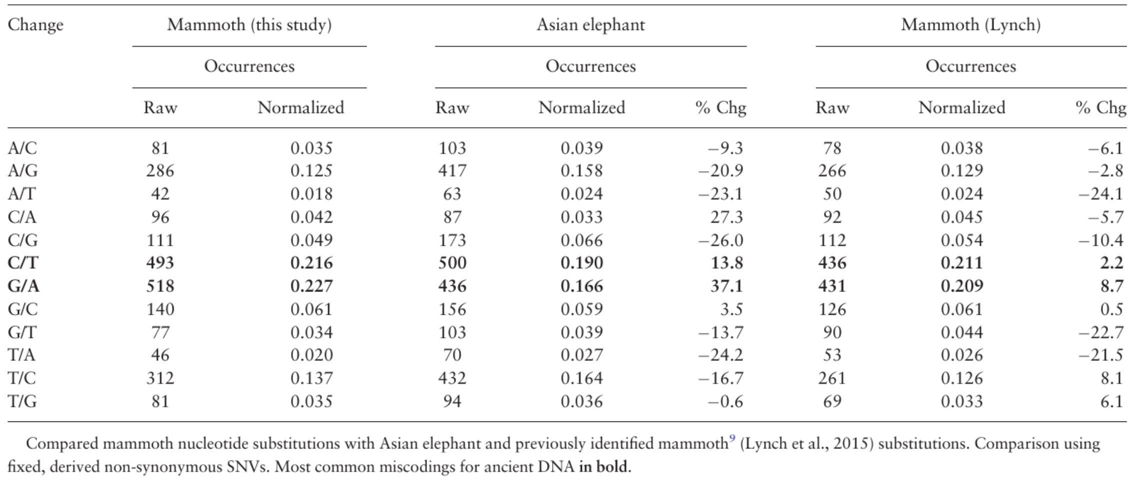


Figure 16.
Occurrences of woolly mammoth SNV combinations in reads for RNase L domain variants. Read counts for Wrangel, M4, Oimyakon, and M25 are left of reads. Read counts in parentheses indicate addition of inferred counts based on the observation that SNVs for residues 675 and 680 always co-occur. Most frequent haplotypes highlighted in gray. Residue number above reads. SNV combinations with no occurrences are not shown.

SNVs

We found 836,806 (606,176 intergenic; 224,483 intronic; 6,147 exonic) fixed, derived woolly mammoth SNVs, including 2,283 fixed, derived amino acid substitutions in 1,888 protein-coding genes, four protein-coding genes with a start loss and 18 protein-coding genes with pre-mature stop codons (Supplementary Material S1). We note that Lynch et al. (2015) has used similar criteria (see Materials and methods) to find a higher total of fixed, derived woolly mammoth SNVs (1.4 million compared with 0.8 million).

This was expected considering our study had more samples (four woolly mammoths, four Asian elephants) compared with the Lynch study (two woolly mammoths, three Asian elephants), likely decreasing false positives. However, Lynch et al. had fewer fixed, derived woolly mammoth amino acid substitutions (2,020 compared with 2,283). Several factors may have contributed to the discrepancy. For instance, we did not hard-mask potential cytosine deamination sites or filter short reads, and instead used a higher mapping quality threshold (>20) compared with Lynch et al.'s threshold (>0). Ancient DNA is known to have miscodings of C to T and G to A (Briggs et al. 2007). However, Briggs et al. (2007) has suggested that with sufficient coverage nucleotide misincorporations should not prevent a reliable Neandertal or mammoth genome sequence from being determined. We reasoned that our potentially less stringent thresholds (to increase sensitivity) would be adequate considering we analysed four mammoth samples, each with ≥10× coverage, and filtered SNVs using four Asian elephants (see Supplementary Table S2 for coverages). We tested for ancient DNA biases by comparing fixed, non-synonymous variants in woolly mammoths (2,283), Asian elephants (2,634), and the Lynch study woolly mammoths (2,020). We found that rates (as a fraction of all substitutions) for fixed, non-synonymous variants in the mammoths were similar for substitutions corresponding to the most common potential miscodings, C → T (0.216, present study; 0.211, Lynch study) and G → A (0.227, present study; 0.209, Lynch study) (Table 9), and thus our variant collection was unlikely to have been biased by these miscodings compared with the variants in the Lynch study. The rates in mammoths were only moderately higher than those in Asian elephants (Table 9).



Change	Mammoth (this study)		Asian elephant			Mammoth (Lynch)		
	Occurrences		Occurrences			Occurrences		
	Raw	Normalized	Raw	Normalized	% Chg	Raw	Normalized	% Chg
A/C	81	0.035	103	0.039	−9.3	78	0.038	−6.1
A/G	286	0.125	417	0.158	−20.9	266	0.129	−2.8
A/T	42	0.018	63	0.024	−23.1	50	0.024	−24.1
C/A	96	0.042	87	0.033	27.3	92	0.045	−5.7
C/G	111	0.049	173	0.066	−26.0	112	0.054	−10.4
C/T	**493**	**0.216**	**500**	**0.190**	**13.8**	**436**	**0.211**	**2.2**
G/A	**518**	**0.227**	**436**	**0.166**	**37.1**	**431**	**0.209**	**8.7**
G/C	140	0.061	156	0.059	3.5	126	0.061	0.5
G/T	77	0.034	103	0.039	−13.7	90	0.044	−22.7
T/A	46	0.020	70	0.027	−24.2	53	0.026	−21.5
T/C	312	0.137	432	0.164	−16.7	261	0.126	8.1
T/G	81	0.035	94	0.036	−0.6	69	0.033	6.1

Compared mammoth nucleotide substitutions with Asian elephant and previously identified mammoth[9] (Lynch et al., 2015) substitutions. Comparison using fixed, derived non-synonymous SNVs. Most common miscodings for ancient DNA **in bold**.

Table 9.
Nucleotide substitution comparisons

In agreement with an earlier analysis of KEGG pathways and KO phenotypes (Lynch et al. 2015), we analysed fixed, derived woolly mammoth amino acid substitutions and found amongst 58 enriched KEGG pathways complement and coagulation cascades ($P = 5.5 \times 10{-}9$, $adjP = 4.2 \times 10{-}7$), fat digestion and absorption ($P = 0.0038$, $adjP = 0.021$), and circadian rhythm – mammal ($P = 0.034$, $adjP = 0.092$) enrichment, and our GO term analysis revealed 284 enrichments including lipid metabolic process ($P = 9.5 \times 10{-}4$) and homeostatic process ($P = 3.2 \times 10{-}4$) (Supplementary Material S1). Additionally, we were interested in identifying genes under positive selection. Given our limited sample population, we identified genes with maximum FST values (top 5%) (Zhou et al. 2016). We found 544 of 2,283 (24%) mammoth substitutions, affecting 446 of 1,888 (24%) protein-coding genes, were in the top 5% of FST values (Supplementary Material S1). Rather than repeat previous observations by Lynch et al. (which we also detected), we report a few interesting findings from our analysis, expanding the list of likely selected variants fixed across the mammoth genomes.

We found three fixed, derived non-synonymous woolly mammoth variants (R381K, A424V, I2640V) in APOB (0.99 FST rank). Multiple alignment, using MUSCLE (Edger 2004), of the APOB protein sequence for mammoth and 15 other species indicated the V substitution at mammoth residue position 424 was shared with the cold-adapted walrus, the cold-blooded Anole lizard, and platypus (Supplementary Fig. S2). Residue I was pre-dominant at mammoth residue position 2,640 except for a V substitution in mammoth, the cold-adapted Alpine marmot, and the cold-blooded Anole lizard. APOB codes an apolipoprotein for chylomicrons and LDL particles. High levels of the APOB protein have been linked to atherosclerotic plaques (McQueen et al. 2008). In a comparison of polar bears and brown bears, a recent study (Liu et al. 2014) noted nine non-synonymous APOB mutations in polar bears and suggested that a carnivorous diet of pre-dominantly fatty acids induced adaptive changes in APOB. Mammoths likely had a plant-based diet similar to modern elephants (Schwartz-Narbonne et al. 2015). Lower ApoB/ApoA1 plasma levels have been noted in cold-adapted human swimmers (Lensa et al. 2015). In carp, cold adaptation included up-regulation of six apolipoprotein genes including APOB (Gracey et al. 2004). We also found three non-synonymous variants in TRPM8, a protein that was briefly noted in Lynch et al. and is responsible for sensitivity to noxious cold. TRPM8 transmembrane domain S2 influences menthol binding and mutation Y745H (mouse) results in a loss of sensitivity to menthol (Montell 2006). In mammoth, mutations were R368H (h364), G710S (h706), and C711S (h707). On the basis of multiple protein sequence alignment, using T-Coffee, with 16 other organisms, mammoth was the only organism with a substitution of the consensus amino acid at these locations (Supplementary Fig. S3). G710S and C711S occur in S1 (residues 697–716).

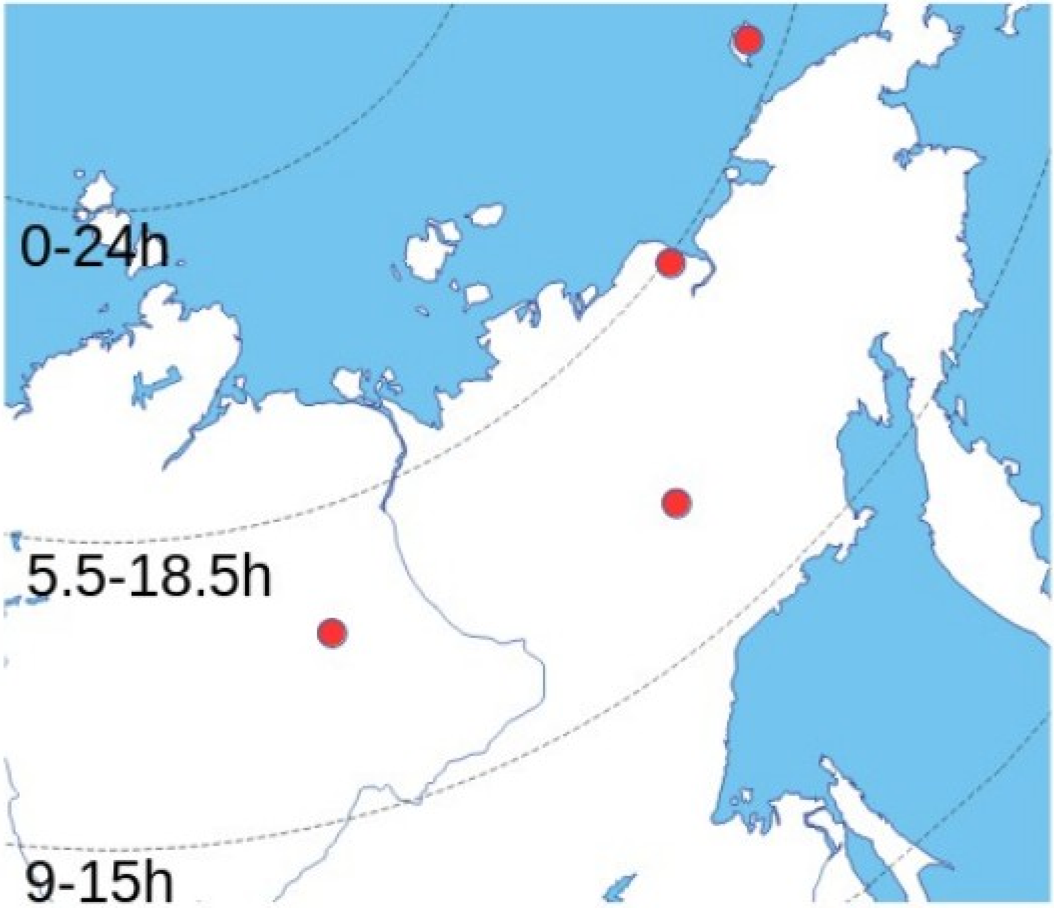
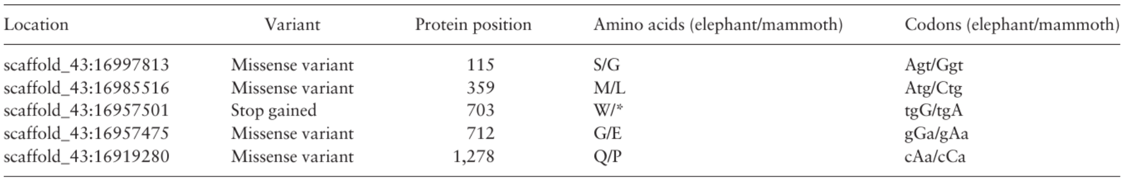


Figure 17. Variation in seasonal light hours at latitudes woolly mammoth samples were located.

Modifications to genes responsible for circadian rhythms have also been identified uniquely in the woolly mammoth. As previously reported by Lynch et al. (2015), woolly mammoth had a fixed, non-synonymous variant in PER2. We also found previously unreported fixed, non-synonymous woolly mammoth variants in several important clock genes: ARNTL (BMAL1) of 0.25 FST rank, CRTC1 (1.0 FST rank), KDM5A (0.86 FST rank), and KMT2A (1.0 FST rank). The ARNTL gene is of particular importance as a core circadian oscillator in mammals. ARNTL controls not only the innate 24-h cycle of the organism, but also the transcription of several circadian dependent genes. BMAL1 KO mice have been shown to lose core circadian control, and therefore become arrhythmic (Lowrey et al. 2011). Although loss of control may not be beneficial, a weakening of this core oscillator could potentially aid the mammoth in adjusting its behaviour and compensating for the large differences of daylight experienced in summer compared with winter (Fig 17.). Another important molecular component of the circadian clock, CRTC1, has a substitution identified in woolly mammoths. This protein is an important member of the signaling pathway for light response and setting the circadian clock of an organism (Jagannath et al. 2013). Alterations of this gene can possibly help to force the innate 24-h cycle of an organisms to adapt to the continually variable light–dark cycle of its environment (Jagannath et al. 2014). This may have enabled the mammoth to more easily entrain to light and elicit a varied wake response throughout the changing seasons. KDM5A acts on the transcription of the BMAL1/CLOCK complex and works to dampen circadian oscillators (Katada et al. 2010). KMT2A also plays an essential role in altering the chromatin state of circadian controlled genes (Katada et al 2010). These adaptations are possibly due to the increased latitude that woolly mammoths inhabited, forcing a more flexible circadian clock as seasonal changes introduced an extremely varied cycle of light and darkness through the course of the year. Another highly mutated gene was NCKAP5 (0.98 FST rank). The function of NCKAP5 is unknown, but SNPs in NCKAP5 have been linked to hypersomnia (Khor et al. 2013).

KEGG analysis indicated enrichment of ‘pathways in cancer’ (P = 9.6 × 10−5, adjP = 0.0021), which lead us to investigate several cancer-related genes that had multiple non-synonymous variants. BRCA1 (0.98 FST rank), BRCA2 (0.52 FST rank), and PARP14 (0.84 FST rank) had three, seven, and seven non-synonymous variants, respectively. Although BRCA1 and BRCA2 have vital functions, rapid evolution of BRCA1 and BRCA2 has been observed in mammals (Lou et al. 2014). Citing that ‘nearly all known cases of recurrent positive selection in primate genomes involve genes in one of three categories: (i) immunity, (ii) environmental perception (such as odorant and taste receptors), or (iii) sexual selection and mate choice’ (Lou et al. 2014, Clark et al. 2003, Vallender et al. 2004), Lou et al. proposed that rapid BRCA1 and BRCA2 evolution is due to adaptation to viruses. A recent study hypothesized that PARP14 is involved in host–virus defense due to the gene’s strong positive selection in primates (Daugherty et al. 2014). This may also explain the high number of fixed, derived woolly mammoth non-synonymous variants we found in PARP14, also involved in DNA repair. Although speculative, possible changes in the repair pathways in mammoth genomes in response to viruses might lend further support to our observation of active development of antiviral mechanisms via amplification of RNase L.

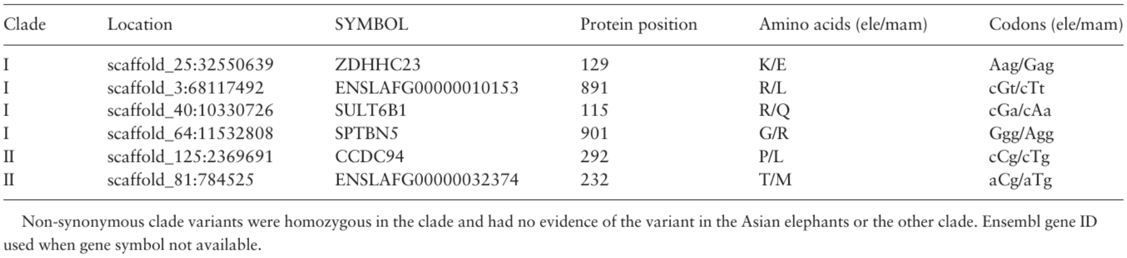
We also investigated the 18 genes with pre-mature stop codons, eight having gene symbols. Genes with pre-mature stop codons had mutations associated with ancient DNA bias (C → T, G → A) in 14 of 18 (78%) cases, considerably higher than the rate observed in Asian elephants (36%). However, one gene with a stop gain, ABCC11 (1.0 FST rank), had four non-synonymous variants, of which two occurred downstream of the stop gain, W703* (Table 10). ABCC11 has been shown to determine wet or dry ear wax (Yoshiura et al. 2006) in human populations, with the ear wax single nucleotide polymorphism, rs17822931-G/A (G180R), shown to be absent in African populations and increase in prevalence with absolute latitude (Ohashi et al. 20111). Ohashi et al. also indicated that rs17822931-A resulted in loss of function. ABCC11 also has roles involving bile acids, conjugated steroids, and cyclic nucleotides (Chen et al. 2005, Guo et al. 2003).



Location	Variant	Protein position	Amino acids (elephant/mammoth)	Codons (elephant/mammoth)
scaffold_43:16997813	Missense variant	115	S/G	Agt/Ggt
scaffold_43:16985516	Missense variant	359	M/L	Atg/Ctg
scaffold_43:16957501	Stop gained	703	W/*	tgG/tgA
scaffold_43:16957475	Missense variant	712	G/E	gGa/gAa
scaffold_43:16919280	Missense variant	1,278	Q/P	cAa/cCa

Table 10. Fixed, derived non-synonymous woolly mammoth variants in ABCC11

Mitochondrial phylogenetic analysis has suggested two primary clades of woolly mammoth, clades I and II, that are thought to have evolved in isolation on opposite sides of the Bering Strait (Palkopoulou et al. 2013). Our dataset contained two mammoths from each major clade, clade I (Wrangel and M4) and clade II (Oimyakon and M25). Clade II disappeared ∼30,000–40,000 years prior to the extinction of clade I. We investigated the differences between the genomes of clades I and II and found four fixed, derived clade I CNVs, three deletions and one duplication, none of which occurred in exonic regions. Similarly, there were two fixed, derived clade II CNVs, both duplications, but neither occurred in exons. We found 57 fixed, derived clade I indels (11 insertions and 46 deletions), and 65 fixed, derived clade II indels (24 insertions and 41 deletions), none of which occurred in exons. For SNVs, we found 1,215 fixed, derived clade I variants, of which 279 and five occurred in introns and exons, respectively. Four were non-synonymous variants in protein-coding genes (Table 11). Clade II had 584 fixed SNVs, 125 and three occurred in introns and exons, respectively. Two non-synonymous clade II variants occurred in protein-coding genes (Table 11).



Clade	Location	SYMBOL	Protein position	Amino acids (ele/mam)	Codons (ele/mam)
I	scaffold_25:32550639	ZDHHC23	129	K/E	Aag/Gag
I	scaffold_3:68117492	ENSLAFG00000010153	891	R/L	cGt/cTt
I	scaffold_40:10330726	SULT6B1	115	R/Q	cGa/cAa
I	scaffold_64:11532808	SPTBN5	901	G/R	Ggg/Agg
II	scaffold_125:2369691	CCDC94	292	P/L	cCg/cTg
II	scaffold_81:784525	ENSLAFG00000032374	232	T/M	aCg/aTg

Non-synonymous clade variants were homozygous in the clade and had no evidence of the variant in the Asian elephants or the other clade. Ensembl gene ID used when gene symbol not available.

Table 11. Fixed, derived non-synonymous clade variants

Clade I genes with fixed, derived amino acid substitutions included ZDHHC23, SULT6B1, and SPTBN5. Of most interest, SULT6B1 is a sulfotransferase that utilizes 3-phospho-5-adenylyl sulphate (PAPS) to catalyze the sulphate conjugation of thyroxine and is involved in the metabolism of thyroxine (Takahashi et al. 2009). Unlike other SULTs, SULT6B1 is uniquely specific for thyroxine, suggesting a role in regulation of thyroxine (Takahashi et al. 2009). Thyroxine (T4) is a thyroid hormone involved in growth, development, differentiation, and basal metabolic homeostasis, as well as the regulation of protein, fat, and carbohydrate metabolism (Bates et al. 1988, Boelen 2009, Pucci et al. 2000). Also, it is involved in facultative thermogenesis (heat production in response to cold or overeating) (Silva 2001). Administering thyroid hormones results in a severe drop in body temperature (Doyle et al. 2007).

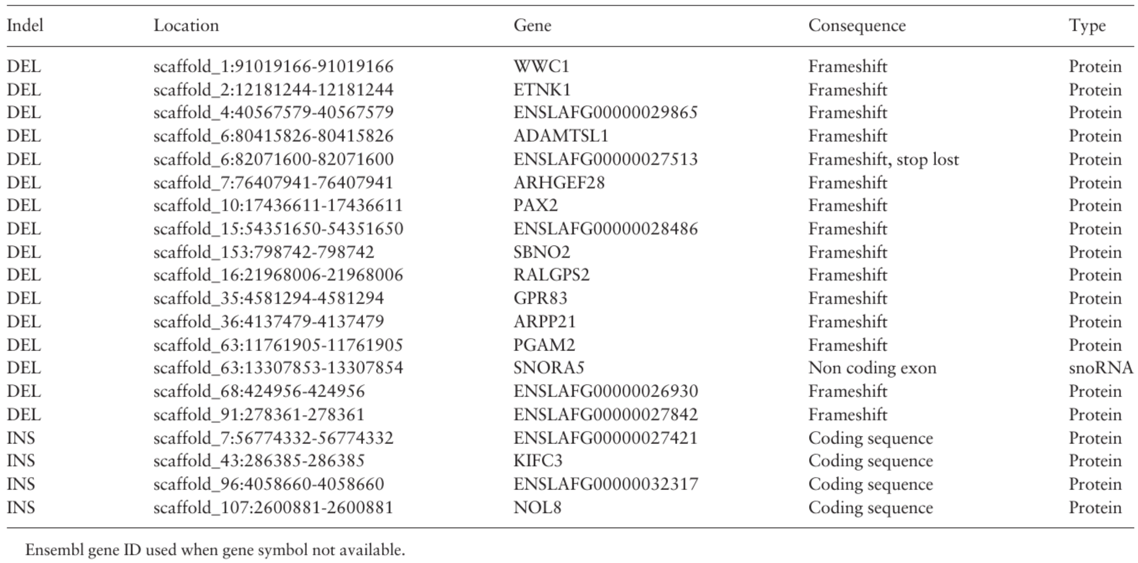
Clade II had a fixed, derived amino acid substitution in CCDC94. CCDC94 knockdown in zebrafish increased sensitivity of ionizing radiation-induced apoptosis and CCDC94 protects cells from ionizing radiation-induced apoptosis by repressing expression of p53 mRNA (Sorrells et al. 2012).

GO terms and KO mammalian phenotypes for genes with fixed, derived non-synonymous clade I and II variants are summarized in Supplementary Tables S3 and S4, respectively. Interestingly, fixed, derived non-synonymous clade I variant ZDHHC23 and clade II variant CCDC94 had KO phenotypes ‘increased caudal vertebrae number’ and ‘decreased caudal vertebrae number’, respectively. At the time of writing, it was unknown to us if clade I and clade II differed in their number of tail bones. The caudal vertebrae number of woolly mammoths has been estimated to be 21 but has not been ‘confidently established’ (Haynes 1993) as few complete skeletons have been found. Modern elephants have (Brennan-Laun et al. 2014, Fabre et al. 2012, Salehzada et al. 2009, Huang et al. 2014, Han et al. 2014, Ezelle et al. 2016) 28-33 caudal bones (Haynes 1993). Woolly mammoth shorter tail adaptation likely reduced heat loss and frostbite (Campbell et al. 2010).

Indels

We found 20,576 fixed, derived woolly mammoth indels, 2,413 insertions and 18,163 deletions. Of these, 597 insertion indels and 5,174 deletion indels overlapped introns. Exons were found to harbor four insertion and 16 deletion indels (Table 12). All fixed, derived woolly mammoth indels were <3 bases in length resulting in frameshifts and were classified as high impact variants by VEP (ensembl.org). A number of protein-coding gene regions disrupted by an indel showed interesting functions, potentially relevant to woolly mammoth adaptation. Supplementary Table S5 lists GO terms and KO phenotypes for fixed, derived woolly mammoth indels affecting exonic regions. WWC1 functions as a tumor suppressor regulating Hippo signaling (Xiao et al. 2011, Yu et al.

2010) and has GO term ‘negative regulation of organ growth’. In agreement with SNV analysis in the present study and by Lynch et al. (2015), we found an abundance of fixed, derived woolly mammoth variants in genes involved with lipid metabolism, possibly an adaptation for storing fat (e.g., ETNK1, lipid metabolic process).



Indel	Location	Gene	Consequence	Type
DEL	scaffold_1:91019166-91019166	WWC1	Frameshift	Protein
DEL	scaffold_2:12181244-12181244	ETNK1	Frameshift	Protein
DEL	scaffold_4:40567579-40567579	ENSLAFG00000029865	Frameshift	Protein
DEL	scaffold_6:80415826-80415826	ADAMTSL1	Frameshift	Protein
DEL	scaffold_6:82071600-82071600	ENSLAFG00000027513	Frameshift, stop lost	Protein
DEL	scaffold_7:76407941-76407941	ARHGEF28	Frameshift	Protein
DEL	scaffold_10:17436611-17436611	PAX2	Frameshift	Protein
DEL	scaffold_15:54351650-54351650	ENSLAFG00000028486	Frameshift	Protein
DEL	scaffold_153:798742-798742	SBNO2	Frameshift	Protein
DEL	scaffold_16:21968006-21968006	RALGPS2	Frameshift	Protein
DEL	scaffold_35:4581294-4581294	GPR83	Frameshift	Protein
DEL	scaffold_36:4137479-4137479	ARPP21	Frameshift	Protein
DEL	scaffold_63:11761905-11761905	PGAM2	Frameshift	Protein
DEL	scaffold_63:13307853-13307854	SNORA5	Non coding exon	snoRNA
DEL	scaffold_68:424956-424956	ENSLAFG00000026930	Frameshift	Protein
DEL	scaffold_91:278361-278361	ENSLAFG00000027842	Frameshift	Protein
INS	scaffold_7:56774332-56774332	ENSLAFG00000027421	Coding sequence	Protein
INS	scaffold_43:286385-286385	KIFC3	Coding sequence	Protein
INS	scaffold_96:4058660-4058660	ENSLAFG00000032317	Coding sequence	Protein
INS	scaffold_107:2600881-2600881	NOL8	Coding sequence	Protein

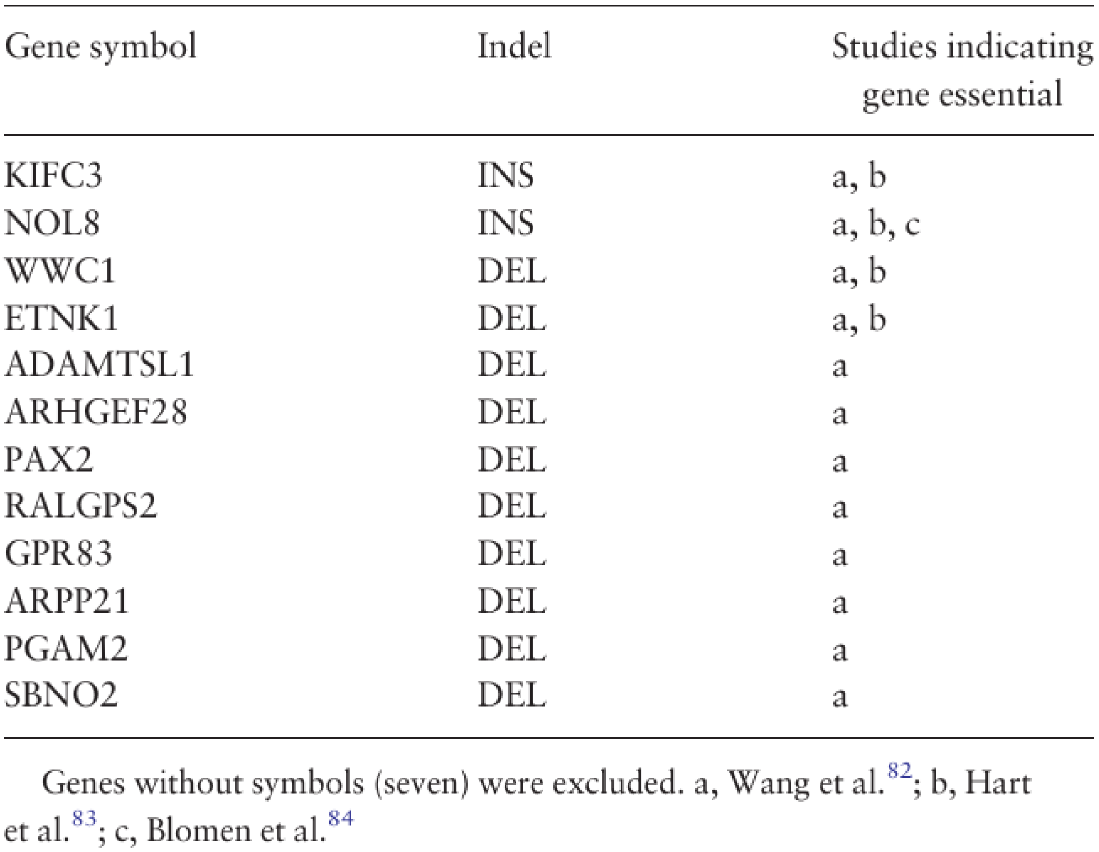
Ensembl gene ID used when gene symbol not available.

Table 12. Fixed, derived woolly mammoth indels occurring in exons

Of particular interest, GPR83 is a member of the G protein-coupled receptor subfamily (Oh et al. 2006) that includes several receptors linked to the regulation of metabolism.GPR83 is expressed in the hypothalamus and regulated by nutrient availability (Muller et al. 2013). GPR83 knock-out mice have shown a significant increase in food intake compared with wild-type mice but appeared to be protected from obesity and glucose intolerance and have normal insulin sensitivity, regardless of diet type (Muller et al. 2013). Another study found that GPR83 knockdown mice had increased levels of weight gain compared with wild-type mice, despite consuming an identical amount of food. Additionally, the knockdown of GPR38 promoted a decrease in core body temperature during the daily activity period (Dubins et al. 2012). Increases in weight gain and lower core body temperatures have shown to be advantageous to mice when presented with lower ambient environmental temperatures. Mice that had developed these traits were shown to have a stable metabolic rate and more easily maintained core body temperatures in ambient cold temperatures compared with their wild-type counterparts (Kaplan et al. 1974). These studies suggest a possible role for modification or loss-of-function of GPR83 transcripts in the survival of mammoths considering the harsh environment of their habitat.

Another affected gene, RALGPS2, had one mammalian KO phenotype, 'increased rib number'. In mammals, the number of cervical vertebrae is highly conserved at seven. Approximately 90 percent of humans with a cervical rib die before reaching reproductive age (Galis et al. 2006), due to a strong association with multiple major congenital abnormalities. However, a recent study (Reumer et al. 2014) found abnormal cervical rib numbers in Late Pleistocene mammoths (33%), a rate 10 times higher than in extant elephants (3.6%), which appears consistent with our finding. The abnormal numbers were due to large cervical ribs on the seventh vertebra and mammoth cervical ribs were relatively large compared with cervical ribs in humans (Reumer et al. 2014).

Several genes with fixed, derived woolly mammoth indels had mammalian KO phenotypes related to decreased body size: decreased birth body size (GPR83), decreased body length (ARHGEF28), and decreased body weight (NOL8), and as noted earlier, CD44 KO was linked to short tibias. Towards the end of their existence mammoths decreased in size (Haynes et al 1993). All samples in this study were near the end of the mammoths existence, from 60,000 years ago or more recent. The cause of the extinction of woolly mammoths is unknown but is often regarded to be due to a warming climate and hunting by humans. We note that several genes with fixed, derived woolly mammoth indels had detrimental reproductive KO phenotypes, such as NOL8 (pre-weaning lethality), SBNO2 (pre-weaning lethality), and PAX2 (postnatal lethality and perinatal lethality). ARHGEF28 had the mammalian KO phenotype ‘reproductive system’. Additionally, we found that all genes with fixed, derived woolly mammoth indels (and with a gene symbol) occurred in one or more recent screens of gene essentiality (Table 13).



Gene symbol	Indel	Studies indicating gene essential
KIFC3	INS	a, b
NOL8	INS	a, b, c
WWC1	DEL	a, b
ETNK1	DEL	a, b
ADAMTSL1	DEL	a
ARHGEF28	DEL	a
PAX2	DEL	a
RALGPS2	DEL	a
GPR83	DEL	a
ARPP21	DEL	a
PGAM2	DEL	a
SBNO2	DEL	a

Genes without symbols (seven) were excluded. a, Wang et al.[82]; b, Hart et al.[83]; c, Blomen et al.[84]

Table 13. Essentiality of genes fixed, derived woolly mammoth indels occuring in protein-coding regions.

Discussion

We performed systematic detection of variants, ranging from single nucleotide changes and short indels to deletions and amplifications of regions encompassing gene fragments and complete genes in four woolly mammoth genomes, comparing them to the elephant genomes. Extending the findings of previous SNV-only analyses, we found many unique woolly mammoth genetic variants involved in processes relevant to living in a harsh, cold environment, such as lipid metabolism and thermogenesis. We also observed changes in the core circadian oscillator genes that may have occurred as woolly mammoths adapted to the extremely varied cycle of light and darkness through the course of the year in the Arctic. Further, we noticed many immunity-related variants, most strikingly the copy number amplification of RNase L, important in antiviral response. Immunity tends to evolve rapidly and woolly mammoth genomes display signs of adaptation to the differences in environmental threats they faced compared with their tropical/subtropical elephant counterparts. The haplotype patterns in RNase L gene described above show similarities in clade II but differences in clade I, likely reflecting the history of development of multiple defense mechanisms, which appear to be absent (or reduced) in the modern elephants. Do multiple RNase L paralogs and multiple non-synonymous changes in the repair pathways indicate that mammoths faced much deadlier viral threats? Were those threats further exacerbated by the cold Arctic conditions? Was this possibly linked to the ultimate demise of mammoths?

African elephants have been shown to have at least 20 copies of the tumor suppressor TP53, possibly explaining their low cancer mortality despite their large size and long life span (Abegglen et al. 2015). Such TP53 locus amplification is also present in woolly mammoth genomes making RNase L their second largest CNV. It is worth noting that RNase L may also play a role in tumor suppression (Chakrabarti et al. 2011) and has been linked to prostate cancer (Meyer et al. 2010).

Recently, other avenues of analysis of ancient DNA have been proposed, related to CpG methylation and corresponding epigenetic patterns (as summarized in a recent review (Morozova et al. 2016)). Sampling based on methylated binding domains (MBD) has been performed in a number of organisms, including two other woolly mammoths (Seguin-Orlando et al. 2015). We analysed these MDB mammoth samples but found the genome coverage to be too fragmentary (as estimated by total SNP counts from all MDB- and MBD+ samples <0.3% of SNPs in a typical mammoth genome and by nearly negligible overlap with a union of SNPs from our study) to draw any meaningful conclusions.

And as for the differences between clades I and II, although only a few are reported here, some variants may represent continuing evolution of adaptations to a cold environment. Fixed non-synonymous variants in clade I (ZDHHC23) and clade II (CCDC94) have been shown to affect the number of tail bones and clade I non-synonymous variant SULT6B1 plays a role in thyroid hormone regulation. Since clade II disappeared during the last glacial period prior to rising temperatures, these variants might have contributed to the longer survival of clade I mammoths.

Taken together, these observations point to a rich population history of these iconic animals extending beyond the description based on three clades.

Chapter 5: Insight into current genetic health of eastern gorilla species with structural variant analysis

Gorilla beringei are endangered great apes with dwindling populations, and drastically differing population sizes between two gorilla subspecies, *G. b. beringei* and *G. b. graueri*. Many studies of genetic health and diversity in endangered populations are limited to SNV analysis, such as the previous work in gorilla by Xue et al. (2015). We performed analysis of structural variation in these two gorilla subspecies to identify additional possible genetic causes of divergent physical characteristics and disease phenotypes. We found subspecies unique genetic evidence related to physical characteristics as well as disease and abnormality.

As a first author of this work, my contributions included the analysis of the gorilla genomes using GROM, with help from S. Smith, and the comparison of variants found between the subspecies. Additionally, I identified all variants within gene exons, performed the enrichment analysis utilizing g:Profiler, and frequency plotting within subspecies. All variant examples within the paper were identified by me and I wrote the paper in its entirety with help from S. Smith and A. Grigoriev who reviewed the data and edited the paper, and M. Ozbas who assisted writing the section on cardiovascular health.

5.1 Analysis of gorilla populations

We analyzed whole genome sequence (WGS) for 12 gorillas from two gorilla subspecies, eastern lowland gorilla, *Gorilla graueri*, and the mountain gorilla, *Gorilla beringei*. Seven mountain gorillas and five eastern lowland gorillas were used in this study. Previous work by Xue et al. (2015) focused on the analysis of SNVs. However, structural variants (or SVs, which include copy number variants, insertions, and inversions) potentially account for more differences between genomes in terms of number of nucleotides, as shown in previous human studies (Baker 2012). Therefore we expanded the comparative analysis of these two gorilla populations to include such structural variation.

We classified mutations as either subspecies unique (no evidence of mutation in reciprocal species) or subspecies shared (mutation potentially occurred in any individual). This allowed us to further analyze mutation enrichments that could be part of the potential causation for the observed phenotypic differences between the subspecies. Furthermore, we checked for enrichment of mutations that crossed genes associated with disease phenotypes, an important concern for gorillas dealing with a high amount of population loss.

A comprehensive list of mutations (Table 14) was detected using GROM. We report on the significance of CNVs, SVs, and indels in the sampled gorilla populations as well as highlight several interesting findings of affected genes in exonic regions. This includes mutations that are prevalent in the sampled gorilla populations that have been previously unreported and associate with known expressed phenotypes.

Variant	Eastern Lowland	Mountain
Deletion	31,316 (2,556)	29,015 (2,617)
Duplication	4,882 (235)	4,969 (554)
Inversion	6,088 (387)	5,874 (552)
Indel (deletion)	584,952 (112,360)	614,611 (140,864)
Indel (insertion)	506,954 (93,706)	527,239 (123,081)
SNV	8,365,879 (706,399)	8,639,860 (875,385)

Table 14. Total (average) number of each type of variant found in each gorilla subspecies.

5.2 Structural variant enrichment

We found that the eastern gorilla populations have significant enrichment of SV-affected genes that associate with digit development and formation, as well as the specific phenotype of syndactyly. Several reported cases have been made of gorillas being born with toe or finger syndactyly (Cooper et al. 2017, Fossey 1983, Mudakikwa et al. 2001, Routh et al. 1997). This adverse phenotype is often prevalent when population size is significantly limited and is indicative of increased inbreeding (Xue et al. 2015). Additionally, the mountain gorilla population shows specific enrichment for toe syndactyly. This mutation is mirrored by findings in the field, where mountain gorilla seems to be particularly susceptible to this abnormality (Cooper et al. 2017, Mudakikwa et al. 2001). In total we found within the sampled populations SVs crossing 48 different named genes that associate with the syndactyly phenotype.

A significant amount of SV enrichment was found in genes associated with physical characteristics, many of which define the two subspecies. This includes subspecies specific mutations in genes that affect body structure such as stature, specifically a shorter thorax and heavier midsection, as well as clavicle development. These phenotype associations could potentially explain the difference in torso size and arm length between the subspecies. There were 21 different named genes that intersected a SV and associated with body structure development. There was also an enrichment of subspecies mutations for development of facial shape, with mutations crossing 45 associated named genes. Some interesting developmental phenotypes that associated with these genes included the choanae, as well as a pronounced mandible and spacing of teeth. These facial enrichments reflect what has been described as key distinguishing traits that separate the closely related eastern lowland gorilla from the mountain gorilla (Schultz 1934, Stanford 2001, Varsha 2010).

5.3 Analysis of mutations within coding genes

A particularly interesting SV was a heterozygous deletion that crossed several exons in the gene KCNU1 (Figure 18). This gene is believed to be critically responsible for the development of healthy sperm and therefore male fertility. Studies in knockout mice have demonstrated that the gene KCNU1 is critical for male fertility (Santi et al. 2010). This could be a potentially important connection to a previous study by Beck (1982) who surveyed several captive gorilla populations and found that only 21% of adult males were able to sire offspring. The heterozygous mutation in KCNU1 spans 7 exons.

It should also be noted that our algorithm, GROM, also found a homozygous deletion, within the larger heterozygous deletion.

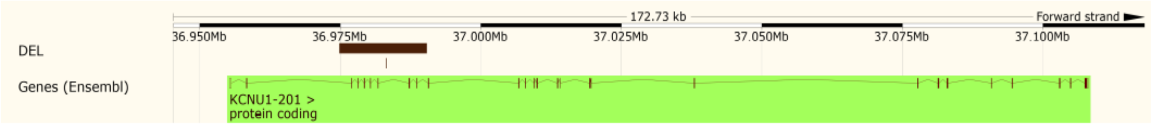


Figure 18. A visual representation of the deletion that is present within the KCNU1 gene.

Additional indel mutations have been found in other genes associated with fertility. This includes a heterozygous indel in PLTP that is unique to mountain gorillas. A deficiency of the plasma phospholipid transfer protein in a mice study was associated with reduced sperm motility as well as impairment of fertility in males (Drounieaud et al. 2006). This could contribute to the low motile or “poor quality” sperm in various studies of gorilla sperm evaluation (Seager et al. 1982, Schaeffer et al. 1992). In particular, it was reported that female mountain gorillas received less sperm during copulation (Watts 1990). Similarly we found indels in ESR1 and NRIP1 of the eastern lowland population. Both ESR1 and NRIP1 have been indicated in a multi-locus analysis to contribute to human male infertility (Galan et al. 2005).

PLTP also plays an important role in transferring phospholipids and cholesterol from triglyceride-rich lipoproteins into high-density lipoprotein (HDL) (Jiang et al. 1999). Mutations in PLTP are linked to an increased risk of high HDL cholesterol and a higher risk for cardiac diseases and atherosclerosis (Vergeer et al. 2010, Jiang et al. 2012). Studies with PLTP knockout mice characterized low high-density levels causing less absorbance of cholesterol (Liu et al. 2007, Jiang et al. 2012). Several unique mutations were identified in the ABC transporter for both gorilla populations, specifically ABCG1, ABCA5, and ABCA12. All three genes play critical roles in cholesterol regulation, accumulation, and efflux (Ye et al. 2010, Fu et al. 2013, Yvan-Charvet et al. 2010). While ABCA12, ABCG1 and ABCA1 deficiencies correlate to the development of atherosclerosis (Yvan-Charvet et al. 2007, Fu et al. 2013).

Mountain gorillas in captivity were found to have a lower prevalence of heart disease than eastern lowland gorilla (Cooper et al., 2017). A cholesterol concentration survey by Schmidt et al. (2006) depicted non-significantly varying cholesterol levels between samples from free-ranging eastern gorillas and free-ranging western lowland gorillas. Even with near equal levels of cholesterol only about 3% of mountain gorilla deaths were cardiovascular disease-related while eastern lowland gorillas have a higher death prevalence along with higher documented incidences of myocardial fibrosis and arteriosclerosis than in an isolated eastern gorilla population (Cooper et al. 2017).

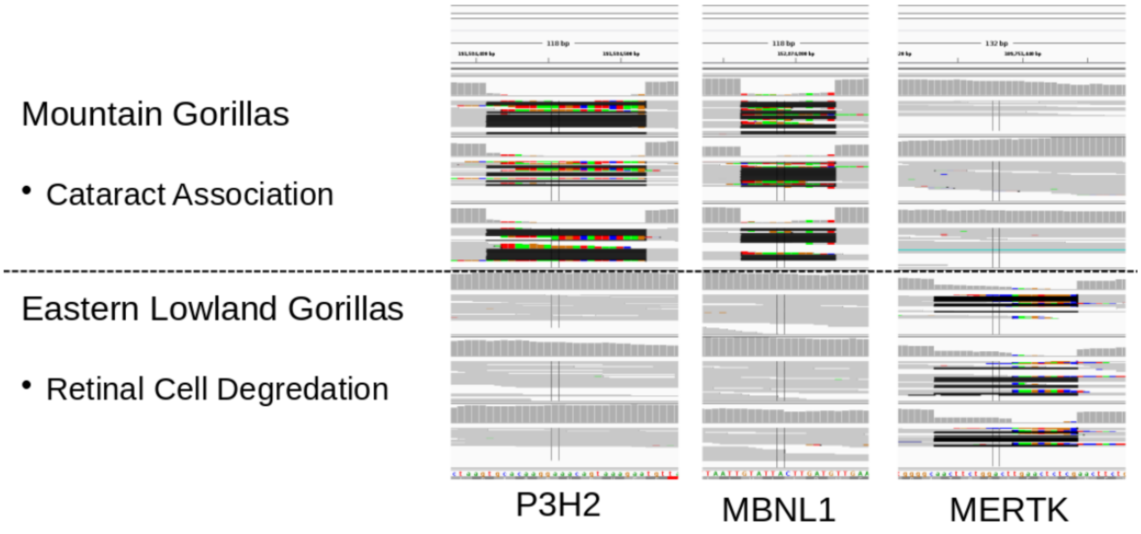


Figure 19. Sub-species specific indels related to vision disease phenotypes.

A subspecies specific indel of particular interest was identified in the coding region of P3H2 (Fig. 19), causing a frameshift mutation in a gene associated with cataract formation. This indel is specific to the mountain gorilla population. There have been reports of gorillas in captivity developing cataracts as juveniles or loss of vision at a younger age than expected. This early cataract formation is believed to be due to a genetic predisposition as both gorillas had the same father (de Faber et al. 2004). Additional indels were found in MERTK (Fig. 19), which also associates with vision degradation due to cataracts.

5.4 Mutation frequency in gorilla populations

An important characteristic present in declining populations is the decrease of genetic diversity and therefore increased frequency of mutations shared within the population. We report the number of unique and overlapping copy number and indel mutations in each of the populations studied here. Due to the sample sizes of eastern lowland gorilla and mountain gorilla being unequal, we downsampled the mountain gorilla population to five individuals through all possible iterations for this comparison.

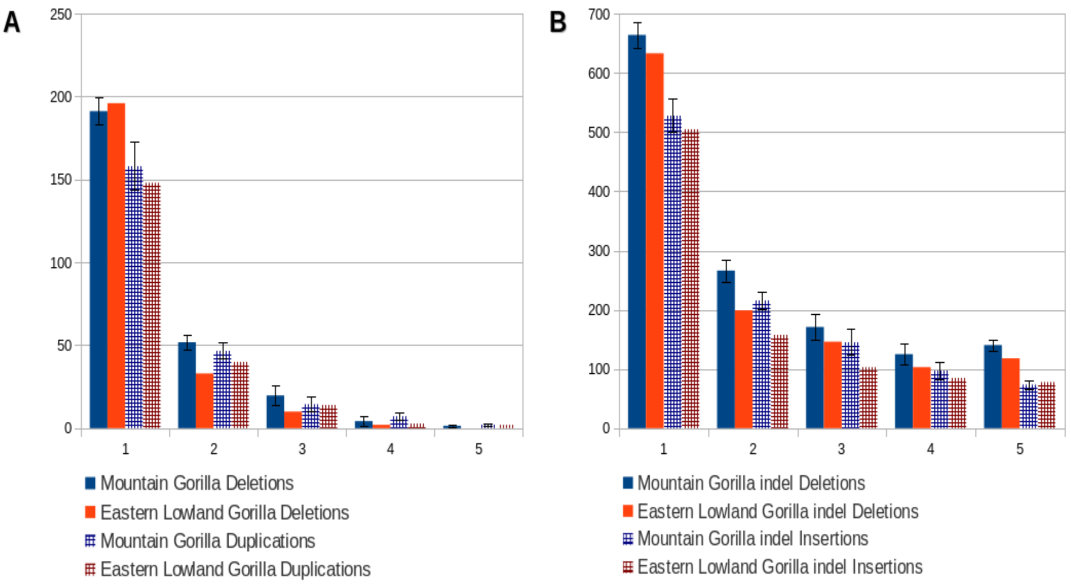


Figure 20. The frequency of duplications, deletions and indels that cross genes in the sampled gorilla populations. Bars represent standard error of all iterations of down sampling to equal population sizes. X-axis is the number of individuals sharing a mutation. Y-axis is the count of mutations. A) Deletions and duplications that cross gene exons. B) Indel deletion and insertions within exons.

We found that the number of unique deletions and duplications present in the exons of genes is equal in the two populations studied. However, there are significant increases in the number of shared mutations in the smaller mountain gorilla population compared to the larger eastern lowland population for both types of CNV (Figure 20). This is to be expected however as it has been previously reported that the mountain gorilla population has gone through a significant reduction and has a high rate of inbreeding (Xue et al. 2015). A similar pattern can be observed in indels as well, where the mountain gorilla population is harboring an increased number of deletion indel mutations across every category compared to that of the eastern lowland population. Nearly the same is true for indel insertions, except for mutations that are shared in every sampled individual.

In our extensive SV analysis of genomes from the populations of eastern lowland gorillas and mountain gorillas, we found an enrichment of phenotypes that associate with identifiable subspecies-specific physical characteristics. Additionally, several individual SVs were identified that are reportedly linked with disease states that are oft reported in various gorilla populations. As evidenced in other populations in decline or of limited breeding diversity, we also found there are an increasing number of SVs and indels that are shared among the smaller mountain gorilla population compared to that of the larger eastern lowland gorilla.

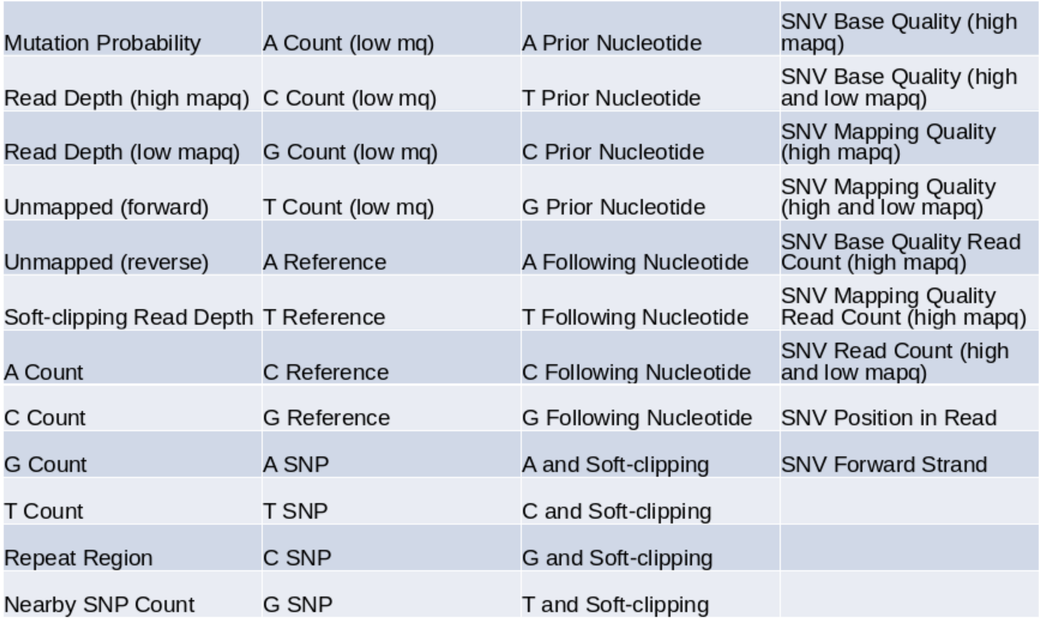
Chapter 6: ARIADNA: Machine Learning Methods for Ancient DNA variant discovery

In the absence of validated datasets, ancient DNA (aDNA) studies often rely on standard methods of mutation calling, optimized for high-quality contemporary DNA not exposed to excessive contamination, time, or environmental damage. Despite showing extreme sensitivity to aDNA quality, these methods have been used in many published studies, sometimes with additions of arbitrary filters or modifications that attempt to overcome aDNA degradation and contamination problems. The general lack of best practices for aDNA mutation calling may lead to inaccurate results. To resolve these problems, we present ARIADNA (ARtificial Intelligence for Ancient DNA), a novel approach based on machine learning techniques, using specific aDNA characteristics as features to yield improved mutation calls. In our comparisons of variant callers across several ancient genomes ARIADNA consistently detected higher-quality genome variants with fast runtimes, while reducing the false positive rate compared to other approaches.

As a first author of this work, I developed and designed the ARIADNA algorithm myself, with help from S. Smith in modifying GROM for this specific purpose. The application and testing of performance in woolly mammoth and Altai neandertal genomes was also performed by myself, with help from S. Smith and A. Grigoriev in experimental design and review. I also wrote the paper and generated all figures, with help in review and revisions from S. Smith and A. Grigoriev.

6.1 Variant discovery in ancient genomes

We developed methodology specifically designed to work with ancient DNA samples and provide more accurate and comprehensive mutation calls than non-adapted methods. This method utilizes a machine learning algorithm for the classification of potential mutation sites from BAM files. The unique features of the corresponding sites are used by our algorithm to determine the difference between a mutation in aDNA compared to noise due to aDNA degradation or contamination. We provide a comparison of our ARIADNA results in woolly mammoth with the most commonly employed mutation caller, GATK, as well as the Altai Neanderthal genome with output from Prüfer et al. in 2014 and 2017.

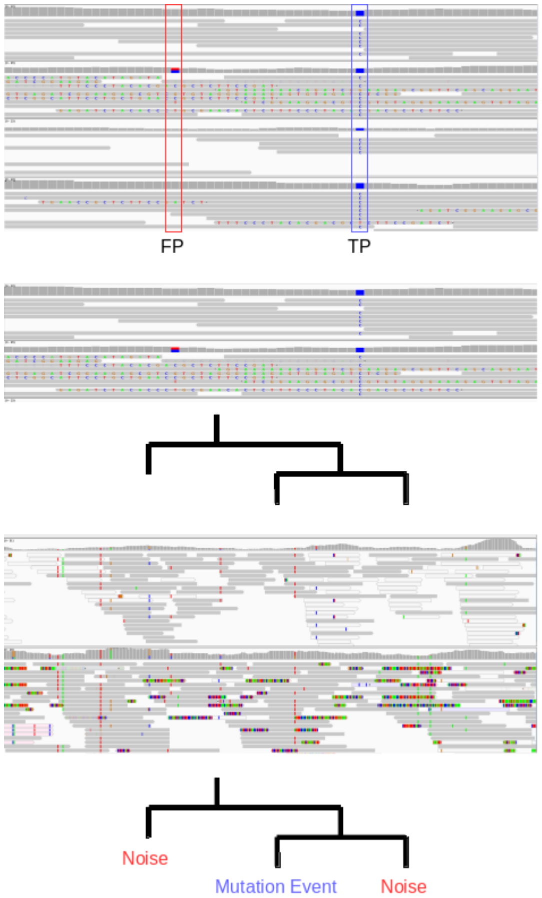


Mutation Probability	A Count (low mq)	A Prior Nucleotide	SNV Base Quality (high mapq)
Read Depth (high mapq)	C Count (low mq)	T Prior Nucleotide	SNV Base Quality (high and low mapq)
Read Depth (low mapq)	G Count (low mq)	C Prior Nucleotide	SNV Mapping Quality (high mapq)
Unmapped (forward)	T Count (low mq)	G Prior Nucleotide	SNV Mapping Quality (high and low mapq)
Unmapped (reverse)	A Reference	A Following Nucleotide	SNV Base Quality Read Count (high mapq)
Soft-clipping Read Depth	T Reference	T Following Nucleotide	SNV Mapping Quality Read Count (high mapq)
A Count	C Reference	C Following Nucleotide	SNV Read Count (high and low mapq)
C Count	G Reference	G Following Nucleotide	SNV Position in Read
G Count	A SNP	A and Soft-clipping	SNV Forward Strand
T Count	T SNP	C and Soft-clipping	
Repeat Region	C SNP	G and Soft-clipping	
Nearby SNP Count	G SNP	T and Soft-clipping	

Table 15. Features provided by GROM at each potential mutation site for use in the ML classification algorithm.

6.2 Performance of ARIADNA in woolly mammoth

We used a modified version of GROM (Smith et al. 2017) to scan the mammoth genomes for any evidence of difference with African elephant the reference genome. This yielded an average of 140bp/PSNV locations, between 18 million and 23 million locations per genome. Of these, 15 million PSNVs shared some evidence in all woolly mammoth genomes, and 6.6 million sites were unique to single woolly mammoth genomes (the remaining PSNVs were shared between 2-3 mammoths and not used for training). Because validation of these unique mutation sites is difficult using NGS of aDNA, our approach of capturing all deviation from reference in PSNVs would intercept a greater proportion of noise for use as FP events in the training set. Although it is certain that some true mutations were picked up in the false positive set, we reasoned that the effects of this mis-classification of events are diminished due to the high frequency of true FP events. Additionally, the validation of TPs across four genomes should alleviate problems with mis-classification during application of the trained model. Mutation events shared between either two or three of the woolly mammoth samples are ignored for the purpose of training to eliminate excessive uncertainty.



Classification of training data, true positive or false positive, using the four available mammoth genomes.

Construction of decision tree based on feature profiles of true positive and false positive mutations from the two mammoths used in training.

Feature collection of potential mutations for testing (pictured) or implementation.

Classification of potential mutations.

Figure 21. The design of the machine learning method for training and implementation.

We utilized two woolly mammoth genomes from separate studies out of four total mammoths available, for the purpose of training our ML model, a specimen from Wrangel Island and an M4 sample (one of the two noisy and potentially contaminated candidates). Two million shared and one million unique PSNVs from each of the two woolly mammoths in our training set were selected at random for training the ML model, resulting in four million training sites total. For the test set we used the two additional woolly mammoths from each study, Oimyakon and M25 (The second noisy woolly mammoth sample), and examined the results from the first largest contig (contig_0). The data from these two genomes is not used in any way as part of the training set in order to keep from over-fitting, or learning the unique characteristics of all available SNVs. In contig_0, the Oimyakon and M25 samples contained 799,849 and 960,816 PSNVs, respectively. Our algorithm utilized a feature set (Fig. 21) from the GROM genome scanner and the boosted regression tree ML module implemented using scikit learn (Pedregosa et al. 2011). This gave the ML portion of our algorithm 45 different features to utilize (Table 15). The parameters of the boosted regression trees algorithm in scikit learn were set to 200 trees in the construction of the classifier, and a learning rate of 0.01.

ARIADNA identified 607,354 and 599,847 mutations in contig_0 of the Oimyakon and M25 woolly mammoth samples respectively. Of these, 569,556 (Oimyakon) and 587,621 (M25) identified mutation sites are shared between all four woolly mammoths. An additional 32,050 (Oimyakon) and 87,130 (M25) mutation sites identified are shared in at least one other woolly mammoth. Only a small number of mutations identified in either woolly mammoth sample are unique, 5,748 (Oimyakon) and 12,226 (M25), making up 1.0% (Oimyakon) and 2.0% (M25) of all mutation calls.

To compare our results with commonly used methods, we also employed the GATK HaplotypeCaller (DePristo et al. 2011) on all available woolly mammoth genomes. Comparatively, GATK made more calls than ARIADNA methods. In the Oimyakon sample GATK made 825,955 calls, a 36% increase over our ML methods. In the M25 sample, 1,214,870 calls are made by GATK, a 102% increase over our ML methods. However, the most drastic increase is in the number of calls made by GATK that had no evidence of a variant in any other woolly mammoth. For Oimyakon this was 47,663 mutations and 280,049 mutations for M25. This comprised 5.8% and 23.1% of the calls GATK made in their respective genomes, a striking discrepancy, suggesting that

GATK over-predicted mutations at a very high rate in ancient genomes. This would be especially true in data sets with a high level of noise, and therefore a large number of false positive calls exist in methodology that utilizes this GATK algorithm. Identical over-prediction behavior of GATK has also been noted in a recent study on Neandertals by Prüfer et al. (2017), despite the comparatively high quality of NGS data in the Neandertal.

Another indication of over-prediction can be observed in the large difference in the counts of nucleotide change type between ARIADNA and GATK for the two mammoth genomes. ARIADNA call sets were robust; we observed very little change in counts or proportion of nucleotide substitution types in either Oimyakon or M25 mammoths (Fig. 22B, D). Conversely, GATK predicted large discrepancies (>2.5-fold) in such counts between the woolly mammoth genomes (Fig. 22D).

We found that in the Oimyakon samples, nearly 94% of the mutations identified using ARIADNA methods were shared in all four woolly mammoth samples, compared to 81% called by GATK. This was even more starkly contrasted in the noisier M25 dataset, where the ARIADNA algorithm made common calls with all woolly mammoths at a rate of 83%, while GATK called these same mutations at a rate of 55%. Calls shared between two or three woolly mammoths in Oimyakon made up 13% and 5% of the calls in GATK and ARIADNA respectively. In M25, these calls were made at a rate of 22% (GATK) and 15% (ARIADNA). More surprisingly, in this noisy dataset the use of GATK resulted in 23% of total calls being unique to M25. When analysed with GATK, there was a large increase in rate for unique calls from less than 6% to 23% in Oimyakon. The difference in the rate of unique calls between Oimyakon and M25 using ARIADNA methods was only 1.1%. Such robustness in predicted mutation rate strongly suggests that ARIADNA likely produced a more reliable call set in aDNA than that of GATK. The mutation rate between the two genomes using ARIADNA was 214bp/SNV in Oimyakon and 216bp/SNV in M25. The mutation rate provided by GATK was 157bp/SNV in Oimyakon and 107bp/SNV in M25.

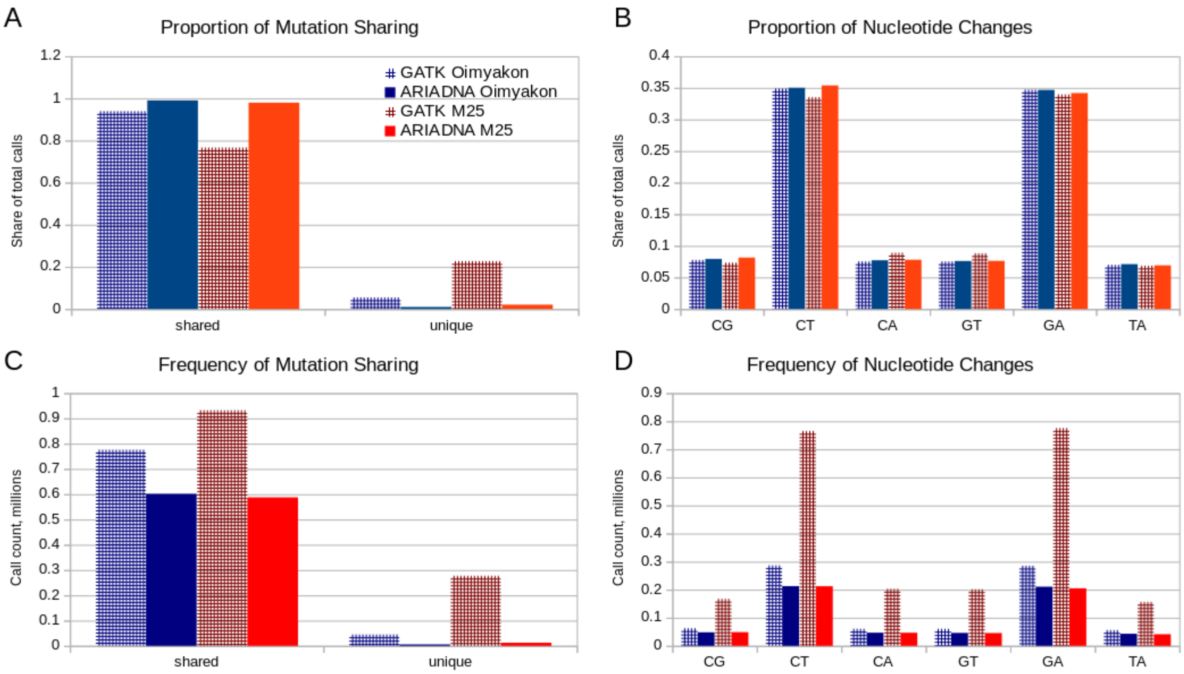


Figure 22. Performance of ARIADNA and GATK on the woolly mammoth genomes. Proportion of shared calls among all calls (A) and total numbers of shared calls (C) are plotted for ARIADNA (solid) and GATK (check pattern) on contig_0 of the woolly mammoth genomes. Spectra of nucleotide changes are also shown as proportions (B) and total numbers (D) are plotted for ARIADNA (solid) and GATK (check pattern) on contig_0 of the woolly mammoth genomes. Oimyakon sample is represented in blue and M25 in red.

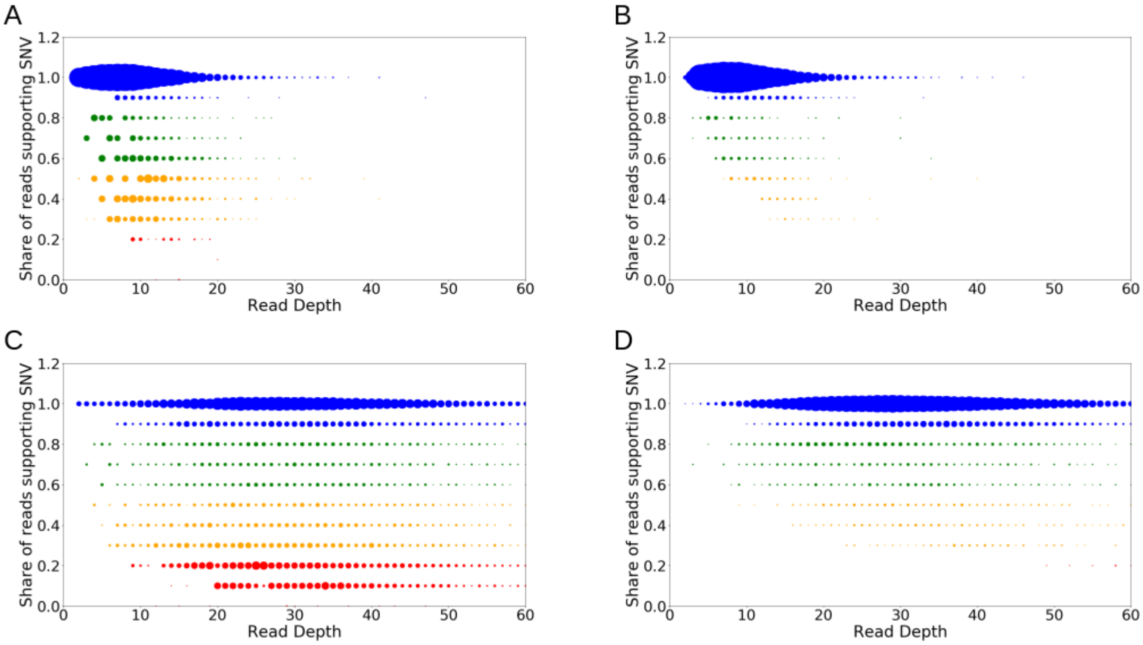


Figure 23. Improvements in calling variants with ARIADNA. Read depth and share of reads supporting SNV are shown for 20,000 randomly sampled calls from Oimyakon genome using GATK (A) and ARIADNA (B), and from M25 genome using GATK (C) and ARIADNA (D). The point size corresponds to the count of calls that were sampled at that coordinate on the plot.

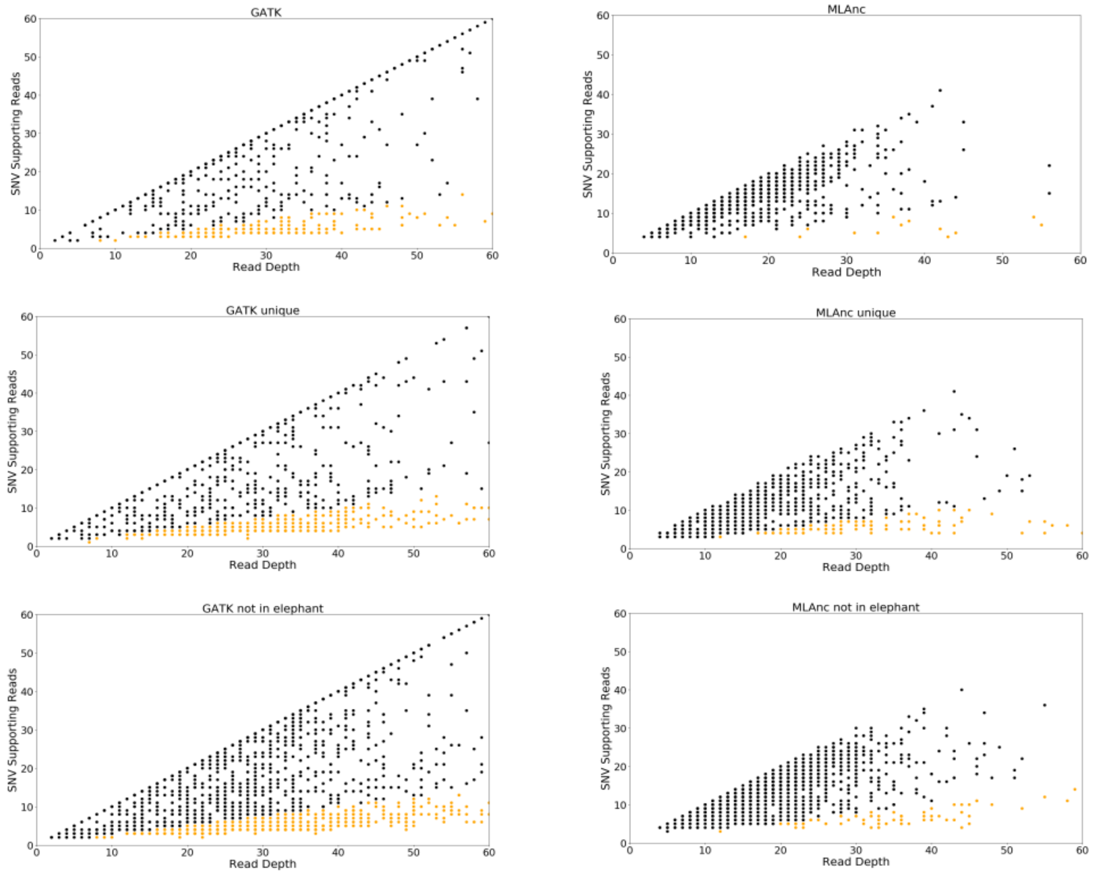


Figure 24. Scatter plots of 1,000 random calls as predicted from GATK (left) and ARIADNA (right) in M25. All calls (A and B), unique calls (C and D), calls that show no evidence in Asian elephant (E and F). Orange points have less than 25% of reads supporting the call.

We further tested the quality of calls produced by GATK and ARIADNA by comparing the distributions of reads supporting called SNVs. In the M25 woolly mammoth ARIADNA SNVs showed generally higher read support (Fig. 24A) vs. those of GATK (Fig. 24B). In contrast, there was little difference in the read support for SNVs in the Oimyakon sample (Fig 25). We observed an increased rate of calling for low supported reads in GATK that became more divergent from ARIADNA in unique calls

(Fig. 24 C,D), and calls that did not have evidence of variants in Asian elephant (Fig. 24 E,F). Thus, ARIADNA was able to filter out many heterozygous SNVs with biased read support in noisy datasets.

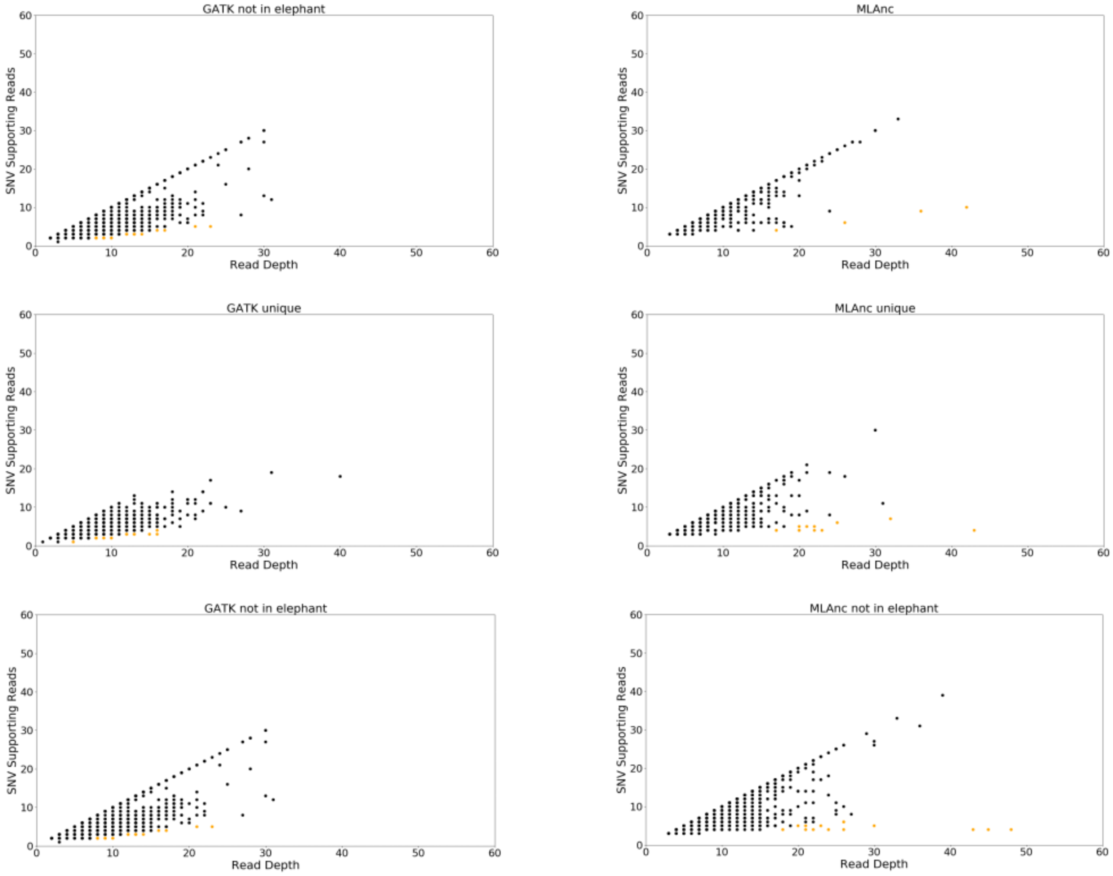
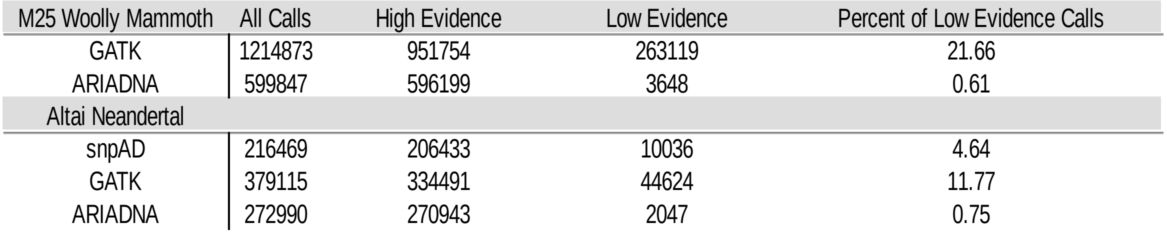


Figure 25. Scatter plots of 1,000 random calls as predicted from GATK (left) and ARIADNA (right) in Oimyakon. All calls (A and B), unique calls (C and D), calls that show no evidence in Asian elephant (E and F). Orange points have less than 25% of reads supporting the call.



M25 Woolly Mammoth	All Calls	High Evidence	Low Evidence	Percent of Low Evidence Calls
GATK	1214873	951754	263119	21.66
ARIADNA	599847	596199	3648	0.61
Altai Neandertal				
snpAD	216469	206433	10036	4.64
GATK	379115	334491	44624	11.77
ARIADNA	272990	270943	2047	0.75

Table 16. High and low evidence calls made by different algorithms.

Following an approach used to establish the high noise levels in the mammoth NGS data (Rogers et al. 2016), we tested the quality of calls produced by GATK and ARIADNA by comparing the share of reads supporting called SNVs in 20,000 randomly sampled calls of Oimyakon and M25 samples. In both cases GATK produced more calls with lower read support and from lower coverage regions than ARIADNA (Fig. 23). While this difference was small in the Oimyakon sample (Fig. 23A, B), it was quite substantial in the noisy M25 sample, where ARIADNA SNVs showed generally higher read support versus those of GATK (Fig. 23C, D). We found that in the M25 mammoth, nearly 22% of the calls made by GATK were of lower quality, i.e., had either <30% of reads supporting the call or originated from regions covered by <1/3 of the median number of reads (Table 16). In contrast, ARIADNA was able to filter out many heterozygous SNVs with biased read support in noisy aDNA data.

6.3 Performance of ARIADNA in Altai Neandertal

We then tested if our model could be applied to other genomes. We used the same ML decision tree that was constructed with the woolly mammoth training set to analyse the Altai Neandertal. As a first benchmark, we used the results of Prüfer at all (2014), who have compared GATK calls on Altai Neandertal to two 1,000 Genomes Project populations (European and East Asian)(Prüfer et al. 2017). As a second benchmark, we used the call set produced by Prüfer et al. using snpAD (Neandertal 2017).

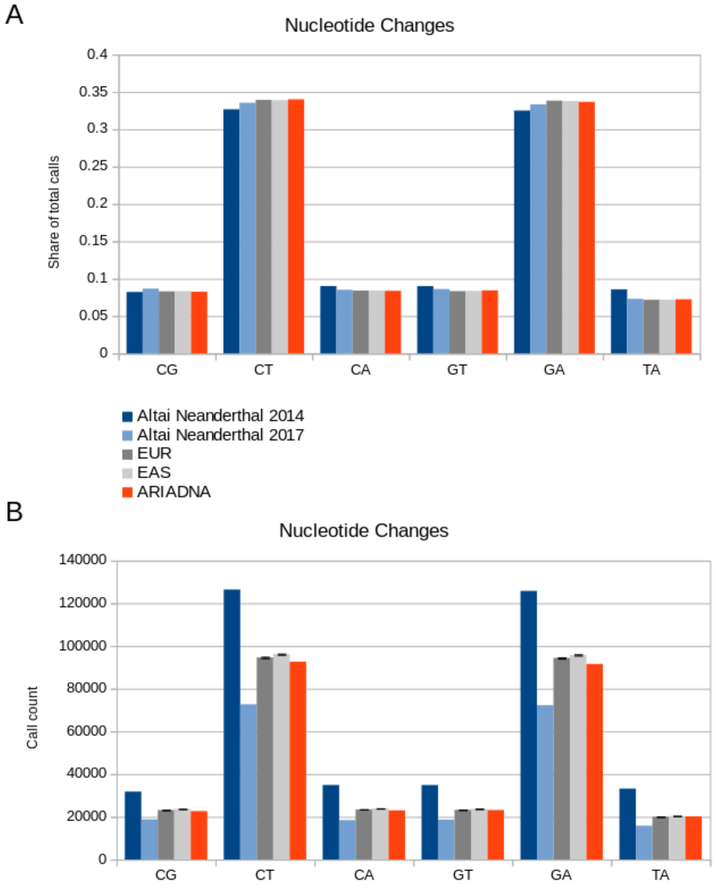


Figure 26. Performance of GATK, snpAD and ARIADNA on the Altai Neandertal genome (chromosome 1). Spectra of nucleotide changes (proportions-A, total numbers-B) in variant call sets are plotted for GATK (dark blue) Prüfer 2014, snpAD (light blue) Prüfer 2017, European individuals (dark grey), East Asian individuals (light grey) and ARIADNA (red).

Compared to these benchmarks of 379,115 GATK calls and 216,469 snpAD calls, ARIADNA made 272,990 calls, much closer to the number of mutations found in the two modern populations, 279,007 (European) and 283,776 (East Asian). This observation for all nucleotide changes also held true for each nucleotide change type, where ARIADNA consistently produced call sets that were most similar to the European and East Asian populations (Fig 26 B)

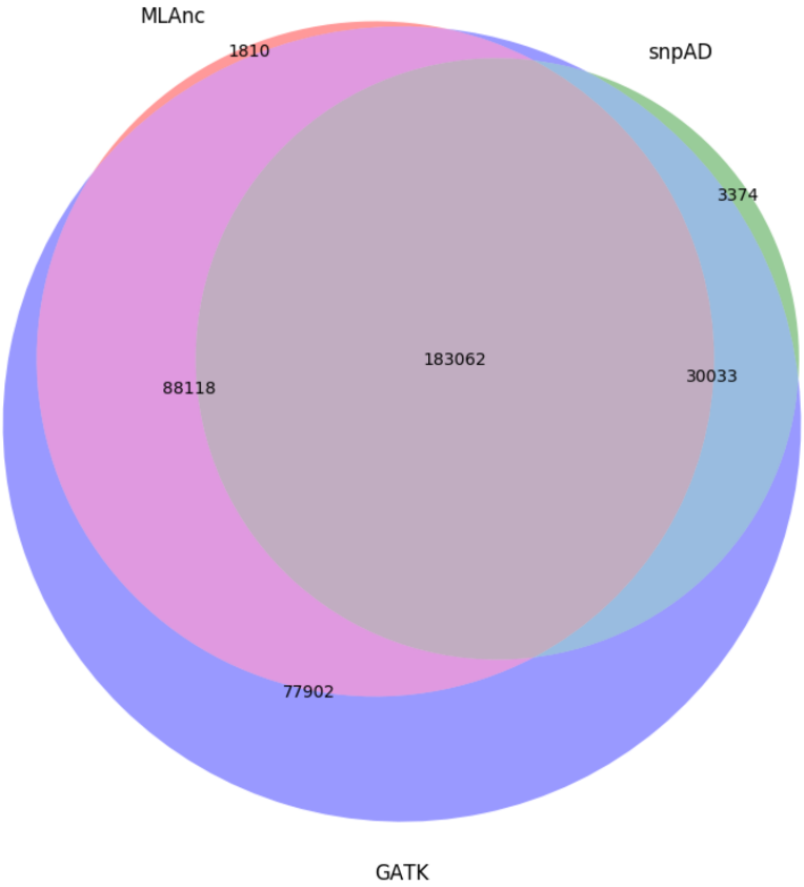


Figure 27. Overlap between calls made by GATK, snpAD, and ARIADNA on chromosome 1 of the Altai neandertal.

Our results are consistent with the earlier observations hat GATK, being sensitive to aDNA noise and degradation, tends to over-predict the aDNA mutations, producing more variants than any other methodology (Prüfer et al. 2017). On the other hand, the approach utilizing snpAD (Prüfer et al. 2017), seems to overcompensate in stringency and therefore to under-predict SNVs, reducing the amount of variation to less than what is otherwise found in the modern human population. In contrast, ARIADNA reported nucleotide substitution frequencies that are most similar to the modern human variation (Fig 26B).

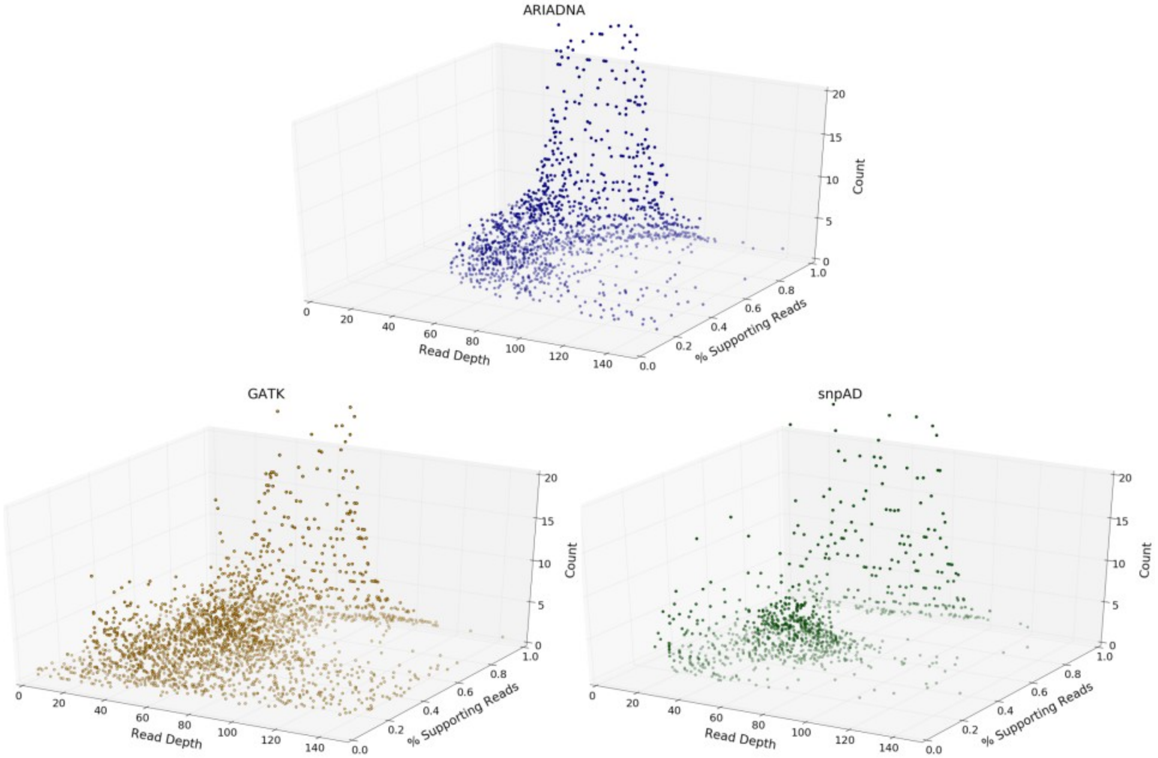
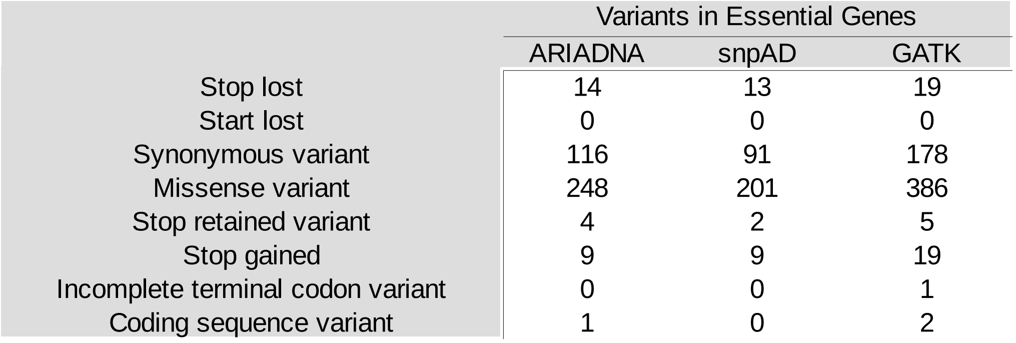


Figure 28. Scatter plots of 1,000 randomly selected calls that are made from GATK, snpAD, and ARIADNA on chromosome 1 of the Altai neandertal.

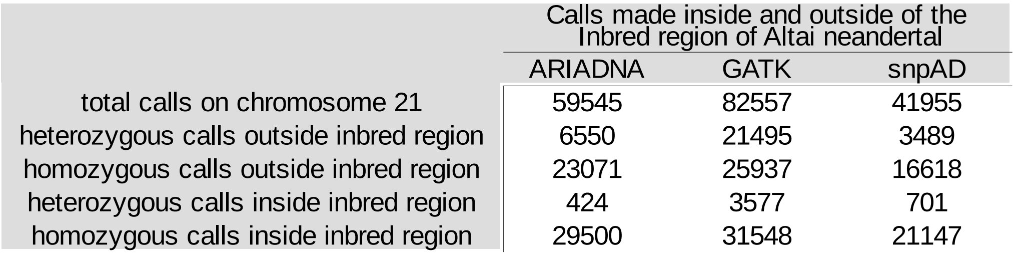
To further compare calls made by ARIADNA, snpAD, and GATK, we analysed the mutations of the Altai neandertal genome in essential genes identified by OGEE (Chen et al. 2011) and catalogued potential affects according to VEP (McLaren et al. 2016) (Table 17). GATK made the greatest number of calls within coding regions of essential genes, 610, including the greatest number of missense mutations, 386, stop loss, 19, and stop gains, 19. Comparatively, ARIADNA made fewer calls in essential genes, 392, as well as fewer missense mutations, 248. snpAD made the fewest calls in essential genes, 316, and fewer missense calls than ARIADNA, 201. ARIADNA made one more stop loss call than snpAD, 14, and both ARIADNA and snpAD made 9 stop gain calls. Enumerating all potentially deleterious mutations (missense, stop gain, start lost, and stop lost) in essential genes, GATK makes the greatest number of these calls, 424, while ARIADNA makes 271 and snpAD makes 223.



	Variants in Essential Genes		
	ARIADNA	snpAD	GATK
Stop lost	14	13	19
Start lost	0	0	0
Synonymous variant	116	91	178
Missense variant	248	201	386
Stop retained variant	4	2	5
Stop gained	9	9	19
Incomplete terminal codon variant	0	0	1
Coding sequence variant	1	0	2

Table 17. Effects of nucleotide changes in essential genes of Altai neandertal.

Prüfer et al. 2017 asserted that the inaccuracy of GATK is exemplified in chromosome 21 of the Altai neandertal, where there is a large inbred region from 17081807bp to 35881807bp. Here it was shown that GATK made an excessive number of heterozygous calls within the inbred region compared to snpAD. We compared the output of all three callers within the same region, using the provided vcf from Prüfer using GATK and snpAD (Table 18). As expected we found GATK made a far greater number of heterozygous calls than ARIADNA or snpAD both in total number, and proportion. However, ARIADNA made fewer heterozygous calls than snpAD, both in total number and proportionally, in this same inbred region, despite making a greater number of calls than snpAD overall. Additionally, outside of this inbred region, ARIADNA identified a greater number of heterozygous calls, both in total count and proportionally, than snpAD.



	Calls made inside and outside of the Inbred region of Altai neandertal		
	ARIADNA	GATK	snpAD
total calls on chromosome 21	59545	82557	41955
heterozygous calls outside inbred region	6550	21495	3489
homozygous calls outside inbred region	23071	25937	16618
heterozygous calls inside inbred region	424	3577	701
homozygous calls inside inbred region	29500	31548	21147

Table 18. Homozygous and heterozygous calls made within and outside of an inbred region of chromosome 21 of the Altai neandertal.

GATK made the greatest number of SNV calls not made in any of the other algorithms tested here, 77,902, far ahead of snpAD, 3,374, or ARIADNA, 1,810 (Fig 27). This was an expected result due to the reports of GATK being excessively sensitive, including noisy calls in its predictions (Prüfer et al. 2017). Neither ARIADNA nor snpAD made any common calls that were not identified by GATK. In order to identify trends of the predicted SNPs for each caller, we sub-sampled 1,000 calls from each set (Fig 28). Similar to the behavior in the woolly mammoth datasets, GATK had the highest proportion of calls made with lower coverage and lower read support, nearly 12% (Table 16), further indicating its sensitivity to aDNA noise, degradation, or contamination. Surprisingly, despite the least number of total calls made by snpAD, 4.6% also had low read coverage and low numbers of supporting reads, while ARIADNA produced the least number of such low quality calls, 0.8% (Table 16). ARIADNA made most of its unique calls with a moderate amount of read support, suggesting that it is better at the identification of heterozygous calls than snpAD, without the inclusion of false positives of GATK.

A combination of GROM and ARIADNA is also much faster than GATK. In a direct comparison (Smith et al. 2017), GROM was 12-25 times faster than GATK on a single thread and more than 70 times faster on 24 threads. Using the output from GROM, ARIADNA classifier run took between 5.5 mins (Oimyakon genome) and 14.5 mins (Neandertal genome) on a single thread, compared to >60 hours of GATK runtime. A one-time ARIADNA training run took 4 hours to generate a model.

ARIADNA utilizes a comprehensive feature set, incorporating several features that are often used in *ad-hoc* methods (Table 15). This includes identifying the position of the SNV within reads, base quality and mapping quality, nucleotide mismatch counts, and nucleotide change. We have also included novel features, such as accounting for nearby SNVs, adjacent nucleotides, and repeat regions, to better define difficult mapping regions or mutation hot-spots. The incorporation of several features as well as a decision tree ML model allows a dynamic level of filtering to compensate for changing NGS quality and read availability that is difficult to do with more static algorithms.

In summary, ARIADNA yielded consistent proportions of shared and unique mutations in the two woolly mammoth data sets compared to GATK. The frequency of nucleotide substitutions was also more stable using ARIADNA on the two woolly mammoth genomes than that of GATK. Utilizing modern human variation from 1,000 Genomes Project to compare results in the Altai Neandertal, we also found that the SNV calls made by ARIADNA were more consistent and potentially more relevant than of either GATK or snpAD. Our testing suggests that ARIADNA approach is superior for variant detection in ancient samples and has the capability to build models that can be utilized across a range of species.

Chapter 7: Discussion & Future Directions

The elucidation of genetic causes for disease, phenotypic variation, and evolutionary adaptation have been made increasingly possible on an ever increasing scale through the advancement of NGS technologies and associated analysis methodologies. Genome sequencing has become an invaluable tool in the broader study of biology, and with it the amount of data is outpacing researchers' ability to thoroughly analyze these data sets in sufficient detail.

Therefore we have developed tools for the researcher to rely on, increasing the accuracy and efficiency with which NGS can be analyzed and interpreted. This includes an algorithm for CNV detection utilizing read depth methods and a second high-throughput comprehensive algorithm that is able to identify all variant types with astounding efficiency while maintaining a high level of accuracy and sensitivity.

We have demonstrated the utility of these algorithms through applying them in two different comparative studies. The first, a comparison of extinct woolly mammoth genomes to that of the modern elephant. We elucidated previously unfound variation in genes responsible for adaptation in the woolly mammoth's unique environment. Furthermore we applied our algorithms to two populations of threatened gorilla subspecies. While previous studies focused on only a small fraction of the available mutation types, in this case SNPs, we found that by incorporating all types of genetic variants, a broader picture can be painted that more comprehensively points to genetic causes of disease, speciation, and possible genetic decline in these gorillas.

Work in both the development of variant calling methods as well as ancient genomes has introduced me to a specific niche of genome analysis with a unique set of problems. With the increasing prevalence of genome sequencing in contemporary organisms, there has also been in influx of comparative studies of ancient and modern genomes. However the analysis methods of these extinct species still lags behind that of its modern counterparts. The analysis of these ancient genomes is no less important, yet despite the additional hurdles that are infrequently accounted for in current methods, little has been done to alleviate these problems. These studies must deal with a comparatively unprecedented levels of degradation and contamination that leaves variant calling a struggle. We were able to apply out methodology of efficient and accurate variant identification to these data sets, and with additional analysis through machine learning, develop an algorithm specifically designed for variant identification in ancient genomes. Because out methodology is constructed specifically to deal with ancient genomes and its associated features and challenges, we find it to be more accurate and consistent in calling variants compared to any available methods.

7.1 Read depth methods of CNV detection

GROM-RD is a novel approach to detect CNVs in NGS data. Our methods of bias correction are more effective than previously developed algorithms. GROM-RD produces fewer false positives calls while maintaining a high level of sensitivity in both high and low coverage data sets. In addition our method of sliding window detection improved break point accuracy compared to other available methods.

Although RD analysis is a specific methodology over the broader scope of genome variant detection, it is still an important tool in variant discovery that is potentially frequently overlooked. CNV detection plays an important role in the explanation for phenotypic variation and evolution, accounting for roughly 4 million base pairs of variation between individuals (Mills et al 2011). CNV are potentially linked with varied gene regulation and expression levels between individuals (Stranger et al. 2007, McCarroll et al. 2006, Kleinjan et al. 2005, Somerville at al. 2005). Differences can arise from duplications or deletion of entire gene and regulatory regions. Differences have been accountable for several disease susceptibilities including; HIV (Gonzalex et al. 2005) , cancer (Berger et al. 2022, Stephens et al, 2009), and neurological disorders (Stefansson et al. 2009, Marshall et al. 2008) and evolutionary adaptation (Perry et al. 2007). Neurological disorders in particular have been linked to CNV in entire genes, potentially altering gene expression, while SNVs in these same genes do not sufficiently explain phenotypic variance (Bucan et al. 2009, Glessner et al. 2009, Nord et al 2011). RD methods are particularly useful in detecting CNVs with ambiguous breakpoints or low coverage where paired read and split read methods frequently struggle.

Our RD method provided improve sensitivity and specificity as well as break point accuracy, while being relatively tolerant of genome wide coverage levels. It has proven to be a valuable tool for comprehensive analysis of NGS data in both our woolly mammoth and eastern gorilla comparative studies.

7.2 Comprehensive variant detection

GROM was developed as a single, efficient, and comprehensive mutation calling algorithm to address the modern needs of NGS analysis. As NGS studies become increasingly popular, and NGS throughput continues to advance, bioinformatic tools are needed to keep up with demands. Previous methodology for complete genome analysis required the use of multiple tools to identify all possible genome variants. This is an inefficient use of computational resources and burdensome for the researchers. Our algorithm identifies SNVs, SVs, indels, and CNVs in a single run. Utilizing multiple detection methods including mismatch, split read, read pair, and read depth evidence also allows GROM to perform detection of all variant types with superior sensitivity and specificity in addition to the incredible reduction of time cost that would otherwise be associated with such a comprehensive NGS analysis.

GROM is an essential tool to match pace up with modern NGS studies. The data utilized in projects is continuously increasing. No longer is it sufficient to analyze small sections of the genome or single mutation type analysis for the identification of causative mutations. It is increasingly evident that phenotypic variation and disease states cannot be explained by single gene mutations or even complex interactions of SNPs alone (Weischenfeldt et al 2013, Visscher 2008). This is exemplified by Visscher, where over 50 nucleotide variants were found to influence human height, yet these variants only explained roughly 5% of the variance. To address this, many population studies have expanded their analysis to include CNVs and SVs (McCarroll et al. 2008, Kidd et al. 2008, Sudmant et al. 2015). As SVs account for a greater amount of variation within a genome than single nucleotide changes (Weischenfeldt et al 2013) they can potentially account for a large amount of phenotypic difference that is otherwise currently unexplained in previous studies.

Additionally, comparative and population level studies are greatly expanding, covering no longer tens or hundreds, but thousands and now millions of genomes. Current research has expanded beyond the 1,000 Genomes Project, as the UK looks to the 100,000 genomes project, the US Million Veterans Genome Program, and China's own Million Genomes project.

Algorithms such as GROM are requisite for the efficient collection of accurate variant information for interpretation. We have begun to prove the utility of this method through various collaborative studies that analyze a large number of genome samples, one such collaboration is the analysis of 3,000 rice genomes with the International Rice Research Institute.

7.3 Woolly mammoth evolutionary adaptation

Our first utilization of GROM for genome analysis was a comparative study between four woolly mammoth genomes and four Asian elephants. Previous studies had been performed on these data sets, but were limited in their analysis to SNVs. The application of our algorithm led to a much more comprehensive survey of potential causative genome variation for woolly mammoth adaptation to its arctic habitat. We identified several variants within the genome that were not previously discovered through SNV analysis alone. Variation was found in genes associated with lipid metabolism and thermogenesis, two important changes for life in the colder climates that woolly mammoth faced. Also observed were large changes in immunity related variants, and a large copy number amplification of RNase L and TP53. We also found several circadian oscillator genes that may have been selected for as mammoth high latitude necessitated adaptation to extremely varied day/night cycles throughout the course of the year.

7.4 Sub-species genetic health in eastern gorilla populations

The second comparative analysis we performed with GROM was between two eastern gorilla subspecies. Eastern lowland gorillas and especially mountain gorillas are facing significant population loss, and several disease phenotypes are being observed that put their genetic health into question. In addition, these two subspecies are closely related both in location and evolutionary distance. Our analysis revealed several variants that associate with phenotypes that differentiate the two subspecies that have not been previously accounted for in the genome. This includes mutations in genes that associate with body structure and size as well as facial structure. As expected with a a population facing such decline, we also found several mutations in genes that led to an enrichment for disease phenotypes. This included enrichment in genes associated with syndactyly, infertility, and vision impairment. Interestingly, our SV and indel analysis showed an equal number of unique mutations in the two populations. However the mountain gorilla harbored several more shared variants than that of the eastern lowland gorilla. This is further evidence of a declining breeding population than what can be found through associated disease phenotypes.

7.5 Variant discovery in ancient genomes

The usage of modern variant calling tools on woolly mammoth genomes gave first hand insight into the unique problems that are faced with variant calling in ancient genomes. These genomes have high frequencies of decomposition, fragmentation, environmentally induced substitutions, and contamination that is not frequently encountered in contemporary NGS studies. An increasing number of WGS is being performed on ancient genomes as DNA extraction and sequencing methods develop. Yet the difficulty in variant calling persists as the quality of the available DNA cannot be changed. At the time there was no available tool for variant identification in these ancient NGS samples, and therefore the tools for contemporary analysis had to be adapted. This proved exceedingly difficult as many of these tools took the inherent noise and fragmentation of the data to be interpreted as variation. Compounding this problem were research groups devising their own methods toward variant detection making validation of accuracy and reproducibility difficult. To this end we developed an SNV caller specifically designed to work with ancient genomes. To overcome these challenges we employed a machine learning method, boosted regression trees, to account for the variation and potential interactions of variant features more effectively and efficiently than the ad-hoc compensatory methods currently in use. The resulting algorithm proved to be much better suited for work in ancient genomes than even the best post-adopted methods on highly processed data. The mammoth data sets were specifically chosen for design and testing as there was a great deal of variation in the DNA sequence quality of the genomes. Previous methods faced a large discrepancy in call rate between these genomes as they were not able to effectively compensate for the inherent noise and degradation. GATK, the most commonly used mutation calling algorithm was especially susceptible to this variation in NGS data of ancient genomes, and made several SNP calls with rather limited supporting evidence. Our method that was employed in ARIADNA effectively eliminated this variation between genome call rates. Additionally, when tested in Neanderthal samples, ARIADNA provided a mutation call rate that most closely mirrored that of the variance found in the human population.

7.6 Future directions

All of the data sets that were used in these studies were publicly available, yet much was left on the table for analysis, discovery, and interpretation. Currently more NGS data is being produced than can be effectively analyzed. This leaves a huge swath of discovery in genome variation yet to be reported on. With our efficient means of variant discovery we expect that comparative analysis will become more comprehensive, elucidating genetic explanations for previously observed but unaccounted phenotypes.

The study of ancient genomes is especially exciting, as the previous hurdle to their study was the extraction and sequencing of DNA. With our newly developed ARIADNA mutation caller, we hope to continue our study of comparative genomics in ancient genomes and expand our variant call methods to lower coverage datasets.

Chapter 8: Materials & methods

8.1 Gorilla genome samples

Details of samples collected and sequencing are provided in Xue et al. Briefly, 12 gorillas (7 mountain and 5 eastern lowland) were selected for WGS. An average of 26x coverage was produced on an Illumina HiSeq 2000 sequencer using standard library preparation. We downloaded fastq files from SRA project PRJEB3220 provided in Xue et al. and mapped the data to the western lowland gorilla reference sequence gorGor3.1 (http://www.ensembl.org/Gorilla_gorilla) using BWA-MEM v0.7.4 (Li, 2013) with default parameters.

8.2 Genome variant detection in gorilla populations

Several methods are available to detect SVs. Read depth methods are useful for detection of copy number variants (CNVs), which comprise deletions and duplications. Paired-read and split-read methods are able to detect deletions and duplications, as well as insertions and inversions. This limitation of SV detection by a singular method necessitates utilizing a combination of approaches to increase sensitivity and breadth of analysis. We used a read depth method, GROM-RD (Smith 2015), and a paired-read, split-read integrated approach GROM (Smith et al. 2017). We identified all structural variants <1 MB in size for every individual. GROM settings included a minimum of 3 high quality (mapping quality ≥ 20 and for SNVs, base quality ≥ 20) reads (SNVs, indels) or read pairs (SVs) supporting the variation.

We analyzed complete lineage-specific variation, wherein a mutation was identified in every individual of a gorilla subspecies and no support of the variant was present in any individual of the reciprocal subspecies. Also identified were non-ubiquitous lineage-specific variants, where any members of the subspecies carried the mutation, with no evidence supporting the variant in any individual of the other gorilla subspecies. A reciprocal overlap of >50% of variant length was required for common variants. We focused specifically on affected gorilla genes and thus only considered variants overlapping gene regions.

8.3 Mutation Enrichment in gorilla genomes

Phenotypic enrichment analysis was performed using g:Profiler (Reimand et al. 2007) with default parameters unless noted below. Sets for both complete lineage-specific variation and non-ubiquitous lineage-specific variation were entered into g:Cocoa (Reimand et al. 2016) for comparison of enrichment analysis between the two subspecies. Our statistical domain size was limited to known genes for gorilla and enrichment profiles from Human Phenotype Ontology, as we were interested in the observable effects these mutations would potentially have.

8.4 ARIADNA algorithm development

The backbone of our method consists of a machine learning algorithm tailored for aDNA mutation calling by utilizing unique features found in aDNA samples. Specifically we implement the use of boosted regression tree models for our methodology (Friedman 2001). In summary, boosted regression trees works through building a succession of additive decision trees to best classify known data (training set). The algorithm assigns thresholds within the trees using the characteristics of given true positive (TP) and false positive (FP) calls from the training set. Through a series of these trees, prediction deviance from known truth is continually reduced. Our feature set is provided by modification of the comprehensive mutation caller GROM (Smith et al. 2017), to act as a genome scanner and output feature information at potential mutation locations. These features include common measures such as read depth, SNV base count, read and base quality as well as features that are unique to aDNA, such as distance from read end, C to T substitutions and neighbouring mutation rates (Table 15). Once this series of trees is built from the training data, a model is constructed and classification of further data (known, for testing, or unknown, for implementation) can take place (Figure 21). Here, a hold back set of known mutations is used to test performance. For our purposes, these known values not used in training become the testing set.

8.5 ARIADNA woolly mammoth genomes

We tested our machine learning method on four woolly mammoth samples M4, M25, Wrangel Island and Oimyakon (Lynch et al. 2015, Palkopoulou et al. 2015). These samples originated from two different studies, with the former being suspected of experiencing high levels of problems associated with aDNA sequencing (Rogers et al. 2016). WGS fastq files for woolly mammoths M4 and M25 were downloaded from the Sequence Read Archive (SRA), http://www.ncbi.nlm.nih.gov/sra (project accession number: PRJNA281811). WGS fastq files for the Wrangel and Oimyakon woolly mammoths were downloaded from the European Nucleotide Archive (ENA), http://www.ebi.ac.uk/ena (accession number: ERP008929). WGS fastq files were mapped to the African Elephant reference genome loxAfr3, downloaded from UCSC (https://genome.ucsc.edu, http://hgdownload.soe.ucsc.edu/goldenPath/loxAfr3/bigZips/), using BWA MEM, version 0.7.4, with default parameters. Duplicates were removed using SAMtools (Li et al. 2009), version 0.1.19. We limited analysis to supercontigs/scaffolds ≥1,000,000 bases. PSNVs were detected using GROM (Smith et al. 2017), customized to include output of additional features from Table 15.

WGS fastq files Asian Elephants Asha, Parvathy, and Uno were downloaded from the Sequence Read Archive (SRA), http://www.ncbi.nlm.nih.gov/sra (project accession number: PRJNA281811). WGS fastq files for the Asian elephant Emelia were downloaded from ENA (accession: ERP004241). WGS fastq files were mapped to the African reference genome loxAfr3 and aligned in the same way as the woolly mammoth.

8.6 ARIADNA woolly mammoth variant identification and construction of training and testing sets

For the development of our training and testing sets, potential mutation locations shared between all woolly mammoth genomes are considered TP locations (we did not take zygosity into account when designing our data sets). Conversely, potential mutation locations that only occurred in a single individual are deemed FP. The shared potential mutation events that occurred in all woolly mammoth samples served as validation that the mutation did not occur as a result of contamination, degradation, or sequencing artefact. Additionally, the use of woolly mammoth training samples for this study (as compared to Neandertal or ancient human) reduced the risk of misrepresented calls due to either misalignment or contamination of closely related samples (Wall et al. 2007, Green 2009).

We used a modified version of GROM to scan the mammoth genomes for any evidence of difference with the reference genome. This yielded an average of 140bp/PSNV locations, between 18 million and 23 million locations per genome. Of these, 15 million PSNVs shared some evidence in all woolly mammoth genomes, and 6.6 million sites were unique to single woolly mammoth genomes (the remaining PSNVs were shared between 2-3 mammoths and not used for training). Because validation of these unique mutation sites is difficult using NGS of aDNA, our approach of capturing all deviation from reference in PSNVs would intercept a greater proportion of noise for use as FP events in the training set. Although it is certain that some true mutations were picked up in the false positive set, the effects of this mis-classification of events are diminished due to the high frequency of true FP events. Additionally, the validation of TPs across four genomes should alleviate problems with mis-classification during application of the trained model. Mutation events shared between either two or three of the woolly mammoth samples are ignored for the purpose of training to eliminate excessive uncertainty.

We utilized two woolly mammoth genomes from separate studies for the purpose of training our ML model, a specimen from Wrangel Island and an M4 sample (a noisy and potentially contaminated candidate). One million shared and one million unique PSNVs from each of the two woolly mammoths in our training set were selected at random for training the ML model, resulting in four million training sites total. For the test set we used the two additional woolly mammoths from each study, Oimyakon and M25, and examined the results from the first largest contig (contig_0). The data from these two genomes is not used in any way as part of the training set in order to keep from over-fitting, or learning the unique characteristics of all available SNVs. In contig_0, the Oimyakon and M25 samples contained 799,849 and 960,816 PSNVs, respectively. Our algorithm utilized a feature set (Fig. 11) from the GROM genome scanner and the boosted regression tree ML module implemented using scikit learn (Pedregosa et al. 2011). This gave the ML portion of our algorithm 45 different features to utilize (Table 15). The parameters of the boosted regression trees algorithm in scikit learn were set to 200 trees in the construction of the classifier, and a learning rate of 0.01.

8.7 ARIADNA Altai neandertal and contemporary genomes

Additional testing was performed on the Altai Neandertal chromosome 1 from Prüfer 2014, using BAM files hosted at Max Planck Institute for Evolutionary Anthropology (http://cdna.eva.mpg.de/neandertal/altai/). Calls and feature information were produced by GROM. The VCF files produced by GATK (Prüfer et al. 2014), and snpAD (Prüfer et al. 2017), from the respective Prüfer publications were used for comparison. These were downloaded from the hosted neandertal files at Max Planck Institute for Evolutionary Anthropology (for the 2014 dataset: http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/, for the 2017 dataset: http://cdna.eva.mpg.de/neandertal/Vindija/VCF/) To better observe mutation rates and nucleotide change frequencies, further comparisons were made from 20 random genomes of the 1,000 Genomes Project (1000 Genomes Project Consortium 2015); ten individuals from the European population group, and ten individuals from the East Asian population group. These two groups are believed to be the contemporary populations that are most related to the Neandertals. Here mutation information was provided through the 1,000 Genomes VCF (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/).

To identify variants affecting essential genes, a list of human essential genes was downloaded from the Online GEne Essentiality database (http://ogee.medgenius.info/browse/). Only genes listed as “essential” were used in the analysis. Variants in essential genes were uploaded to the Ensembl Variant Effect Predictor (https://www.ensembl.org/vep) to categorize impact. The inbred region of the neandertal genome analysed (chromosome 21:17081807-35881807) was identified by Prüfer et al. 2017 (30).

Chapter 9: Works Cited

1000 Genomes Project Consortium. "A global reference for human genetic variation."*Nature* 526, no. 7571 (2015): 68-74.

1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing."*Nature* 467, no. 7319 (2010): 1061-1073.

Abegglen, Lisa M., Aleah F. Caulin, Ashley Chan, Kristy Lee, Rosann Robinson, Michael S. Campbell, Wendy K. Kiso et al. "Potential mechanisms for cancer resistance in elephants and comparative cellular response to DNA damage in humans." *Jama* 314, no. 17 (2015): 1850-1860.

Abyzov, Alexej, Alexander E. Urban, Michael Snyder, and Mark Gerstein. "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing."*Genome Research* 21, no. 6 (2011): 974-984.

Aglipay, Jason A., Sam W. Lee, Shinya Okada, Nobuko Fujiuchi, Takao Ohtsuka, Jennifer C. Kwak, Yi Wang et al. "A member of the Pyrin family, IFI16, is a novel BRCA1-associated protein involved in the p53-mediated apoptosis pathway."*Oncogene* 22, no. 55 (2003): 8931-8938.

Aird, Daniel, Michael G. Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B. Jaffe, Chad Nusbaum, and Andreas Gnirke. "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries."*Genome Biology*12, no. 2 (2011): R18.

“All of Us." National Institutes of Health. Accessed September 30, 2017. https://allofus.nih.gov/.

Ameyar-Zazoua, Maya, Christophe Rachez, Mouloud Souidi, Philippe Robin, Lauriane Fritsch, Robert Young, Nadya Morozova et al. "Argonaute proteins couple chromatin silencing to alternative splicing." *Nature Structural & Molecular Biology* 19, no. 10 (2012): 998-1004.

Atchison, C. J., C. C. Tam, S. Hajat, W. Van Pelt, J. M. Cowden, and B. A. Lopman. "Temperature-dependent transmission of rotavirus in Great Britain and The Netherlands." *Proceedings of the Royal Society of London B: Biological Sciences* 277, no. 1683 (2010): 933-942.

Baker, Monya. "Structural variation: the genome's hidden architecture."*Nature Methods* 9, no. 2 (2012): 133-137.

Bartenhagen, Christoph, and Martin Dugas. "RSVSim: an R/Bioconductor package for the simulation of structural variations."*Bioinformatics* 29, no. 13 (2013): 1679-1681.

Bates, P. C., and A. T. Holder. "The anabolic actions of growth hormone and thyroxine on protein metabolism in Snell dwarf and normal mice." *Journal of Endocrinology* 119, no. 1 (1988): 31-41.

Beck, Benjamin B. "Fertility in North American male lowland gorillas." *American Journal of Primatology* 3, no. S1 (1982): 7-11.

Benjamini, Yuval, and Terence P. Speed. "Summarizing and correcting the GC content bias in high-throughput sequencing."*Nucleic Acids Research* 40, no. 10 (2012): e72-e72.

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall et al. "Accurate whole human genome sequencing using reversible terminator chemistry." *Nature* 456, no. 7218 (2008): 53-59.

Berger, Michael F., Michael S. Lawrence, Francesca Demichelis, Yotam Drier, Kristian Cibulskis, Andrey Y. Sivachenko, Andrea Sboner et al. "The genomic complexity of primary human prostate cancer." *Nature* 470, no. 7333 (2011): 214-220.

Blomen, Vincent A., Peter Májek, Lucas T. Jae, Johannes W. Bigenzahn, Joppe Nieuwenhuis, Jacqueline Staring, Roberto Sacco et al. "Gene essentiality and synthetic lethality in haploid human cells." *Science* 350, no. 6264 (2015): 1092-1096.

Boelen, Anita. "Thyroid hormones and glucose metabolism: the story begins before birth." *Experimental Physiology* 94, no. 10 (2009): 1050-1051.

Boeva, Valentina, Andrei Zinovyev, Kevin Bleakley, Jean-Philippe Vert, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. "Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization." *Bioinformatics* 27, no. 2 (2010): 268-269.

Brennan-Laun, Sarah E., Heather J. Ezelle, Xiao-Ling Li, and Bret A. Hassel. "RNase-L control of cellular mRNAs: roles in biologic functions and mechanisms of substrate targeting." *Journal of Interferon & Cytokine Research* 34, no. 4 (2014): 275-288.

Briggs, Adrian W., Udo Stenzel, Philip LF Johnson, Richard E. Green, Janet Kelso, Kay Prüfer, Matthias Meyer et al. "Patterns of damage in genomic DNA sequences from a Neandertal." *Proceedings of the National Academy of Sciences* 104, no. 37 (2007): 14616-14621.

Brotherton, Paul, Phillip Endicott, Juan J. Sanchez, Mark Beaumont, Ross Barnett, Jeremy Austin, and Alan Cooper. "Novel high-resolution characterization of ancient DNA reveals C> U-type base modification events as the sole cause of post mortem miscoding lesions." *Nucleic Acids Research* 35, no. 17 (2007): 5717-5728.

Bucan, Maja, Brett S. Abrahams, Kai Wang, Joseph T. Glessner, Edward I. Herman, Lisa I. Sonnenblick, Ana I. Alvarez Retuerto et al. "Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes." PLoS genetics 5, no. 6 (2009): e1000536.

Campbell, Kevin L., Jason EE Roberts, Laura N. Watson, Jörg Stetefeld, Angela M. Sloan, Anthony V. Signore, Jesse W. Howatt et al. "Substitutions in woolly mammoth hemoglobin confer biochemical properties adaptive for cold tolerance." *Nature Genetics* 42, no. 6 (2010): 536-540.

Campbell, Peter J., Shinichi Yachida, Laura J. Mudie, Philip J. Stephens, Erin D. Pleasance, Lucy A. Stebbings, Laura A. Morsberger et al. "The patterns and dynamics of genomic instability in metastatic pancreatic cancer." *Nature* 467, no. 7319 (2010): 1109-1113.

Chakrabarti, Arindam, Babal Kant Jha, and Robert H. Silverman. "New insights into the role of RNase L in innate immunity." *Journal of Interferon & Cytokine Research* 31, no. 1 (2011): 49-57.

Chakrabarti, Arindam, Shuvojit Banerjee, Luigi Franchi, Yueh-Ming Loo, Michael Gale, Gabriel Núñez, and Robert H. Silverman. "RNase L activates the NLRP3 inflammasome during viral infections." *Cell Host & Microbe* 17, no. 4 (2015): 466-477.

Chen, Ken, John W. Wallis, Michael D. McLellan, David E. Larson, Joelle M. Kalicki, Craig S. Pohl, Sean D. McGrath et al. "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation." *Nature Methods* 6, no. 9 (2009): 677-681.

Chen, Wei, Vera Kalscheuer, Andreas Tzschach, Corinna Menzel, Reinhard Ullmann, Marcel Holger Schulz, Fikret Erdogan et al. "Mapping translocation breakpoints by next-generation sequencing." *Genome Research* 18, no. 7 (2008): 1143-1149.

Chen, Wei-Hua, Pablo Minguez, Martin J. Lercher, and Peer Bork. "OGEE: an online gene essentiality database." *Nucleic acids research* 40, no. D1 (2011): D901-D906.

Chen, Xiaoyu, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. "Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications." *Bioinformatics* 32, no. 8 (2015): 1220-1222.

Chen, Zhe-Sheng, Yanping Guo, Martin G. Belinsky, Elena Kotova, and Gary D. Kruh. "Transport of bile acids, sulfated steroids, estradiol 17-β-D-glucuronide, and leukotriene C4 by human multidrug resistance protein 8 (ABCC11)." *Molecular Pharmacology* 67, no. 2 (2005): 545-557.

Chiang, Colby, Ryan M. Layer, Gregory G. Faust, Michael R. Lindberg, David B. Rose, Erik P. Garrison, Gabor T. Marth, Aaron R. Quinlan, and Ira M. Hall. "SpeedSeq: ultra-fast personal genome analysis and interpretation." *Nature Methods* 12, no. 10 (2015): 966-968.

Chiang, Derek Y., Gad Getz, David B. Jaffe, Michael JT O'kelly, Xiaojun Zhao, Scott L. Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S. Lander. "High-resolution mapping of copy-number alterations with massively parallel sequencing." *Nature Methods* 6, no. 1 (2009): 99-103.

Clark, Andrew G., Stephen Glanowski, Rasmus Nielsen, Paul D. Thomas, Anish Kejariwal, Melissa A. Todd, David M. Tanenbaum et al. "Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios." *Science* 302, no. 5652 (2003): 1960-1963.

Cooper, John E., and Gordon Hull. *Gorilla Pathology and Health: With a Catalogue of Preserved Materials*. Academic Press, 2017.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker et al. "The variant call format and VCFtools." *Bioinformatics* 27, no. 15 (2011): 2156-2158.

Dastjerdi, Akbar, Christelle Robert, and Mick Watson. "Low coverage sequencing of two Asian elephant (Elephas maximus) genomes." *GigaScience* 3, no. 1 (2014): 12.

Database of Genomic Variants. Accessed September 30, 2017. http://dgv.tcag.ca/dgv/app/home

“Database of Genotypes and Phenotypes." National Center for Biotechnology Information. Accessed April 24, 2017. https://www.ncbi.nlm.nih.gov/gap.

Daugherty, Matthew D., Janet M. Young, Julie A. Kerns, and Harmit S. Malik. "Rapid evolution of PARP genes suggests a broad role for ADP-ribosylation in host-virus conflicts." *PLoS Genetics* 10, no. 5 (2014): e1004403.

David Cyranoski, Nature. "Nature News Feature: China's bid to be a SNA Superpower." (2016).

de Faber, JT HN, J. H. Pameijer, and W. Schaftenaar. "Cataract surgery with foldable intraocular lens implants in captive lowland gorillas (Gorilla gorilla gorilla)." *Journal of Zoo and Wildlife Medicine* 35, no. 4 (2004): 520-524.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis et al. "A framework for variation discovery and genotyping using next-generation DNA sequencing data." *Nature genetics* 43, no. 5 (2011): 491-498.

DiTacchio, Luciano, Hiep D. Le, Christopher Vollmers, Megumi Hatori, Michael Witcher, Julie Secombe, and Satchidananda Panda. "Histone lysine demethylase JARID1a activates CLOCK-BMAL1 and influences the circadian clock." *Science* 333, no. 6051 (2011): 1881-1885.

Doyle, Kristian P., Katherine L. Suchland, Thomas MP Ciesielski, Nikola S. Lessov, David K. Grandy, Thomas S. Scanlan, and Mary P. Stenzel-Poore. "Novel thyroxine derivatives, thyronamine and 3-iodothyronamine, induce transient hypothermia and marked neuroprotection against stroke injury." *Stroke* 38, no. 9 (2007): 2569-2576.

Drouineaud, Véronique, Laurent Lagrost, Alexis Klein, Catherine Desrumaux, Naig Le Guern, Anne Athias, Franck Ménétrier et al. "Phospholipid transfer protein deficiency reduces sperm motility and impairs fertility of mouse males." *The FASEB Journal* 20, no. 6 (2006): 794-796.

Dubins, Jeffrey S., Manuel Sanchez-Alavez, Victor Zhukov, Alejandro Sanchez-Gonzalez, Gianluca Moroncini, Santos Carvajal-Gonzalez, John R. Hadcock, Tamas Bartfai, and Bruno Conti. "Downregulation of GPR83 in the hypothalamic preoptic area reduces core body temperature and elevates circulating levels of adiponectin." *Metabolism* 61, no. 10 (2012): 1486-1493.

Durbin, R. M., G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler et al. "An integrated map of genetic variation from 1,092 human genomes." (2012).

Ebbert, Mark TW, Mark E. Wadsworth, Lyndsay A. Staley, Kaitlyn L. Hoyt, Brandon Pickett, Justin Miller, John Duce, John SK Kauwe, Perry G. Ridge, and Alzheimer’s Disease Neuroimaging Initiative. "Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches." *BMC Bioinformatics* 17, no. 7 (2016): 239.

Eberle, Michael A., Epameinondas Fritzilas, Peter Krusche, Morten Källberg, Benjamin L. Moore, Mitchell A. Bekritsky, Zamin Iqbal et al. "A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree." *Genome Research* 27, no. 1 (2017): 157-164.

Edgar, Robert C. "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic Acids Research* 32, no. 5 (2004): 1792-1797.

Ezelle, Heather J., Krishnamurthy Malathi, and Bret A. Hassel. "The roles of RNase-L in antimicrobial immunity and the cytoskeleton-associated innate response." *International Journal of Molecular Sciences* 17, no. 1 (2016): 74.

Fabre, O., T. Salehzada, K. Lambert, Y. Boo Seok, A. Zhou, J. Mercier, and C. Bisbal. "RNase L controls terminal adipocyte differentiation, lipids storage and insulin sensitivity via CHOP10 mRNA regulation." *Cell Death & Differentiation* 19, no. 9 (2012): 1470-1481.

Fisher, Daniel C., Alexei N. Tikhonov, Pavel A. Kosintsev, Adam N. Rountrey, Bernard Buigues, and Johannes van der Plicht. "Anatomy, death, and preservation of a woolly mammoth (Mammuthus primigenius) calf, Yamal Peninsula, northwest Siberia." *Quaternary International* 255 (2012): 94-105.

Fossey, Dian. *Gorillas in the Mist*. Boston: Houghton Mifflin, 1983. 72.

Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* (2001): 1189-1232.

Fromer, Menachem, Jennifer L. Moran, Kimberly Chambert, Eric Banks, Sarah E. Bergen, Douglas M. Ruderfer, Robert E. Handsaker et al. "Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth." *The American Journal of Human Genetics* 91, no. 4 (2012): 597-607.

Fu, Ying, Nigora Mukhamedova, Sally Ip, Wilissa D’Souza, Katya J. Henley, Tia DiTommaso, Rajitha Kesani et al. "ABCA12 regulates ABCA1-dependent cholesterol efflux from macrophages and the development of atherosclerosis." *Cell Metabolism* 18, no. 2 (2013): 225-238.

Galan, Jose J., Belen Buch, Natalio Cruz, Ana Segura, Francisco J. Moron, Lluis Bassas, Luis Martinez-Pineiro, Luis M. Real, and Agustin Ruiz. "Multilocus analyses of estrogen-related genes reveal involvement of the ESR1 gene in male infertility and the polygenic nature of the pathology." *Fertility and Sterility* 84, no. 4 (2005): 910-918.

Galis, Frietson, Tom JM Van Dooren, Johan D. Feuth, Johan AJ Metz, Andrea Witkam, Sebastiaan Ruinard, Marc J. Steigenga, and Liliane CD Wijnaendts. "Extreme selection in humans against homeotic transformations of cervical vertebrae." *Evolution* 60, no. 12 (2006): 2643-2654.

Gansauge, Marie-Theres, and Matthias Meyer. "Selective enrichment of damaged DNA molecules for ancient genome sequencing." *Genome Research* 24, no. 9 (2014): 1543-1549.

“Genomics England." Genomics England. Accessed September 30, 2017. https://www.genomicsengland.co.uk/.

Gilbert, M. Thomas P., Daniela I. Drautz, Arthur M. Lesk, Simon YW Ho, Ji Qi, Aakrosh Ratan, Chih-Hao Hsu et al. "Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes." *Proceedings of the National Academy of Sciences* 105, no. 24 (2008): 8327-8332.

Glessner, Joseph T., Kai Wang, Guiqing Cai, Olena Korvatska, Cecilia E. Kim, Shawn Wood, Haitao Zhang et al. "Autism genome-wide copy number variation reveals ubiquitin and neuronal genes." Nature 459, no. 7246 (2009): 569-573.

Gonzalez, Enrique, Hemant Kulkarni, Hector Bolivar, Andrea Mangano, Racquel Sanchez, Gabriel Catano, Robert J. Nibbs et al. "The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility." *Science* 307, no. 5714 (2005): 1434-1440.

Gracey, Andrew Y., E. Jane Fraser, Weizhong Li, Yongxiang Fang, Ruth R. Taylor, Jane Rogers, Andrew Brass, and Andrew R. Cossins. "Coping with cold: an integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate." *Proceedings of the National Academy of Sciences of the United States of America* 101, no. 48 (2004): 16970-16975.

Gray, Maryke, Justin Roy, Linda Vigilant, Katie Fawcett, Augustin Basabose, Mike Cranfield, Prosper Uwingeli, Innocent Mburanumwe, Edwin Kagoda, and Martha M. Robbins. "Genetic census reveals increased but uneven growth of a critically endangered mountain gorilla population." *Biological Conservation* 158 (2013): 230-238.

“GRCh38 Human Reference Genome." Genome Browser. 2017. Accessed September 30, 2017. https://genome.ucsc.edu/.

Green, Richard E., Adrian W. Briggs, Johannes Krause, Kay Prüfer, Hernán A. Burbano, Michael Siebauer, Michael Lachmann, and Svante Pääbo. "The Neandertal genome and ancient DNA authenticity." *The EMBO Journal* 28, no. 17 (2009): 2494-2502.

Guo, Yanping, Elena Kotova, Zhe-Sheng Chen, Kun Lee, Elizabeth Hopper-Borge, Martin G. Belinsky, and Gary D. Kruh. "MRP8, ATP-binding cassette C11 (ABCC11), is a cyclic nucleotide efflux pump and a resistance factor for fluoropyrimidines 2′, 3′-dideoxycytidine and 9′-(2′-phosphonylmethoxyethyl) adenine." *Journal of Biological Chemistry* 278, no. 32 (2003): 29509-29514.

Gusnanto, Arief, Henry M. Wood, Yudi Pawitan, Pamela Rabbitts, and Stefano Berri. "Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data." *Bioinformatics* 28, no. 1 (2011): 40-47.

Han, Yuchen, Jesse Donovan, Sneha Rath, Gena Whitney, Alisha Chitrakar, and Alexei Korennykh. "Structure of human RNase L reveals the basis for regulated RNA decay in the IFN response." *Science* 343, no. 6176 (2014): 1244-1248.

Handsaker, Robert E., Vanessa Van Doren, Jennifer R. Berman, Giulio Genovese, Seva Kashin, Linda M. Boettger, and Steven A. McCarroll. "Large multiallelic copy number variations in humans." *Nature Genetics* 47, no. 3 (2015): 296-303.

Hart, Traver, Megha Chandrashekhar, Michael Aregger, Zachary Steinhart, Kevin R. Brown, Graham MacLeod, Monika Mis et al. "High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities." *Cell* 163, no. 6 (2015): 1515-1526.

Haynes, Gary. *Mammoths, mastodonts, and elephants: biology, behavior and the fossil record*. Cambridge University Press, 1993.

“hg19 Human Reference Genome.” Data. GATK | Resource Bundle. Accessed September 30, 2017. https://software.broadinstitute.org/gatk/download/bundle.

Hill, Christopher L. "Mammoths: Giants of the Ice Age." *Geoarchaeology-An International Journal* 24, no. 1: 117-119.

Hofreiter, Michael, David Serre, Hendrik N. Poinar, Melanie Kuch, and Svante Pääbo. "Ancient DNA." *Nature Reviews Genetics* 2, no. 5 (2001): 353-359.

Höss, Matthias, Amrei Dilling, Andrew Currant, and Svante Pääbo. "Molecular phylogeny of the extinct ground sloth Mylodon darwinii." *Proceedings of the National Academy of Sciences* 93, no. 1 (1996): 181-185.

Hu, Xuesong, Jianying Yuan, Yujian Shi, Jianliang Lu, Binghang Liu, Zhenyu Li, Yanxiang Chen et al. "pIRS: Profile-based Illumina pair-end reads simulator." *Bioinformatics* 28, no. 11 (2012): 1533-1535.

Huang, Hao, Elton Zeqiraj, Beihua Dong, Babal Kant Jha, Nicole M. Duffy, Stephen Orlicky, Neroshan Thevakumaran et al. "Dimeric structure of pseudokinase RNase L bound to 2-5A reveals a basis for interferon-induced antiviral activity." *Molecular Cell* 53, no. 2 (2014): 221-234.

Ivakhno, Sergii, Tom Royce, Anthony J. Cox, Dirk J. Evers, R. Keira Cheetham, and Simon Tavaré. "CNAseg—a novel framework for identification of copy number changes in cancer from second-generation sequencing data." *Bioinformatics* 26, no. 24 (2010): 3051-3058.

Jagannath, Aarti, Rachel Butler, Sofia IH Godinho, Yvonne Couch, Laurence A. Brown, Sridhar R. Vasudevan, Kevin C. Flanagan et al. "The CRTC1-SIK1 pathway regulates entrainment of the circadian clock." *Cell* 154, no. 5 (2013): 1100-1111.

Jiang, Xian-cheng, Can Bruce, Jefferson Mar, Min Lin, Yong Ji, Omar L. Francone, and Alan R. Tall. "Targeted mutation of plasma phospholipid transfer protein gene markedly reduces high-density lipoprotein levels." *Journal of Clinical Investigation* 103, no. 6 (1999): 907.

Jiang, Xian-Cheng, Weijun Jin, and Mahmood M. Hussain. "The impact of phospholipid transfer protein (PLTP) on lipoprotein metabolism." *Nutrition & Metabolism* 9, no. 1 (2012): 75.

Johnstone, Ricky W., Wu Wei, Alison Greenway, and Joseph A. Trapani. "Functional interaction between p53 and the interferon-inducible nucleoprotein IFI 16." *Oncogene* 19, no. 52 (2000): 6033-6042.

Kaplan, Murray L., and Gilbert A. Leveille. "Core temperature, O2 consumption, and early detection of ob-ob genotype in mice." *American Journal of Physiology--Legacy Content* 227, no. 4 (1974): 912-915.

Katada, Sayako, and Paolo Sassone-Corsi. "The histone methyltransferase MLL1 permits the oscillation of circadian gene expression." *Nature Structural & Molecular Biology* 17, no. 12 (2010): 1414-1421.

Khor, Seik-Soon, Taku Miyagawa, Hiromi Toyoda, Maria Yamasaki, Yoshiya Kawamura, Hisashi Tanii, Yuji Okazaki et al. "Genome-wide association study of HLA-DQB1* 06: 02 negative essential hypersomnia." *PeerJ* 1 (2013): e66.

Kidd, Jeffrey M., Gregory M. Cooper, William F. Donahue, Hillary S. Hayden, Nick Sampas, Tina Graves, Nancy Hansen et al. "Mapping and sequencing of structural variation from eight human genomes." Nature 453, no. 7191 (2008): 56-64.

Kim, Kyungjin, Hyuk Bang Kwon, and Jae Young Seong. "Cellular and Molecular Biology of Orphan G Protein- Coupled Receptors." *International Review of Cytology* 252 (2006): 163-218.

Kim, Tae-Min, Lovelace J. Luquette, Ruibin Xi, and Peter J. Park. "rSW-seq: algorithm for detection of copy number alterations in deep sequencing data." *BMC Bioinformatics* 11, no. 1 (2010): 432.

Klambauer, Günter, Karin Schwarzbauer, Andreas Mayr, Djork-Arné Clevert, Andreas Mitterecker, Ulrich Bodenhofer, and Sepp Hochreiter. "cn. MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate." *Nucleic Acids Research* 40, no. 9 (2012): e69-e69.

Kleinjan, Dirk A., and Veronica van Heyningen. "Long-range control of gene expression: emerging mechanisms and disruption in disease." The American Journal of Human Genetics 76, no. 1 (2005): 8-32.

Korbel, Jan O., Alexander Eckehart Urban, Jason P. Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M. Kim et al. "Paired-end mapping reveals extensive structural variation in the human genome." *Science* 318, no. 5849 (2007): 420-426.

Krumm, Niklas, Peter H. Sudmant, Arthur Ko, Brian J. O'Roak, Maika Malig, Bradley P. Coe, Aaron R. Quinlan, Deborah A. Nickerson, Evan E. Eichler, and NHLBI Exome Sequencing Project. "Copy number variation detection and genotyping from exome sequence data." *Genome Research* 22, no. 8 (2012): 1525-1532.

Lam, Hugo YK, Xinmeng Jasmine Mu, Adrian M. Stütz, Andrea Tanzer, Philip D. Cayting, Michael Snyder, Philip M. Kim, Jan O. Korbel, and Mark B. Gerstein. "Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library." *Nature Biotechnology* 28, no. 1 (2010): 47-55.

Lao, Xianjun, Qiliu Peng, Yu Lu, Shan Li, Xue Qin, Zhiping Chen, and Junqiang Chen. "Glutathione S-transferase gene GSTM1, gene-gene interaction, and gastric cancer susceptibility: evidence from an updated meta-analysis." *Cancer Cell International* 14, no. 1 (2014): 127.

Layer, Ryan M., Colby Chiang, Aaron R. Quinlan, and Ira M. Hall. "LUMPY: a probabilistic framework for structural variant discovery." *Genome Biology* 15, no. 6 (2014): R84.

Lazaridis, Iosif, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant et al. "Ancient human genomes suggest three ancestral populations for present-day Europeans." *Nature* 513, no. 7518 (2014): 409-413.

Lesna, I. Kralova, J. Rychlikova, L. Vavrova, and S. Vybiral. "Could human cold adaptation decrease the risk of cardiovascular disease?." *Journal of Thermal Biology* 52 (2015): 192-198.

Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform." *Bioinformatics*25, no. 14 (2009): 1754-1760.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. "The sequence alignment/map format and SAMtools." *Bioinformatics* 25, no. 16 (2009): 2078-2079.

Li, Heng. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM." *arXiv preprint arXiv:1303.3997* (2013).

Li, Jason, Richard Lupat, Kaushalya C. Amarasinghe, Ella R. Thompson, Maria A. Doyle, Georgina L. Ryland, Richard W. Tothill, Saman K. Halgamuge, Ian G. Campbell, and Kylie L. Gorringe. "CONTRA: copy number analysis for targeted resequencing." *Bioinformatics* 28, no. 10 (2012): 1307-1313.

Lin, Che-Hsin, Gwo-Bin Lee, Lung-Ming Fu, and Shu-Hui Chen. "Integrated optical-fiber capillary electrophoresis microchips with novel spin-on-glass surface modification." *Biosensors and Bioelectronics* 20, no. 1 (2004): 83-90.

Liu, Ruijie, Jahangir Iqbal, Calvin Yeang, David Q-H. Wang, M. Mahmood Hussain, and Xian-Cheng Jiang. "Phospholipid transfer protein–deficient mice absorb less cholesterol." *Arteriosclerosis, Thrombosis, and Vascular Biology* 27, no. 9 (2007): 2014-2021.

Liu, Shiping, Eline D. Lorenzen, Matteo Fumagalli, Bo Li, Kelley Harris, Zijun Xiong, Long Zhou et al. "Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears." *Cell* 157, no. 4 (2014): 785-794.

Lou, Dianne I., Ross M. McBee, Uyen Q. Le, Anne C. Stone, Gregory K. Wilkerson, Ann M. Demogines, and Sara L. Sawyer. "Rapid evolution of BRCA1 and BRCA2in humans and other primates." *BMC Evolutionary Biology* 14, no. 1 (2014): 155.

Lowen, Anice C., Samira Mubareka, John Steel, and Peter Palese. "Influenza virus transmission is dependent on relative humidity and temperature." *PLoS Pathogens* 3, no. 10 (2007): e151.

Lowrey, Phillip L., and Joseph S. Takahashi. "Genetics of circadian rhythms in Mammalian model organisms." *Advances in Genetics* 74 (2011): 175.

Lynch, Vincent J., Oscar C. Bedoya-Reina, Aakrosh Ratan, Michael Sulak, Daniela I. Drautz-Moses, George H. Perry, Webb Miller, and Stephan C. Schuster. "Elephantid genomes reveal the molecular bases of woolly mammoth adaptations to the Arctic." *Cell reports* 12, no. 2 (2015): 217-228.

MacDonald, G. M., D. W. Beilman, Y. V. Kuzmin, L. A. Orlova, K. V. Kremenetski, B. Shapiro, R. K. Wayne, and B. Van Valkenburgh. "Pattern of extinction of the woolly mammoth in Beringia." *Nature Communications* 3 (2012): 893.

MacDonald, Jeffrey R., Robert Ziman, Ryan KC Yuen, Lars Feuk, and Stephen W. Scherer. "The Database of Genomic Variants: a curated collection of structural variation in the human genome." *Nucleic Acids Research* 42, no. D1 (2013): D986-D992.

Magi, Alberto, Ants Kurg, Betti Giusti, Cristina Battaglia, Elena Bonora, Eleonora Mangano, Gian Franco Gensini et al. "EXCAVATOR: detecting copy number variants from whole-exome sequencing data." *Genome Biology* 14, no. 10 (2013): R120.

Magi, Alberto, Matteo Benelli, Seungtai Yoon, Franco Roviello, and Francesca Torricelli. "Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm." *Nucleic Acids Research* 39, no. 10 (2011): e65-e65.

Malmström, Helena, Jan Storå, Love Dalén, Gunilla Holmlund, and Anders Götherström. "Extensive human DNA contamination in extracts from ancient dog bones and teeth." *Molecular Biology and Evolution* 22, no. 10 (2005): 2040-2047.

Marshall, Christian R., Abdul Noor, John B. Vincent, Anath C. Lionel, Lars Feuk, Jennifer Skaug, Mary Shago et al. "Structural variation of chromosomes in autism spectrum disorder." *The American Journal of Human Genetics* 82, no. 2 (2008): 477-488.

McCarroll, Steven A., Tracy N. Hadnott, George H. Perry, Pardis C. Sabeti, Michael C. Zody, Jeffrey C. Barrett, Stephanie Dallaire et al. "Common deletion polymorphisms in the human genome." Nature genetics 38, no. 1 (2006): 86-92.

McCarroll, Steven A. "Extending genome-wide association studies to copy-number variation." Human molecular genetics 17, no. R2 (2008): R135-R142.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. "The ensembl variant effect predictor." *Genome biology* 17, no. 1 (2016): 122.

McQueen, Matthew J., Steven Hawken, Xingyu Wang, Stephanie Ounpuu, Allan Sniderman, Jeffrey Probstfield, Krisela Steyn et al. "Lipids, lipoproteins, and apolipoproteins as risk markers of myocardial infarction in 52 countries (the INTERHEART study): a case-control study." *The Lancet* 372, no. 9634 (2008): 224-233.

Meyer, Mara S., Kathryn L. Penney, Jennifer R. Stark, Fredrick R. Schumacher, Howard D. Sesso, Massimo Loda, Michelangelo Fiorentino et al. "Genetic variation in RNASEL associated with prostate cancer risk and progression." *Carcinogenesis* 31, no. 9 (2010): 1597-1603.

Miller, Christopher A., Oliver Hampton, Cristian Coarfa, and Aleksandar Milosavljevic. "ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads." *PloS One* 6, no. 1 (2011): e16327.

Miller, Webb, Daniela I. Drautz, Aakrosh Ratan, Barbara Pusey, Ji Qi, Arthur M. Lesk, Lynn P. Tomsho et al. "Sequencing the nuclear genome of the extinct woolly mammoth." *Nature* 456, no. 7220 (2008): 387-390.

Mills, Ryan E., Klaudia Walter, Chip Stewart, Robert E. Handsaker, Ken Chen, Can Alkan, Alexej Abyzov et al. "Mapping copy number variation by population-scale genome sequencing." *Nature* 470, no. 7332 (2011): 59-65.

Montell, Craig. "A mint of mutations in TRPM8 leads to cool results." *Nature Neuroscience* 9, no. 4 (2006): 466-468.

Morozova, Irina, Pavel Flegontov, Alexander S. Mikheyev, Sergey Bruskin, Hosseinali Asgharian, Petr Ponomarenko, Vladimir Klyuchnikov et al. "Toward high-resolution population genomics using archaeological samples." *DNA Research* 23, no. 4 (2016): 295-310.

Mudakikwa, Antoine B., Michael R. Cranfield, Jonathan M. Sleeman, and Ute Eilenberger. "Clinical medicine, preventive health care and research on mountain gorillas in the Virunga Volcanoes region." Mountain Gorillas, 2001, 341-60. doi:10.1017/cbo9780511661631.014.

Müller, Timo D., Anne Müller, Chun-Xia Yi, Kirk M. Habegger, Carola W. Meyer, Bruce D. Gaylinn, Brian Finan et al. "The orphan receptor Gpr83 regulates systemic energy metabolism via ghrelin-dependent and ghrelin-independent mechanisms." *Nature Communications* 4 (2013):1968.

Noonan, James P., Michael Hofreiter, Doug Smith, James R. Priest, Nadin Rohland, Gernot Rabeder, Johannes Krause, J. Chris Detter, Svante Pääbo, and Edward M. Rubin. "Genomic sequencing of Pleistocene cave bears." *Science* 309, no. 5734 (2005): 597-599.

Nord, Alex S., Wendy Roeb, Diane E. Dickel, Tom Walsh, Mary Kusenda, Kristen Lewis O'connor, Dheeraj Malhotra et al. "Reduced transcript expression of genes affected by inherited and de novo CNVs in autism." European Journal of Human Genetics 19, no. 6 (2011): 727-731.

Nørskov, M. S., R. Frikke-Schmidt, S. E. Bojesen, B. G. Nordestgaard, S. Loft, and A. Tybjærg-Hansen. "Copy number variation in glutathione-S-transferase T1 and M1 predicts incidence and 5-year survival from prostate and bladder cancer, and incidence of corpus uteri cancer in the general population." *The Pharmacogenomics Journal* 11, no. 4 (2011): 292-299.

Notredame, Cédric, Desmond G. Higgins, and Jaap Heringa. "T-Coffee: A novel method for fast and accurate multiple sequence alignment." *Journal of Molecular Biology* 302, no. 1 (2000): 205-217.

Nour, Adel M., Mike Reichelt, Chia-Chi Ku, Min-Yin Ho, Thomas C. Heineman, and Ann M. Arvin. "Varicella-zoster virus infection triggers formation of an interleukin-1β (IL-1β)-processing inflammasome complex." *Journal of Biological Chemistry* 286, no. 20 (2011): 17921-17933.

“Office of Research & Development." Million Veteran Program (MVP). Accessed September 30, 2017. https://www.research.va.gov/mvp/.

Ohashi, Jun, Izumi Naka, and Naoyuki Tsuchiya. "The impact of natural selection on an ABCC11 SNP determining earwax type." *Molecular Biology and Evolution* 28, no. 1 (2010): 849-857.

Orzalli, Megan H., Sara E. Conwell, Christian Berrios, James A. DeCaprio, and David M. Knipe. "Nuclear interferon-inducible protein 16 promotes silencing of herpesviral and transfected DNA." *Proceedings of the National Academy of Sciences* 110, no. 47 (2013): E4492-E4501.

Ou, Zhishuo, Paweł Stankiewicz, Zhilian Xia, Amy M. Breman, Brian Dawson, Joanna Wiszniewska, Przemyslaw Szafranski et al. "Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes." *Genome Research* 21, no. 1 (2011): 33-46.

Pääbo, Svante, Hendrik Poinar, David Serre, Viviane Jaenicke-Després, Juliane Hebler, Nadin Rohland, Melanie Kuch, Johannes Krause, Linda Vigilant, and Michael Hofreiter. "Genetic analyses from ancient DNA." *Annual Review of Genetics* 38 (2004): 645-679.

Pääbo, Svante. "Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification." *Proceedings of the National Academy of Sciences* 86, no. 6 (1989): 1939-1943.

Palkopoulou, Eleftheria, Love Dalén, Adrian M. Lister, Sergey Vartanyan, Mikhail Sablin, Andrei Sher, Veronica Nyström Edmark et al. "Holarctic genetic structure and range dynamics in the woolly mammoth." *Proceedings of the Royal Society of London B: Biological Sciences* 280 (2013): 20131910.

Palkopoulou, Eleftheria, Swapan Mallick, Pontus Skoglund, Jacob Enk, Nadin Rohland, Heng Li, Ayça Omrak et al. "Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth." *Current Biology* 25, no. 10 (2015): 1395-1400.

Pang, Andy Wing Chun, Ohsuke Migita, Jeffrey R. MacDonald, Lars Feuk, and Stephen W. Scherer. "Mechanisms of formation of structural variation in a fully sequenced human genome." *Human Mutation* 34, no. 2 (2013): 345-354.

Parikh, Hemang, Marghoob Mohiyuddin, Hugo YK Lam, Hariharan Iyer, Desu Chen, Mark Pratt, Gabor Bartha et al. "svclassify: a method to establish benchmark structural variant calls." *BMC Genomics* 17, no. 1 (2016): 64.

Parks, Matthew, and David Lambert. "Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study." *BMC Genomics* 16, no. 1 (2015): 19.

Pendleton, Matthew, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M. Stütz et al. "Assembly and diploid architecture of an individual human genome via single-molecule technologies." *Nature Methods* 12, no. 8 (2015): 780-786.

Perry, George H., Nathaniel J. Dominy, Katrina G. Claw, Arthur S. Lee, Heike Fiegler, Richard Redon, John Werner et al. "Diet and the evolution of human amylase gene copy number variation." *Nature genetics* 39, no. 10 (2007): 1256-1260.

Pilbrow, Varsha. "Dental and phylogeographic patterns of variation in gorillas." *Journal of Human Evolution* 59, no. 1 (2010): 16-34.

Plumptre, Andrew J., Stuart Nixon, Deo K. Kujirakwinja, Ghislain Vieilledent, Rob Critchlow, Elizabeth A. Williamson, Andrew E. Kirkby, and Jefferson S. Hall. "Catastrophic decline of world's largest primate: 80% loss of Grauer's gorilla (Gorilla beringei graueri) population justifies Critically Endangered status." *PloS One* 11, no. 10 (2016): e0162697.

Prüfer, Kay, Cesare de Filippo, Steffi Grote, Fabrizio Mafessoni, Petra Korlević, Mateja Hajdinjak, Benjamin Vernot et al. "A high-coverage Neandertal genome from Vindija Cave in Croatia." *Science* (2017): eaao1887.

Prüfer, Kay, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze et al. "The complete genome sequence of a Neanderthal from the Altai Mountains." *Nature* 505, no. 7481 (2014): 43-49.

Prüfer, Kay, Udo Stenzel, Michael Hofreiter, Svante Pääbo, Janet Kelso, and Richard E. Green. "Computational challenges in the analysis of ancient DNA." *Genome Biology*11, no. 5 (2010): R47.

Pucci, E., L. Chiovato, and A. Pinchera. "Thyroid and lipid metabolism." *International journal of obesity* 24, no. S2 (2000): S109-S112.

Rasmussen, Morten, Yingrui Li, Stinus Lindgreen, Jakob Skou Pedersen, Anders Albrechtsen, Ida Moltke, Mait Metspalu et al. "Ancient human genome sequence of an extinct Palaeo-Eskimo." *Nature* 463, no. 7282 (2010): 757-762.

Rausch, Tobias, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. "DELLY: structural variant discovery by integrated paired-end and split-read analysis." *Bioinformatics* 28, no. 18 (2012): i333-i339.

Reimand, Jüri, Meelis Kull, Hedi Peterson, Jaanus Hansen, and Jaak Vilo. "g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments." *Nucleic Acids Research* 35, no. suppl_2 (2007): W193-W200.

Repin, V.E., O.S. Taranov, Ryabchikova, A.N. Tikhonov, and V.G. Pugachev. 2004. "Sebaceous glands of the woolly mammoth, Mammothus primigenius Blum: histological evidence." *Doklady Biological Sciences : Proceedings Of The Academy Of Sciences Of The USSR, Biological Sciences Sections / Translated From Russian* 398, 382-384.

Reumer, Jelle WF, Clara MA ten Broek, and Frietson Galis. "Extraordinary incidence of cervical ribs indicates vulnerable condition in Late Pleistocene mammoths." *PeerJ* 2 (2014): e318.

Richman, Laura K., Jian-Chao Zong, Erin M. Latimer, Justin Lock, Robert C. Fleischer, Sarah Y. Heaggans, and Gary S. Hayward. "Elephant endotheliotropic herpesviruses EEHV1A, EEHV1B, and EEHV2 from cases of hemorrhagic disease are highly diverged from other mammalian herpesviruses and may form a new subfamily." *Journal of Virology* 88, no. 23 (2014): 13523-13546.

Rizzi, Ermanno, Martina Lari, Elena Gigli, Gianluca De Bellis, and David Caramelli. "Ancient DNA studies: new perspectives on old samples." *Genetics Selection Evolution* 44, no. 1 (2012): 21.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. "Integrative genomics viewer." *Nature Biotechnology*29, no. 1 (2011): 24-26.

Rogers, Rebekah L., and Montgomery Slatkin. "Genomic disintegration in woolly mammoths on Wrangel island." *arXiv preprint arXiv:1606.06336* (2016).

Ross, Michael G., Carsten Russ, Maura Costello, Andrew Hollinger, Niall J. Lennon, Ryan Hegarty, Chad Nusbaum, and David B. Jaffe. "Characterizing and measuring bias in sequence data." *Genome Biology* 14, no. 5 (2013): R51.

Routh A, Sleeman J. Proceedings of the British Veterinary Zoological Society. Howletts and Port Lympne Wild Animal Parks; Kent: Jun, 1997. 14–15. 22–25.

Salehzada, Tamim, Linda Cambier, Nga Vu Thi, Laurent Manchon, Laëtitia Regnier, and Catherine Bisbal. "Endoribonuclease L (RNase L) regulates the myogenic and adipogenic potential of myogenic cells." *PloS One* 4, no. 10 (2009): e7563.

Santi, Celia M., Pablo Martínez-López, José Luis de la Vega-Beltrán, Alice Butler, Arturo Alisio, Alberto Darszon, and Lawrence Salkoff. "The SLO3 sperm- specific potassium channel plays a vital role in male fertility." *FEBS Letters* 584, no. 5 (2010): 1041-1046.

Sathirapongsasuti, Jarupon Fah, Hane Lee, Basil AJ Horst, Georg Brunner, Alistair J. Cochran, Scott Binder, John Quackenbush, and Stanley F. Nelson. "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV." *Bioinformatics* 27, no. 19 (2011): 2648-2654.

Sawyer, Susanna, Johannes Krause, Katerina Guschanski, Vincent Savolainen, and Svante Pääbo. "Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA." *PloS One* 7, no. 3 (2012): e34131.

Schaffer, Nan, R. S. Jeyendran, and Bruce Beehler. "Improved sperm collection from the lowland gorilla: recovery of sperm from bladder and urethra following electroejaculation." *American Journal of Primatology* 24, no. 3- 4 (1991): 265-271.

Schmidt, Debra A., Mark R. Ellersieck, Michael R. Cranfield, and William B. Karesh. "Cholesterol values in free-ranging gorillas (Gorilla gorilla gorilla and Gorilla beringei) and Bornean orangutans (Pongo pygmaeus)." *Journal of Zoo and Wildlife Medicine* 37, no. 3 (2006): 292-300.

Schuenemann, Verena J., Alexander Peltzer, Beatrix Welte, W. Paul van Pelt, Martyna Molak, Chuan-Chao Wang, Anja Furtwängler et al. "Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods." *Nature Communications* 8 (2017).

Schultz, Adolph H. "Some distinguishing characters of the mountain gorilla." *Journal of Mammalogy* 15, no. 1 (1934): 51-61.

Schwartz-Narbonne, Rachel, Fred J. Longstaffe, Jessica Z. Metcalfe, and Grant Zazula. "Solving the woolly mammoth conundrum: amino acid 15N-enrichment suggests a distinct forage or habitat." *Scientific Reports* 5 (2015): 9791.

Seager, S. W. J., D. E. Wildt, N. Schaffer, and C. C. Platz. "Semen collection and evaluation in Gorilla gorilla gorilla." *American Journal of Primatology* 3, no. S1 (1982): 13-13.

Seguin-Orlando, Andaine, Cristina Gamba, Clio Der Sarkissian, Luca Ermini, Guillaume Louvel, Eugenia Boulygina, Alexey Sokolov et al. "Pros and cons of methylation-based enrichment methods for ancient DNA." *Scientific Reports* 5 (2015): 11826.

“Sequence Read Archive." National Center for Biotechnology Information. Accessed April 24, 2017. https://www.ncbi.nlm.nih.gov/sra.

“Shi PacBio and Shi IrysChip validated SVs.” Wang Genomics Lab. Accessed September 30, 2017. http://hx1.wglab.org/data/cnv sv/.

Shi, Lingling, Yunfei Guo, Chengliang Dong, John Huddleston, Hui Yang, Xiaolu Han, Aisi Fu et al. "Long-read sequencing and de novo assembly of a Chinese genome." *Nature Communications* 7 (2016).

Silva, J. Enrique. "The multiple contributions of thyroid hormone to heat production." *Journal of Clinical Investigation*108, no. 1 (2001): 35.

Sims, David, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, and Chris P. Ponting. "Sequencing depth and coverage: key considerations in genomic analyses." *Nature Reviews Genetics* 15, no. 2 (2014): 121-132.

Sindi, Suzanne S., Selim Önal, Luke C. Peng, Hsin-Ta Wu, and Benjamin J. Raphael. "An integrative probabilistic model for identification of structural variation in sequencing data." *Genome Biology* 13, no. 3 (2012): R22.

Singh, H. O., S. Lata, M. Angadi, S. Bapat, J. Pawar, V. Nema, M. V. Ghate, S. Sahay, and R. R. Gangakhedkar. "Impact of GSTM1, GSTT1 and GSTP1 gene polymorphism and risk of ARV-associated hepatotoxicity in HIV-infected individuals and its modulation." *The Pharmacogenomics Journal* 17, no. 1 (2017): 53-60.

Smith, Sean D., and Andrey Grigoriev. "GROM." OSF | GROM. 2017. Accessed September 30, 2017. http://doi.org/10.17605/OSF.IO/6RTWS.

Smith, Sean D., Joseph K. Kawash, and Andrey Grigoriev. "GROM-RD: Resolving genomic biases to improve read depth detection of copy number variants." *PeerJ* 3 (2015): e836.

Smith, Sean D., Joseph K. Kawash, and Andrey Grigoriev. "Lightning-fast genome variant detection with GROM." *GigaScience* 6, no. 10 (2017): 1-7.

Sneath, R. J., and D. C. Mangham. "The normal structure and function of CD44 and its role in neoplasia." *Molecular Pathology* 51, no. 4 (1998): 191.

Somerville, Martin J., Carolyn B. Mervis, Edwin J. Young, Eul-Ju Seo, Miguel del Campo, Stephen Bamforth, Ella Peregrine et al. "Severe expressive-language delay related to duplication of the Williams–Beuren locus." New England Journal of Medicine 353, no. 16 (2005): 1694-1701.

Sorrells, Shelly, Seth Carbonneau, Erik Harrington, Aye T. Chen, Bridgid Hast, Brett Milash, Ujwal Pyati et al. "Ccdc94 protects cells from ionizing radiation by inhibiting the expression of p53." *PLoS Genetics* 8, no. 8 (2012): e1002922.

Stanford, Craig B. "The subspecies concept in primatology: The case of mountain gorillas." *Primates* 42, no. 4 (2001): 309-318.

Stefansson, Hreinn, Roel A. Ophoff, Stacy Steinberg, Ole A. Andreassen, Sven Cichon, Dan Rujescu, Thomas Werge et al. "Common variants conferring risk of schizophrenia." *Nature* 460, no. 7256 (2009): 744-747.

Stephens, Philip J., David J. McBride, Meng-Lay Lin, Ignacio Varela, Erin D. Pleasance, Jared T. Simpson, Lucy A. Stebbings et al. "Complex landscapes of somatic rearrangement in human breast cancer genomes." *Nature* 462, no. 7276 (2009): 1005-1010.

Stephens, Zachary D., Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. "Big data: astronomical or genomical?." *PLoS Biology* 13, no. 7 (2015): e1002195.

Stranger, Barbara E., Matthew S. Forrest, Mark Dunning, Catherine E. Ingle, Claude Beazley, Natalie Thorne, Richard Redon et al. "Relative impact of nucleotide and copy number variation on gene expression phenotypes." *Science* 315, no. 5813 (2007): 848-853.

Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang et al. "An integrated map of structural variation in 2,504 human genomes." Nature 526, no. 7571 (2015): 75-81.

Takahashi, Saki, Yoichi Sakakibara, Emi Mishiro, Haruna Kouriki, Rika Nobe, Katsuhisa Kurogi, Shin Yasuda, Ming-Cheh Liu, and Masahito Suiko. "Molecular cloning, expression and characterization of a novel mouse SULT6 cytosolic sulfotransferase." *Journal of Biochemistry* 146, no. 3 (2009): 399-405.

Tani, Naoto, Yoshiko Dohi, Norio Kurumatani, and Kunio Yonemasu. "Seasonal distribution of adenoviruses, enteroviruses and reoviruses in urban river water." *Microbiology and Immunology* 39, no. 8 (1995): 577-580.

Tarasov, Artem, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. "Sambamba: fast processing of NGS alignment formats." *Bioinformatics* 31, no. 12 (2015): 2032-2034.

Telenti, Amalio, Levi CT Pierce, William H. Biggs, Julia di Iulio, Emily HM Wong, Martin M. Fabani, Ewen F. Kirkness et al. "Deep sequencing of 10,000 human genomes." *Proceedings of the National Academy of Sciences* (2016): 201613365.

Thomas, Paul D., Michael J. Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. "PANTHER: a library of protein families and subfamilies indexed by function." *Genome Research* 13, no. 9 (2003): 2129-2141.

Vartanyan, Sergey L., Khikmat A. Arslanov, Juha A. Karhu, Göran Possnert, and Leopold D. Sulerzhitsky. "Collection of radiocarbon dates on the mammoths (Mammuthus primigenius) and other genera of Wrangel Island, northeast Siberia, Russia." *Quaternary Research* 70, no. 1 (2008): 51-59.

Vergeer, Menno, S. Matthijs Boekholdt, Manjinder S. Sandhu, Sally L. Ricketts, Nicholas J. Wareham, Morris J. Brown, Ulf de Faire et al. "Genetic variation at the phospholipid transfer protein locus affects its activity and high-density lipoprotein size and is a novel marker of cardiovascular disease susceptibility." *Circulation* 122, no. 5 (2010): 470-477.

Visscher, Peter M. "Sizing up human height variation." Nature genetics 40, no. 5 (2008): 489-490.

Wagner, Franz F., and Willy A. Flegel. "RHD gene deletion occurred in the Rhesus box." *Blood* 95, no. 12 (2000): 3662-3668.

Wang, Tim, Kıvanç Birsoy, Nicholas W. Hughes, Kevin M. Krupczak, Yorick Post, Jenny J. Wei, Eric S. Lander, and David M. Sabatini. "Identification and characterization of essential genes in the human genome." *Science* 350, no. 6264 (2015): 1096-1101.

Watts, David P. "Mountain gorilla life histories, reproductive competition, and sociosexual behavior and some implications for captive husbandry." *Zoo Biology* 9, no. 3 (1990): 185-200.

Weir, Bruce S., and C. Clark Cockerham. "Estimating F- statistics for the analysis of population structure." *Evolution* 38, no. 6 (1984): 1358-1370.

Weischenfeldt, Joachim, Orsolya Symmons, Francois Spitz, and Jan O. Korbel. "Phenotypic impact of genomic structural variation: insights from and for human disease." Nature Reviews Genetics 14, no. 2 (2013): 125-138.

Wu, Yinghua, Lifeng Tian, Mario Pirastu, Dwight Stambolian, and Hongzhe Li. "MATCHCLIP: locate precise breakpoints for copy number variation using CIGAR string by matching soft clipped reads." *Frontiers in Genetics* 4 (2013).

Xiao, Ling, Yuanhong Chen, Ming Ji, and Jixin Dong. "KIBRA regulates Hippo signaling activity via interactions with large tumor suppressor kinases." *Journal of Biological Chemistry* 286, no. 10 (2011): 7788-7796.

Xie, Chao, and Martti T. Tammi. "CNV-seq, a new method to detect copy number variation using high-throughput sequencing." *BMC Bioinformatics* 10, no. 1 (2009): 80.

Xue, Yali, Javier Prado-Martinez, Peter H. Sudmant, Vagheesh Narasimhan, Qasim Ayub, Michal Szpak, Peter Frandsen et al. "Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding." *Science* 348, no. 6231 (2015): 242-245.

Yang, Haiyan, Siyu Yang, Jing Liu, Fuye Shao, Haiyu Wang, and Yadong Wang. "The association of GSTM1 deletion polymorphism with lung cancer risk in Chinese population: evidence from an updated meta-analysis." *Scientific reports* 5 (2015).

Yang, Tie-Lin, Xiang-Ding Chen, Yan Guo, Shu-Feng Lei, Jin-Tang Wang, Qi Zhou, Feng Pan et al. "Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis." *The American Journal of Human Genetics* 83, no. 6 (2008): 663-674.

Ye, Dan, Illiana Meurs, Megumi Ohigashi, Laura Calpe-Berdiel, Kim LL Habets, Ying Zhao, Yoshiyuki Kubo et al. "Macrophage ABCA5 deficiency influences cellular cholesterol efflux and increases susceptibility to atherosclerosis in female LDLr knockout mice." *Biochemical and Biophysical Research Communications* 395, no. 3 (2010): 387-394.

Ye, Kai, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads." *Bioinformatics* 25, no. 21 (2009): 2865-2871.

Yi, Ming, Yongmei Zhao, Li Jia, Mei He, Electron Kebebew, and Robert M. Stephens. "Performance comparison of SNP detection tools with illumina exome sequencing data—an assessment using both family pedigree information and sample-matched SNP array data." *Nucleic Acids Research* 42, no. 12 (2014): e101-e101.

Yoon, Seungtai, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, and Jonathan Sebat. "Sensitive and accurate detection of copy number variants using read depth of coverage." *Genome Research* 19, no. 9 (2009): 1586-1592.

Yoshiura, Koh-ichiro, Akira Kinoshita, Takafumi Ishida, Aya Ninokata, Toshihisa Ishikawa, Tadashi Kaname, Makoto Bannai et al. "A SNP in the ABCC11 gene is the determinant of human earwax type." *Nature Genetics* 38, no. 3 (2006): 324-330.

Yu, Jianzhong, Yonggang Zheng, Jixin Dong, Stephen Klusza, Wu-Min Deng, and Duojia Pan. "Kibra functions as a tumor suppressor protein that regulates Hippo signaling in conjunction with Merlin and Expanded." *Developmental Cell* 18, no. 2 (2010): 288-299.

Yvan-Charvet, Laurent, Mollie Ranalletta, Nan Wang, Seongah Han, Naoki Terasaka, Rong Li, Carrie Welch, and Alan R. Tall. "Combined deficiency of ABCA1 and ABCG1 promotes foam cell accumulation and accelerates atherosclerosis in mice." *The Journal of Clinical Investigation*117, no. 12 (2007): 3900.

Yvan-Charvet, Laurent, Nan Wang, and Alan R. Tall. "Role of HDL, ABCA1, and ABCG1 transporters in cholesterol efflux and immune responses." *Arteriosclerosis, Thrombosis, and Vascular Biology* 30, no. 2 (2010): 139-143.

Zao, Chih-Ling, John A. Ward, Lisa Tomanek, Anthony Cooke, Ron Berger, and Karyn Armstrong. "Virological and serological characterization of SRV-4 infection in cynomolgus macaques." *Archives of Virology* 156, no. 11 (2011): 2053.

Zhang, Bing, Stefan Kirov, and Jay Snoddy. "WebGestalt: an integrated system for exploring gene sets in various biological contexts." *Nucleic Acids Research* 33, no. suppl_2 (2005): W741-W748.

Zhou, Xuming, Xuehong Meng, Zhijin Liu, Jiang Chang, Boshi Wang, Mingzhou Li, Pablo Orozco-ter Wengel et al. "Population genomics reveals low genetic diversity and adaptation to hypoxia in snub-nosed monkeys." *Molecular Biology and Evolution* 33, no. 10 (2016): 2670-2681.

Zöller, Margot. "CD44, hyaluronan, the hematopoietic stem cell, and leukemia-initiating cells." *Frontiers in Immunology* 6 (2015): 235.

Zook, Justin M., Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. "Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls." *Nature Biotechnology* 32, no. 3 (2014): 246-251.

Zook, Justin M., David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng et al. "Extensive sequencing of seven human genomes to characterize benchmark reference materials." *Scientific Data* 3 (2016).