Lossy Compression of Quality Values in Next-Generation Sequencing Data

Veronica Suaste Morales

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Master of Science

of

Brock University.

Department of Computer Science

Brock University

I, Veronica Suaste Morales, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

In recent years costs for sequencing human DNA have dropped drastically. This fact has allowed a fast development of several projects around the world that are generating large amounts of DNA sequencing data. This deluge of data has caused the problem of limited storage space that researchers are trying to solve through compression techniques for DNA sequencing files.

In this work we address the compression of SAM files which is the standard output file for DNA alignment. We specifically studied lossy compression techniques used for quality values reported in the SAM file and we analysed the impact of such lossy techniques in the CRAM format. We present a series of experiments using a data set corresponding to individual NA12878 with three different fold coverages. For these data sets we applied lossy techniques: QVZ [1], LEON [2], Illumina binning [3], and we also introduced a new lossy model, dynamic binning technique. We analysed the compression ratio when using CRAM format and we also studied the impact of all these lossy techniques in the SNP calling process. Our results show that lossy techniques allow a better CRAM compression ratio. We also show that SNP calling performance is not negatively affected. Moreover we confirmed that this process can even boost the SNP calling performance.

Acknowledgements

I would like to thank everyone who helped to complete this work:

- I thank my supervisor Sheridan Houghten for her time and her interest in this research.
- I extend my gratitude to Cale Fairchild who provided me with technical support for storage space in several occasions. Your help is really appreciated.
- I also thank Consejo Nacional de Ciencia y Tecnología (CONACYT) for its financial support for my master program.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years the sequencing of human DNA has become a notable field of research from many different areas such as genetics, biology, chemistry, computer science, bioinformatics, etc. The importance that this topic has achieved is due to the broad impact that it has directly on humanity. The influence that it is having on humans goes from understanding DNA, and therefore life and evolution, to health care. In particular, DNA sequencing is being largely used for understanding genetic diseases and their prevention. It also allows the design of specific drugs for specific groups of patients based on their genetic profiles.

In the last twenty years the costs of sequencing human DNA was a barrier for researchers, but since 2015 it became possible to sequence a complete human genome for $1000 [7]. This huge advance in the field implies a large amount of data being generated every day from many big projects around the world such as the 1000 Genomes Project.

It is expected that by the year 2025 [6] two billion human genomes will be sequenced. In general this deluge of data represents a huge challenge from the storage space point of view. Researchers have studied several compression techniques for next-generation output data in an effort to face this problem. The studied approaches for compression of these files vary from general text compression techniques to some more specialized models where the particular properties of DNA strands make the compression more efficient (for example, DNA strands have repetitive content given that the alphabet consists of only 4 letters — A, C, T and G). Nevertheless, there is also much more information reported in the output files than just the DNA sequence (see Section 3.4.1) which makes the compression even more challenging.

Our main target for the study is the quality values reported in FASTQ files, the standard format for storing the output of high-throughput sequencing instruments, as well as SAM (Sequence Alignment/Map format) files, the standard format for storing read alignments against

reference sequences. These quality values report a score per-base associated with the nucleotide sequence and can be understood as the probability of an error in the base calling. The alphabet used for these quality scores consists of approximately 40 characters, which represents another barrier for achieving better compression. See Section 3.5 for further information on quality values.

Recently, new lossy models of compression for quality values have been studied [8], [2], [9], [1], [10]. We focus our work in analysing three of them: [2], [1] and [3]. We will also suggest new ideas for adjusting the quality scores, (Section 4.6).

The analysis we will perform is related to the random access CRAM format, see Section 3.4.4. The CRAM format is a new format designed by the European Bioinformatics Institute that compresses SAM/BAM files and achieves 40-50% space saving over the alternative BAM format. The objective of this format is to replace the BAM format and become the standard compression model for sequencing data. We will study the impact of the lossy models mentioned above in this CRAM compression format.

When using lossy compression techniques it is expected that the loss of information should be measured, and the effect of this loss on subsequent tasks performed with the compressed data should be also analysed. With the intention of analysing the possible loss of information when adjusting the quality scores, we study the effect that these adjustments will have on the SNP calling performance, which is the process of finding single nucleotide variants in sequencing data. Some recent results [11] suggest that SNP calling performance is not negatively affected, and it can even be boosted when adjusting the quality scores.

1.1 Thesis Structure

This thesis is structured as follows: In Chapter 2 we present the background required to understand this work. Chapter 3 intends to summarize the state of the art related to compression of next-generation output data. Chapter 4 explains all the methodology for the experimental process. Chapter 5, Chapter 6 and Chapter 7 contain the results and the discussion for experiments performed with 5x, 6x and 50x fold coverage data sets, respectively. Chapter 8 concludes this study, and suggests some possible ideas to continue exploring compression of quality values.

Chapter 2

Background

In this chapter we introduce the basic biological concepts required to understand this work. Secondly, we introduce DNA sequencing, its history and the main technologies used for this purpose. Next, we summarize the most important projects related to human genome sequencing around the world and its importance for humanity.

2.1 Basic Biological Concepts

Bioinformatics, according to the Oxford English Dictionary [12], consists of conceptualising biology in terms of molecules and applying informatics techniques to understand and organise the information with these molecules, on a large scale. In other words, this area is about applying knowledge from areas such as applied mathematics, computer science and statistics to analyse physical chemistry of molecules. The recent deluge of biological data, and the problem it represents [13], has made bioinformatics an essential tool for many related applications. In this work we address one of these applications.

2.1.1 Genetics and DNA

Genetics is considered a field of biology for which the principal targets of study are genes. The objective of this area is to understand biological heredity and genetic variation in living organisms. The origin of this science dates back to 1865 when Johann Gregor Mendel (1822-1884) published his results about fundamental laws of inheritance that he discovered by working on pea plants for about eight years. The experiments that Mendel performed pointed to the existence of one of the most important biological elements that nowadays we know as genes.

Around the same time Friedrich Miescher in 1869 at the University of Tübingen was the first to isolate a substance that he called *nuclein* and that one hundred years after would be

known as DNA. It was in 1953 when James Watson and Francis Crick at the University of Cambridge, based on so many other results and research, discovered and proposed the double helix model of DNA structure. They published their results in the journal Nature [14], [15], and because of this discovery they were awarded the Nobel prize.

After years of research, now we know that DNA, or deoxyribonucleic acid, is a double helix molecule that contains most of the genetic information that makes each individual from each species unique. This structure enables DNA to carry biological information from one generation to the next and it is considered as the blueprint for each living thing. Most DNA is located in the cell nucleus and it is called nuclear DNA. Also, a small portion of DNA can be found in the mitochondria and it is known as mitochondrial DNA or mtDNA. All organisms, in sexual reproduction, inherit half of their nuclear DNA from the male parent and the other half from the female parent. However mitochondrial DNA is inherited in all organisms from the female progenitor.

The two-stranded shape of DNA chemical structure allows biological instructions to be passed along with great precision. Each strand is made up of building blocks known as nucleotides and each nucleotide has three parts: a phosphate group, a sugar group and one of four different types of nitrogen bases. The four possible chemical bases are adenine (A), guanine (G), cytosine (C) and thymine (T). The way these bases are ordered dictates the biological instruction contained in the DNA. Another particular property of these bases is that they pair up with each other, A with T and C with G, to form units called base pairs (see Fig 2.1). This property becomes relevant when DNA copies itself during cell division. In this process the double strand structure is split so each one of the strands can be used as template for the production of the opposite strand. As a result two new double structures are created by pairing up the corresponding bases.

The human genome is built of about 3.2 billion bases considering only one set of chromosomes. Between each individual, no matter what race, less than 1% of the DNA is different. Research has shown that among these variations of DNA we can find diseases and changes in the cell functions. This is why understanding the human genome is of vital importance.

2.1.2 Genes, Chromosomes and Proteins

The definition of *gene* can be complicated. We simplify the concept by considering a gene as a specific region of DNA that contains instructions usually on how to produce molecules called *proteins* or for a particular function. A gene is also known as “the functional unit of heredity”

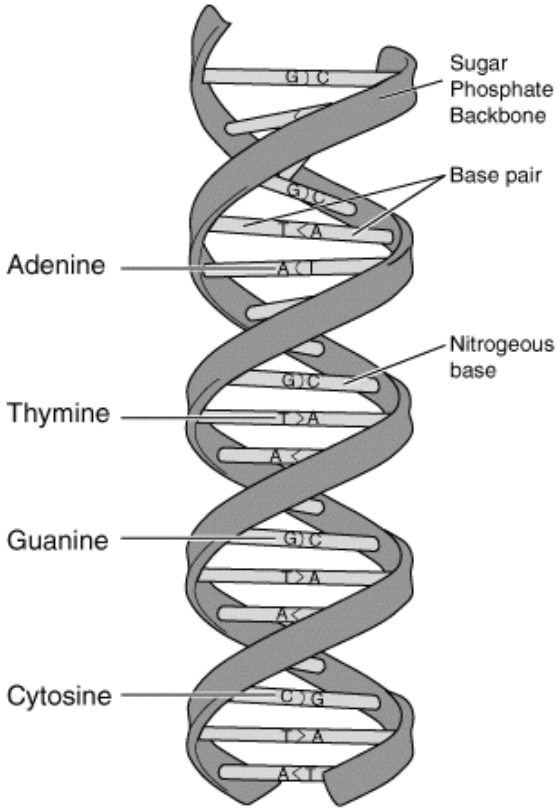


Figure 2.1: DNA 3D structure. (a) DNA double helix structure. (b) Base pairs formed by A-T and C-G [4]

as per the Oxford English Dictionary [16].

Recent research sets the number of human genes between 20,000 and 25,000 [17] and gene size can vary from a few hundred to more than 2 million bases. Each gene consists of three types of nucleotide sequence (see Fig 2.2):

- Coding regions, called exons, which specify a sequence of amino acids.
- Non-coding regions, called introns, which do not specify amino acids.
- Regulatory sequences, which play a role in determining when and where the protein is made as well as how much of the protein is produced.

In human DNA there are always two copies of each gene, one inherited from the mother and the other one from the father. The different forms of the same gene can have differences in their sequence of DNA; these forms are called alleles. The small differences in people's DNA are what make each person unique in the sense of physical features, although most of the genes are the same in all people.

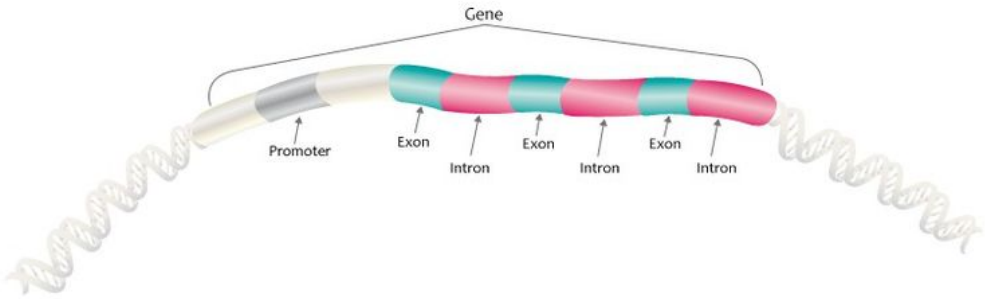


Figure 2.2: Gene structure, [5]

A long set of nucleotides form genes, and groups of genes are packaged tightly to form important structures in life called *chromosomes*. These structures and their DNA are copied as part of the cell cycle and they are passed to daughter cells as part of the process called mitosis and meiosis. Human beings have 23 pairs of chromosomes; 22 pairs of *autosomes*, which means they look the same in males and females, and one sex pair which differ between males and females.

The process in which genes determine the behaviour of the cell is complex and is controlled within each cell. This task consists mainly of two steps called *transcription* and *translation* and is also known as *gene expression*. This bring us to another important concept, proteins. When a gene is expressed it generates a copy of itself in the form of messenger RNA and then is translated to generate a protein, which is a molecule consisting of long chains of amino acids. Protein structure dictates where it will act and what it will do.

2.2 DNA Sequencing

DNA sequencing is the process in which the precise order of nucleotide bases, adenine, guanine, cytosine, and thymine within a DNA molecule, is determined. The importance of this process lies in the fact that the order of the nucleotides determines the instructions for the hereditary properties of life, as well as the biochemical properties. Nowadays knowledge of DNA sequences is indispensable for basic biological research, diagnostics, biotechnology, forensic biology and many other applications.

The first time researchers sequenced DNA molecules was in 1970, by using a series of complicated and time consuming methods. Since then, DNA sequencing methods have evolved and became easier and faster. In 1977 Frederick Sanger introduced the *dideoxy* se-

quencing method [18], which nowadays is known as *Sanger sequencing* and until approximately 2005, before the emergence of next-generation sequencing technologies, it was the most used method of DNA sequencing. The Sanger method was the choice of technology to produce the first human genome in 2001. Of course, the time and costs were huge, but the progress in science that this represented was huge as well.

2.2.1 Next-Generation Sequencing

The term *next-generation sequencing (NGS)* is used to describe a set of modern sequencing technologies such as Illumina (Solexa), Roche 454, Ion Torrent, Oxford Nanopore and others. As a result of these new technologies, DNA sequencing is done more cheaply and quickly than the previous Sanger sequencing technology. Next-generation sequencing is also known as *high-throughput sequencing* and it has revolutionised the study of genomics and molecular biology. NGS is mainly characterized by its improved speed, reduced manpower and reduced cost. All these properties are gained because all these methods are massively parallel, which means that the number of sequence reads for a single experiment is greater than for the experiment based on Sanger sequencers. All these DNA sequencing technologies share the property that they cannot read whole genomes in one step. Instead, these machines read small pieces, varying in size from 30 bases to even 30000 bases. These small sequences are called *reads*.

2.2.1.1 Illumina (Solexa)

What nowadays we know as Illumina sequencing technologies [19] started in the mid-1990s in the Chemistry Department of Cambridge University with experiments from scientists Shankar Balasubramanian and David Klenerman. The Illumina next-generation sequencing approach differs from the classic Sanger chain-termination method. With these instruments the sequencing is done by synthesis (SBS) technology, tracking the addition of labeled nucleotides as the DNA chain is copied, in a massively parallel fashion. The amount of data that Illumina sequencing systems can deliver went from 300 kilobases up to one terabase in only a single run, depending on the configuration and instrument used.

As part of the Illumina technologies, there is the most powerful sequencing platform ever created, HiSeq X System, which was released in 2015. The system made it possible to sequence the human genome for only $1000, and is capable of delivering 18,000 of these per year. This sequencing solution is the world's first to break the thousand dollar human genome barrier.

2.2.1.2 Ion Torrent

This sequencing technology [20] is based on the detection of hydrogen ions that are released during polymerization of DNA. When a nucleotide is incorporated into a DNA strand by a polymerase, a hydrogen ion is released, so if there are two identical bases on the DNA strand the voltage will be double, and the chip will record two identical bases. This detection can be done directly so each nucleotide incorporation is recorded in seconds, because there is no scanning, no cameras and no light like other methods. These technologies were released for the first time in 2010 and up to now they have kept improving, marketing their machines as a rapid, compact and economical sequencer that can be afforded by a large number of laboratories all around the world.

2.2.1.3 Nanopore Technologies

DNA sequencing which works with nanopore-based technology is considered as part of fourth-generation DNA sequencing technologies. It has been under development since 1995 but it was not until February of 2012 that Oxford Nanopore Technologies (ONT) released preliminary experimental results from GridION system [21] using this technology.

A nanopore is a tiny hole with internal diameter at the order of 1 nanometer. In this type of technology the idea is to pass DNA molecules through the nanopore. The scale of the nanopore forces the DNA to pass as a long string, one base at the time. Depending on which base is blocking the nanopore, A, C, T or G, the amount of current needed is different.

2.2.2 DNA Assembly

Given the fact that DNA sequencing machines cannot read the whole genome, the output of this technology is a large set of reads, which are small sub-sequences. The process in which this large set of reads is organized and put back together to re-build the genome is called *DNA assembly*. This process has represented a big challenge because of the large amount of data that needs to be processed. There are two main methods of DNA assembly. The first one is *reference alignment* consisting of the alignment of all reads against one reference genome. The second one is called *de novo assembly*, which does not need a reference genome, making use of the overlaps information contained in the read themselves.

2.3 Human Genome Project

After DNA sequencing became widespread since 1977, the idea of large-scale sequencing started to be discussed by human geneticists. It took years for the idea to become an action plan, but finally in 1990 the Human Genome Project (HGP) was announced. The main goal of this project was to complete mapping and understanding of all genes in human beings. The time for doing this was expected to be 15 years.

This project stands as the world's largest collaborative biological project. At the start it was funded by the Department of Energy and US National Institutes of Health, and later in the UK from the Medical Research Council and Wellcome Trust which helped to run this project on a huge scale, with sequencing centres in France, Germany, China and Japan also joining the big project. The large collaboration around the world made it possible to finish two years before expected. In April 2003, the Human Genome Project was declared complete [22].

The final sequence released by HGP of about 3 billion DNA bases, for the haploid genome, covers about 99% of the human genome for regions that contain genes and it was sequenced to an accuracy of 99.99 percent, which helps to understand better the organization and structure of genes. Besides the human genome, this project included the sequencing of the mouse genome, and the identification of more than 3 million human genetic variations. All sequenced data generated by HGP is available in public databases for all scientists around the world, a fact that has led to an outstanding advance in health science discoveries. The complete human genome sequence started a series of more in-depth comparative studies and generated a large amount of raw data that required specific computing infrastructures, software implementation and biological data analysis, problems that bioinformatics is trying to solve.

2.3.1 The 1000 Genomes Project

The 1000 Genomes Project was a joint effort among several research groups from the US, UK, China and Germany to produce a catalogue of human genetic variations. This project ran from 2007 to 2015 and its principal goal was to find most genetic variations with frequencies of at least 1% in the populations studied. The large catalogue created has allowed medical researchers to find genetic differences that contribute to rare and common diseases. Locating and analysing these genetic variations leads to discovery of new diagnostic tests and in some cases treatments. The whole project was divided into three phases and respective publications

were made public for research purposes: Pilot Analysis [23], Phase 2 Analysis [24] and Phase 3 Analysis [25], [26]. The final version released by the project contains low coverage and exome sequence data for 2,504 individuals from 26 different populations and data for 24 individuals that were sequenced to high coverage for validation purposes. As part of the 1000 Genomes Project the Data Coordination Center (DCC) was set up to manage project-specific data flow, to ensure archival sequence data and to manage community access. This represented a fundamental challenge for bioinformatics considering that as of March 2012 there were more than 260 terabytes of raw data [27].

2.3.2 Platinum Genomes Project

The Platinum Genomes Project is a data set publicly available released by Illumina performed on Illumina HiSeq systems [28]. It has data from the 17 member CEPH pedigree 1463 that was sequenced to 50x depth and one trio sequenced to 200x depth. This project also made public a set of high-confidence variant calls for NA12877 and NA12878 members by taking into account the inheritance constraints in the pedigree and the concordance of variant calls across different methods.

2.4 Genome Analysis

DNA variations are linked to genetic disorders, so they are the target of health researchers. Nowadays, with all of the data obtained from sequencing technologies they are making considerable advances and allowing the discovery of treatments and in some cases cures for genetic diseases.

2.4.1 Mutations and Polymorphisms

Mutation is the natural process in which the DNA sequence is changed. More regularly only a single base, A, T, G, C, is substituted for another, but also sometimes a base can be deleted or an extra base can be added. However the cell is able to naturally repair most of these changes and also not all mutations are responsible for something bad or dangerous for the living organism.

Genetic differences that occur in more than 1 percent of the population are called *polymorphisms*, and these changes in DNA are common enough to be considered normal variation. Normally these kinds of polymorphisms are the cause for common differences between people such as blood type, complexion, skin color and eye color.

Although many polymorphisms in human genomes are not related to a person's health, some of these variations may influence the risk of developing certain disorders as well as genetic diseases such as cancer among others. What happens is that if a mutation occurs in a functional part of DNA, which is called the *coding area*, it may prevent one or more proteins from working properly causing the genetic disorder.

2.4.2 Variant Calling

Variant calling is known as a set of processes for finding variations in data from next-generation sequencing technologies. These computational methods are based on known population single nucleotide polymorphisms (SNP). The large amount of data today is making these techniques more challenging and a wide variety of algorithms have been specifically designed.

For understanding how the variant calling algorithms work, first we have to remember how the main NGS technologies work (Section 2.2.1). They generally do sequencing by synthesis. The synthesis process is captured in a series of fluoresce images and base-calling algorithms infer the actual nucleotide information. Then they assign a measure of uncertainty or *quality score* to each base call. Of course base-calling procedures vary depending on the sequencing system. Once having the base-calling and its respective quality score, the process of obtaining a set of genotypes— DNA sequences which determines a specific characteristic (phenotype)— for each individual in a sample is divided into two steps: SNP calling and genotype calling.

SNP calling or variant calling is part of the process that determines where polymorphisms exist or where there is a difference from a reference sequence in at least one of the bases. Genotype calling consists of determining the genotype for each individual which is highly related to the position of a SNP or a variant that has already been called.

Normally the SNP calling process will consist of a series of steps which includes filtering before and after, but the main algorithm is usually a statistical model or some heuristics to predict the likelihood of variation at each locus, based on the quality scores and counts of the aligned reads at that locus.

Chapter 3

Literature Review of Compression for Sequencing Data

In this chapter we introduce the problem of Big Data related to next-generation sequencing data. We also present an overview of general and particular compression techniques for sequencing data that already exist. We introduce the quality values, which are the main focus of our research. We summarize the state of the art of lossy compression techniques for quality values and we provide a brief discussion about it.

3.1 Big Data in Genomics

Next-generation sequencing technologies and their continuous improvements have revolutionized several areas in biology and health sciences by reducing the time and cost required for sequencing. As of 2015 [7] the cost of sequencing the human genome has been reduced to $1000. This fact has led to the generation of a large amount of data, as a typical data file varies from tens to hundreds of gigabytes of disk space. And although, so far, the main successful focus has been investing in data generation [13], the final goal of all this sequencing development is to be able to analyse and understand DNA, which means that a collateral challenge is not only the storage but the distribution and data analysis as well. For example, the raw data obtained by the 1000 Genomes Project (Section 2.3.1) after six months of activity exceeded the sequence data in NCBI Genbank database accumulated in the preceding 21 years [13].

Big Data in genomics has been compared with three of the largest generators of Big Data: astronomy, YouTube, and Twitter [6], which makes evident the huge challenge that scientists are facing with this issue. For a wider perspective refer to Figure 3.1.

According to the analysis in [6], projecting for the year 2025 they estimate there will be at

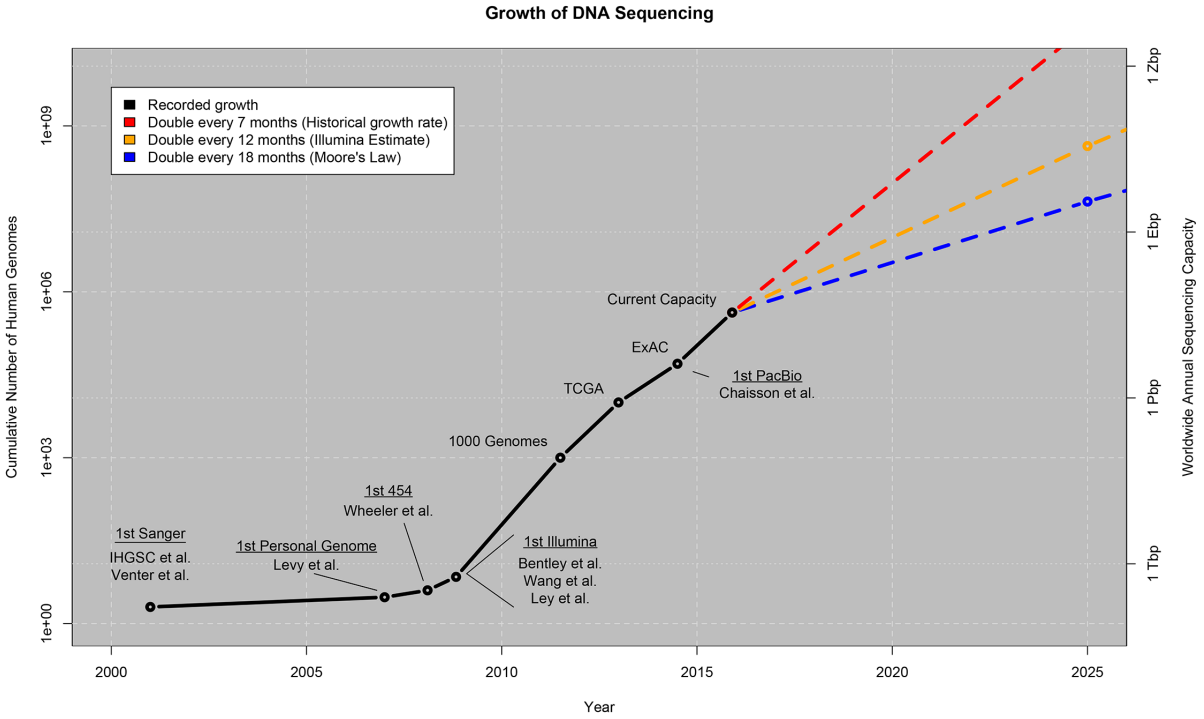


Figure 3.1: Growth of DNA sequencing [6]

least 2.5 million plant and animal genome sequences, as the result of massive projects around the world [29], [30]. Also, for the same year, they estimate between 1 billion and as many as 2 billion human genomes. Translating all this to disk space, they have determined that until today, even only considering 20 of the largest institutions, the storage required is more than 100 petabytes and they predict that just for human genomes the storage capacity that will be needed in 2025 is as much as 2-40 exabytes.

Considering the challenge of storing such large amounts of data, it is natural to think about data compression. Within the next pages the objective is to show how the Big Data problem in genomics has been tackled, the different techniques that have been used, and all the possibles approaches that still need further research and experiments.

3.2 General Data Compression

In general data compression refers to the set of different techniques for handling huge data of any kind by reducing the space needed to store this data and speed up the data circulation. Strictly speaking, a compression algorithm takes an input X and generates a representation X_c that requires fewer bits. There is also the inverse process, which operates on the compressed representation X_c to generate the reconstruction, Y [31]. Depending on the reconstruction Y, data compression schemes can be either a *lossless* compression technique if Y is identical to

X or *lossy* compression technique if Y is only an approximation for X; the latter implies that the original data cannot be recovered exactly.

3.2.1 Measures of Performance

Evaluation of compression algorithms can be done from many different approaches; it can be evaluated by the relative complexity, the memory required by the algorithm implementation, the speed given by how fast the algorithm performs on a given machine, the amount of compression and how closely the reconstruction resembles the original data.

For measuring the amount of compression the most common technique is called *compression ratio* which is simply the ratio of the number of bits required to represent the original data over the number of bits required to represent the data after compression.

In lossy compression, the reconstruction differs from the original data. This difference is called *distortion*. The measure of this distortion is an indicator of the algorithm efficiency. There are many distortion metrics (refer to [31] for more details).

Nowadays data compression is present in almost every application, and is used for text, images, sounds and video. Below we present a broad overview of the most commonly used techniques.

Huffman Coding

Huffman coding is a lossless compression data technique based on Huffman codes developed in 1952 [32], which are a specific type of prefix code.
This algorithm assigns binary codes to symbols of a given alphabet in such a way that the overall number of bits used to encode a typical string formed with those symbols is optimally minimized.

Lempel-Ziv

The Lempel-Ziv algorithm was published for the first time in 1977 [33]. After that, the authors published several variations, all based on the main idea of a sliding window during compression. As result of this algorithm the transmission of the data consists of a set of addresses and the length of the copied segment.

Burrows-Wheeler Transform

Burrows-Wheeler's Transform [34] is a lossless data compression algorithm, mainly based on block sorting of the input. For each block in the input data, this algorithm applies a reversible

transformation, which reorders the data. The idea for reordering is to group characters together. Although the transformation itself does not compress the data, standard compression algorithms can be applied with much more efficient results to the reordered data. The mentioned efficiency is given because, after the transformation, the probability of finding two instances of the same character close to each other has been substantially increased and this specific property is exploited by a combination of other common algorithms such as move-to-front, Huffman and arithmetic coding.

Arithmetic Coding

The arithmetic coding algorithm for data compression was popularized in 1987 [35] and can be used in lossless and lossy compression models.
The important process of this algorithm consists of the conversion of input data consisting of a set of symbols into a floating-point number in the interval $[0, 1)$. The conversion relies on a model to characterize each symbol during the time of processing.

Golomb Coding

In 1960 Solomon W. Golomb [36] invented the data compression codes used for lossless data compression. This algorithm is based on a model of the probability of the values: one natural number is assigned to each value according to its probability, with small values more likely than big ones. One important parameter for the code is the divisor which captures the relationship between size and probability.

3.3 Sequencing Data Compression

Compression of DNA sequencing data started even before the big data revolution that whole human genome sequencing and NGS technologies brought in latter years. The first approaches for sequencing data were based on text compression techniques which were adapted to exploit obvious properties of DNA sequences such as the 4-letter alphabet, regularities and presence of palindromes [37]. All these previous techniques were characterized by using a combination of two different methods: firstly, substitutional or dictionary based, where most repetitive sub-sequences are identified and encoded with a representation of smaller size, and secondly, statistical based methods, in which a prediction model is established which assigns probabilities to each base based on the data and then uses an encoding scheme that will perform more efficiently based on the probability distribution.

Among the most relevant within this path of research we have XM [38] which trains a second-order Markov model on the full data and uses arithmetic encoding[35] relying on the calculated probabilities, CTW-LZ [39] which uses the context tree weighting prediction model and Lempel-Ziv [33] algorithms for compression. BioCompress [40] and BioCompress2 [41] use both Lempel-Ziv and arithmetic encoding, and there are others following the same idea such as GenCompress, [42] DNACompress [43] and GeNML [44].

3.4 Next-Generation Sequencing Data Compression

With all the advances from NGS technologies new challenges also emerged for compression of its output data because these technologies, along with the sequence or read itself, also report additional metadata needed for downstream DNA analysis. This metadata includes a larger alphabet than the 4-letter alphabet, the one that had been considered for sequencing data compression until that time. Therefore new research and techniques for sequencing data were necessary. We will give an overview of the research that has been developed. First we introduce the data formats which are the target for compression in all these problems.

3.4.1 Data formats

3.4.1.1 Sequence Alignment/Map format (SAM)

The Sequence Alignment/Map (SAM) [45] format is a generic alignment format for storing read alignments against reference sequences. It has been developed with the main purpose of allowing DNA analysis and the exchange of information from various sequencing platforms. This format consists of an optional header and then each line represents the linear alignment of one read. Each line has 11 mandatory fields described in Table 3.1. For an example of this file see Figure 3.2. There is also BAM format which is the binary version of a SAM file and is designed to compress reasonably well. These two formats are nowadays the industry standards for reporting alignment/mapping information. Also all the important tools for analysis of high-throughput sequencing data require these formats as the input. Examples of these tools are GATK [46], Samtools [47] and FreeBayes [48].

3.4.1.2 FASTQ Format

FASTQ format [49] was originally developed at the Wellcome Trust Sanger Institute and it has become the standard format for storing the output of high-throughput sequencing instruments. This format is used to store the nucleotides sequence and its corresponding quality scores,

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A- ˜] {1,254}	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	*|[!-()+-¡¿- ˜][!- ˜] *	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	* |([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=—[!-()+-¡¿- ˜][!- ˜] *	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+]	segment SEQuence
11	QUAL	String	[!-\]+	ASCII of Phred-scaled base QUALity+33

Table 3.1: SAM format mandatory fields

```
SRR622461.74266141      99      20      60000   60      28S73M  =       60270   371     AAACTGAA
AGTTAATAGAGAGGTGACTCAGATCCAGAGGTGGAAGAGGAAGGAAGCTTGGAACCCTATAGAGTTGCTGAGTGCCAGGACCAGATCCTGGCC
HHHHHHHHHHHHHHHHHHHHHHHHEHHHHHHHHHHHHHHHHHGHHHHHHHHHHHHHHHHHHHHHHGHHHHHHHHDHHHHHHHHHGHEHHHHHHHHHHDHH
EHHHH     MD:Z:73 PG:Z:MarkDuplicates     RG:Z:group1     NM:i:0  AS:i:73 XS:i:0
SRR622461.74266142      99      20      60000   60      19S82M  =       60272   373     GTTAATAG
AGAGGAGACTCAGATCCAGAGGTGGAAGAGGAAGGAAGCTTGGAACCCTATAGAGTTGCTGAGGGACAGGACCAGATCCTGGCCCTAAACAGG
FCEEFA@<AA.44++44544F<=C2544447@A/5514448?>CC####################################################
#####     MD:Z:52T1C27    PG:Z:MarkDuplicates     RG:Z:group1     NM:i:2  AS:i:72 XS:i:21
SRR622461.84716310      121     20      60000   60      33S68M  =       60000   0       AGAAAAAA
CTGAAAGTTAATAGAGAGGTGACTCAGATCCAGAGGTGGAAGAGGAAGGAAGCTTGGAACCCTATAGAGTTGCTGAGTGCCAGGACCAGATCC
GHHHHHHIHHHHHHHHHHHHGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHIHHHHHHHHHHHHHHHHHHHHHHHH
HHHHH     MD:Z:68 PG:Z:MarkDuplicates     RG:Z:group1     NM:i:0  AS:i:68 XS:i:21
SRR622461.84716310      181     20      60000   0       *       =       60000   0       TCAATGTT
TTCTAAGTTTCTGTGAACAGATCTTTTATATCCTTGGTTAAATTATTTCTAAGAATTTTTTGTACCAATTGTAAATAGTTTCTCTTGATTTCT
HHHHHHHHHHEHIHHHHEHHHGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHHH     PG:Z:MarkDuplicates     RG:Z:group1     AS:i:0  XS:i:0
SRR622461.84716299      177     20      60033   60      101M    1       219044260       0
ACCCTATAGAGTTGCTGAGTGCCAGGACCACACCCTGCCCCTAAACAGGTGGTAAGGAAGGAGAGAGTGAAGGAACTGCCAGGTGACACACTCCCA
CCATG   ##########################################F?AEEFHFFHEBG@DCFFG??DAADD>@DE555355555-4404440
4-49FDD=@@?@@     MD:Z:30G1T4G63  PG:Z:MarkDuplicates     RG:Z:group1     NM:i:3  AS:i:86 XS:i:2
4
SRR622461.74266140      163     20      60048   60      101M    =       60242   295     TGAGTGCC
AGGACCAGATCCTGGCCCTAAACAGGTGGTAAGGAAGGAGAGAGTGAAGGAACTGCCAGGTGACACACTCCCACCATGGACCTCTGGGATCCT
HHHHHHHHHHHHHHHHHHHHHHHHHGHHHHHHHHHHFHHGHHHHFHHHBE?H:>+>>FF@DFHGIHHFGA>F35::7CFGEGGGGECDAAEB######
#####     MD:Z:101        PG:Z:MarkDuplicates     RG:Z:group1     NM:i:0  AS:i:101        XS:i:2
3
```

Figure 3.2: Partial Sam file

each of them encoded with a single ASCII character. Each line in the file uses four lines per sequence, containing the read id (always starts with '@'), the sequence, an optional description or default '+' sign and a last line for the quality values. See Fig 3.3 for a better idea of how a FASTQ file looks.

3.4.2 Reference-based Compression

Considering that DNA strings contain only four possible letters, A, C, G and T, it is expected that there are many repetitions between the sequences, and this property has been broadly exploited for compression techniques of DNA sequencing. The methods grouped as *reference-based* follow two essential steps. First they choose a reference sequence and then only encode

```
@ERR001291.19 080901_HWI-EAS301_0003_30APFAAXX:2:1:1108:802/1
GAGAAGGAAGTGACTATACAAGGGCAGCACAAGGAA
+
IIIIIIIIIIIIIIIIIIIIIII2IIIII2IIIII
@ERR001291.20 080901_HWI-EAS301_0003_30APFAAXX:2:1:1182:146/1
GTAATGTCTGTTGTTGTTTTTTTTTTGAAAAATATT
+
IIIIIIIIIIIIIII+IIIIII3II6)6?>IIB+I.
@ERR001291.21 080901_HWI-EAS301_0003_30APFAAXX:2:1:1330:1638/1
GTTTTCAAGTAATACAACTTAGATAATTTTTAGTGA
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIBI@I
@ERR001291.22 080901_HWI-EAS301_0003_30APFAAXX:2:1:1516:206/1
GAACCAGATCTTTGTTTCATTGATTTTATCTATTGT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII4I
```

Figure 3.3: Partial Fastq file

the differences between the sequence to be compressed and the reference. For a good performance of this method the choice of the reference sequence is important. There are some cases in which instead of one reference a set of possible references is allowed [50], [51], [52]. As another example, in [53] the genome is divided into blocks and for each position the longest matching block is compressed. Also, in [54], only differences between similar or overlapping reads are encoded. Also, Deorowicz and Grabowski [52] propose a technique which has a reference genome in addition to other short sequences taken from the data to be compressed. There are other approaches focused only on output of NGS technologies, namely SAM and FASTQ formats. For FASTQ format there is Fastqz [55] which encodes the DNA alphabet and uses Lempel-Ziv encoding for matched and mismatched positions against a reference. Gencompress [56], also reference-based, calculates statistics on the mismatches and performs the encoding based on these statistics. For SAM format mzip [57] uses Huffman coding for compressing only position and read length; Samcomp [55] and NGC [58], SlimGene [59] and Quip [60] are also in this category.

3.4.3 Reference-free Read Compression

De novo or reference-free compression is performed without an external reference genome. Instead, this kind of method exploits similarities between reads themselves [55], [61], [62], [63], [64], [65]. Most commonly this technique will use a context-model to predict the bases and then use an arithmetic encoder or they will re-order reads to maximize similarities for consecutive reads allowing a better compression with standard methods. FQZCOMP [55] and DSRC [62] are examples of the context-model compressors; PATHENC [66] is also of

this type complemented with arithmetic encoding. According to [2] read re-ordering methods are the ones that achieve a better compression ratio. Compressors of this type are BEETL [64] which uses Burrows-Wheeler transform, ORCOM [63], MINCE [65] and more recently, LEON [2], which proposes a method based on a probabilistic de Bruijn graph stored in a Bloom filter [67].

3.4.4 Random Access and CRAM Format

Although compression is mainly focused on solving the problem of storage and distribution of Big Data, it is also possible to make more practical the step of analysing the data, by applying particular compression techniques for specific files. This type of compression allows us to access part of the file, from its compressed version, without going through the entire decompression process. For achieving random access compression generally the input is split into blocks and even different compression algorithms can be applied to different blocks within the same file.

Today, BAM format[45] is the standard compression model with random access property achieving compressions of 50-80% of their original SAM file and allowing an accessible and practical analysis of sequencing data using the compressed file. Although it is a huge success and it is supported by most sequencing data analysis tools, the compression ratio achieved is not sustainable in the long run as sequencing data is growing at a big rate. Because of this, researchers have explored new options, and as one of the best results CRAM framework technology has been developed. CRAM [68], based on the work of Fritz et al. [57], is a new format designed by the European Bioinformatics Institute that compresses SAM/BAM files and achieves 40-50% space saving over the alternative BAM format. The objective of this format is to replace the BAM format and become the standard compression model for sequencing data. In recent years it has gained huge popularity in the area and its popularity is expected to grow even more. By now it is supported for the main tools for analysis of sequencing data and also big initiatives such as the 1000 genomes project have their data available for public use stored in this format.

3.5 Quality Values

The base calling process performed for NGS technologies to determine the bases of a DNA string is prone to a different type of error. In order to report the probability of base calling mistakes, sequencers generate a quality score for each nucleotide in the read.

A quality score indicates the level of confidence of a particular read base and is represented as the probability of the base being correctly determined during sequencing. The higher the score, the lower the probability that the base at that position has been incorrectly called. It is generally computed using *Phred score* [69] which is the number $Q = -10log_{10}P$, where P is the estimated probability of the corresponding nucleotide being incorrect, calculated by specific software running in the sequencing machine. Usually the quality values are represented in a file with a printable ASCII alphabet [33:73] or [64:104], with each value corresponding to $Q+33$ or $Q+64$, respectively. The importance of maintaining quality scores as part of the data relies on the fact that they are directly used in next-generation sequencing analysis, such as Single Nucleotide Polymorphism(SNP) detection [70]. Quality scores comprise a significant percentage of sequencing data and they became a bottle neck for compression because the alphabet required to represent all quality values is larger —about 40 characters— than the one required for the read sequence. Therefore compression algorithms designed for the sequence itself will not perform quite as well when applied to quality values. For this purpose, specific algorithms must be provided taking into account particular properties and information that quality scores have themselves.

3.6 Lossy compression for sequencing data

As we mentioned before, there are lossless and lossy techniques for data compression. Nevertheless lossy techniques are not allowed for DNA sequences due to the fact that losing or changing one single nucleotide represents a big impact on the possible encoded protein. However researchers have applied this technique not to the DNA sequence itself but to the quality values reported for each base, which represents almost half of the total data in the output file. There are in the literature several techniques trying to solve the problem of quality values compression by using lossy techniques [8],[2],[9], [1],[10].

When lossy compression techniques are applied to any type of data, the natural path to follow is to measure the loss of data and in good cases the missed data should not affect or mislead any other possible output derived from the compressed data. Any lossy technique in this field should show that downstream analysis, SNP calling, is not affected. A good starting point for this type of analysis is [10], which suggests that quality values data is noisy data, and which presents an analysis where losing precision in quality score was actually beneficial for SNP calling.

Quality Score Bins	Mapped Quality score
N(no call)	N(no call)
2-9	6
10-19	15
20-24	22
25-29	27
30-34	33
35-39	37
≥ 40	40

Table 3.2: Q-score Bins for an Optimized 8-level mapping

Cánovas *et al.* [8] proposed a lossy technique where the quality values were separated into blocks of variable size, each block with only one representative value which depends on the distortion measure selected for the user from a set of possible measures. Concurrently, Illumina [3] suggested what today is known as *illumina binning*, in which the resolution of quality scores was reduced by employing a quality scoring scheme with only eight levels of quality or less. Their complete binning is presented in Table 3.2.

The same year, Ochoa *et al.* [10] presented QualComp, a lossy compressor for quality scores based on rate distortion. In this framework the user is allowed to specify the number of bits per quality score prior to compression. This compressor works with FASTQ files and performs in clustering model.

More recently, researchers from Stanford University published a new approach, *Quality Values Zip* QVZ [1], based on the same ideas as [10]. In this work the quality score sequence is modelled as a Markov chain of order one. Then empirical transition probabilities are computed from the data, a collection of Lloyd-Max quantizers are constructed (one for each possible base in each position) and finally an arithmetic encoder is run over the result.

Benoit *et al.* [2] proposed a reference-free compression model for sequencing data; they present the model for reads compression and they use the same construction for implementation of lossy compression for quality values. Their process consists of building a de Bruijn Graph of the most recurrent *k-mers* in the sequences, with each read encoded as a path in this graph. For quality score compression, they truncate all quality values above a given threshold. Also all positions covered by at least a certain number of recurrent *k-mers* are replaced by one representative value, with this number computed based on the quality value so that the lower the quality the higher the number of covering *k-mers* is required.

3.7 Comparison and Discussion

Lossy compression techniques for quality values is a very recently explored area, with the first ideas presented in 2011. During our research on the state of the art of quality values compression we realized that the main ideas have evolved so that nowadays they are used in more new discoveries. We have analyzed results mainly from the most recent research [2], [1], [10], as we consider it contains also previous research and the comparison they presented suggests more worthy results to consider when it comes to studying this topic.

When comparing two or more lossy compression techniques, there is not a general rule about how to proceed. There are many factors involved in the result and there is always a trade off between compression ratio and any other measure, that can be distortion rate, impact on downstream analysis, speed or memory requirements. For measuring distortion rate many different metrics are used as can be appreciated in [8], [1] and [10]. For rating impact on downstream analysis —SNP calling— the scores are normally presented using F-score which considers both precision and recall measures. This measures are defined in Section 4.5.

Qualcomp [10] focusses its comparisons and performance measurement on rate distortion metric, specifically working with mean square error. Qualcomp compression allows the user to specify the number of bits per quality score, so they present the relationship between number of bits per quality score and mean square error.

For downstream analysis they study SNP calling but we note that they consider as ground truth the data resulting from variant calling performed with the original quality values. They also accept that the running time for their algorithm is longer than the ones they are comparing against[61], [55]. Nevertheless they achieve better compression ratio minimizing mean square error and only a little is compromised in SNP calling.

Cánovas *et. al* [8] presented their work and compared it against Qualcomp. In this case they used several fidelity measures and showed that their method outperformed Qualcomp when considering *Max : Min Distance* as the measure. They also based their SNP analysis considering variant calling with original quality values as benchmark. They did not report data for running time and memory storage.

The next year, in 2015, Malysa *et al.* [1] released QVZ for quality values compression. They also measured their performance with distortion rate metrics, including mean square error, average L1, where $d(x,y) = |x-y|$ and average Lorentzian where $d(x+y) = log_2(1+|x-y|)$. They showed that their method outperformed Qualcomp and the algorithm

of Cánovas *et al.* for all three choices of distortion metric. Also a few months later they published an exhaustive analysis on the effect of lossy compression of quality values using QVZ on variant calling [71]. This analysis included SNP calling compared against two different sets for benchmark recently released, one by GIAB (Genome in a Bottle) and adapted by the National Institute of Standardizations and Technology (NIST) and the other one by Illumina as part of the Platinum Genomes project. For their comparisons, along with the previously mentioned algorithms, they also included the Illumnina binning method, see Table 3.2. The main contribution of their work is not only the reduction of storage space but also their proof that SNP calling is not affected. Moreover, they confirm with their experiments that smoothing quality values can improve it.

Benoit *et al.* [2] presented a model for sequencing data. Their algorithm was based on a probabilistic de Bruijn graph that was designed originally for compression of the sequence (read). Once they built the graph, they also used it to perform a lossy transformation of the quality values. We have to consider this for comparisons because the probabilistic de Bruijn graph has high memory requirements. For comparison against other algorithms they selected models with option to compress with lossy techniques as well [55], [61], [65]. Their new algorithm performed better when considering compression ratio, compression time and decompression time. They also presented SNP calling analysis considering as bench mark set the variants provided by the 1000 genomes project and also showed that lossy techniques on quality values can improve SNP calling.

3.8 Objective of this work

In this work we address the compression of SAM files which is the standard output file for DNA alignment. We specifically study lossy compression techniques used for quality values reported in the SAM. We selected three of the most promising lossy techniques: QVZ [1], LEON [2], Illumina binning [3], and we also introduce a new lossy model, dynamic binning technique. The objective of this study is to analyse and discuss how each of these lossy techniques will perform when using the CRAM compression format for SAM files. Because we are analysing lossy techniques for quality values we also want to provide evidence that these kinds of methods will not impact negatively in the SNP calling process. For such purpose we provide an analysis of SNP calling performance.

Chapter 4

Methodology

In this chapter we present all toolkits and software used for our research. We introduce the data sets for the experiments as well as the sets used as ground truth for evaluation of SNP calling. We also present the metrics for SNP calling performance. And we explain the experimental process followed in our work.

4.1 Toolkits

4.1.1 GATK

GATK stands for Genome Analysis Toolkit [46]. It was developed by the Data Science and Data Engineering group at the Broad Institute. This toolkit is a collection of command-line tools for analyzing high-throughput sequencing data in formats such as SAM/BAM/CRAM (see Section 3.4.1) and VCF (see Section 4.2) with a primary focus on variant discovery and genotyping.

When using GATK for variant calling we follow GATK Best Practices [72], [73], which is the recommended workflow for variant discovery analysis with GATK. This process intends to maximize the technical correctness of the data. The first steps start from the raw reads indicating how to do the mapping to a reference genome, marking duplicates with Picard tools (see subsection 4.1.4) and performing a base quality score recalibration. Once the data has been pre-processed it is ready to continue with the variant discovery process. In this second part the process considers the fact that some of the variation might be caused by mapping and sequencing artifacts. Finding a good trade-off between sensitivity (minimizing false negatives) and specificity (minimizing false positives) can be very difficult, and can also be dependant on the project, Instead, the process maximize sensitivity but they also report a variant quality score recalibration (VQSR) which further allows the user to customize specificity for each

project.

4.1.2 Samtools

Samtools [47] is a suite of programs for interacting with high-throughput sequencing data in formats such as SAM/BAM/CRAM. Samtools makes it possible to work directly with a compressed BAM/CRAM file, without having to uncompress the whole file. We also use this toolkit for making SNP calling through *mpileup* command, which calculates genotype likelihoods supported by the aligned reads and does the SNP calling based on those likelihoods.

Bcftools is another module included in samtools and we use it for handling VCF files, see Section 4.2.

4.1.3 HTSlib

HTSlib is a C-library for manipulating file formats such as SAM, CRAM and VCF. It is used for studying high-throughput sequencing data and is the core library used by samtools.

4.1.4 Picard

Picard [74], also created by Broad Institute developers, is an open source under MIT license set of command line tools for manipulating high-throughput sequencing data. We specifically use this toolkit for marking duplicate reads as part of the GATK Best Practices.

4.1.5 BWA

Burrows-Wheeler Aligner (BWA) [75] is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. When performing alignments with this software, we considered the reference genome GRCh37 called human_g1k_v37.fasta available at [76], which corresponds to the reference genome used in phase1 and phase3 of 1000 Genomes Project.

4.2 VCF Format

The Variant Call Format (VCF) [77] is a text file format for storing gene sequence variations. It consists of a couple of lines for meta-information, a header line, and then is followed by a number of lines that each contains information about a variation in the genome. It is also possible to store genotype information on samples for each position in this format. There are eight mandatory fields for each reported variation and they have to follow a specific order as well, see Table 4.1 for further explanation. A VCF file will be the output of variant calling

Name	Brief description
CHROM	The chromosome in which the variation is being called.
POS	The 1-based position of the variation on the given sequence.
ID	The identifier of the variation.
REF	The reference base(s).
ALT	The list of alternative allele(s).
QUAL	A quality score associated with the inference of the given alleles.
FILTER	A flag indicating which of a given set of filters the variation has passed.
INFO	An extensible list of information describing the variation.
FORMAT	An optional field for describing the samples.

Table 4.1: VCF format specifications

performed with either GATK tools or Samtools.

4.3 Datasets For SNP Calling

For analyzing the impact of lossy compression models for the quality values we use datasets from 1000 Genomes project. All datasets correspond to the *Homo Sapiens* individual NA12878. This individual is the daughter in one of the trios sequenced from Utah residents of northern and western European ancestry (CEU). Specifically we use the low coverage alignment (6x) to perform the experiments with the whole genome (22 chromosomes). This alignment is provided by 1000 Genomes project and available at their public repository [78]. We also extracted chromosome 11 and 20 from the whole genome with high coverage alignment (50x), available at the same repository. And finally, for a complete process experiment, we consider the read dataset SRR622461 for the same individual (NA12878) with 5x coverage; for this raw dataset we performed the alignment with bwa software [75]. We have selected individual NA12878 because it is the only one for which a well analysed ground truth set of variants has been developed and has been publicly released.

4.4 Quality Benchmark for SNP Calling

For having a measure of how lossy models can affect SNP calling, we first need to set the baseline that will serve as a reference when comparing the performance of lossless compression against the different lossy compression models. For this purpose we use two *ground truth* sets of variants that have been developed and refined specifically for individual NA12878. The first set of variants that we consider as ground truth was released by the Genome in a Bottle consortium (GIAB) [79] and it has been adapted by the National Institute of Standardizations

and Technology (NIST). In their work [79] they integrated and arbitrated between 14 data sets from five sequencing technologies, seven read mappers and three different variant callers resulting in a set of variants which allows high-confidence SNP calling without depending on specific caller or sequencing technologies. The second gold standard we used is the one released by Illumina as part of the Platinum Genomes project.

4.5 SNP Calling Performance Metrics

When comparing two different lossy models for quality values, we will evaluate how each one of them affects SNP calling. For this purpose, we consider this problem as a binary classifier for the variants. This will allow us to divide each variant in the resulting VCF file, as True Positive (TP) when the same variant is also part of the ground truth and False Positive (FP) when the variant in the resulting VCF file can not be found in the ground truth. We also examine False Negatives (FN) which are the variants in the ground truth that cannot be found in the resulting VCF file.

By considering this segregation we can do the performance evaluation with typical metrics such as sensitivity, precision and F-score.

1. **Sensitivity**, also known as the true positive rate or recall, measures the proportion of correctly identified positives and is given by the expression:

$$Sensitivity = \frac{TP}{TP+FN}. \tag{4.1}$$

2. **Precision**, also known a positive predictive value, measures the proportion of identified positives that are true and is given by the expression:

$$Precision = \frac{TP}{TP+FP}. \tag{4.2}$$

3. **F-score** is a metric for accuracy in binary classification. This score considers both **precision** and **sensitivity** by computing the harmonic mean of them. It can be interpreted as a weighted average and is computed as:

$$F-score = 2 \times \frac{Sensitivity \times Precision}{Sensitivity + Precision}. \tag{4.3}$$

When analyzing results given by these metrics we interpret them considering that a perfect **Sensitivity** score of 1.0 indicates that all variants from the ground truth were correctly identified, although it does not indicate how many irrelevant variants were also called as positive. On the other hand, a total **Precision** score of 1.0 means that all obtained variants are relevant but indicates nothing about the total of possible positive variants. Generally there is always a trade-off between these two metrics and depending on the case one could prefer to increase one of them by causing a decrease in the other one. Because of this, F-score, which combines both metrics, will give us a better overall evaluation.

4.5.1 ROC Curve

Receiver Operating Characteristic (ROC) curve is used to visualize the performance of a binary classifier while varying a certain discrimination threshold. This curve is the result of plotting the true positive rate (TPR) against the false positive rate (FPR) with different thresholds.

When evaluating variant calling performance with metrics sensitivity, precision and F-score, we considered all variants in an output VCF file to be correct. With this second approach, using ROC as metric, we vary the quality threshold and consider variants in an output VCF file to be correct only if they are above the threshold.

This metric, in this specific problem of variant calling performance, was introduced in the work of William et al. [11]. As in their original work, we follow their design and when comparing different sets of variants we take the union of them as the domain. This rescaling is done with the purpose of addressing the fact that the true negative rate of correctly called variants will be so much larger, as most of the genome will not be variant, and so could cause misleading results. The implication of performing this rescaling is that ROC curves between different plots are not comparable as they will have different domains.

For our analysis we also look at the AUC, area under the curve, which indicates the probability that the binary classifier will rank a random positive case higher than a random negative case. For the AUC, the closer to 1 the better.

For plotting ROC curve and computing AUC, we use ROCR package [80].

4.6 Dynamic Binning

In order to explore new ideas for reducing the alphabet used for quality values, we developed a dynamic binning. As in Illumina binning, this method splits the alphabet into bins. But in

Quality Score Bins	Mapped Quality score
1-10	c_1
11-20	c_2
21-30	c_3
31-40	c_4
41-256	c_5

Table 4.2: Q-score Bins for dynamic binning, value c_i, for a given bin with range $[l,r]$, is such that $H(c_i) \geq H(c) \forall c \in [l,r]$.

this case it will have 5 bins and the value representing each bin will be the one with the largest number of occurrences belonging to that bin. We apply this method block-wise, which means that the whole file will be split into blocks and for each block the 5-binning will be different depending on its histogram. In our experiment we considered blocks of 1000 reads each, an empirically selected parameter.

For a given block, let H denote the histogram of the quality values, i.e. for any character c in that block, $H(c)$ is the number of occurrences of c in the block. The 5-binning for each block is performed according to Table 4.2 where each representative value c_i, for a given bin with range $[l,r]$, is such that $H(c_i) \geq H(c)\ \forall c \in [l,r]$.

In Figure 4.1 we show an example of a histogram for chromosome 20, 5x coverage, with representative values coloured in red for each bin.

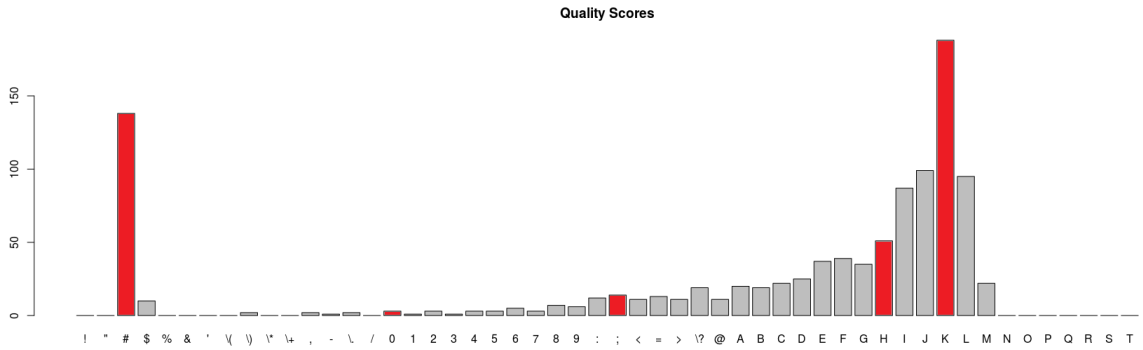


Figure 4.1: Quality values histogram, chromosome 20. Red values are the representatives of each bin.

4.7 Experiments Process

Our purpose with these experiments is to study how we can improve CRAM compression by modifying quality values with four different techniques: QVZ [1], LEON [2], dynamic binning and Illumina binning [3] and also analyze how the lossy models impact on SNP calling.

In order to perform CRAM compression we need as input a SAM file with the quality values modified by each of the techniques to analyze. The general work flow of the experiments can be split into the following steps:

1. The very first step is the alignment of the reads to a reference genome. In this case we use BWA MEM command from BWA software and as for the reference genome we work with genome GRCh37. The output of this alignment will be a SAM file that, from now on, we will refer to as the *original* SAM and the quality values in it will be the *original* or *raw* quality scores, as they are the ones provided for the sequencer, in this case, Ilumina technologies.

2. In the second step we create a SAM file for each one of the lossy models that we are analyzing, in which these new files will have the quality scores updated according to each model. Then we convert each SAM file to CRAM format using samtools in order to compare their *compression ratio*.

    - For *Illumina binning* the process is straight forward and we only modify the code in *htslib* so when converting the BAM file to a CRAM file it will apply the corresponding transformation to each quality value according to Table 3.2.
    - For applying LEON transformation, their software requires a FASTQ file. In this case we create a temporal FASTQ file with the reads and quality values extracted from the *original* SAM. We run LEON algorithm with this created file, then we decompress their output and create the new SAM file with the modified quality values.
    - On the other hand, the input for QVZ algorithm is only a file with the quality scores, with one read per line. This file is created directly by extracting the quality values from the *original* SAM file. Again, once the quality values are processed we decompress the output and create a SAM file with the new quality scores.
    - As for dynamic binning, we read the file twice. In the first round we compute the histogram for each block and we create a dictionary of the representative values for each bin corresponding to each block. In the second round we only apply the transformation to the quality values according to the dictionary created in the first round.

3. Other than only comparing the *compression ratio*, we also intend to compare how each of the techniques impact on SNP calling performance. In this step, we follow *Best Practices* [72], [73] to improve the data.

4. Lastly, we perform the variant calling with two different software tools and compare each one of the lossy models with the mentioned metrics, sensitivity, precision and F-score:

    - From Samtools we use *samtools mpileup* pipeline.
    - From GATK toolkit we perform variant calling with *Haplotype Caller* command.

Chapter 5

Results and Analysis I

In this chapter we present the results of the experiment performed with the data set of reads SRR622461 which corresponds to the individual NA12878 with a fold coverage of 5x. We specifically provide a discussion of the results related to compression ratio and SNP calling performance.

5.1 5x Coverage Experiments

For this part of the experiments we selected the data set of reads SRR622461 which corresponds to the individual NA12878 with a fold coverage of 5x.

We performed the alignment with BWA software, obtaining the SAM file with size 49GB. From this file we extracted only chromosome 11 (2.3 GB) and chromosome 20 (1.0GB). For each of these chromosomes we applied the four lossy techniques to the quality values: Illumina binning, QVZ, LEON and dynamic binning. After modifying the quality values with each one of the techniques and converting each output to proper format (SAM), as explained in Section 4.7, we compressed the resulting SAM file by converting it into a CRAM file.

5.1.1 Compression Ratio

One of the main aspects of this work is to study and analyse how we can improve compression of SAM files when converting them into CRAM files. Tables 5.1 and 5.2, for chromosome 11 and 20 respectively, show a summary of the file sizes, displayed in bytes, for each one of the lossy models as well as for the raw SAM file, i.e. without modifying the quality scores at all.

These tables contain SAM and CRAM file size in order to analyse the compression ratio, which we also include in the last row. For a wider analysis we also integrated the BAM file size to appreciate how even the lossy models for the quality scores also have an impact on the

	Raw	Illumina	QVZ	LEON	Dynamic bin
SAM	2357302048	2357302048	2357302048	2357302048	2357302048
BAM	452482490	363185570	445223159	310363489	340863258
CRAM	168488810	101719649	159262731	72954322	87290416
Compression	13.99	23.17	14.80	32.31	27.00

Table 5.1: Chromosome 11 (5x fold coverage), files size(bytes) and compression ratio

	Raw	Illumina	QVZ	LEON	Dynamic bin
SAM	1037326297	1037326297	1037326297	1037326297	1037326297
BAM	206153342	165303409	195703116	136871744	154713228
CRAM	77390459	46623719	70289403	32371234	39748470
Compression	13.40	22.24	14.75	32.04	26.09

Table 5.2: Chromosome 20 (5x fold coverage), files size(bytes) and compression ratio

BAM format. As we can notice in Table 5.2 all lossy techniques will improve the size of the BAM file, by dropping from 206MB, with raw quality values, to 165MB, 195MB, 136MB and 154MB for each one of the corresponding techniques for chromosome 20. Similar results for chromosome 11 in Table 5.1 show that BAM file size drops from 452MB to 363MB, 445MB, 310MB and 340MB correspondingly.

With regard to the compression ratio we can start by noticing that the CRAM format by itself already performs a very good compression, as displayed in the Raw column. Without adjusting quality values, the compression ratio goes to 13.40 for chromosome 20 and to 13.99 for chromosome 11, which means that SAM file is approximately 13 times larger than the CRAM file in both cases, even when they contain exactly the same information.

As for the set of lossy techniques that we are analysing, all of them prove to have a favorable impact on the compression ratio. According to Table 5.1 and Table 5.2 we observed that, when using Illumina binning, the SAM file is roughly 22 times larger than the CRAM file. For the QVZ technique we have a compression ratio of approximately 14 which, even though, it is very close to the result that we obtain without adjusting quality scores, still represents an improvement for both chromosomes. LEON technique, in both cases, for chromosome 11 and 20, has the best impact on the compression ratio as it reaches more than twice the compression ratio than the one obtained with raw quality values. Lastly, the dynamic binning also reports a considerable improvement in this matter with a compression ratio of around 26 for the pair of chromosomes.

5.1.2 Variant Calling Performance

In all lossy compression models, as obviously expected, there will be a loss of information that we need to be aware of and we need to analyse all kinds of possible impact that this loss of information can have on the data.

In this case we have to consider that quality values are used when performing variant calling in a given alignment. So, when we transform the quality scores, it is anticipated that variant calling results will be influenced by these transformations. In an effort to measure and evaluate these possible changes we performed variant calling with the raw quality values and also with each one of the lossy models that we are studying. We then computed three different scores, as explained in section 4.5, sensitivity, precision and F-score, to evaluate the variant calling performance against two separate ground truth sets of variants.

In Tables 5.3 and 5.4 the evaluation with sensitivity, precision and F-score is summarized for chromosomes 11 and 20 respectively. Each table also contains the results obtained for two different callers, GATK and Samtools, with the purpose of presenting evidence about the fact that these results do not depend on any specific caller. Also, it is worthwhile to mention that these two tables are the result of the experiment executed with the ground truth released by Illumina, nevertheless the same tables are detailed in Appendix A for the same experiment with GIAB-NIST ground truth.

As our main metric for evaluation we will focus on the F-score because this one combines both sensitivity and precision. Otherwise the natural trade-off between sensitivity and precision will not allow us to conclude anything important, for example, we can note in Table 5.3, for chromosome 11, the LEON technique outperforms Raw quality values in sensitivity, 0.7486 *vs* 0.7486, but Raw quality values outperforms the LEON model in precision, 0.9488 *vs* 0.9429. From this we can not argue one is better than the other, but F-score combine both metrics and shows that both of them are very close, although, for Raw quality values F-score is slightly higher, 0.8363 *vs* 0.8346.

What is intended to be presented in these tables is the evidence that, even though the compression ratio is considerably better when applying lossy techniques (more for some than for others), the overall F-score does not change drastically. And what is even more important, these tables show that variant calling performance can even be improved. As an indication of this fact, we can look at chromosome 20 in Table 5.4. With the GATK caller, the F-score for Illumina binning (0.7621) and dynamic binning(0.7606) are both slightly better than the

	GATK			Samtools		
	Sensitivity	Precision	F-Score	Sensitivity	Precision	F-Score
Raw	0.7201	0.9600	0.8229	0.7476	0.9488	0.8363
Illumina	0.7228	0.9605	0.8249	0.7550	0.9472	0.8402
QVZ	0.7200	0.9601	0.8229	0.7476	0.9487	0.8362
LEON	0.7165	0.9598	0.8205	0.7486	0.9429	0.8346
Dynamic bin	0.7206	0.9587	0.8228	0.7483	0.9455	0.8354

Table 5.3: Variant calling performance with Illumina ground truth, chromosome 11 (5x fold coverage).

	GATK			Samtools		
	Sensitivity	Precision	F-Score	Sensitivity	Precision	F-Score
Raw	0.6763	0.8683	0.7604	0.7063	0.8614	0.7762
Illumina	0.6787	0.8688	0.7621	0.7139	0.8606	0.7804
QVZ	0.6514	0.8644	0.7430	0.6770	0.8578	0.7567
LEON	0.6481	0.8629	0.7402	0.6780	0.8508	0.7546
Dynamic bin	0.6775	0.8670	0.7606	0.7072	0.8586	0.7756

Table 5.4: Variant calling performance with Illumina ground truth, chromosome 20 (5x fold coverage).

F-score achieved with the raw quality values (0.7604). In the same table for the Samtools caller we can also notice that Illumina binning F-score (0.7804) outperforms the Raw quality values F-score (0.7762).

For chromosome 11 in Table 5.3 we have similar results. With GATK caller the QVZ F-score is tied with raw quality values F-score, and the Illumina binning F-score is even higher than both of them.

5.1.3 ROC Curve Analysis

In order to go deeper with the analysis and study the behaviour of each lossy technique when varying the threshold to consider a variant as correctly called, we plotted ROC curves. Figures 5.1 and 5.2 display ROC curves for chromosomes 11 and 20 respectively. The results in these plots were based on the Illumina ground truth with the Samtools caller. We can note that at each point the raw quality values curve is always overlapped or dominated by at least one of the other techniques. The clearest message is that the Illumina binning curve(blue) is above the Raw one(red) in both chromosomes. This confirms once again the fact that by changing the quality values with all different techniques, the variant calling performance is not negatively affected.

In Tables 5.5 and 5.6 we also present, for both chromosomes, the area under the curve (AUC) as a metric for each technique and for each caller. From looking at these numbers it

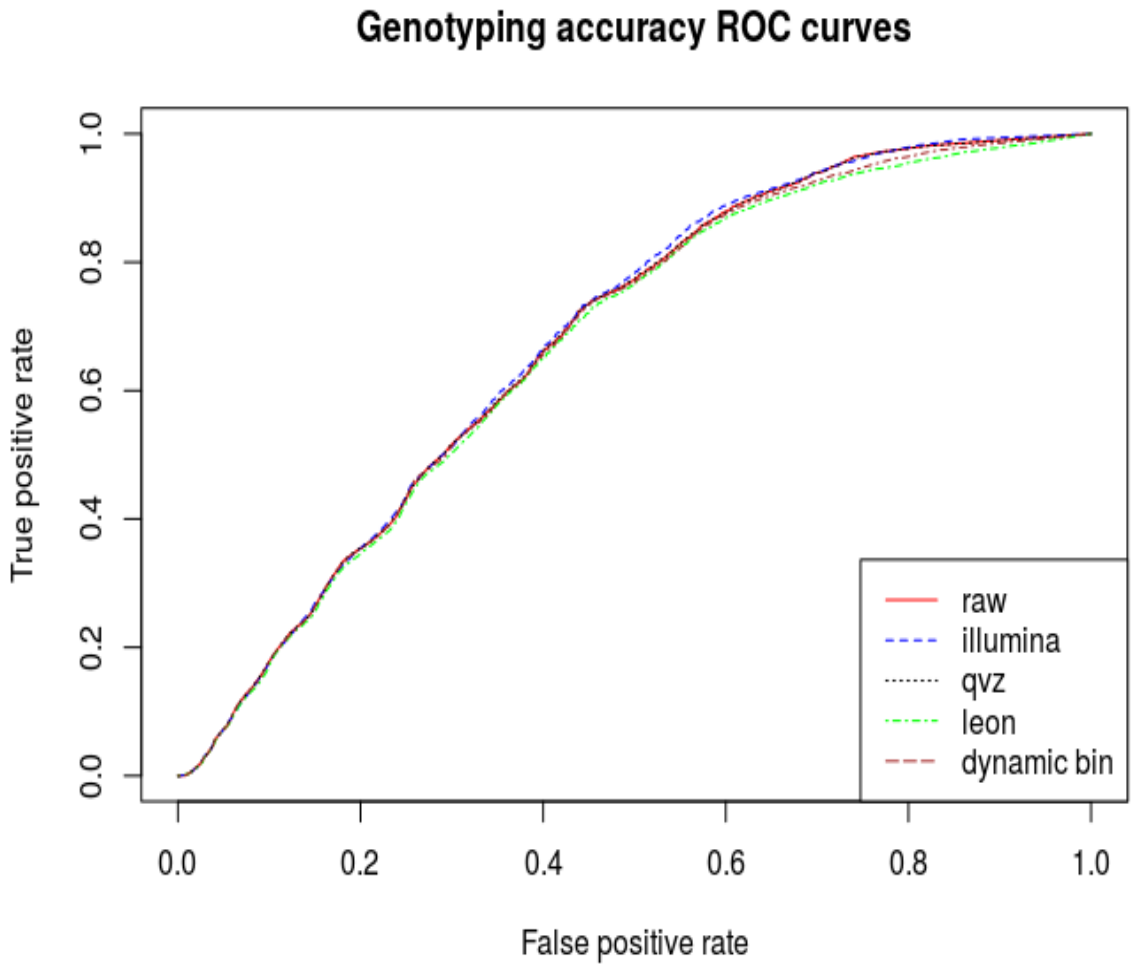


Figure 5.1: ROC,chromosome 11 (5x fold coverage).

is clearer that the Raw quality values curve is dominated in all cases by the Illumina binning curve. Also, in all cases QVZ technique has higher AUC than the raw quality values one. Nevertheless, one important point to discuss about the AUC results is the fact that all of them are very close to 0.5 and therefore it can be argued that the classification is not necessarily good. This is completely reasonable because from a data set with only 5x fold coverage it is not expected to have high accuracy, not even for the alignment and therefore neither for the variants. In Chapter 7 we show the same results for high coverage (50x) and there we can notice the differences. In this chapter we have nonetheless evidence that lossy techniques do not really affect variant calling performance even for low coverages.

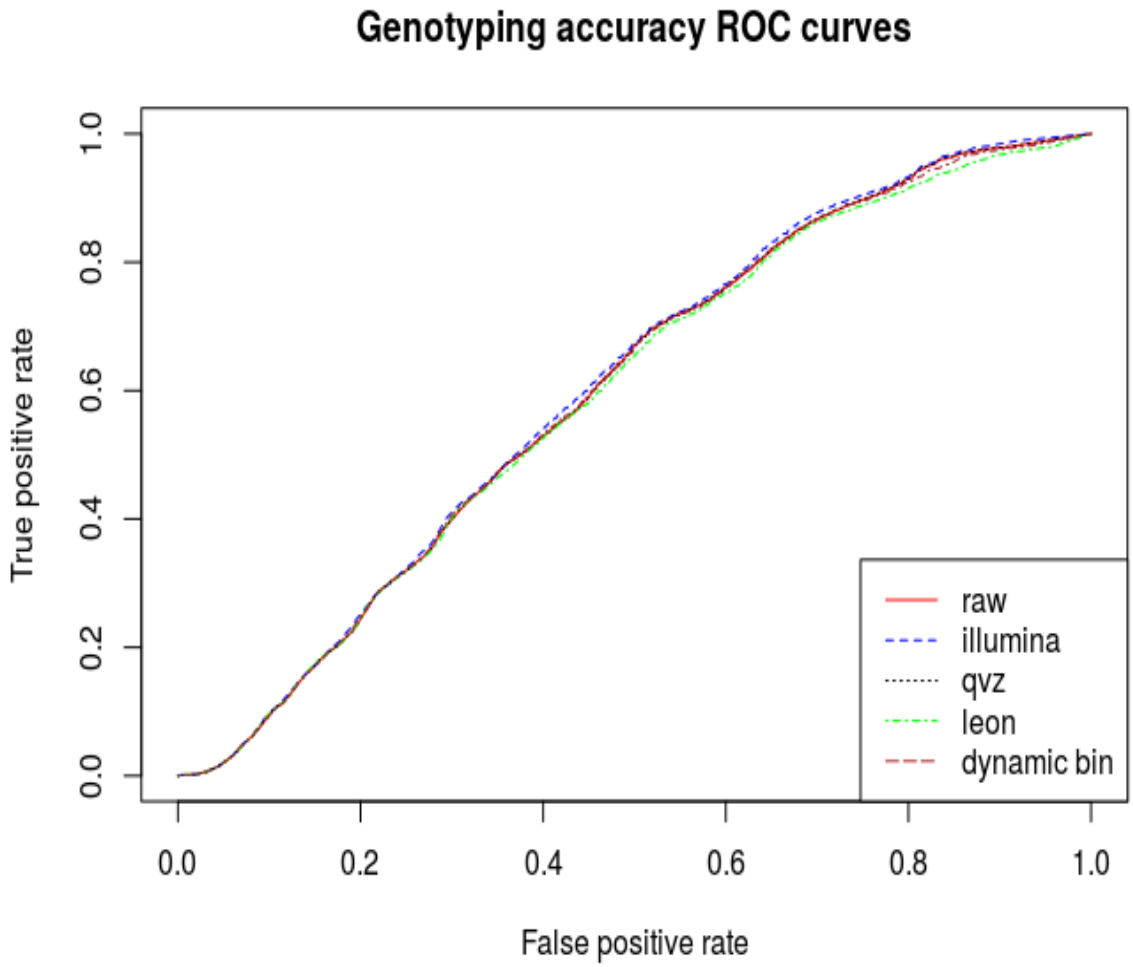


Figure 5.2: ROC,chromosome 20 (5x fold coverage).

	Raw	Illumina	QVZ	LEON	Dynamic bin
GATK	0.5149	0.5395	0.5173	0.5223	0.5105
Samtools	0.6766	0.6800	0.6766	0.6664	0.6724

Table 5.5: AUC, Chromosome 11 (5x fold coverage).

	Raw	Illumina	QVZ	LEON	Dynamic bin
GATK	0.5007	0.5094	0.5012	0.5033	0.4983
Samtools	0.5965	0.6020	0.5967	0.5902	0.5958

Table 5.6: AUC, Chromosome 20 (5x fold coverage).

Chapter 6

Results and Analysis II

In this chapter we report the results of the experiment performed with low coverage (6x) data set of the individual NA12878. We specifically present a discussion about the results obtained for the compression ration and SNP calling performance.

6.1 6x Coverage Experiment

In this section of experiments we used the low coverage (6x) alignment provided by 1000 Genomes Project [78]. They provide the BAM file (16.2GB), for the whole genome, containing the alignment performed with BWA software. This data set also corresponds to the individual NA12878.

The experiments here consisted of applying the four already presented lossy techniques; Illumina binning, QVZ, LEON and dynamic binning, to each SAM file and analysing their behaviour when converting to CRAM format, as well as studying the possible impact in the variant calling performance. In this particular case we run the experiment with the first 22 chromosomes, in order to get better insights with no bias to any specific chromosome. Also, because of storage constraints and in order to have greater control of the results, the experiments were performed for each chromosome separately and results for each one of the chromosomes are in corresponding tables in Appendix B. Below we present only results for chromosome 11 and chromosome 20 in order to have consistency with the discussion in the previous experiment with 5x coverage.

6.1.1 Compression Ratio

In the interest of analysing how each one of the lossy models affects the size of the SAM file when converting it to CRAM format, we present Table 6.1 and Table 6.2, for chromosome 11

and chromosome 20, respectively. These tables contain the file size, including the BAM file size, as well as the compression ratio which is computed as the SAM file size over the CRAM file size.

In both tables we can appreciate how every single lossy technique improves the compression in the BAM file. As we can see, in Table 6.1 with raw qulity values the BAM file is about 705MB while for Illumina it is only 477MB, 663MB for QVZ, 351MB for LEON and 420MB for dynamic binning. We can observe similar results in Table 6.2 for chromosome 20 and for all other chromosomes in Appendix B.

Regarding the CRAM format, we also observe that compression ratio is always better for Illumina, LEON and dynamic binning techniques. For QVZ, the compression ratio is also better but not significantly better. For evidence of this we can look at Table 6.2, for chromosome 20, which shows that the compression ration with raw quality scores is 8.55 and the compression ratio with QVZ technique is only 8.72. If we look at the actual file size we confirm that the improvement goes only from around 164MB to 161MB, while, for instance, with LEON technique having compression ratio of 24.51, the file size decreases to only 57MB.

As a conclusion of the results shown in Tables 6.1 and 6.2 and combined with the same results for all other chromosomes in Appendix B, we realize that the raw quality compression ratio is between 8.0 - 9.0, Illumina compression ratio is between 15.0 - 17.0, QVZ 8.0 - 9.0, LEON 25.0 - 27.0 and dynamic binning compression ratio between 17.0 - 19.0. Roughly speaking, with raw quality values and QVZ techniques the CRAM file will be around 8 times smaller than the original SAM file, for Illumina technique the CRAM file will be around 16 times smaller, while for LEON the compressed file will be approximately 26 times smaller and with dynamic binning it will be 18 times smaller. As a summary for all chromosomes, we present Table 6.3 which contains the total size of all 22 chromosomes when converted to CRAM files after applying each one of the lossy techniques. For appreciation of how well compression is performed with each one of them, we have to consider that the total size of the SAM files goes to 66.11GB. The best size obtained is with LEON technique which reduces all files to a size of 2.58GB.

6.1.2 Variant Calling Performance

After we analyse compression ratio in our experiment, we obtain positive results in the sense that, when applying lossy models for the quality values, we observe considerable improvements regarding the file size. Nevertheless, we still have the necessity of measuring how the

	Raw	Illumina	QVZ	LEON	Dynamic bin
SAM	3323718560	3323718560	3323718560	3323718560	3323718560
BAM	705296161	477193614	663864702	351036712	420715141
CRAM	399026425	214238552	357220844	135091577	174185183
Compression	8.32	15.51	9.30	24.60	19.08

Table 6.1: Chromosome 11(6x fold coverage), files size(bytes) and compression ratio

	Raw	Illumina	QVZ	LEON	Dynamic bin
SAM	1409441215	1409441215	1409441215	1409441215	1409441215
BAM	300302069	205177256	297843978	153751094	189692794
CRAM	164791886	89437663	161609469	57501251	79082759
Compression	8.55	15.75	8.72	24.51	17.82

Table 6.2: Chromosome 20(6x fold coverage), files size(bytes) and compression ratio

Raw	Illumina	QVZ	LEON	Dynamic bin
7.68GB	4.31GB	7.49GB	2.58GB	3.61GB

Table 6.3: Total size after compression, original SAM files size is 66.11GB

adjusted quality values affect the variant calling process, as this task relies broadly on them.

In Table 6.4 for chromosome 11 and Table 6.5 for chromosome 20 we display the score of three different metrics; sensitivity, precision and F-score, for variant calling performance of two different callers; GATK and Samtools, considering Illumina set as ground truth. In Appendix B the results performed against the NIST-GIAB ground truth are also reported.

As mentioned before and to avoid talking about cases in which only one measure, precision or sensitivity, is boosted at the cost of decreasing the other, we focus our attention mainly on the F-score for each case, because this measure combines both of them.

In general, in both chromosomes, and also in all of them in Appendix B, we can observe that F-score with raw quality values is not much different than with all other lossy models, except in the case of QVZ, when using Samtools caller, in which all of them are the same until precision of two decimals. And even for the QVZ case the F-score only decreases from 0.8110 to 0.8069 in chromosome 20 and from 0.8424 to 0.8386 in chromosome 11.

Following with the discussion about improving variant calling performance with some of the lossy models, we can notice that in both chromosomes, with Samtools caller, F-score reports an improvement when using Illumina and LEON technique. For instance in chromosome 11, the raw quality values F-score is 0.8428 while Illumina technique reports a F-score of 0.8444 and LEON a higher F-score of 0.8482.

	GATK			Samtools		
	Sensitivity	Precision	F-Score	Sensitivity	Precision	F-Score
Raw	0.7137	0.9856	0.8279	0.7388	0.9809	0.8428
Illumina	0.7123	0.9881	0.8278	0.7414	0.9805	0.8444
QVZ	0.7135	0.9814	0.8263	0.7339	0.9780	0.8386
LEON	0.7108	0.9858	0.8260	0.7511	0.9742	0.8482
Dynamic bin	0.7125	0.9859	0.8272	0.7353	0.9810	0.8406

Table 6.4: Variant calling performance with Illumina ground truth, chromosome 11 (6x fold coverage).

	GATK			Samtools		
	Sensitivity	Precision	F-Score	Sensitivity	Precision	F-Score
Raw	0.6687	0.9408	0.7951	0.6963	0.9710	0.8110
Illumina	0.6661	0.9826	0.7940	0.7010	0.9709	0.8142
QVZ	0.6687	0.9755	0.7935	0.6917	0.9680	0.8069
LEON	0.6656	0.9804	0.7929	0.7037	0.9658	0.8142
Dynamic bin	0.6673	0.9807	0.7942	0.6929	0.9716	0.8089

Table 6.5: Variant calling performance with Illumina ground truth, chromosome 20 (6x fold coverage).

6.1.3 ROC Curve Analysis

In our previous analysis with sensitivity, precision and F-score metrics we considered every single variant in an output VCF file as a correctly called variant. For this analysis we want to study the behaviour that each lossy technique would present when setting different thresholds for considering each single variant called correctly in an output VCF file.

Figures 6.1 and 6.2 display all ROC curves for each one of the lossy techniques when performing variant calling with Samtools pipeline and considering Illumina variant set as ground truth. In the same plots we also display the ROC curve for the case of variant calling with raw quality scores. In both figures we can observe how the Illumina binning curve (blue) outperforms the raw quality values one (red) and also, LEON (green) is visibly above if looking a critical points. For instance, in Figure 6.2 when FPR is around 0.4 the rate for true positives is considerably better for Illumina and LEON, and as for QVZ and dynamic binning, both of them overlap the one representing the raw quality values. But in general we can appreciate how variant calling performance is not negatively affected and even for some cases it is improved.

	Raw	Illumina	QVZ	LEON	Dynamic bin
GATK	0.6990	0.7492	0.6991	0.7120	0.6923
Samtools	0.8257	0.8320	0.8262	0.8273	0.8200

Table 6.6: AUC, Chromosome 11 (6x fold coverage).

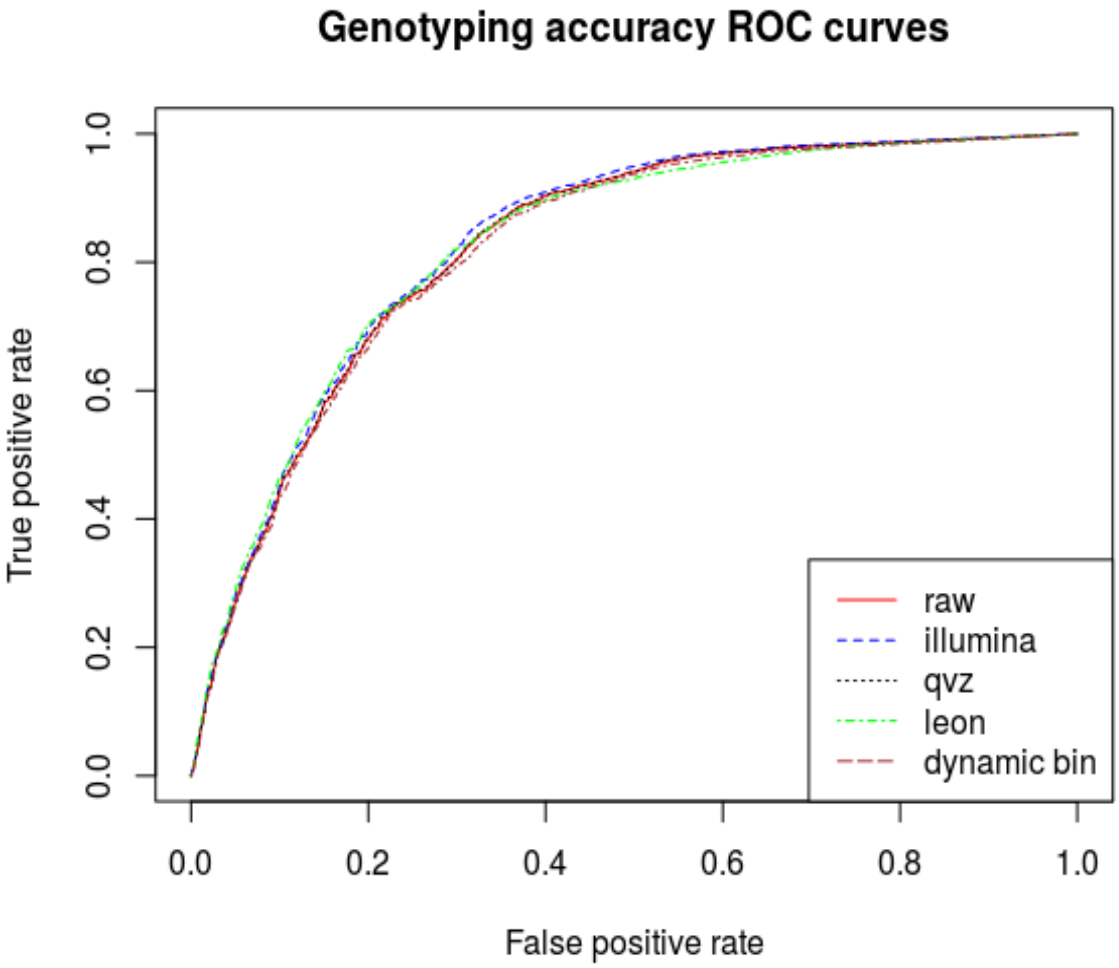


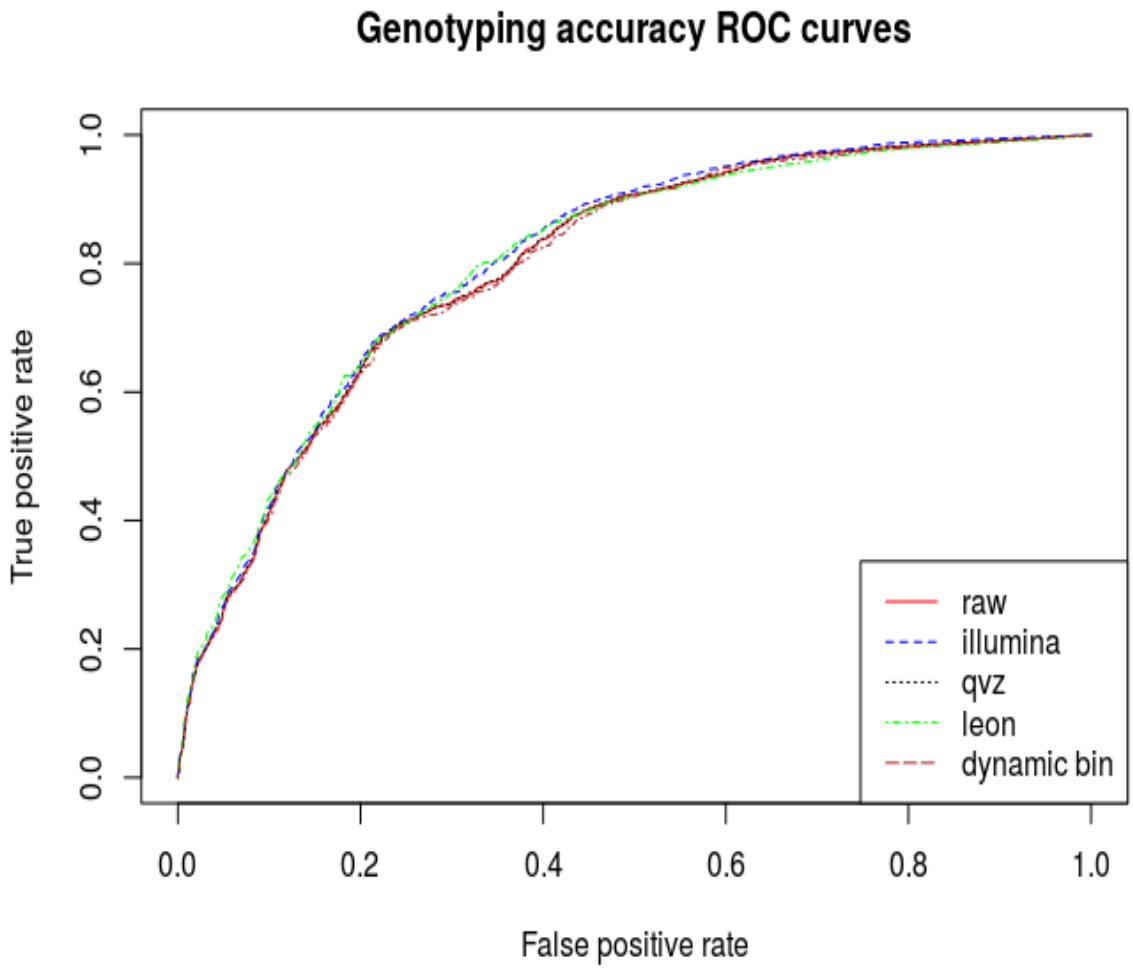
Figure 6.1: ROC,chromosome 11 (6x fold coverage).

	Raw	Illumina	QVZ	LEON	Dynamic bin
GATK	0.6880	0.7195	0.6890	0.7005	0.6857
Samtools	0.7988	0.8068	0.7991	0.8028	0.7950

Table 6.7: AUC, Chromosome 20 (6x fold coverage).

In Tables 6.6 for chromosome 11 and in Table 6.7 for chromosome 20 we present the values for area under the curve (AUC) for the plots in Figures 6.1 and 6.2 correspondingly. In these two tables it becomes more evident how Illumina, LEON and QVZ techniques outperform the case with the raw quality values.

Also in these two tables and comparing with the previously discussed experiment for 5x fold coverage we can observe that with this data set (6x) we obtain better accuracy for variant calling, which is not surprising because the more coverage, the more likely we are to have the correct information about each one of the bases and therefore to find the variants.



5

Figure 6.2: ROC,chromosome 20 (6x fold coverage).

Chapter 7

Results and Analysis III

In this chapter we report the results of the experiment performed with the high coverage (50x) data set of the individual NA12878. We specifically present a discussion about the results obtained for the compression ratio and SNP calling performance.

7.1 High Coverage (50x) Experiment

As in the previous experiments, the chromosomes here analyzed are part of the individual NA12878.

The alignment performed with BWA is provided by the 1000 Genomes project [78] and in this case the fold coverage for the genome is of 50x. The BAM file for this data set is 254 GB, therefore, due to storage constraints, we only performed all experiments on chromosome 11 and chromosome 20. As shown in the experiment performed with the whole genome with 6x coverage, the behaviour in all other chromosomes is expected to be similar in terms of compression ratio.

With regards to the size of the data set, in this case the SAM files are 44.7 GB for chromosome 11 and 20.0 GB for chromosome 20.

7.1.1 Compression Ratio

For analysing the compression ratio and the influence that each lossy technique has on the BAM and the CRAM files we present Table 7.1 for chromosome 11 and Table 7.2 for chromosome 20, summarizing the sizes of the files in bytes. Last row in these tables display the compression ratio which is our main target for discussion.

According to these tables the BAM files are also affected positively by reducing the size. For chromosome 11, See Table 7.1, for instance, with raw quality values the BAM file is 10.6

	Raw	Illumina	QVZ	LEON	Dynamic bin
SAM	44741080610	44741080610	44741080610	44741080610	44741080610
BAM	10668938373	8830246318	10531487823	7883319664	8680171268
CRAM	6684954326	5297101246	6527041230	4726468838	5191134746
Compression	6.69	8.44	6.85	9.46	8.61

Table 7.1: Chromosome 11 (50x fold coverage), files size(bytes) and compression ratio

	Raw	Illumina	QVZ	LEON	Dynamic bin
SAM	19978222245	19978222245	19978222245	19978222245	19978222245
BAM	4763279583	3951732501	4703215963	3529714410	3881179898
CRAM	2982788014	2370470617	2913269311	2116033897	2321367241
Compression	6.69	8.42	6.85	9.44	8.60

Table 7.2: Chromosome 20 (50x fold coverage), files size(bytes) and compression ratio

GB and with each technique this size goes lower: 8.8 GB with the Illumina binning, 10.5 GB with the QVZ technique, 7.8 GB with LEON and 8.6 GB with dynamic binning. In this case we can notice that once again the QVZ model is the one which makes the least improvement in the size.

Concerning the compression ratio we can observe that in both chromosomes the results are pretty similar. If we only consider one decimal precision the compression ratio is 6.6 when using the original quality values, 8.4 with the Illumina binning, 6.8 with the QVZ technique, 9.4 with LEON model and 8.6 with dynamic binning. From this we can note that applying the LEON technique to the quality values we obtain better compression ratio than the others. As for Illumina binning and dynamic binning we can point out that results for both of them are close in terms of compression ratio, for example for chromosome 11 the CRAM file is reduced to 5.29 GB when applying Illumina binning and to 5.19 GB when applying dynamic binning.

Although with each one of the lossy models that we are studying we obtain an improvement in the compression, we can note that in this experiment with high coverage, the compression ratio in general is lower than with the one obtained with 5x and 6x coverage. Roughly speaking, the compression ratio for high coverage case is half that obtained with 6x coverage and this property holds for the file with raw quality values as well as for the ones with each lossy model applied. Another thing to note is that the pair of compression ratio values belonging to each chromosome and obtained with each technique, and even for the raw file, they are pretty similar and we would expect them to be so in every chromosome.

	GATK			Samtools		
	Sensitivity	Precision	F-Score	Sensitivity	Precision	F-Score
Raw	0.9778	0.9669	0.9723	0.9785	0.9590	0.9686
Illumina	0.9752	0.9718	0.9735	0.9791	0.9588	0.9688
QVZ	0.9778	0.9668	0.9723	0.9781	0.9592	0.9685
LEON	0.9776	0.9634	0.9704	0.9815	0.9512	0.9661
Dynamic bin	0.9777	0.9668	0.9722	0.9783	0.9580	0.9680

Table 7.3: Variant calling performance with Illumina ground truth, chromosome 11 (50x fold coverage)

7.1.2 Variant Calling Performance

For analysing the impact that adjusted quality values can have in the variant calling process we present Tables 7.1 and 7.2 for chromosome 11 and chromosome 20 respectively. These tables summarize the variant calling performance scores: sensitivity, precision and F-score for each lossy model as well as for the variant calling performed with the original quality values. The column (Raw) contains the scores obtained when quality scores are not modified. Also it is important to mention that the scores displayed there are for each caller, GATK and Samtools pipeline. For the results in these tables we considered Illumina ground truth. For the same results with GIAB-NIST ground truth refer to Appendix C. The tables in the Appendix also contain the actual number of true positives, false positives and false negatives in more detail.

By looking at the F-scores in each case we provide evidence that in general the variant calling performance is not affected by any of the lossy techniques. For instance, in chromosome 11, Table 7.1 and when using GATK caller we note that for the case when the quality values are not adjusted, the F-score obtained is 0.9723. In comparison the performance when previously applying Illumina binning to the quality values gives a F-score of 0.9735, for the QVZ technique the score is 0.9723, for the LEON model 0.9704 and for dynamic binning 0.9722. There are two facts to highlight in this. Firstly, according to the F-score measure, the variant calling performance, when applying Illumina binning to the quality scores, is improved: the F-score is: 0.9723 for the no modified quality values and 0.9735 for the Illumina binning. The other fact to highlight is that when the LEON model is applied to the quality values, the F-score obtained is the lowest one and this evidence holds for both chromosomes and both callers. It is important to mention this because as for the compression ratio results, the LEON model reports a considerably better compression ratio than the other techniques. We can see this fact as an evidence of the natural trade-off between compression ratio and variant calling performance.

	GATK			Samtools		
	Sensitivity	Precision	F-Score	Sensitivity	Precision	F-Score
Raw	0.9549	0.9586	0.9568	0.9577	0.9469	0.9523
Illumina	0.9513	0.9653	0.9582	0.9586	0.9469	0.9527
QVZ	0.9550	0.9587	0.9569	0.9581	0.9465	0.9522
LEON	0.9547	0.9549	0.9548	0.9622	0.9369	0.9494
Dynamic bin	0.9548	0.9586	0.9567	0.9579	0.9459	0.9518

Table 7.4: Variant calling performance with Illumina ground truth, chromosome 20 (50x fold coverage).

7.1.3 ROC Curve Analysis

In Figure 7.1 and Figure 7.2 for chromosome 11 and for chromosome 20 respectively we display the ROC curves which correspond to the experiment with Illumina ground truth and Samtools caller. With these plots we intend to analyse the behaviour of each one of the techniques when having different thresholds for recognizing a variant as correctly called. In general we can appreciate that with high coverage the variant calling performance is considerably better than the one obtained with 5x and 6x coverage.In both images, when false positive rate is around 0.2, the true positive rate is above 0.8 already.

As for comparing each one of the lossy techniques versus the curve when having the original quality values, by looking at Figure 7.1 for chromosome 11, we observe that before all of them converge (i.e before false positive rate of 0.2) the curve corresponding to the LEON technique(green) is below the one corresponding to the raw quality values. As for the Illumina technique(blue), the curve is above the one for the raw quality values, which confirms the slight improvement previously noted with the F-score as well for the Illumina binning. In the same image we see that both the QVZ technique and dynamic binning mostly overlap the performance with raw quality values. This we consider to be a good result because this implies that the variant calling performance is not negatively affected even though both of them achieved a better compression ratio than the one obtained with the original quality scores.

For summarizing the information from ROC curves, we consider the AUC as another metric for the variant calling performance. In Table 7.5 for chromosome 11 and in Table 7.6 for chromosome 20 we report the AUC values for both callers, GATK and Samtools.

By comparing the AUC for each technique we can notice that there is no indication that lossy models affect the variant calling performance as all of them are either greater than the AUC with raw quality values or significantly closer.

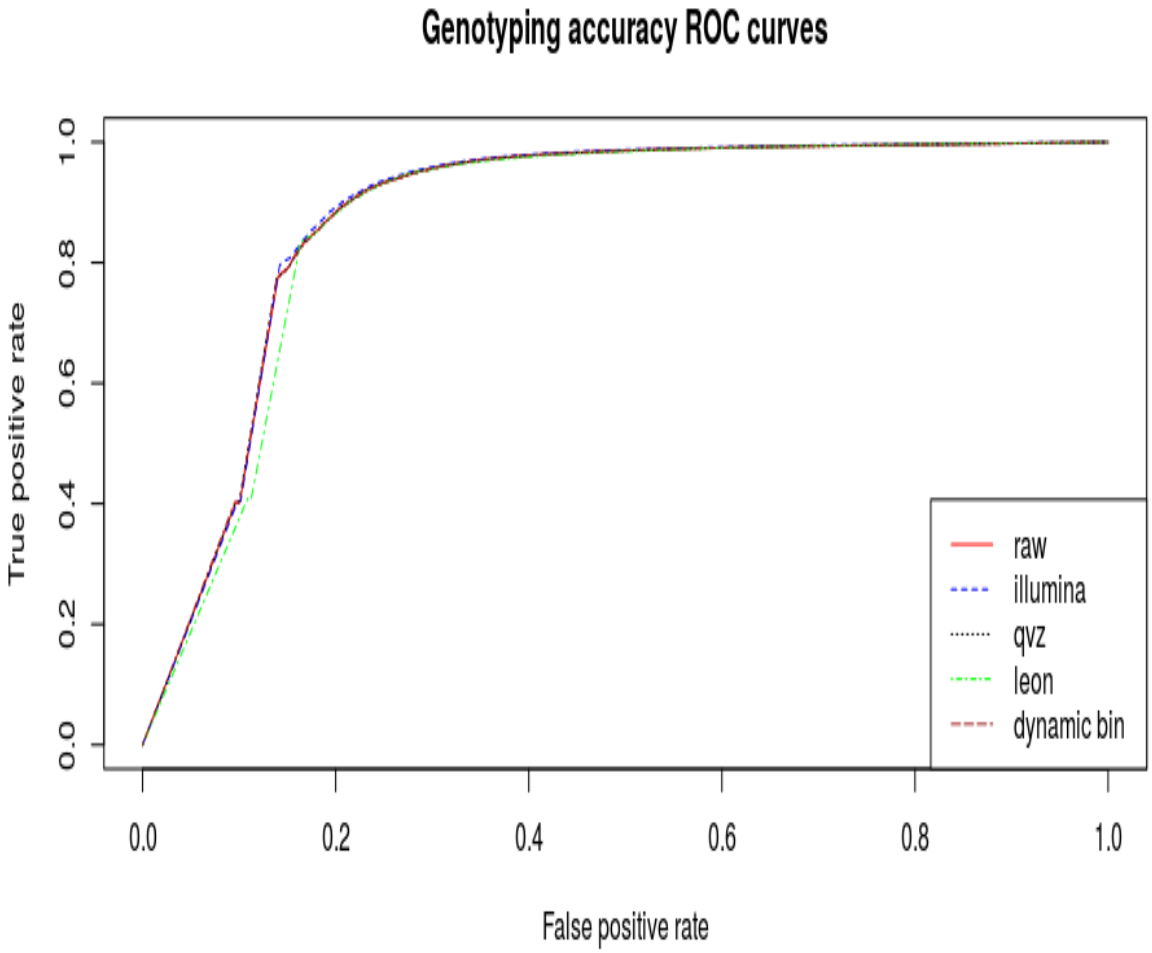


Figure 7.1: ROC,chromosome 11 (50x fold coverage).

	Raw	Illumina	QVZ	LEON	Dynamic bin
GATK	0.8205	0.8439	0.8192	0.8250	0.8191
Samtools	0.8792	0.8804	0.8797	0.8713	0.8797

Table 7.5: AUC, Chromosome 11 (50x fold coverage).

	Raw	Illumina	QVZ	LEON	Dynamic bin
GATK	0.7858	0.8129	0.7878	0.7946	0.7865
Samtools	0.8952	0.8988	0.8954	0.8926	0.8954

Table 7.6: AUC, Chromosome 20 (50x fold coverage).

7.2 Discussion and Analysis

According to the three experiments that we have presented with different coverage data sets, the four lossy techniques that we described in Section 4.7 help to improve the compression ratio. Nevertheless, a general behaviour that we can note is that the compression ratio will be

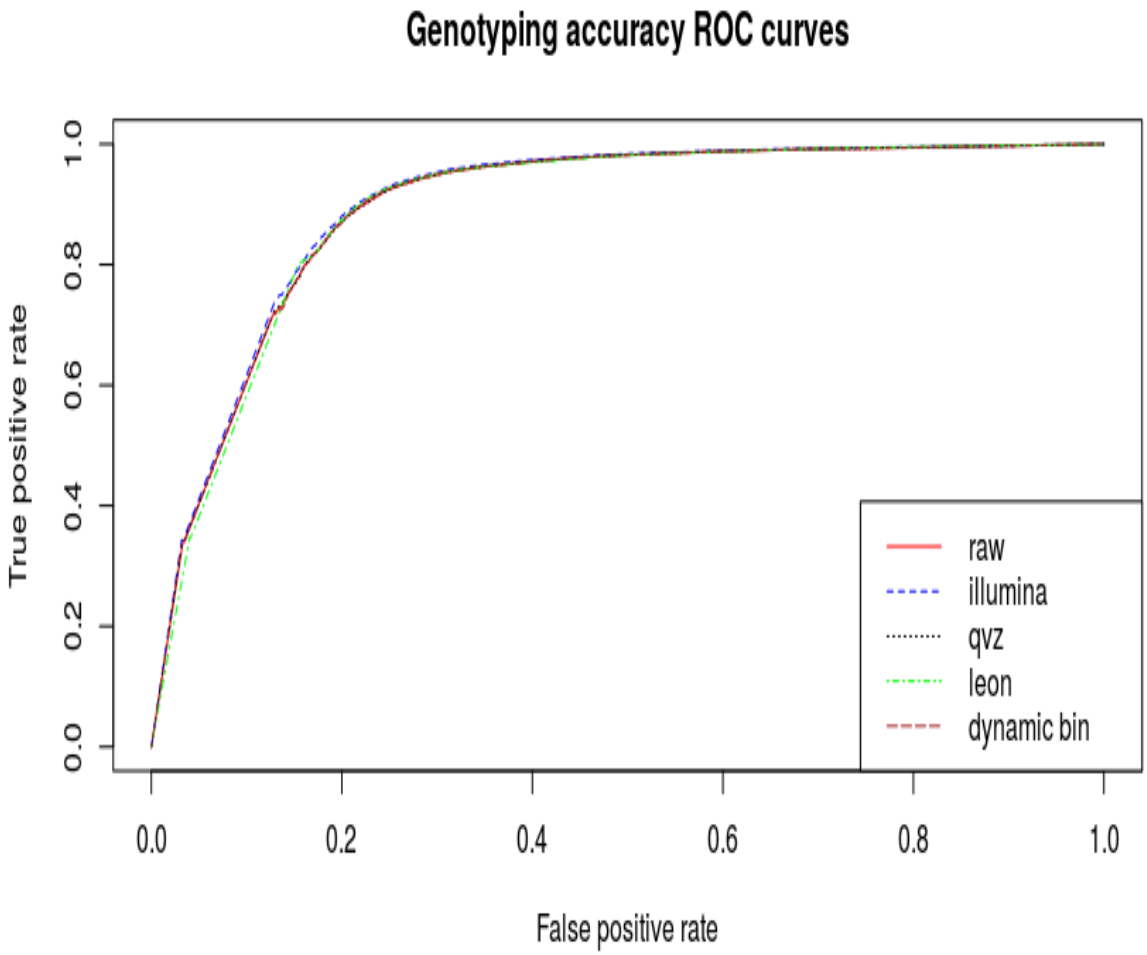


Figure 7.2: ROC,chromosome 20 (50x fold coverage).

lower when the coverage is higher, which means that in this case more information implies higher entropy but also implies better variant calling performances as was confirmed with the F-score and AUC measures. For instance, for chromosome 20 with raw quality values the compression ratio varies from 13.40 with 5x coverage to 8.32 with 6x coverage and to 6.69 with high coverage. And for each lossy model, the compression ratio varies proportionally.

In terms of compression ratio the LEON technique reports the best results by more than double the compression ratio obtained with raw quality values in the case of 5x coverage and 6x coverage. As for the experiment with a high coverage, the LEON technique also has the better compression ratio but the improvement only goes from 6.69 to 9.44 for chromosome 11 and from 6.69 to 9.46 for chromosome 20. Nevertheless with regards to variant calling performance, except in the case with Samtools and 6x coverage, the LEON technique reports a lower F-score than any of the other techniques.

Another general behaviour that we can note is that the QVZ technique in all the experiments showed the lowest compression ratio of all lossy techniques and is only slightly above the compression ratio obtained with the raw quality values. For the high coverage experiment the compression ratio improvement is from 6.69 obtained with raw quality values to 6.85 with the QVZ. Although it is worthwhile to mention that the variant calling performance remained considerably close to the performance obtained with no adjusted quality scores.

With regards to Illumina binning and dynamic binning, both improved the compression ratio without compromising variant calling performance. In general dynamic binning obtained a better compression ratio than the one achieved by Illumina binning. And although with dynamic binning the improvement of variant calling performance was possible, this can be confirmed in the experiment with 5x fold coverage and GATK caller with chromosome 20, Illumina binning proved to be more consistent in terms of boosting variant calling performance as can be seen in Tables 7.3 and 7.4.

At this point of the work we have shown that by adjusting quality values the compression ratio can be improved without compromising the SNP calling performance. To demonstrate that quality values are still necessary and that setting all of them to a constant value will lead to a very poor SNP calling performance we present Table 7.7 where we display the comparison of the F-scores achieved against the raw quality values. In this table we observe that when setting all quality values to a constant value the F-score drops drastically, indicating that the information provided by the quality scores is still necessary.

	5x	6x	50x
Raw	0.7762	0.8237	0.9523
No quality values	0.7180	0.7682	0.8796

Table 7.7: F-score comparison for no quality values. Chromosome 20, ground truth Illumina and Samtools caller.

Chapter 8

Conclusions and Future Work

In this work we have presented the problem that researchers in the field of bioinformatics are facing today regarding the storage space for next-generation sequencing output data files. We discussed the importance of DNA sequencing for humanity as well as the main technological challenges associated with this topic. Around the world there are several important projects for DNA sequencing and today, after just defeating the bottle neck that the costs of DNA sequencing represented for the last ten years, researchers are confronting another bottle neck in terms of storage space due to the large amount of data that is constantly generated everyday and that is expected to grow exponentially within the next few years.

We presented the main approaches for compression of next-generation DNA sequencing output data files and we analysed the most recent studies in the literature related to this problem. We specifically introduced the quality values which comprise roughly half of the file in the most common output formats used for reporting DNA sequencing and DNA alignment, FASTQ and SAM formats. We explained the challenges that these quality scores represent for the compression of SAM and FASTQ file formats. Also, we addressed the influence that quality values have in the variant calling process with the purpose of analysing lossy models for compression of such quality scores. We studied four different lossy techniques for quality scores, Illumina binning, the Qvz model, the LEON algorithm and the dynamic binning. In particular we analysed the effect that each one of these techniques introduces to the CRAM compression format. In our experiments performed with three data sets, each with different fold coverage and all belonging to the individual NA12878, we found that each one of the lossy techniques improved the compression ratio of the SAM files. The LEON technique is the one that achieved the best compression ratio, nevertheless the variant calling performance improved when using either Illumina binning or dynamic binning. Both Illumina and dynamic

binning also achieved a considerable improvement in the compression ratio and concerning the variant calling performance, the results showed that a boost with respect to the results obtained with raw quality values was possible. We also showed that the lack of information provided by the quality values leads to very poor variant calling performance.

Nowadays many approaches for compression of next-generation DNA sequencing data output are being studied. We believe that lossy techniques can be a considerable insight in this area, as it has been proved that it is possible that variant calling performance remains the same or even better in some cases when adjusting the quality values.

Although lossy models for compression of quality scores have proved to be a convenient option for tackling the problem regarding storage space for next-generation output data, there remain several studies to continue improving these techniques. The noise that quality values present in their distribution needs to be well understood by making further analysis of quality values behaviour and statistics.

We also believe that rather than finding a standard and unique lossy technique it is more likely to develop several options which the user will be able to select depending on the project as we observed different behaviours in our experiments by varying only the fold coverage. And still there are many other factors to vary such as, for example sequencing technology.

For the exploratory idea we developed with dynamic binning, there are several other paths to continue studying such as the number of bins as well as different block lengths.

Bibliography

[1] Malysa G., Hernaez M., Ochoa I., Rao M., Ganesan K., and Weissman T. Qvz: lossy compression of quality values. *Bioinformatics*, 31(19):3122–3129, October 2015.

[2] Benoit G., Lemaitre C., Lavenier D., Drezen E., Dayris T., Uricaru R., and Rizk G. Reference-free compression of high throughput sequencing data with a probabilistic de bruijn graph. *BMC Bioinformatic*, 16:288, April 2015.

[3] Illumina White Paper: Informatics. Reducing whole-genome data storage footprint.

[4] The structure of dna. *https://commons.wikimedia.org/wiki/File:DNA-structure-and-bases.png*. Accessed: 2016-06-15.

[5] Genomics Education Programme. Gene structure. *https://commons.wikimedia.org/wiki/File:Gene_structure_(13080962024).jpg*. Accessed: 2016-06-15.

[6] Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., and et al Efron M.J. Big data: Astronomical or genomical? *PLoS Biol*, 13(7):e1002195, 2015.

[7] Wetterstrand K.A. Dna sequencing costs: Data from the nhgri genome sequencing program (gsp). *www.genome.gov/sequencingcostsdata*. Accessed: 2016-06-15.

[8] Cánovas R., Moffat A., and Turpin A. Lossy compression of quality scores in genomic data. *Bioinformatics*, 30(15):2130–2136, August 2014.

[9] Yu Y.W, Yorukoglu D., and Berger B. Traversing the k-mer landscape of ngs read datasets for quality score sparsification. *Research in Computational Molecular Biology. RECOMB 2014. Lectures Notes in Computer Science*, 8394:385–399, 2014.

[10] Ochoa I., Asnani H., Bharadia D., Chowdhury M., Weissman T., and Yona G. Qualcomp: a new lossy compressor for quality scores based on rate distortion theory. *BMC Bioinformatics*, 14:187, June 2013.

[11] Yu Y.W, Yorukoglu D., Peng J., and Berger B. Quality score compression improves genotyping accuracy. *Nature Biotechnology*, pages 240–243, 2015.

[12] Bioinformatics, n. OED Online. Oxford University Press, December 2016. Web. 2 January 2017.

[13] Pennisi E. Will Computers Crash Genomics? *Science*, 331(6018):666–668, February 2011.

[14] Watson J.D and Crick F.H.C. A structure for deoxyribose nucleic acid. *Nature*, (171):737–738, 1953.

[15] Watson J.D and Crick F.H.C. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, (171):964–967, 1953.

[16] Gene, n.2. OED Online. Oxford University Press, December 2016. Web. 2 January 2017.

[17] Pertea M. and Salzberg S.L. Between a chicken and a grape: estimating the number of human genes. *Genome Biology*, 11(5):206, May 2010.

[18] Sanger F., Nicklen S., and Coulson A.R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA*, 74(12):5463–5467, December 1970.

[19] Bennett S. Solexa ltd. *Pharmacogenomics*, 5(4):433–8, June 2004.

[20] Rothberg J.M., Hinz W., Rearick T.M., Schultz J., Mileski W., Davey M., Leamon J.H., Johnson K., Milgrew M.J., Edwards M., Hoon J., Simons J.F., Marran D., Myers J.W., Davidson J.F., Branting A., Nobile J.R., Puc B.P., Light D., Clark T.A., Huber M., Branciforte J.T., Stoner I.B., Cawley S.E., Lyons M., Fu Y., Homer N., Sedova M., Miao X., Reed B., Sabina J., Feierstein E., Schorn M., Alanjary M., Dimalanta E., Dressman D., Kasinskas R., Sokolsky T., Fidanza J.A, Namsaraev E., McKernan K.J., Williams A., Roth G.T., and Bustillo J. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475:348–352, July 2011.

[21] Erika Check Hayden. Nanopore genome sequencer makes its debut. *Nature News*, February 2012.

[22] Noble I. Human genome finally complete. *BBC News*, April 2003. Retrieved February 2017.

[23] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, October 2010.

[24] 1000 Genomes Project Consortium, Abecasis G.R., Auton A., Brooks L.D., DePristo M.A., Durbin R.M., Handsaker R.E., Kang H.M., Marth G.T., and McVean G.A. An integrated map of genetic variation from 1,902 human genomes. *Nature*, 491:56–65, November 2012.

[25] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:68–74, October 2015.

[26] Sudmant P.H., Rausch T., Gardner E.J., Handsaker R.E., Abyzov A., Huddleston J., Zhang Y., Ye K., Jun G., Hsi-Yang Fritz M., Konkel M.K., Malhotra A., Stütz A.M., Shi X., Paolo Casale F., Chen J., Hormozdiari F., Dayama G., Chen K., Malig M., Chaisson M., Walter K., Meiers S., Kashin S., Garrison E., Auton A., Lam H.Y., Jasmine Mu X., Alkan C., Antaki D., Bae T., Cerveira E., Chines P., Chong Z., Clarke L., Dal E., Ding L., Emery S., Fan X., Gujral M., Kahveci F., Kidd J.M., Kong Y., Lameijer E.W., McCarthy S., Flicek P., Gibbs R.A., Marth G., Mason C.E., Menelaou A., Muzny D.M., Nelson B.J., Noor A., Parrish N.F., Pendleton M., Quitadamo A., Raeder B., Schadt E.E., Romanovitch M., Schlattl A., Sebra R., Shabalin A.A., Untergasser A., Walker J.A., Wang M., Yu F., Zhang C., Zhang J., Zheng-Bradley X., Zhou W., Zichner T., Sebat J., Batzer M.A., McCarroll S.A., 1000 Genomes Project Consortium, Mills R.E., Gerstein M.B., Bashir A., Stegle O., Devine S.E., Lee C., Eichler E.E., and Korbel J.O. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526:75–81, October 2015.

[27] Clarke L., Zheng-Bradley X., Smith R., Kulesha E., Xiao C., Toneva I., Vaughan B., Preuss D., Leinonen R., Shumway M., Sherry S., Flicek P., and The 1000 Genomes Project Consortium. The 1000 Genomes Project: data management and community access. *Nature Methods*, 9:459–462, April 2012.

[28] Platinum genomes. *http://www.illumina.com/platinumgenomes/*. Accessed: 2016-06-15.

[29] Li J.Y, Wang J., and Zeigler R.S. The 3,000 rice genomes project: new opportunities and challenges for future rice research. *Genome medicine*, 3:8, 2014.

[30] Zhu J. A year of great leaps in genome research. *Genome medicine*, 4(1):4, 2012.

[31] Khalid Sayood. *Introduction to Data Compression, Third Edition*. Morgan Kaufmann Publishers Inc., 2005.

[32] David A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, September 1952.

[33] IJ. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 2003.

[34] M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. 1994.

[35] Ian H. Witten, Radford M. Neal, and John G. Cleary. Arithmetic coding for data compression. *Communications of the ACM*, 30(6):520–540, June 1987.

[36] S.Golomb. Run-length encodings. *IEEE Transactions on Information Theory*, 12(3):399–401, July 1966.

[37] Stéphane Grumbach and Fariza Tahi. Compression of dna sequences (extended abstract). 1994.

[38] Duc Cao M., Dix T.I., Allison L., and Mears C. A simple statistical algorithm for biological sequence compression. *Proc. Data Compression Conference*, pages 43–52, 2007.

[39] Matsumoto T., Sadakane K., and Imai H. Biological sequence compression algorithms. *Genome Informatics*, 11(2000):43–42, July 2011.

[40] Stéphane Grumbach and Fariza Tahi. Compression of dna sequences. *In: Proceedings of the 1993 IEEE Data Compression Conference*, pages 340–350, 1993.

[41] Stéphane Grumbach and Fariza Tahi. A new challenge for compression algorithms. *Genet Seq Inform Process Manag*, 30:875–886, 1994.

[42] Chen X., Kwong S., and Li M.A. A compression algorithm for dna sequences and its applications in genome comparison. *Genome Informat Ser*, 10:51–61, 1999.

[43] Chen X., Li M., and et al Ma B. Dnacompress:fast and effective dna sequence compression. *Bioinformatics*, 18:1696–1698, 2002.

[44] Tabus I., Korodi G., and Rissanen J. Dna sequence compression using the normalized maximum likelihood model for discrete regression. In *Data Compression Conference, 2003. Proceedings. DCC 2003*, pages 253–262, March 2003.

[45] The sam format specification. *http://samtools.github.io/hts-specs/SAMv1.pdf*. Accessed: 2016-06-15.

[46] McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., and DePristo M.A. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *GENOME RESEARCH*, 20:1297–1303, 2010.

[47] Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009.

[48] Garrison E. and Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN] 2012.

[49] Cock P.J.A., Fields C.J., Goto N., Heuer M.L., and Rice P.M. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Research*, 38(6):1767–1771, April 2010.

[50] Afify H., Islam M., and Wahed M.A. Dna lossless differential compression algorithm based on similarity of genomic sequence database. *International Journal of Computer Science & Information Technology (IJCSIT)*, 3(4):145–154, August 2011.

[51] Bhola V., Bopardikar A.S., Narayanan R., and Lee K. et al. No-reference compression of genomic data stored in fastq format. *International Journal of Computer Science & Information Technology (IJCSIT)*, pages 147 – 150, November 2011.

[52] Deorowicz S. and Grabowski S. Robust relative compression of genomes with random access. *Bioinformatics*, 27(21):2979–2986, September 2011.

[53] Wandelt S. and Leser U. Adaptive efficient compression of genomes. *Algorithms for Molecular Biology*, 7(1):30, October 2012.

[54] W. Timothy, J. White, and Michael D. Hendy. Compressing dna sequence databases with coil. *BMC Bioinformatics*, 9:242, May 2008.

[55] Bonfield J.K. and Mahoney M.V. Robust relative compression of genomes with random access. *PLoS One*, 8:e59190, March 2013.

[56] Daily K., Rigor P., Christley S., Xie X., and Baldi P. Data structures and compression algorithms for high-throughput sequencing technologies. *BMC Bioinformatics*, 11:514, October 2010.

[57] Hsi-Yang Fritz M., Leinonen R., Cochrane G., and Birney E. Efficient storage of high throughput dna sequencing data using reference-based compression. *Genome Research*, 21(5):734–740, May 2011.

[58] Popitsch N. and von Haeseler A. Ngc: lossless and lossy compression of aligned high-throughput sequencing data. *Nucleic Acids Research*, 41(1):e27, January 2013.

[59] Kozanitis C., Saunders C., Kruglyak S., Bafna V., and Varghese G. Compressing genomic sequence fragments using slimgene. *Journal of computational biology*, 18(3):401–413, March 2008.

[60] Jones D.C., Ruzzo W.L., Peng X., and Katze M.G. Compression of next-generation sequencing reads aided by highly efficient de novo assembly. *Nucleic Acids Research*, 40(22):e171, December 2012.

[61] Hach F., Numanagic I., Alkan C., and Sahinalp S.C. Scalce: boosting sequence compression algorithms using locally consistent encoding. *Bioinformatics*, 28(23):3051–3057, December 2012.

[62] Deorowicz S. and Grabowski S. Compression of dna sequence reads in fastq format. *Bioinformatics*, 27(6):860–862, January 2011.

[63] Grabowski S., Deorowicz S., and Roguski L. Disk-based compression of data from genome sequencing. *Bioinformatics*, 31:844, December 2014.

[64] Janin L., Schulz-Trieglaff O., and Cox A.J. Beetl-fastq: a searchable compressed archive for dna reads. *Bioinformatics*, 30(19):2796–2801, October 2014.

[65] Patro R. and Kingsford C. Data-dependent bucketing improves reference-free compression of sequencing reads. *Bioinformatics*, 32(14):248, April 2015.

[66] Kingsford C. and Patro R. Reference-based compression of short-read sequences using path encoding. *Bioinformatics*, 31(12):1920–1928, June 2015.

[67] Kirsch A. and Mitzenmacher M. Less hashing, same performance: Building a better bloom filter. *Random Struct. Algorithms*, 33(2):187–218, September 2008.

[68] Cram format specification. *https://samtools.github.io/hts-specs/CRAMv3.pdf*. Accessed: 2016-06-15.

[69] Ewing B., Hillier L., Wendl M.C., and Green P. Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Research*, 8(3):175–185, March 1998.

[70] DePristo M.A., Banks E., Poplin R., Garimella K.V, Maguire J.R., Harl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernytsky A.M., Sivachenko A.Y., Cibulskis K., Gabriel S.B., Altshuler D., and Daly M.J. A framework for variation discovery and genotyping using next-generation dna sequencing data. *NATURE GENETICS*, 43:491–498, August 2011.

[71] Ochoa I., Hernaez M., Goldfeder R., Weissman T., and Ashley E. Effect of lossy compression of quality scores on variant calling. *Brief Bioinform*, March 2016.

[72] DePristo M., Banks E., Poplin R., Garimella K., Maguire J., Hartl C., Philippakis A., del Angel G. Rivas MA, Hanna M., McKenna A., Fennell T., Kernytsky A., Sivachenko A., Cibulskis K., Gabriel S., Altshuler D., and Daly M. A framework for variation discovery and genotyping using next-generation dna sequencing data. *NATURE GENETICS*, 43:491–498, 2011.

[73] Van der Auwera G.A., Carneiro M., Hartl C., Poplin R., del Angel G., Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J., Banks E., Garimella K., Altshuler D., Gabriel S., and DePristo M. From fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43:11.10.1–11.10.33, 2013.

[74] Picard tools. *http://broadinstitute.github.io/picard*. Accessed: 2016-06-15.

[75] Li H. and Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25:1754–60, 2009.

[76] Grch37 reference genome. *http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/*. Accessed: 2016-06-15.

[77] The vcf format specification. *http://samtools.github.io/hts-specs/VCFv4.2.pdf*. Accessed: 2016-06-15.

[78] 1000 genomes project repository. *ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/data/NA12878*. Accessed: 2016-06-15.

[79] Zook J.M., Chapman B., Wang J., Mittelman D., Hofmann O., Hide W., and Salit M. Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls. *Nature Biotechnology*, pages 246–251, February 2014.

[80] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):7881, 2005.

Appendix A

Results of 5x fold coverage experiment

The next tables summarize the results for the 5x fold coverage experiment for chromosome 20 and chromosome 11. The first block in the table presents the file size for each lossy technique as well as the compression ratio.

The second block in the table corresponds to the results of SNP calling with GIAB-NIST ground truth and GATK caller.

The third block in the table corresponds to the results of SNP calling with Illumina ground truth and GATK caller.

The fourth block in the table corresponds to the results of SNP calling with GIAB-NIST ground truth and Samtools' caller.

The last block in the table corresponds to the results of SNP calling with Illumina ground truth and Samtools' caller.

CHROMOSOME 20

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	1037326297	1037326297	1037326297	1037326297	1037326297
BAM	206153342	165303409	195703116	136871744	154713228
CRAM	77390459	46623719	70289403	32371234	39748470
Compression	13.40380081	22.24889647	14.75793296	32.04469428	26.09726354
True positives	45815	46008	44104	43893	45878
False positives	15999	15992	15701	15713	16137
False negatives	17233	17040	18944	19155	17170
Sensitivity	0.72666857	0.7297297297	0.6995305164	0.6961838599	0.7276678087
Precision	0.7411751383	0.7420645161	0.7374634228	0.7363855988	0.7397887608
F-Score	0.7338501706	0.7358454354	0.7179963045	0.7157206451	0.7336782262
True positives	53679	53870	51704	51439	53770
False positives	8139	8133	8105	8171	8245
False negatives	25687	25496	27662	27927	25592
Sensitivity	0.6763475544	0.6787541265	0.651462843	0.6481238818	0.6775282881
Precision	0.8683393186	0.8688289276	0.8644852781	0.8629256836	0.8670482948
F-Score	0.7604119447	0.7621189936	0.7430070056	0.7402573106	0.7606612108
True positives	47616	48117	45675	45705	47665
False positives	17457	17725	16959	17539	17711
False negatives	15432	14931	17373	17343	15383
Sensitivity	0.7552341073	0.763180434	0.7244480396	0.7249238675	0.756011293
Precision	0.7317320548	0.7307949333	0.7292365169	0.72267725	0.729090186
F-Score	0.7432973517	0.7466366669	0.7268343916	0.7237988154	0.7423067339
True positives	56062	56666	53733	53811	56134
False positives	9013	9178	8903	9435	9242
False negatives	23304	22700	25633	25555	23230
Sensitivity	0.7063730061	0.7139833178	0.6770279465	0.6780107351	0.7072980193
Precision	0.8614982712	0.8606099265	0.8578612938	0.8508206053	0.8586331375
F-Score	0.7762615878	0.7804696646	0.7567921579	0.7546489776	0.7756528948

CHROMOSOME 11

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	2357302048	2357302048	2357302048	2357302048	2357302048
BAM	452482490	363185570	445223159	310363489	340863258
CRAM	168488810	101719649	159262731	72954322	87290416
Compression	13.99085226	23.17450042	14.80134136	32.31202735	27.00527911
True positives	102111	102423	102083	101595	102090
False positives	38182	38339	38185	38045	38508
False negatives	31516	31204	31544	32032	31537
Sensitivity	0.7641494608	0.7664843183	0.7639399223	0.7602879658	0.7639923069
Precision	0.7278410184	0.7276324576	0.7277711238	0.7275494128	0.726112747
F-Score	0.7455534463	0.746553251	0.7454170394	0.7435584977	0.7445710639
True positives	134696	135214	134677	134035	134793
False positives	5602	5553	5596	5610	5805
False negatives	52350	51832	52369	53011	52248
Sensitivity	0.7201223229	0.7228916951	0.7200207436	0.7165884328	0.7206601761
Precision	0.9600707066	0.9605518339	0.960106364	0.9598267034	0.9587120727
F-Score	0.8229629992	0.8249459295	0.8229097608	0.8205613255	0.8228141338
True positives	105422	106399	105414	105474	105481
False positives	41950	42692	41997	43032	42548
False negatives	28205	27228	28213	28153	28146
Sensitivity	0.7889273874	0.7962387841	0.7888675193	0.7893165303	0.789368915
Precision	0.7153461987	0.7136513941	0.7151026721	0.7102339299	0.7125698343
F-Score	0.7503371898	0.7526864225	0.7501761328	0.7476899193	0.7490058795
True positives	139838	141223	139853	140031	139976
False positives	7534	7869	7558	8475	8053
False negatives	47208	45823	47193	47015	47070
Sensitivity	0.7476128867	0.7550174823	0.7476930808	0.7486447184	0.7483506731
Precision	0.9488776701	0.9472205081	0.9487283853	0.9429315987	0.9455984976
F-Score	0.8363066581	0.8402679852	0.8362988366	0.8346306981	0.8354905618

Appendix B

Results of 6x fold coverage experiment

The next tables summarize the results for the 6x fold coverage experiment for chromosome 1 to chromosome 22. The first block in the table presents the file size for each lossy technique as well as the compression ratio.

The second block in the table corresponds to the results of SNP calling with GIAB-NIST ground truth and GATK caller.

The third block in the table corresponds to the results of SNP calling with Illumina ground truth and GATK caller.

The fourth block in the table corresponds to the results of SNP calling with GIAB-NIST ground truth and Samtools' caller.

The last block in the table corresponds to the results of SNP calling with Illumina ground truth and Samtools' caller.

CHROMOSOME 1

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	5538094286	5538094286	5538094286	5538094286	5538094286
BAM					735115893
CRAM	646168716	346487598	632573092	218836304	307099389
Compression	8.570662969	15.98352818	8.754868577	25.30701801	18.03355684
True positives	161206	160871	161179	160513	160952
False positives	49616	48765	49578	49457	49489
False negatives	51949	52284	51976	52642	52203
Sensitivity	0.7562853323	0.754713706	0.7561586639	0.753034177	0.7550937111
Precision	0.7646545427	0.7673825106	0.7647622618	0.7644568272	0.7648319481
F-Score	0.760446911	0.7609953854	0.7604361283	0.7587025111	0.759931633
True positives	201207	200685	201164	200362	200892
False positives	9612	8958	9600	9615	9549
False negatives	85681	86213	85734	86536	85999
Sensitivity	0.7013433814	0.6994994737	0.7011690566	0.698373638	0.7002380695
Precision	0.9544063865	0.9572702165	0.9544514243	0.9542092705	0.9546238613
F-Score	0.8085359458	0.8083320411	0.8084362479	0.8064885535	0.8078788415
True positives	166251	166735	166329	169505	165543
False positives	51875	52229	51932	55711	51425
False negatives	46904	46420	46826	43650	47612
Sensitivity	0.7799535549	0.782224203	0.7803194858	0.7952194413	0.7766320283
Precision	0.7621787407	0.7614722055	0.762064684	0.7526330278	0.7629834814
F-Score	0.7709637104	0.7717087191	0.7710840581	0.7733403898	0.7697472583
True positives	208001	208718	208118	212456	207040
False positives	10125	10246	10143	12762	9928
False negatives	78897	78180	78780	74442	79858
Sensitivity	0.7249998257	0.7274989718	0.7254076362	0.7405279925	0.7216502032
Precision	0.9535818747	0.9532069199	0.9535281154	0.9433349022	0.9542421002
F-Score	0.8237271892	0.8251973858	0.8239702747	0.829718267	0.8218057976

CHROMOSOME 2

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	5883006418	5883006418	5883006418	5883006418	5883006418
BAM		832272936	1229033084	613327655	770668285
CRAM	680850795	360209660	666237992	225097212	319104584
Compression	8.640669088	16.33217282	8.830187544	26.1354033	18.43598216
True positives	175547	175312	175572	174788	175330
False positives	46151	45291	46167	45730	46034
False negatives	53540	53775	53515	54299	53757
Sensitivity	0.7662896629	0.7652638517	0.7663987917	0.7629765111	0.7653424245
Precision	0.7918294256	0.7946945418	0.7917957599	0.7926246384	0.7920438734
F-Score	0.7788502279	0.7797015722	0.7788903036	0.7775180436	0.7784642503
True positives	217173	216704	217206	216142	216888
False positives	4534	3908	4542	4385	4476
False negatives	84780	85249	84747	85811	85056
Sensitivity	0.7192278268	0.717674605	0.7193371154	0.7158133882	0.7183053811
Precision	0.9795495857	0.9822856418	0.9795172899	0.9801158135	0.97977991
F-Score	0.8294427682	0.829385818	0.8295038581	0.8273694687	0.8289114632
True positives	180256	181314	180325	183294	179701
False positives	48907	49344	48899	52150	48734
False negatives	48831	47773	48762	45793	49386
Sensitivity	0.7868451724	0.7914635051	0.787146368	0.8001065098	0.784422512
Precision	0.7865842217	0.7860728871	0.7866759153	0.7785035932	0.7866614135
F-Score	0.7867146754	0.788758986	0.7869110713	0.789157236	0.7855403675
True positives	223537	224955	223621	227720	222835
False positives	5626	5703	5603	7725	5600
False negatives	78416	76998	78332	74233	79118
Sensitivity	0.7403039546	0.7450000497	0.7405821436	0.7541571039	0.7379790895
Precision	0.975449789	0.9752750826	0.9755566607	0.9671897895	0.9754853678
F-Score	0.8417633813	0.8447253249	0.8419829925	0.8474910588	0.8402716502

CHROMOSOME 3

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	4817225818	4817225818	4817225818	4817225818	4817225818
BAM	1013168278	677112325	1004601440	497944854	627077093
CRAM	555080753	291506410	543500772	181052220	258028482
Compression	8.67842344	16.52528264	8.863328382	26.60683099	18.66935689
True positives	149241	149049	149258	148667	149082
False positives	35091	34516	35079	34768	34986
False negatives	45145	45337	45128	45719	45304
Sensitivity	0.7677559083	0.7667681829	0.7678433632	0.7648030208	0.7669379482
Precision	0.8096315344	0.8119685125	0.8097017962	0.8104614714	0.8099289393
F-Score	0.7881378757	0.7887212893	0.7882172458	0.7869705495	0.7878474002
True positives	181839	181484	181860	181040	181628
False positives	2498	2086	2482	2400	2440
False negatives	68703	63051	68682	69502	68909
Sensitivity	0.7257825035	0.7421596091	0.7258663218	0.7225934175	0.7249547971
Precision	0.9864487325	0.9886364874	0.9865358952	0.986916703	0.9867440294
F-Score	0.8362739981	0.8478480747	0.8363609606	0.8343203174	0.8358302367
True positives	152944	153580	153036	155995	152568
False positives	3363	37340	37092	39839	36993
False negatives	41442	40806	41350	38391	41818
Sensitivity	0.7868056342	0.7900774747	0.7872789193	0.8025012089	0.7848713385
Precision	0.9784846488	0.8044206998	0.8049103762	0.796567501	0.8048490987
F-Score	0.8722386817	0.7971845754	0.7959970248	0.7995233458	0.7947346899
True positives	186641	187491	186766	190630	186173
False positives	37060	3429	3362	5204	3388
False negatives	63901	63051	63776	59912	64369
Sensitivity	0.7449489507	0.7483415954	0.745447869	0.7608704329	0.7430810004
Precision	0.8343324348	0.9820395977	0.9823171758	0.9734264734	0.9821271253
F-Score	0.7871112489	0.8494094622	0.8476456305	0.8541229815	0.8460428581

CHROMOSOME 4

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	4664114095	4664114095	4664114095	4664114095	4664114095
BAM	979308567	651267366	971370105	477802536	603603127
CRAM	536049470	279395852	525045326	172734515	247605833
Compression	8.700902353	16.69356958	8.883259909	27.00163366	18.83685064
True positives	157429	157230	157445	156853	157202
False positives	44721	44198	44695	44494	44612
False negatives	45098	45297	45082	45674	45325
Sensitivity	0.7773235174	0.7763409323	0.7774025192	0.7744794521	0.7762026791
Precision	0.7787731882	0.7805766825	0.7788908677	0.7790183117	0.7789449691
F-Score	0.7780476775	0.7784530455	0.7781459818	0.7767422513	0.7775714063
True positives	198121	197773	198111	197363	197822
False positives	4036	3662	4036	3991	3992
False negatives	74373	74721	74383	75131	74665
Sensitivity	0.7270655501	0.7257884577	0.727028852	0.7242838374	0.7259869278
Precision	0.9800353191	0.9818204384	0.9800343315	0.9801791869	0.98021941
F-Score	0.8348070477	0.8346102475	0.8347824988	0.8330224038	0.8341622725
True positives	161137	161664	161197	163940	160489
False positives	47415	47702	47422	50173	47149
False negatives	41390	40863	41330	38587	42038
Sensitivity	0.7956321873	0.7982343095	0.7959284441	0.8094723173	0.7924326139
Precision	0.7726466301	0.7721597585	0.7726860928	0.7656704637	0.7729269209
F-Score	0.7839709642	0.7849805653	0.7841350761	0.7869623656	0.7825582388
True positives	203293	204001	203351	207161	202415
False positives	5259	5365	5268	6952	5223
False negatives	69201	68493	69143	65333	70079
Sensitivity	0.7460457845	0.7486440068	0.7462586332	0.7602405925	0.7428236952
Precision	0.9747832675	0.9743750179	0.9747482252	0.9675311635	0.9748456448
F-Score	0.8452123082	0.8467231146	0.8453357112	0.8514509656	0.8431639632

CHROMOSOME 5

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	4418276086	4418276086	4418276086	4418276086	4418276086
BAM	928642406	620549383	920276692	456086794	574749521
CRAM	509638053	267829235	498618629	166452120	237226888
Compression	8.669439144	16.49661616	8.86103292	26.54382585	18.6246851
True positives	110768	110578	110765	110207	110605
False positives	56087	55516	56078	55658	55959
False negatives	35586	35776	35589	36147	35749
Sensitivity	0.7568498299	0.7555516077	0.7568293316	0.7530166582	0.7557360919
Precision	0.6638578406	0.665755536	0.6638876069	0.6644379465	0.6640390481
F-Score	0.707310454	0.7078169807	0.707318397	0.7059595989	0.7069264152
True positives	164389	164033	164385	163501	164151
False positives	2471	2066	2463	2369	2413
False negatives	65713	66069	65717	66601	65946
Sensitivity	0.7144179538	0.7128708138	0.7144005702	0.7105587957	0.7133991317
Precision	0.9851911782	0.9875616349	0.985238061	0.9857177308	0.9855130761
F-Score	0.8282354482	0.8280292074	0.8282403325	0.8258210176	0.8276639246
True positives	113370	113882	113395	115623	113070
False positives	58609	58902	58629	61566	58495
False negatives	32984	32472	32959	30731	33284
Sensitivity	0.7746286401	0.7781270071	0.7747994588	0.7900228214	0.7725788157
Precision	0.6592083917	0.6591003797	0.659181277	0.6525405076	0.6590505056
F-Score	0.7122729971	0.7136849889	0.7123293695	0.714730345	0.711313259
True positives	168817	169586	168871	172418	168407
False positives	3163	3199	3154	4772	3158
False negatives	61285	60516	61231	57684	61694
Sensitivity	0.7336615936	0.7370035897	0.7338962721	0.749311175	0.7318829557
Precision	0.9816083265	0.9814856614	0.9816654556	0.9730684576	0.9815929823
F-Score	0.8397142871	0.8418539193	0.8398888908	0.8466554708	0.8385424706

CHROMOSOME 6

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	4161777636	4161777636	4161777636	4161777636	4161777636
BAM	874941907	867391686	867391686	430505825	542062421
CRAM	480374599	470100459	470100459	157104003	223763482
Compression	8.663608868	8.852953781	8.852953781	26.4905894	18.59900283
True positives	130430	130227	130417	129794	130250
False positives	41666	41091	41640	41361	41567
False negatives	42255	42458	42268	42891	42435
Sensitivity	0.7553059038	0.754130353	0.7552306222	0.7516228972	0.7542635434
Precision	0.7578909446	0.7601477953	0.7579871787	0.7583418539	0.7580739973
F-Score	0.7565962161	0.7571271181	0.7566063897	0.7549674267	0.75616397
True positives	168452	168053	168415	167560	168206
False positives	3652	3273	3650	3603	3611
False negatives	73404	73803	73441	74296	73642
Sensitivity	0.6964970892	0.6948473472	0.6963441056	0.6928089442	0.6955029605
Precision	0.9787802724	0.9808960695	0.9787870863	0.9789498899	0.9789834533
F-Score	0.8138564112	0.8134575078	0.8137543154	0.8113912435	0.8132474345
True positives	133508	134005	133524	135975	132959
False positives	44691	44919	44697	47036	44426
False negatives	39177	38680	39161	36710	39726
Sensitivity	0.7731302661	0.7760083389	0.7732229203	0.787416394	0.769951067
Precision	0.7492073468	0.7489492746	0.7492046392	0.7429881264	0.7495504129
F-Score	0.760980837	0.7622387368	0.7610243199	0.7645573748	0.7596137915
True positives	173350	174037	173372	176688	172574
False positives	4850	4888	4850	6325	4811
False negatives	68506	67819	68484	65168	69281
Sensitivity	0.7167488092	0.7195893424	0.7168397724	0.7305504102	0.7135432387
Precision	0.9727833895	0.972681291	0.9727867491	0.9654396136	0.9728782028
F-Score	0.8253661417	0.8272094035	0.8254276587	0.8317293095	0.8232706803

CHROMOSOME 7

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	3793108020	3793108020	3793108020	3793108020	3793108020
BAM	802991080	541553974	795990185	400498554	501527840
CRAM	440951556	235258121	431737784	148057116	208508100
Compression	8.602096916	16.1231757	8.785675381	25.6192213	18.19165788
True positives	113320	113089	113290	112836	113184
False positives	36113	35605	36102	35859	36013
False negatives	37842	38073	37872	38326	37978
Sensitivity	0.7496593059	0.7481311441	0.7494608433	0.746457443	0.7487596089
Precision	0.758333166	0.760548509	0.7583404734	0.7588419247	0.7586211519
F-Score	0.7539712903	0.7542887253	0.7538745117	0.7525987387	0.7536581224
True positives	146323	145944	146278	145669	146120
False positives	3115	2755	3119	3031	3077
False negatives	62824	63203	62869	63478	63022
Sensitivity	0.699617972	0.6978058495	0.6994028124	0.6964909848	0.6986640656
Precision	0.9791552349	0.9814726394	0.9791227401	0.9796166779	0.9793762609
F-Score	0.8161133344	0.8156804883	0.8159556428	0.8141412391	0.8155405915
True positives	116711	117421	116756	118320	116285
False positives	39289	39578	39316	41210	39082
False negatives	34451	33741	34406	32842	34877
Sensitivity	0.7720921925	0.7767891401	0.7723898863	0.782736402	0.7692740239
Precision	0.7481474359	0.7479092224	0.7480906248	0.7416786811	0.7484536613
F-Score	0.7599312415	0.7620756682	0.7600460886	0.761654629	0.7587210346
		152397			
True positives	151433	152399	151488	153729	150841
False positives	4567	4602	4585	5802	4526
False negatives	57714	56748	57659	55418	58306
Sensitivity	0.7240505482	0.7286693091	0.7243135211	0.7350284728	0.7212200032
Precision	0.970724359	0.9706880848	0.9706227214	0.9636308931	0.9708689748
F-Score	0.8294358163	0.8324448037	0.8295712173	0.8339472385	0.8276280198

CHROMOSOME 8

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	3566420984	3566420984	3566420984	3566420984	3566420984
BAM	748753557	502027774	742522181	369535188	464764741
CRAM	412506313	217672843	403669605	135739002	192754279
Compression	8.645736736	16.38431756	8.835000059	26.27410642	18.50242185
True positives	112134	111930	112095	111716	111967
False positives	31764	31241	31752	31552	31681
False negatives	35654	35858	35693	36072	35821
Sensitivity	0.7587490189	0.7573686632	0.7584851273	0.7559206431	0.7576190218
Precision	0.7792603094	0.7817924021	0.7792654696	0.7797693833	0.779453943
F-Score	0.7688678922	0.7693867521	0.7687348912	0.7676598318	0.7683813942
True positives	141314	140965	141263	140746	141095
False positives	2588	2210	2588	2526	2553
False negatives	55692	56041	55743	56260	55907
Sensitivity	0.7173081023	0.7155365826	0.7170492269	0.7144249414	0.7162110029
Precision	0.9820155384	0.9845643443	0.9820091623	0.9823691998	0.9822273892
F-Score	0.8290447863	0.828764687	0.8288695846	0.8272412557	0.8283869074
True positives	115258	115674	115306	117195	114840
False positives	33879	34080	112095	35671	33689
False negatives	32530	32114	32482	30593	32948
Sensitivity	0.7798874063	0.7827022492	0.7802121958	0.7929940185	0.7770590305
Precision	0.7728330327	0.7724267799	0.5070602152	0.7666518389	0.7731823415
F-Score	0.7763441947	0.7775305671	0.6146555469	0.779600471	0.7751158388
True positives	145644	146219	145687	148256	145072
False positives	3493	3535	3492	4612	3457
False negatives	51362	50787	51319	48750	51934
Sensitivity	0.7392871283	0.7422058211	0.7395053958	0.7525456077	0.7363836634
Precision	0.9765785821	0.9763946205	0.9765918796	0.9698301803	0.976725084
F-Score	0.8415250345	0.8433440997	0.8416713607	0.8474822365	0.8396949658

CHROMOSOME 9

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	2962665999	2962665999	2962665999	2962665999	2962665999
BAM	629611414	426263380	623846100	316103437	394906024
CRAM	346462373	186505869	339052891	118398006	165676433
Compression	8.551191211	15.88510868	8.738064407	25.0229383	17.88224158
True positives	88171	88001	88175	87847	87984
False positives	26743	26355	26750	26676	26663
False negatives	29255	29425	29251	29579	27022
Sensitivity	0.7508643742	0.7494166539	0.7508984382	0.7481051897	0.7650383458
Precision	0.7672781384	0.7695354857	0.767239504	0.7670686238	0.7674339494
F-Score	0.7589825256	0.7593428308	0.7589810244	0.7574682366	0.7662342752
True positives	110879	110573	110888	110459	110650
False positives	4040	3788	4042	4069	3997
False negatives	54829	55135	54820	55249	55053
Sensitivity	0.6691227943	0.6672761725	0.6691771067	0.6665882154	0.6677609941
Precision	0.9648448037	0.9668768199	0.9648307666	0.9644715703	0.9651364624
F-Score	0.7902233213	0.7896125598	0.790256487	0.7883284089	0.7893704298
True positives	90941	91266	90944	92780	90404
False positives	28450	28635	28434	30282	28099
False negatives	26485	26160	26482	24646	29442
Sensitivity	0.7744536985	0.777221399	0.7744792465	0.7901146254	0.7543347296
Precision	0.7617073314	0.7611779718	0.7618154099	0.7539289139	0.7628836401
F-Score	0.7680276332	0.7691160298	0.7680951335	0.7715977512	0.7585851
True positives	114789	115251	114807	117325	114047
False positives	4605	4653	4574	5740	4456
False negatives	50919	50457	50901	48383	51658
Sensitivity	0.6927185169	0.6955065537	0.6928271417	0.7080225457	0.6882532211
Precision	0.9614302226	0.9611939552	0.9616856954	0.9533579816	0.9623975764
F-Score	0.8052486479	0.8070459224	0.8054116434	0.812575968	0.8025600968

CHROMOSOME 10

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	3168997057	3168997057	3168997057	3168997057	3168997057
BAM	671557510	454306096	665767852	337195949	420551471
CRAM	368271123	197133249	360675979	124664704	174489951
Compression	8.605065288	16.07540622	8.78627145	25.42016269	18.16148746
True positives	99115	98919	99100	98710	98944
False positives	32547	32022	32541	32290	32437
False negatives	33226	33422	33241	33631	33397
Sensitivity	0.7489364596	0.7474554371	0.748823116	0.7458761835	0.747644343
Precision	0.7527988334	0.7554471098	0.7528049772	0.7535114504	0.7531073747
F-Score	0.7508626796	0.7514300256	0.7508087673	0.7496743766	0.7503659156
True positives	129086	128729	129066	128505	128835
False positives	2581	2216	2580	2499	2546
False negatives	56052	56409	56072	56633	56298
Sensitivity	0.6972420573	0.6953137659	0.6971340298	0.6941038577	0.6959051061
Precision	0.980397518	0.9830768643	0.9804019871	0.9809242466	0.9806212466
F-Score	0.8149240069	0.8145265642	0.8148517602	0.8129574685	0.8140872126
True positives	102275	102893	102323	103369	101849
False positives	35080	35380	35051	36419	34838
False negatives	30066	29448	30018	28972	30492
Sensitivity	0.7728141695	0.7774839241	0.7731768688	0.7810806931	0.7695952124
Precision	0.7446033999	0.7441293673	0.744849826	0.7394697685	0.7451257252
F-Score	0.7584465472	0.760441071	0.7587490499	0.7597058748	0.7571628232
True positives	133716	134563	133750	135353	133101
False positives	3640	3711	3625	4436	3586
False negatives	51422	50575	51388	49785	52036
Sensitivity	0.7222504294	0.7268253951	0.7224340762	0.7310924824	0.7189324662
Precision	0.9734995195	0.9731619827	0.9736123749	0.9682664587	0.9737648789
F-Score	0.8292619398	0.8321459933	0.8294239302	0.833128672	0.8271664015

CHROMOSOME 11

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	3323718560	3323718560	3323718560	3323718560	3323718560
BAM	705296161	477193614	663864702	351036712	420715141
CRAM	399026425	214238552	357220844	135091577	174185183
Compression	8.329570053	15.51410112	9.304380234	24.60344778	19.08152291
True positives	101882	101751	101795	101485	101723
False positives	33553	33097	34186	33386	33448
False negatives	31745	31876	31832	32142	31904
Sensitivity	0.7624357353	0.7614553945	0.7617846693	0.7594647788	0.761245856
Precision	0.7522575405	0.7545606906	0.7485972305	0.7524597578	0.7525504731
F-Score	0.757312441	0.7579923643	0.7551333788	0.7559460406	0.756873191
True positives	133495	133247	133461	132967	133276
False positives	1943	1604	2523	1907	1895
False negatives	53551	53799	53585	54079	53767
Sensitivity	0.7137014424	0.7123755654	0.713519669	0.7108786074	0.7125420358
Precision	0.9856539524	0.9881053904	0.9814463466	0.9858608776	0.9859807207
F-Score	0.8279170439	0.8278859387	0.8263071541	0.8260872266	0.827251454
True positives	105107	105446	104455	106675	104643
False positives	35773	35988	35923	37536	35556
False negatives	28520	28181	29172	26952	28984
Sensitivity	0.7865700794	0.7891069919	0.781690826	0.7983042349	0.7830977273
Precision	0.7460746735	0.7455491607	0.744098078	0.7397147236	0.7463890613
F-Score	0.7657873934	0.7667099298	0.7624313425	0.7678935207	0.7643028785
True positives	138198	138685	137291	140493	137545
False positives	2682	2749	3087	3718	2654
False negatives	48848	48361	49755	46553	49501
Sensitivity	0.73884499	0.7414486276	0.7339959154	0.7511146991	0.7353538702
Precision	0.9809625213	0.9805633723	0.9780093747	0.9742183329	0.9810697651
F-Score	0.8428608893	0.84440453	0.83861293	0.8482416975	0.8406239973

CHROMOSOME 12

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	3189223619	3189223619	3189223619	3189223619	3189223619
BAM	673459159	453567164	667479162	335357141	419928886
CRAM	369188653	196203259	361432544	123148425	173704914
Compression	8.63846598	16.25469238	8.823841881	25.89739673	18.36000805
True positives	99876	99684	99861	99474	99729
False positives	25697	25174	25702	25528	25615
False negatives	32426	32618	32441	32828	32573
Sensitivity	0.7549092228	0.7534579976	0.7547958459	0.75187072	0.7537981285
Precision	0.7953620603	0.7983789585	0.795305942	0.7957792675	0.7956423921
F-Score	0.7746078526	0.7752683154	0.774521552	0.7732021267	0.7741552363
True positives	123590	123184	123581	123071	123381
False positives	1990	1679	1989	1938	1963
False negatives	52230	52636	52239	52749	52432
Sensitivity	0.7029348197	0.7006256399	0.702883631	0.6999829371	0.701774044
Precision	0.9841535276	0.9865532624	0.9841602294	0.9844971162	0.9843390988
F-Score	0.8201061712	0.8193612542	0.8200736587	0.8182123399	0.8193799248
True positives	102945	103449	102885	104226	102510
False positives	27669	27855	27684	29142	27481
False negatives	29357	28853	29417	28076	29792
Sensitivity	0.7781061511	0.7819156173	0.7776526432	0.7877885444	0.7748182189
Precision	0.7881620653	0.7878587096	0.7879741746	0.7814918121	0.7885930564
F-Score	0.7831018272	0.7848759133	0.7827793861	0.7846275455	0.7816449543
True positives	127739	128425	127681	129517	127157
False positives	2877	2881	2890	3853	2834
False negatives	48081	47395	48139	46303	48661
Sensitivity	0.7265328177	0.7304345353	0.7262029348	0.7366454328	0.723230841
Precision	0.977973602	0.9780588854	0.9778664481	0.9711104446	0.9781984907
F-Score	0.8337075278	0.8363017133	0.8334513742	0.8377825932	0.8316105805

CHROMOSOME 13

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	2365133667	2365133667	2365133667	2365133667	2365133667
BAM	496734164	330881629	492783714	243313768	306550817
CRAM	271687684	141851585	266176807	87985054	125568497
Compression	8.705340015	16.67329743	8.885573817	26.88108445	18.83540636
True positives	87377	87215	87364	87035	87251
False positives	19220	18903	19227	19071	19171
False negatives	24562	24724	24575	24904	24688
Sensitivity	0.7805769214	0.7791297046	0.7804607867	0.7775216859	0.7794513083
Precision	0.8196947381	0.8218681091	0.8196189172	0.820264641	0.8198586758
F-Score	0.7996577223	0.7999284591	0.799560701	0.7983214474	0.7991445359
True positives	105429	105164	105422	104982	105270
False positives	1172	958	1173	1128	1152
False negatives	36878	37143	36885	37325	37033
Sensitivity	0.7408560366	0.7389938654	0.7408068472	0.7377149402	0.7397595272
Precision	0.9890057317	0.9909726541	0.9889957315	0.9893695222	0.9891751705
F-Score	0.8471322738	0.846632237	0.8470964476	0.8452078561	0.8464770329
True positives	89483	89779	89532	91297	89156
False positives	20433	20562	20458	21864	20349
False negatives	22456	22160	22407	20642	22783
Sensitivity	0.7993907396	0.8020350369	0.799828478	0.8155959942	0.7964695057
Precision	0.8141034972	0.813650411	0.8140012728	0.8067885579	0.8141728688
F-Score	0.8066800388	0.8078009717	0.8068526421	0.8111683696	0.8052238941
True positives	108277	108666	108342	110658	107868
False positives	1639	1675	1648	2503	1637
False negatives	34030	33641	33965	31649	34439
Sensitivity	0.7608691069	0.7636026337	0.7613258659	0.7776005397	0.7579950389
Precision	0.9850886131	0.9848197859	0.9850168197	0.9778810721	0.9850509109
F-Score	0.8585814934	0.8602165859	0.8588449328	0.8663159378	0.8567343891

CHROMOSOME 14

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	2171763375	2171763375	2171763375	2171763375	2171763375
BAM	458360455	308612285	454302062	228343254	285714222
CRAM	251564565	133786262	246273919	84210571	118480598
Compression	8.63302578	16.2330821	8.818487089	25.78967639	18.33011828
True positives	67225	67077	67199	66929	67109
False positives	17838	17502	17843	17672	17781
False negatives	21670	21818	21696	21966	21786
Sensitivity	0.7562292592	0.7545643737	0.7559367793	0.7528994882	0.754924349
Precision	0.7902966037	0.7930692016	0.7901860257	0.7911135802	0.7905406997
F-Score	0.7728877085	0.7733377913	0.772682063	0.7715336377	0.7723221222
True positives	82929	82677	82907	82529	82779
False positives	2138	1906	2139	2076	2111
False negatives	35080	35332	35102	35480	35226
Sensitivity	0.7027345372	0.7005991068	0.7025481107	0.6993449652	0.7014872251
Precision	0.9748668696	0.977465921	0.9748489053	0.9754624431	0.9751325244
F-Score	0.8167287124	0.8161921497	0.8165964886	0.8146426209	0.8159787082
True positives	69186	69510	69229	70290	68895
False positives	18814	18950	18830	19984	18737
False negatives	19709	19385	19666	18605	20000
Sensitivity	0.7782889926	0.7819337421	0.7787727094	0.7907081388	0.7750154677
Precision	0.7862045455	0.7857788831	0.7861660932	0.7786295057	0.7861854117
F-Score	0.7822267447	0.7838515971	0.7824519367	0.7846223398	0.7805604808
True positives	85432	85840	85480	86940	85058
False positives	2569	2621	2580	3335	2574
False negatives	32577	32169	32529	31069	32950
Sensitivity	0.7239447839	0.7274021473	0.7243515325	0.7367234702	0.7207816419
Precision	0.9708071499	0.970371124	0.9707017942	0.9630573248	0.9706271682
F-Score	0.8293966312	0.8315009444	0.8296250285	0.8348216858	0.8272515075

CHROMOSOME 15

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	1983354324	1983354324	1983354324	1983354324	1983354324
BAM	421751453	286384538	418139350	212666430	265074790
CRAM	231643029	124937182	226924544	79266838	110665451
Compression	8.562115305	15.87481238	8.74014899	25.0212368	17.92207329
True positives	60036	59971	60022	59851	59977
False positives	17111	16755	17115	17004	17074
False negatives	19243	19308	19257	19428	19302
Sensitivity	0.7572749404	0.7564550511	0.7570983489	0.7549414095	0.7565307332
Precision	0.7782026521	0.7816255246	0.7781220426	0.7787521957	0.77840651
F-Score	0.7675961797	0.7688343322	0.7674662439	0.7666619698	0.7673127359
True positives	75080	74893	75068	74819	74990
False positives	2068	1834	2070	2037	2061
False negatives	33437	33624	33449	33698	33526
Sensitivity	0.6918731627	0.6901499304	0.691762581	0.6894680096	0.6910501677
Precision	0.9731943796	0.9760970714	0.9731649771	0.9734958884	0.9732514828
F-Score	0.8087684809	0.8085875926	0.8086827718	0.8072265109	0.8082256005
True positives	61957	62276	61983	62810	61676
False positives	18187	18339	18200	19133	18045
False negatives	17322	17003	17296	16469	17603
Sensitivity	0.7815058212	0.7855295854	0.7818337769	0.7922652909	0.7779613769
Precision	0.7730709722	0.7725113192	0.7730192185	0.7665084266	0.7736480977
F-Score	0.7772655138	0.778966065	0.7774015126	0.7791740581	0.7757987421
True positives	77808	78236	77835	79027	77423
False positives	2336	2379	2348	2916	2298
False negatives	30709	30281	30682	29490	31094
Sensitivity	0.7170120811	0.7209561636	0.71726089	0.728245344	0.7134642498
Precision	0.9708524656	0.970489363	0.9707169849	0.964414288	0.971174471
F-Score	0.8248445625	0.8273163716	0.8249602544	0.8298540376	0.82260755

CHROMOSOME 16

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	2094381303	2094381303	2094381303	2094381303	2094381303
BAM	435143773	301677383	433305866	224640092	278905563
CRAM	248056436	137018943	242092151	88015032	121649900
Compression	8.443164535	15.28534126	8.65117392	23.79572279	17.21646547
True positives	57078	56931	57074	56830	56946
False positives	22678	22245	22670	22566	22580
False negatives	22214	22361	22218	22462	22346
Sensitivity	0.7198456339	0.7179917268	0.7197951874	0.7167179539	0.718180901
Precision	0.7156577562	0.7190436496	0.7157152889	0.7157791324	0.7160677011
F-Score	0.7177455862	0.7185173032	0.7177494404	0.7162482355	0.7171227443
True positives	75195	74861	75186	74839	74997
False positives	4561	4315	4558	4557	4529
False negatives	38108	38442	38117	38464	38306
Sensitivity	0.6636629215	0.6607150737	0.6635834885	0.6605209041	0.661915395
Precision	0.9428130799	0.945501162	0.9428420947	0.9426041614	0.9430500717
F-Score	0.7789846627	0.777861481	0.7789398437	0.7767450791	0.7778601766
True positives	59394	59603	59399	59843	59159
False positives	23772	23912	23772	24580	23608
False negatives	19898	19689	19893	19449	20133
Sensitivity	0.749054129	0.7516899561	0.7491171871	0.7547167432	0.7460904
Precision	0.7141620374	0.7136801772	0.7141792211	0.7088471151	0.7147655467
F-Score	0.7311920619	0.7321921048	0.7312311111	0.731063128	0.7300921269
True positives	78526	78817	78534	79309	78185
False positives	4640	4698	4637	5114	4582
False negatives	34777	34486	34769	33994	35118
Sensitivity	0.6930619666	0.6956303011	0.6931325737	0.6999726397	0.6900523375
Precision	0.9442079696	0.9437466323	0.9442473939	0.9394240906	0.9446397719
F-Score	0.7993729291	0.8009125182	0.7994340218	0.8022111407	0.7975212934

CHROMOSOME 17

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	1826109217	1826109217	1826109217	1826109217	1826109217
BAM	392382465	271963361	389153211	205001979	251420614
CRAM	216217077	120228288	211969538	78423813	106419989
Compression	8.445721505	15.18868186	8.614960594	23.28513684	17.15945692
True positives	48124	47995	48103	47901	48037
False positives	16570	16213	16567	16465	16504
False negatives	19623	19752	19644	19846	19710
Sensitivity	0.7103487977	0.7084446544	0.7100388209	0.7070571391	0.709064608
Precision	0.7438711472	0.7474925243	0.7438224834	0.744197247	0.7442865775
F-Score	0.7267235977	0.7274449623	0.7265381333	0.7251519533	0.7262487905
True positives	62687	62439	62657	62358	62544
False positives	2010	1772	2016	2011	1997
False negatives	33067	33315	33097	33396	33207
Sensitivity	0.654667168	0.6520771978	0.6543538651	0.6512312802	0.6531942225
Precision	0.9689320989	0.9724034823	0.9688277952	0.9687582532	0.969058428
F-Score	0.7813849711	0.780658269	0.781127865	0.7788762389	0.7803758141
True positives	50755	51045	50783	51214	50630
False positives	18562	18745	18584	19328	18494
False negatives	16992	16702	16964	16533	17117
Sensitivity	0.7491844657	0.7534650981	0.7495977682	0.7559596735	0.7473393656
Precision	0.7322157624	0.7314085112	0.7320916286	0.7260072014	0.7324518257
F-Score	0.74060293	0.7422729884	0.740741281	0.7406807483	0.7398207071
True positives	66502	66936	66544	67209	66305
False positives	2815	2854	2823	3333	2819
False negatives	29252	28818	29210	28545	29449
Sensitivity	0.6945088456	0.6990412933	0.6949474696	0.7018923491	0.6924514903
Precision	0.9593894716	0.9591058891	0.9593034152	0.9527515523	0.9592182165
F-Score	0.8057381369	0.8086792635	0.8060028706	0.8083056718	0.8042916581

CHROMOSOME 18

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	1835915113	1835915113	1835915113	1835915113	1835915113
BAM	387109096	259117101	383709427	191219968	239927520
CRAM	212041234	111593346	207605481	69725785	98757243
Compression	8.65829291	16.45183319	8.843288261	26.33050475	18.59018192
True positives	64288	64160	64272	64032	64153
False positives	14896	14613	14878	14793	14835
False negatives	19719	19847	19735	19975	19854
Sensitivity	0.7652695609	0.7637458783	0.7650791006	0.7622221958	0.7636625519
Precision	0.8118811881	0.8144922753	0.8120277953	0.8123311132	0.8121866613
F-Score	0.7878865869	0.7883032314	0.7878546431	0.7864793161	0.7871775208
True positives	78227	78008	78197	77888	78053
False positives	959	767	955	939	935
False negatives	30480	30699	30510	30819	30652
Sensitivity	0.7196132724	0.7175986827	0.7193373012	0.7164947979	0.7180258498
Precision	0.9878892734	0.9902634084	0.9879346068	0.9880878379	0.9881627589
F-Score	0.8326760443	0.8321652212	0.8325073592	0.8306547079	0.8317092273
True positives	65991	66225	66033	67216	65736
False positives	1513	16137	16089	17088	15967
False negatives	18016	17782	17974	16791	18271
Sensitivity	0.7855416811	0.7883271632	0.7860416394	0.8001237992	0.7825062197
Precision	0.9775865134	0.8040722663	0.8040841675	0.7973049915	0.8045726595
F-Score	0.8711050683	0.7961218737	0.7949605427	0.7987119083	0.7933860358
True positives	80549	80843	80602	82152	80218
False positives	16070	1520	1521	2153	1485
False negatives	28158	27864	28105	26555	28489
Sensitivity	0.7409734424	0.74367796	0.7414609915	0.7557195029	0.7379285603
Precision	0.8336766061	0.9815451113	0.981479001	0.9744617757	0.9818244128
F-Score	0.7845962031	0.8462134296	0.8447518734	0.8512631339	0.8425817972

CHROMOSOME 19

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	1253005372	1253005372	1253005372	1253005372	1253005372
BAM	272741119	192775421	270558389	147464534	178109148
CRAM	150373015	86041747	147619948	57586392	76256531
Compression	8.332647796	14.5627607	8.488049135	21.75870598	16.43144994
True positives	31795	31686	31769	31679	31722
False positives	20362	20031	20342	20332	20290
False negatives	17373	17482	17399	17489	17446
Sensitivity	0.6466604295	0.6444435405	0.6461316303	0.6443011715	0.645175724
Precision	0.6096017792	0.6126805499	0.6096409587	0.609082694	0.6098977159
F-Score	0.6275845053	0.6281607771	0.6273561153	0.6261971358	0.6270409172
True positives	50553	50274	50509	50361	50412
False positives	1606	1445	1604	1652	1600
False negatives	34822	35101	34866	35014	34961
Sensitivity	0.5921288433	0.5888609078	0.59161347	0.5898799414	0.5904911389
Precision	0.9692095324	0.972060558	0.9692207319	0.9682387096	0.9692378682
F-Score	0.7351345849	0.7334237822	0.7347404864	0.7331207966	0.7338792445
True positives	34317	34498	34323	34682	34121
False positives	22550	22705	22562	23333	22393
False negatives	14851	14670	14845	14486	15047
Sensitivity	0.6979539538	0.7016352099	0.6980759844	0.7053774813	0.6939676212
Precision	0.6034607066	0.603080258	0.6033752307	0.597810911	0.6037618997
F-Score	0.6472768426	0.6486354364	0.6472801335	0.647154866	0.6457296418
True positives	54453	54761	54467	55117	54137
False positives	2415	2443	2419	2899	2377
False negatives	30922	30614	30908	30258	31237
Sensitivity	0.6378096633	0.6414172767	0.6379736457	0.6455871157	0.634115773
Precision	0.9575332349	0.9572931963	0.9574763562	0.9500310259	0.9579396256
F-Score	0.7656334582	0.7681495872	0.7657334055	0.7687651247	0.7630948354

CHROMOSOME 20

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	1409441215	1409441215	1409441215	1409441215	1409441215
BAM	300302069	205177256	297843978	153751094	189692794
CRAM	164791886	89437663	161609469	57501251	79082759
Compression	8.552855661	15.75892267	8.721278671	24.51148785	17.82235765
True positives	45818	45664	45796	45617	45715
False positives	8312	8140	8607	8268	8285
False negatives	17230	17384	17252	17431	17333
Sensitivity	0.7267161528	0.7242735693	0.7263672123	0.7235281056	0.7250824768
Precision	0.8464437465	0.8487101331	0.841791813	0.846562123	0.8465740741
F-Score	0.7820239294	0.781569849	0.7798315893	0.780224573	0.7811325268
True positives	53076	52873	53074	52833	52962
False positives	1056	933	1331	1054	1038
False negatives	26290	26493	26292	26533	26402
Sensitivity	0.6687498425	0.6661920722	0.6687246428	0.665688078	0.6673302757
Precision	0.9804921303	0.9826599264	0.9755353368	0.9804405515	0.9807777778
F-Score	0.7951579799	0.7940558075	0.7935053188	0.7929727661	0.7942473231
True positives	47503	47819	47222	47931	47295
False positives	9409	9482	9493	9895	9305
False negatives	15545	15229	15826	15117	15753
Sensitivity	0.7534418221	0.7584538764	0.7489849004	0.7602303007	0.7501427484
Precision	0.8346745853	0.8345229577	0.8326192365	0.8288832013	0.8356007067
F-Score	0.7919806602	0.7946721618	0.7885908002	0.7930737793	0.7905690024
True positives	55264	55639	54903	55855	54997
False positives	1649	1664	1813	1973	1603
False negatives	24102	23727	24463	23511	24368
Sensitivity	0.6963183227	0.7010432679	0.6917697755	0.7037648363	0.692962893
Precision	0.9710259519	0.9709613807	0.9680337118	0.9658815799	0.9716784452
F-Score	0.8110420534	0.8142153671	0.8069105392	0.8142484365	0.8089876071

CHROMOSOME 21

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	895609370	895609370	895609370	895609370	895609370
BAM	189659020	128222195	188137866	95303991	118733993
CRAM	104544430	56089302	102395173	35722367	49773997
Compression	8.566782276	15.96756134	8.746597557	25.07138931	17.99351919
True positives	29618	29566	29635	29491	29556
False positives	13369	13303	13383	13440	13306
False negatives	10204	10256	10187	10331	10266
Sensitivity	0.7437597308	0.7424539199	0.7441866305	0.7405705389	0.7422028025
Precision	0.6889989997	0.6896825212	0.6888976707	0.6869395076	0.6895618497
F-Score	0.7153328744	0.7150959597	0.7154756156	0.7127475741	0.7149146147
True positives	37806	37735	37823	37657	37718
False positives	5184	5137	5198	5277	5144
False negatives	18755	18826	18738	18904	18840
Sensitivity	0.668411096	0.6671558141	0.6687116564	0.665776772	0.6668906256
Precision	0.8794138172	0.8801782049	0.8791752865	0.8770904179	0.8799869348
F-Score	0.759530291	0.7590035501	0.7596352755	0.7569626614	0.7587608127
True positives	30448	30567	30464	30681	30297
False positives	13810	13927	13809	14300	13657
False negatives	9374	9255	9358	9141	9525
Sensitivity	0.764602481	0.767590779	0.765004269	0.7704535182	0.7608106072
Precision	0.6879660174	0.6869915045	0.6880943239	0.6820879927	0.6892888019
F-Score	0.724262607	0.7250581147	0.7245139426	0.7235828921	0.7232859053
True positives	39104	39263	39114	39476	38886
False positives	5154	5231	5159	5505	5068
False negatives	17457	17298	17447	17085	17675
Sensitivity	0.6913597709	0.6941708951	0.6915365711	0.6979367409	0.687505525
Precision	0.8835464775	0.8824335866	0.8834729971	0.877614993	0.8846976384
F-Score	0.7757267975	0.7770619959	0.7758097467	0.77753048	0.7737352634

CHROMOSOME 22

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	790790801	790790801	790790801	790790801	790790801
BAM	171202099	120180923	169776659	91960953	111075593
CRAM	94461457	53601311	92689013	35946465	47538423
Compression	8.371571074	14.75319887	8.531656292	21.99912567	16.63477144
True positives	17010	16963	17022	16960	16975
False positives	12958	12793	12951	12863	12921
False negatives	8983	9030	8971	9033	9018
Sensitivity	0.6544069557	0.6525987766	0.6548686185	0.6524833609	0.6530604393
Precision	0.5676054458	0.5700699019	0.56791112	0.5686885961	0.5678017126
F-Score	0.6079233752	0.6085490323	0.6082978952	0.6077110506	0.6074540607
True positives	28827	28731	28831	28702	28760
False positives	1142	1027	1143	1123	1136
False negatives	21256	21352	21252	21381	21322
Sensitivity	0.5755845297	0.5736677116	0.5756643971	0.5730886728	0.5742582165
Precision	0.9618939571	0.9654882721	0.9618669514	0.9623470243	0.9620016056
F-Score	0.7202068655	0.7197054145	0.7202618135	0.7183761326	0.7191977794
True positives	18167	18289	18190	18351	18095
False positives	14373	14480	14398	14749	14251
False negatives	7826	7704	7803	7642	7898
Sensitivity	0.6989189397	0.7036125111	0.6998037933	0.7059977686	0.6961489632
Precision	0.55829748	0.5581189539	0.5581809255	0.5544108761	0.559420021
F-Score	0.6207438539	0.6224771111	0.6210204674	0.6210887922	0.6203397384
True positives	30974	31197	31014	31332	30810
False positives	1566	1573	1574	1769	1536
False negatives	19109	18886	19069	18751	19273
Sensitivity	0.6184533674	0.6229059761	0.6192520416	0.6256015015	0.6151788032
Precision	0.9518746159	0.9519987794	0.9517000123	0.9465575058	0.9525134483
F-Score	0.749767014	0.7530686879	0.7502993795	0.7533179458	0.7475524391

Appendix C

Results of 50x fold coverage experiment

The next tables summarize the results for the 50x fold coverage experiment for chromosome 11 and chromosome 20. The first block in the table presents the file size for each lossy technique as well as the compression ratio.

The second block in the table corresponds to the results of SNP calling with GIAB-NIST ground truth and GATK caller.

The third block in the table corresponds to the results of SNP calling with Illumina ground truth and GATK caller.

The fourth block in the table corresponds to the results of SNP calling with GIAB-NIST ground truth and Samtools' caller.

The last block in the table corresponds to the results of SNP calling with Illumina ground truth and Samtools' caller.

CHROMOSOME 20

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	19978222245	19978222245	19978222245	19978222245	19978222245
BAM	4763279583	3951732501	4703215963	3529714410	3881179898
CRAM	2982788014	2370470617	2913269311	2116033897	2321367241
Compression	6.697835096	8.427956078	6.857664058	9.441352652	8.606230799
True positives	62964	62898	62965	62961	62964
False positives	16089	15311	16087	16379	16083
False negatives	84	150	83	87	84
Sensitivity	0.9986676818	0.9976208603	0.9986835427	0.998620099	0.9986676818
Precision	0.796478312	0.8042296922	0.7965010373	0.7935593648	0.7965387681
F-Score	0.8861865856	0.8905470171	0.8862068966	0.8843582324	0.8862240051
True positives	75783	75496	75792	75767	75775
False positives	3270	2713	3260	3573	3272
False negatives	3573	3860	3564	3589	3581
Sensitivity	0.9549750491	0.9513584354	0.9550884621	0.9547734261	0.9548742376
Precision	0.9586353459	0.9653108977	0.9587613217	0.9549659692	0.9586069048
F-Score	0.9568016969	0.9582838828	0.9569213676	0.954869688	0.9567369305
True positives	62903	62921	62905	62917	62906
False positives	17358	17421	17429	18581	17458
False negatives	145	127	143	131	142
Sensitivity	0.997700165	0.9979856617	0.9977318868	0.997922218	0.9977477477
Precision	0.783730579	0.7831644719	0.7830432942	0.772006675	0.7827634264
F-Score	0.877865312	0.8776204756	0.8774462624	0.8705464005	0.8772766575
True positives	76005	76077	76038	76363	76021
False positives	4256	4265	4296	5135	4343
False negatives	3357	3285	3323	2995	3341
Sensitivity	0.9577001588	0.958607394	0.9581280478	0.9622596336	0.9579017666
Precision	0.9469730006	0.9469144408	0.9465232654	0.9369923188	0.9459583893
F-Score	0.9523063719	0.9527250413	0.9522903034	0.9494579002	0.9518926161

CHROMOSOME 11

	Raw	Illumina	Qvz	Leon	dynamic bin
SAM	44741080610	44741080610	44741080610	44741080610	44741080610
BAM	10668938373	8830246318	10531487823	7883319664	8680171268
CRAM	6684954326	5297101246	6527041230	4726468838	5191134746
Compression	6.692802737	8.446332915	6.854726213	9.466069098	8.618747692
True positives	133497	133395	133493	133468	133492
False positives	55639	54295	55660	56328	55653
False negatives	130	232	134	159	135
Sensitivity	0.9990271427	0.9982638239	0.9989972086	0.9988101207	0.9989897251
Precision	0.7058254378	0.7107198039	0.7057408553	0.7032181922	0.7057654181
F-Score	0.8272137761	0.8303015402	0.8271454241	0.825346373	0.8271597288
True positives	182884	182414	182886	182853	182874
False positives	6252	5276	6267	6943	6271
False negatives	4151	4621	4149	4182	4161
Sensitivity	0.9778062929	0.9752933943	0.9778169861	0.9776405486	0.977752827
Precision	0.9669444209	0.9718898183	0.9668680909	0.9634186179	0.9668455418
F-Score	0.9723450239	0.9735886317	0.9723117165	0.9704774819	0.9722685948
True positives	133405	133431	133398	133401	133399
False positives	57434	57568	57325	59603	57619
False negatives	222	196	229	226	228
Sensitivity	0.9983386591	0.9985332306	0.9982862745	0.998308725	0.998293758
Precision	0.6990447445	0.6985952806	0.6994332094	0.6911825662	0.6983582699
F-Score	0.8223049565	0.8220598473	0.822555881	0.8168300008	0.8218145975
True positives	183021	183140	182952	183588	182998
False positives	7818	7859	7771	9416	8020
False negatives	4020	3902	4092	3452	4044
Sensitivity	0.9785073861	0.9791383753	0.9781227946	0.9815440547	0.9783791876
Precision	0.9590335309	0.9588531877	0.9592550453	0.9512134464	0.958014428
F-Score	0.9686725945	0.9688896178	0.9685970453	0.9661407626	0.9680897212