South Dakota State University

Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange

Electronic Theses and Dissertations

2020

Assembly of *Triticum boeoticum ssp. aegilopoides and Triticum monoccocum* Genomes

Mustafa Aljadi
South Dakota State University

Follow this and additional works at: https://openprairie.sdstate.edu/etd

 Part of the Agronomy and Crop Sciences Commons, and the Plant Breeding and Genetics Commons

Recommended Citation

Aljadi, Mustafa, "Assembly of *Triticum boeoticum ssp. aegilopoides and Triticum monoccocum* Genomes" (2020). *Electronic Theses and Dissertations*. 4086.
https://openprairie.sdstate.edu/etd/4086

This Dissertation - Open Access is brought to you for free and open access by Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. For more information, please contact michael.biondo@sdstate.edu.

ASSEMBLY OF *TRITICUM BOETICOUM SSP. AEGILOPOIDES AND TRITICUM MONOCCOCUM G*ENOMES

BY

Mustafa Aljadi

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Plant Science

South Dakota State University

2020

DISSERTATION ACCEPTANCE PAGE

MUSTAFA ALJADI

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Jose Gonzalez Hernandez
Advisor Date

David Wright
Department Head Date

Dean, Graduate School Date

ACKNOWLEDGEMENTS

This work would not have been possible without the contributions from a multitude of individuals. Thanks to Dr. Jose Gonzalez, Dr. Sunish Seghal , Dr. Hani Ghosheh , and Dr. Don Auger for sharing their knowledge throughout my project. Additionally, thanks to the many people who have helped me through the ups and downs of my project. Similarly, a special thanks to my friends and people who were nice to me during the years I spent here. Lastly, thanks to my family (my mom and my siblings) for their loving support and throughout this entire experience.

We acknowledge use of the SDSU for IT support especially Dr. Brian Moore, Kevin Brandt, Chad Julius, Luke Gassman, and Roberto Diaz. We acknowledge the fund of high ministry of education of Libya that supported me throughout my journey. The authors do not have any conflict of interest to declare.

TABLE OF CONTENTS

ABBREVIATIONS

AFLP amplified fragments length polymorphism

BACs Bacterial Artificial Chromosomes

BioNano genome mapping

CIMMYT International Maize and Wheat Improvement Center

ddNTP dideoxythynucleotide triphosphate

Gbp Giga base pairs

gDNA chromosomal DNA of an organism

DNA Deoxyribonucleic acid

SOAPdenovo novel short-read assembly method that build a de novo draft assembly

FAO The Food and Agriculture Organization of the United Nations

Illumina Technique used to determine the series of base pairs in DNA

ISSR inter simple sequence repeats

Lr Leaf rust

Mbp megabase pairs

Mbp millions of base pairs

W2RAP Wheat/Whole-genome Robust Assembly Pipeline

Uredinia reddish or black mass of hyphae and spores of a rust fungus

RNA Ribonucleic acid

RFLP restriction fragment length polymorphism

RAPD random amplified polymorphism DNA

ssDNA single-stranded DNA

SOLiD next-generation DNA sequencing technology

SINEs and LINEs Short and long interspersed retrotransposable elements

SMRT Single molecule real-time

SSR microsatellites

PCR Polymerase chain reaction

Pt P. triticina

Pgt Puccinia graminis f. sp. Tritici

Pst *P. striiformis f. sp. Tritici*

PacBio single molecule real time

pH a scale used to specify how acidic or basic a water-based solution

mRNA Mature RNA

Kbp kilo base pairs

NGS massively parallel sequencing that describe a DNA sequencing technology

ONT Oxford Nanopore sequencing

QTL	quantitative traits loci
WGS	Whole-genome sequencing
WGS	Whole-genome shotgun

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

ASSEMBLY OF *TRITICUM BOEOTICUM SSP. AEGILOPOIDES AND TRITICUM MONOCCOCUM G*ENOMES

MUSTAFA AL-JADI

2020

Wheat is ancient cereal. Wheat has n = 7 chromosomes which is belong to genus *Triticum*. The group *Triticum monococcum sp. aegilopoides*(boeoticum) is a diploid wheat (2n = 14). It has a morphology aspect of having narrow, flat spike which usually shatters before harvesting time. The domesticated type of *T.boeoticum* known as a diploid wheat of *Triticum monococcum* L. In this research we want to generate a high quality assembly of wild and domesticated type of Einkorn wheat. Our method is using Illumina short sequence reads assembled by both CLC, and W2RAP(Wheat/Whole-genome Robust Assembly Pipeline) software's. In our approach we will use *Triticum monococcum sp. aegilopoides*(boeoticum) sequence data. We will use 4 four lanes of Paired end (PE) data with an insert size of 180 bp, 300 bp, and 400 bp and we have three lanes of mate pair (MP) data for 2 kb, 4 kb, and 8 kb insert. This project would help breeders around the world to get the detailed genomic information that will help them fight diseases and increase the overall yield in new varieties. After the annotation of boeoticum genome, we predicted 658 coding genes and 1.122 transcriptomes . In addition, when we did a BLAST, we predicted 463 coding genes. On the other hand, after

the annotation of *monococum* genome we predicted 31.000 coding genes and 70 transcripts. By BLAST we predicted 49.538 coding genes

INTRODUCTION

Developing countries, are often suffering from drought, which causes a dire shortage of food supplies. Drought may cause severe damages of crops that leads to a famine in some African countries. To overcome these dire problems there is a need to have substantial breeding programs of crops to survive in drought and become resistant to certain diseases. The appearance of modern technologies in molecular biology, genetics, and cytogenetics, has helped to dramatically enhance plants yield throughout breeding programmers. A good understanding of the structure of DNA and proteins, and regulation mechanisms at single cell could provide a vital information for crop improvement and adaptation to certain weather conditions and enhance yield.

Bread wheat (*Triticum aestivum*) is a hexaploid grass species within the *Poaceae* family. Along with rice and maize, wheat is one of the three most consumed crops in the world, providing 21% of the food calories and 20% of the protein to 4.5 billion people in 94 countries (http://www.fao.org/home/en). Wheat production has been decreased globally because of diseases, for instance, black stem rust race Ug99 have destroyed wheat harvests in Eastern Africa, as well as it's spreading to the Middle East and Asia. New generation breeding techniques are important to enhance wheat yield and reference sequence of wheat genome can lead to a platform for the application of genomics mechanism in breeding.

This dissertation mainly focuses on wheat genome assembly of both the domesticated and the wild type of einkorn wheat that is considering to be the oldest type

of wheat that has been existed. We will use W2RAP assemblers. Their expected genome size will be (5Gb).

1.2 Origin and domestication of wheat

The process of domestication took place around 12,000 years ago in the Diyarbakir region in Southeast Turkey[1, 2]. Wild einkorn considered as a one-grained wheat T. monococcum L. ssp. aegilopoides was first described under the name Crithodium aegilopoides, that found it in Greece between Nauplia and Corinth in 1833. Thellung (1918), decided to name einkorn as T. boeticum, to distinguish it from the domesticated type. Wheat composed from different species with different ploidy levels such as diploid, tetraploid and hexaploid. Einkorn wheat (*Triticum monococcum*, A^mA^m genome 2n = 14) is a domesticated diploid species that come from the wild type *T. boeoticum*.

The domesticated type of einkorn wheat is located in Southeast Turkey[3]. After 5000 years, cultivation of einkorn was replaced by tetraploid and hexaploid wheats[4].

This replacement has decreased the selection pressure on the domesticated einkorn wheat varieties. The second wild diploid wheat (*T. urartu,* A^uA^u) has not been domesticated, which helped in the evolution of wheat by donating the A genome to all tetraploid and hexaploid species[4]. The genome of *T. urartu* has been recently sequenced and assembled, revealing a larger gene content than its counterparts in the tetraploid and hexaploid wheat [5].

Tetraploid wheats (2n = 28) occurred in the Middle East region such as Iraq and Syria. Two wild tetraploid species are known *T. turgidum subsp. dicoccoides* (wild Emmer with AABB genome) and *T. araraticum* (AAGG genome?). The domesticated species *T. turgidum subsp. dicoccom* and *T. timopheevii* (AAGG) come from their wild relatives.

Both wild tetraploids come from allopolyploidization events between T. urartu (and a species from the lineage of the wild wheat *Aegilops speltoides Tausch* [5]. Both domesticated emmer wheat existed in Eastern Syria, Jordan, Lebanon, and Central Eastern Turkey.

As a result, many cultivated wheat species were derived from the domesticated emmer wheat such as the polish wheat, durum wheat and the khurasan wheat [5].

Durum wheat is the second most important species after bread wheat. It comes from domesticated emmer wheat in the eastern Mediterranean region [2, 6]. It has big a genome: about (12Gbp), Moreover; it has a high percentage of paralogous genes [6].

Hexaploid wheat originated in either northwestern Iran or Turkey by hybridization between tetraploid wheat and diploid *Ae. Tauschii*, as well as by chromosome doubling. The outcome of this combination was bread wheat with a genome of AABBDD.

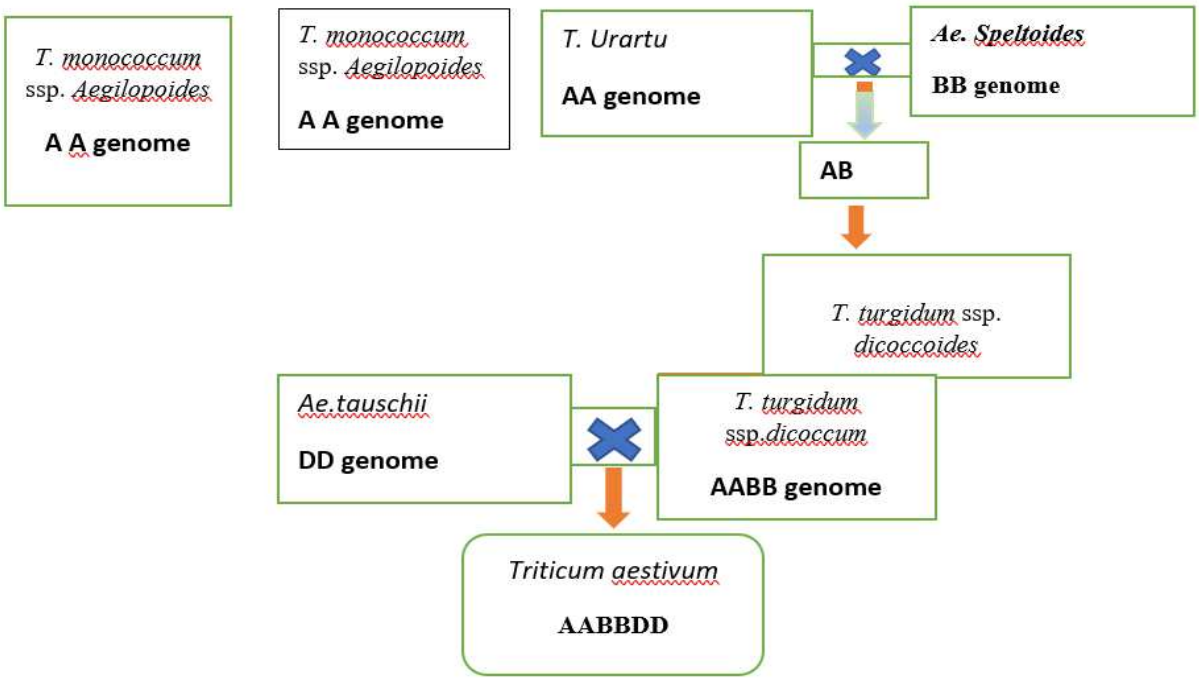


Fig 1 Simplified scheme of the probable origin of wheat[7]

1.2.1 Wheat Domestication

What domestication occurred as a loss of shattering of the spike at maturity, which led to a seed loss at harvesting. As a result, seed dispersal was enhanced in natural populations. This mutation occurred in the *Br* (brittle rachis) locus. The second domestication is the change from hulled forms, so that glumes adhere is attached to the grain, to free-threshing naked forms. This dominant mutation occurred at the *Q* locus. As a result, change the influence of recessive mutations at the *Tg* (tenacious glume) locus.

Durum and bread wheat behaves as a diploid chromosome during meiosis because of the effect of a dominant gene *Ph*1 in chromosome 5B that controls the paring of homologous chromosomes, which prevents pairing between the homeologous chromosomes [8].

The expansion of wheat agriculture from the Fertile Crescent to new places such as Europe, and Africa is demonstrated by archaeological studies. The main influx of wheat spread to Europe occurred about 8,000 BC from Anatolia to Greece.

1.2.2 Why sequence the genome of *T. boeoticum*?

The genome of *T. boeoticum* is more diverse than common wheat because it has a wild type. Other wheat species such as *T. urartu* and *T. astivum* are less diverse because hybridization occurred only a few times among those species to come up with bread wheat. It is very important to have those genes in common wheat so as to increase the yield and resistance to different types of diseases. Genome A is a very important element of the bread wheat genome, as well as *T. turgidum* (AABB), and *T. timopheevii* (AAGG*). T. boeoticum* genome assembly will contribute as a diploid reference for analysis of polyploid wheat genomes. Also, it provides an important resource for the genetic improvement of wheat. Moreover, sequencing of *T. boeoticum* helps to discover any other diploid homologous relationships. The *T. boeoticum* assembly would significantly help in the development of genetic markers. Those molecular markers are used often in breeding to produce new varieties. Thus, sequencing the genome of *T. boeoticum* would accelerate our understanding of genomic and breeding studies of bread wheat, increase the yield of wheat to match the dramatic increase of world population, and sustain food security globally[9].

1.3 Agronomic importance

Common wheat, (*Triticum aestivum*) is very important for food crops throughout the world (FAO, 2016). It is widely used as a cereal, taking up one sixth of the crop acreage in the world (Gupta et al., 2008). In 2017 about 750 million tons of wheat was

produced worldwide, and has the best yield in Mediterranean basin, southern Russia and the central United States. Wheat is the main source of protein and calories for 35% of the world population, providing one fifth of the total calories consumed by humans [10]. Wheat has starch (60-80%), proteins (8-15%), vitamins A, B12, C, and iodine[11].

Bread wheat makes up 93% of total cultivated area compared to durum wheat. Wheat can survive under different climatic zones. According to archaeological studies and DNA fingerprinting, wheat first appeared in the Neolithic south-eastern Turkey as well as northern Syria about10,000 years ago[3, 12]. Einkorn and emmer wheat are the oldest cultivated wheat species.

1.3.1 Plant Genomes:

The first known genome for conception was by Hans Winkler in 1920. He defined genome as "all genetic materials that are highly necessary to form and maintain an organism" (Winkler 1920). In eukaryotes, most of the genetic information is in the nucleus. Eukaryote species tend to have a single nucleus. In contrast, protozoa, fungi, and some plant tissue can have multinucleated cells [13, 14]. Besides nuclear DNA, DNA can also be found in cytoplasm organelles such as mitochondria and plastids of plants.

The genome size of a plant species ranges from 63 megabase pairs (Mbp) in carnivorous plants to 149,000 Mbp in canopy plants (9– 11). (see Table 1.1). For example, grass genomes, such as maize (2,300 Mbp), barley (5,428 Mbp), rye (8,093 Mbp), and wheat (17,100 Mbp) are proximately 5.5-fold the size of the human genome.

Table 1. Shows the genome sizes and chromosome numbers of selected plant and non-plant organisms. The organisms are ordered based on genome size in Mbp per haploid set of chromosomes [19]

Common name	Scientific name	Chromosome number (2n)	Genome size (1C in Mbp)	ploidy level
E. coli	*Escherichia coli*	2	4.6	diploid
yeast	*Saccharomyces cerevisiae*	32	12.1	diploid
carnivorous Genlisea	*Genlisea margaretae*	52	63	diploid
thale cress	*Arabidopsis thaliana*	10	150	diploid
duckweed	*Spirodela polyrhiza*	80	158	diploid
purple false brome	*Brachypodium distachyon*	10	355	diploid
rice	*Oryza sativa*	24	489	diploid
sorghum	*Sorghum bicolor*	20	730	diploid
human	*Homo sapiens*	46	3,100	diploid
barley	*Hordeum vulgare*	14	5,428	diploid
rye	*Secale cereale*	14	8,093	diploid
wheat	*Triticum aestivum*	42	17,100	hexaploid
Norway spruce	*Picea abies*	24	19,570	diploid
canopy plant	*Paris japonica*	40	149,000	octoploid

Fig 2: Genome size can dramatically increase due to polyploidy and segmental duplications, which can make up a large fraction of entire plant genomes(10-90) [15, 16]

1.3.2 Plant nuclear genome organization

The main content of plant nuclear genomes are genes, regulatory sequences, other non-coding sequences, and different types of repetitive DNA [17, 18].

The term gene is understood to be a segment of DNA that is the basic functional unit of inheritance controlling the transmission and expression of functional product. Genes have different variant forms known as alleles. The chromosome term comes from the Greek words for color (chroma) and body (soma). Chromosomes were first observed

in plant cells by Swiss botanist Karl Wilhelm von Nägeli in 1842[19]. Chromosome structure has a long linear DNA molecule, that is coiled tightly around proteins [29]. All chromosomes have a centromeric region which divides chromosome into shorter arm p and the longer arm q. (CI), which relates the ratio of short arm length to the total length of a chromosome. There are four types of chromosomes: metacentric, submetacentric, acrocentric and telocentric.

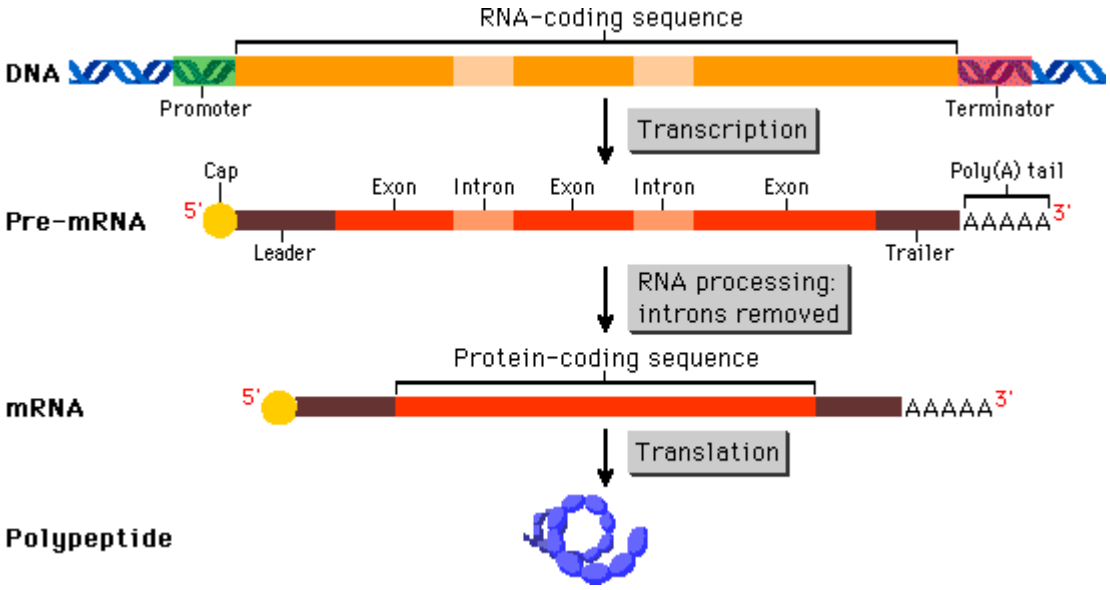


Fig 3 : Eukaryotic protein-coding genes have introns and exons. The transcript of these segments is known as pre-mRNA. This pre-mRNA mainly occurs in the nucleus to eliminate the introns and splice the exons together to form a translatable mRNA. The pre-mRNA exits the nucleus then moves to be translated in the cytoplasm (Adapted from http://www.phschool.com.)

Gene structure and gene expression in eukaryotes are very complicated compared to the genomes of simple prokaryotes such as bacteria. Eukaryotic genes contain coding

sequences (exons) and noncoding sequences (introns). Both exons and introns are transcribed into the precursor mRNA (pre-mRNA). Throughout the Splicing process all introns are discarded. All exons combine to form the messenger RNA (mRNA). Then the mRNA is modified by the addition of the 7methylguanosine 'cap' to the 5′ end 14 and a poly(A)tail to the 3′ end, so that it becomes ready to move from nucleus. The promoter has short regions which simulate and regulate the transcription of a gene [20].

The term "gene family" refers to a group of genes that come from a common ancestor. These genes, known as homologous genes, can be divided into orthologs and paralogs. Orthologs generally have a similar function and are present in many species. On the other hand, paralog genes have a new function.

1.4 Repetitive DNA in eukaryotic genome

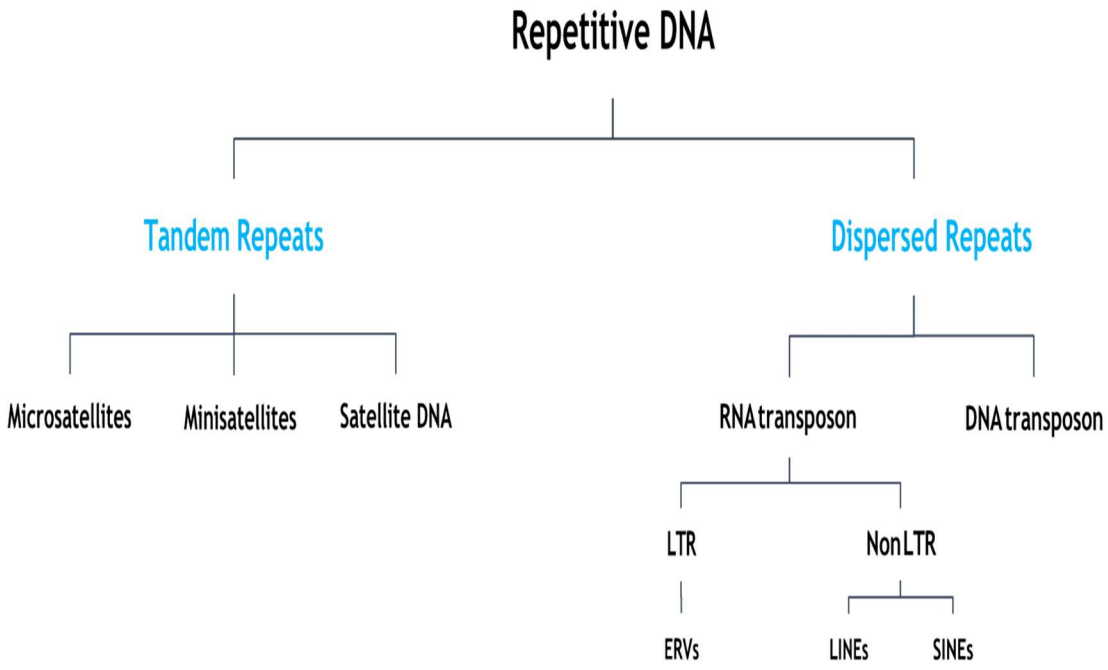


Fig 2: Schematic diagram of repetitive DNA classification[21]

The repetitive DNA sequences cause variation in genome size among organisms [22, 23]. Repetitive sequences stabilize chromosome movement and pairing [24]. Repetitive sequences can be divided into two groups: tandem repeats and dispersed repeats.

1- **Tandem repeats** are made of short (≥ 2bp in length) non-coding consecutive sequences, their units organized in a head to tail orientation. Tandem repeats can be classified based on their copy number of the basic repeat units, length and genomic location, as follows:
2- **Satellites DNA** can be varied in length from 5-300 bp depending on the organism and repeats about 105- 106 times. The second type is Minisatellite, which has average length of 10-60 bp, and repeats from 20-50 times. They can be used for DNA fingerprinting as well as for genetic markers in linkage analysis and population studies. The third type is Microsatellite, known as Simple Sequence Repeats (SSRs), its length is 1-6 bp, the number of the repeats varies from 10-100 times[21]. SSRs are used as cytogenetic and molecular markers due to their high frequency in eukaryotic genomes [23, 25].

Dispersed repeats are known as Transposable elements, TEs, or transposons. They are sequences of DNA scattered within the genome, which can jump from one locus to another in the genome. Dispersed repeats are classified based on transposition methods into two classes called **DNA transposons and RNA transposons**

DNA transposable elements class II (DNA transposons) use a cut-and-paste mechanism by a double-stranded DNA break, and do not depend upon RNA intermediate to move. While, **RNA transposable elements** of class I, known as retrotransposons (RNA transposons), expand by copy-and-paste through an RNA intermediate, and increase in genome size due to polyploidy and segmental duplications. They make up an enormous fraction of plant genomes, their amount varies widely from 10% up to 90% of the entire genome [21].

RNA transposons have two categories, LTR and non-LTR. LTR that repeats several hundreds of times and links both ends of the genomes. LTR retrotransposons are responsible for many of genetic variations. The second type is Non LTR elements, which can be classified into two sections: long interspersed elements (LINEs) and short interspersed elements (SINEs). Non-LTR retroposons are prevalent in eukaryotic genomes. LINEs have a length of (6-8 kbp) that covers up to 21% of human genome. On the other hand, SINEs do not encode a reverse transcriptase, but they depend on LINE encoded enzymes for transposition[26].

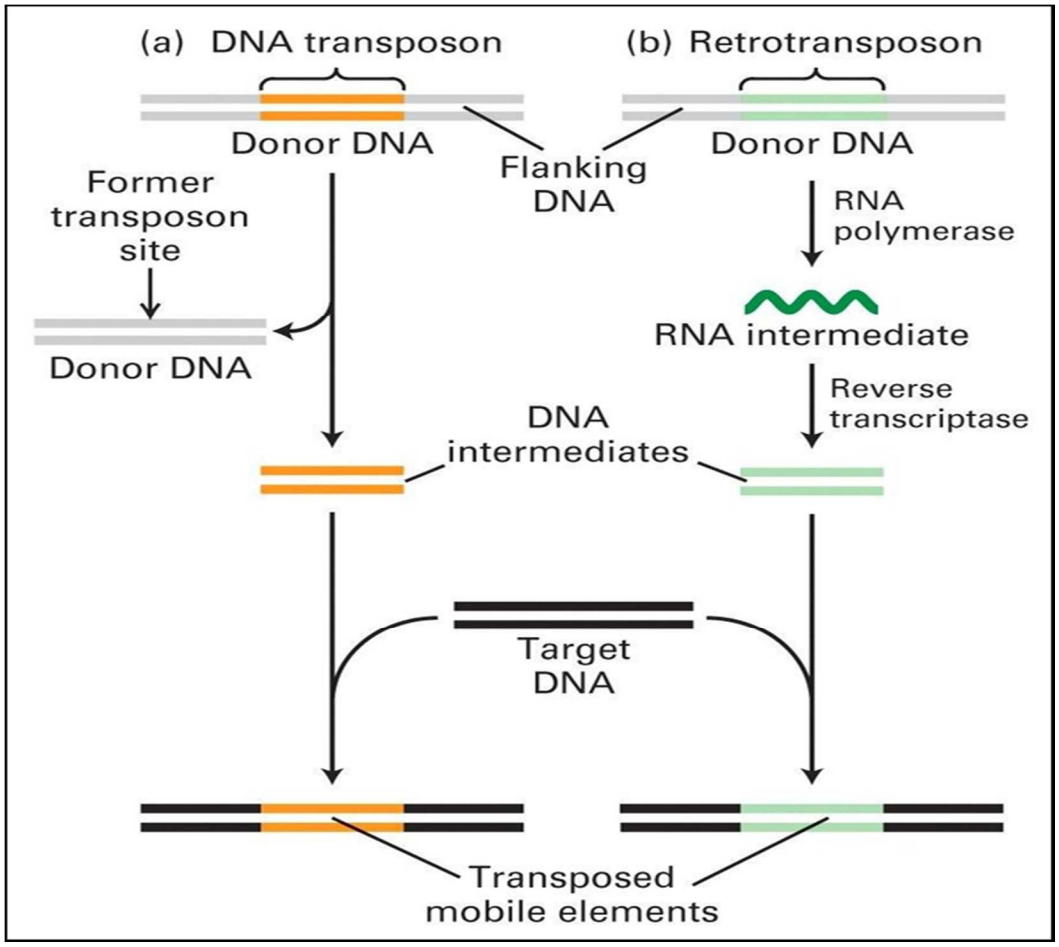


Fig 4: Shows the DNA and RNA transposons and how they move through the genome. Lodish et al., Molecular Cell Biology, 7th ed[27].

1.4.1 Plant genome sequencing and assembly

Sequencing of some plant genomes can be challenging due to a large genome size, which occurs in Maize and Wheat. In contrast, with smaller genomes, such as *Arabidopsis thaliana* (~150 Mbp) [28]. Sequencing becomes difficult because of their highly repetitive nature (>80%). This challenge has been solved because of new technologies. For example, next- generation sequencing and flow cytometry [29-31] has paved the way for sequencing complex genomes, increasing the number of sequenced base pairs (bp) from thousands to millions. Next- generation sequencing (NGS) can

provide read length (26-150) bp (Illumina sequencing) NGS based on detection of light and produce short (75-1,000 bp) reads. While read length can be expanded to (10-20) kbp using PacBio sequencing [29], the disadvantage of these new mechanisms is storing and assembling the large amounts of data generated by NGS[32].

1.4.2 Genome Sequencing Approaches

Sequencing approaches can be classified into two methods: The first one is whole-genome shotgun, while the second one is clone-by-clone. The whole-genome shotgun sequencing (WGS) depends on sequencing and assembly of randomly selected fragments of genomic DNA; WGS is widely used for plants with small genomes like black cottonwood (~500 Mb) and woodland strawberry (240 Mb) [33]. It can be useful for sequencing of wheat genome with a low coverage for gene assembly and SNP detection [34]. WGS becomes increasingly challenging for the genomes that have high content of repetitive sequences and complex genomes such as wheat, which has a genome of 17 Gb.

Interestingly, in the clone-by-clone method the genomic DNA is broken up into large fragments, 150 kilobases long. The location of these chunks on each chromosome is mapped to help with assembling them in order after sequencing. Those chunks are inserted into Bacterial Artificial Chromosomes (BACs) and put inside bacterial cells to grow. The DNA in the individual bacterial clones is broken down into smaller, overlapping fragments. Every clone has about 500 base pairs for sequencing. Then those fragments are put into a vector, which has a known DNA sequence, the DNA fragments are then sequenced, starting with the known sequence of the vector and unknown sequence of the DNA. The small Illumina fragments of DNA are aligned together by

identifying areas of overlap to reform the big fragments that are inserted into the BACs. This assembly is carried out by computers that identify areas of overlap [35]. Clone-by-clone technique was used to sequence the human genome, as well as rice 400 Mbp and maize 2.3 Gb [36].

1.4.1 Sanger sequencing

The first sequencing method was discovered by Sanger in 1977, and was based on synthesis by primer and DNA polymerase I. It incorporates deoxyribonucleoside triphosphates and 2',3'-dideoxythynucleotide triphosphate (ddNTP), which can terminate the reaction. Those four ddNTP are labelled with different fluorescent dyes, then the fragments get separated by capillary electrophoresis based on their sizes. The average length of reads is up to 800 bp[37].

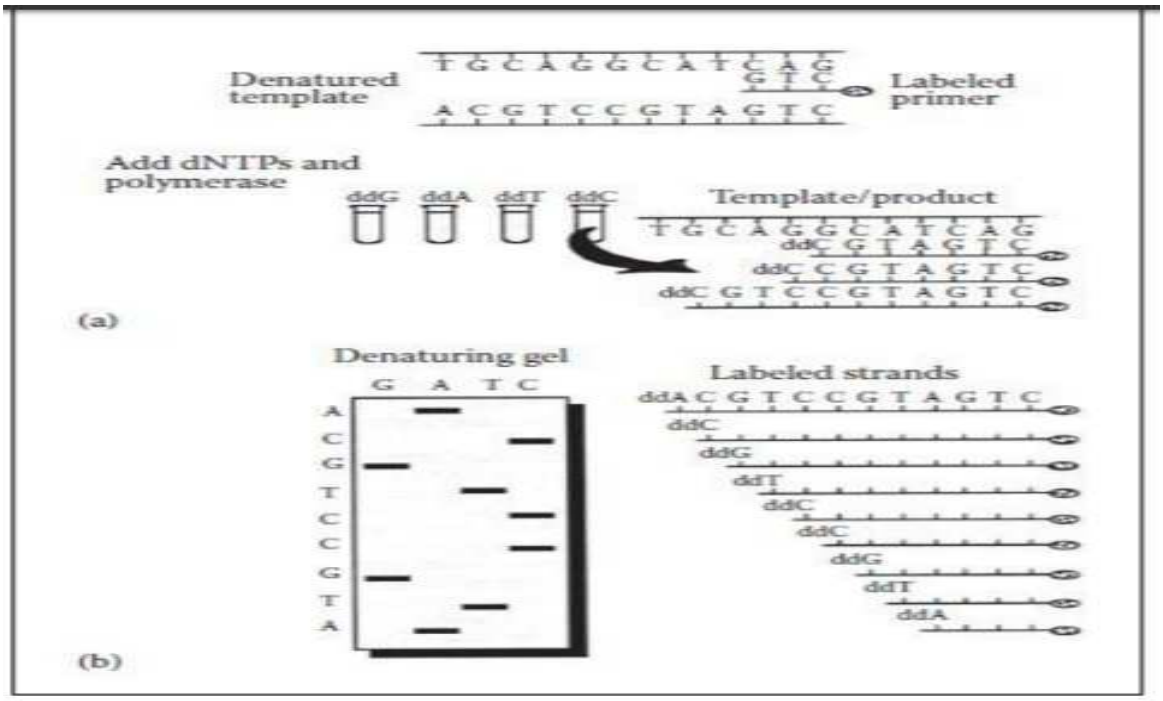


Fig 5 performed by the presence of denatured DNA template, radioactively labeled primer, DNA polymerase, and dNTPs. The DNA polymerase helps in incorporate the

dNTPs into the elongating DNA strand. The four dNTPs are then run in a separate reaction so the polymerization can randomly terminate at every single base position. The result of each reaction is a population of DNA fragments with different lengths, with the length of each fragment dependent on where the dNTPs are incorporated. (b) shows the separation of these DNA fragments by using a gel electrophoresis.

1.4.2 454 sequencing:

454 sequencing is one of the first next generation technologies. The single strand DNA fragments are annealed to capture beads. Beads with PCR reagents are emulsified to create a small micro-reactor. Beads with bound amplified fragments are released from the emulsion and loaded onto Picotiter Plate. 454 can make long reads which are easier to map to a reference genome. But this approach has disadvantages because it is unable to identify errors detected of sequencing insertions and deletions because of homopolymer regions. Signals with too high or too low intensity cause under or over estimation of the number of nucleotides[37].

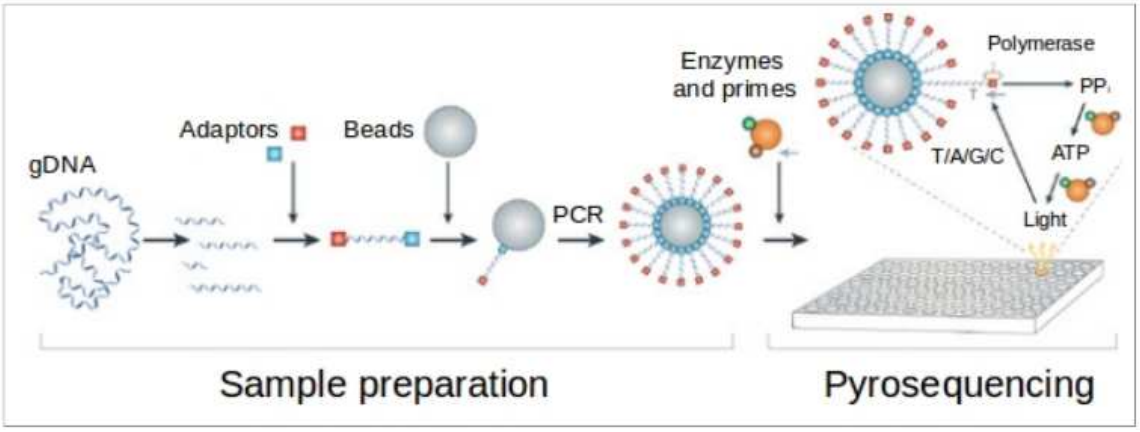


Fig 6 Sample for sequencing is prepared by fragmentation of gDNA and ligation of specific adapters. Resin beads are added to reaction and DNA sequences bounded on the

beads are complementary to the adaptor sequence. Captured fragments are amplified in micro-reactors, containing enzymes and primers. Beads are then moved to a plate. The deoxy nucleotides (dNTPP).

are alternately added to the reaction and a light pulse is detected because of pyrophosphate release, if the nucleotide is incorporated into synthetized strand.[38]

1.4.3 Illumina

During Illumia sequencing, tagged fragments are amplified by primers that include adapters for sequencing. The library is then diluted, denatured and ssDNA is hybridized to oligonucleotides. Bridge amplification is used for clonal multiplication of DNA fragments. Then hundreds of the same strands of DNA are formed into clusters on the top of the chip. Those clusters are sequenced. Every base is detected in each sequencing cycle. Each one of the nucleotides has different color, then primers and DNA polymerase are added. Laser induced fluorescence is captured, a nucleotide in every cluster is identified, and reversible terminators with dye are cleaved. For this process, the error rate is low. Illumina short read sequencing starts with ligation of DNA fragments onto flow cell fixed adapters. Bridge amplification of a single DNA fragment results in a dense cluster of sequences. Next, the sequencing by synthesis step is initiated. Therefore, fluorescence labeled nucleotides are synthesized onto single stranded fragments. Upon integration, clusters emit a distinct light for each nucleotide. Fluorescent emissions are measured by specific lasers and the sequence can automatically be inferred. The fluorescence labels are cleaved and washed away so that the process can be repeated until the complete sequence of a fragment is known.[37].

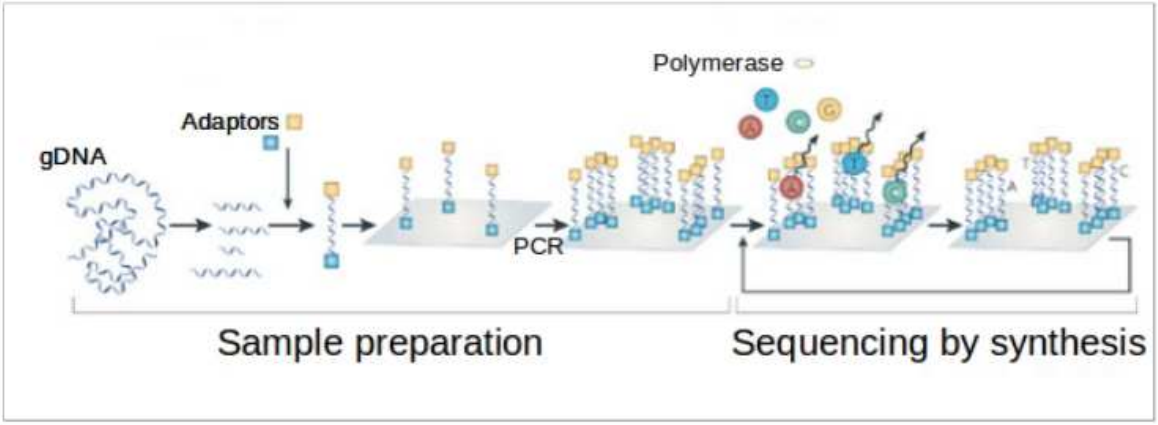


Fig 7 The sample is prepared by fragmentation of gDNA and ligation of specific adapters that are complementary to oligos on a solid surface. Clusters are formed by bridge amplification and sequencing is done by synthesis with fluorescently labelled nucleotides. The reversible terminators of synthesis with dye are cleaved after each sequencing cycle.[38]

1.4.4 SOLiD Sequencing:

Prepared libraries are captured to the beads and amplified in microreactors using an emulsion PCR template that has the beads bound to a glass slide. Sequencing runs on the slide. After hybridization of primer to DNA, ligase join one of four labelled probes with defined di-nucleotide based on its compatibility with the template. Signal detection, cleavage of a part of the probe and another ligation follow. After the cycles process is done, the primers and probes are removed; simultaneously new primer is annealed to template. The new primer anneals to the template but is shifted by one nucleotide towards 5' end compared to previous one. SOLiD sequencing has a very low rate of error. Moreover; it's a powerful tool to detect mutation and variant discovery by whole-genome resequencing or sequence/exome[37].

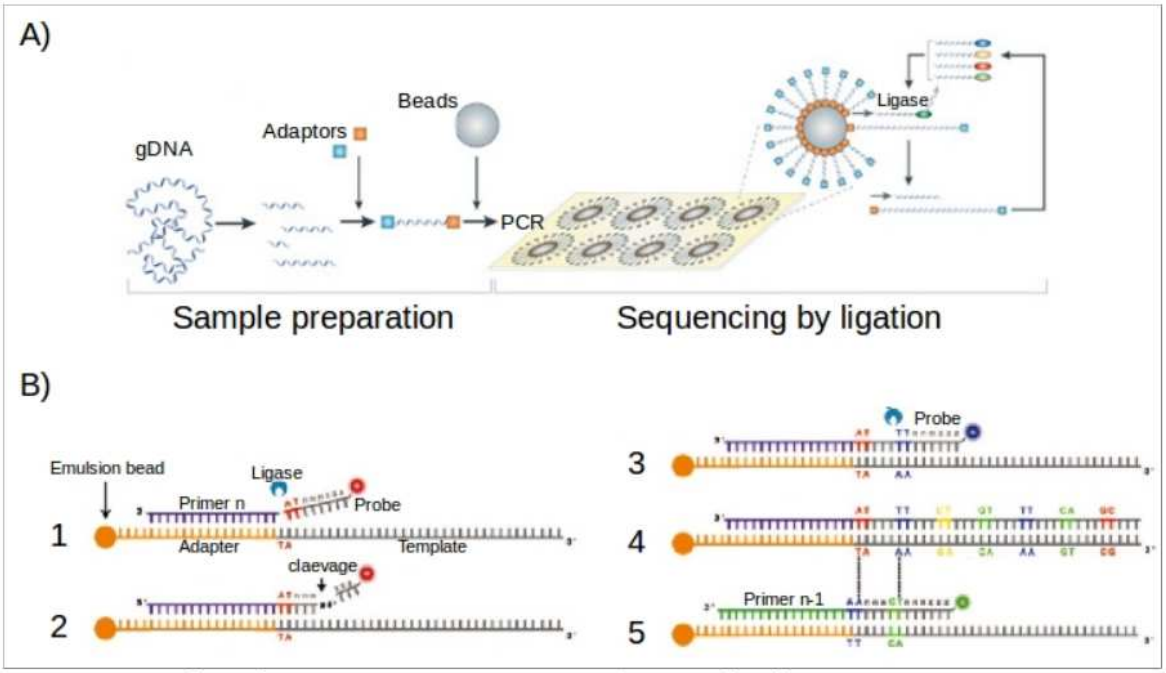


Fig 8 **A)** a sample for sequencing is prepared by fragmentation and ligation of adaptors. The fragments are amplified on the surface of beads, and the beads are moved onto glass slide and fragments are sequenced by ligation. **B)** Sequencing is initiated by ligation of labelled oligonucleotide probes to primer (1). Part of probe is removed (2) and the other oligonucleotide probes are successively ligated (3-4). The library is then denaturized and the whole process is repeated with new primer moved one base towards 5'end. [38]

1.4.5 Ion Torrent:

Ion Torrent has a similar method to 454; the only difference is a change in pH is detected. After emulsion PCR, beads with amplified DNA fragments are loaded into proton-sensing wells on a semi-conductor chip. Four bases are added sequentially, then monitor any fluctuation in voltage in every well. These changes occur because of the change of pH caused by hydrogen ions that attach the free nucleotides to DNA. Ion

Torrent can be useful for sequencing small genome as well as exome sequencing and whole transcriptome analysis[37].

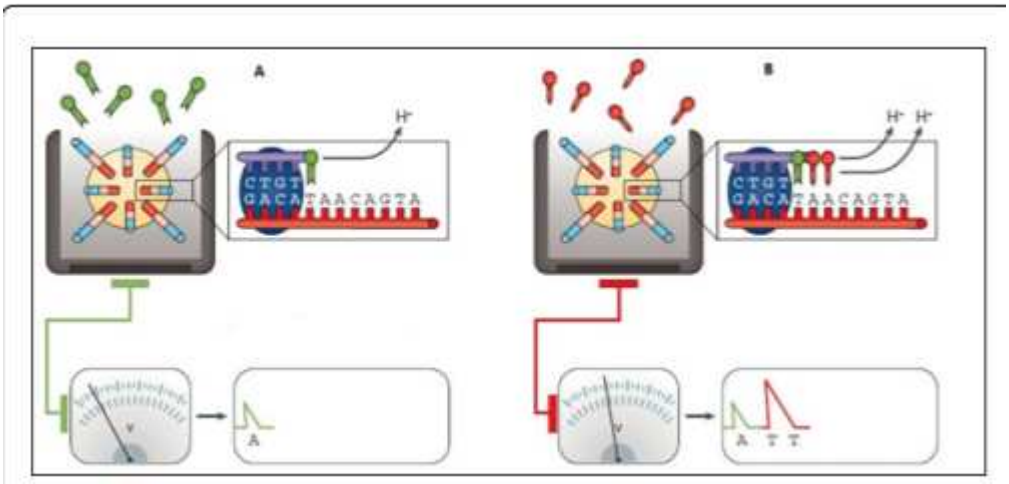


Fig 9 Ion Torrent uses a chip that contains a set of micro wells. Each has a bead with several identical fragments. Each nucleotide is incorporated as a fragment in the pearl, a hydrogen ion is released which changes the pH of the solution. This change is detected by a sensor attached to the bottom of the micro well, then can be read by a voltage signal which represents the number of nucleotides incorporated [40]

1.4.5 PacBio:

PacBio is known as single molecule real time (SMRT) sequencing. Libraries are prepared from amplicons or sheared genomic DNA. Single-stranded hairpin adapters are then ligated on the repaired ends and closed circles are formed. This kind of sequencing is done by synthesis. A single molecule of DNA polymerase is anchored at 28, the bottom of every nanophotonic visualization chamber, and a single molecule of DNA is sequenced in every well. Then the fluorescent tag is released[37].

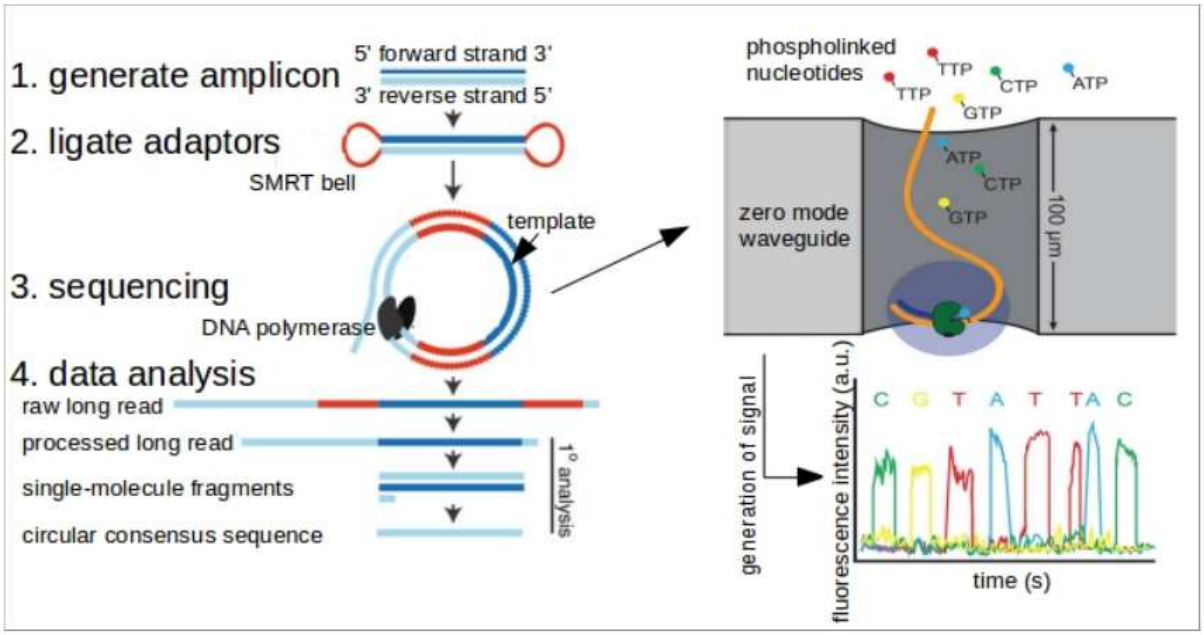


Fig 10: To the amplicon fragments (1,) hairpin adapters are ligated (2). Sequencing is done by synthesis with fluorescently labelled nucleotides and polymerase is anchored at the bottom of nanophotonic visualization chamber. (3). The adaptors are removed, and strands are resolved during data analysis (4).[39]

1.5 Oxford nanopore sequencing

Oxford Nanopore sequencing (ONT) is used to determine the order of nucleotides in a DNA sequence. A single molecule of DNA or RNA can be sequenced without PCR amplification.

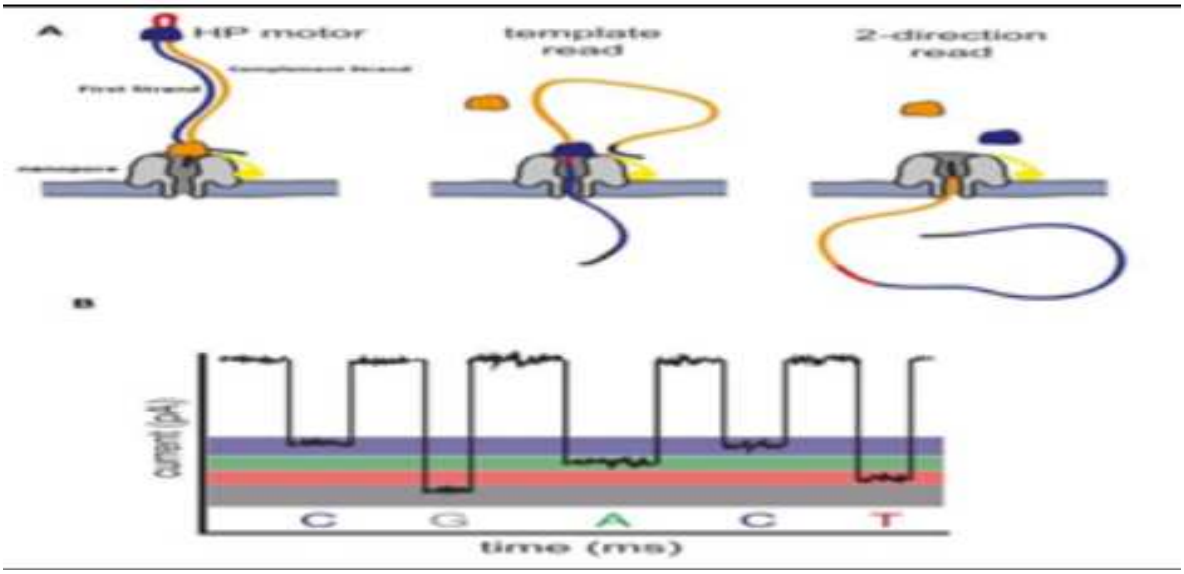


Fig 10 To the amplicon fragments (1,) hairpin adapters are ligated (2). Sequencing is done by synthesis with fluorescently labelled nucleotides and polymerase is anchored at the bottom of nanophotonic visualization chamber. (3). The adaptors are removed, and strands are resolved during data analysis (4).[39]

1.5.1 Challenges of assembling the wheat genome

The high repetitively and ploidy level of plant genomes make it hard to have an accurate assembly, because the identical sequences might collapse on top of one another[41]. As a result; an accurate assembly is often restricted to the low-copy regions, which results in highly fragmented draft genome assemblies[30] [42] . The assembly quality depends on very low DNA contamination and sequencing errors.

There are many assembly algorithms, such as SOAPdenovo [43] and Abyss, but they have failed to overcome these obstacles[32] [44] [45].

At the beginning of sequencing, big genomes, such as barley [46] [47] and wheat [30] [34], at low-coverage (~1 to 5-fold), are generated by using the Roche 454 sequencing

platform. Due to the low sequence coverage, the assembly was is good. As a result, new approaches to evaluate the gene content should be explored.

1.5.2 The advantages of new Annotated reference genome

Facilitating an easy access sequence-level information to scientists' community to identify any development in the genomes. In addition, finding the exact regulatory regions, can make QTL easier and faster. The full annotated genome will help make more DNA marker platforms to enhance the process of plans. Also, made it easy to identify the coding and noncoding DNA that exist inside the A, B, and D sub genomes. The assembly covered 94% of the wheat genome: discovering 107,891 high-confidence gene models. Moreover, it helped to identify all developmental stages of wheat by using co-expression networks[48].

1.5.3 Current methods to assemble wheat genome

1- The most common methods, as I mentioned above are second generation sequencing NGS combined along with whole genome shotgun or BAC- BAC strategies.

2- There is another approach, which is using single molecule real-time (SMRT) sequencing to generate longer sequence reads about 40–50 kb long. This will help to generate assembly with very few gaps, as well as longer contigs, which can be a fantastic pick to assemble wheat.

3- To enhance the mapping and assembly performance of BioNano genome mapping, as well as 10× Genomics linked reads. Three-dimensional architecture of chromosomes (Hi-C) data helps in terms of assessing the quality of genome assembly and scaffold order on chromosomes[49].

ABSTRACT

Wild Einkorn wheat *(Triticum monococcum subsp. boeoticum*), and its domesticated counterpart (*T. monococcum subsp. monococcum*), is one of the progenitors of wheat. We executed whole genome de novo sequencing using an assembler W2rap to evaluate the differences between wild and domesticated Einkorn while, discovering protein-coding genes. In this paper, we present the generation, assembly, and analysis of a whole-genome shotgun draft sequence of both wild and domesticated wheat. To explore the overall conservation between modern and wild wheat genomes. After the annotation of *T.boeoticum* genome, we predicted 658 coding genes and 1.122 transcriptomes . In addition, when we did a BLAST, we predicted 463 coding genes. On the other hand, after the annotation of *T. monococum* genome we predicted 31.000 coding genes and 70 transcripts. By BLAST we predicted 49.538 coding genes.

Introduction

Common Wheat (*Triticum aestivum L.*) is one of the top 5 crops globally. Wheat is widely known as one of the oldest domesticated plants in history. Wheat is an allohexaploid (AABBDD) genome that comes from two hybridization events. The first event happened about (0.5–-3.0) million years ago, when two diploid ancestral species form *tirticea* family, one with the A genome (*T. Urartu*) and B (an unknown species) genome, became hybridized following chromosome doubling. As a result, the wild emmer tetraploid wheat (*Triticum turgidum ssp. dicoccoides,* AABB) occurred. The domestication of Emmer Wheat gave rise to (*T. turgidum ssp. dicoccum,* AABB). The second hybridization happened 9000 million years ago when cultivated emmer and diploid goat grass (*Aegilops tauschii*, DD) hybridized together to create the allohexaploid common Wheat (AABBDD) genome 17 Gb [50]. Wheat is widely considered the first domesticated crop and is the prime polyploid species among other crops. Since ancient civilizations in areas such as Egypt and Iraq, wheat was a primary staple source of food [1].

Polyploidy is a crucial milestone of the domestication of cultivated crop species. About 25–35% of plant species are polyploids [2]. Wheat production demands improving wheat genetic diversity, which remains a massive setback due to the lack of knowledge of wheat biology and the molecular basis of use of agronomic traits. By overcoming those obstacles, wheat production will dramatically increase to meet ever increasing human consumption demands generated by world population increase [3].

Wild einkorn is one of Wheat varieties *T. monococcum* L. ssp. aegilopoides. Wild einkorn has a domesticated type known as, *T. monococcum* ssp. *monococcum,* which has

a brittle rachis trait. *Triticum monococcum* spp, *aegilopoides* widely occurred in the Middle East and some parts of Europe, such as Greece and southern Bulgaria. Currently, more attention has been focused on wild and cultivated einkorn wheat because it may be a reliable source for wheat genetics and utilized to detect any mutant from Einkorn wheat to identify mutant alleles [51] [52]. Sequencing cultivated einkorn wheat is a significant step in identifying agronomically essential genes based on genome-wide association with molecular markers [53]. Moreover, it would significantly enhance human health by producing new varieties of wheat that contain a desired content of nutrients [50].

It has been stated that wild einkorn wheat can be used as resource to enhance disease resistance and grain quality in common wheat [54, 55] [56, 57] . The *triticeae* family has a unique structure based on high DNA repeats that can be up to 80%–90% of the whole genome; these help in building of a structural level in plants. The wheat genome contains three sub genomes A, B and D that have similar TE compositions. Our genome assembly provides a vital source of diploid reference for analysis of polyploid wheat genomes that would help in improving wheat. Lack of genome sequence for the three homeologous and highly similar bread wheat genomes (A, B, and D) has impeded expression analysis of the grain transcriptome.

Using wild Wheat relatives to enhance cultivars quality is commonly used term. For instance, *Triticum turgidum ssp. dicoccoides* (2n = 4x = 28, AABB), the wild relative of durum wheat.

T. dicoccoides populations possess an outstanding genetic diversity for agronomic traits such as grain micronutrient content and biotic stress resistance [58, 59]. The genome of the domesticated wheat *T. monococcum* is useful as a model for the A genome of

hexaploid wheat [60]. In the current study, we report a comparison in terms of genetic diversity between wild and domesticated Einkorn.

Material and Method

The sequence for the *T. boeticoum* wheat gene assembly was generated using Illumina technology. DNA sequence was generated by University of Nebraska . Sequencing libraries were constructed and sequenced on Illumina next-generation sequencing platforms (GAII and HiSeq -2000). High-quality reads were assembled with SOAPdenovo3.

Results

The genome coverage of wild and domesticated Einkorn:

We sequenced the wild Einkorn wheat using a whole genome shotgun on the Illumina HiSeq-2000 platform by using denovo sequencing method and an assembler called W2rap (Wheat /Whole-genome Robust Assembly Pipeline). We estimated the genome size of *T.boeticoum* to be 4.9 Gb. The genome assembly reached about with contigs of N50 size 2354 kilobases (kb). After gap closure, the assembly was covered 81% with a scaffold. The N50 length of 6351 kilobases (kb). Genome annotation of the assembly was performed using the CLC Workbench.

Table 1 Assembly Results of *T. monococoum*

Genome assembly	Estimated genome size	4.5 Gb
	GC content	45.99%
	N50 length (contig)	2920
	Longest contig	39557
	Total length of contigs	322652
	N50 length (scaffolds)	9744
	Longest scaffold	124868
Protein-coding genes	Predicted genes	31.000
	Average transcript length	70.194

Table 2 Assembly Results of *T.boeoticum*

Genome assembly	Estimated genome size	4.5Gb
	GC content	42.64%
	N50 length (contig)	2293
	Longest contig	56404
	Total length of contigs	657809209
	N50 length (scaffolds)	8565
	Longest scaffold	87188
Protein-coding genes	Predicted genes	658
	Average transcript length	1.112

The quality of the *T.boeticoum* assembly was representing 1440 of single copy orthologs genes. Of the BUSCO_ v3_v2 genes, 66.4 % (951) classified as a complete. Also, there are 15 % (216) consider as Complete+ Partial genes. The number of missing genes was 273 (19 %). In addition, the average number of ortholog core genes is 1.11 which indicates that *boeoticum* is diploid genome. The percentage of genes coverage is 81.04%. On the other hand, out of 1440 total number of *T.monoccoum* , there was 951(66.04%) as complete genes, 1359 (4.38%) as a Complete+ Partial, and 81 (5.62%) as messing genes**.** While the percentage of genes coverage is 94 % which is a significant result. As *T. boeoticoum* genome, 1.11 was the number of ortholog in core genes that proves *T.monoccocum* is a diploid genome.

Table 3: Completeness Assessment Results of boeticoum:

Total number of core genes queried	1440
Number of core genes detected	
Complete	951 (66.04%)
Complete+ Partial	216 (15 %)
Number of missing core genes	273 (19 %)
Average number of ortholog per core genes	1.11
% of detected core genes that have more than 1 ortholog	10.62
Scores in BUSCO format	C:66.0% [S:59.0%, D:7.0%],F:15.0%,M19.0%

BUSCO of *boeoticum*

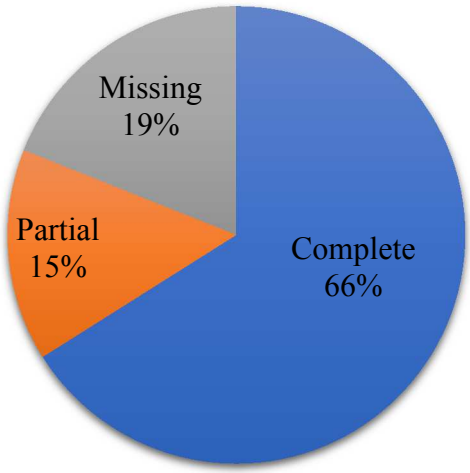


Fig 11: BUSCO (Benchmarking Universal Single-Copy Orthologs) of *T.* boeoticum

Table 4: Completeness Assessment Results of *T.monococcum*:

Total number of core genes queried	1440
Number of core genes detected	
Complete	1280 (88.89%)
Complete+ Partial	1359 (94.38%)
Number of missing core genes	81 (5.62%)
Average number of ortholog per core genes	1.11
% of detected core genes that have more than 1 ortholog	10.62
Scores in BUSCO format	C:66.0% [S:59.0%, D:7.0%],F:15.0%,M19.0%

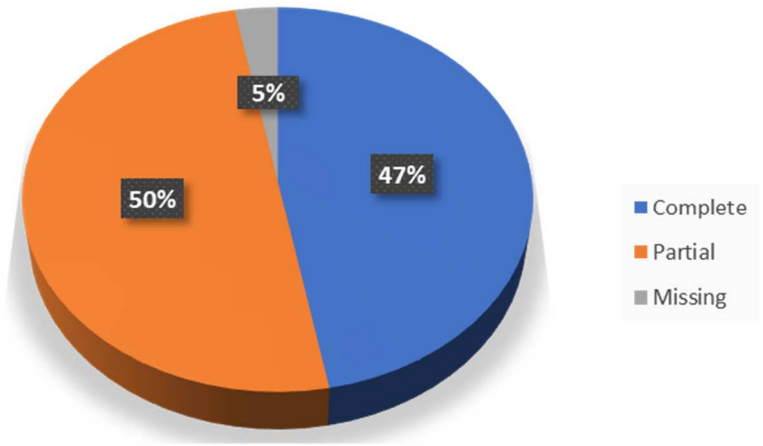


Fig: 12 BUSCO (Benchmarking Universal Single-Copy Orthologs) of *T.monococcum*

Assembly statistics of *T.monoccocum* and *T.boeoticum*

In terms of scaffolding assembly both, *T.monoccocum* and *T.boeoticum* include N and without N are close to each other because of the trimming we did before the assembly. Also, the number of scaffold number, longest seq and Shortest seq is close as it shown in the table (5). In terms of the minimum number of contig length such as N 50, N 80, and N 90, both of them have close results as it shown in the table(6). The contig assembly of *monoccocum* showed 124868 of long seq, which is less value comparing to the scaffold assembly because of the using the mate pairs to connect the contigs together. The number of contigs in scaffolds is 171534. In contrast, the number of contigs not in scaffolds is 322652. The average number of contigs per scaffold is 3.4

Interestingly, *T. monoccocum* and *T.boeoticum* genomes have very close number of GC_Content 46 and 43% respectively. The contig assembly of *T.boeoticum* showed 56404 longest-seq more than 39557 of *monoccocum.* Both genomes have the same shortest contig number of 200. In terms of how much those contigs cover the genome, there were not much difference between the two genomes as shown in table (7 and 8). In the *T.boeoticum* assembly, number of contigs in scaffolds is 174299 whereas the number of contigs not in scaffolds is 179674. The average number of contigs per scaffold is 3.4 as it shown in tables (9,10 11, 12).

Table 5 : assembly Scaffold *T.monoccocum*

Size_includeN	**924383959**
Size_withoutN	**809771790**

Scaffold_Num	**373505**
Mean_Size	**2474**
Median_Size	**282**
Longest_Seq	**124868**
Shortest_Seq	**200**
Singleton_Num	**322652**
Average_length_of_break in_scaffold	**306**

Table 6: The coverage of *T.monococoum* genome

Number of N	**Minimum contig length**	**Number of Scaffolds**
N50	**9744**	**25409**
N80	**2360**	**81707**
N90	**1416**	**133822**

Table 7: **assembly Contig 'monoccocum.contig**

Size_includeN	**817428522**
Size_withoutN	**817342722**
Scaffold_Num	**494186**
Mean_Size	**1654**
Median_Size	**1307**
Longest_Seq	**39557**

Shortest_Seq	**200**
GC_Content	**45.99%**

Table 8 The coverage of *T.monococum* genome

Number of N	**Minimum contig length**	**Number of Scaffolds**
N50	**2920**	**79442**
N80	**1547**	**197432**
N90	**1273**	**255856**

Number_of_contigs_in_scaffolds	**171534**
Number_of_contigs_not_in_scaffolds(Singleton)	**322652**
Average_number_of_contigs_per_scaffold	**3.4**

Table 9: assembly Scaffold 'boeoticum genome

Size_includeN	**765161908**
Size_withoutN	**649326860**
Scaffold_Num	**231495**
Mean_Size	**3305**
Median_Size	**1485**
Longest_Seq	**124868**
Shortest_Seq	**200**

Singleton_Num	**179674**
Average_length_of_break in_scaffold	**500**

Table 10: **The coverage of *T.boeoticum* genome**

Number of N	**Minimum contig length**	**Number of Scaffolds**
N50	**8565**	**25451**
N80	**2303**	**74983**
N90	**1464**	**117827**

Table 11: assembly Contig *T.boeoticum*

Size_includeN	**657809209**
Size_withoutN	**657506009**
Scaffold_Num	**353973**
Mean_Size	**1858**
Median_Size	**1542**
Longest_Seq	**56404**
Shortest_Seq	**200**
GC_Content	**42.64%**

Table 12:The genome coverage of *T.boeoticum*

Number of N	**Minimum contig length**	**Number of Scaffolds**
N50	**2293**	**84649**
N80	**1460**	**194282**
N90	**1280**	**242477**

Number_of_contigs_in_scaffolds	**174299**
Number_of_contigs_not_in_scaffolds(Singleton)	**179674**
Average_number_of_contigs_per_scaffold	**3.4**

Mapping on the *T. monoccocum* genome:

We mapped the linkage map markers on assembled monoccocum genome . The number of matching reads is 6,557 that made up of 93.42 % . While the unmapped reads are very low 462 with 6.58%, the average length is 63. The total number of reads are 442.197 as it shown in Table(13).

Table 13: Mapping *T.monoccocum* genome

	Counts	Percentage of reads	Average length	Number of bases
References	373,505	-	2,474.89	924,383,959
Mapped reads	6,557	93.42%	63.00	413,091
Not mapped reads	462	6.58%	63.00	29,106
Total reads	7,019	100%	63.00	442,197

Mapping of *T.boeoticum* Genome :

The mapped linkage map markers on assembled of *T. boeoticum* genome is less significant because the assembly did not cover the genome as *T.monococcum* , so the number of mapped reads is 5,241 that made 74.67%. The unmatched reads are 462 with 6.58%. The total number of reads are 442,197 as it shown in table (14).

Table 14 Mapping of *T.boeoticum* Genome

	Counts	Percentage of reads	Average length	Number of bases
References	231,495	-	3,305.31	765,161,908
Mapped reads	5,241	74.67%	63.00	330,183
Not mapped reads	462	6.58%	63.00	112,014
Total reads	7,019	100%	63.00	442,197

The annotation results of *T.* boeoticum and *T. monoccocum*

After the assembly is done, we annotated the assembly to figure out how many genes we covered in our genome assembly of both genomes. Firstly, *T. boeoticum* annotation, the number of genes is 658 genes, while the number of transcripts is 1.122. on the other hand, the annotation of *T. monoccocum* found 31.116 genes, which a substantially higher than the number of genes in the *T.boeoticum* and the number of transcripts is 70. 194. The gene length ranged from few hundred to 12500 base pairs (bp), while the most common sizes are between 2500 to 5700 (bp). The number of transcripts it is ranged from (1-8) transcripts, more than 11.000 genes are having 2 transcripts , which is roughly 40% of the number of the genes. Moreover, the rest of genes have 3 or more transcripts basically. To validate our assembly we align the RNA

seq data of *T. monoccocum* to *T.boeoticum* assembly to test of we could get more genes, the results were quite significant with 25.779 genes and the number of transcripts is 75.165 which is close to the number of genes we found in the *T. monoccocum* annotation. The annotation results showed exact same outcome as the annotation of *T. monoccocum.*

Discussion:

Wheat has high agronomic importance traits, so tremendous efforts emphasized achieving whole-genome sequencing in wheat with the help of molecular breeding programs [12] [13]. *T. monococcum* has a genome size of 5.7 3 109 bp [14]. In recent years, an extensive library of *T. monococcum* was constructed to detect genomic regions in wheat [61].

The genome of *T. urartu was* sequenced and the final assembly result was 3.92 Gb with a contig N50 of 3.42 kilobases (kb). Also, about 66.88% of the *T. urartu* assembly were repetitive elements, such as long terminal repeat retrotransposons (49.07%), DNA transposons (9.77%) and unclassified elements (8.04%) [16]. A genome of common wheat has 28,000 genes. On the other hand, the assembly results of *T. urartu* indicates that there 6,800 more genes than hexaploid wheat. The most reasonable explanation is that gene numbers difference is due to the loss of genes in the hexaploid A genome during the domestication process [9].

An analysis of the DNA sequence of *Triticum monococcum L.* found five putative genes, two have similarity to disease resistance genes [61]. A new approach called chromosome walking, which uses bacterial artificial chromosome (BAC) clones, was successfully used in *Triticum monococcum L* to identify the *Lr10* leaf rust disease resistance gene in bread wheat [62]. The annotation results of *T. boeoticum* finding less significant than the *T. monoccocum* **Assembly assessment:**

We used two common ways of evaluating the quality of *T.monococcoum* and *T.boeoticum* genomes, the N50 which representing the average coverage of 50% assembled genome. The N50 of *T.monococcoum* is 9744 while the N50 of *T.boeoticum* is 8565. It is obvious that *T. monococoum* results is better than *boeoticum* because in *T.monococoum* we used 6 lanes of mate pairs (MP) such as 4 lines of 2000 kb and 2 lanes of 5000 kb than, while in *boeoticum* assembly we used 3 lanes of mate pair (MP) data for 2 kb, 4 kb, and 8 kb insert . As a result, the genome coverage of *T.monococoum* is more than *T. boeoticum.* Another way of evaluating the quality of genome assembly is using Benchmarking Universal Single-Copy Orthologs (BUSCO), which is a vital method of quantifications of completeness of genes that might be only found in single-copy. The results of (BUSCO) both genomes, shows that *T.monococoum* assembly covered 94.38% of 1440 conserved genes. While the number of missing genes is very low (5.62%). On other hand, the (BUSCO) of *T.boeoticum* shows (81.4%) genome coverage in terms of gene numbers. While the missing genes is higher than *T. monococoum* (19%), which due to the many gaps we have in our *T.boeoticum* assembly comparing to the *T. monoccocum* assembly. In terms of minimum number of contig length, such as N50 that represents the number of contigs that cover 50 % of the genome, both genomes have close N50 outcome. In details, N50 length of *T. monoccocum* is 2290, while the N50 length of *T.boeoticum* is 2920the percentage of GC in both genomes is close to 45%. Moreover, both of them have the same shortest contig number of 200. The best explanation is the library of both genomes are very close to each other. Also, they have same number of shortest contigs 200. The annotation results were quite significant between two genomes, because when we did the annotation for *T.boeoticum* we only

found 4 available RNA seq data, while we used 75 RNA seq data from different types of tissues, to annotate *T.monococoum* genome, we found 31.000 genes in *T.monococoum*, whereas the number of transcripts is 70.194. In addition, the functional annotation (BLAST) we found 49.538 genes with the lowest expectation number (-5), on the other hand, we found 658 genes and 1.112 transcripts in *T.boeoticum* genomes. The functional annotation of *T.boeoticum* genome we found 463 genes by the same lowest expectation number (-5). We did a mapping reads to both genomes to validate our assembly's results, so we aligned the assembly results of *T.boeoticum* and *T.monococoum* to genotyping by sequencing marker that has 7.109 markers of one population of *T.monococoum* with length of 63 base pairs (bp). Firstly, the mapping results of *T.monococoum* is substantial with 6557 reads, which make 93%, the not mapped reads are 462 reads with 6.58%, the number of total reads is 7.019 reads. In contrast, the mapped reads of *T.boeoticum* is 5.241 with 74.67% , the missing mapped reads is the same as *T.monococoum* 6.58% . The less significant results of *T.boeoticum* contributed to the less regions that covered in the assembly. In contrast, the mapping results of *T.monococoum* is sufficient due to more regions covered in the assembly.

Conclusion

This Einkorn draft genome sequence provides new insights into the A genome. This improved genome assembly combined with high quality sequencing data will extremely help scientists to identify genes and areas of the genome with interesting functions more accurately. Also, to identify more complete sets of similar genes that known as a gene family that are important for yield, disease resistance or other qualities

that are so important for agriculture. By using breeding programs to produce new variety that have increasing of the yield and resist the diseases. That is shared by many polyploid wheat species.

Leaf Rust of wild and domesticated of Einkorn Wheat

Introduction

Wheat (*Triticum aestivum* L.) is the most cultivated crop in the world, taking third place in total production. It is believed to have originated in the Middle East, particularly Syria and Turkey [65]. It is considered one of the main staple foods globally. There is dire need to increase the wheat production globally to meet its high demand as the world population grows [66]. According to FAO (2014), wheat production has increased to 659.7 million tons between 2012 and 2013 and 715.1 million tons between 2013 and 2014. Wheat is facing a lot of challenges in terms of production due to biotic and abiotic stresses. Biotic stresses cause loss of 31 to 42 percent of all crops annually [67]. Nearly, 14 percent of that damage is because of diseases, a specific type of biotic stress, which is estimated to cause loss of $220 billion USD per year [67].

Wheat, like other crops, can being attacked by different types of pathogens, such as parasitic fungi and bacteria. These pathogens cause huge reduction of yield. Rusts diseases are the most common disease of cereals, causing significant yield losses internationally [68]. The three rust diseases of wheat are stem (black) rust, leaf (brown) rust, and stripe (yellow) rust, which are caused by the pathogens *Puccinia graminis f. sp. tritici* (Pgt), *P. triticina* (Pt), and *P. striiformis f. sp. tritici* (Pst) [69]. Leaf rust that is caused by Puccinia triticina is a serious threat in countries that produce wheat. This diseases might cause high yield losses in susceptible cultivars [70]. Yield losses in wheat are due to decreased numbers of kernels per head as well as lower kernel weights [71]. The losses can rise up to 14 percent, according to reports from the University of Nebraska, located in the Great Plains.

One of the main symptoms of rust diseases is having a big uredinia without chlorosis or necrosis in the host tissues. On the other hand, resistant wheat varieties have small hypersensitive flecks in order to minimize the size of uredinia that is usually surrounded by either chlorotic or necrotic zones [71]. Leaf rust is easily identified by the uredinial stage. The uredinia size is 1.5 mm in diameter, erumpent, round to ovoid, with brown uredinia that are spotted on both the upper and the lower leaf surfaces of the host [72]. Leaf rust mainly occurs due to Puccinia triticina Eriks. This type of rust is a very common disease in wheat (Triticum aestivum L.). The fungus is heteroecious, which means it needs a telial/uredinial host (wheat) and an alternative (pycnial/aecial) host (Thalictrum speciosissimum or Isopyrum fumaroides) to finish its life cycle. Rust leaf causes yield losses in wheat due to low kernels per head. Therefore, Puccinia triticina is an important pathogen in wheat production globally[71].

There are many strategies to control the disease, such as using a certified, fungicide-treated seed. These seeds reduce loss using seed-transmitted and soilborne fungal diseases of wheat. They can protect against fall diseases and insects like aphids, which play a role as vectors. A second strategy is to control the grassy weeds before planting. A third is to select disease resistance varieties. Lastly, one can plant varieties that have different genetic makeups, so yield loss can be minimized.

Using genes to provide resistance, especially to rust diseases, is an old approach. However, this method cannot last forever because the pathogen might become virulent due to mutations. To overcome this issue, the genetic base for resistance among cultivars

should be broadened and made more diverse by using wild types cultivars through a plant breeding program [73].

Wild relatives are useful sources of disease resistance, which are available in the wild relatives of common wheat [74]. Scientists used wild relatives of Aegilops spp to introgress it into cultivated common wheat [75]. To date, 56 leaf rust resistance genes have been identified. About 51 of these genes have been mapped [76]. The use of molecular markers was a tool in identifying 28 of the leaf rust resistant genes. The most common markers are RFLP (restriction fragment length polymorphism), RAPD (random amplified polymorphism DNA), ISSR (inter simple sequence repeats), and AFLP (amplified fragments length polymorphism).

Literature review

1.1 Wheat

Wheat is a very important crop globally. It can be classified into spring and winter types based on the growing season. Winter wheat planting time is late in winter, which needs cold temperatures to grow. Spring wheat planting time in the spring [77]. Economically, world wheat trade was 149.5 million tons between 2014 and 2015, according to (FAO 2014). Due to the dramatic increase in world population, there is a need to increase yield annually by 2 percent [50]. To achieve this goal, wheat cultivars must be improved [78].

1.2 Wheat taxonomy

Wheat belongs to the genus *Triticum,* which is associated with the *Poaceae* family. Wheat is composed of three different genomes. The first is diploid (2n=2x=14), like the wild type einkorn wheat A genome. The second type is tetraploid (2n=4x=28) durum wheat AB genome. The third type is hexaploid (2n=6x=42), like bread wheat AABBDD genome. Bread wheat (*T. aestivum*) is an allohexaploid, which is composed of three similar genomes A, B and D, every genome has 7 chromosomes to form (AABBDD, 2n = 42) [79]. The bread wheat genome originated from spontaneous hybridization between cultivated emmer (*T. turgidum* AABB, 2n = 28) and *T. tauschii* (DD, 2n = 14) during the evolution of wheat which occurred hundreds of years ago [79].

1.3 The Rusts

1.3.1 Origin and distribution

Rusts are very damaging diseases to many cereals, which had a significant role during the domestication process for a lot of cereal crops [80]. Rusts were present on

grasses before the creation of cereals. There was a specific mutation to the rusts, making them able to attack many cereals [81]. The Pucciniales are believed to cause rust diseases in different types of cereals. Based on some studies, there are 7,000 rust species that attack many plants [82]. There are three main species of rust pathogens in wheat, such as *Puccinia graminis f. sp. tritici* (Pgt), *P. triticina* (Pt), and *P. striiformis f. sp. tritici* (Pst), which are damaging to many plant species [83].

The three rust pathogens have different preferred environments in which to develop. For example, Pgt grows well in warmer regions, whereas Pst exists in cooler, wet places. Interestingly, Pt could survive in intermediate temperatures [81]. The main method of disease distribution is through their mutations and the ability to defeat the existing resistant variants by switching from virulence to virulence. Moreover, they have the ability to move very long distances, which make rust diseases a big threat to many cereal crops [81, 84].

The life cycles of rust fungi are complicated, but they can be easily diagnosed in the uredinial and telial stages using the naked eye. For instance, in stem rust, the telial stage is clear whereas brown rust has a brown uredinial stage. In contrast, stripe rust has a yellow uredinial stage. (Figure 14).

Fig 13 Uredinial stages of leaf rust, stem rust, and stripe rust

1.3.2 Taxonomy and nomenclature

Rust fungi is considered Basidiomycetes, which belongs to the order Pucciniales. To complete the cycle of the pathogen, a third host is needed [85]. The Pucciniales family has about 100 genera as well as 7,000 species [86].

1.3.3 Life cycles and host range

All kinds of pathogens share the same spore stages, such as the five spore types: basidiospores, pycniospores, aeciospores, urediniospores, and teliospores. In addition, some of the pathogens need just one host, some need two, and others need three hosts to complete the life cycle [87]. For example, P. triticina is considered a macrocyclic and heteroecious rust fungus with five spore stages [72]. The final part of life cycle, in which a uredinial stage disappears, is usually called a demicyclic. Microcyclic takes place after

the uredinial and aecial stages. Interestingly, microcyclic forms are aoutoecious because of their short life cycle [88].

1.4 Variation in the rust pathogens

All kind of pathogens are usually produced through sexual process. As a result, there seem to be similarities among them, because the degree of variability among those individuals are reduced [89].

1.4.1. Mechanisms of variability

Mutation: considered the main source of producing new alleles in plant pathogens (CIMMYT 1988). For example, if a mutation has happened on avirulence function of the plant, then the plant defense system would be unable to identify the pathogen elicitor, which makes the pathogen invisible and ready to attack the plant [90]. Therefore, mutation could cause a wide variation to take place in rust pathogens [91].

Selection: a natural evolutionary mechanism to have phenotypic variability of rust pathogen populations throughout a specific kind of environment [92]. Generally speaking, the genetic makeup of survived pathogens would most likely prevail over thoer genetics makeup due to selection (CIMMYT 1988).

Sexual recombination: alternate hosts that are located close to wheat fields provide a high chance of evolution of new rust pathotypes [91]. For example, a specific kind of P. graminis evolved to a rust which infected barberry. These alternate hosts played a siginficant role in producing many pathogen pathotypes [93].

Somatic hybridisation: provides exchange of nuclei between fusing hyphae in the dikaryotic stage [94]. It occurs depending on the type of environment. It may also happen as an alternative mechanism of sexual recombination like in Australia [95].

Migration: helps organisms relocate to new regions. The spores of many fungal pathogens of rusts have the ability to migrate to new places by wind [96]. The rust fungi are spread as a clonally produced dikaryotic urediniospores make them spread thousands of kilometers from the original land. As a result, it may cause a huge infection that destroys millions of plants [97]. One good example of how the pathogen can spread long distances is distribution of stem rust "Ug99" from Uganda to different West African countries and Yemen [98].

2.1 Host-pathogen interactions

As the infection happens, plants usually defend themselves against pathogens using the waxy substances around the locus of infection to produce anti-microbial compounds [99]. This relationship is controlled by the gene-for-gene[100]. To explain this relationship, resistance to specific pathogen occurs when the plant-pathogen interaction is incompatible, if resistance allele in the plant and the corresponding functional avirulence allele in the pathogen occur simultaneously [101].

2.2 Plant disease resistance genes

1- Race-specific resistance: Widely known as as major gene or gene-for-gene resistance, this type of resisitance does not last for very long because of the mutation of pathogens strains. This is also because it works just for some pathotypes and can be broken down quickly [102].

2-Race-non-specific resistance: This type of genetic resistance is based on additive interaction of a few or several genes. These genes have minor to intermediate effects. This is known as non-differential interaction [103]. It is difficult to identify which type of pathogen is attacking the plants under this type of resistance [104].

3-Slow rusting: This type of resistance occurs when the pathogen grows inside of the plants slowly, which leads to low disease levels against all pathotypes of a pathogen [105].

4-Partial resistance: This is known as incomplete resistance, which causes reduction of disease spread. It is more durable than hypersensitive resistance [104].

5- Durable resistance: Durable resistance can still be effective in a plant during its widespread cultivation for a long sequence of generations under preferable environment condition to a specific disease [103]. Durable resistance is formed by combinations of many genes with minor effects in an additive manner [106].

2.3 Analysis of resistance genes

2.3.1 Gene postulation

When discovering new resistance genes, it is very important to include those genes in the breeeding program. In identifying the unknown resistance, two common approaches can be used, namely multipathotype testing and molecular marker analysis [107]. These methods are based on different test cultivars with unknown genes for resistance with control culitvars that have identified resistance genes [108].

2.3.2 Genetic analysis

Genetic analysis is usually used to evaluate resistance gene postulations. This method would provide exact numbers of resistance genes and their identity in wheat cultivars [71]. The disadvantage of this method is that it is time consuming [109].

3.1 Molecular markers for resistance genes

Molecular markers are important in identifing the differences in genetic makeup. They also help in evaluating the genetic diversity and the mapping of quantitative traits loci (QTL). QTL is vital in identifiying complex traits, including quantitative disease resistance [110]. In simple words, QTL is a genomic location that governs different quantitative traits of interest [111]. There are many molecular markers used to identify those resistance genses, such as PCR, microsatellites (SSR), RFLP, AFLP, and RAPD. The most common markers are RFLPs and RAPDs [112].

3.2 Wheat stem rust

This disease is caused by the fungus *Pgt,* which leads to huge damage to yield in wheat fields globally. The fungus grows in the stems, which blocks nutrient from reaching out the developing heads and leads to shriveled grain [113]. This disease affects many regions around the world, such as Africa, the Middle East, Europe, Australia, and North and South America [114]. This disease causes tremendous yield loss. For example, from 1986 to 1999, a *Pst* pathotype virulent on *Yr9* caused a big yield loss in East Africa, near southeast Asia [115].

3.3 Life cycle and host range

Pgt is considered a heteroecious between telial and aecial hosts and macrocyclic with five spore stages that differ in morphology and function [116]. Teliospore cell usually called a dikaryotic, the next step is forming the teliospore [117]. This type of disease has the ability to travel more than 8,000 Km, which is roughly equivalent to the distance between southern Africa and Australia [93].

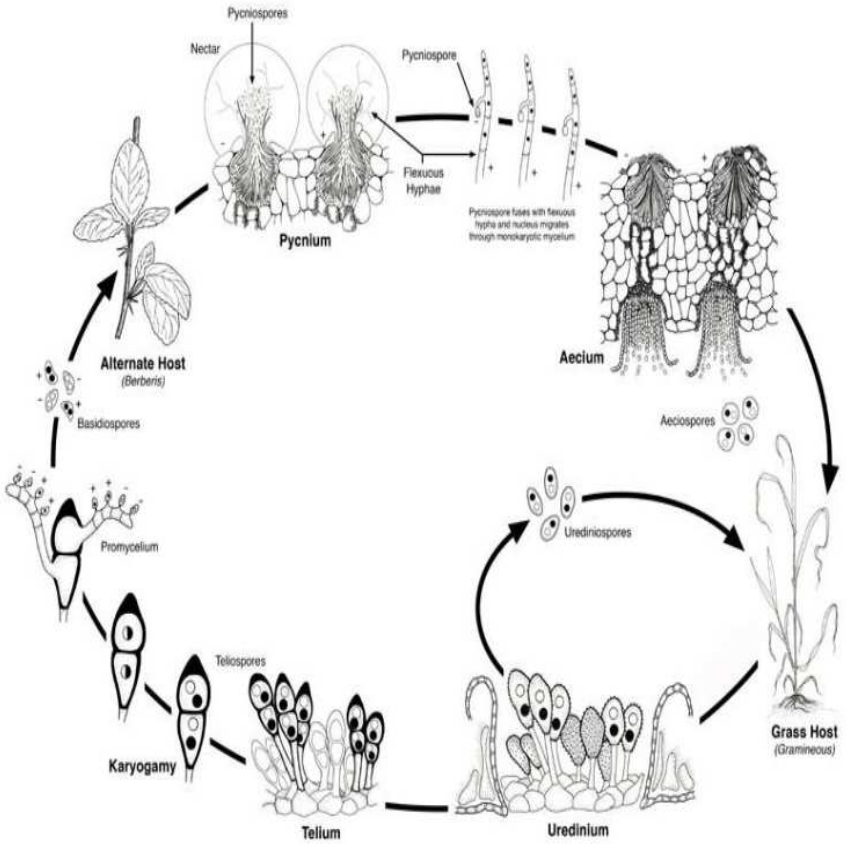


Fig 14: shows the life cycle of *Puccina graminis f. sp. tritici*, showing both primary and alternate hosts [117]

3.3.1 Management

One of the primary management methods is to remove common barberry from wheat fields, which leads to a reduction of the initial source of inoculum in many regions, such as North America [71]. Fungicides have been used as another method for many years to reduce the attack of *Pgt*. However, the affect on the disease is very limited due to external factors, including hosts of the disease and environmental condition [91]. The third method is to use resistance genes. These genes rely on the relationship between a single R gene product in the host, which has the ability to recognize a single pathogen. The more durable resistance genes are race-non-specific because they are governed by more than one gene. There are two common durable resistance genes: *Sr26* and Sr 31 [118].

3.4 Wheat stripe rust

3.4.1 Nature of the pathogen

Stripe rust, which is caused by *Pst,* is responsible for the loss of thousands of wheat and other crop yields annually in many parts of the world [70]. For example, in China in 1985, *Pst* caused significant yield loss of 2.65 million tons. Transcaucasia is the center of origin for *Pst,* because it provided the perfect environment to grow due to the abundance of the grasses in that region. Then it spread to other parts of the world [119].

3.4.2 Life cycle and host range

The strip rust disease usually attacks the green tissues of wheat at early growth stages until its maturity phase [70]. In affected wheat plants, the pathogen starts to grow from yellow to orange in pustules, leaves and leaf sheaths, and also infects glumes [120].

Uredinia stripes are formed by the elongation of stem. As a result, chlorosis and necrosis appear in plants [70].

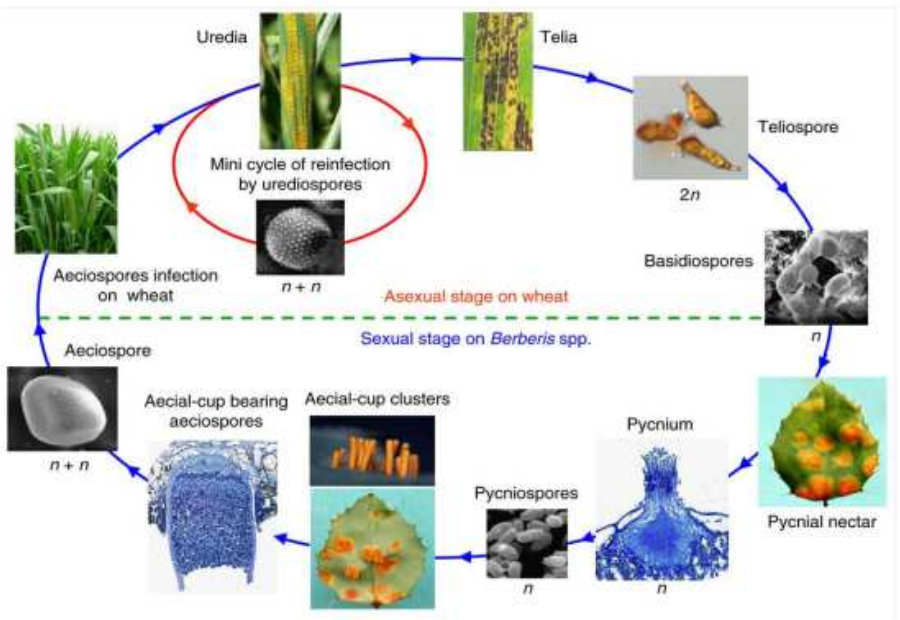


Fig 15 shows the life cycle of stripe rust disease [121]

3.4.3 Management methods

It has been reported that using an alternate host would reduce the attack of *Pst* because sexual reproduction increases the production of variation in the *Pst* population [122]. Fungicides can control stripe rust in Western Europe. The disadvantage to using fungicides is the high expense for farmers. Resistance genes of *Pst* are believed to be race specific, such as Yr 6 and Yr 9 [83].

3.4.4Wheat leaf rust

The cause of leaf rust is *Puccinia triticina* (Pt*).* The pathogen is considered heteroecious, meaning that it needs a telial/uredinial host and an alternate host to regenerate [72]. The severity of leaf rust depends on the growth of the wheat plant. However, the disease mainly infects the leaves and sheaths of wheat, which leads to enormous yield loss of up to 70 percent. The total loss of yield was 3 million tons, which cost about $350 million from 2000 to 2004 [123]. *Pt* typically uses susceptible wheat plants as well as alternate hosts for a life cycle, as is shown in Figure 2.3. This type of disease mainly depends on the bread wheat as a host [108].

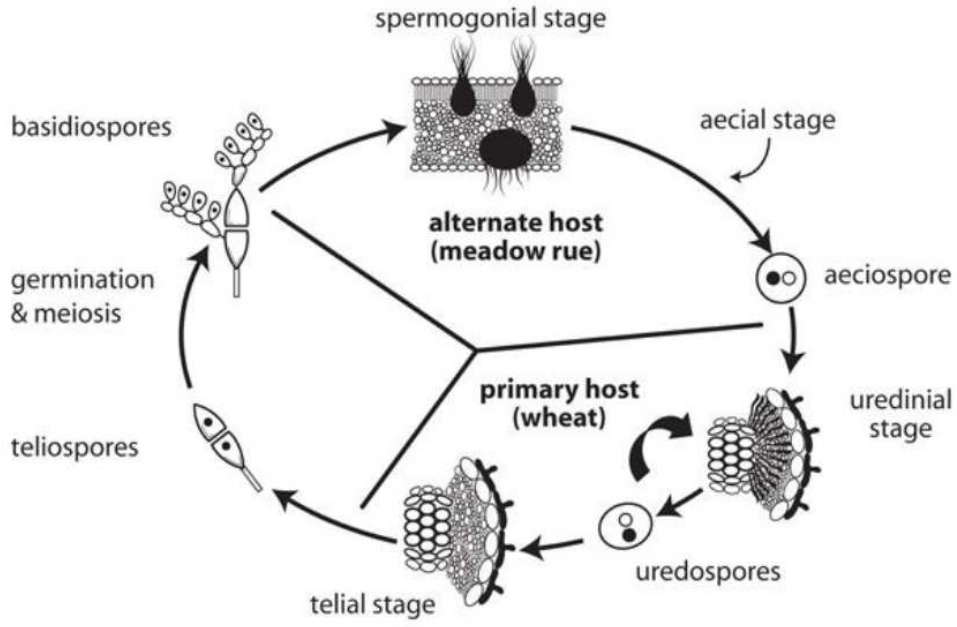


Fig 16 Life cycle of leaf rust, showing primary and alternate hosts [116]**.**

4.1 Management

The use of alternate hosts *Clematis* and *Isopyrum* can enormously decrease the disease inoculum [108]. Fungicides can be used as a backup when new *Pt* pathotypes occur and

resistance genes can be used [81]. Resistance genes of leaf rust have more than 70 designated loci. These genes are located in bread wheat and durum wheat [124]. Most of the current resistance genes come from common wheat. In addition, there are some resistance genes derived from wild types, such as *Lr9* from *Aegilops* and *Lr28*, and *Lr47* from *Aegilops speltoid* wild type [125].

References

1. Lev-Yadun, S., A. Gopher, and S. Abbo, *The cradle of agriculture.* Science, 2000. **288**(5471): p. 1602-1603.
2. Hegerl, G.C., et al., *Understanding and attributing climate change.* 2007.
3. Heun, M., et al., *Site of einkorn wheat domestication identified by DNA fingerprinting.* Science, 1997. **278**(5341): p. 1312-1314.
4. Dvořák, J., et al., *The evolution of polyploid wheats: identification of the A genome donor species.* Genome, 1993. **36**(1): p. 21-31.
5. Sarkar, P. and G. Stebbins, *Morphological evidence concerning the origin of the B genome in wheat.* American Journal of Botany, 1956. **43**(4): p. 297-304.
6. Feldman, M. and M.E. Kislev, *Domestication of emmer wheat and evolution of free-threshing tetraploid wheat.* Israel Journal of Plant Sciences, 2007. **55**(3-4): p. 207-221.
7. Bálint, A., G. Kovács, and J. Sutka, *Origin and taxonomy of wheat in the light of recent research.* Acta Agronomica Hungarica, 2000. **48**(3): p. 301-313.
8. Martinez-Perez, E., P. Shaw, and G. Moore, *The Ph1 locus is needed to ensure specific somatic and meiotic centromere association.* Nature, 2001. **411**(6834): p. 204.
9. Ling, H.-Q., et al., *Draft genome of the wheat A-genome progenitor Triticum urartu.* Nature, 2013. **496**(7443): p. 87.
10. Jacob, D., et al., *EURO-CORDEX: new high-resolution climate change projections for European impact research.* Regional Environmental Change, 2014. **14**(2): p. 563-578.

11. Zohary, D., M. Hopf, and E. Weiss, *Domestication of Plants in the Old World: The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin*. 2012: Oxford University Press on Demand.

12. Harlan, J.R. and D. Zohary, *Distribution of wild wheats and barley.* Science, 1966. **153**(3740): p. 1074-1080.

13. Zettler, J.-T., *Characterization of epitaxial semiconductor growth by reflectance anisotropy spectroscopy and ellipsometry.* Progress in Crystal Growth and Characterization of Materials, 1997. **35**(1): p. 27-98.

14. Horton, T.R., *The number of nuclei in basidiospores of 63 species of ectomycorrhizal Homobasidiomycetes.* Mycologia, 2006. **98**(2): p. 233-238.

15. Bennetzen, J.L., J. Ma, and K.M. Devos, *Mechanisms of recent genome size variation in flowering plants.* Annals of botany, 2005. **95**(1): p. 127-132.

16. Wicker, T., et al., *Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives.* The Plant Cell, 2011: p. tpc. 111.086629.

17. Wang, X., et al., *The genome of the mesopolyploid crop species Brassica rapa.* Nature genetics, 2011. **43**(10): p. 1035.

18. Biscotti, M.A., E. Olmo, and J.P. Heslop-Harrison, *Repetitive DNA in eukaryotic genomes*. 2015, Springer.

19. Hammond, C.M., et al., *Histone chaperone networks shaping chromatin function.* Nature Reviews Molecular Cell Biology, 2017. **18**(3): p. 141.

20. Whetten, D.A., *Albert and Whetten revisited: Strengthening the concept of organizational identity.* Journal of management inquiry, 2006. **15**(3): p. 219-234.

21. López-Flores, I. and M. Garrido-Ramos, *The repetitive DNA content of eukaryotic genomes*, in *Repetitive DNA*. 2012, Karger Publishers. p. 1-28.

22. Pearce, S.R., et al., *TheTy1-copia group retrotransposons ofAllium cepa are distributed throughout the chromosomes but are enriched in the terminal heterochromatin.* Chromosome Research, 1996. **4**(5): p. 357-364.

23. Schwarzacher, T. and J. Heslop-Harrison, *In situ hybridization to plant telomeres using synthetic oligomers.* Genome, 1991. **34**(3): p. 317-323.

24. Mehrotra, S. and V. Goyal, *Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function.* Genomics, proteomics & bioinformatics, 2014. **12**(4): p. 164-171.

25. Ramel, C., *Mini-and microsatellites.* Environmental Health Perspectives, 1997. **105**(suppl 4): p. 781-789.

26. Britten, R.J. and D.E. Kohne, *Repeated sequences in DNA.* Science, 1968. **161**(3841): p. 529-540.

27. Lodish, H., et al., *Molecular cell biology*. 2008: Macmillan.

28. Initiative, A.G., *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* nature, 2000. **408**(6814): p. 796.

29. Claros, M.G., et al., *Why assembling plant genome sequences is so challenging.* Biology, 2012. **1**(2): p. 439-459.

30. Hernandez, P., et al., *Next□generation sequencing and syntenic integration of flow□sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content.* The Plant Journal, 2012. **69**(3): p. 377-386.

31. Hamilton, J.P. and C. Robin Buell, *Advances in plant genome sequencing.* The Plant Journal, 2012. **70**(1): p. 177-190.

32. Schatz, M.C., J. Witkowski, and W.R. McCombie, *Current challenges in de novo plant genome sequencing and assembly.* Genome biology, 2012. **13**(4): p. 243.

33. Sargent, D.J., et al., *Simple sequence repeat marker development and mapping targeted to previously unmapped regions of the strawberry genome sequence.* The Plant Genome, 2011. **4**(3): p. 165-177.

34. Brenchley, R., et al., *Analysis of the bread wheat genome using whole-genome shotgun sequencing.* Nature, 2012. **491**(7426): p. 705.

35. Sanger, F., et al., *Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing.* Journal of molecular biology, 1980. **143**(2): p. 161-178.

36. Schnable, P.S., et al., *The B73 maize genome: complexity, diversity, and dynamics.* science, 2009. **326**(5956): p. 1112-1115.

37. Kchouk, M., J.-F. Gibrat, and M. Elloumi, *Generations of sequencing technologies: From first to next generation.* Biology and Medicine, 2017. **9**(3).

38. Medini, D., et al., *Microbiology in the post-genomic era.* Nature Reviews Microbiology, 2008. **6**(6): p. 419.

39. Fichot, E.B. and R.S. Norman, *Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform.* Microbiome, 2013. **1**(1): p. 10.

40. Jain, M., et al., *Improved data analysis for the MinION nanopore sequencer.* Nature methods, 2015. **12**(4): p. 351.

41. Meyers, L.A. and D.A. Levin, *On the abundance of polyploids in flowering plants.* Evolution, 2006. **60**(6): p. 1198-1206.

42. Feuillet, C., et al., *Crop genome sequencing: lessons and rationales.* Trends in plant science, 2011. **16**(2): p. 77-88.

43. Kajitani, R., et al., *Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads.* Genome research, 2014: p. gr. 170720.113.

44. Earl, D.A., et al., *Assemblathon 1: a competitive assessment of de novo short read assembly methods.* Genome research, 2011: p. gr. 126599.111.

45. Bradnam, K.R., et al., *Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species.* GigaScience, 2013. **2**(1): p. 10.

46. Mayer, K.F., et al., *Gene content and virtual gene order of barley chromosome 1H.* Plant Physiology, 2009. **151**(2): p. 496-505.

47. Mayer, K.F., et al., *Unlocking the barley genome by chromosomal and comparative genomics.* The Plant Cell, 2011: p. tpc. 110.082537.

48. Appels, R., et al., *Shifting the limits in wheat research and breeding using a fully annotated reference genome.* Science, 2018. **361**(6403): p. eaar7191.

49. Shi, X. and H.-Q. Ling, *Current advances in genome sequencing of common wheat and its ancestral species.* The Crop Journal, 2018. **6**(1): p. 15-21.

50. Gill, B.S., et al., *A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium.* Genetics, 2004. **168**(2): p. 1087-1096.

51. Yan, L., et al., *The wheat VRN2 gene is a flowering repressor down-regulated by vernalization.* Science, 2004. **303**(5664): p. 1640-1644.

52. Murai, K., et al., *A large-scale mutant panel in wheat developed using heavy-ion beam mutagenesis and its application to genetic research.* Nuclear Instruments

and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms, 2013. **314**: p. 59-62.

53. Jing, H.-C., et al., *Identification of variation in adaptively important traits and genome-wide analysis of trait–marker associations in Triticum monococcum.* Journal of experimental botany, 2007. **58**(13): p. 3749-3764.

54. Rogers, W., et al., *Introduction to bread wheat (Triticum aestivum L.) and assessment for bread-making quality of alleles from T. boeoticum Boiss. ssp. thaoudar at Glu-A1 encoding two high-molecular-weight subunits of glutenin.* Euphytica, 1997. **93**(1): p. 19-29.

55. Shi, A., S. Leath, and J. Murphy, *A major gene for powdery mildew resistance transferred to common wheat from wild einkorn wheat.* Phytopathology, 1998. **88**(2): p. 144-147.

56. Anker, C.C. and R.E. Niks, *Prehaustorial resistance to the wheat leaf rust fungus, Puccinia triticina, in Triticum monococcum (ss).* Euphytica, 2001. **117**(3): p. 209-215.

57. Hovhannisyan, N.A., et al., *Tracking of powdery mildew and leaf rust resistance genes in Triticum boeoticum and T. urartu, wild relatives of common wheat.* Czech Journal of Genetics and Plant Breeding, 2011. **47**(2): p. 45-57.

58. Budak, H., et al., *Proteome changes in wild and modern wheat leaves upon drought stress by two-dimensional electrophoresis and nanoLC-ESI–MS/MS.* Plant molecular biology, 2013. **83**(1-2): p. 89-103.

59. Ergen, N.Z. and H. Budak, *Sequencing over 13 000 expressed sequence tags from six subtractive cDNA libraries of wild and modern wheats following slow drought stress.* Plant, cell & environment, 2009. **32**(3): p. 220-236.

60. Dubcovsky, J., M. Luo, and J. DvORAK, *Differentiation between homoeologous chromosomes 1A of wheat and 1Am of Triticum monococcum and its recognition by the wheat Ph1 locus.* Proceedings of the National Academy of Sciences, 1995. **92**(14): p. 6645-6649.

61. Wicker, T., et al., *Analysis of a contiguous 211 kb sequence in diploid wheat (Triticum monococcum L.) reveals multiple mechanisms of genome evolution.* The Plant Journal, 2001. **26**(3): p. 307-316.

62. Stein, N., et al., *Subgenome chromosome walking in wheat: a 450-kb physical contig in Triticum monococcum L. spans the Lr10 resistance locus in hexaploid wheat (Triticum aestivum L.).* Proceedings of the National Academy of Sciences, 2000. **97**(24): p. 13436-13441.

63. Michikawa, A., et al., *Genome-wide polymorphisms from RNA sequencing assembly of leaf transcripts facilitate phylogenetic analysis and molecular marker development in wild einkorn wheat.* Molecular Genetics and Genomics, 2019: p. 1-15.

64. Avni, R., et al., *Wild emmer genome architecture and diversity elucidate wheat evolution and domestication.* Science, 2017. **357**(6346): p. 93-97.

65. Gibson, L. and G. Benson, *Origin, history, and uses of oat (Avena sativa) and wheat (Triticum aestivum).* Iowa State University, Department. of Agronomy: United States, 2002.

66. Edmeades, G., R. Fischer, and D. Byerlee. *Can we feed the world in 2050*. in *Proceedings of the New Zealand grassland association*. 2010.

67. Agrios, G.N., *Plant pathology*. 2005: Academic press.

68. Vurro, M., B. Bonciani, and G. Vannacci, *Emerging infectious diseases of crop plants in developing countries: impact on agriculture and socio-economic consequences.* Food Security, 2010. **2**(2): p. 113-132.

69. Bahri, B., et al., *Genetic diversity of the wheat yellow rust population in Pakistan and its relationship with host resistance.* Plant Pathology, 2011. **60**(4): p. 649-660.

70. Chen, W., T. Liu, and L. Gao, *Suppression of stripe rust and leaf rust resistances in interspecific crosses of wheat.* Euphytica, 2013. **192**(3): p. 339-346.

71. Kolmer, J.A., *Tracking wheat rust on a continental scale.* Current opinion in plant biology, 2005. **8**(4): p. 441-449.

72. Bolton, M.D., J.A. Kolmer, and D.F. Garvin, *Wheat leaf rust caused by Puccinia triticina.* Molecular plant pathology, 2008. **9**(5): p. 563-575.

73. Valkoun, J., *Wheat pre-breeding using wild progenitors.* Euphytica, 2001. **119**(1-2): p. 17-23.

74. Zaharieva, M., et al., *Evaluation of a collection of wild wheat relative Aegilops geniculata Roth and identification of potential sources for useful traits*, in *Wheat in a Global Environment*. 2001, Springer. p. 739-746.

75. Assefa, S. and H. Fehrmann, *Evaluation of Aegilops tauschii Coss. for resistance to wheat stem rust and inheritance of resistance genes in hexaploid wheat.* Genetic Resources and Crop Evolution, 2004. **51**(6): p. 663-669.

76. Liu, Z., et al., *Molecular characterization of a novel powdery mildew resistance gene Pm30 in wheat originating from wild emmer.* Euphytica, 2002. **123**(1): p. 21-29.

77. Curtis, B.C., S. Rajaram, and M. Gómez, *Bread wheat: improvement and production*. 2002: Food and Agriculture Organization of the United Nations (FAO).

78. Tilman, D., et al., *Global food demand and the sustainable intensification of agriculture.* Proceedings of the National Academy of Sciences, 2011. **108**(50): p. 20260-20264.

79. Gupta, P., et al., *Genetic mapping of 66 new microsatellite (SSR) loci in bread wheat.* Theoretical and applied genetics, 2002. **105**(2-3): p. 413-422.

80. Schafer, J., A. Roelfs, and W. Bushnell, *Contributions of early scientists to knowledge of cereal rusts.* The cereal rusts, 1984. **2**: p. 3-38.

81. Johnson, T., G. Green, and D. Samborski, *The world situation of the cereal rusts.* Annual Review of Phytopathology, 1967. **5**(1): p. 183-200.

82. McKenzie, E., *Rust fungi of New Zealand—an introduction, and list of recorded species.* New Zealand Journal of Botany, 1998. **36**(2): p. 233-271.

83. Boyd, L.A., *Can Robigus defeat an old enemy?–Yellow rust of wheat.* The Journal of Agricultural Science, 2005. **143**(4): p. 233-243.

84. Nsabiyera, V., et al., *Molecular markers for adult plant leaf rust resistance gene Lr48 in wheat.* Molecular breeding, 2016. **36**(6): p. 65.

85. Voegele, R.T., M. Hahn, and K. Mendgen, *The uredinales: cytology, biochemistry, and molecular biology*, in *Plant relationships*. 2009, Springer. p. 69-98.

86. Hawksworth, D., et al., *Ainsworth & Bisby's Dictionary of the fungi. CAB International.* International Mycological Institute, Kew, UK, 1995.

87. Hiratsuka, Y. and G.B. Cummins, *Illustrated genera of rust fungi*. 1983: American Phytopathological Society.

88. Petersen, R.H., *The rust fungus life cycle.* The Botanical Review, 1974. **40**(4): p. 453-513.

89. Kiyosawa, S., *Genetics and epidemiological modeling of breakdown of plant disease resistance.* Annual Review of phytopathology, 1982. **20**(1): p. 93-117.

90. McDonald, D., et al., *Cytogenetical studies in wheat XIX. Location and linkage studies on gene Yr27 for resistance to stripe (yellow) rust.* Euphytica, 2004. **136**(3): p. 239-248.

91. Roelfs, A., *Wheat and rye stem rust*, in *Diseases, Distribution, Epidemiology, and Control*. 1985, Elsevier. p. 3-37.

92. Brown, G., *The inheritance and expression of leaf chlorosis associated with gene Sr2 for adult plant resistance to wheat stem rust.* Euphytica, 1997. **95**(1): p. 67-71.

93. Watson, I. and N. Luig, *SR 15—a new gene for use in the classification of Puccinia graminis var. tritici.* Euphytica, 1966. **15**(2): p. 239-247.

94. Cheng, P., *Development and Use of CDNA-derived SSR Markers for Studying Puccinia Striiformis Populations and Molecular Mapping of New Genes for Effective Resistance to Stripe Rust in Durum Wheat.* 2012.

95. Luig, N. and S. Rajaram, *The Effect of Temperature and Genetic Background on Host Gene Expression and Interaction to.* Puccinia graminis tritici. Phytopathol., 1972. **62**: p. 1171-1174.

96. Nagarajan, S. and D. Singh, *Long-distance dispersion of rust pathogens.* Annual review of phytopathology, 1990. **28**(1): p. 139-153.

97. Roelfs, A., *Epidemiology of the cereal rusts in North America.* Canadian Journal of Plant Pathology, 1989. **11**(1): p. 86-90.

98. Singh, R., et al., *Spread of a highly virulent race of Puccinia graminis tritici in Eastern Africa*, in *Wheat Production in Stressed Environments*. 2007, Springer. p. 51-57.

99. Dangl, J.L. and J.D. Jones, *Plant pathogens and integrated defence responses to infection.* nature, 2001. **411**(6839): p. 826-833.

100. Kolmer, J., *Genetics of resistance to wheat leaf rust.* Annual review of phytopathology, 1996. **34**(1): p. 435-455.

101. Ansan□Melayah, D., et al., *Genes for race□specific resistance against blackleg disease in Brassica napus L.* Plant Breeding, 1998. **117**(4): p. 373-378.

102. Qayoum, A. and R. Line, *High-temperature, adult-plant resistance to stripe rust of wheat.* Phytopathology, 1985. **75**(10): p. 1121-1125.

103. Johnson, R., *A critical analysis of durable resistance.* Annual review of phytopathology, 1984. **22**(1): p. 309-330.

104. Parlevliet, J., *Resistance of the non-race-specific type*, in *Diseases, Distribution, Epidemiology, and Control*. 1985, Elsevier. p. 501-525.

105. Caldwell, R. *1 968. Breeding for general and/or specific plant disease resistance*. in *Proc. 3rd Int. Wheat Genet. Symp., Eds. Finlay, KW and Shepherd, DW, Aust. A cad. Sci., Canbera*. 1968.

106. McIntosh, R., *Pre-emptive breeding to control wheat rusts.* Euphytica, 1992. **63**(1-2): p. 103-113.

107. Kolomiets, T. *Postulated resistance genes in cultivars and lines with alien genes to wheat leaf rust*. in *Proceeding of 11th International Cereal Rusts and Powdery Mildews Conference, John Innes Centre, Norwich, England: 22nd to 27th August*. 2004.

108. Roelfs, A.P., *Rust diseases of wheat: concepts and methods of disease management*. 1992: Cimmyt.

109. Simmonds, N.W. and S. Rajaram, *Breeding strategies for resistance to the rusts of wheat*. 1988: CIMMYT.

110. Young, N.D., *QTL mapping and quantitative disease resistance in plants.* Annual review of phytopathology, 1996. **34**(1): p. 479-501.

111. Doerge, R.W., *Mapping and analysis of quantitative trait loci in experimental populations.* Nature Reviews Genetics, 2002. **3**(1): p. 43-52.

112. William, M., et al., *Molecular marker mapping of leaf rust resistance gene Lr46 and its association with stripe rust resistance gene Yr29 in wheat.* Phytopathology, 2003. **93**(2): p. 153-159.

113. Jin, Y., L.J. Szabo, and M. Carson, *Century-old mystery of Puccinia striiformis life history solved with the identification of Berberis as an alternate host.* Phytopathology, 2010. **100**(5): p. 432-435.

114. Saari, E.E. and J. Prescott, *World distribution in relation to economic losses*, in *Diseases, Distribution, Epidemiology, and Control*. 1985, Elsevier. p. 259-298.

115. Singh, R.P., et al. *Wheat rust in Asia: meeting the challenges with old and new technologies*. in *Proceedings of the 4th International Crop Science Congress*. 2004. The Regional Institute Ltd Gosford, Australia.

116. Alexopoulos, C.J., C.W. Mims, and M. Blackwell, *Introductory mycology*. 1996: John Wiley and Sons.

117. Leonard, K.J. and L.J. Szabo, *Stem rust of small grains and grasses caused by Puccinia graminis.* Molecular plant pathology, 2005. **6**(2): p. 99-111.

118. McIntosh, R., *Close genetic linkage of genes conferring adult□plant resistance to leaf rust and stripe rust in wheat.* Plant Pathology, 1992. **41**(5): p. 523-527.

119. Stubbs, R., *Stripe rust. In 'Cereal rusts. Vol. II. Disease, distribution, epidemiology, and control'.(Eds AP Roelfs, WR Bushnell) pp. 61–101*. 1985, Academic Press: New York.

120. Line, R.F., *Stripe rust of wheat and barley in North America: a retrospective historical review.* Annual review of phytopathology, 2002. **40**(1): p. 75-118.

121. Zheng, W., et al., *High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus.* Nature communications, 2013. **4**(1): p. 1-11.

122. Zhao, J., et al., *Identification of eighteen Berberis species as alternate hosts of Puccinia striiformis f. sp. tritici and virulence variation in the pathogen isolates from natural infection of barberry plants in China.* Phytopathology, 2013. **103**(9): p. 927-934.

123. Huerta-Espino, J., et al., *Global status of wheat leaf rust caused by Puccinia triticina.* Euphytica, 2011. **179**(1): p. 143-160.

124. McIntosh, R. and Z. Pretorius, *Borlaug Global Rust Initiative provides momentum for wheat rust research*. 2011, Springer.

125. Todorovska, E., et al., *Biotic stress resistance in wheat—breeding and genomic selection implications.* Biotechnology & Biotechnological Equipment, 2009. **23**(4): p. 1417-1426.

APPENDIX

Contents

W2RAP Pipeline

---

1 QC PE read files

1.1 Run FASTQC to check read metrics

Listing 1: Executed code

```
1 module load FastQC/0.11.7
2 mkdir fastqc
3 fastqc -o fastqc TA4342-L95_R1_001.fastq.gz TA4342-L95_R2_001.fastq.gz
&> ←-
1a.log &
```

Output contained in a directory called fastqc

Listing 2: Command output

```
1 -rw-r--r--. 1 software 240K May 18 09:49 TA4342-L95_R1_001_fastqc.html
2 -rw-r--r--. 1 software 325K May 18 09:49 TA4342-L95_R1_001_fastqc.zip
3 -rw-r--r--. 1 software 240K May 18 11:06 TA4342-L95_R2_001_fastqc.html
4 -rw-r--r--. 1 software 324K May 18 11:06 TA4342-L95_R2_001_fastqc.zip
```

Check the output on Appendix 9.1.1 and 9.1.2

1.2 Use KAT hist to estimate kmer coverage

Listing 3: Executed code

```
1 module load KAT/2.4.1
2 kat hist -o scer_pe_hist -h 80 -t 8 -m 27 -H 100000000 TA4342-
  L95_R?_001.←fastq.gz &> 1b.log &
```

Listing 4: Command output

```
1 -rw-r--r--. 1 software 1.2K May 18 12:45 scer_pe_hist
2 -rw-r--r--. 1 software 1.2K May 18 12:45 scer_pe_hist.dist_analysis.json
3 -rw-r--r--. 1 software 158K May 18 12:45 scer_pe_hist.png
```

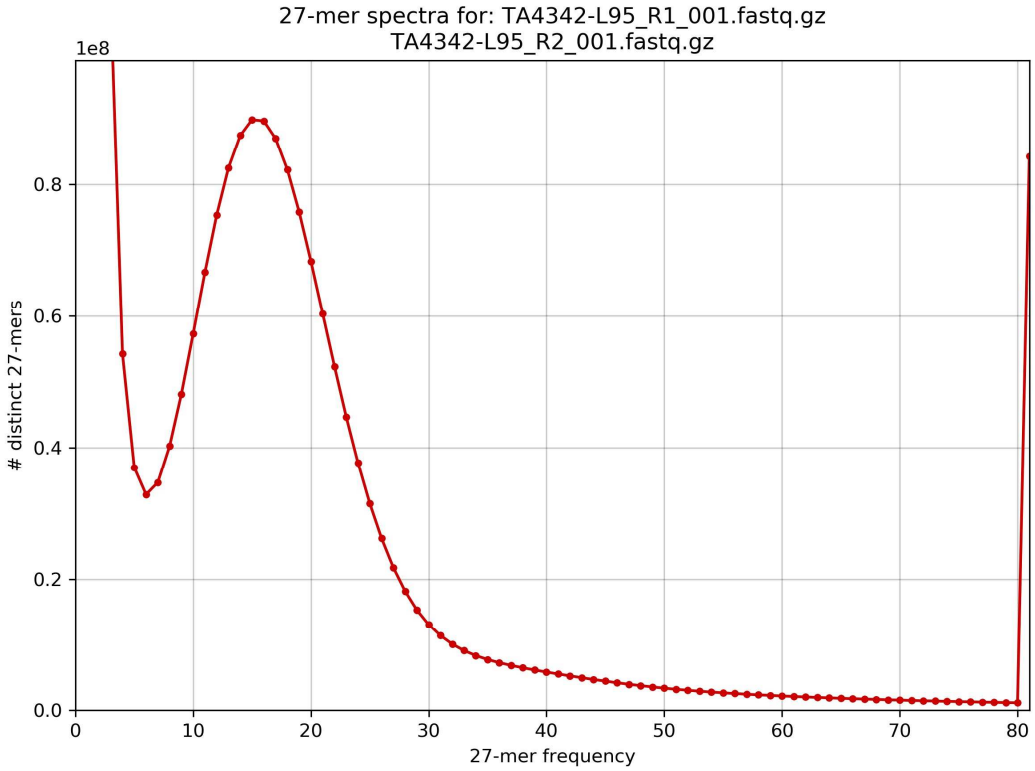


Figure 1: scer_pe_hist.png

1.3 Check insert size and distribution

Not executed as no references existed.

2 Contigging

Executed on XSEDE-COMET large memory node

Listing 5: Submission file

```
1  #!/bin/bash
2  #SBATCH --job-name="W2RAP"
3  #SBATCH --output="boeticum.%j.%N.out"
```

```
4  #SBATCH --partition=large-shared
5  #SBATCH --nodes=1
6  #SBATCH --ntasks-per-node=50
7  #SBATCH --mem=1200G
8  #SBATCH --export=ALL
9  #SBATCH -t 15:00:00
10 #SBATCH -A dsu103
11
12 #This job runs with 1 node, 50 cores per node for a total of 50 cores.
13 module use /home/villegar/modules
14 module load w2rap
15
16 module list
17
18 export OMP_PROC_BIND=spread
19 export MALLOC_PER_THREAD=1
20 mkdir -p contigs
21 w2rap-contigger -t 50 -m 1200 -d 50 -r TA4342-L95_R1_001.fastq.gz,TA4342-
   ←-
L95_R2_001.fastq.gz -o contigs -p scer_k200 --dump_all 1
```

Output contained in a directory called contigs

Listing 6: Command output

```
1  -rw-r--r--. 1 software 643M May 24 10:15 a.lines.efasta
2  -rw-r--r--. 1 software 644M May 24 10:15 a.lines.fasta
3  -rw-r--r--. 1 software 6.0M May 24 10:16 a.lines.src
4  -rw-r--r--. 1 software 50G May 24 18:46 frag_reads_orig.fastb
5  -rw-r--r--. 1 software 250G May 24 21:24 frag_reads_orig.qualp
6  -rw-r--r--. 1 software 983 May 24 10:16 large_K.frags.dist
7  -rw-r--r--. 1 software 101 May 24 10:16 large_K.frags.dist.png.FAIL
8  -rw-r--r--. 1 software 3.2M May 24 10:16 scer_k200_assembly.covs
9  -rw-r--r--. 1 software 21M May 24 10:16 scer_k200_assembly.lines
10 -rw-r--r--. 1 software 2.7M May 24 10:16 scer_k200_assembly.lines.npairs
11 -rw-r--r--. 1 software 37M May 24 10:16 scer_k200_assembly.lines.stats
12 -rw-r--r--. 1 software 21M May 24 10:16 scer_k200.fin.lines
13 -rw-r--r--. 1 software 2.8M May 24 10:16 scer_k200.fin.lines.npairs
14 -rw-r--r--. 1 software 1.8G May 24 10:13 scer_k200.large_K_clean.hbv
15 -rw-r--r--. 1 software 8.9G May 24 20:37 scer_k200.large_K_clean.paths
```

```
-rw-r--r--. 1 software 355M May 24 10:16 scer_k200.large_K_expanded.hbv
-rw-r--r--. 1 software 9.3G May 24 19:05 scer_k200.large_K_expanded.paths
-rw-r--r--. 1 software 355M May 24 10:16 scer_k200.large_K_final.hbv
-rw-r--r--. 1 software 9.3G May 24 19:24 scer_k200.large_K_final.paths
-rw-r--r--. 1 software 1.8G May 24 10:16 scer_k200.large_K.hbv
-rw-r--r--. 1 software 1.6G May 24 10:14 scer_k200.large_K_patched.hbv
-rw-r--r--. 1 software 9.1G May 24 19:47 scer_k200.large_K_patched.paths
-rw-r--r--. 1 software 8.9G May 24 20:18 scer_k200.large_K.paths
-rw-r--r--. 1 software 12G May 24 18:58 scer_k200.small_K.hbv
-rw-r--r--. 1 software 65G May 24 16:36 scer_k200.small_K.paths
-rw-r--r--. 1 software 983 May 24 10:16 small_K.frags.dist
-rw-r--r--. 1 software 101 May 24 10:16 small_K.frags.dist.png.FAIL
-rw-r--r--. 1 software 2.9K May 24 10:16 small_K.freqs
-rw-r--r--. 1 software 177 May 24 10:16 stats
```

---

3 Contig assessment

3.1 Check assembly contiguity

Listing 7: Executed code

```
module load abyss/2.0.1 module load
gcc/5.5.0 abyss-fac
contigs/a.lines.fasta
```

Listing 8: Command output

```
n       n:500   L50     min     N80     N50     N20     E-size  max
353973  284777  80465   500     1509    2354    4335    3640    56004

sum     name
637.6e6 contigs/a.lines.fasta
```

3.2 Compare PE reads to contigs

Listing 9: Executed code

```
module purge
module load KAT/2.4.1
kat comp -o scer_pe_v2_ctgs -t 8 -m 27 -H 100000000 -I 100000000 TA4342-←-
L95_R?_001.fastq.gz contigs/a.lines.fasta &> 3b.log &
```

Listing 10: Command output

```
1  -rw-r--r-- 1 software 456 Aug 21 10:55 scer_pe_v2_ctgs.dist_analysis.json
2  -rw-r--r-- 1 software 2.0M Aug 21 10:55 scer_pe_v2_ctgs-ends.mx
3  -rw-r--r-- 1 software 2.2M Aug 21 10:55 scer_pe_v2_ctgs-main.mx
4  -rw-r--r-- 1 software 96K Aug 21 10:55 scer_pe_v2_ctgs-main.mx.spectra-cn←-
.png
5  -rw-r--r-- 1 software 2.0M Aug 21 10:55 scer_pe_v2_ctgs-middle.mx
6  -rw-r--r-- 1 software 2.0M Aug 21 10:55 scer_pe_v2_ctgs-mixed.mx
7  -rw-r--r-- 1 software 1.0K Aug 21 10:55 scer_pe_v2_ctgs.stats
```

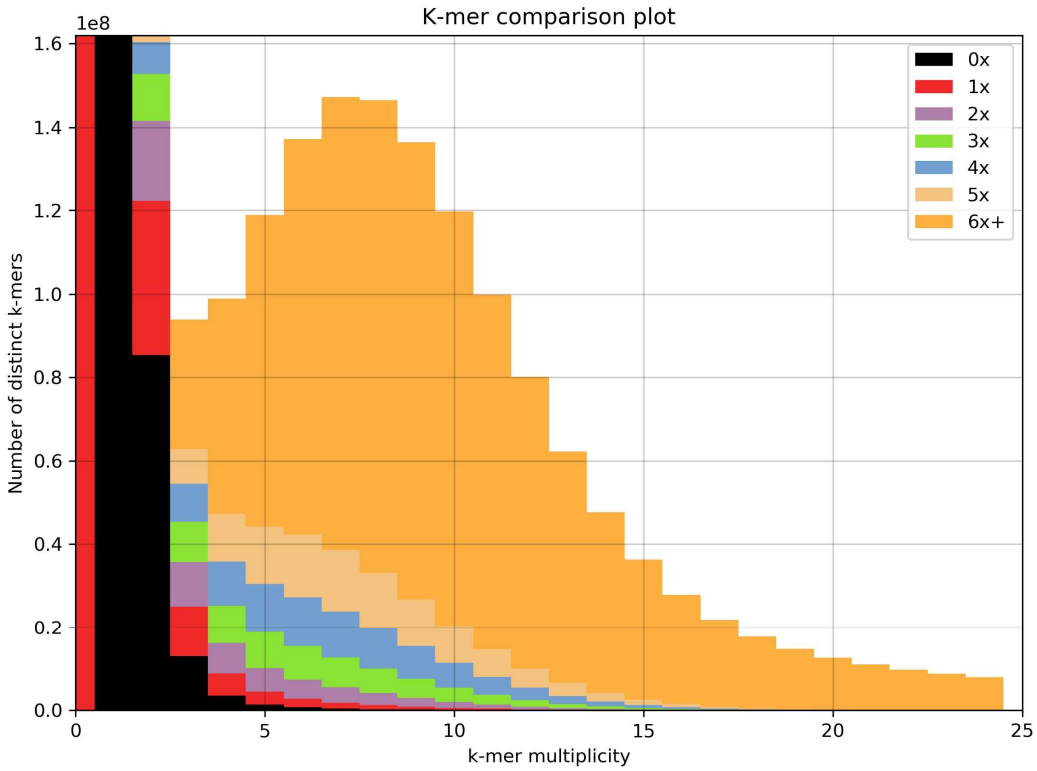


Figure 2: scer_pe_v2_ctgs-main.mx.spectra-cn.png

3.3 Assess assembly accuracy using QUAST

Not executed as no references existed.

3.4 Assess assembly completeness by aligning BUSCO genes

Listing 11: Executed code

```
1 module load w2rap/2018.4
2 CPUS=64
3 BUSCO.py -o busco_pe --in contigs/a.lines.efasta -l
  $BUSCO_PLANTS/←embryophyta -m genome -f --cpu $CPUS
```

Listing 12: Command output

```
1 drwxr-xr-x. 7 software2 191 May 25 21:09 augustus_output
2 drwxr-xr-x. 2 software2 239 May 25 21:08 blast_output
3 -rw-r--r--. 1 software2 29K May 25 21:09 full_table_busco_pe.tsv
4 drwxr-xr-x. 2 software2 10 May 25 10:47 hmmer_output
5 -rw-r--r--. 1 software2 18K May 25 21:09 missing_busco_list_busco_pe.tsv
```

```
6 -rw-r--r--. 1 software2 687 May 25 21:09 short_summary_busco_pe.txt
7 drwxr-xr-x. 2 software2 10 May 25 21:09 single_copy_busco_sequences
```

4 LMP processing

All the configuration and output files can be found inside the directory:

/stor02/boeoticum/MP/step4

4.1 Run FastQC to check read metrics

Listing 13: Executed code

```
1 #!/usr/bin/env bash
2
3 module load FastQC/0.11.7
4 mkdir fastqc
5 fastqc -o fastqc 2kb_R1.fastq 2kb_R2.fastq
6 fastqc -o fastqc 4kb_R1.fastq 4kb_R2.fastq
7 fastqc -o fastqc 8kb_R1.fastq 8kb_R2.fastq
```

Output stored inside the directory fastqc

Listing 14: Command output

```
-rw-r--r-- 1 software 243K Aug 23 11:04 2kb_R1_fastqc.html
-rw-r--r-- 1 software 324K Aug 23 11:04 2kb_R1_fastqc.zip
-rw-r--r-- 1 software 244K Aug 23 11:05 2kb_R2_fastqc.html
-rw-r--r-- 1 software 329K Aug 23 11:05 2kb_R2_fastqc.zip
-rw-r--r-- 1 software 235K Aug 23 11:23 4kb_R1_fastqc.html
-rw-r--r-- 1 software 279K Aug 23 11:23 4kb_R1_fastqc.zip
-rw-r--r-- 1 software 239K Aug 23 11:23 4kb_R2_fastqc.html
-rw-r--r-- 1 software 286K Aug 23 11:23 4kb_R2_fastqc.zip
-rw-r--r-- 1 software 239K Aug 23 11:50 8kb_R1_fastqc.html 10 -rw-r--r-- 1
software 325K Aug 23 11:50 8kb_R1_fastqc.zip
-rw-r--r-- 1 software 240K Aug 23 11:51 8kb_R2_fastqc.html
-rw-r--r-- 1 software 325K Aug 23 11:51 8kb_R2_fastqc.zip
```

See the following appendix to see FastQC reports:

- 2kb-R1: 9.1.3
- 2kb-R2: 9.1.4
- 4kb-R1: 9.1.5
- 4kb-R2: 9.1.6
- 8kb-R1: 9.1.7

- 8kb-R2: 9.1.8

4.2 Identify good LMP reads

Listing 15: Executed code

```
1 #!/bin/bash
2 module load w2rap/2018.4
3 module load python/2.7
4 CPUS=64 5 lmp_processing libs_list $CPUS 6 unset CPUS
```

A file called libs_list was create with the following contents:

Listing 16: libs_list

```
1 /stor02/boeoticum/MP/2kb_R1.fastq
2 /stor02/boeoticum/MP/2kb_R2.fastq
3 /stor02/boeoticum/MP/4kb_R1.fastq
4 /stor02/boeoticum/MP/4kb_R2.fastq
5 /stor02/boeoticum/MP/8kb_R1.fastq
6 /stor02/boeoticum/MP/8kb_R2.fastq
```

Output stored inside the directory nextclip

Listing 17: Command output

---

1 %-rw-r--r--. 1 software 0 Jul 19 14:25 TA-4342-L95-2←-
kb_NoIndex_L005_nc_ABC_R1.fastq

2 -rw-r--r--. 1 software 0 Jul 19 14:25 2kb_nc_ABC_R1.fastq

3 -rw-r--r--. 1 software 0 Jul 19 14:25 2kb_nc_ABC_R2.fastq

4 -rw-r--r--. 1 software 0 Jul 19 14:25 4kb_nc_ABC_R1.fastq

5 -rw-r--r--. 1 software 0 Jul 19 14:25 4kb_nc_ABC_R2.fastq

6 -rw-r--r--. 1 software 0 Jul 19 14:25 8kb_nc_ABC_R1.fastq

7 -rw-r--r--. 1 software 0 Jul 19 14:25 8kb_nc_ABC_R2.fastq

---

5 QC processed LMPs

All the configuration and output files can be found inside the directory:

/stor02/boeoticum/MP/step5

5.1 Use KAT comp to check for LMP representation issues

Not executed as the output of Nextclip (Listing 14) was empty.

5.2 Check the LMP insert size distribution

Listing 18: Executed code

```
1 module load bwa/0.7.17
2 bwa index -p boeoticum /stor02/boeoticum/PE/contigs/a.lines.fasta
```

Output stored inside the directory bwa

Listing 19: Command output

```
1 -rw-r--r--. 1 software 44K Jul 23 21:46 boeoticum.amb
2 -rw-r--r--. 1 software 17M Jul 23 21:46 boeoticum.ann
3 -rw-r--r--. 1 software 628M Jul 23 21:46 boeoticum.bwt
4 -rw-r--r--. 1 software 157M Jul 23 21:46 boeoticum.pac
5 -rw-r--r--. 1 software 314M Jul 23 21:50 boeoticum.sa
```

5.3 Calculate the read and fragment coverage

Manual calculation required.

6 **Scaffolding**

All the configuration and output files can be found inside the directory:

/stor02/boeoticum/MP/step6

6.1 Make a SOAPdenovo config file

The file called soap.config was create with the following contents:

Listing 20: soap.config file

```
1  [LIB]
2  avg_ins=180
3  q1=/stor02/boeoticum/PE/180bp_R1.fastq.gz
4  q2=/stor02/boeoticum/PE/180bp_R2.fastq.gz 5
6  [LIB]
7  avg_ins=300
8  q1=/stor02/boeoticum/PE/300bp_R1.fastq.gz
9  q2=/stor02/boeoticum/PE/300bp_R2.fastq.gz
10
11 [LIB]
12 avg_ins=400
13 q1=/stor02/boeoticum/PE/400bp_R1.fastq.gz
14 q2=/stor02/boeoticum/PE/400bp_R2.fastq.gz 15
16 [LIB]
```

```
17 avg_ins=2000
18 reverse_seq=1
19 q1=/stor02/boeoticum/MP/2kb_R1.fastq
20 q2=/stor02/boeoticum/MP/2kb_R2.fastq 21
22 [LIB]
23 avg_ins=4000
24 reverse_seq=1
25 q1=/stor02/boeoticum/MP/4kb_R1.fastq
26 q2=/stor02/boeoticum/MP/4kb_R2.fastq 27
28 [LIB]
29 avg_ins=8000
30 reverse_seq=1
31 q1=/stor02/boeoticum/MP/8kb_R1.fastq
32 q2=/stor02/boeoticum/MP/8kb_R2.fastq
```

6.2 Run the "prepare -> map -> scaff" pipeline

Listing 21: Executed code

```
1 #!/bin/bash
2 module load w2rap/2018.4
```

```
3 CPUS=64
4
5 s_prepare -g boeoticum -K 71 -c contigs/a.lines.fasta 2>&1
6 s_map -k 31 -s soap.config -p $CPUS -g boeoticum > boeoticum.map.log 2>&1
7 s_scaff -p 16 -g boeoticum > boeoticum.scaff.log 2>&1
8
9 unset CPUS
```

Output stored inside the directory scaffolds

Listing 22: Command output

```
1 -rw-r--r-- 1 software        0 Jul 27 09:15 boeoticum.Arc
2 -rw-r--r-- 1 software 25K Jul 27 13:29 boeoticum.bubbleInScaff
3 -rw-r--r-- 1 software 641M Jul 27 09:15 boeoticum.contig
4 -rw-r--r-- 1 software 4.6M Jul 27 09:15 boeoticum.ContigIndex
5 -rw-r--r-- 1 software 3.8M Jul 27 14:09 boeoticum.contigPosInscaff
6 -rw-r--r-- 1 software 12M Jul 27 09:15 boeoticum.conver
7 -rw-r--r-- 1 software        0 Jul 27 13:29 boeoticum.gapSeq
8 -rw-r--r-- 1 software 69M Jul 27 13:29 boeoticum.links
9 -rw-r--r-- 1 software 2.7K Jul 27 13:24 boeoticum.map.log
```

```
-rw-r--r-- 1 software 11M Jul 27 13:24 boeoticum.newContigIndex
-rw-r--r-- 1 software 143 Jul 27 13:24 boeoticum.peGrads
-rw-r--r-- 1 software  77 Jul 27 09:15 boeoticum.preGraphBasic
-rw-r--r-- 1 software 13G Jul 27 13:24 boeoticum.readInGap.gz
-rw-r--r-- 1 software 5.7G Jul 27 13:24 boeoticum.readOnContig.gz
-rw-r--r-- 1 software 26M Jul 27 13:29 boeoticum.scaf
-rw-r--r-- 1 software 19K Jul 27 14:10 boeoticum.scaff.log
-rw-r--r-- 1 software 4.9M Jul 27 13:29 boeoticum.scaf_gap
-rw-r--r-- 1 software 741M Jul 27 14:09 boeoticum.scafSeq
-rw-r--r-- 1 software 1.8K Jul 27 14:10 boeoticum.scafStatistics
-rw-r--r-- 1 software 12M Jul 27 09:15 boeoticum.updated.edge
```

6.3 Recover gaps from contigging stage

Listing 23: Executed code

```
#!/bin/bash
module load w2rap/2018.4
module load python/2.7
SOAP_n_remapper.py boeoticum.contigPosInscaff boeoticum.scafSeq
boeoticum.←contig boeoticum.fasta
```

Output stored inside the directory scaffolds

Listing 24: Command output

```
1 -rw-r--r-- 1 software 733M Aug 7 17:05 boeoticum.fasta
```

6.4 Collapse repeats surrounding gaps

Listing 25: Executed code

```
1 #!/bin/bash
2 module load w2rap/2018.4
3 module load python/2.7
4
5 SOAP_n_collapser.py boeoticum.fasta boeoticum-collapsed.fasta &> collapser←-
.log &
```

Output stored inside the directory scaffolds

Listing 26: Command output

```
1 -rw-r--r-- 1 software 732M Aug 23 11:37 boeoticum-collapsed.fasta
2 -rw-r--r-- 1 software  70 Aug 23 11:40 collapser.log
```

7 Scaffold validation

7.1 Check assembly contiguity

All the configuration and output files can be found inside the directory:

/stor02/boeoticum/MP/step7

Listing 27: Executed code

```
1 module load abyss/2.0.1 module load gcc/5.5.0 abyss-fac
2 /stor02/boeoticum/MP/step6/scaffolds/boeoticum.fasta
3
```

Listing 28: Command output

```
   n          n:500   L50     min   N80   N50   N20        E-size  max
1  231495 165359 27068   500   2152  6351  13783  8566    71545
2
3  sum        name
4  630.2e6 /stor02/boeoticum/MP/step6/scaffolds/boeoticum.fasta
5
```

7.2 Compare PE reads to scaffolds

Listing 29: Executed code

```
1 module load KAT/2.4.1
2 kat comp -t 16 -m 31 -H10000000000 -I10000000000 -o reads_vs_scaffolds
  ’./←TA4342-L95_R1_001.fastq.gz ./TA4342-L95_R2_001.fastq.gz’
  ./boeoticum.←scafSeq &> 7b.log &
```

Listing 30: Command output

```
1  -rw-r--r-- 1 software 4.8K Aug 23 12:43 7b.log
2  -rw-r--r-- 1 software 6.5K Aug 23 12:42 reads_vs_scaffolds.dist_analysis.←json
3  -rw-r--r-- 1 software 2.0M Aug 23 12:42 reads_vs_scaffolds-main.mx
4  -rw-r--r-- 1 software 95K Aug 23 12:42 reads_vs_scaffolds-main.mx.spectra←-
-cn.png
5  -rw-r--r-- 1 software 929 Aug 23 12:42 reads_vs_scaffolds.stats
```

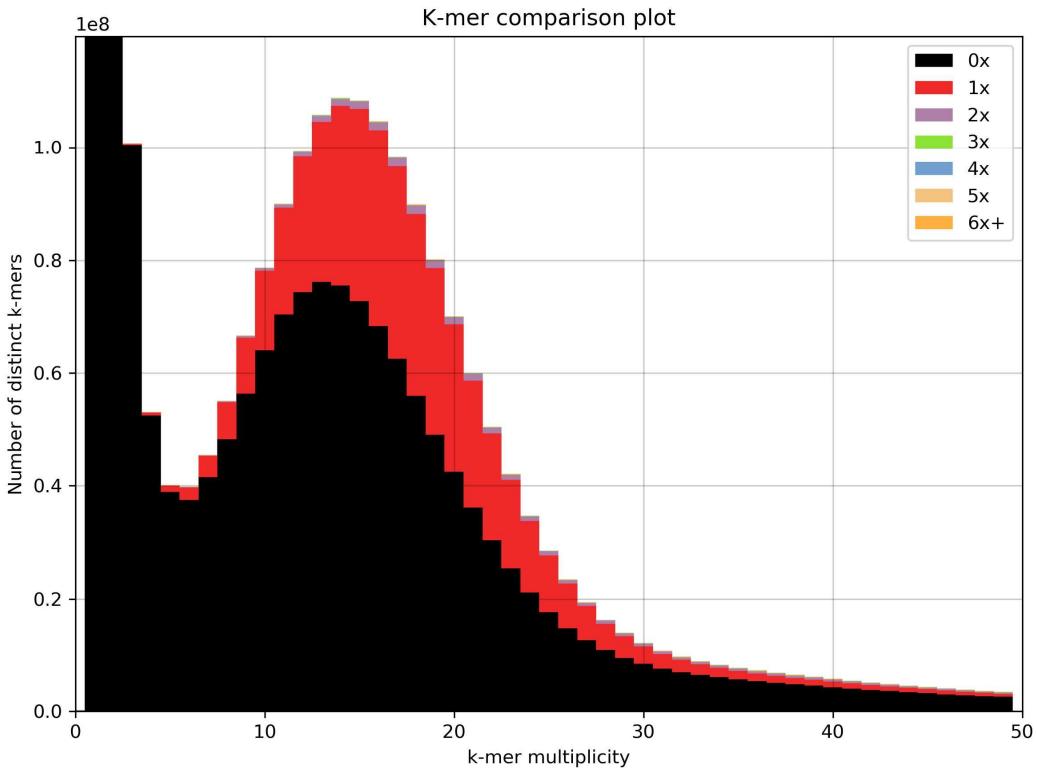


Figure 3: reads_vs_scaffolds-main.mx.spectra-cn.png

7.3 Assess assembly accuracy using QUAST

Not executed as no references existed.

7.4 Check assembly completeness by aligning BUSCO genes

Listing 31: Executed code

```
1 module load w2rap/2018.4
2 CPUS=38
3 BUSCO.py -o busco_lmp --in scaffolds/boeoticum.fasta -l
  $BUSCO_PLANTS/←embryophyta -m genome -f --cpu $CPUS
```

Output stored inside the directory run_busco_lmp

Listing 32: Command output

```
1 drwxr-xr-x 7 software 192 Aug 9 00:12 augustus_output
2 drwxr-xr-x 2 software 219 Aug 9 00:12 blast_output
3 -rw-r--r-- 1 software 29K Aug 9 00:12 full_table_busco_lmp.tsv
4 drwxr-xr-x 2 software  6 Aug 8 13:10 hmmer_output
5 -rw-r--r-- 1 software 18K Aug 9 00:12 missing_busco_list_busco_lmp.tsv
6 -rw-r--r-- 1 software 694 Aug 9 00:12 short_summary_busco_lmp.txt
7 drwxr-xr-x 2 software  6 Aug 9 00:12 single_copy_busco_sequences
```

8 **Create release FASTA**

9 Appendices

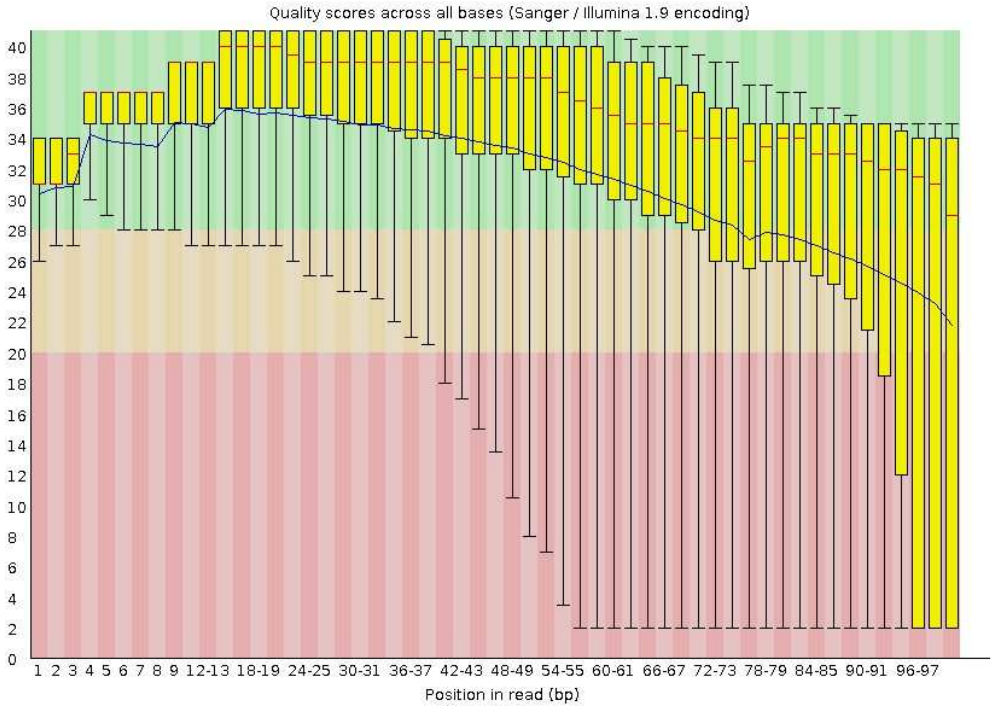
9.1 FastQC output

9.1.1 PE-R1

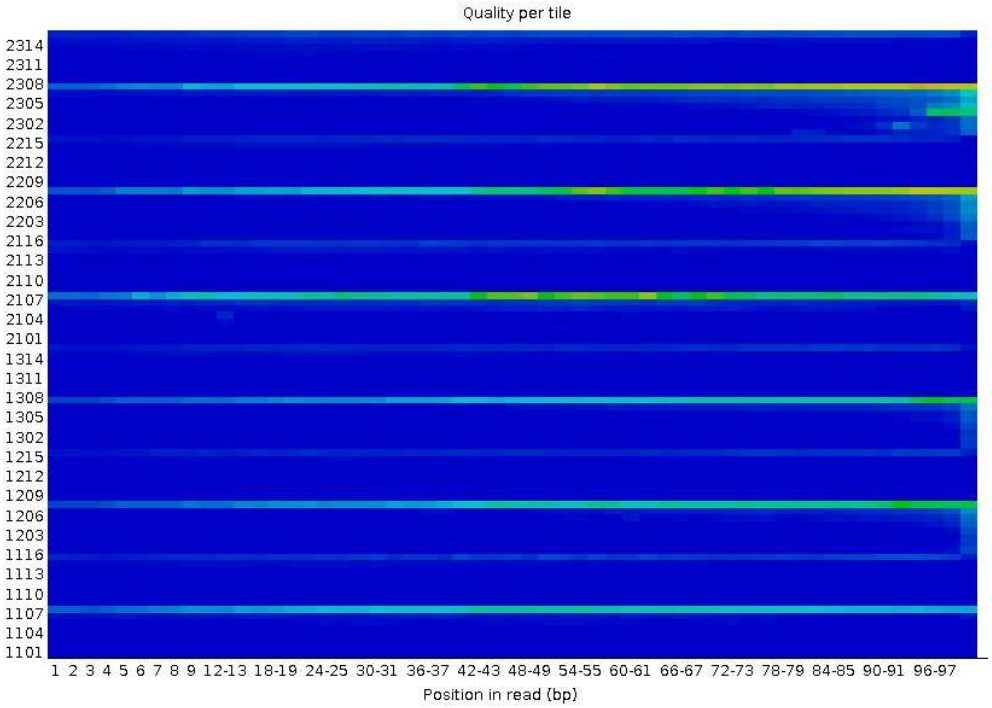
- **Basic Statistics**

Measure	Value
Filename	TA4342- L95_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	716189767
Sequences flagged as poor quality	0
Sequence length	100
%GC	45

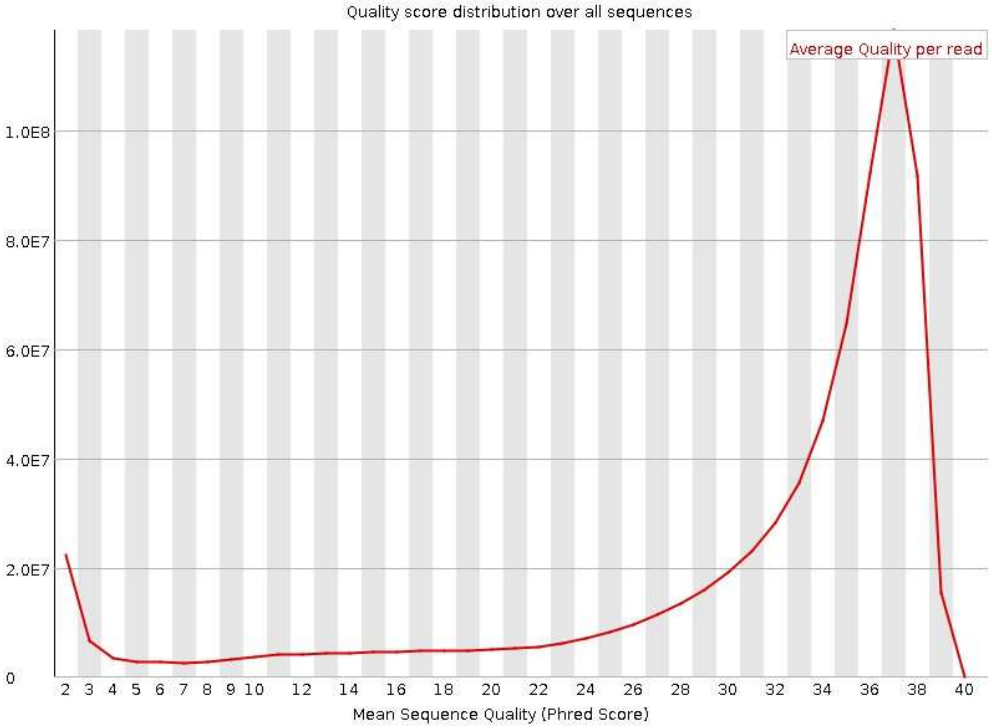
- **Per base sequence quality**


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85 90-91 96-97
Position in read (bp)

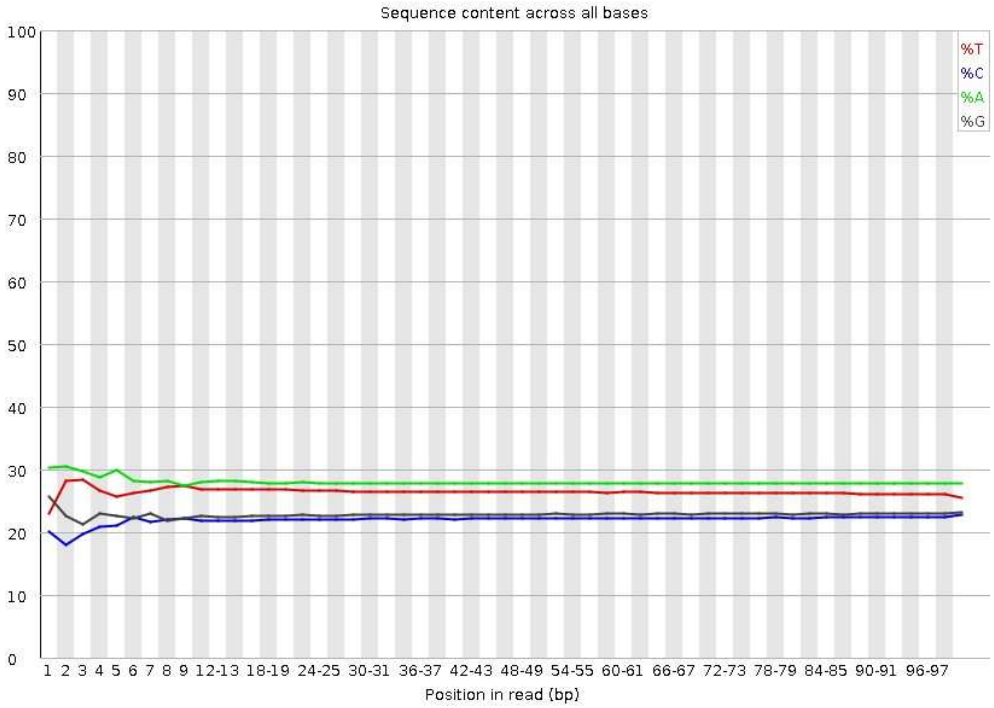
Per tile sequence quality



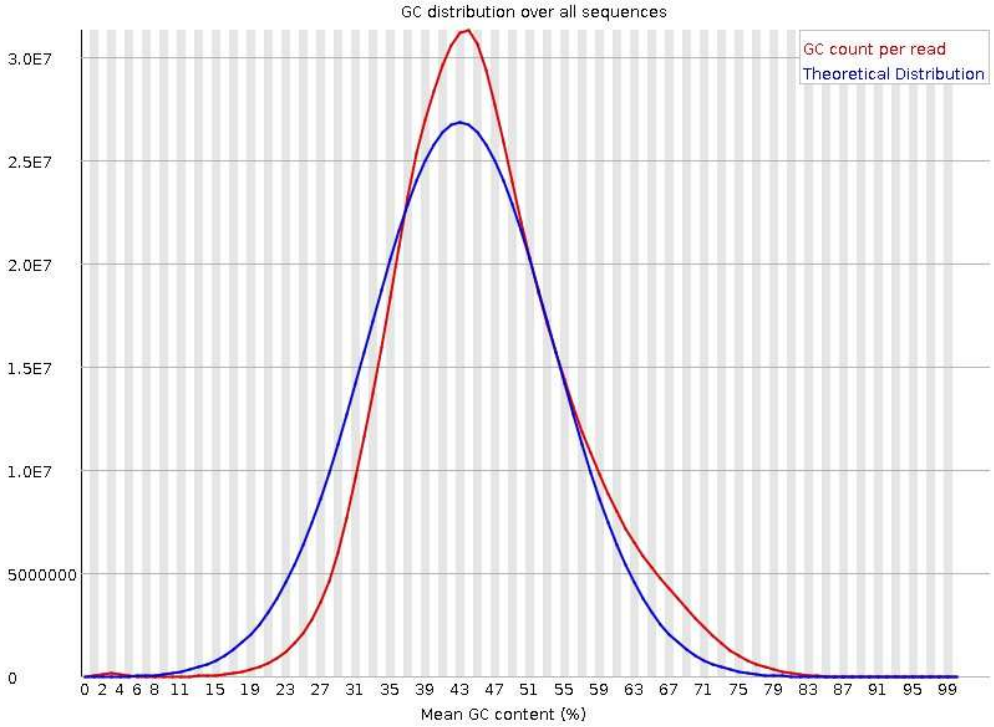
- **Per sequence quality scores**



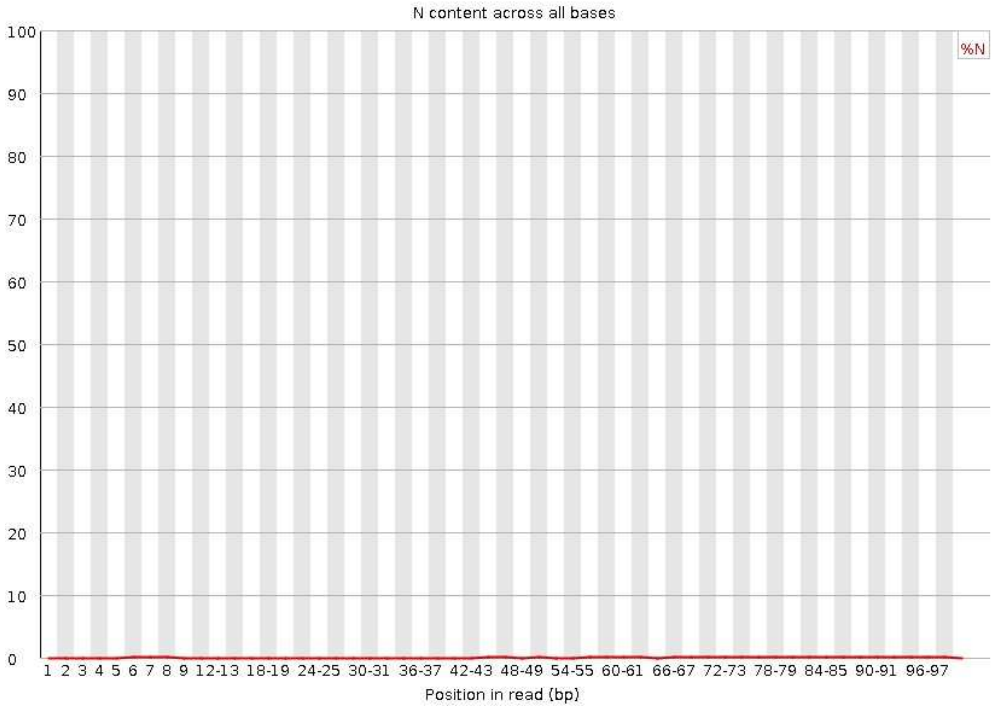
Per base sequence content



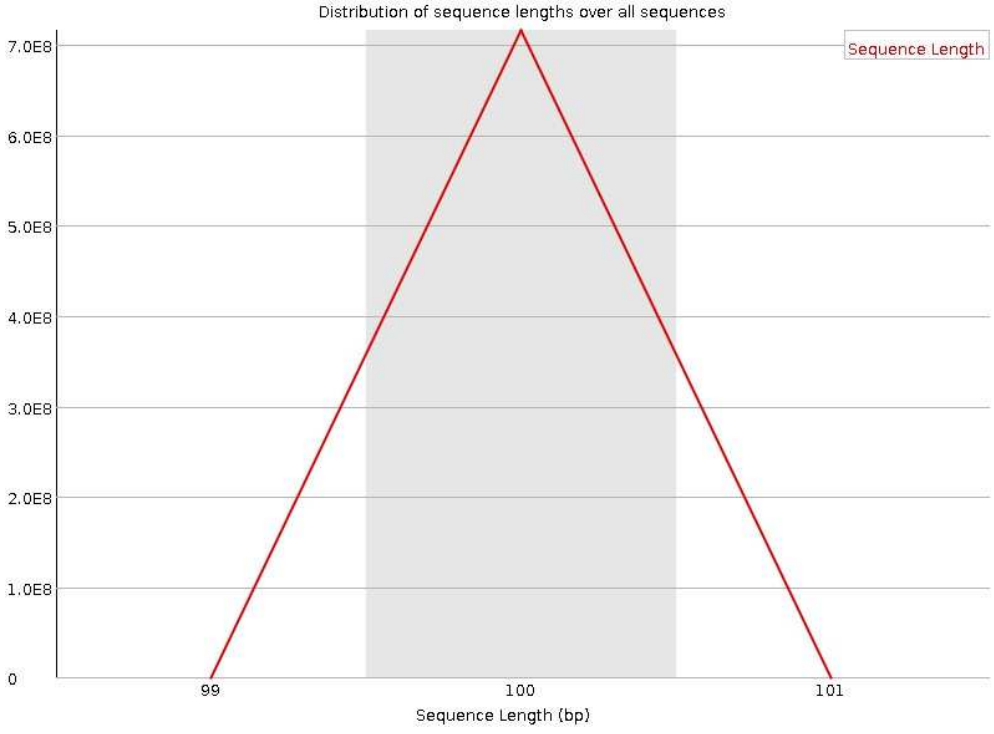
- **Per sequence GC content**



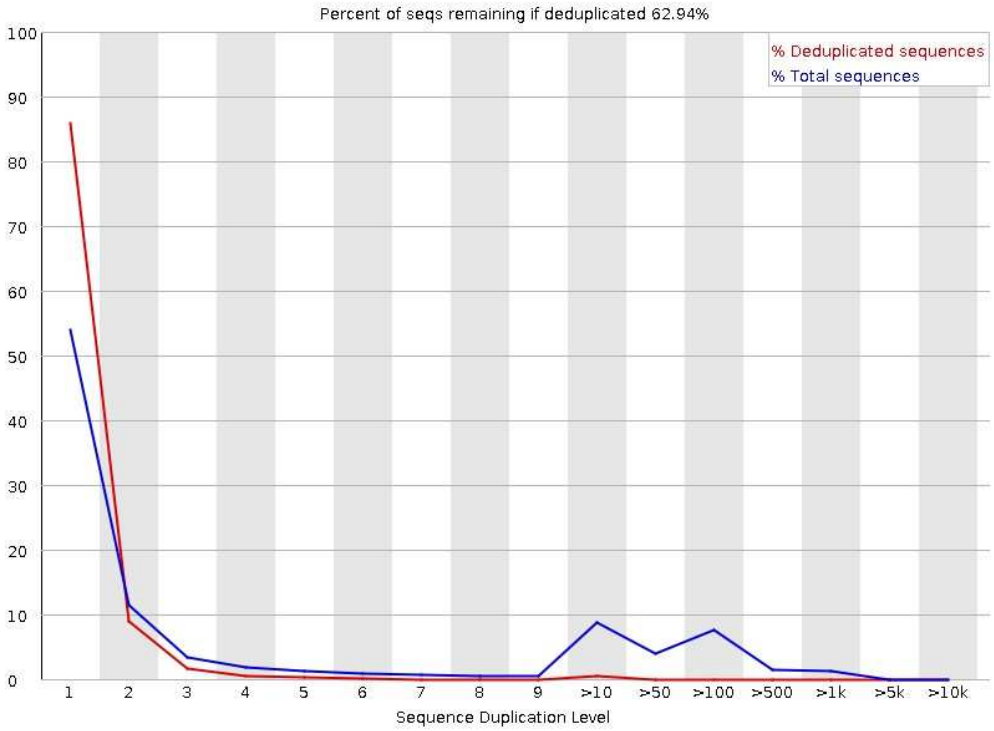
Per base N content



- **Sequence Length Distribution**



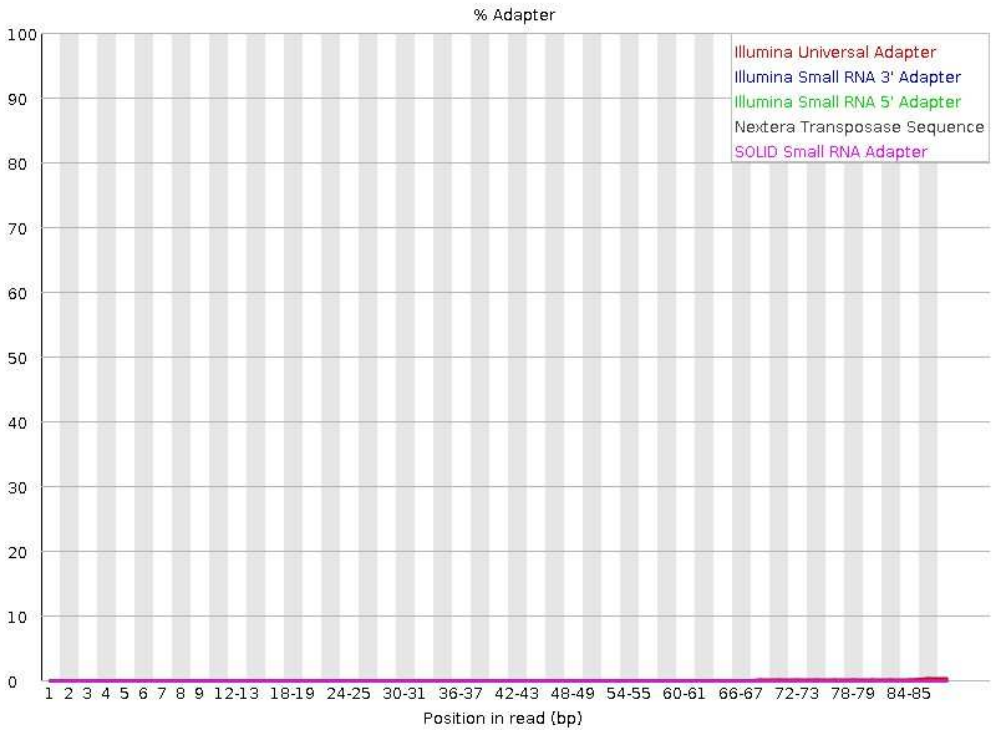
Sequence Duplication Levels



- **Overrepresented sequences**

No overrepresented sequences

- **Adapter Content**

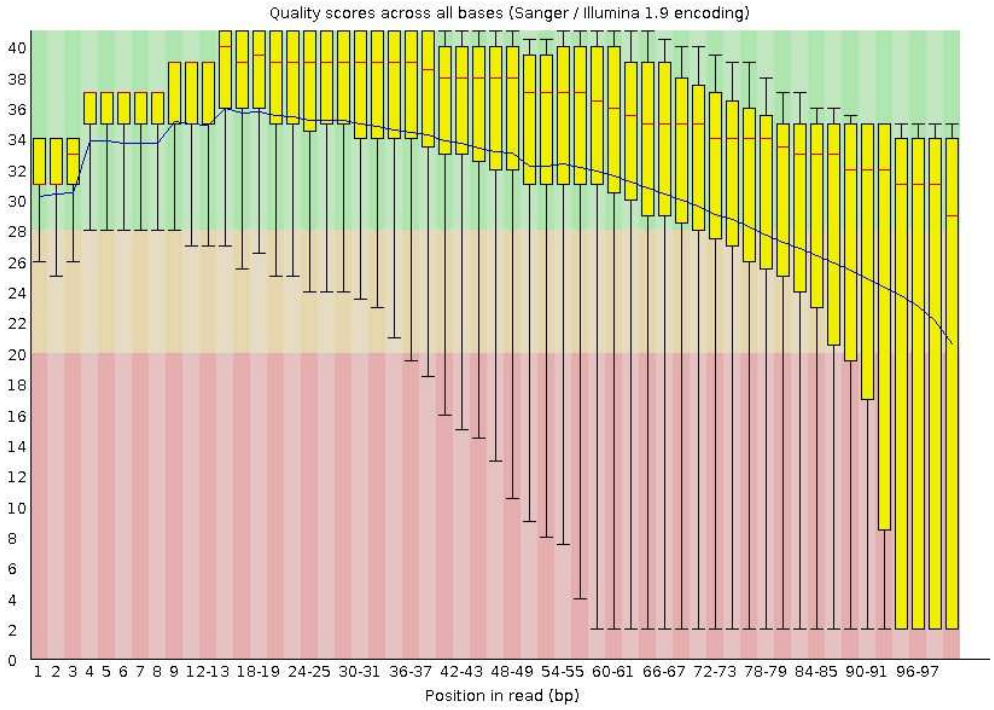

% Adapter
100
90
80
70
60
50
40
30
20
10
0
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85
Position in read (bp)

9.1.2 PE-R2

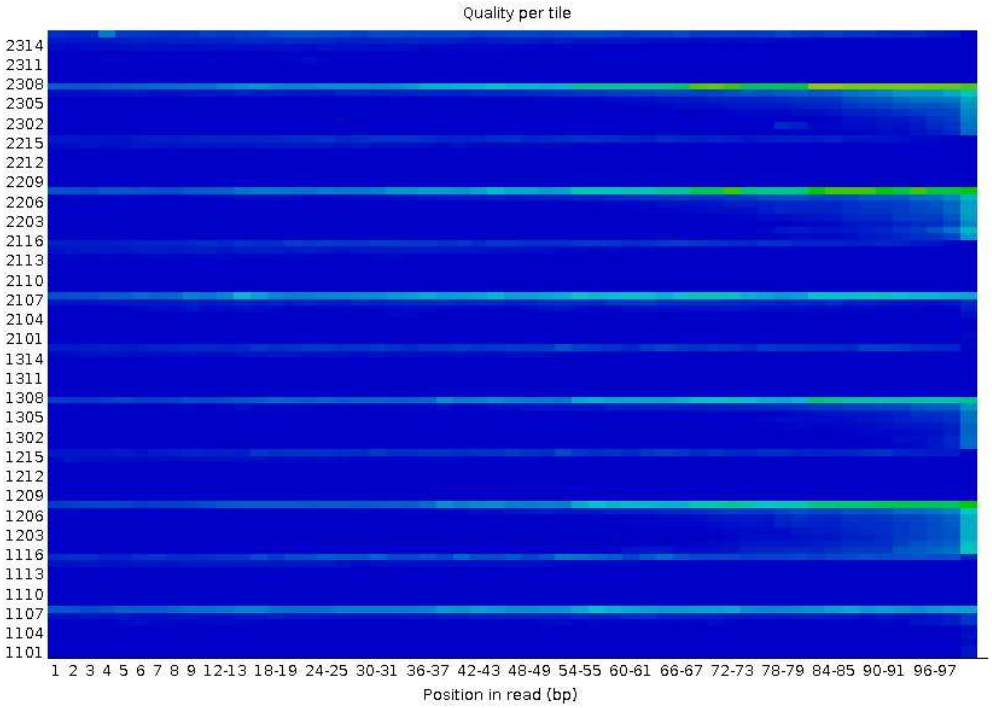
- **Basic Statistics**

Measure	Value
Filename	TA4342- L95_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	716189767
Sequences flagged as poor quality	0
Sequence length	100
%GC	45

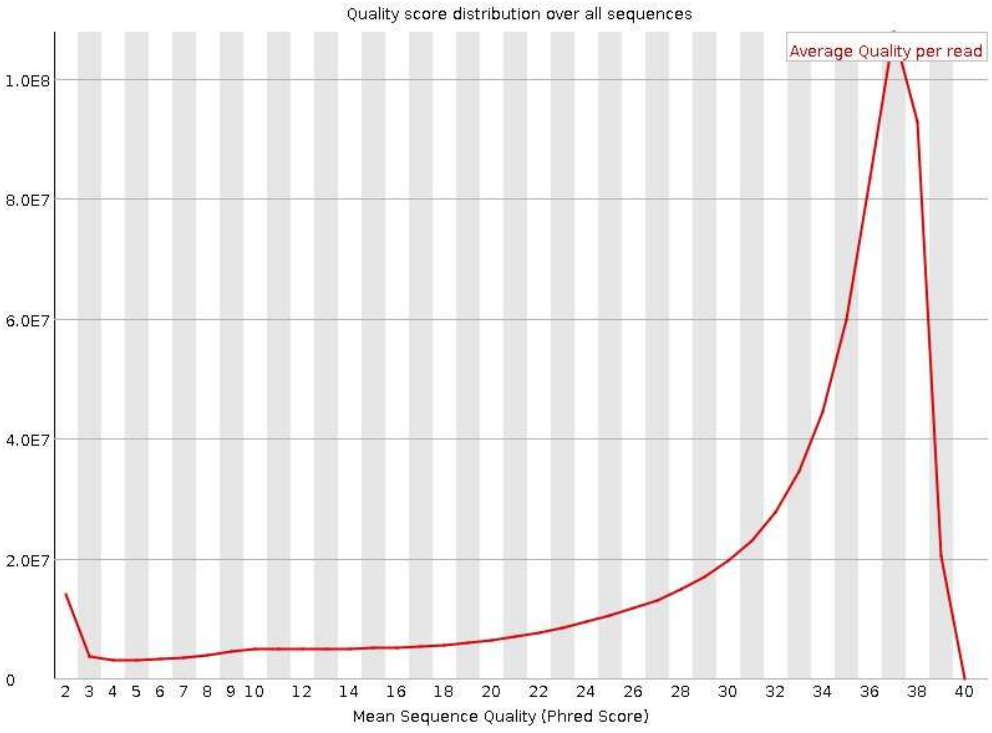
- **Per base sequence quality**



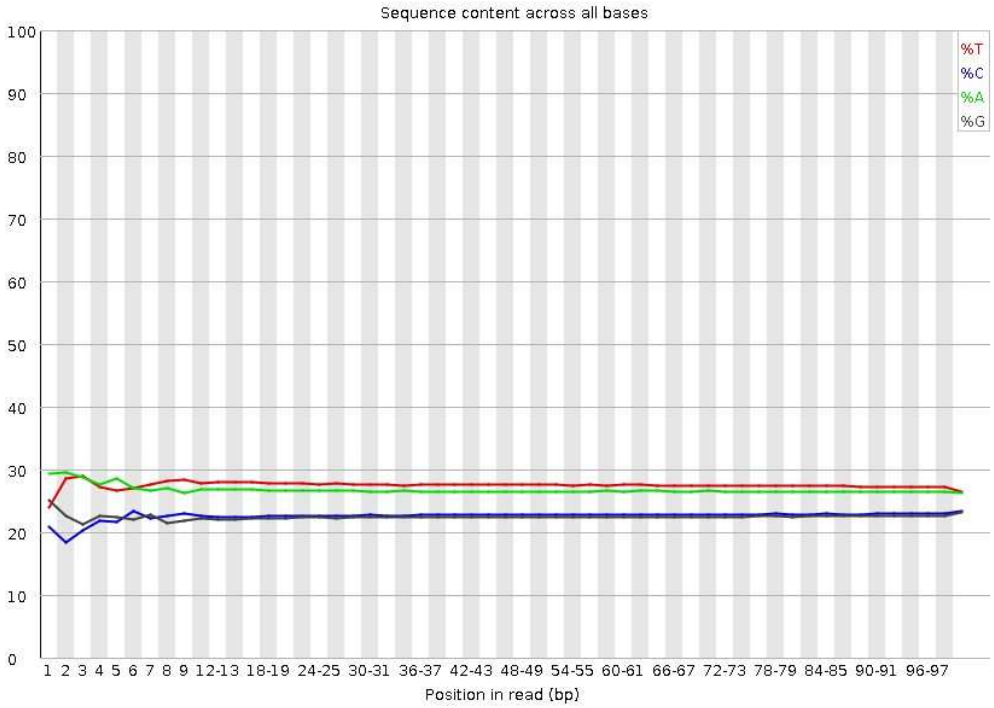
Per tile sequence quality



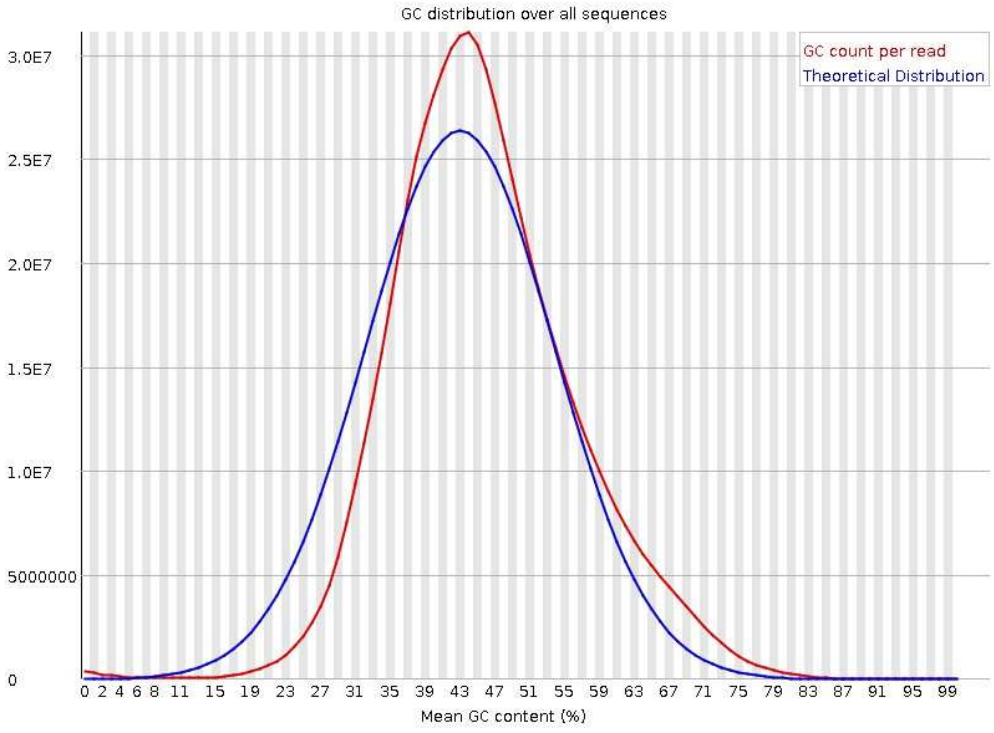
- **Per sequence quality scores**



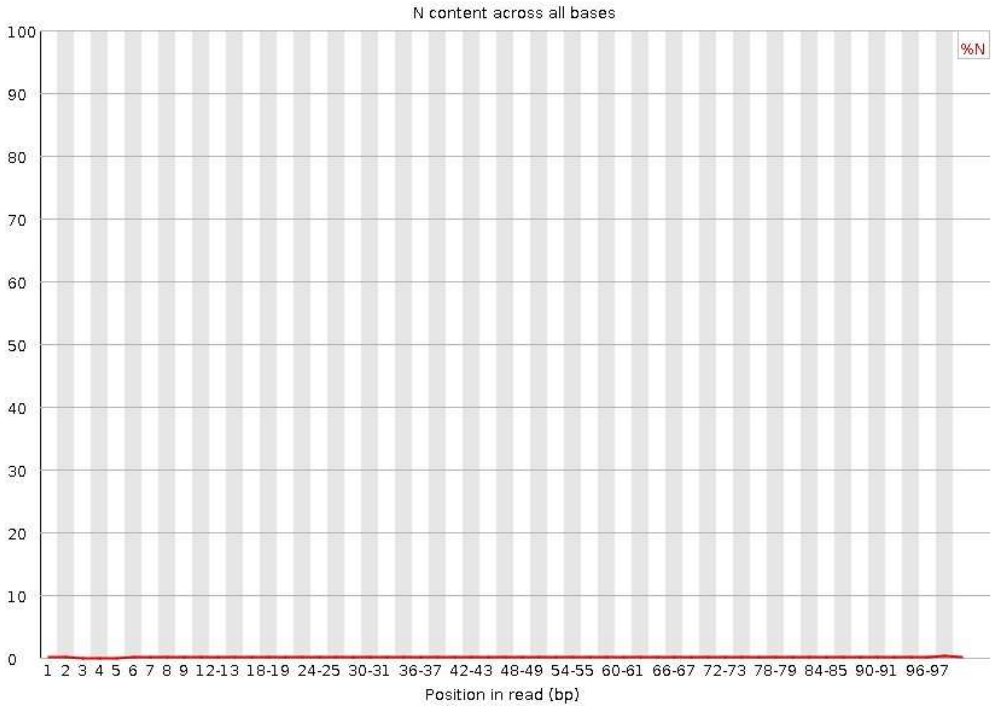
Per base sequence content



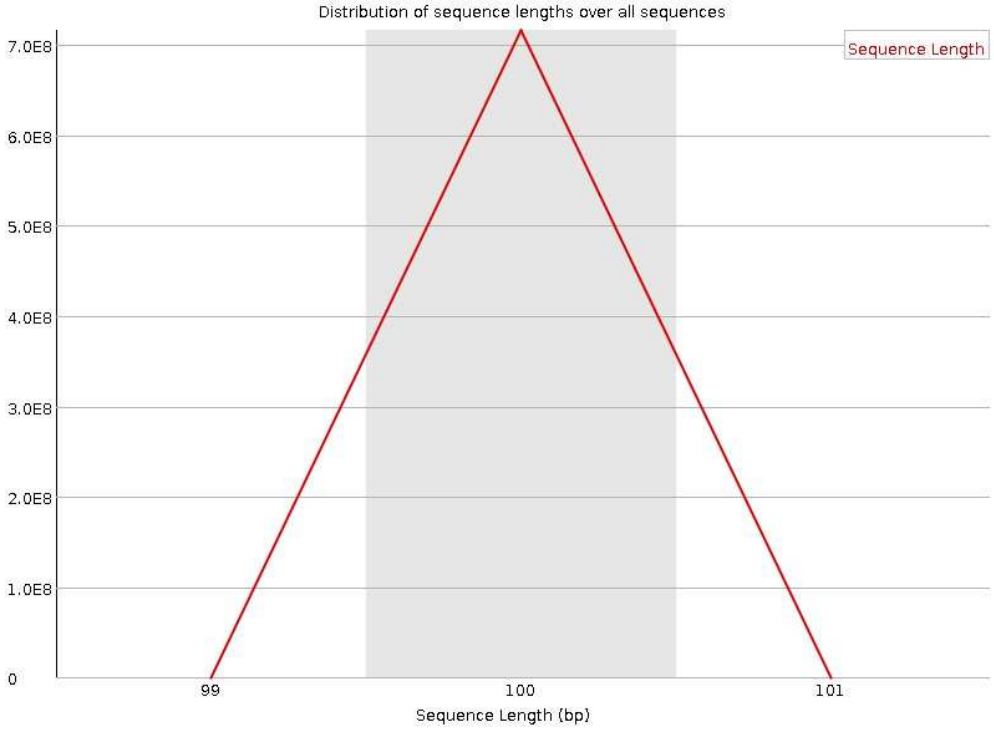
- **Per sequence GC content**



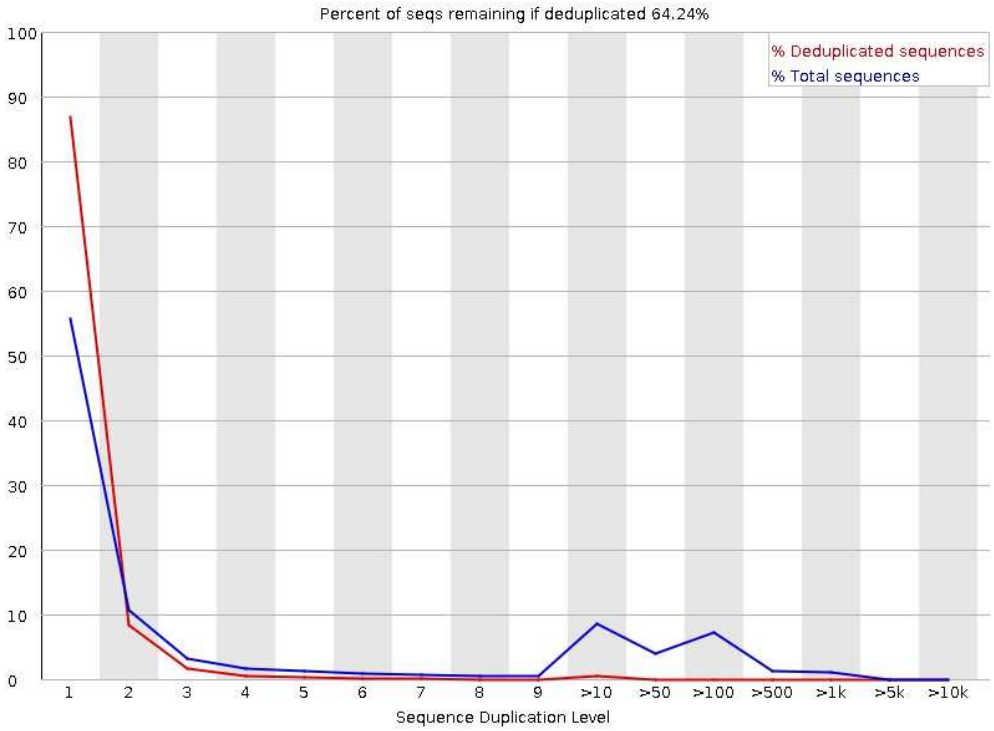
Per base N content



- **Sequence Length Distribution**



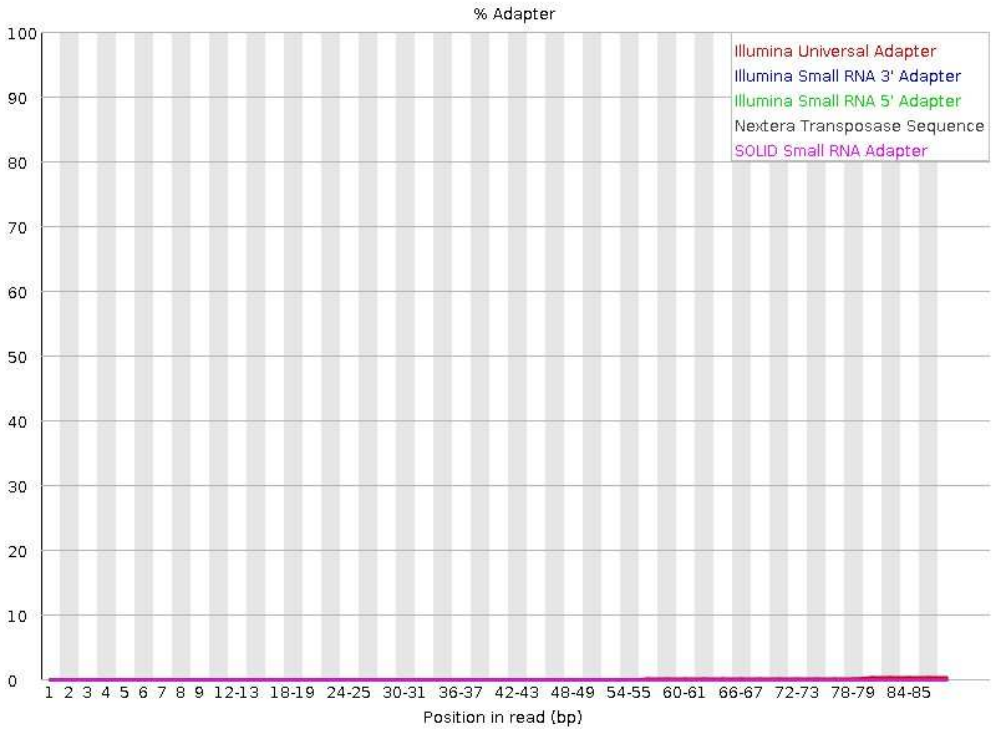
Sequence Duplication Levels



- **Overrepresented sequences**

No overrepresented sequences

- **Adapter Content**

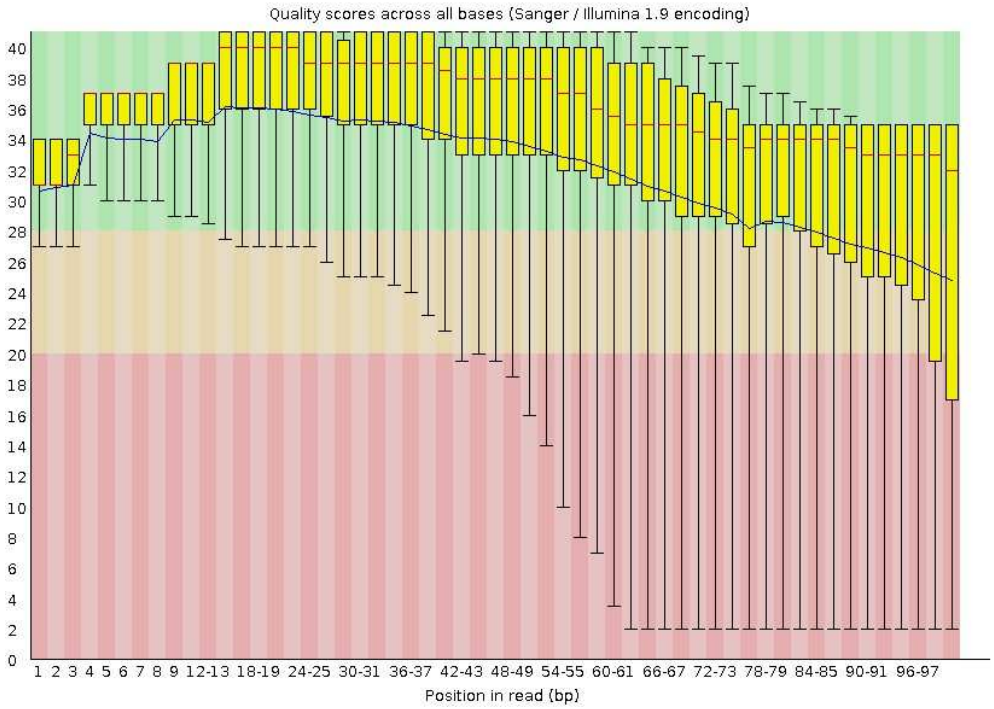

% Adapter
100
90
80
70
60
50
40
30
20
10
0
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85
Position in read (bp)

9.1.3 2kb-R1

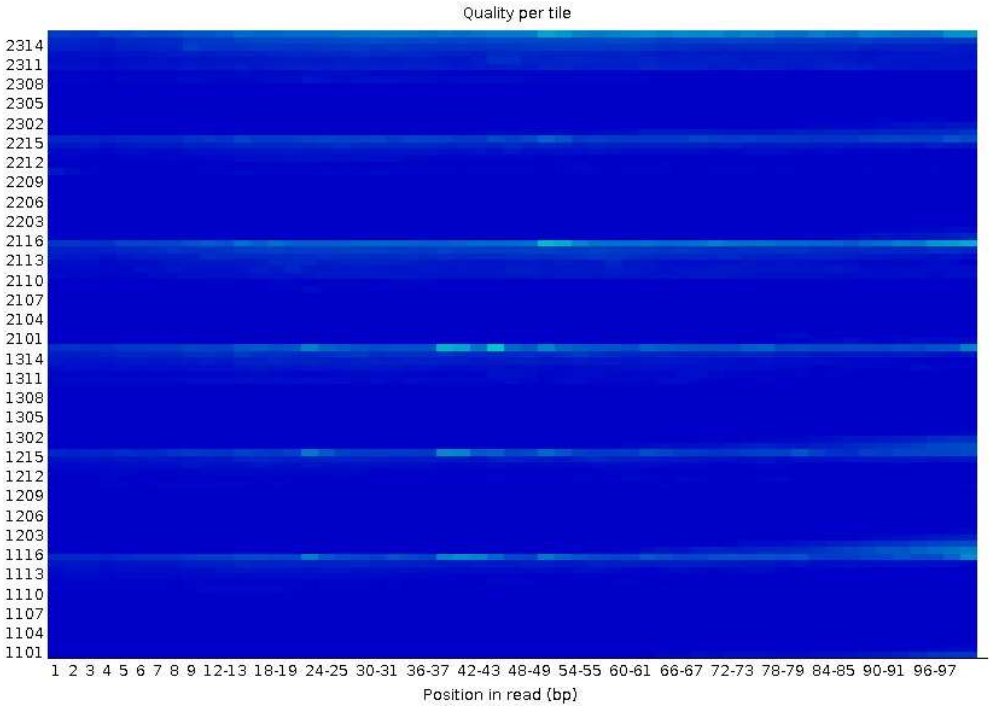
- **Basic Statistics**

Measure	Value
Filename	2kb_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	198461816
Sequences flagged as poor quality	0
Sequence length	100
%GC	46

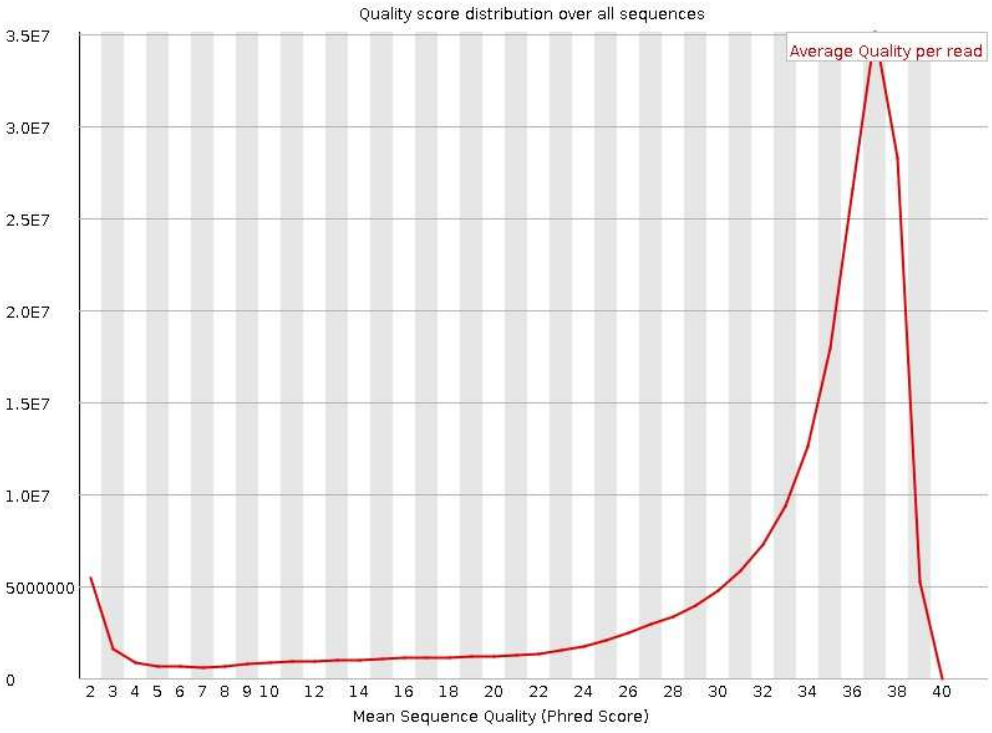
- **Per base sequence quality**



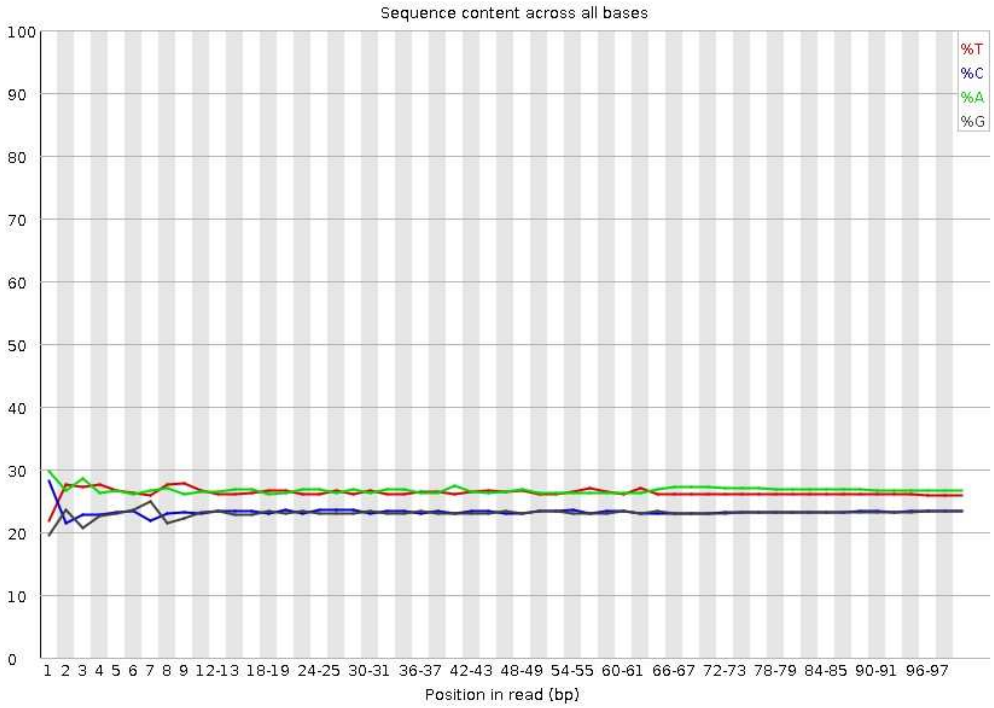
Per tile sequence quality



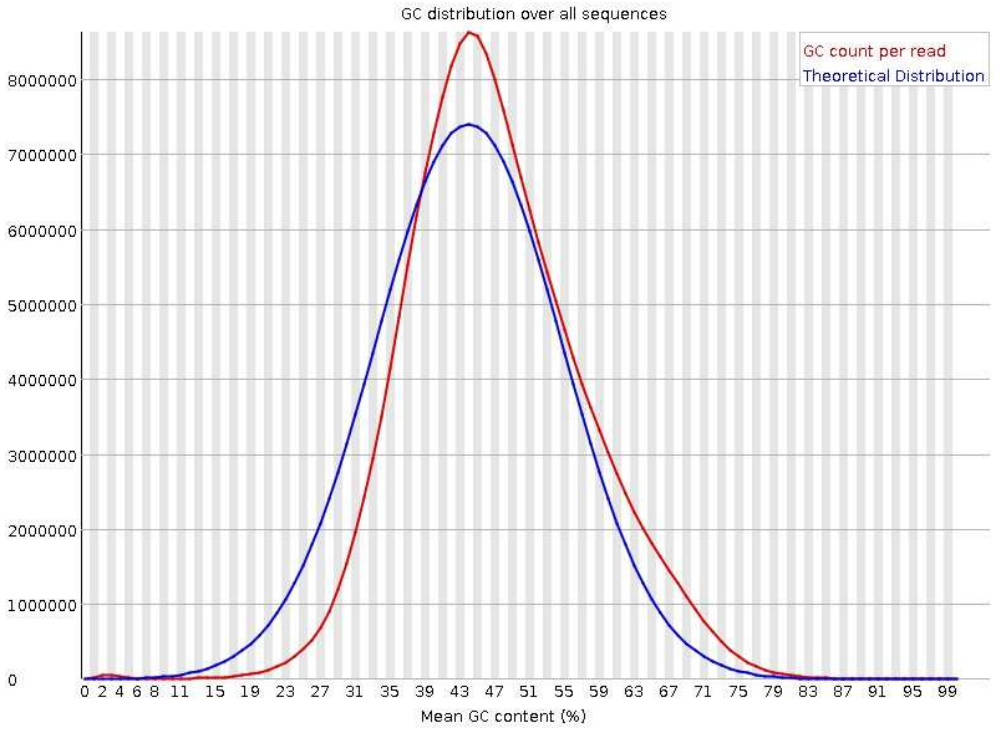
- **Per sequence quality scores**



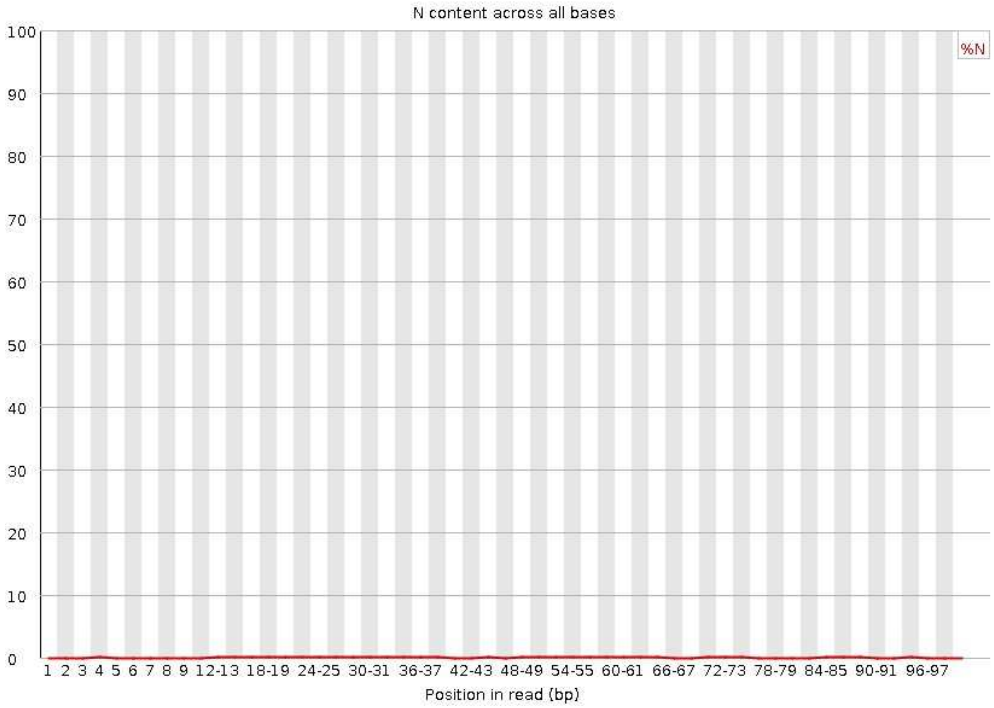
Per base sequence content



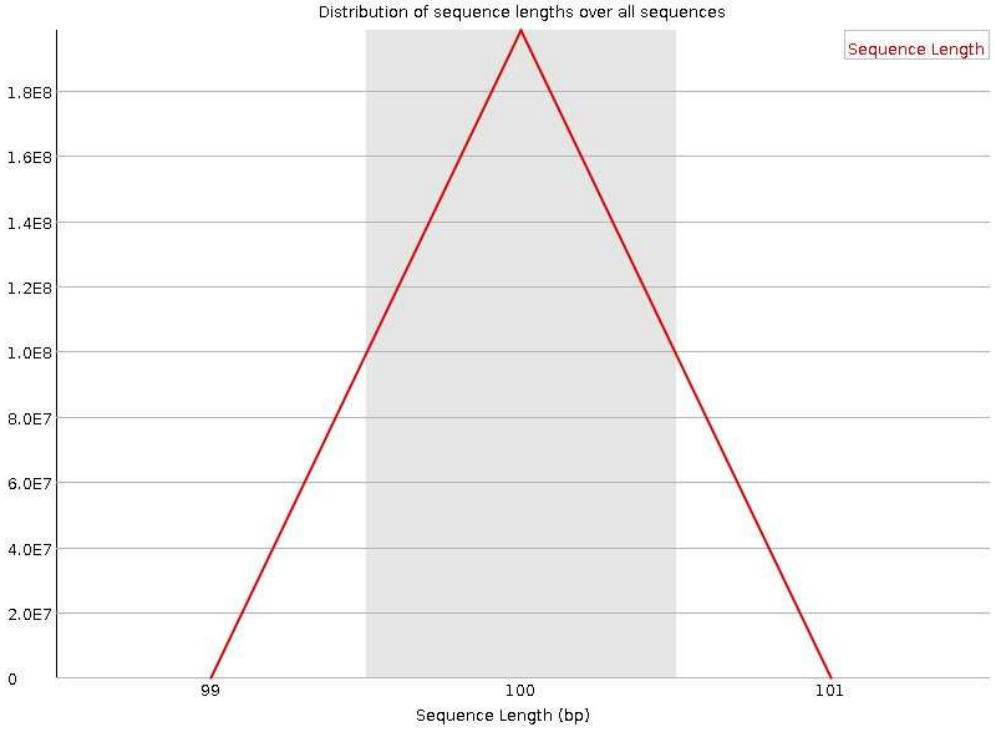
- **Per sequence GC content**



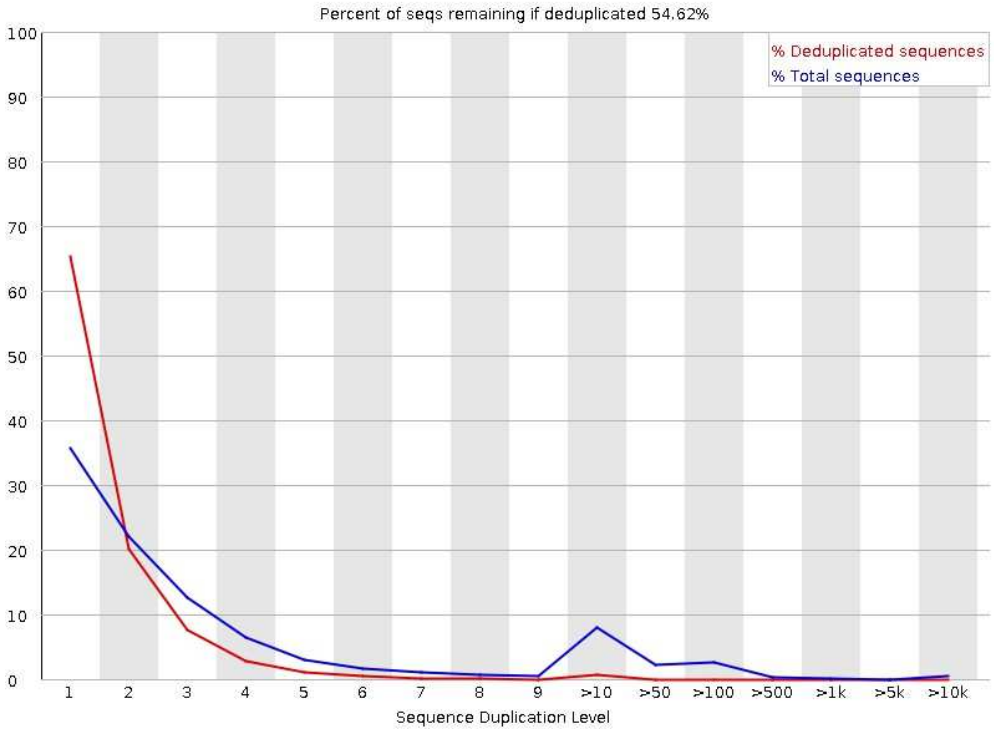
Per base N content



- **Sequence Length Distribution**



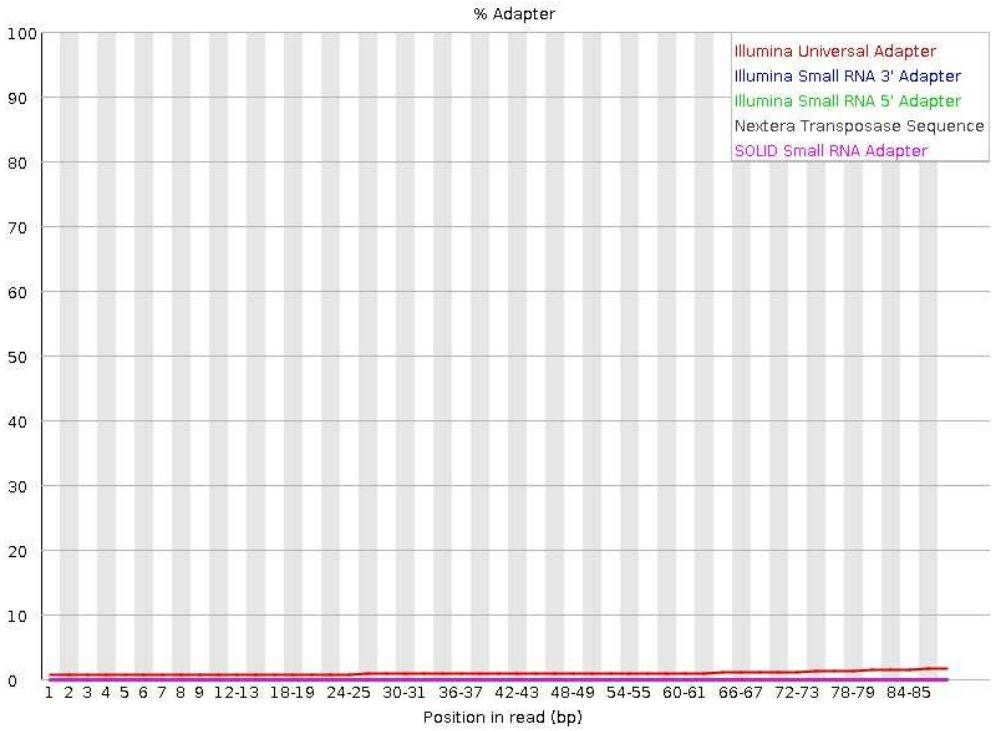
Sequence Duplication Levels



- **Overrepresented sequences**

Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACGTCTGAA CTCCAGTCACATGTCAATCTCGTATG	1359001	0.6847669881243049	TruSeq Adapter, Index 15 (97% over 49bp)

Adapter Content

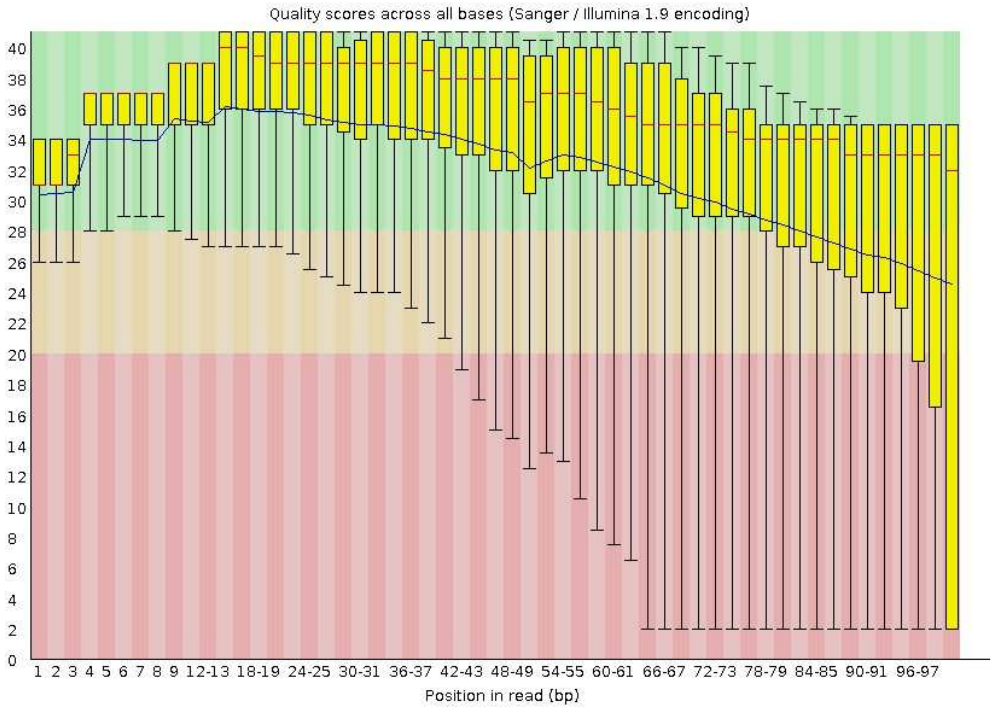

% Adapter
100
90
80
70
60
50
40
30
20
10
0
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85
Position in read (bp)

9.1.4 2kb-R2

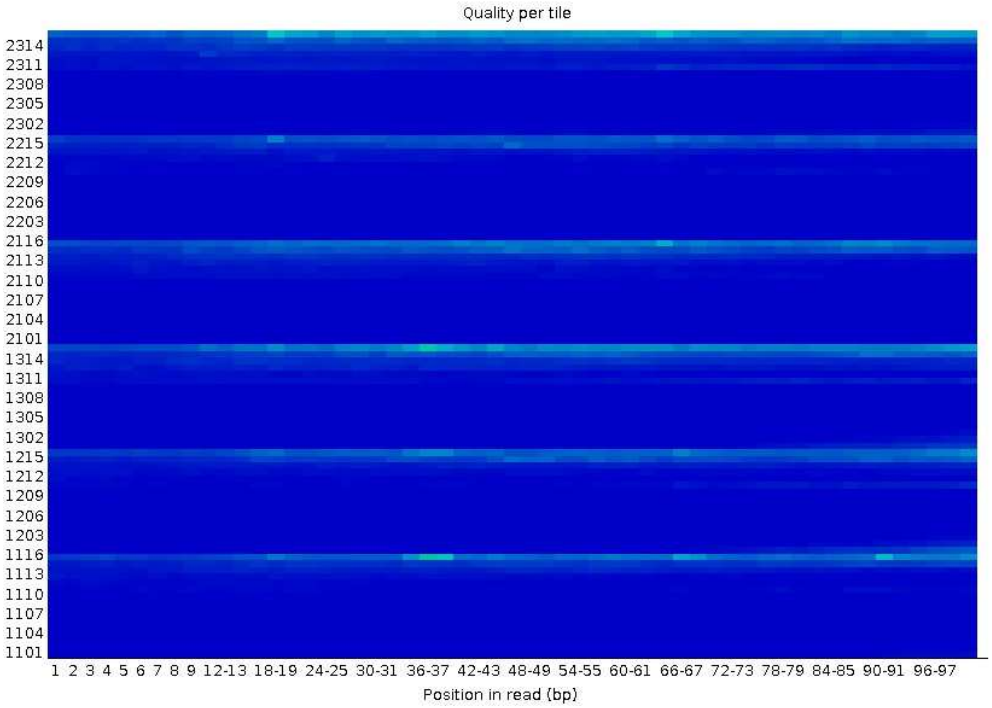
- **Basic Statistics**

Measure	Value
Filename	2kb_R2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	198461816
Sequences flagged as poor quality	0
Sequence length	100
%GC	46

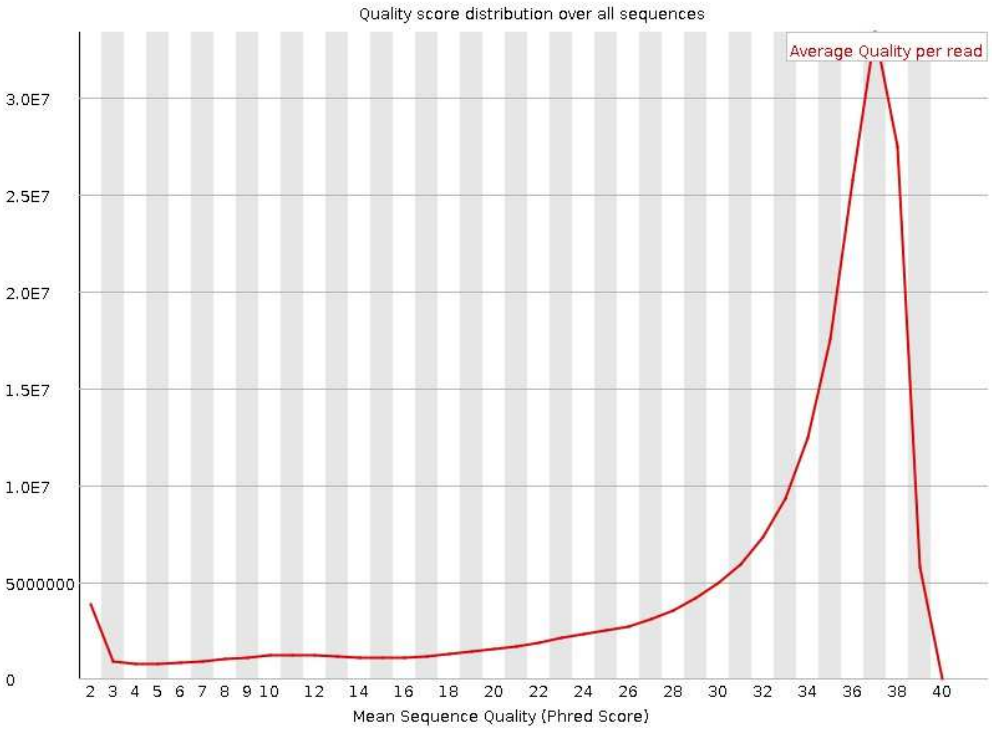
- **Per base sequence quality**



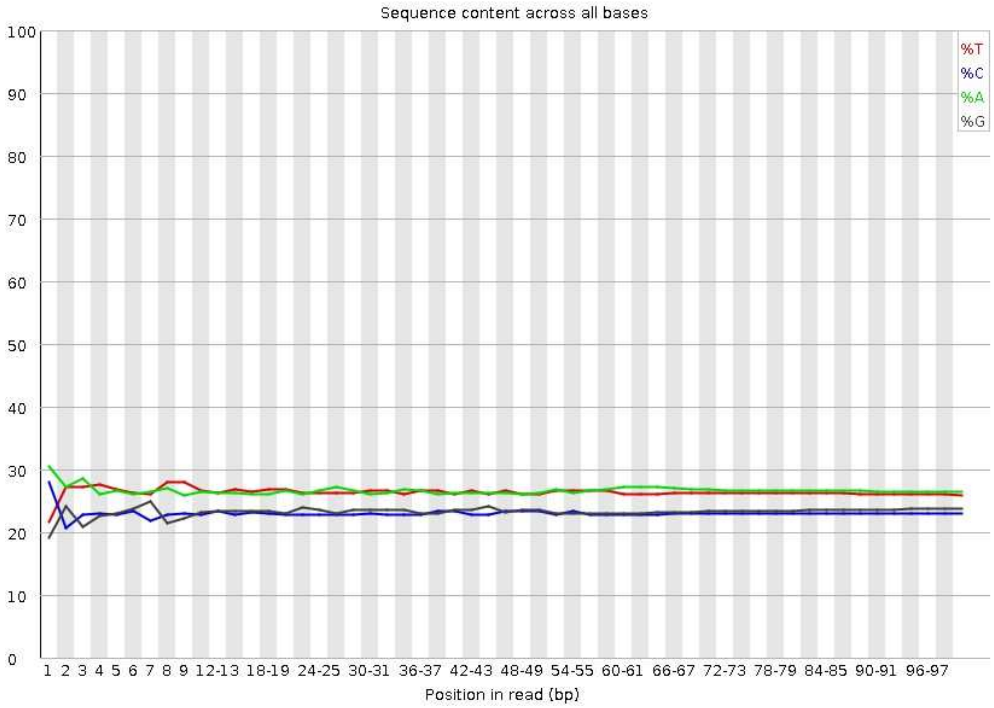
Per tile sequence quality



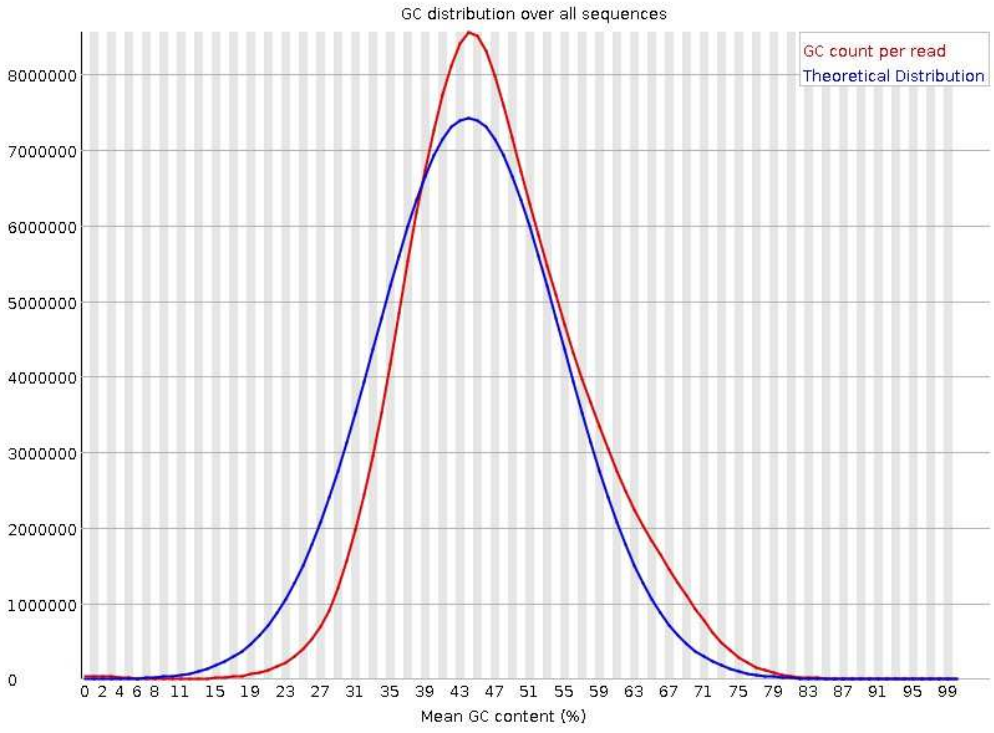
- **Per sequence quality scores**



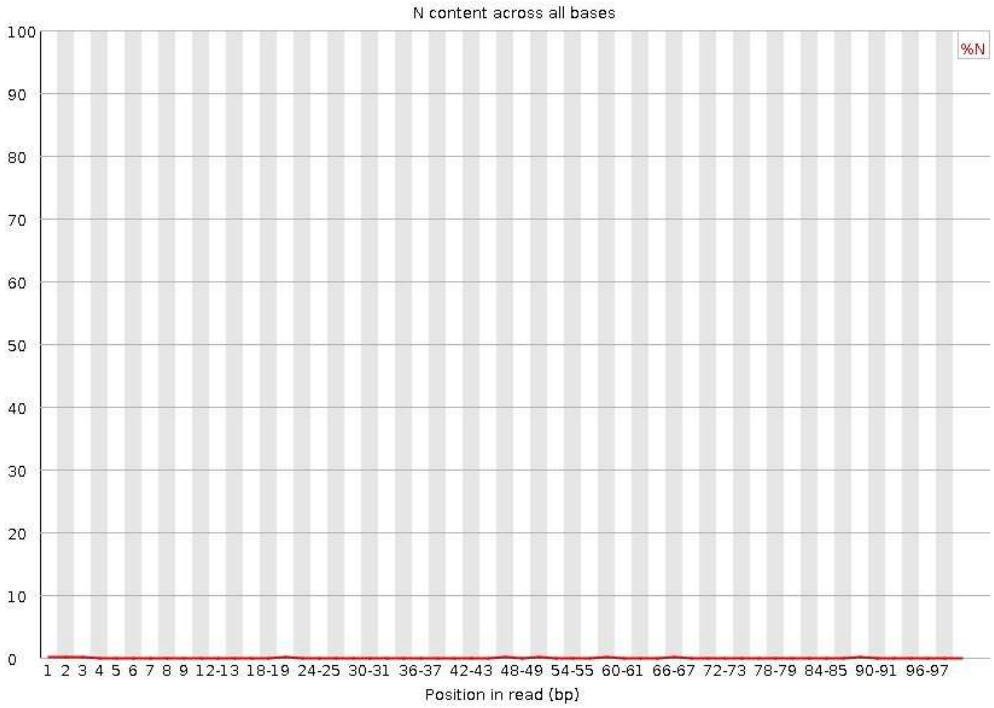
Per base sequence content



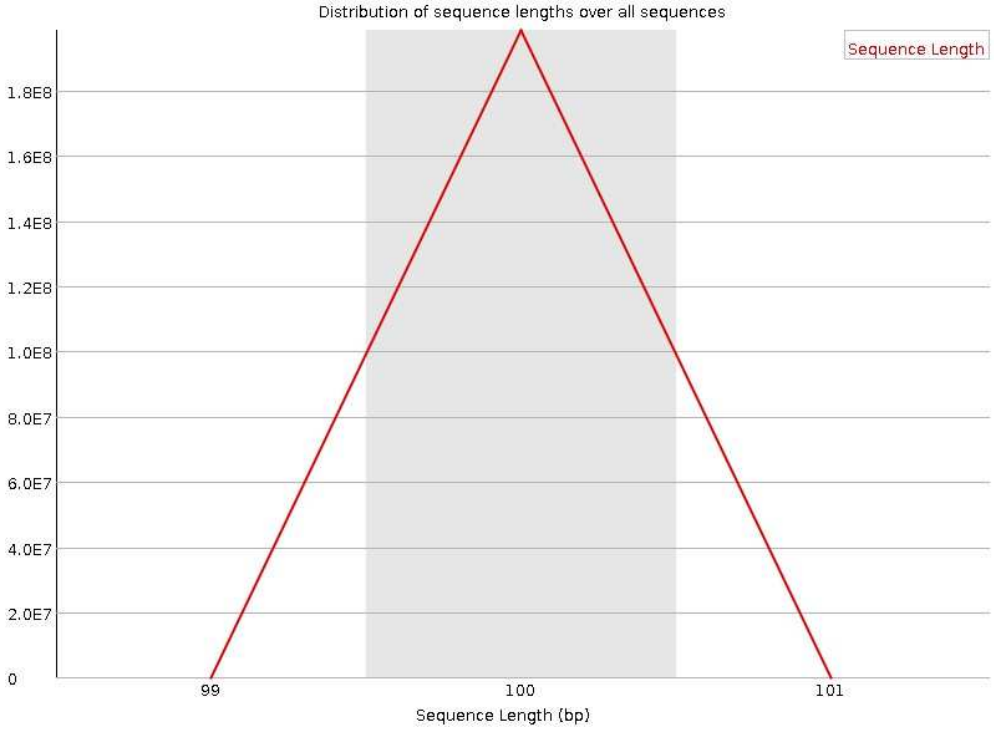
- **Per sequence GC content**



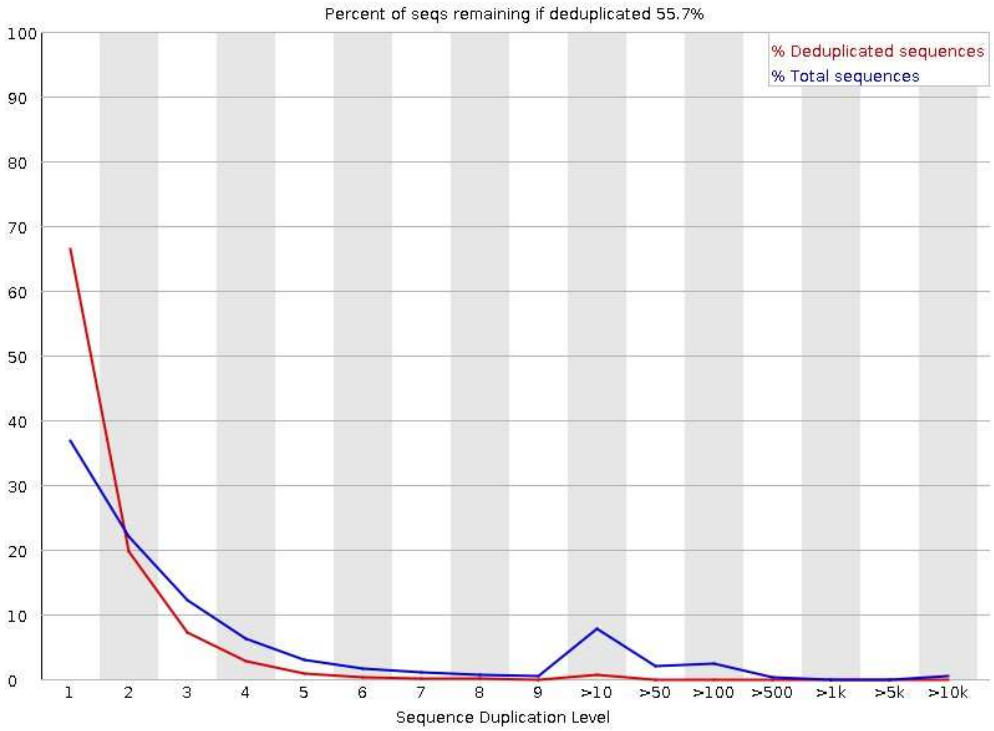
Per base N content



- **Sequence Length Distribution**



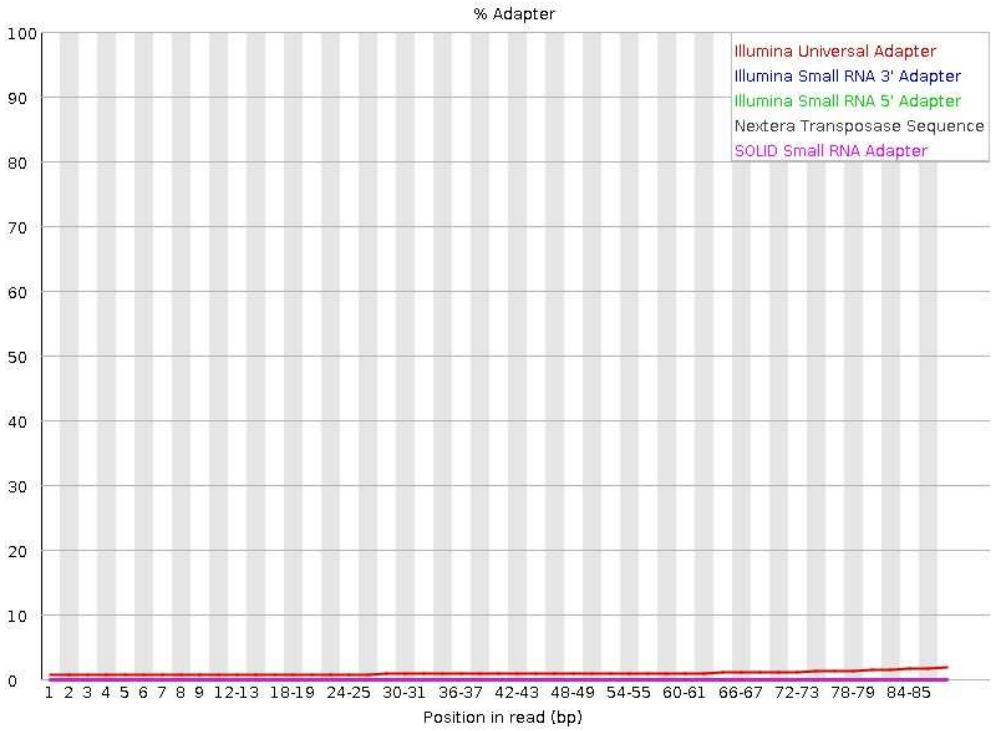
Sequence Duplication Levels



- **Overrepresented sequences**

Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCGTCGTGTAGGGA AAGAGTGTAGATCTCGGTGGTCGCC	1275223	0.6425533262277515	Illumina Single End PCR Primer 1 (100% over 50bp)

Adapter Content

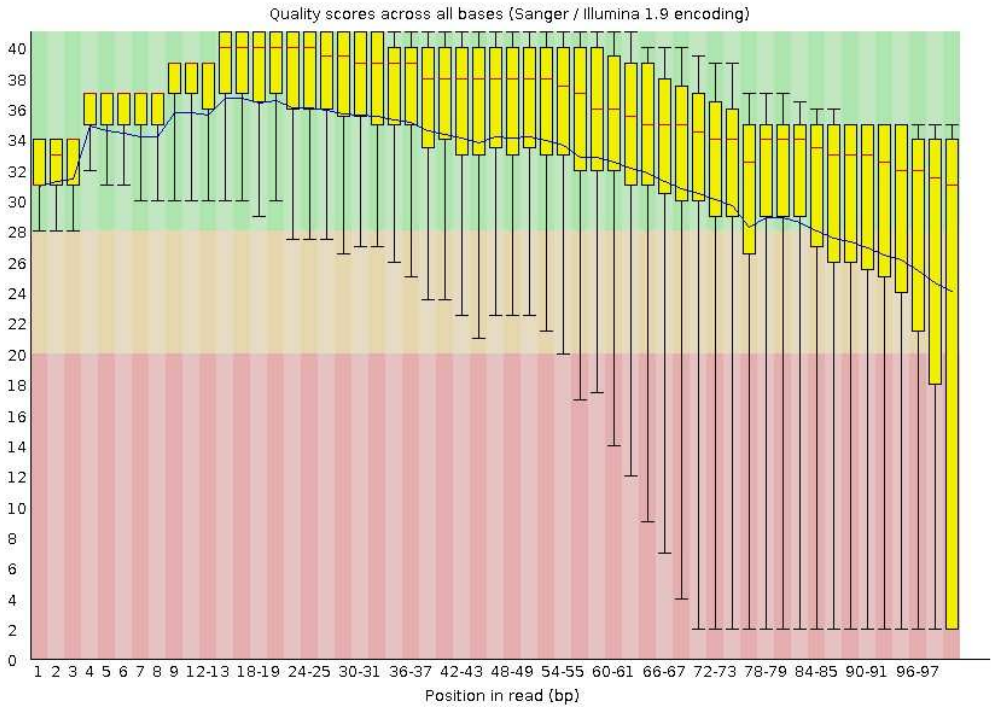

% Adapter
100
90
80
70
60
50
40
30
20
10
0
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85
Position in read (bp)

9.1.5 4kb-R1

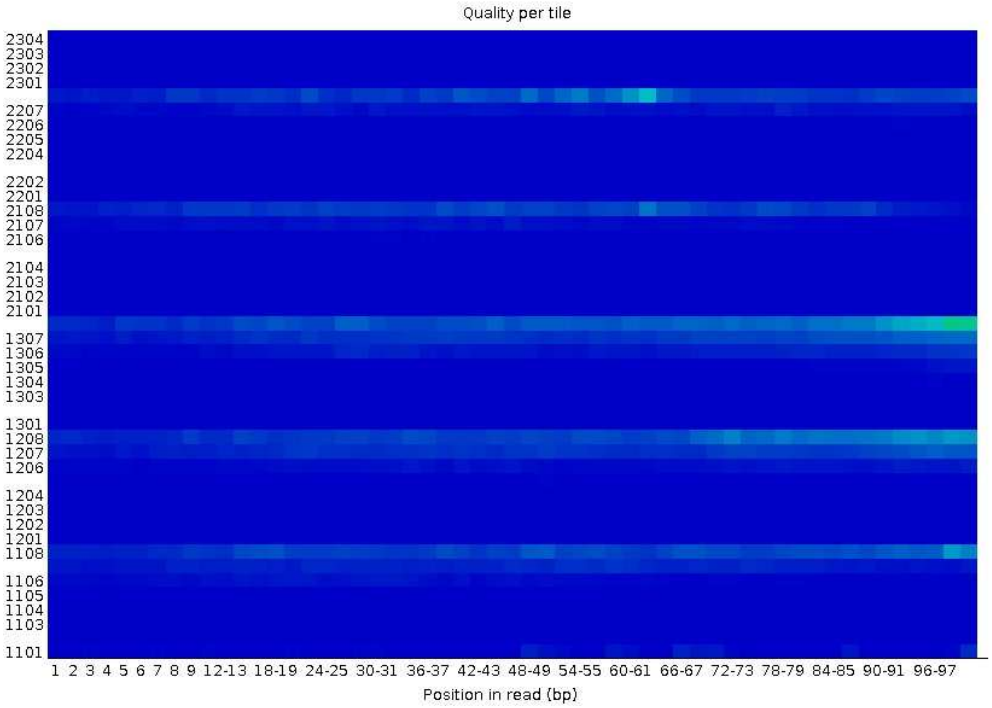
- **Basic Statistics**

Measure	Value
Filename	4kb_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	144000000
Sequences flagged as poor quality	0
Sequence length	100
%GC	46

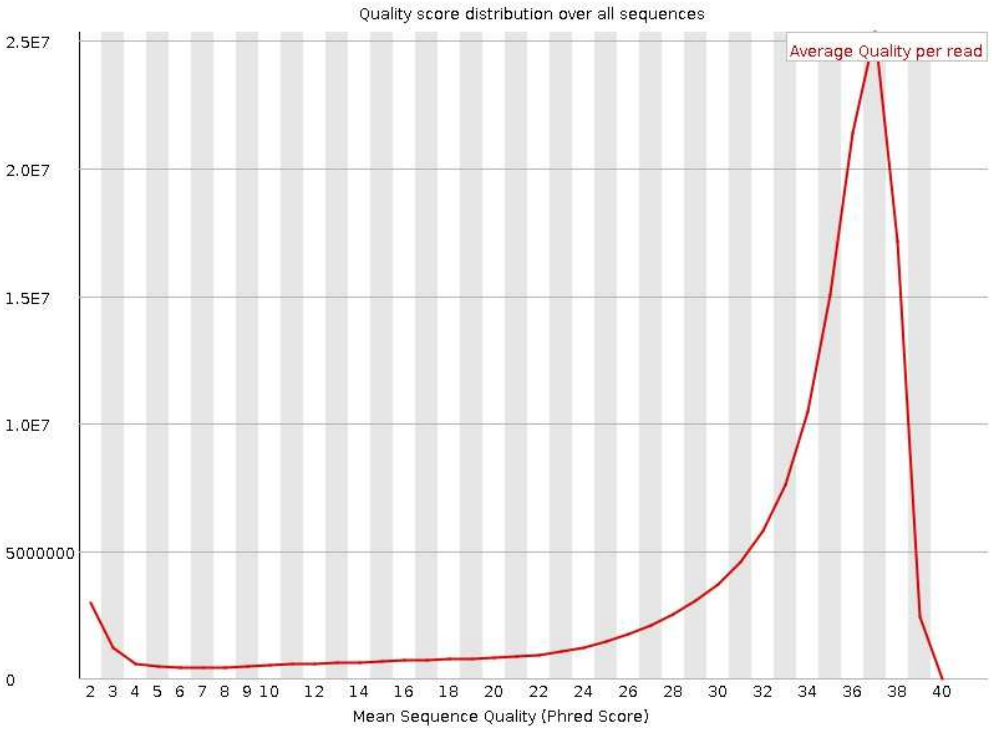
- **Per base sequence quality**



Per tile sequence quality

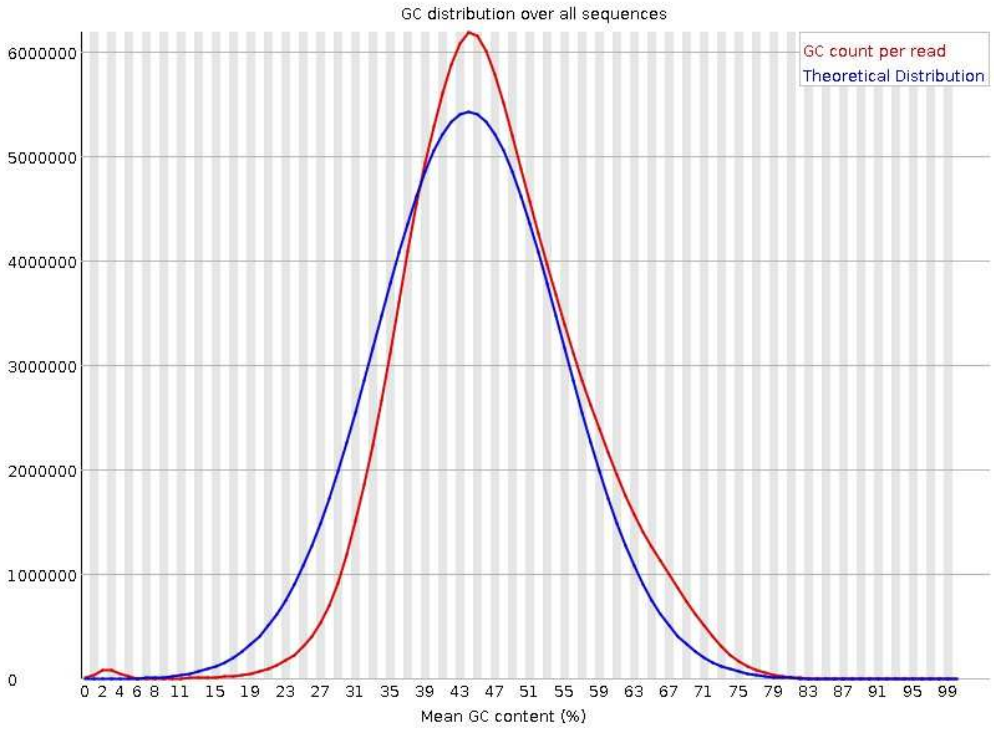


- **Per sequence quality scores**

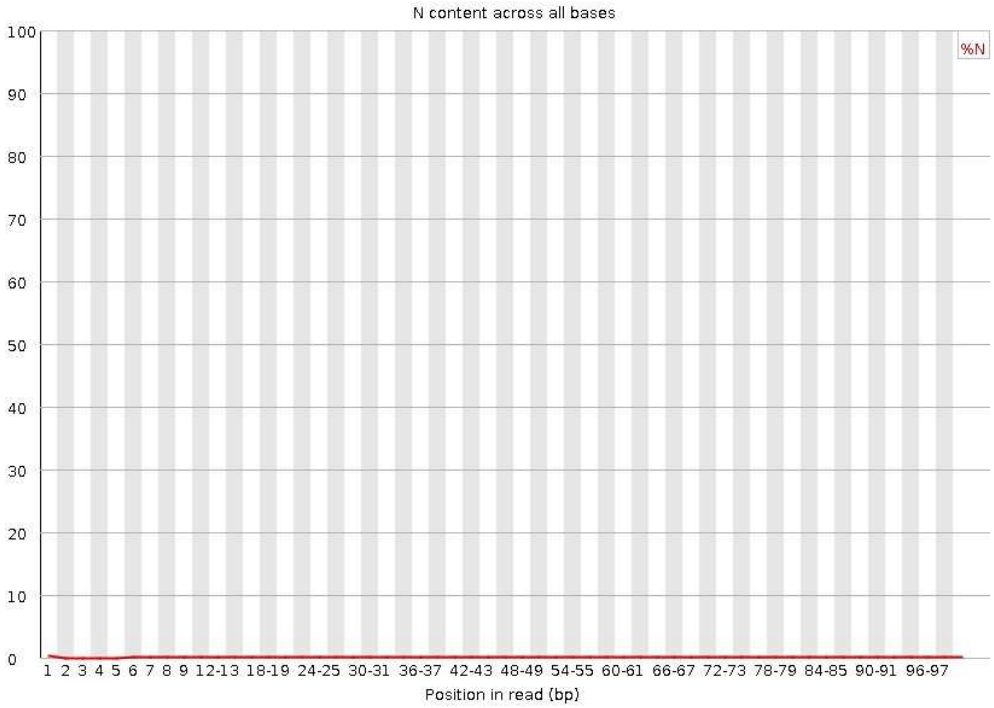


Per base sequence content

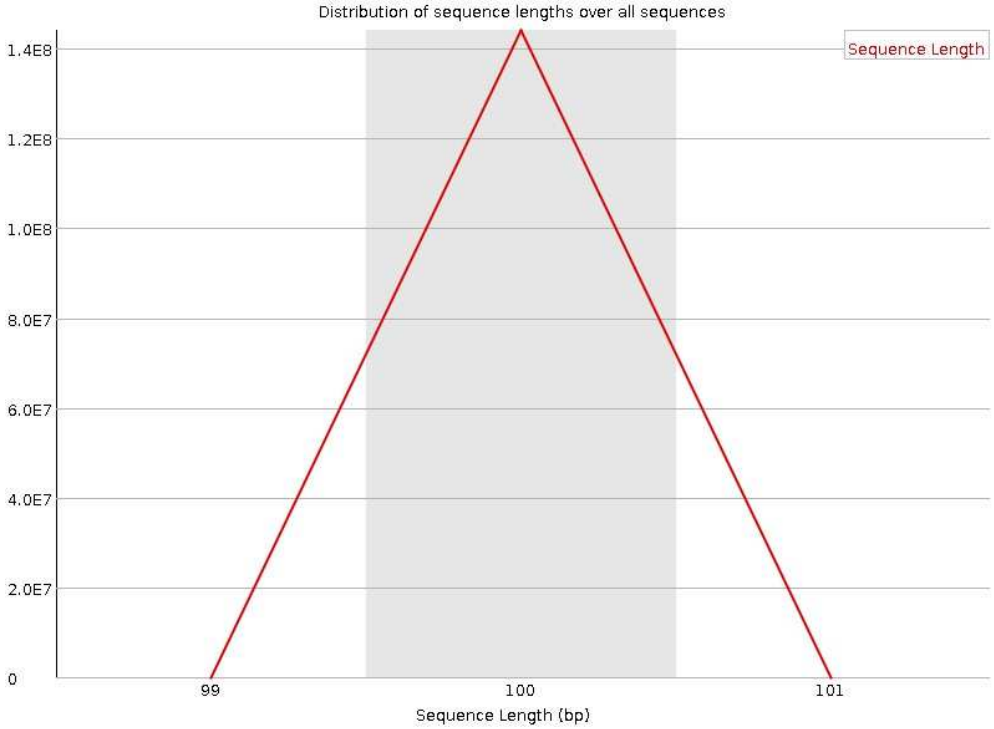
- **Per sequence GC content**



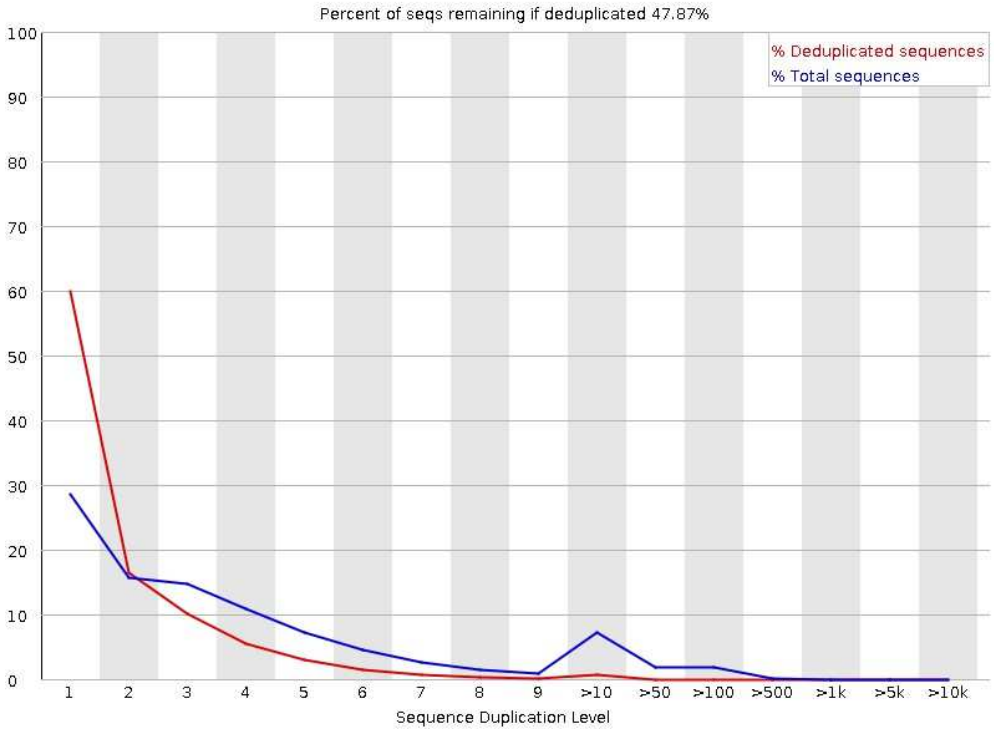
Per base N content



- **Sequence Length Distribution**



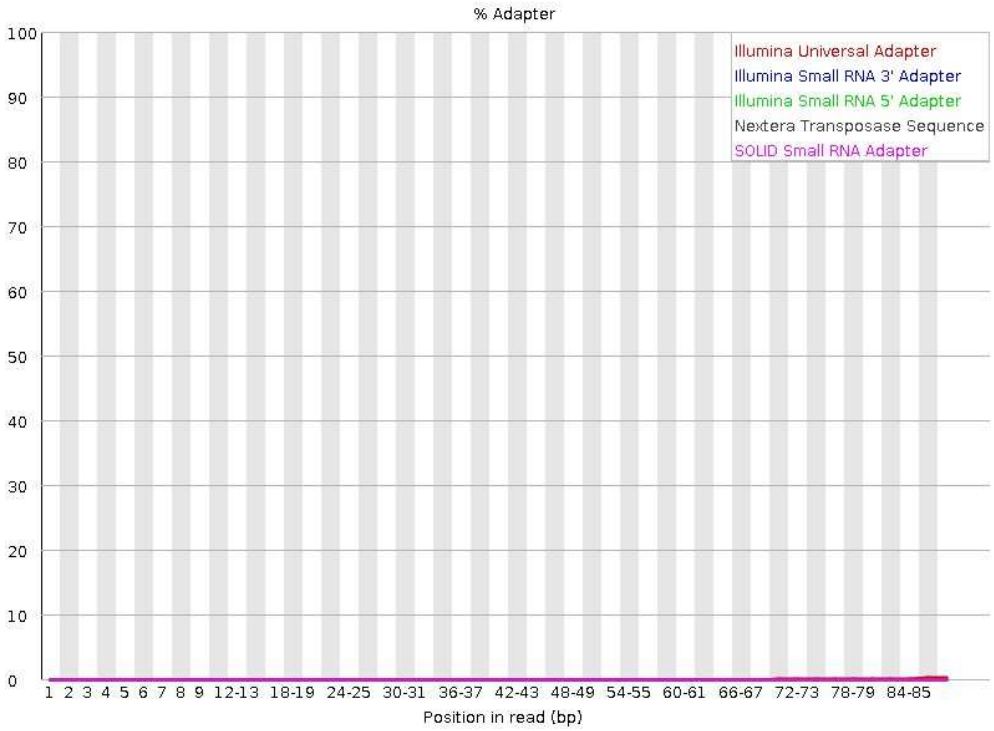
Sequence Duplication Levels



- **Overrepresented sequences**

No overrepresented sequences

- **Adapter Content**

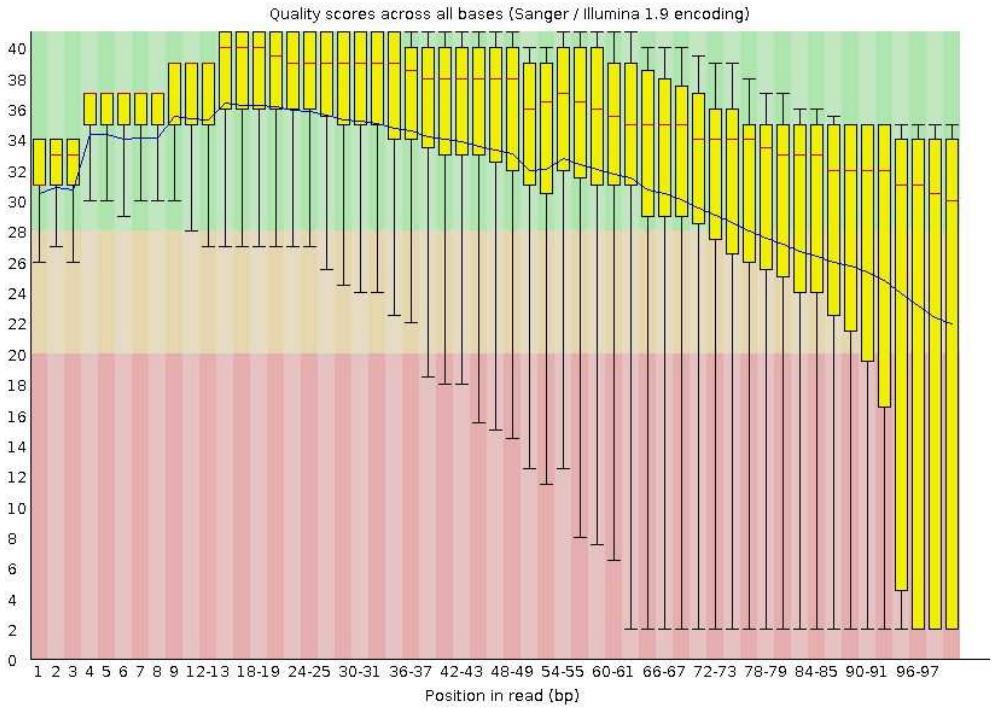

% Adapter
100
90
80
70
60
50
40
30
20
10
0
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85
Position in read (bp)

9.1.6 4kb-R2

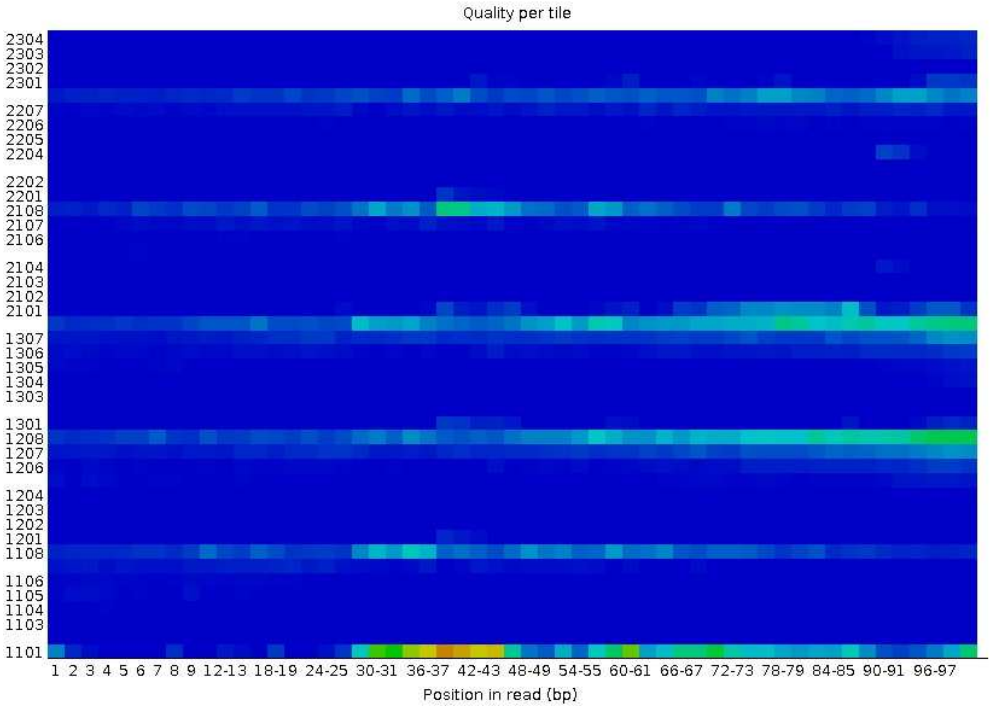
- **Basic Statistics**

Measure	Value
Filename	4kb_R2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	144000000
Sequences flagged as poor quality	0
Sequence length	100
%GC	46

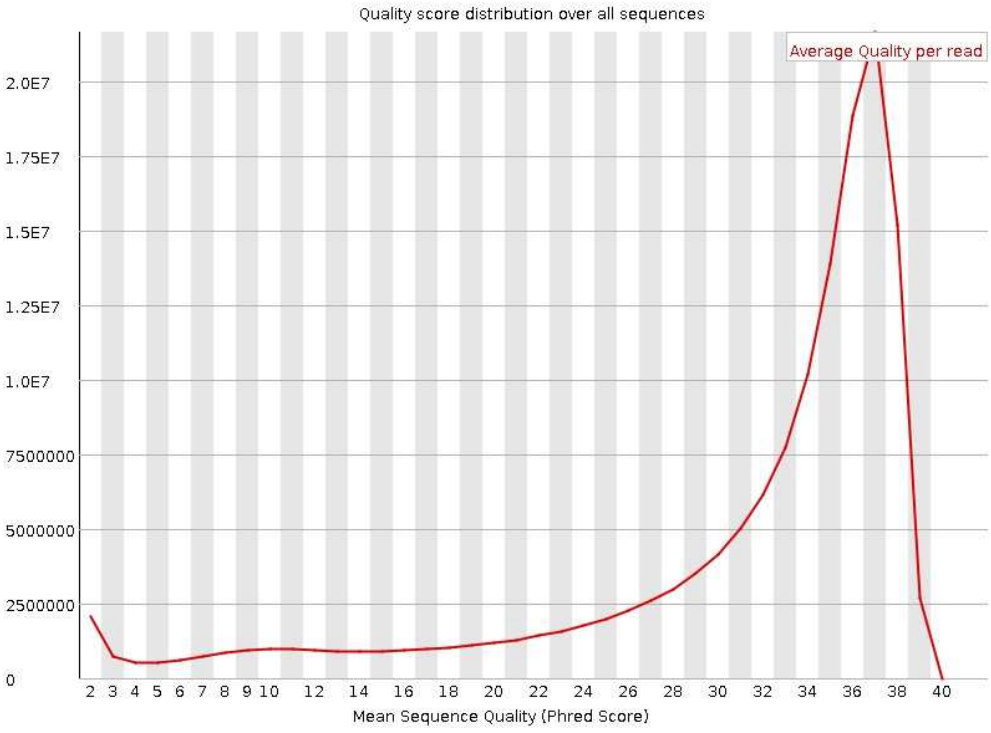
- **Per base sequence quality**



Per tile sequence quality

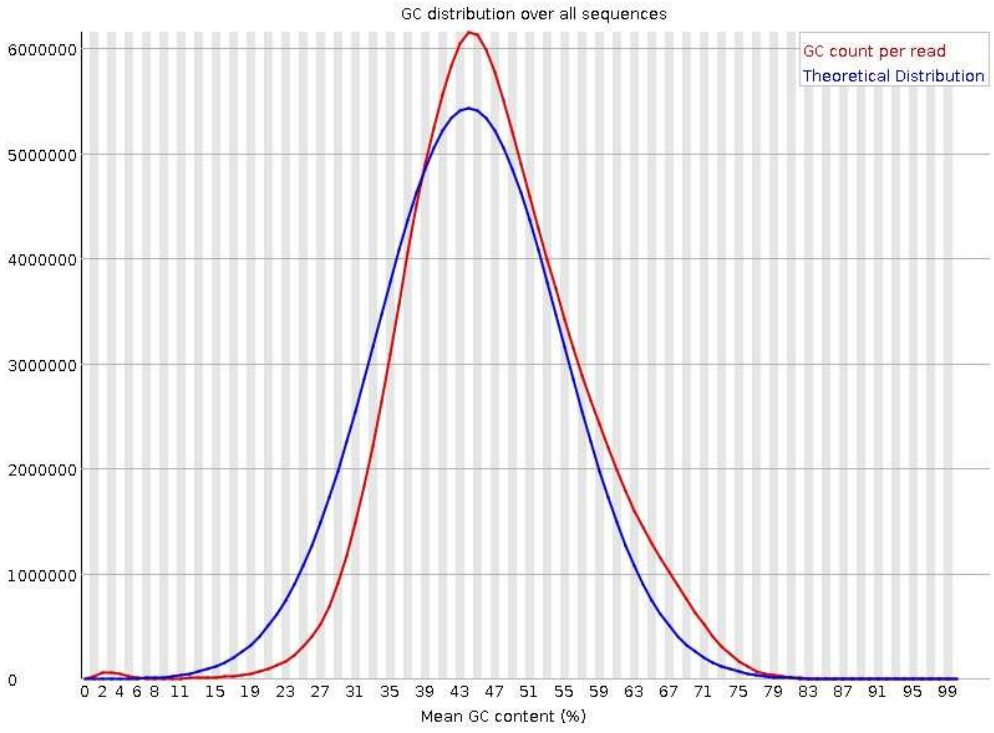


- **Per sequence quality scores**

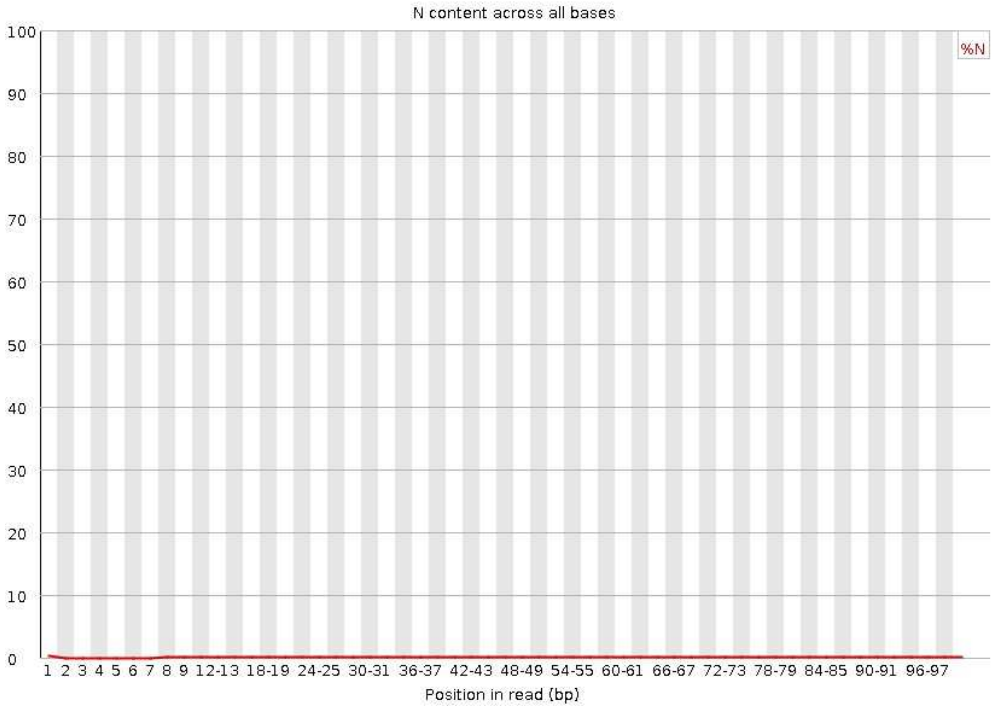


Per base sequence content

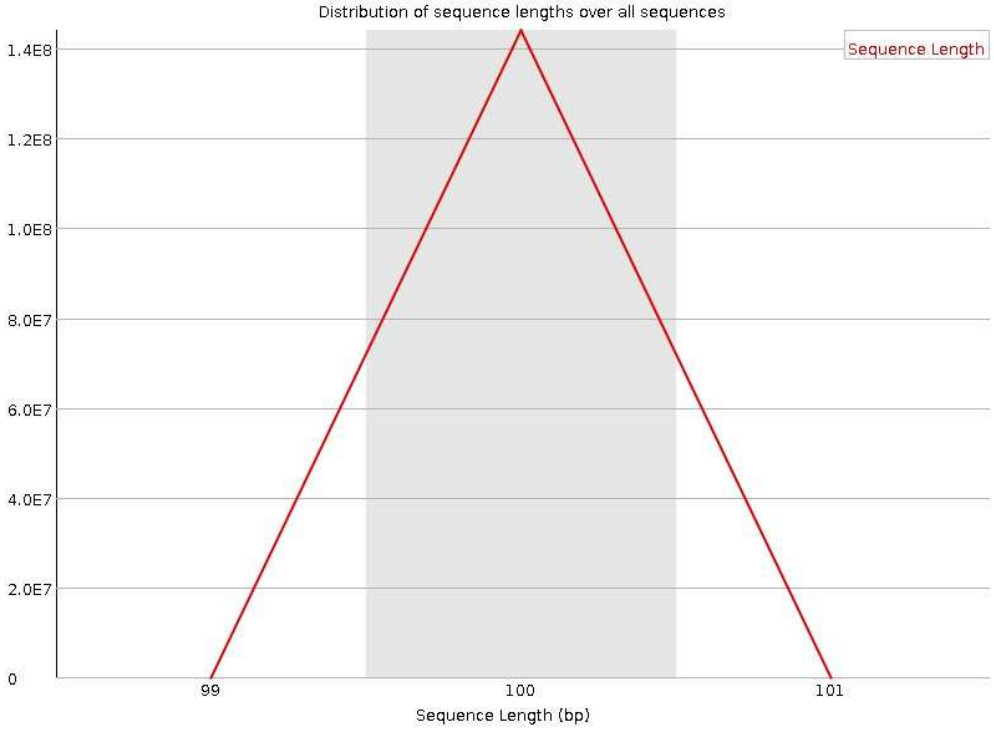
- **Per sequence GC content**



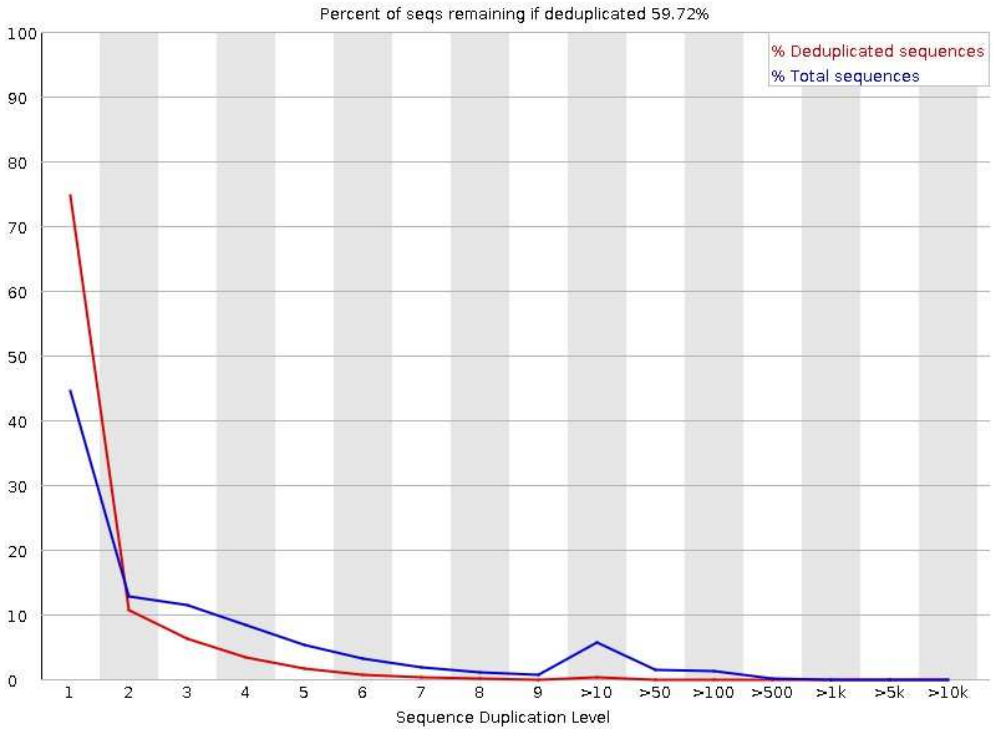
Per base N content



- **Sequence Length Distribution**



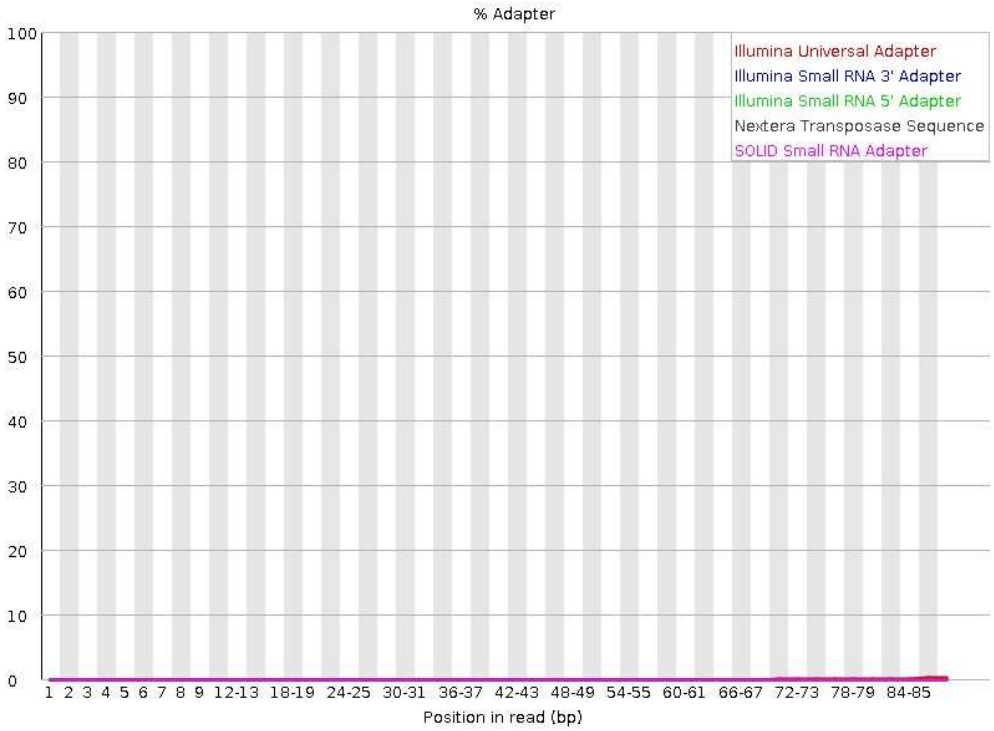
Sequence Duplication Levels



- **Overrepresented sequences**

No overrepresented sequences

- **Adapter Content**

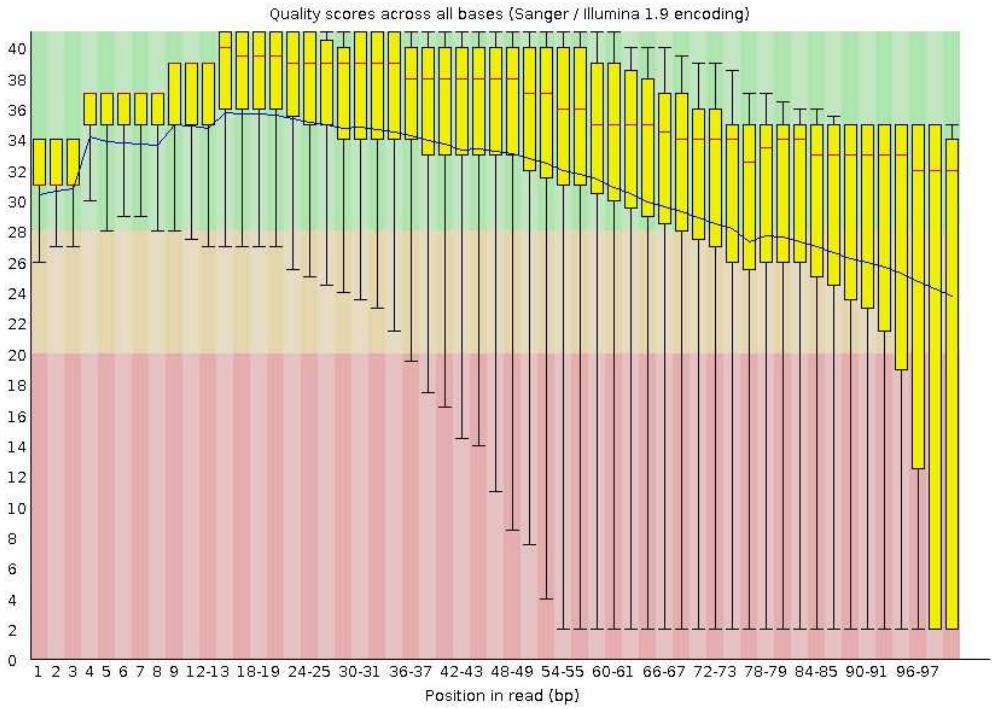

% Adapter
100
90
80
70
60
50
40
30
20
10
0
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85
Position in read (bp)

9.1.7 8kb-R1

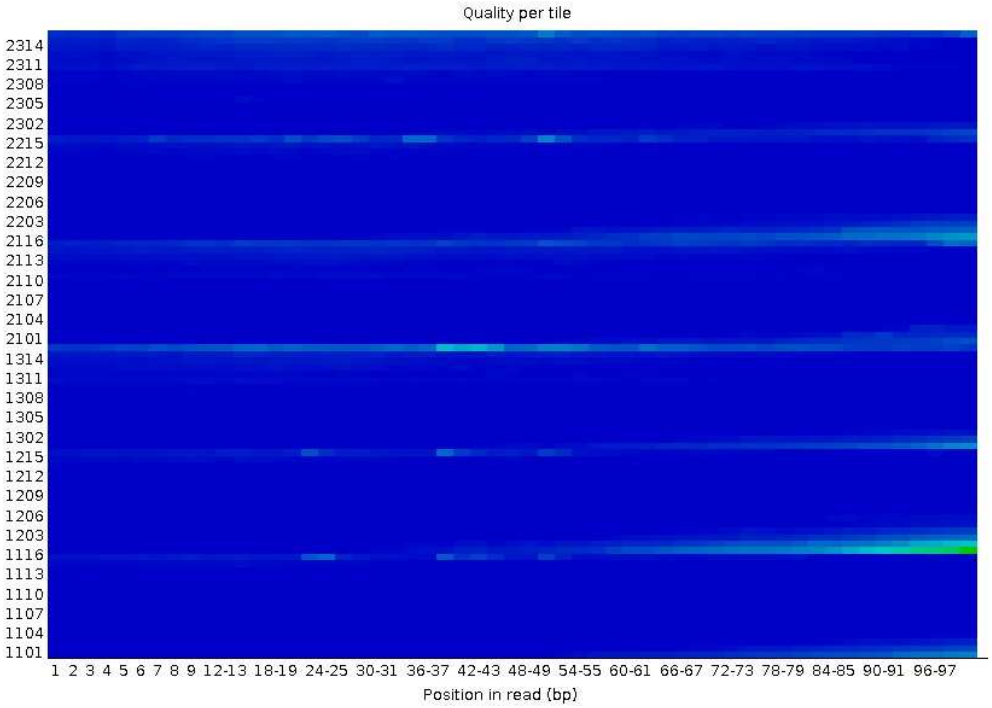
- **Basic Statistics**

Measure	Value
Filename	8kb_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	213300808
Sequences flagged as poor quality	0
Sequence length	100
%GC	48

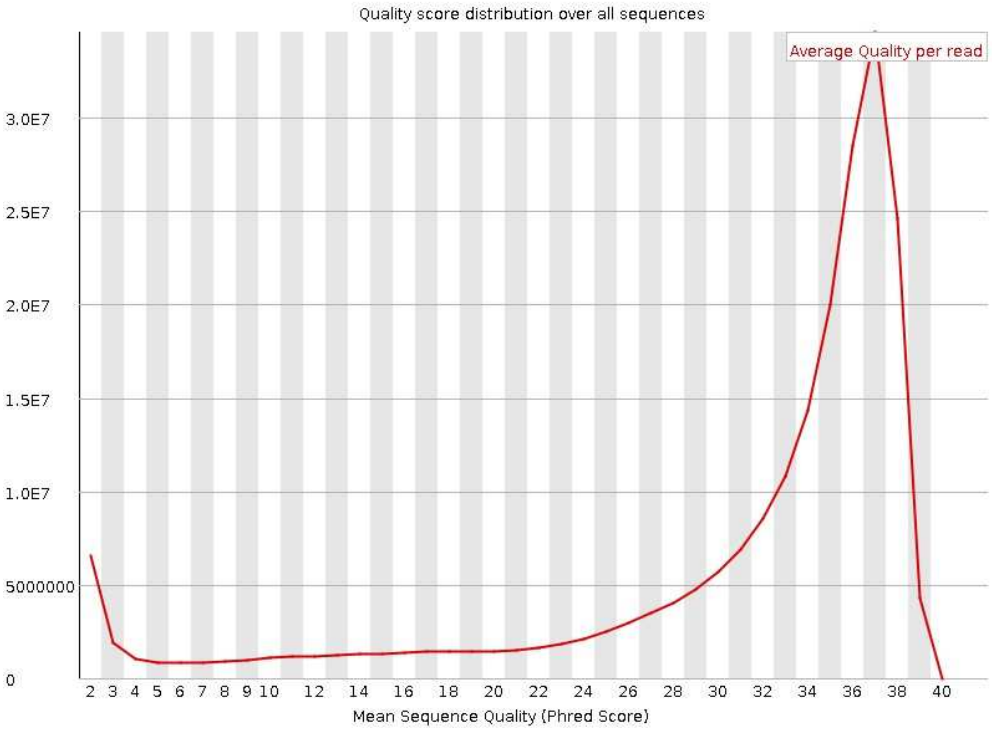
- **Per base sequence quality**



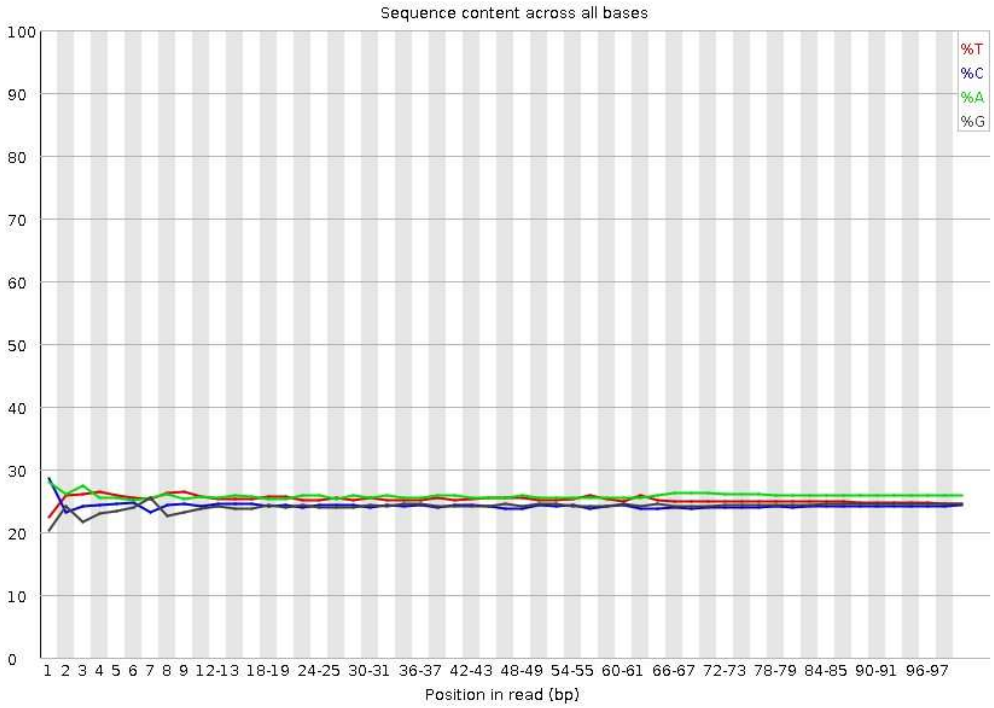
Per tile sequence quality



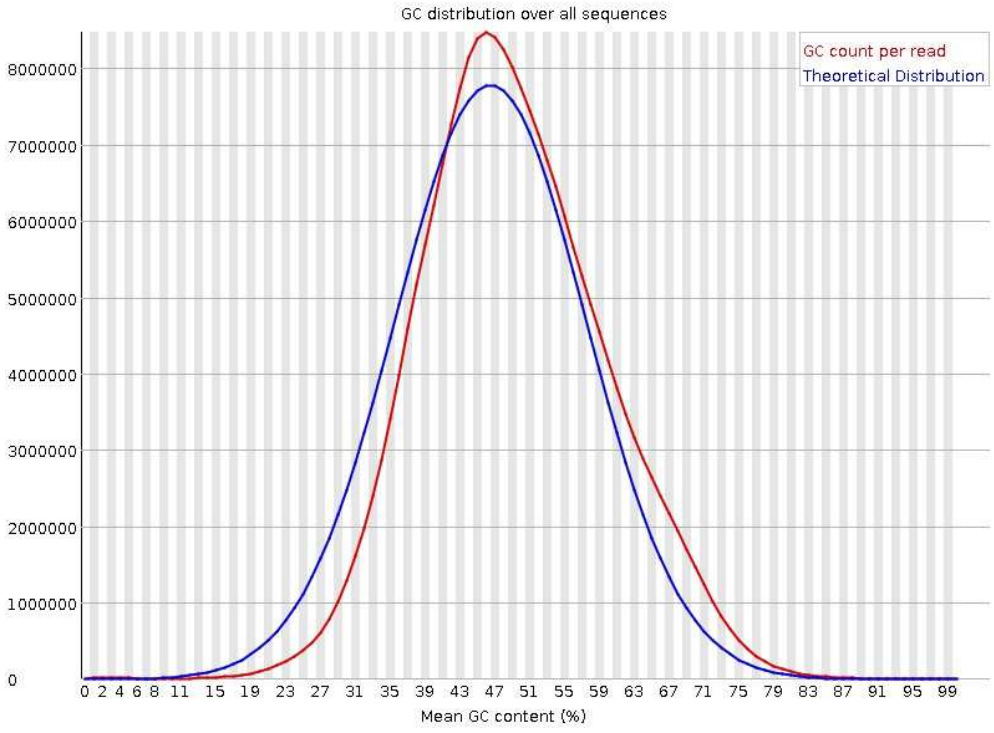
- **Per sequence quality scores**



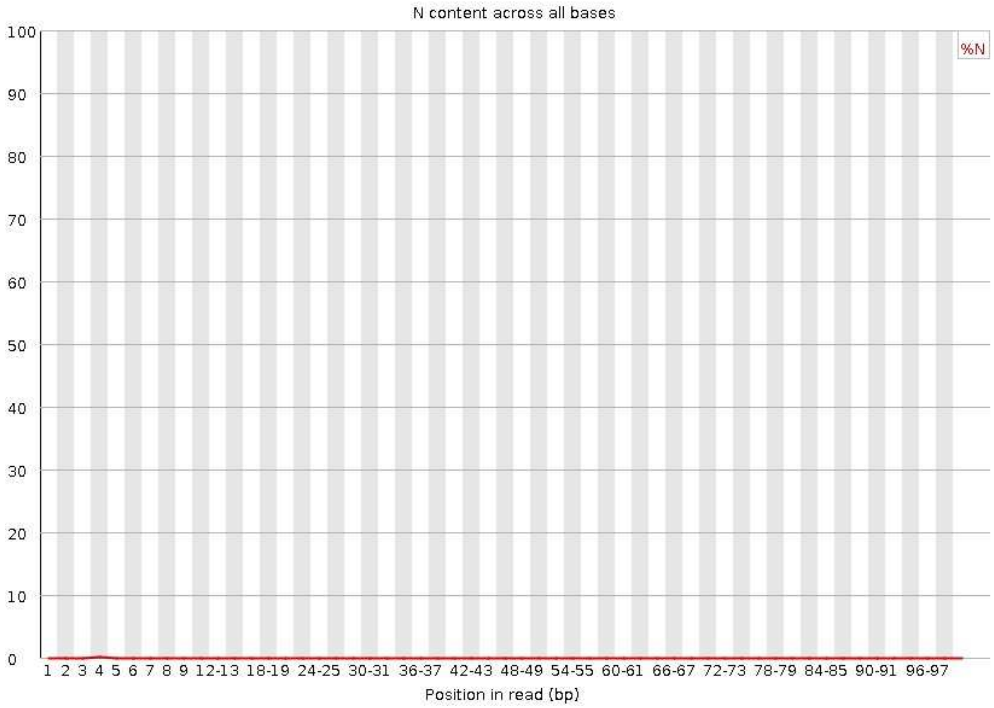
Per base sequence content



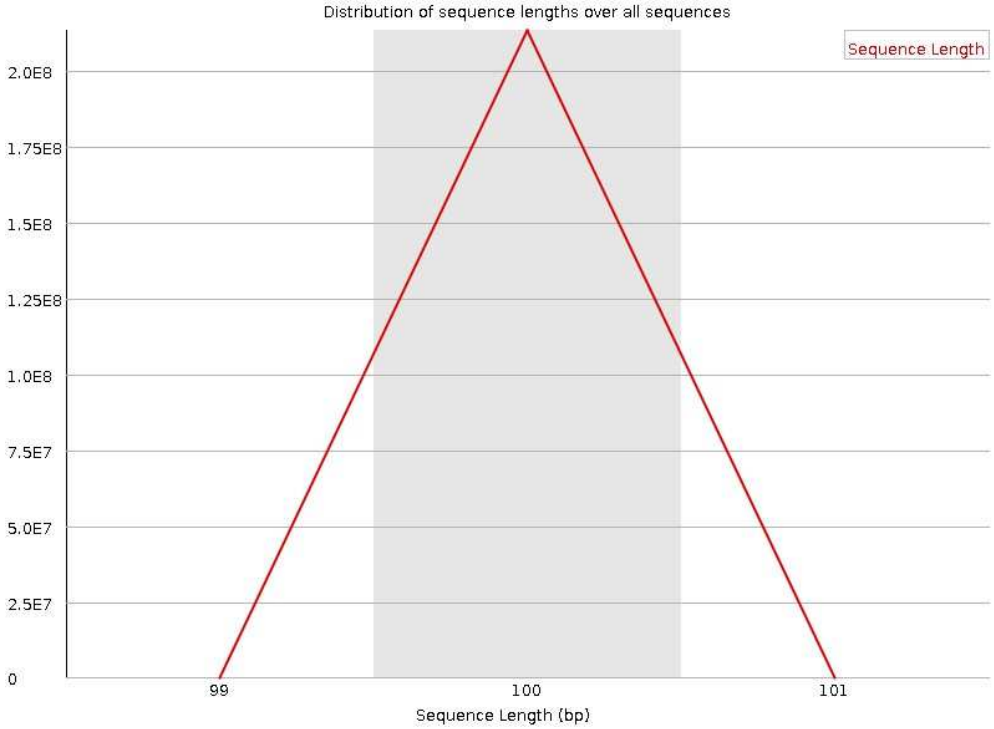
- **Per sequence GC content**



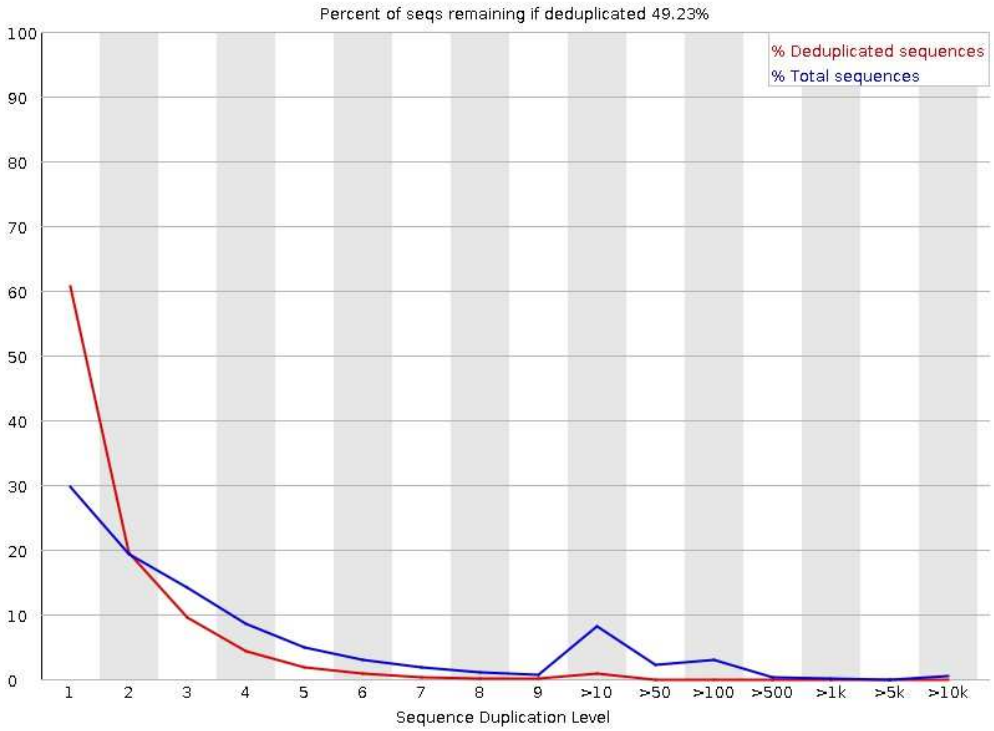
Per base N content



- **Sequence Length Distribution**



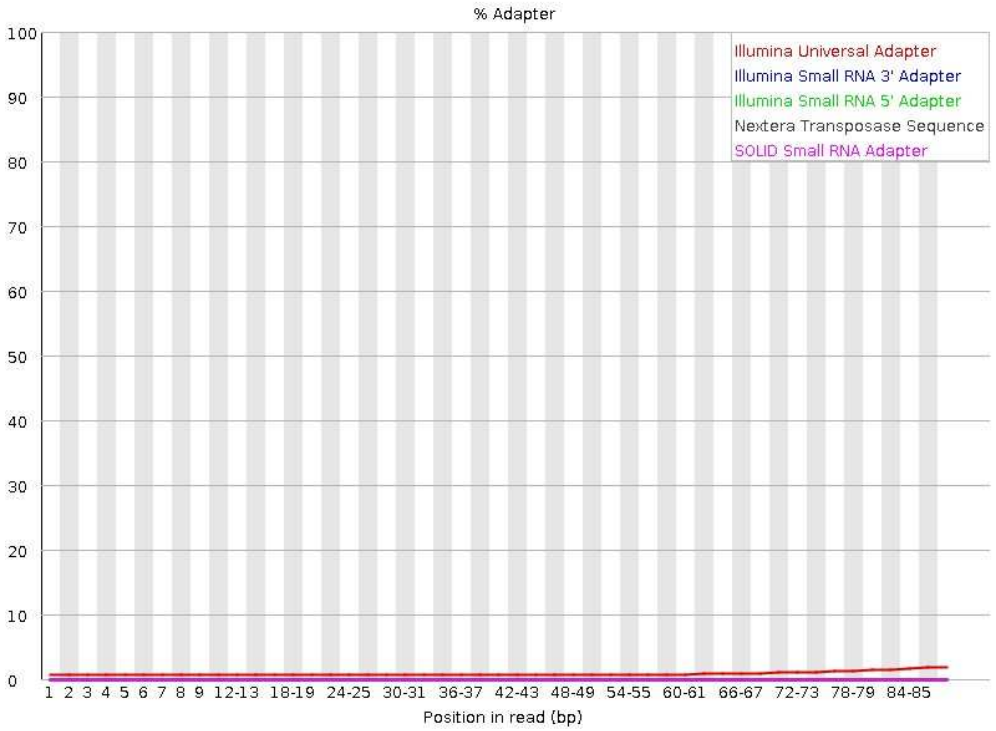
Sequence Duplication Levels



- **Overrepresented sequences**

Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACGTCTGAAC TCCAGTCACGGCTACATCTCGTATG	1108393	0.5196384441262876	TruSeq Adapter, Index 11 (100% over 49bp)

Adapter Content

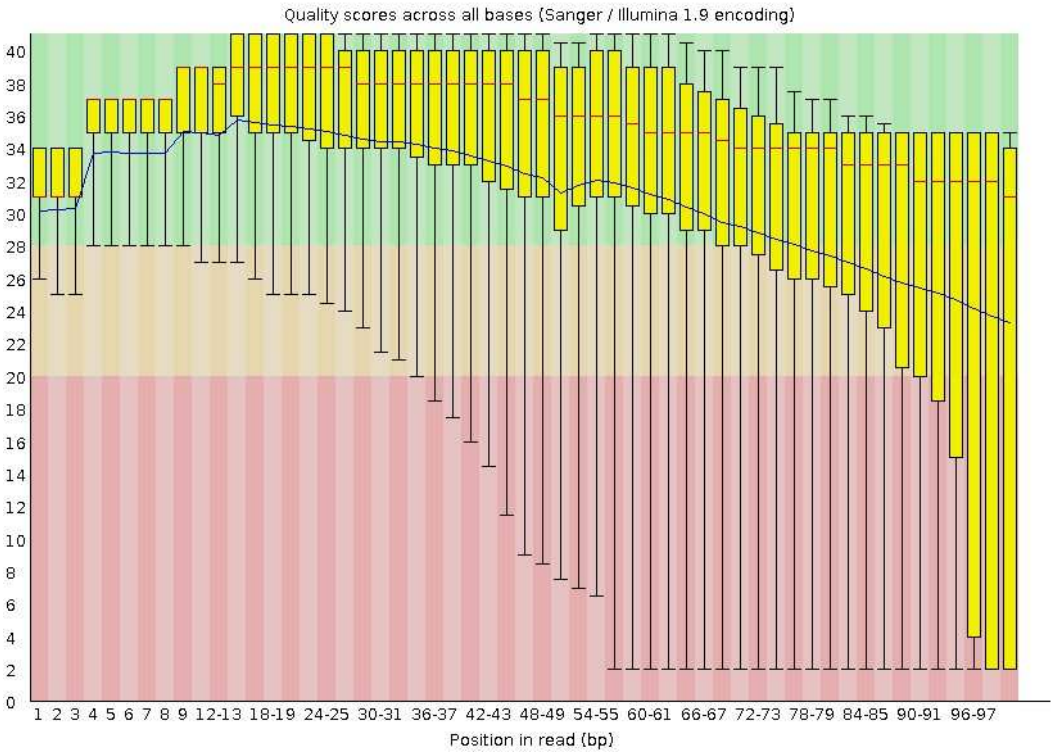

% Adapter
100
90
80
70
60
50
40
30
20
10
0
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85
Position in read (bp)

9.1.8 8kb-R2

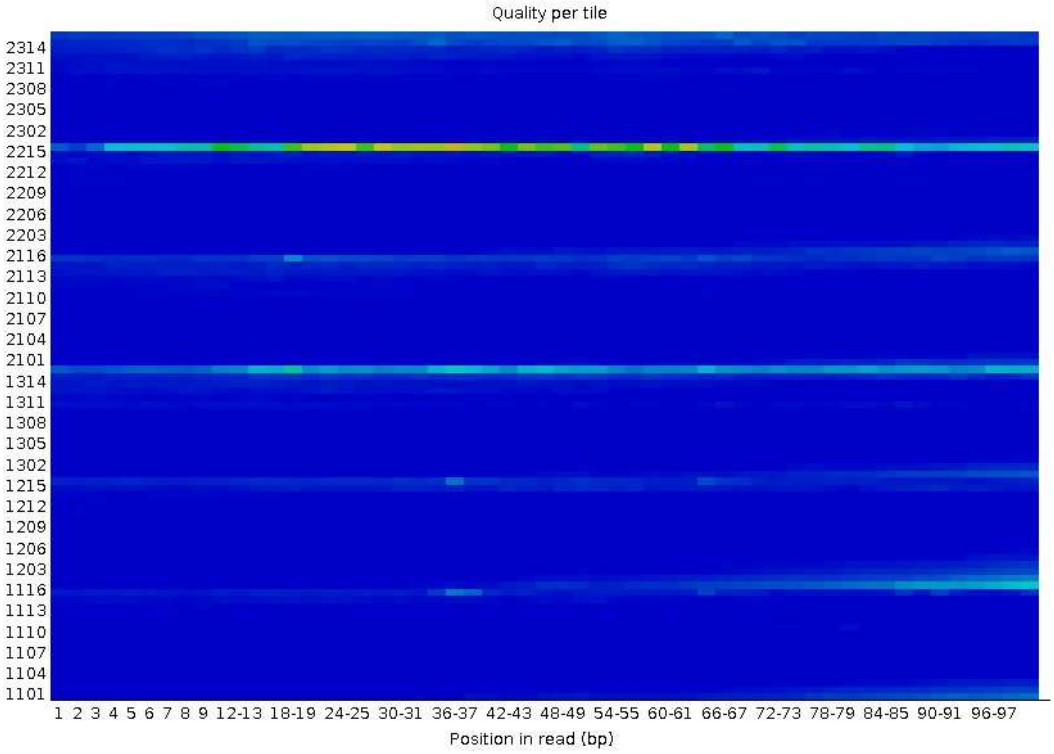
- **Basic Statistics**

Measure	Value
Filename	8kb_R2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	213300808
Sequences flagged as poor quality	0
Sequence length	100
%GC	48

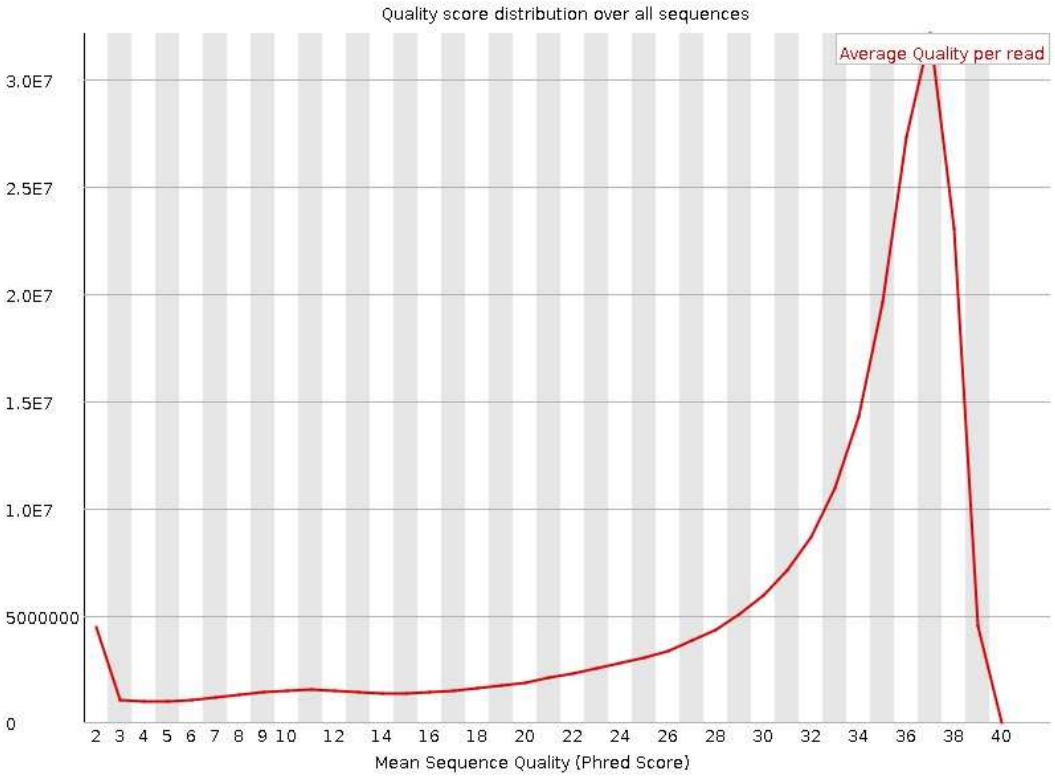
- **Per base sequence quality**



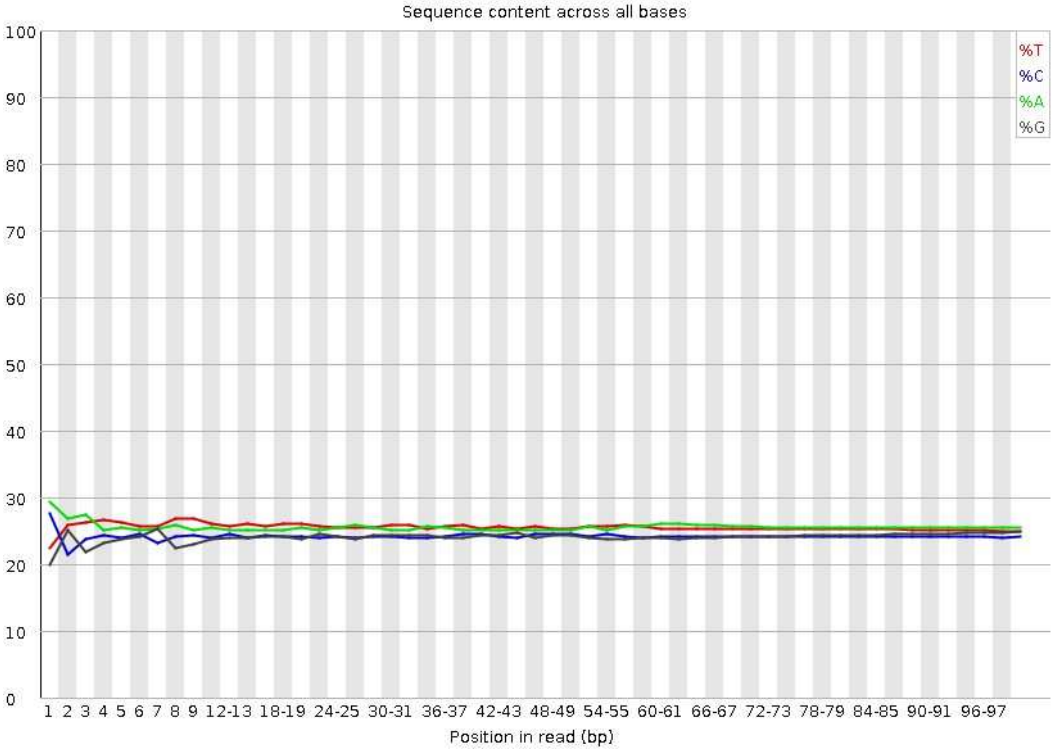
Per tile sequence quality



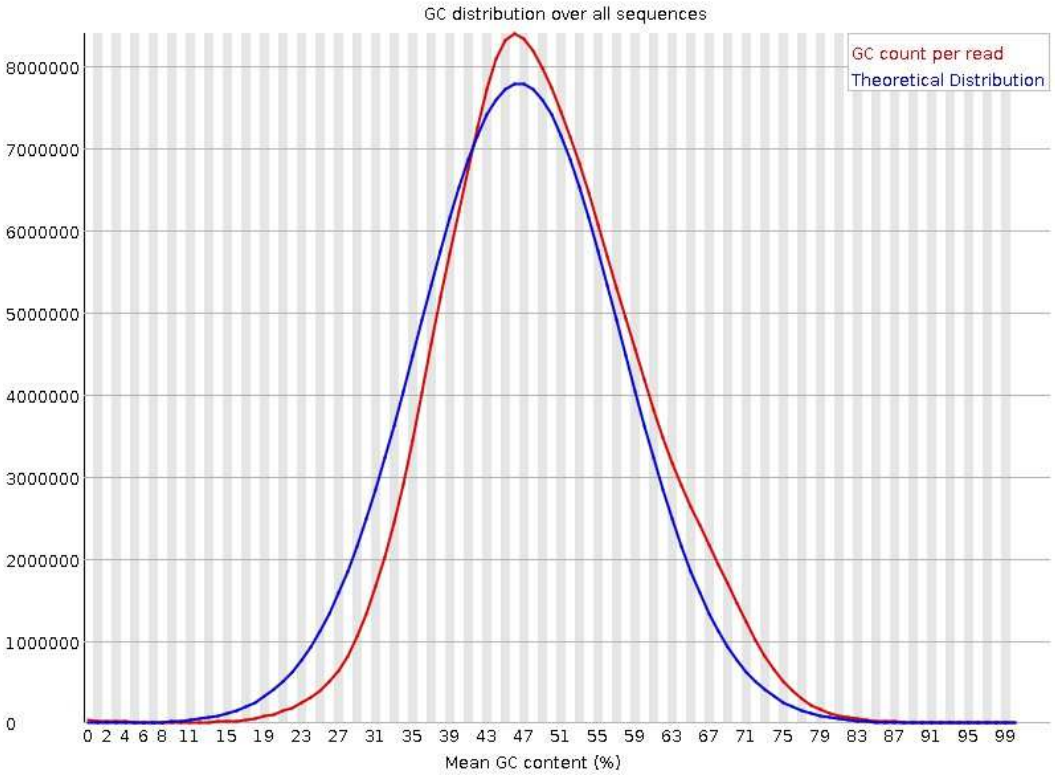
- **Per sequence quality scores**



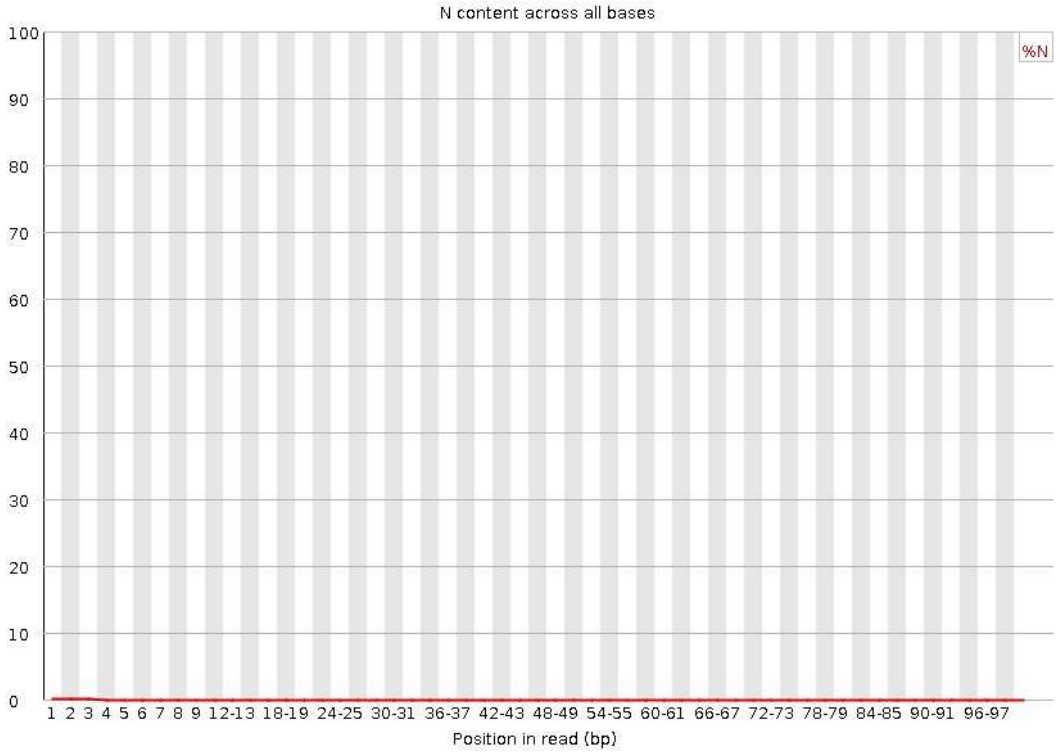
Per base sequence content



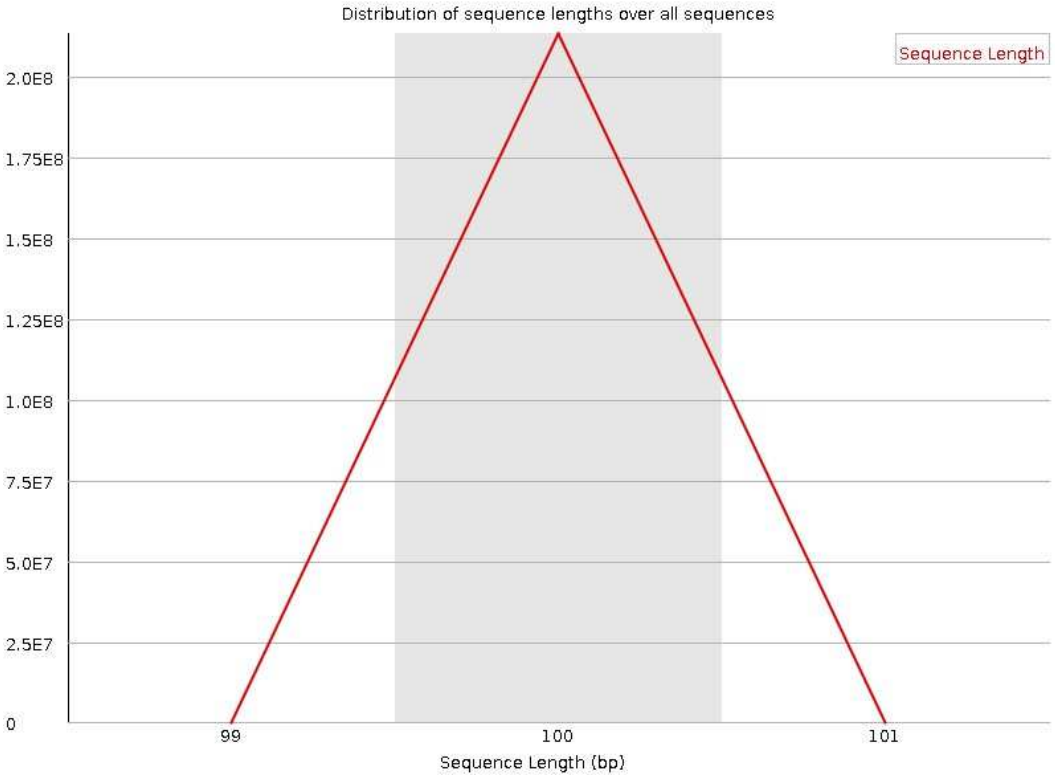
- **Per sequence GC content**



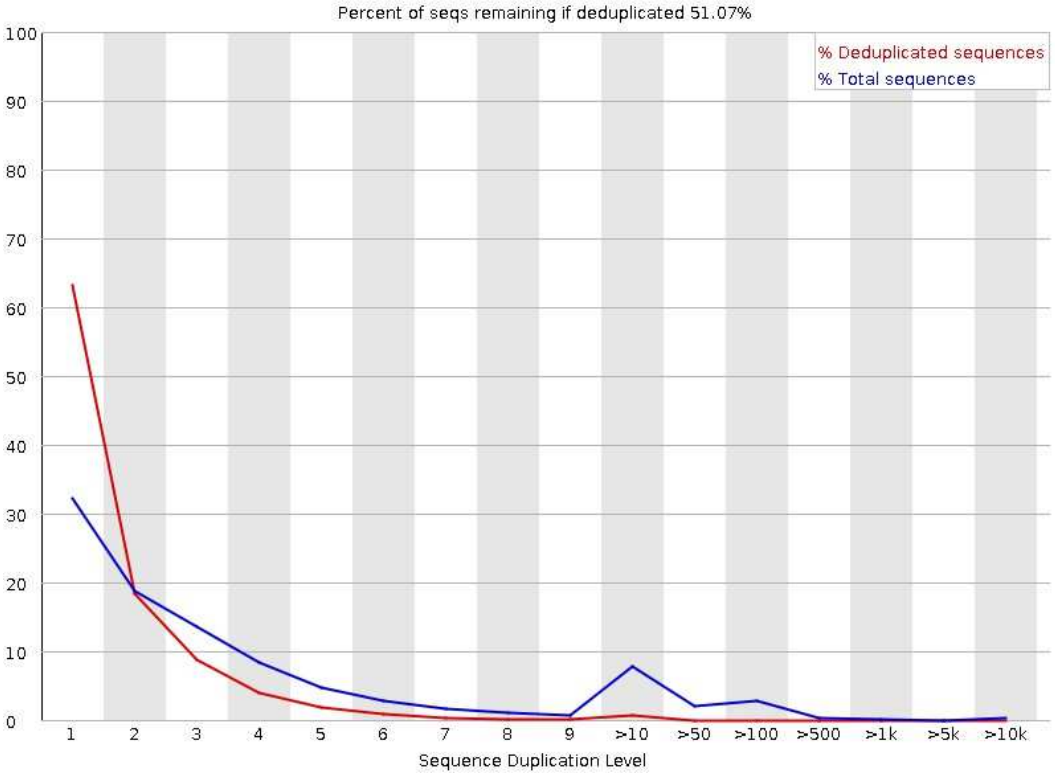
Per base N content



- **Sequence Length Distribution**



Sequence Duplication Levels



- **Overrepresented sequences**

Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCGTCGTGTAGGGA AAGAGTGTAGATCTCGGTGGTCGCC	1013649	0.4752204220435958	Illumina Single End PCR Primer 1 (100% over 50bp)

Adapter Content